行政院國家科學委員會補助專題研究計畫■ 成 果 報 告
□期中進度報告

室內用寬頻無線資訊接取系統
Broadband Wireless Information Access for e-Building

計畫類別：□ 個別型計畫　■ 整合型計畫
計畫編號：NSC－93－2213－E－002－001
執行期間：　93 年　1 月　1 日至　93 年　12 月　31 日

計畫主持人：汪重光
共同主持人：闕志達、吳安宇、周世傑、蘇朝琴、廖婉君、吳曉光
計畫參與人員：張景祺、侯鈞鏘、丁姿吟、雷貽晴、沈文中、游志強、曹亞嵐、黃伯翔、魏庭楨、陳筱筠、陳鴻愷、劉得煌、陳一強、江致和、陳俊達、張中原、楊惟仁

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
■出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立臺灣大學電機工程學系暨研究所

中　華　民　國　94 年　3 月　29 日

目錄

附錄一：論文投稿成果發表：

子計畫一：

- ✧ Chun-Chih Hou, Ching-Chi Chang, and Chorng-Kuang Wang, "*A Dual-Band IEEE802.11a/b/g Receiver Front-End Using Half-IF and Dual Conversion,*" AP-ASIC 2004, Fukuoka, Japan, Aug., 2004.
- ✧ Tsung-Chieh Pao, Ching-Chi Chang, and Chorng-Kuang Wang, "*A Variable-Length DHT-Based FFT/IFFT Processor for VDSL/ADSL Systems,*" APCCAS 2004, Taiwan, R.O.C., Dec., 2004.
- ✧ Wei-Hsiang Tseng, Ching-Chi Chang, and Chorng-Kuang Wang," *IEEE 802.11a WLAN Transceiver Using Self-Correcting Digital Timing Recovery Design And Implementation,*" VLSI Design/CAD Symposium 2004, Pingtong, Taiwan, R. O. C.
- ✧ Chun-Chih Hou, Ching-Chi Chang, and Chorng-Kuang Wang, "*Inductor-Aided Gain, Noise Figure, and IIP3 Enhancement Techniques for Active Mixer and the Application in Dual-Band Receiver,*" VLSI Design/CAD Symposium 2004, Pingtong, Taiwan, R. O. C.

子計畫二：

- ✧ C. J. Jan, P. H. Lin, and T. D. Chiueh, "*An OFDM WLAN Baseband*

Receiver with Space Diversity, "AP-ASIC Conference, Fukuoka, Japan, August 2004.

- ✧ Sang-Jung Yang, Yi-Ching Lei, and Tzi-Dar Chiueh, "*Design Design and simulation of a Baseband Transceiver for IEEE 802.16a OFDM-Mode Subscriber Stations,*" APCCAS 2004, Taiwan, R.O.C., Dec., 2004.
- ✧ Chi-Yeh Yu, Zih-Yin Ding, and Tzi-Dar Chiueh, "*Design and Simulation of a MIMO OFDM Baseband Transceiver for High Throughput Wireless LAN,*" APCCAS 2004, Taiwan, R.O.C., Dec., 2004.

子計畫三:

- ✧ Jih-Chiang Yeo, Huai-Yi Hsu, and An-Yeu Wu, "*A Scalable Reed-Solomon Decoding Processor based on Unified Finite-field Processing Element Design,*" IEEE Workshop on Signal Processing Systems (SiPS-2004), Austin, USA, Oct. 2004.

子計畫四:

- ✧ Ya-Lan Tsao, Wei-Hao Chen and Shyh-Jye Jou, "*Buffering Hardware Nested Loop of Parameterized and Embedded DSP Core,*" The 12th Workshop on Synthesis And System Integration of Mixed Information technologies, 2004.
- ✧ Ya-Lan Tsao, Jun-Xian Teng, Maw-Ching Lin and Shyh-Jye Jou, "*Low Power Correlator of DSP Core for Communication System ,*" The 2004 IEEE Asia-Pacific Conference on Circuits and Systems
- ✧ Ya-Lan Tsao, Yu-Chun Lin, Wei-Hao Chen, Bo-Shiang Huang and Shyh-Jye Jou, "*A module generator for parameterized DSP core,*" The 2004 IEEE Asia-Pacific Conference on Circuits and Systems
- ✧ S. J. Jou, K. Y. Jheng, H. Y. Chen H. Y. Chen, and A. Y. Wu, "*Multiplierless Multirate Decimator/Interpolator Module Generator,*" IEEE Asia-Pacific Conference on Advanced System Integrated Circuits (AP-ASIC 2004), Aug. 2004 pp. 58-61.

子計畫五:

- ✧ Hung-Kai Chen, and Chau-Chin. Su, "*Convolusion Based RF Transceiver Testing Methodology,*" Proc. IEEE VLSI Test Symposium 2004, Monterey CA, USA.
- ✧ Hung-Kai Chen, and Chau-Chin. Su, "*Convolusion Based RF Transceiver Testing Methodology,*" VLSI Design/CAD Symposium 2004, Pingtong, Taiwan R. O. C.

子計畫六:

- ✧ Hsi-Lu Chao and Wanjiun Liao, "*Fair Scheduling with QoS Support in Ad*

Hoc Wireless Networks," accepted by IEEE Transactions on Wireless Communications, 2004.

- ✧ Hsi-Lu Chao and Wanjiun Liao, "*Channel Compensation for Multimedia Wireless Ad Hoc Networks with Channel Errors*," IEEE Transactions on Wireless Communications 2004.
- ✧ Wanjiun Liao, Chang-Jung Kao, and Chin-Hei Chien, "*Improving TCP Performance in Heterogeneous Mobile Networks*," IEEE Transactions on Communications, April 2005.
- ✧ Wanjiun Liao, "*The Behavior of TCP over DOCSIS-based CATV Networks*," accepted by IEEE Transactions on Communications 2004.
- ✧ Jiunn-Ru Lai and Wanjiun Liao, "*Modeling User Mobility for Reliable Packet Delivery in Mobile IP Networks*," Proc. IEEE PIMRC 2004, Barcelona, Spain, Sept. 2004.
- ✧ Hsi-Lu Chao and Wanjiun Liao, "*Fair Scheduling on Ad Hoc Networks with Channel Errors*," Proc. IEEE VTC-Fall 2004, Los Angeles, CA, Sept. USA.
- ✧ Jia-Chun Kuo and Wanjiun Liao, "*Modeling the Behavior of Flooding on Target Location Discovery in Mobile Ad Hoc Networks*," accepted by IEEE ICC 2005, Korea, June. 2005.
- ✧ Chong-Wei Bao and Wanjiun Liao, "*Performance Analysis of Reliable MAC-Layer Multicast for IEEE 802.11 Wireless LANs*," accepted by IEEE ICC 2005, Korea, June. 2005.

子計畫七:

- ✧ Chung-Yuan Knight Chang and Eric Hiao-Kuang Wu, "*SARS : A Linear Source Model Based Adaptive Rate-control Scheme for TCP-friendly Real-time MPEG-4 Video Streaming*," WCMC 2004
- ✧ Eric Hsiaokuang Wu, Hsu-Te Lai, and Meng-feng Tsai," *Low Latency and Efficient Packet Scheduling for Streaming Applications*", Proceedings of IEEE ICC 2004.
- ✧ Eric Hsiaokuang Wu, Chien-Shao Tseng, ChungYuan Chang, and Meng-feng Tsai,"*TMRC: A Load-Adaptive TCP-Friendly Rate Control Protocol for Real Time Multimedia Applications*" Proceedings of IEEE ICC2004.
- ✧ Yu-Liang Kuo, Eric Hsiaokuang Wu, and Gen-Huey Chen, "*Noncooperative Admission Control for Differentiated Services in IEEE 802.11 WLANs: Game Theoretic Framework*", Proceedings of IEEE Globecom 2004, Dallas, Texas.

附錄二：計畫成果海報展示

附錄三：計畫成果技術轉移項目

- ✧ 應用 MIMO-OFDM 技術之高速無線區域網路基頻訊號處理架構(C 程式) - 台灣大學電機系 闕志達

附錄四：出席國際學術會議心得報告及發表之論文

附錄五：會議記錄

一、中文摘要

本 IT 整合計劃研究範圍包括有線/無線寬頻網路的通訊架構與無線區域網路之架構；前者包含多媒體與多重服務(如 VoIP、Video Streaming、Internet Access)技術、離散多載波(Discrete Multi-Tone, DMT)傳輸技術、802.16 分頻多工(OFDM)技術、DMT/OFDM 傳收機系統架構及其超大型積體電路設計與晶片製作。前者包括 2 到 11GHz 射頻電路技術及類比前端訊號處理技術、DMT/OFDM 基頻處理技術;後者包括 5GHz 射頻前端、802.11a 與 802.11g 射頻信號處理及電路架構、無線區域網路 Soft-radio DSP 架構，與網路協定之製作。

本計畫為三年計畫之第二年度，在子計畫一的部份為設計及實現一個適用於IEEE802.11a所訂定的高速無線區域網路標準的基頻傳輸系統，包括發射機及接收機。同時針對射頻前端設計一適用於IEEE802.11a/b/g三模雙頻(2.4及5.8GHz)的電路。同時也針對系統整合的研究做模擬與分析，以期單晶片系統(SoC)的實現。另外，子計畫二的部份的研究目標是發展使用正交分頻多工調變之高速無線區域網路及無線都會區域網路收發機實體層基頻電路的演算法與架構設計。其中高速無線區域網路採用多重輸入多重輸出正交分頻多工技術，而無線都會區域網路則採用IEEE 802.16-2004標準，並針對標準的要求做系統的模擬驗證並評估效能。最後將系統轉成可合成的硬體程式語言，以FPGA驗證通過。今年度的研究重點以軟體模擬為主，下年度將著手進行硬體設計及FPGA驗證。在子計畫三的部份是發展無線區域網路(IEEE802.11i)進階加密標準系統架構及進階加密標準數位模組化設計以及前端電路設計。另外，對於以AMBA介面標準整合於系統晶片架構的考量，針對系統晶片的設計環境做整合性的模擬驗證。同時配合本整合計畫內其他子計畫的設計做完整積體電路的整合模擬與評估。在子計畫四的部份，我們完成了河流式架構DSP核心設計、無線網路在多重(Multi-DSP)DSP架構之評估以及DSP核心之IP化實現。除此之外，我們完成了數位有限脈衝響應多階多速率降頻器之模組產生器。使用者能夠藉由此產生器，自動設計出高速低複雜度的數位有限脈衝響應多階多速率濾波器。在子計畫五內，推導和實作一種利用產生以及接收中頻信號來測試與偵測混合信號電路中之中頻與射頻電路模組的方法。利用數位信號處理器產生測試訊號，經由數位類比轉換器產生較低頻的中頻測試訊號輸入至待測的發射器之中頻與射頻模組。在接收端，利用類比數位轉換器擷取波形，接著傳送到數位信號處理器作分析。利用這個方法可以大幅節省測試成本，因為所需之各項元件均已內建在原本之電路中。且我們不需要額外的射頻自動化測試設備。

子計畫六在802.11e WLANs提出兩個方法BIWF-SP及IDFQ-SP，分別以backoff interval及interframe space為基礎以提供不同AC不同的服務品質，包括絕對的優先順序及依權重分配頻寬。在這份報告中，我們仔細的描述這兩個方法的作法，及透過模擬比較它們和原來EDCA的比較。模擬結果顯

示BIWF-SP及IDFQ-SP在支援絕對的優先順序及依權重分享頻寬方面都表現的比EDCA好。和IDFQ-SP比較起來，BIWF-SP在實際系統比較容易實作；但和BIWF-SP比較的話，IDFQ-SP有較好的aggregate throughput及較穩定的優點。更重要的是，這兩個方法都和現有的802.11e標準相容，表示它們都很適合在802.11 WLANs裡對每個AC提供不同的服務品質。子計畫七之研究成果分別為應用於MPEG-4 即時視訊平台中以線性模型為基礎的可調適性速率控制機制，一個可以幫助視訊串流服務，適合於隨時在變動的可用頻寬現制下的可調適性速率控制機制，必須要被發展設計出來。另一成果為在無線網路環境下的可靠多點傳輸的壅塞控制，壅塞控制的機制使得網路傳輸，可以調整傳送速率符合網路上可用的頻寬，並且公平地與目前盛行的網路傳輸方式。

關鍵詞：
分散式多載波調變、收發機、寬頻、正交多頻分工、無線區域網路、無線都會區域網路、多重輸入多重輸出、高速無線區域網路、收發機、進階加密標準、系統晶片、AMBA、數位模組化設計、DSP 核心、內嵌式、可參數化、河流式、多階多速率、降頻器、混合信號、射頻電路模組、可靠群播服務、可調適服務品質保證、MPEG-4、調整速率控制、線性來源模型、壅塞控制、無線網路

Abstract

The NSC-IT integrated project includes wired and wireless wideband communication systems. In the wired system, multimedia and multi-service, e.g. VoIP, Video streaming, and Internet Access are included. For the physical layer, discrete multi-tone (DMT) modulation technique, IEEE802.16 OFDM technique, DMT/OFDM transceiver architecture, and VLSI circuits design and chip implementations are taken into account. In the wireless systems, it includes RF designs, between 2 and 11 GHz, RF signal processing and circuit designs for 802.11a/b and g, Soft-radio DSP for WLAN, and protocols for WLAN.

This report is for the second year of the three-year project. In the sub-project 1, the goal is to design baseband system architecture of IEEE802.11a and RF circuits of dual band RF system for IEEE802.11a/b/g. Besides, the studies and emulation of mixed mode signal integration simulation are included. Meanwhile, the results of other groups will be taken into account for the integrated emulation. In the sub-project 2, the goal is to design baseband OFDM transceiver algorithm and architecture including the high throughput wireless LAN and wireless MAN baseband transceiver. The high throughput wireless LAN and the wireless MAN systems adopt respectively MIMO-OFDM technology and IEEE 802.16-2004 standard to implement and evaluate the performance. Finally, the C code will be translated to synthesizable Verilog code and then pass the FPGA verification. This year, the research target is software simulation. Next year, we will start hardware design and FPGA verification. In the sub-project 3, the goal is to design Advanced Encryption Standard (AES) system architecture of IEEE802.11i and IP design of AES for IEEE802.11i. Besides, the studies and simulation of System on Chip integration simulation with AMBA compliant Interface are included. Meanwhile, the results of other groups will be taken into account for the integrated simulation. In the sub-project 4, the goal is the design a stream-based DSP core, performance analysis of the multi-DSPs architecture for wireless network and implementation the IP of the DSP core. Besides, a module generator, which can automate the process of designing high-speed low-complexity multirate multistage digital decimator, is carried out. In the sub-project 5, To derive and implement a methodology to test and debug IF and RF modules using the IF signal generated and received at the mixed signal modules. Low IF test signal is generated by DAC from DSP module and fed to the IF/RF module at transmitting end. At the receiving end, the response waveform is captured by the ADC and transported to DSP for analysis. By this method, the test cost is

minimal because the AD/DA converters and DSP modules are built in already. There is no external RF ATE needed.

For the sub-project 6, we propose two mechanisms, called BIWF-SP and IDFQ-SP, based on backoff interval (BI) and inter-frame space (IFS), respectively, to provide per-class QoS at the MAC layer for IEEE 802.11e Enhanced Distributed Channel Access (EDCA) WLANs. In our mechanisms, both strict priority and proportional fair service among different service classes are supported. We describe the operations of the proposed mechanisms in details, and compare their performance with the original EDCA mechanism via simulation. The simulation results show that both BIWF-SP and IDFQ-SP outperform the original EDCA in terms of the support for both strict priority and weighted fair service. Compared to IDFQ-SP, BIWF-SP is easier to implement in real systems; compared to BIWF-SP, IDFQ-SP has better aggregate throughput and is more stable. More importantly, both mechanisms conform with the IEEE 802.11e EDCA standard, rendering both good candidates to provide per-class QoS service for IEEE 802.11 WLANs. In the sub-project 7, We have two research results. One of them is "A Linear Source Model Based Adaptive Rate Control Scheme for TCP-friendly Real-time MPEG-4 Video Streaming" The increasing demand for streaming video applications on the Internet or wireless motivates the problem of building an adaptive rate control scheme which is adapted to the time-varying network condition. Another research is congestion control that makes the transmissions adapt the sending rate to the available bandwidth and fairly share bandwidth with the popular traffic in the networks, TCP.

Keywords:

Discrete multi-tone modulation (DMT), transceiver, wideband, OFDM, wireless LAN, MIMO, transceiver, Advanced Encryption Standard (AES), System on Chip, AMBA compliant, IP, DSP core, embedded, parameterized, stream-based, multirate, multistage, decimator, mixed-signal, RF module, QoS, Video streaming, protocol, MPEG-4, adaptive rate control, linear source model, congestion control, wireless network

◆ 子計畫一：系統架構、射頻、中頻及類比前端電路設計（汪重光）

一、簡介

本報告為設計及實現一個適用於IEEE802.11a所訂定的高速無線區域網路標準的基頻傳輸系統，包括發射機及接收機。同時針對射頻前端設計一適用於IEEE802.11a/b/g三模雙頻(2.4及5.8GHz)的電路。同時也針對系統整合的研究做模擬與分析，以期單晶片系統(SoC)的實現。

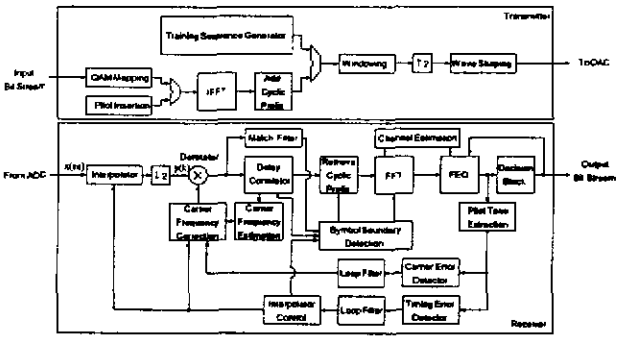
二、計畫緣由與目的

由於對寬頻通訊的要求增加，在1997年IEEE802.11標準被制定，以提供1或2Mbps。但這樣的速率不夠使用，因此IEEE802.11a被提出來做為高速無線區域網路的新標準，它是採用Orthogonal Frequency Division Multiplexing (OFDM)的調變方式，OFDM是被證實用來對付像載頻雜訊及多通道fading最好的傳方式。另外，IEEE802.11射頻載波的位置是在2.4GHz而IEEE802.11a射頻載波的位置是在5GHz，且IEEE802.11g是結合IEEE802.11a與b的規格，是傳輸在2.4 GHz。因此能針對這三種模式的兩個頻段應用來設計此雙頻三模的射頻電路是具挑戰性。另外，介由對混合訊號系統的瞭解來實現單晶片系統(SoC)是目前許多高科技研究人員的目標。

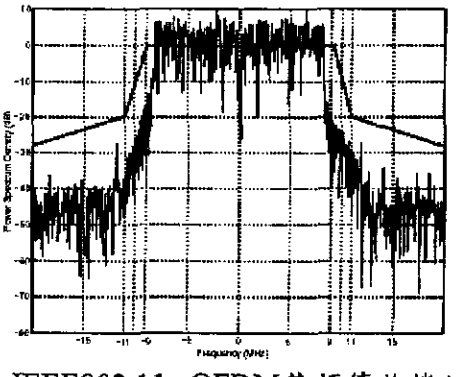
三、基頻系統設計與成果討論

圖一所示為所提出的IEEE802.11a OFDM基頻傳收機的系統架構圖，在發射機方面主要包含：training sequence generator, QAM mapping, pilot insertion, IFFT, CP extension, windowing, 及wave shaping。這個wave shaping filter 的roll-off factor是0.22，所送出的訊號可完全符合規格要求，如圖二所示。

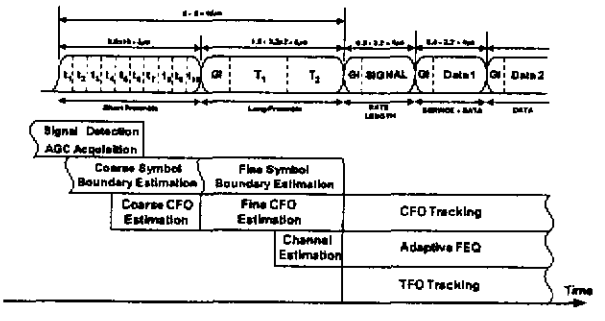
在接收機方面，除了interpolator、symbol boundary detector、de-rotator、delay correlator、match filter及FFT外還包含了一個多階的載波回復電路(multi-step carrier recovery loop)、一個自我修正的時序回復電路(self-correcting timing recovery loop)及對一般頻域等化器加上了自我更新的機制(additional adaptation aided frequency domain equalizer)。其中 FFT and IFFT 是共用一硬體來實現。



圖一、IEEE802.11a OFDM基頻傳收機的系統架構圖。



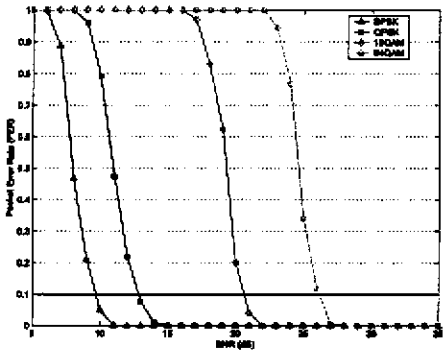
圖二、IEEE802.11a OFDM基頻傳收機的功率遮罩及本系統的發射訊號。



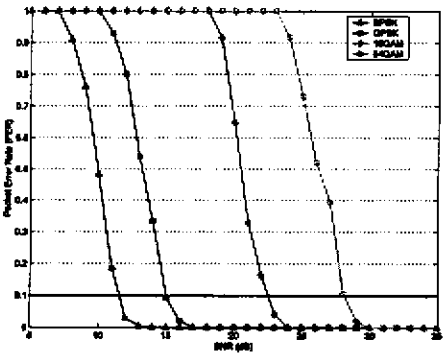
圖三、所設計的訓練機制(training process)。

圖三為所設計的訓練機制(training process)，在short preamble前端約4.8 μs是

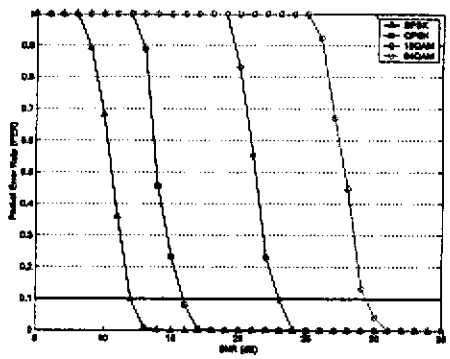
保留給signal arriving detection及AGC acquisition所使用。接著是做為coarse symbol boundary detection及carrier frequency offset (CFO) compensation。然後fine symbol boundary detection、fine CFO estimation及channel estimation則在long preamble內完成。其餘的如residual CFO、timing frequency offset (TFO)及FEQ的參數則以自我適應的方式做修正補償。



圖四、PER在delay spread 為0ns時的表現。



圖五、PER在delay spread 為50ns時的表現。



圖六、PER在delay spread 為100ns時的表現。

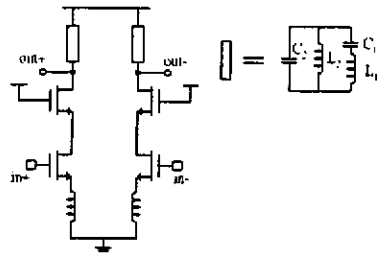
在此基頻系統裡，最高支援的資料傳輸速率可達54 Mbps。在特定的信號雜訊比下，系統可提供低於10%的封包錯誤率的要求。如圖四、五及六所示為packet error rate (PER)在delay spread 為0、50及100ns的表現。表一所示為Packet Error Rate (PER) 符合10%要求下在delay spread 為0、50及100ns時所需的SNR值。

表一、Packet Error Rate (PER) 符合10%要求下在delay spread 為0、50及100ns時所需的SNR值。

SNR Requirement for PER < 10% (without coding)			
Modulation / Delay Spread	0 *ns*	50 *ns*	100 *ns*
BPSK	9.7 dB	11.6 dB	12.0 dB
QPSK	12.8 dB	15.0 dB	15.9 dB
16QAM	20.7 dB	22.6 dB	23.0 dB
64QAM	26.2 dB	28.0 dB	29.3 dB

四、射頻電路設計與成果討論

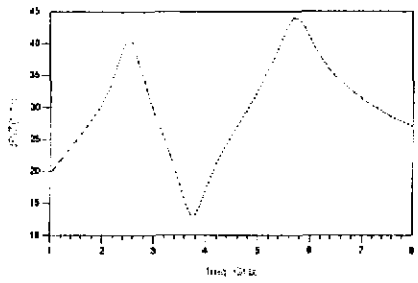
圖七是雙頻低雜訊放大器的電路圖。其基本架構是以電感當source degeneration的差動放大器。為了達到dual-band的input matching，所使用的匹配電路為多級的open stubs。在負載端的LC tanks部分，因為要達到兩個頻帶皆有高阻抗，用兩個電感和電容構成，圖八為其模擬結果，兩個poles的位置分別在2.4 GHz與5.7 GHz；zero的位置則在3.8 GHz。



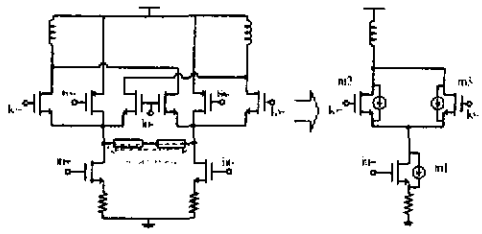
圖七、雙頻低雜訊放大器架構圖。

圖九是接於低雜訊放大器之後的混波器架構圖。降頻的動作顯示在圖十，針對Half-IF的部份，在switching pairs的源端接上LC tanks，這樣會使對flicker noise而言的等效半電路如圖九(b)所示，在低頻的flicker noise不會經第一顆混波器後被升頻到與訊號相同的

地方，而干擾到訊號。

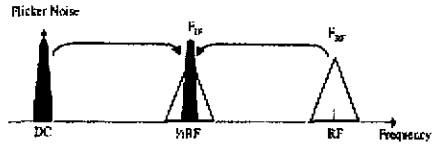


圖八、雙頻負載的阻抗。



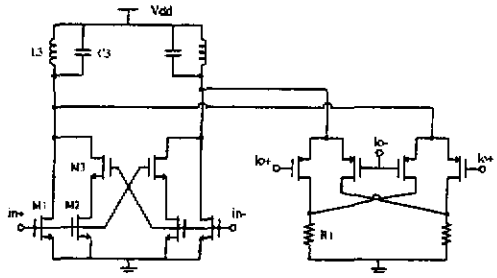
(a) (b)

圖九、雙頻混波器架構圖。



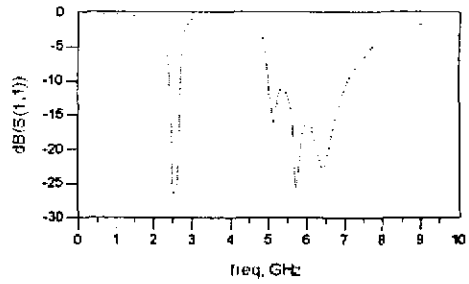
圖十、Flicker noise 的升頻。

圖十一為第二顆混波器的電路架構圖。LNA的S11模擬如圖十二，達到兩個頻帶的matching。Stability factor的模擬如圖十三，從10 MHz到50 GHz都穩定。

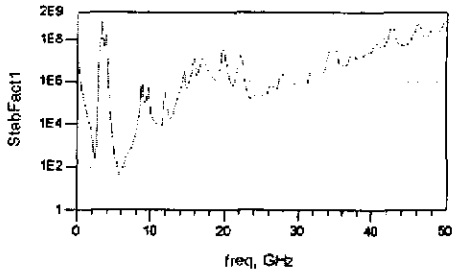


圖十一、第二個混波器。

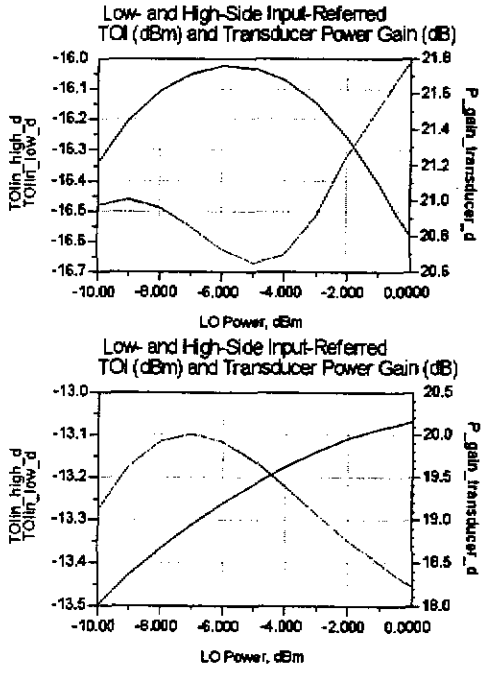
接下來是LNA加第一顆mixer的模擬結果。圖十四為在2.4 GHz的頻帶的IIP3和Conversion gain對LO power作圖。上面是沒有加noise filtering tank的情形，下面是有加。Gain的下降是因為此LC tank在2.4 GHz的Q較差的緣故。因gain下降，所以linearity變好。圖十五為在5.7 GHz的IIP3和Conversion gain對LO power作圖。上面是沒有加noise filtering tank的情形，下面是有加。因在5.7 GHz的Q較高，所以gain較好，相對的IIP3較低。



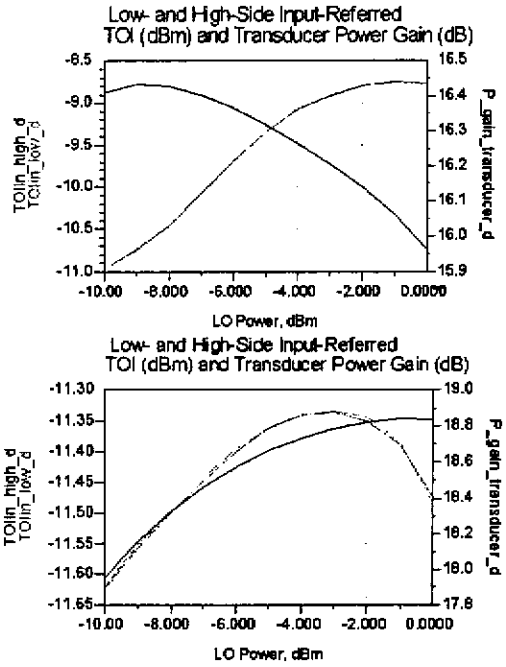
圖十二、LNA 的 S11。



圖十三、LNA 的穩定係數。

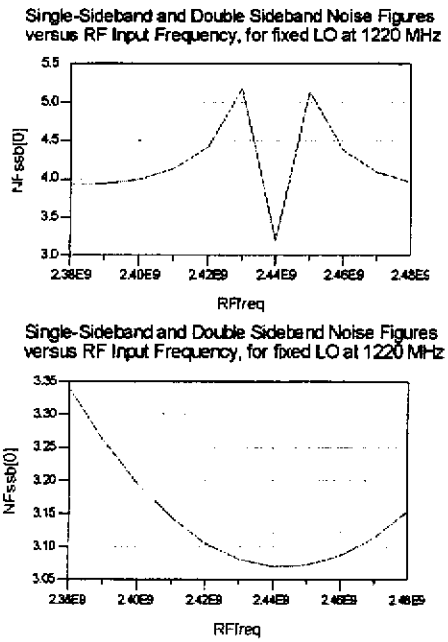


圖十四、2.4 GHz LNA+1st Mixer 的 gain 與 IIP3。

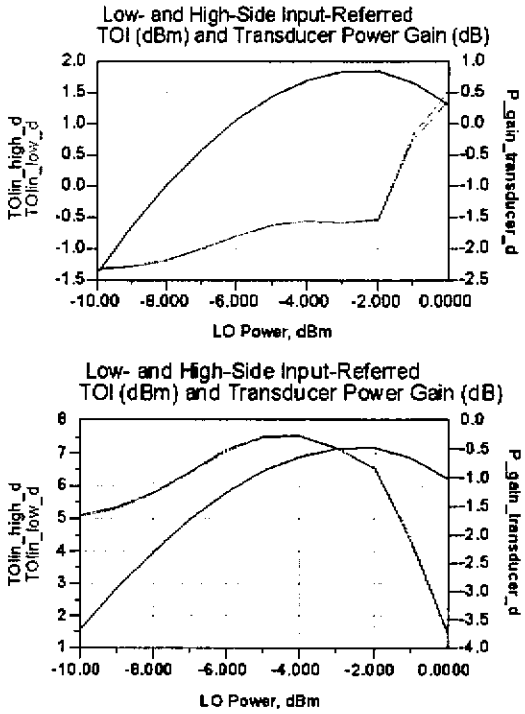


圖十五、5.7 GHz LNA+1st Mixer 的 gain 與 IIP3。

接下來是noise figure的模擬。圖十六是在2.4 GHz的頻帶，LO power為-2 dBm，頻率固定在1.22 GHz，掃RF的頻率，Single side band noise figure的圖。上方為沒加noise filtering circuit時，下方為有加時。可以明顯看到當沒加時，訊號受到flicker noise的影響甚大。

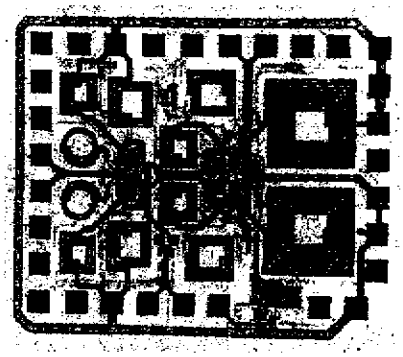


圖十六、2.4 GHz LNA+1st Mixer 的 NF。



圖十七、2nd Mixer 的 gain 與 IIP3。

晶片圖如圖十八所示，包含LNA和第一個mixer。使用on PCB的量測方式，成品如圖十九所示。



圖十八、晶片圖。

圖十九、PCB 板的圖。

圖二十為S11。在2.44GHz為-8.3 dB。圖二十一為LNA加output driver的增益。圖二十

二為其noise figure。圖二十三為全部電路的conversion gain對LO作圖，可知mixer的部分已經沒有增益了。圖二十四為two tones test，IIP3為-5 dBm。圖二十五為noise figure。因為mixer的gain不足，且又有LO至IF的feedthrough，所以會影響noise figure在LO=IF時的量測。

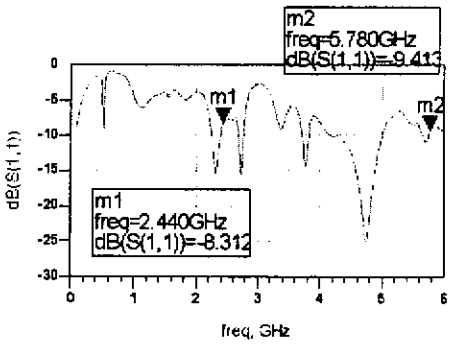
為了知道flicker noise到底有沒有up-converted到IF，在DUT前面外加了一顆放大器，量放大器加DUT的noise figure，最後再經由

$$NF = NF_1 + \frac{NF_2 - 1}{A_{p1}}$$

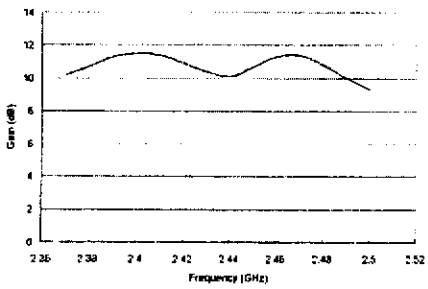
$$\Rightarrow NF_2 = 1 + (NF - NF_1) \times A_{p1}$$

推回DUT的noise figure，結果為圖二十六。雖然還受LO-IF feedthrough的影響，但可以清楚看到此圖並沒有如模擬所呈現的大於10 MHz的flicker noise corner，即沒有up-converted的flicker noise。

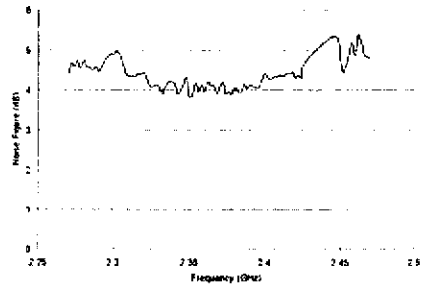
接下來是5.8GHz的頻帶的量測結果。圖二十七為S11，為-7.4 dB。圖二十八為LNA的S21，可知LNA提供的IRR為50 dB。圖二十九為整個晶片conversion gain。因為LNA和mixer的gain不足，所以量測noise figure並沒有意義。全部的量測與模擬的結果整理於表二。



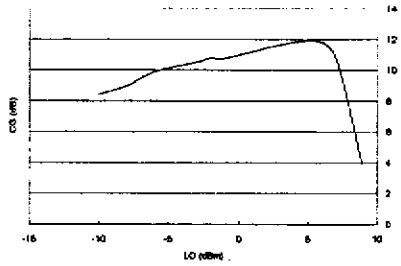
圖二十、2.4 GHz LNA 的 S11。



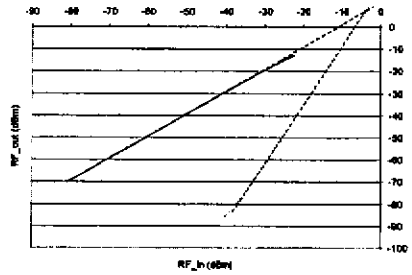
圖二十一、2.4 GHz LNA 的增益。



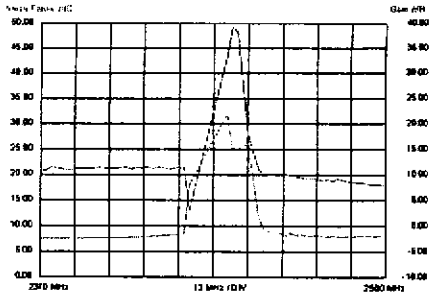
圖二十二、2.4 GHz LNA 的 NF。



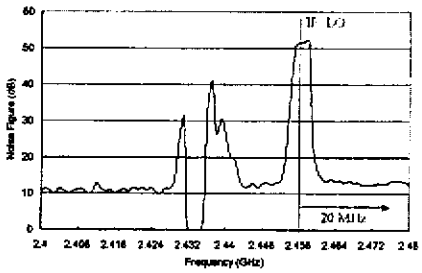
圖二十三、2.4 GHz LNA + mixer 的增益。



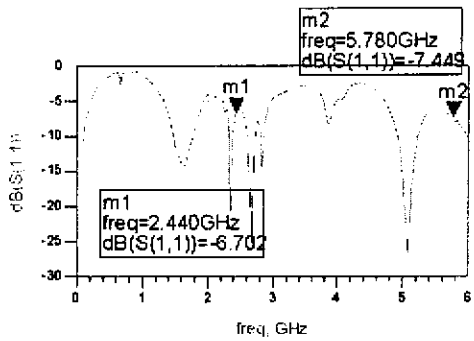
圖二十四、2.4 GHz LNA + mixer 的 IIP3。



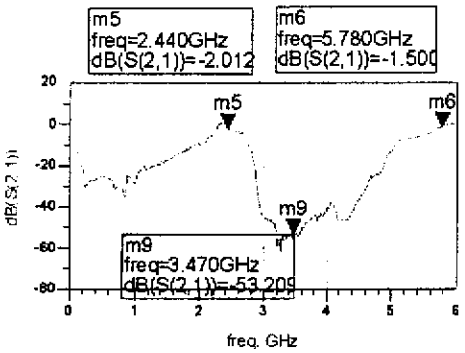
圖二十五、2.4 GHz LNA + mixer 的 NF。



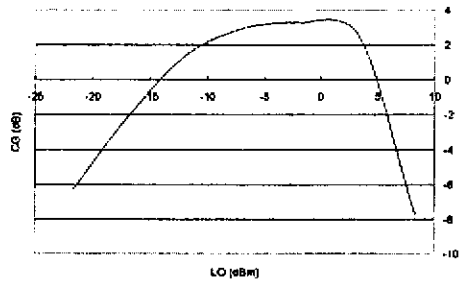
圖二十六、2.4 GHz LNA + mixer 的 NF。



圖二十七、5.7 GHz LNA 的 S11。



圖二十八、5.7 GHz LNA 的增益。



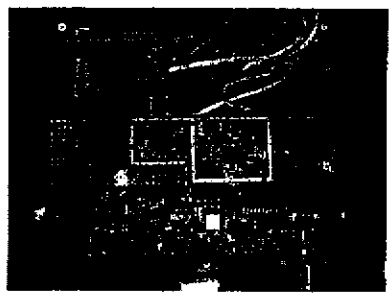
圖二十九、5.7 GHz LNA + mixer 的增益。

五、系統整合考量

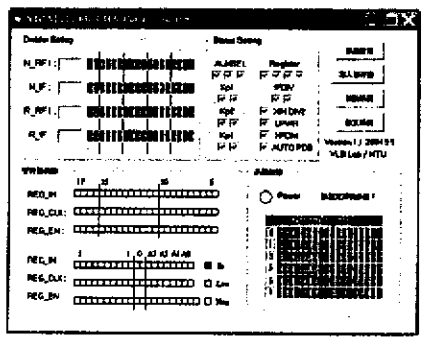
在射頻前端電路架構整合的部份，由於自行研發的射頻前端電路經量測後發現其特性並未達到理想，因此採用了廠商所提供的試驗性整合商用電路板，如圖三十所示，其中包括了可程式化頻率合成器si4136。而我們也自行設計了一套軟體程式，如圖三十一所示，經列印埠來控制其振盪頻率，使其提供一中頻振盪頻率為748MHz及一射頻振盪頻率為2.5GHz，後者再經倍頻器後產生5GHz的振盪頻率來完成一超外差的射頻架構電路。圖三十二所示為傳送機發出的訊號頻譜分析，該訊號是由電腦軟體GPIC產生數位基頻訊號再經AMIQ儀器轉成基頻類比I/Q訊號，接著由射頻調變器將之混波到5GHz的載波位置，可見完全符合IEEE802.11a功率遮罩的要求。圖三十三為接收機解調後的I/Q訊號，是由接收機前端輸入一經混波於5GHz的單頻7MHz的訊號，經射頻兩次解調後所得到的基頻I/Q訊號，由於電路本身有I/Q不匹配的影響加上電路板也有同樣的潛在問題，同時量測儀器也會有I/Q不匹配的現象，所以從圖上可看到兩解調出的基頻訊號有振幅及相位的偏差，各別是43mV及0.6度。

表二、Performance Summary

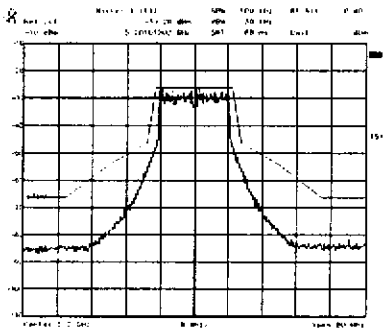
		Simulation	Measurement
2.4 GHz	LNA+Buffer Gain	14 dB	11 dB
	LNA+Mixer Gain	20 dB	12 dB
	LNA+Buffer NF	4 dB	4 dB
	LNA+Mixer NF	3.1 dB	8 dB
	LNA+Mixer IIP3	-13 dBm	-5 dBm
	S11	-20 dB	-8.3 dB
	IRR	30 dB	15 dB
	Up-Converted 1/f noise	No	No
	Power (core only)	24 mW	24 mW
5.7 GHz	LNA+Mixer Gain	18.8 dB	3.5 dB
	LNA+Mixer NF	3.6 dB	25 dB
	LNA+Mixer IIP3	-11.4 dBm	4 dBm
	S11	-15 dB	-7.5 dB
	IRR	60 dB	55 dB
	Power (core only)	24 mW	24 mW



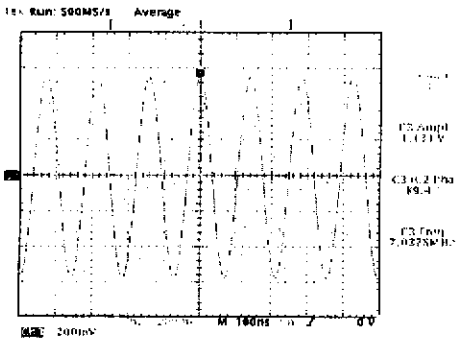
圖三十、射頻前端傳收機電路板。



圖三十一、振盪器控制介面之軟體開發。



圖三十二、傳送機發出訊號的頻譜。



圖三十三、接收機解調的 I/Q 訊號。

六、未來計畫

本期末報告含基頻系統與射頻前端電路的設計，也包含這二個主題的研究成果加以測試與效能分析。再進一步對整合系統加以評估及模擬驗證，透過這樣的分析，我們發現在射頻接收機架構裡I/Q不匹配的影響是必須嚴正去面對的，因此，基於現有的研究成果，在未來我們除了繼續進行混合訊號整合的研究分析外，亦會朝解決I/Q不匹配解決方案的領域邁進，這些都將在往後的研究報告中呈現。此外，基於第二年的研究經驗與所獲知識，在第三年的研究主題：IEEE802.16a(WiMAX)部份，也將會有豐富的成果展現。

七、參考文獻

[1] Wei-Hsiang Tseng, Ching-Chi Chang, and Chorng-Kuang Wang," *IEEE 802.11a WLAN Transceiver Using Self-Correcting Digital Timing Recovery Design And Implementation*, " VLSI Design/CAD Symposium 2004, Pingtong, Taiwan, R. O. C.

[2] Chun-Chih Hou, Ching-Chi Chang, and Chorng-Kuang Wang, "*A Dual-Band IEEE802.11a/b/g Receiver Front-End Using Half-IF and Dual Conversion,*" AP-ASIC 2004, Fukuoka, Japan, Aug., 2004.

◆ 子計畫二：基頻處理電路設計 (闕志達)

一．簡介：

隨著網際網路的蓬勃發展，越來越多相應而生的應用，使得對它的傳輸速度需求也不斷持續的提升，而無線傳輸所帶來的便利性，使得現今的通訊產品漸漸有以無線取代有線的趨勢。本報告分別設計及模擬了適用於IEEE所制定的高速無線區域網路標準802.11n及無線都會區會網路標準802.16-2004的基頻處理電路。

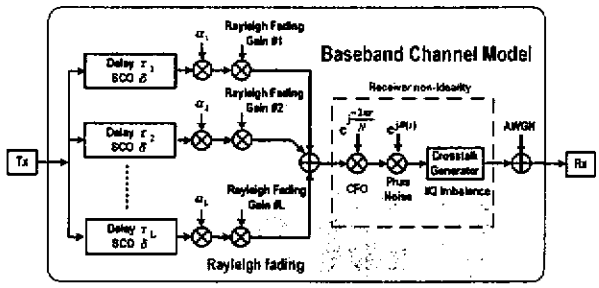
二．計畫緣由與目的：

隨著各式多媒體應用的出現，傳輸速度的需求已變得越來越高，因此IEEE 802.11 task group n在2003年成立，開始制定下一代的高速無線區域網路的規格。目前最具潛力應用於下一代的高速無線區域網路的技術為多重輸入多重輸出正交分頻多工技術(MIMO-OFDM)。由於IEEE 802.11n預計到2005年7月完成草案的制定，目前仍在提案及表決的階段，本計劃目前參考TGn Sync聯盟的提案[1]設計下一代的高速無線區域網路IEEE 802.11n MIMO-OFDM的基頻處理電路。

雖然我們可以利用 IEEE 802.11 的技術無線上網，但是在鄉村地區或人口稀疏之地，住戶密度低，架設有線網路如 DSL 或 cable modem 的成本高昂，並不符合經濟效益。因此 IEEE 製訂了 802.16 的通訊標準，稱為無線都會區域網路，利用無線寬頻上網來解決這「最後一哩」的問題。它還能有效地與 IEEE 802.11 結合，使用者可透過無線區域網路與當地的 802.11 之接收機連結，該接收機再透過 802.16 之收發機與基地台溝通。最後基地台才需要鋪設光纖或電纜來與網際網路骨幹連結。有鑑於此，本計畫亦將設計及實現一個符合 IEEE 802.16- 2004 OFDM 模組的基頻接收機電路。

三．基頻系統設計之通道模型：

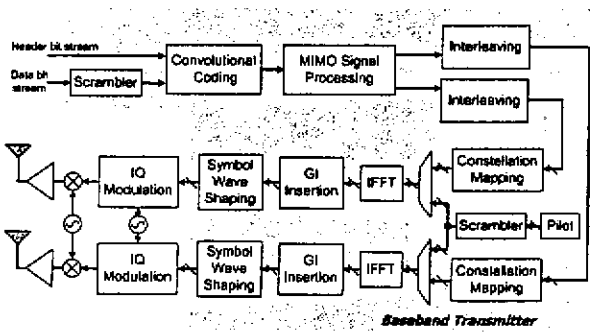
圖一為我們模擬所使用的通道模型，由傳送機送出的信號將加入多路徑(multi-path)、加成性白色高斯雜訊(AWGN)、載波頻率偏移(CFO)、取樣時脈偏移(SCO)等效應之後，再送到接收機。而接收機必須對這些效應做補償及處理方能解回傳送機所送出的資料。



圖一、通道模型

四．高速無線區域網路之基頻系統設計：

所提出的2x2 MIMO-OFDM基頻傳送機的系統架構圖如圖二所示。欲傳送的資料串流首先會經過擾亂器(scrambler)改變信號中位元的分布，加上訊號參數欄位後，一起送入迴旋編碼器 (convolutional encoder) 作編碼，編碼結果再經過 MIMO 訊號處理，我們使用的 MIMO 傳輸技術為空間多工技術(spatial multiplexing)以達到最大傳輸速度的提升，並得到空間的多樣性。分流之後的兩路資料串流經過交錯器(interleaver)，再以 QAM Mapper 對應至星座圖上，並加上經過攪亂器的領航碼(Pilots)。最後經過 OFDM 調變(IFFT、GI insertion)，再對輸出符元作平滑處理(Symbol wave shaping)之後，將會分別經由兩根不同的天線傳送出去，藉此增加傳輸的速度及頻寬使用效率。



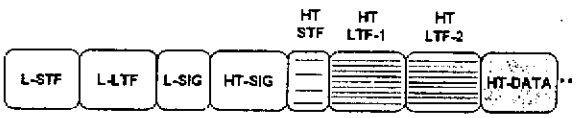
圖二、802.11n 基頻傳送機的系統架構圖

系統使用的 OFDM 參數如表一所示。從表中可以看到大多數的參數均與 IEEE 802.11a 相符合，以達到向下相容的規定，Coding rate 也支援至 7/8，以提供更高的傳輸效率。

Modulation type	BPSK,QPSK, 16QAM,64QAM
FFT size	64
Subcarrier in use	52
Bandwidth	20MHz
Subcarrier spacing	312.5kHz
FFT period	3.2us
GI duration	0.8us/0.4us
OFDM symbol duration	4.0us(=3.2us+0.8us) 3.6us(=3.2us+0.4us)
Error correcting code	Convoultional code(K=7)
Coding rate	1/2, 2/3, 3/4, 7/8
Operating frequency	UNII band

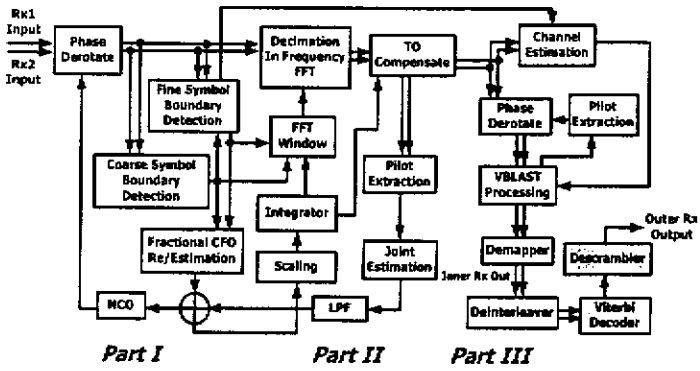
表一、802.11n OFDM系統參數

參考TGn Sync的提案，傳送的封包格式如圖三所示，為了能夠達到向下相容的要求，一開始是一段和IEEE 802.11a完全相同的短前置碼(Legacy Short Training Field, L-STF)、長前置碼(Legacy Long Training Field, L-LTF)與訊號參數欄位(Legacy Signal Field, L-SIG)。接下來是高速訊號參數欄位(High Throughput Signal Field, HT-SIG)，用來存一些要告知接收端關於之後的高速資料欄位(HT-DATA)的資訊，例如資料長度、調變方式、coding rate，並用CRC保護。在高速訊號參數欄位之後則是高速訓練欄位(HT-STF /HT-LTF)以及高速資料欄位。



圖三、TGn Sync提出的802.11n封包格式

接收機的架構則如圖四所示，架構可大約分為三部份，第一部份負責初始同步，包括符元邊界粗調與細調(Coarse/Fine Symbol Timing Detection)，分數載波頻率偏移估計與重估計(Fractional CFO Re/Estimation)。第二個部份負責快速傅立葉轉換(FFT)，並追蹤殘餘的CFO與取樣時脈偏移(Sampling Clock Offset, SCO)。第三個部份則是負責通道估計與資料回復(data recovery)。解出來的資料會送入外接收機進行反交錯、解迴旋碼以及反擾亂以得到傳送的資料串流。以下會分別介紹這三部份的運作方式。

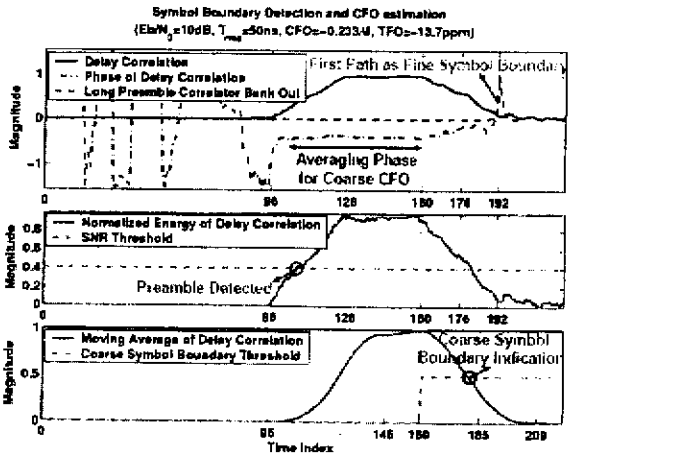


圖四、802.11n 接收機架構圖

✷ 初始同步：

接收機在一開始會進行封包偵測，當偵測到延遲相關器(delay correlation)達到門檻時，作為開始收到封包的根據。由於短前置碼的週期性，延遲相關值會逐漸增加到短前置碼結束時才漸漸下降，藉此現象就可以找到大略的符元邊界，並利用延遲相關值的相位估計fractional CFO。

接著在長前置碼處會再一次估計CFO，接收的長前置碼經過匹配濾波器的輸出，再作移動平均(moving average)後，可利用移動平均的最大值找到最佳符元邊界及FFT window。初始同步的過程如圖五所示。



圖五、初始同步的過程

- 追蹤迴路：

在初始同步之後，估計得到的同步參數可能會與實際值還有一點差距，所以當時域訊號經由FFT轉到頻域之後，我們採用聯合權重最小方差估計(JWLSE)[4]演算法來追蹤殘餘的CFO和SCO，計算連續兩個符元之間的領航碼相位差，並以領航碼的振幅大小作為權重，來估計殘餘的CFO和SCO，估計值會經過一迴路濾波器以去除雜訊帶來的微小變動。

- 通道估測與資料回復：

封包中的HT-LTF是用於MIMO通道估測，由於此段的訊號使用tone interleaving的技術[1]，兩路傳送訊號互相正交，使用one tap equalizer即可得出MIMO通道估計值。資料回復使用VBLAST[5]演算法，以ordered successive cancellation的方式來分離出兩路的傳送資料串流。得到的傳送資料串流會分別進行soft demapping及反交錯，再送入soft decision Viterbi decoder以解迴旋碼，解碼出來的結果再送到反擾亂器得到傳送的資料位元串流。

表三所示為模擬在通道模型D下，分別在使用hard/soft decision Viterbi decoding時，系統符合10%PER所需的SNR值大小。

SNR Requirement for PER < 10%		
Data Rate	Soft Viterbi	Hard Viterbi
108 Mbps	23.3 dB	32 dB
96 Mbps	21.4dB	26dB
72 Mbps	18.5dB	22.3dB
48 Mbps	15.7dB	19.3dB

表三、在通道模型 D 符合 10%PER 的 SNR 值

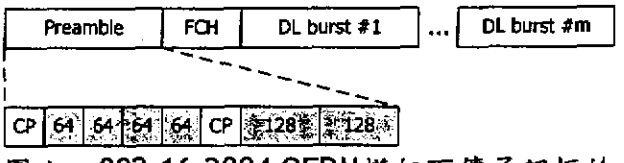
五. 無線都會區域網路之基頻系統設計：

IEEE 802.16-2004是2004年10月才正式定案推出的標準，所以我們將原本設計適用於IEEE 802.16a OFDM模組的程式修改成適用於IEEE 802.16-2004 OFDM模組。表四是此標準的主要參數列表[2]。

RF Frequency (GHz)	2-11
FFT Size	256
Effective subcarriers	192
Bandwidth(MHz)	1.75, 3.5, 7, 14, 28
Sampling rate (MHz)	8/7 x Bandwidth
Guard Interval Ratio	1/32, 1/16, 1/8, 1/4
Index of Pilots	±88, ±63, ±38, ±13
Maximum SCO	±8 ppm

表四、802.16-2004 OFDM模組主要參數列表

基地台與用戶端進行上下傳的動作是以訊框為單位，每個訊框包含上傳與下傳兩個子訊框。標準中下傳子訊框的封包格式如圖六所示，前置碼(preamble)共兩段，在時間上分別是四段64點重覆和二段128點重覆信號，供接收機做封包同步和通道估測使用。前置碼後的FCH(Frame Control Header)含有調變、封包長度等資料，接收機需根據解讀FCH的資料將其他信號做相對應的處理。

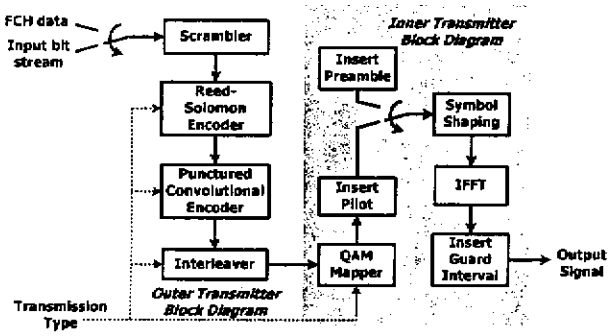


圖六、802.16-2004 OFDM模組下傳子訊框的封包格式

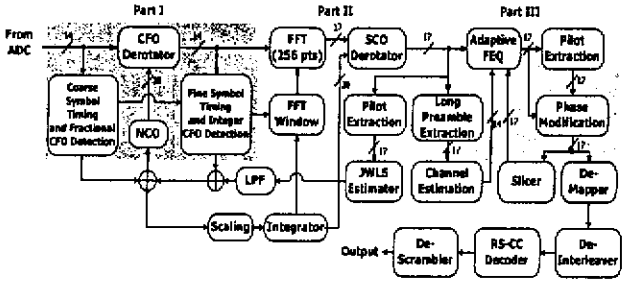
圖七為IEEE 802.16-2004 OFDM模組傳送機的架構圖。首先，資料經過擾亂器將資料打亂，避免一連串0或1的序列產生，以降低PAPR。接著，經過通道編碼和交錯器，藉由增加一些多餘的位元數來降低位元錯誤率(BER)。然後，加碼後的資料位元被送入內傳送機進行OFDM 調變，包括QAM-Mapper，領航碼及前置碼的加入，IFFT和保護區間(guard interval)的加入。最後，將調變後的信號送到RF模組送出。

圖八為我們所設計的接收機的架構圖，依其功能主要可分為三大部分。第一部分是時域方面的信號處理，包括初始同步和載波頻率偏移的補償。第二部分是FFT及通過FFT之後載波頻率偏移及取樣時脈偏移的追蹤迴

路。第三部分是通道估測及頻域等化器(FEQ)等在頻域的信號處理，通過de-mapper後將位元序列送到外接收機進行解碼。



圖七、802.16-2004 OFDM模組傳送機架構圖



圖八、802.16-2004 OFDM模組接收機架構圖

※ 符元邊界粗估及分數CFO的偵測：

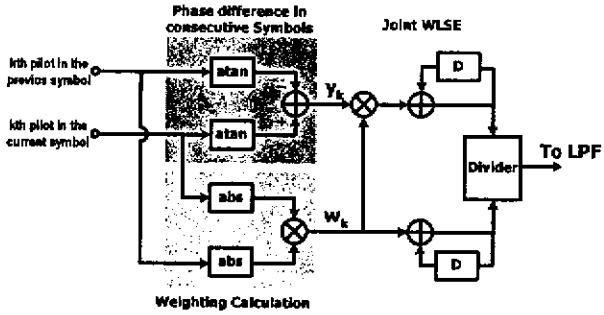
利用延遲相關器出現最大值的時候，視為符元邊界，由於短前置碼在時域上有四段重覆信號，所以最大值處會出現平台，找到的符元邊界變化較大，所以我們改採用延遲相關器出現最大值，而後其值降為原來一半的地方做為細估的符元邊界。分數CFO則利用延遲相關器出現最大值時的輸出相位即可估得。

※ 符元邊界細估及整數CFO的偵測：

由於最大的SCO為16 ppm，最大載波頻率偏移可達到10.9個子載波間距(subcarrier spacing)，因此我們需要做整數CFO的偵測。利用7套匹配濾波器，分別對應短前置碼受到不同的整數載波頻率偏移量，與補償過分數CFO的信號相乘，其中只有一套匹配濾波器會有峰值出現，我們只要找出峰值出現在那一套，即可偵測出整數CFO的量。接著，只要再利用一套匹配濾波器，使長前置碼與補償完整數及分數CFO之後的信號相乘，出現最大值的所在即符元邊界。不過由於有多路徑效應，出現最大值位置不一定就是第一條路徑，所以得到的位置需再做些微的調整，才得更有效地避免ISI的發生。

※ CFO及SCO追蹤迴路：

同樣地，我們利用聯合權重最小方差估測器來同時估測出殘餘CFO及SCO的量[4]，其架構圖如圖九所示。我們一樣在估測器輸出端接一個低通濾波器(LPF)使迴路的表現較穩定。根據標準訂定，接收端的次載波降頻及取樣時脈皆來自同一個振盪器，所以將CFO的總量，經過簡單的運算即可得到SCO的值，再利用SCO-derotator做補償，將信號反轉回正確的星座圖位置上。



圖九、聯合加權最小方差估測器電路圖

※ 通道估測：

由於接收端已知長前置碼的理論值，所以將收到的長前置碼除以各對應次載波上的長前置碼理論值，即能得到各次載波的頻域通道增益。但由於長前置碼只有偶數頻才有理論值，所以奇數頻的通道增益需經由內插而得。我們利用6階的Raised-Cosine內插器，並將其係數乘以一旋轉量，相當於此內插器在時域上涵蓋的範圍向右移動，而能包含更多的延遲訊號的能量，以提升內插的準確度。

標準中規定在AWGN通道下，加上最大可能之CFO及SCO，各傳輸頻型達到位元錯誤率為10^{-6}所需之訊雜比如表六所示。表中同時列出我們所設計的接收機在同樣的測試環境下所需之訊雜比，可以看到模擬結果均比標準中要求之訊雜比還低，因此本系統符合標準的要求。

Type	Modulation	Code Rate	Required SNR (dB)	Resultant SNR (dB)
1	QPSK	1/2	9.4	7.4
2		3/4	11.2	11.0
3	16-QAM	1/2	16.4	13.7
4		3/4	18.2	17.1
5	64-QAM	2/3	22.7	21.4
6		3/4	24.4	23.0

表六、802.16-2004 OFDM模組接收機在AWGN通道下可達10^{-6} BER所需之SNR列表

六、未來計畫：

本期末報告參考 TGn Sync 的 IEEE 802.11n 提案設計出 MIMO-OFDM 系統基頻收發機，之後將繼續完成其他的附加功能，例如 transmit beamforming，研究 3~4 根天線的傳送接收演算法簡化，並作到系統的定點數模擬及驗證。在 IEEE 802.16-2004 方面，本期末報告包含 OFDM 模組基頻收發機的演算法及架構設計與系統模擬結果，未來會將此系統改寫為可合成的硬體描述語言 (Verilog)，然後以 FPGA 驗證其效能。

七、 參考資料：

[1] http://www.tgnsync.org/techdocs

[2] IEEE Std 802.16-2004

[3] 楊尚融, IEEE 802.16a 用戶端基頻收發機之設計, 93 年七月

[4] Pei-Yun Tsai, H. Y. Kang, and Tzi Dar Chiung, "Joint weighted least squares estimation of frequency and timing offset for OFDM systems over fading channels," in *Proc. of 2003 IEEE 57th Vehicular Technology Conference*, 2003, pp. 2543-2547

[5] P.W. Wolniansky, G.J. Foschini, G.D. Golden and R.A. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *1998 URSI International Symposium on Signals, Systems, and Electronics*, Sept.-2 Oct. 1998, pp. 295-300

◆ 子計畫三：適用於無線區域網路之加解密引擎 IP 設計及實現（吳安宇）

一．簡介

在 WLAN (Wireless Local Area Network) 無線區域網路中，安全性的問題非常重要;在 IEEE802.11i 標準中選擇進階加密演算法來保障其安全性。在網路的分層架構中，加解密標準介於軟體和硬體實現之間；因此我們採用系統晶片(System on Chip)中軟硬體協同設計(HW/SW Co-design)的方式進行實現。系統晶片因為面積小以及可重覆使用性高的特點，逐漸成為未來晶片開發的主流。因此我們以 ARM 公司所提供之現有平台(Example AMBA System)和 ARM922T 精簡化指令集處理器(RISC Processor)進行開發。將進階加密標準以硬體方式實現成模組化 IP，以符合進階高效能滙排流標準(Advanced High-Performance Bus)後；配合軟體實現 IEEE802.11i 所規定之計數器及串列編碼協定 (Counter mode with Cipher block chain-MAC Protocol)。

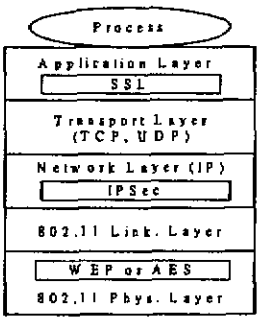


圖 一：Full description of security in 802.11

二．計劃緣由與目的

隨著無線區域網路技術的發達，擁有可隨身攜帶性的特點使得無線區域網路被廣泛利用。但是無線區域網路是以無線電波在空氣中進行傳播，使得資料容易被竊聽；因此無線區域網路需要比有線網路更高的安全性。

在無線區域網路中，為了因應無線電波在空中傳播時造成的安全性影響。因此 IEEE 制定了 IEEE 802.11i 取代了原有的 WEP(Wired Equivalence Privacy)演算法 [1,2]。在 IEEE 802.11i 中訂立了 Temporal Key Integrity Protocol (TKIP) 以及 Counter mode with Cipher block chain-MAC Protocol (CCMP)兩種演算法。其中 TKIP 演算法被用於在新標準採用的過渡期，因此 TKIP 仍以 WEP 演算法作為核心。而另一方面，CCMP 演算法則用於改良原有的加密安全性。CCMP 採用新提出的進階加密演算法 (Advanced Encryption Standard)作為核心 [3]。

三．設計與成果討論

在這個子計劃中，首先 CCMP 的加解密流程可以視為 AES 的兩種模式(counter mode，Cipher-block-chain mode)結合而產生的加密法。因此我們的架構設計為實現一個符合 AMBA(Advanced Micro- processor Bus Architecture)介面標準的雙模式進階加密標準 IP。如圖 六所示，雙模式進階加密標準 IP 包含了基本的進階加密標準 IP、計數器、以及密文回授電路。如圖 七所示，為了符合 AMBA 介面標準，包含了 address decoder、FIFO、freq. divider。

表格 1 wireless LAN requirement

Spec	802.11	802.11a	802.11b	802.11g
Time	Aug-99	Jul-99	Jul-99	Jul-01
Freq Band	2.4~2.5 GHz	5.25, 5.7 GHz	2.4~2.5 GHz	2.4~2.5 GHz
Multiplex	FHSS/DSSS	OFDM	DSSS	DSSS
Modulation	B/QPSK	B/QPSK~64QAM	CCK	PBCC
Bit rate	1.2 Mb/s	6~54 Mb/s	5.5~11 (22) Mb/s	~54 Mb/s

在表格 1中可以得知，我們所實現的無線區域網路加解密系統，必須達到 54Mb/s 的 through-put rate。依表格 2中，我們實做出來的 AES IP 可達到的 CCMP through-put=83MHz*1024bit/18cycle/2times=2.36Gb/s

其中 1024bits 是每個 block 的資料長度，而 18cycles 指資料搬移和資料進行加密所需的週期數；最後，2times 則是代表每筆資料均需經過 counter mode 以及 CBC mode 的演算。因此，我們所實做的 AES IP 可以符合無線區域網路的需求。

表格 2 AES IP performance

	Area	Power	Speed
AES	593,367um2	351.24mW	83MHz

- 系統晶片之必要性

由上述的內容可以得知，我們所實現進階加密標準的加密速度遠高於無線區域網路的要求。近年來資料的安全性越來越重要的同時，進階加密標準必然會被大量採用；另外，系統晶片的推廣促使 IP 的可重覆使用性受到重視。因此我們將 AES IP 模組化，使 AES IP 在單一晶片能被更多的系統重覆使用。藉以提高晶片面積使用率。

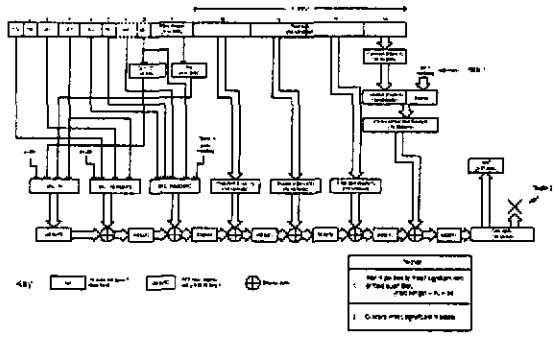


圖 二 CCMP MIC computation block diagram

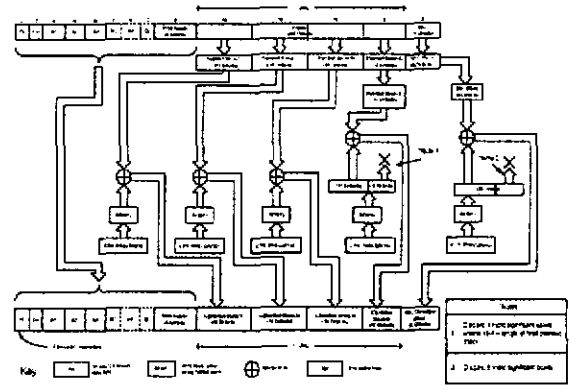


圖 三 CCMP encapsulation block diagram

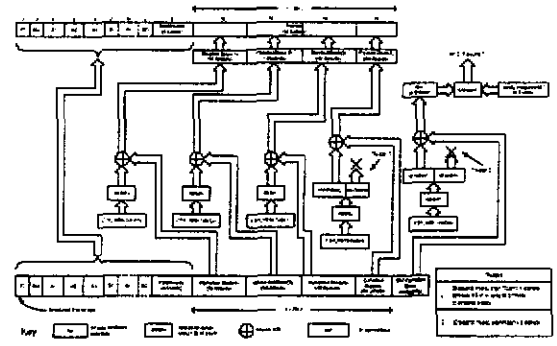


圖 四 CCMP decapsulation block diagram

我們以前一年計劃的成果-AES IP(圖 五)為基礎開發 CCMP 協定的流程。我們依據 CCMP 的特性決定軟硬體分割的界限(圖 六 HW/SW partition format for CCMP)。

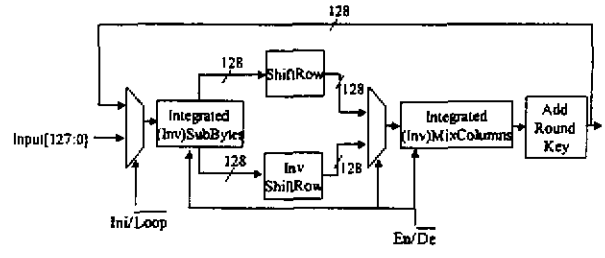


圖 五 Enc/Dec AES Hardware Architecture

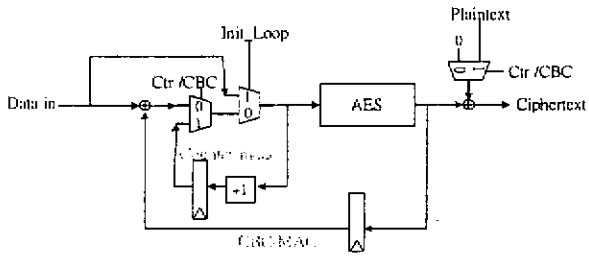


圖 六 HW/SW partition format for CCMP

- 以 ARM 爲基礎之 SoC 平台實現

近年來，設計的複雜性隨著製程技術的進步而增加；為了縮短設計時花費的時間，數位模組的重覆使用越來越受重視。SoC 設計方式的發展建基於大量數位矽智產的重覆使用。因此，我們將 AES 的加解密引擎更進一步放到 SoC 的平台上實現；可以更進一步的加強既有數位模組的重覆使用性，而且可以大幅減少移植時的困難度。

為了加強數位模組的重覆使用性，我們以嵌入式系統常用的 AHB(Advanced High-performance Bus)做為矽智財之介面；使 AES IP 更易於和其它使用共同介面的 IP 進行整合。AES IP 的介面經由 AHB Wrapper 進行轉換，使得 IP 的訊號易於和滙排流訊號進行交換(圖 七)。

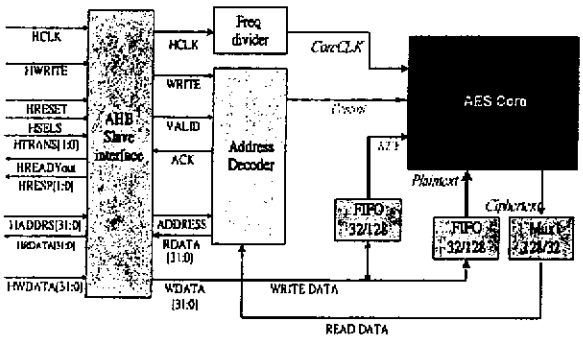


圖 七：AHB bus wrapper implementation

最後，我們在 SoC 的平台上進行實現和驗證，系統架構圖如圖 九，模擬結果如圖 八所示。並以 VLSI 進行電路實現，實體佈局如圖 一一所示。模擬規格如 Table 1 所示。

圖 八：AES IP interface timing diagram

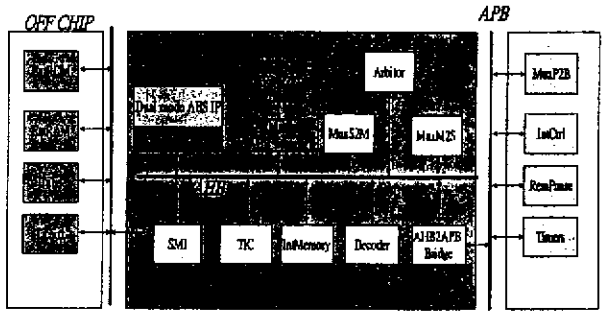


圖 九：EASY platform ARM922T verification

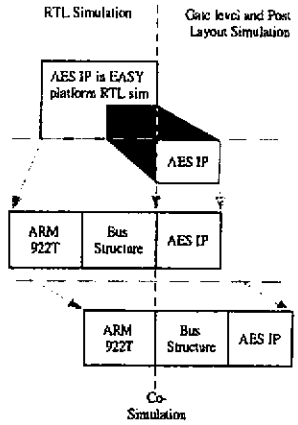


圖 一〇：AES IP RTL/GateLevel verification flow

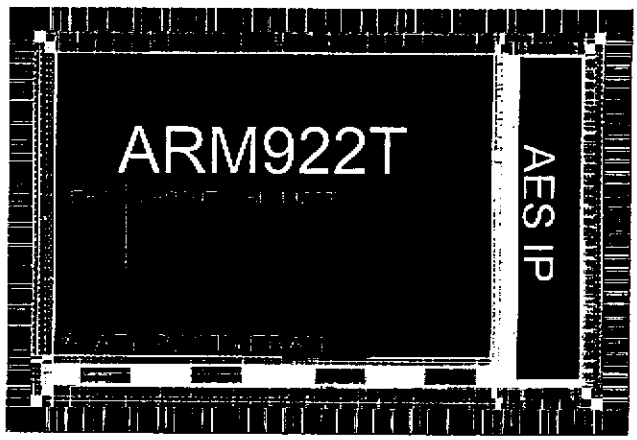


圖 一一：ARM based AES encryption engine

Table 1：layout specification

Cell Library	Artisan 0.18um 1p6m
Voltage	1.8v
Core size	4438x3105um^2
Die size	5X3.5mm^2
i/o pad	128pin package
speed	125MHz
Ram size	4X256Bytes

四、結果與討論

在本子計劃中，我們針對 WLAN Security Engine 來設計高效能/高重覆使用率的數位 IP 模組。藉以提高數位 IP 模組的廣泛的可用性，最後以 VLSI 電路進行實現。

五、參考文獻

[1] "LAN MAN Standards of the IEEE Computer Society. Wireless LAN medium access control (MAC) and physical layer (PHY) specification. IEEE Standard 802.11, 1997 Edition, " 1997.

[2] "LAN MAN Standards of the IEEE Computer Society. Wireless LAN medium access control (MAC) and physical layer (PHY) specification. IEEE Standard 802.11, 1999 Edition, " 1999.

[3] "Advanced Encryption Standard (AES)," Federal Information Processing Standards Publication 197, November 26, 2001.

[4] "AES Cryptographic Engine Design for 802.11i Applications," Hsin-Chung Wang, and An-Yeu Wu, Master Thesis, July 2001.

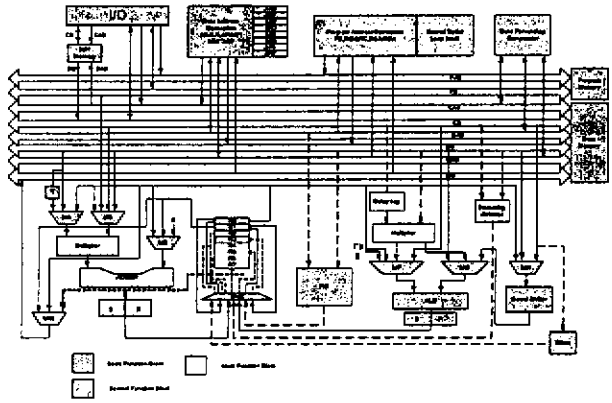
[5] "New algorithms for min-cut replication in partitioned circuits," Hannah Honghua Yang, D.F. Wong, IEEE/ACM International Conference on Computer Aided Design, 1995.

[6] "AHB EASY technology reference manual" in ARM University Kit.

◆ 子計畫四：適用於無線網路之河流式數位信號處理核心 (周世傑)

一. 簡介

本計畫為三年計畫之第二年度。在子計畫四的部份，圖一顯示我們所完成的適用於通訊系統的 embedded DSP core。在這一年，我們完成了河流式架構 DSP 核心設計、無線網路在多重(Multi-DSP)DSP 架構之評估以及 DSP 核心之 IP 化實現。除此之外，我們完成了數位有限脈衝響應多階多速率降頻器之模組產生器。使用者能夠藉由此產生器，自動設計出高速低複雜度的數位有限脈衝響應多階多速率濾波器。



圖一 Embedded DSP core之方塊圖。

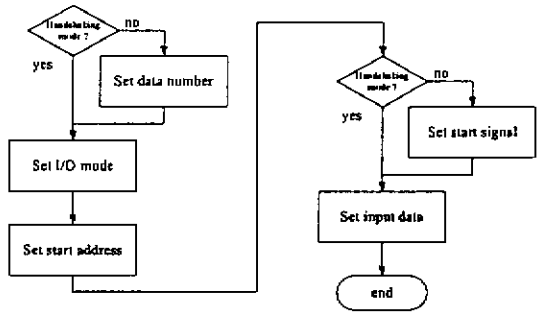
二. 河流式數位信號處理核心架構設計與討論

有鑒於系統複雜度日益提升，在相同時間內，需要完成更多的運算，只靠一顆 DSP 勢必無法完成全部的運算要求，在系統中整合兩顆以上的數位訊號處理器，將整個系統分配至個別的處理器是必然的現象，而如何把資料在多顆處理器中有效率的流動，就成為十分重要的課題。

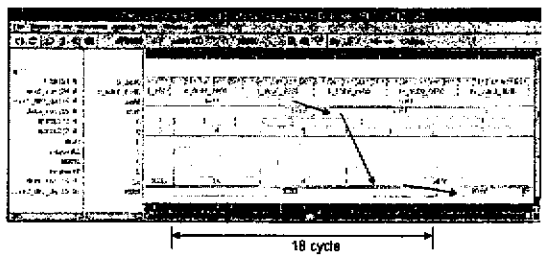
在原本的 NCU_DSP 中，透過輸出入介面，提供了三種模式，茲說明如下[7]:

1. 直接記憶體存取模式(DMA mode)，直接把外部資料藉由 data_in 埠直接存取到資料記憶體內，可以一次寫入多筆資料，由於處理器對於資料記憶體的優先權較直接記憶體存取模式低，所以在做此類存取時，處理器是不能做運算的，所以通常用於設定個別處理器一開始的資料及所需執行的程式碼。
2. 交握式傳輸模式(handshaking mode)，適用於傳輸時間不固定的一筆資料，存取到主控端記憶體中，可由指令集中的 IN、OUT 再傳到資料記憶體做運算，由於並未對資料記憶體動作，所以處理器仍可做運算。
3. 合併模式(merge mode)，結合以上兩種傳輸模式，可以一次寫入多筆資料，存取到主控端記憶體中，處理器的運算不需中斷。

三種模式的流程圖如圖二所示，而利用輸出入介面把資料 20FF 由 core1_HPI_QA 傳遞至 core2_HPI_QA 的波形如圖三所示，可發覺在利用輸出入介面傳遞資料時，必須要依序設定使用的模式、存取的位置等資訊，最後才傳遞資料，造成多餘的時脈浪費，使得多顆處理器所能帶來的效能改進被傳遞資料拖累，整體效率大打折扣。



圖二 輸出入模式流程圖



圖三 輸出入介面資料傳遞波形圖

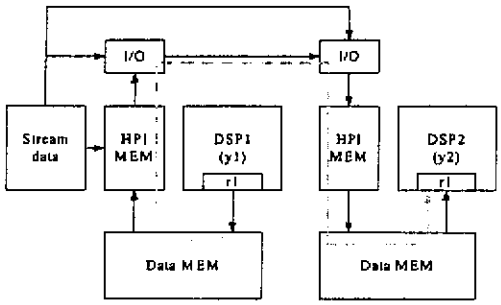
(一)解決方案

為了改進原本的資料傳遞模式，提出了河流式輸出入介面，利用累加暫存器連線以及記憶體連線，縮短資料傳遞所需時脈，減少不必要的傳遞路徑。

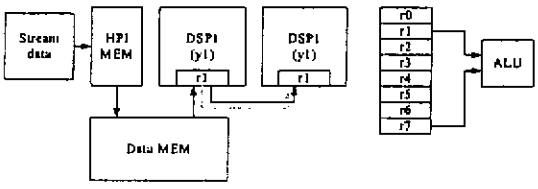
1.1 累加暫存器連線

在 NCU_DSP 中，有許多指令是直接對累加暫存器做讀取，把得到的值放入運算單元做處理後，再回存至累加暫存器中，所以如果能將累加暫存器內的值直接傳遞至別的

處理器中，對於需要大量運算的應用，可以節省不少的時間。可是在原本的系統架構上，如圖四所示，要把第一個累加暫存器的內容傳到下一顆處理器，要藉由指令先存到資料記憶體，再存至主控端記憶體，接著透過輸出入介面寫入下一顆處理器的主控端記憶體之中，最後以指令透過資料記憶體寫入累加暫存器之內，如圖四虛線路徑所示，經過的路徑實在太長。如果將累加暫存器陣列中選一個作為溝通的橋樑，當處理器對該暫存器寫入時，並不把值寫入該處理器的暫存器，而是將該值寫入至下一顆處理器的暫存器中，而下一顆處理器再對該暫存器做讀取的動作，便可完成累加暫存器傳遞資料的動作。其使用方式如圖五所示，r1 內為要傳遞至 DSP2 的資料，r7 是暫存器橋樑，當執行完 SHIFT0 r1,r7,0 後，r1 的值寫入到 DSP2 的 r7 之中，DSP2 便可直接執行 ADD2 r7,0,r1 的指令，將上一級 r1 的值與本身的 r1 作相加的動作，完成資料的傳遞。由於將資料寫入累加暫存器前，運算單元已經造成一定的延遲了，為了避免該延遲導致寫入下一級累加暫存器失敗，還需額外加一級暫存器作區隔，所以使用累加暫存器做為河流式輸出入是會需要多一個時脈的延遲。



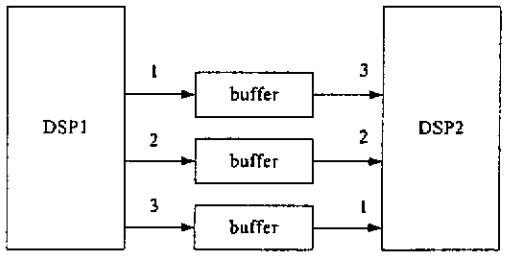
圖四 兩顆傳統處理器間資料傳遞區塊圖



圖五 提出之新型處理器間累加暫存器連線架構圖

1.2 記憶體連線

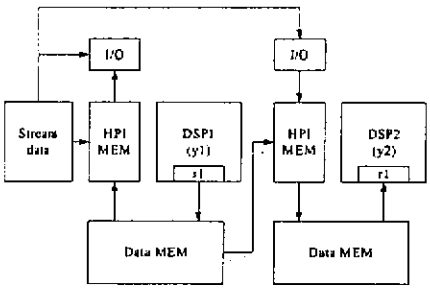
在某些應用上，處理器間的資料並不是依序傳遞，如圖六所示，DSP1 的輸出順序為 1、2、3，但對於 DSP2 而言，所需的順序剛好是相反的，在此情況下直接相連是不可行的。為了解決此問題，需要在連接通道上增加緩衝區，存放先到的資料。在河流式輸出入介面中，此緩衝區以主控端記憶體負責，其架構如圖七虛線所示，增加一條連線把資料記憶體與下一級的主控端記憶體相連接，完成資料的傳遞。



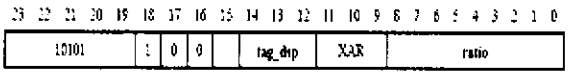
圖六 兩個 DSP 間資料做非循序傳遞

為了完成以上所提的動作，原本設計只用來對自身主控端記憶體做儲存的 OUT 指令，必須做出相對應的更動，來配合新的傳輸方式，其語法如圖八所示。由於記憶體連線並不只限於一對一的連接，也就是下一級的處理器不只一個，為了能區分出傳遞的目標，在指令中利用三個位元來指定要寫入的處理器編號(tag_dsp)，對於下一級的處理器而言，在啟動時也需要個別指定本身的處理器編號(dsp_id)，當接收到前一級傳來的資料時，如果處理器編號不符合，便不做儲存的動作。除此之外，將處理器編號等於 0 定義為不使用記憶體連線傳輸模式，此時輸出入控制單元的行為模式跟原本相同。

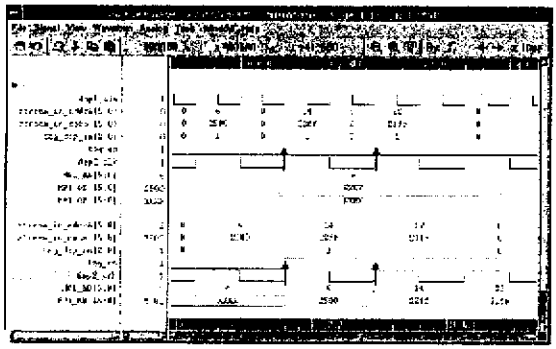
另外一個增加的欄位 ratio 是用來調整輸出時所需延遲的時脈數，由於處理器之間的速度並不一定相同，當資料由高速處理器傳給低速處理器時，可能會發生圖九中上半部的情形：對 DSP2 而言，由 DSP1 傳來的資料剛好位於時脈負緣的位置，剛好無法抓取。雖然可以在程式碼中調整 OUT 指令的位置來避開此問題，但如果利用 ratio 這欄位設為 1，將 DSP1 的輸出每筆多延遲一個時脈週期，如圖九中下半部所示，可使 DSP2 抓到正確的資料，完成傳輸，如此對於程式的寫作上會更加的方便。



圖七提出之新型處理器間記憶體連線架構圖



圖八 OUT 指令格式



圖九 資料延遲對不同速度處理器間溝通的影響

(二)無線網路在多重(Multi-DSP)DSP 架構之評估：

利用多顆 NCU_DSP 的計算能力，以及河流式輸出入介面有效率的連結，可以完成從前用單顆處理器做不完的工作。在此處以一個載波相位追蹤系統(carrier phase tracking system)來展現其能力。

在一個正交分頻多工(OFDM)通訊系統中，傳送端與接收端往往會有頻率上的誤差。在 802.11a 的系統中，利用已知的訓練信號配合最大近似法(maximum likelihood)即可估計此誤差並加以修正[8]，但其修正結果往往會剩下一些十分微小的錯誤。隨著時間的流逝，錯誤累積而造成星座圖旋轉的效應越來越明顯，最終旋轉量超過邊際值，導致接下來的符元(symbol)全部解錯。為了解決此問題，利用系統中已知的指導信號($P_{n,k}$)、接收端所收到的資料($R'_{n,k}$)以及傳輸通道的頻率響應(H_k)，可將估計出來的錯誤表示為公式 1[7]

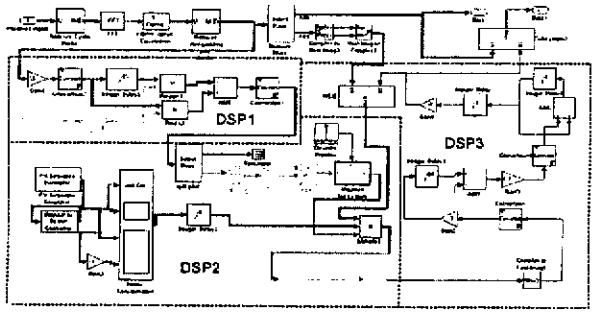
$$\phi = \angle\left[\sum_{k=1}^{N_c} R'_{n,k}(\hat{H}_k P_{n,k})^*\right]$$
$$= \angle\left[\sum_{k=1}^{N_c} H_k P_{n,k} e^{j2\pi\varepsilon f}(\hat{H}_k P_{n,k})^*\right]$$
$$= \angle\left[\sum_{k=1}^{N_c} |H_k|^2 |P_{n,k}|^2 e^{j2\pi\varepsilon f}\right] \quad (1)$$

其中頻率響應利用已知的訓練信號(X_k)以及接收端所收到的兩筆資料(R_1、R_2)由公式 2 求得，其中 X^*_k 為訓練信號取共軛數。

$$\hat{H}_k = \frac{1}{2}(R_{1,k} + R_{2,k})X^*_k \quad (2)$$

得知估計出來的錯誤之後，透過一鎖相迴路以及解相位旋轉器(derotator)將錯誤角度鎖定並加以反轉，就可以解決錯誤累積的問題。

系統架構確定後，以 matlab 做系統模擬並依據所需運算量切割成三塊，分別給三顆處理器去做運算，如圖十所示。DSP1 負責計算傳輸通道的頻率響應，利用內建記憶體以及乘加器來實現公式 2。由於此部分對於 52 個頻道都需要做計算，所以評估後將其操作速度設在 110MHz。在 DSP2 的區塊中只需對頻道中屬於指導信號的部分共 4 個頻道做運算，但由於還需要求取完美的指導信號以及乘上由 DSP3 所得的補償值，所以將其操作速度設在 55MHz。DSP3 負責計算迴路濾波器(loop filter)的輸出並利用數位控制震盪器(NCO)計算需要補償回來的旋轉量。



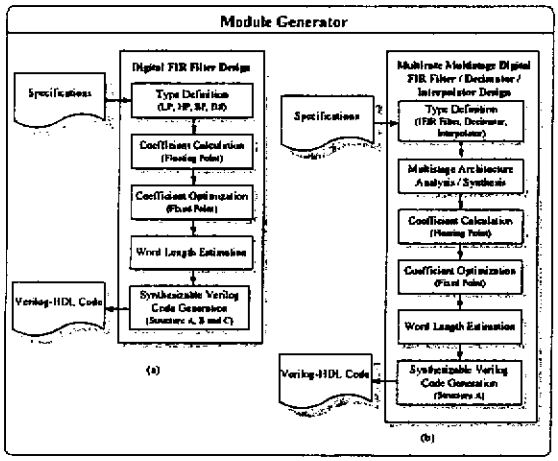
圖十 頻率誤差回復系統架構圖

四. 數位有限脈衝響應多階多速率之降頻器設計與成果討論

近年來 Digital signal processing(DSP)在電子產品上的應用是非常廣泛的，如多媒體技術，數據機以及手機等個人通訊設備。對於所有的電子產品而言，電路的低複雜度是電路設計的目標，主要的作用在於降低成本。而對於攜帶式的電子產品（如 notebook,手機等）則需減少 power 的

消耗。因此一個低複雜度及低功率消耗的設計是非常重要的。

在本計劃中，我們實現了一個數位有限脈衝響應多階多速率之濾波器/降頻器/升頻器模組產生器。使用者能夠藉由此產生器，自動設計出高速低複雜度的數位有限脈衝響應多階多速率濾波器。此產生器利用線性相位濾波器的對稱性架構，並運用多階多速率 IFIR [9]濾波器的方法，以達到低複雜度之目的。運用 polyphase representation [10] 將濾波器分解成多個子濾波器。所產生的濾波器，利用 CSD 乘法器、transposed direct 式架構、和 CSA 以達到高速的要求。此產生器只需要系統規格為了擁有良好的適應性，輸出的程式碼將以可合成的行為階層硬體描述語言撰寫，讓合成工具軟體能依據使用者所指定的條件選擇最適合的架構。圖十一，為整個模組產生器之流程圖。



圖十一 模組產生器之流程圖

表格一 輸入信號之系統規格

Filter Type (LP, HP, BP, BS, Decimator, Interpolator)	
Normalized Passband and Stopband Edge Frequencies	ω_P , ω_S
Passband Ripple in Magnitude or dB	δ_P / A_P
Stopband Attenuation in Magnitude or dB	δ_S / A_S
Input and/or Output Data Word Length (bit)	W_{in} / W_{out}
Signal to Noise Ratio (dB)	SNR
Up Conversion Ratio	L
Down Conversion Ratio	M

在此模組中，輸入信號為 system-level specifications 並且列舉在表格一中。此外，使用 symmetric transposed direct form 作為 Filter 之架構並搭配 MSB Fixed 技巧進一步加快速度效能。圖十二所示數位多階多速率降頻器,主要採用之技巧有:

(1)IFIR:我們使用多級的IFIR技術並且跟隨取樣率轉換率(SRCR)的關係：$M1 \geq M2 \geq \ldots \geq Mk$把decimator分解成可能的級數，有效的減小硬體的複雜度。

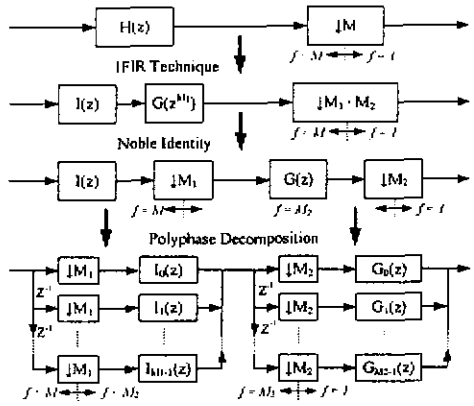
(2)Noble identity:將periodic model filter $G(z^L)$往低頻移動去除頻率之次方項，有效減小硬體的複雜度及降低功率消耗。

(3)Polyphase representation:將一個filter拆解成多個sub-filters，可以在較低頻率下運作且有效避免不需要的運算，有效的降低功率消耗。

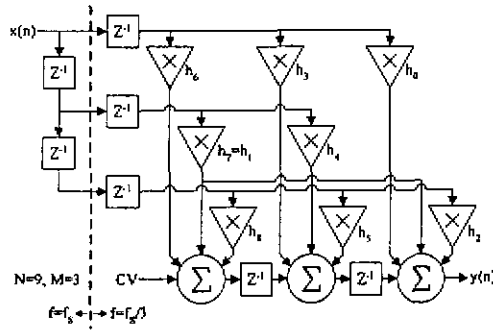
在最後合成電路，採用 Transposed direct form decimator。從圖十三，可以看到 Transposed direct form decimator 具有 memory saving 的特性減少對 register 的需要，進一步改善硬體複雜度。除此之外，模組產生器會產生 3 種架構之 synthesizable verilog。第一種為可合成的行為階層硬體描述語言撰寫，讓合成工具軟體能依據使用者所指定的條件選擇最適合的架構。此外，第二種為配合 CSA 的使用有效的改善整體架構之操作速度，第三種則為第二種架構在進一步做 pipelined 設計以達到更高速的要求。架構二與架構三如圖十四所示。

我們提供了一個用於 64-QAM 基頻解調器的濾波器設計實例。使用 Synopsys 的合成工具並採用 TSMC 0.25μm 製程設計晶片。結果在低複雜度應用方面，減少了 1.64 倍的面積並節省 1.95 倍的功率消耗(表格二)。而對於高速應用方面，此晶片能操作在 714 MHz。除此之外，我們以此模組產生器設計一個 CDMA 規格之多階濾波器(IFIR filters)的例子(如圖十五與表格三)，與傳統濾波器設計相比較，結果減少了 1.72 倍的硬體面積，節省 13.10 倍的功率消耗。並且以相同之規格設計一個多階多速率之降頻器，用運 IFIR 與 polyphase representation 之設計技巧，與傳統的降頻器設計相比較，結果減少了 3.56 倍的硬體面積，節省 1.96 倍的功率消耗。最後，我們還實現一個窄頻的多階多速率之升頻器，與傳統的升頻器設計相比較，結果

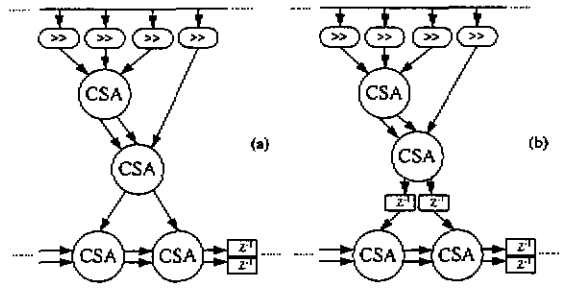
減少了 3.06 倍的硬體面積，節省 1.36 倍的功率消耗。



圖十二 數位多階多速率降頻器設計



圖十三 Transposed direct form 之降頻擁有節省記憶體之使用效果



圖十四 (a)以 CSA 之設計架構(b)CSA 搭配 pipelined 之設計架構

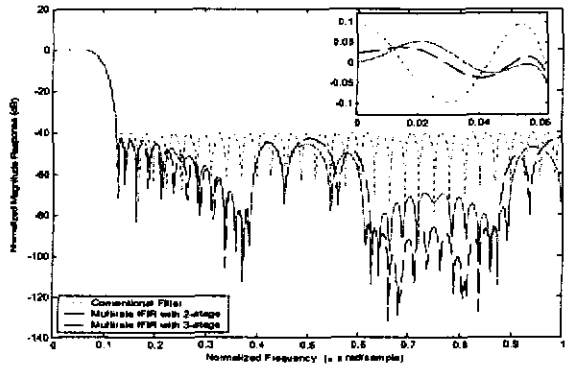
表格二 64-QAM 基頻解調器設計實例之合成結果 .

	Work#1	Work#2	[11]
Technology(um)	0.25	0.25	0.25
Max. Operating Frequency	714 MHz	146 MHz	72 MHz
Total Gate Count	11117	5155	8477
Combination Area	5011	2496	5938
Noncombination Area	6106	2659	2539

Work#1 Specifications under high-speed constraint.
Work#2 Specifications under low-complexity constraint.

表格三 傳統與多速率降頻器之比較 [12] (降頻倍率, M=8)

Technology: TSMC 0.25um Input Frequency: 200 MHz	Conventional Decimator (M=8)	Multirate Decimator with two-stage	
		Stage#1 (M1=4)	Stage# 2 (M2=2)
Total Gate Count	13742	1580	2554
Combination Area	12567	1162	1792
Noncombination Area	1174	418	761
Technology: TSMC 0.25um Input Frequency: 200 MHz	Multirate Decimator with three-stage		
	Stage#1 (M1=2)	Stage#2 (M2=2)	Stage# 3 (M1=2)
Total Gate Count	638	808	2417
Combination Area	336	496	1706
Noncombination Area	301	312	710



圖十五 傳統與多階多速率實現之頻率響應

五、成果

(一)第二版適合通訊系統之 Embedded DSP core 設計完成晶片。

(二)串流式架構(Stream-Based) DSP。

(三)IP 化 DSP 核心。

(四) 一篇 SASIMI2004 Conference Paper。

(五)兩篇 APCCAS2004 Conference Paper。

(六) 一篇 SASIMI2004 Conference Paper。

六、 參考文獻

[1] C.M. Moerman, R. Woudsma, P. Kievits, "*Embedded DSP Technologies In Consumer Applications*," DSP World workshops, September 1998

[2] I. VERBAUWHEDE and M. TOURIGUIAN, "*A Low Power DSP Engine for Wireless Communications*," Journal of VLSI Signal Processing, No.18, pp.177-186, 1998.

[3] B.-W. Kim, J.-H. Yang, C.-S. Hwang, Y.-S. Kwon, K.-M. Lee, I.-H. Kim,

Y.-H. Lee, C.-M. Kyung, "*MDSP-II: A16-Bit DSP with Mobil Communication Accelerator*," IEEE 1998 Custom Integrated Circuits Conference, pp. 2.1.1-2.1.4, 1998.

[4] M. Kuulusa, J. Nurmi, J. Takala, P. Ojala, H. Herranen, "*A Flexible DSP Core for Embedded Systems*," IEEE Design & Test of Computers, Vol. 14, NO. 4, pp.60-68, Oct.-Dec., 1997.

[5] H. Yang, B.-W. Kim, S.-W. Seo, S.-J. Nam, C.-H. Ryu, J.-H. Cho and C.-M. Kyung ," *MetaCore A configurable & Instruction-Level Extensible DSP Core* " , ASP-DAC'98, pp. 325 - 326, Feb. 1998

[6] P. Paulin, C. Liem, M. Cornero, F. Nacaba, and G. Gossens, "*Embedded software in real-time signal processing systems: Application and architecture trends*," Proc. of the IEEE, vol. 85, no. 3, pp. 419-435, Mar. 1997.

[7] Y.T. Chen, "*Embedded DSP Module generators for Communication System*," MS thesis, Dep. Elec. Eng., National Central University, Taiwan, June, 2001

[8] IP Qualification Alliance, "*IP Qualification Guidelines 2003 v1.0*," Industrial Technology Research Institute, Dec. 19, 2003

[9] Y. Neuvo, C. Y. Dong, and S. K. Mitra, "Interpolated finite impulse response filters," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 563-570, June 1984.

[10] M. Bellanger, G. Bonnerot, and M. Coudreuse, "Digital filtering by polyphase network: application to sample rate alteration and filter banks," IEEE Trans. ASSAP, vol. ASSP-24, pp. 109-114, April 1976.

[11] S. J. Jou, C. H. Kuo, M. T. Shiau, J. Y. Heh and C. K. Wang, "VLSI implementation of timing recovery and carrier recovery for QAM/VSB dual mode," International Symp. on VLSI Technology, Systems and Applications, Taipei, R. O. C. June 1999, pp.159-162.

[12] "The CDMA network engineering handbook, volume 1: concepts in CDMA," Qualcomm Inc., March 1993.

[13] K.Y. Jeng, S.J. Jou and A.Y. Wu, "A Design Flow for Multiplierless Linear-Phase FIR Filters: From System Specification to Verilog Code," IEEE Circuits and Systems, 2004, ISCAS '04.

◆ 子計劃五：射頻和混合訊號系統晶片之測試技術(蘇朝琴)

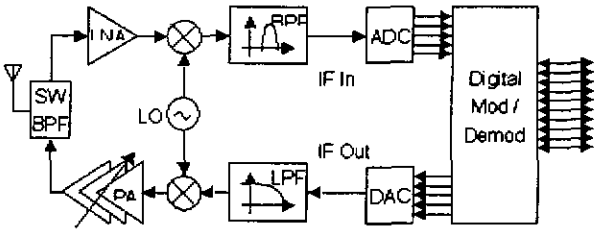
一、簡介

推導和實作一種利用產生以及接收中頻信號來測試與偵測混合信號電路中之中頻與射頻電路模組的方法。利用數位信號處理器產生測試訊號，經由數位類比轉換器產生較低頻的中頻測試訊號輸入至待測的發射器之中頻與射頻模組。在接收端，利用類比數位轉換器擷取波形，接著傳送到數位信號處理器作分析。利用這個方法可以大幅節省測試成本，因為所需之各項元件均已內建在原本之電路中。且我們不需要額外的射頻自動化測試設備。

二、計畫的緣由與目的

無線通信在今日已經蔚為風氣，在行動通訊百家齊放，有 GSM、DECT、PHS、GPRS、WCDMA、CDMA2000 等不同的系統問世。在無線網路上，亦有 Bluetooth、802.11、等不同的規範與應用。由於有非常大量的需求與高附加價值，全球的設計業者，晶圓廠，與系統業者，無一不傾全力往此一方向發展，並已獲致相當可觀的成就與成果。如今，頻電路的應用，已經非常的廣泛與普及。更進一步，在有線傳輸上，光纖與高速傳輸也都屬於射頻電路的應用範疇。無線通訊的設備越來越普及，可謂是後 PC 時代最重要的產業。雖然射頻電路的設計、製作、與應用日趨成熟，然而，射頻電路的測試並無法如其他技術般如此順利的發展。為了解射頻測試的複雜度，我們需要先來看一般射頻電路系統的架構。

在發射端，數位信號輸入 Digital_In 經過數位調變 Digital_Mod 後成為數位中頻信號，再經由數位類比轉換器 DAC (Digital Analog Converter)將之轉換成類比中頻信號，然後再經射頻升頻 RF Up Convert 成為射頻信號，最後再以功率放大器 PA (Power Amplifier) 將之放大後送至天線傳送出去。在接收部分則走相反但卻相似的路徑。射頻信號由天線接收後，先經低雜訊放大器 LNA (Low Noise Amplifier) 加以放大，再經由射頻降頻 RF Down Convert 至中頻類比信號，再由類比數位轉換器 ADC (Analog Digital Converter) 轉換成數位中頻信號，再經數位解調 Digital Demodulation 成數位輸出信號。以上說明僅以射頻電路與中頻數位調變電路為例子，基頻調變電路也有相似的系統，惟 ADC/DAC 的位置不同。而數位輸出與輸入可以為經過編碼的影像語音信號，也可以是直接的數位資料 (Digital Data)。



圖一、射頻與混合信號電路架構示意圖

測試如上圖之無線通訊電路測試的困難如下。首先，此一電路包含有不同的電路組合。射頻電路、類比電路、混合信號電路、數位電路各有其設計、製程技術、與測試技術。僅就測試技術而言，不僅錯誤模型、測試信號、測試設備、就連測試工程師所需要的訓練都不同。面對此一電路，資深的工程師都會有相當的困難度，更遑論一般的初級工程師。就算測試的技術、設備、人才得以解決，就測試經濟層面而言，此一電路也會遭遇到一些問題。無線通訊已經走向低價大眾化的不歸路，為了因應一般的消會大眾，手機售價越來越低。手機的成本勢必要有相當程度的壓低，才能因應一元或免費手機的市場。然而，測試的成本要壓低到足以承受如此複雜的測試技術將會相當的困難。因此測試技術的突破的需求是相當急迫的。現在我們先來看一下測試的技術。

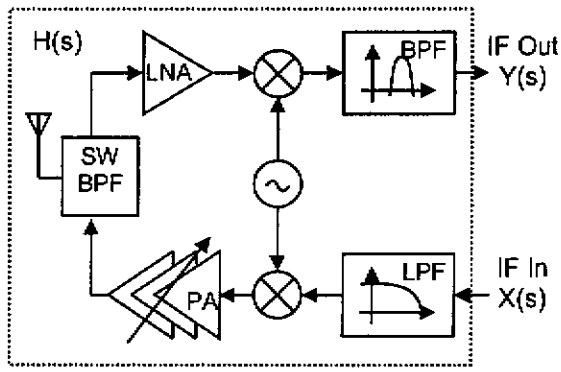
傳統上射頻電路的測試以兩類方式為之。第一類型是以儀表量測的方式為之[3, 4,

5]，利用射頻訊號專用的精密儀表送出或接收射頻訊號，並利用儀表的內建功能，量測電路的參數，以決定電路的正確與否。此一類型的最大優點在於能準確的量測電路特性，正確的判斷電路的好壞，誤放 Miss 與誤殺 Kill 的機率較低，通過測試的電路的可靠性 Reliability 較佳。但是此一方法的缺點是，測試成本非常的高，似乎無法滿足低價手機與無線網路的要求。第二類型則是利用迴路 Loop Back 的方式[1, 2]將傳送出的訊號接回並與予解碼，在檢查回收資料的位元錯誤率 Bit Error Rate (BER) ，如下圖所示。此一方法與第一類型剛好相反，它的優點在於測試的成本非常的低。由於測試的資料長度往往要十倍甚至百倍於 BER(10-7~10-12)的倒數，因此它的缺點則是測試時間非常的長。再者，位元錯誤率雖受到電路參數的影響，但是其關係並非線性，雜訊的大小也對位元錯誤率有同等的重要性影響。因此，並無法由位元錯誤率來推算電路參數，因此誤放 Miss 與誤殺 False Kill 的機率相當的大。此兩類型的測試對射頻電路都有力有未逮之處。

如果我們仔細的觀察並評估圖一中所示之 RF 模組，我們不難發現在整體架構上，類比數位轉換 AD/DA 模組數位信號處理 DSP 模組為可用之測試資源 Test Resources。基於此一觀察，本計畫要提出的是利用現今無線電路架構中，內建之類比數位轉換單元與數位信號處理單元作為設頻電路測試的基礎。在測試的架構上，我們將利用 DSP 模組產生數位測試信號，經由 DAC 轉換送出中頻信號，再由射頻發射電路將之升頻至射頻，經過射頻接收模組將之接收並降頻至中頻，再經 ADC 將之轉換成為數位信號，送至 DSP 模組作數位信號處理。在數位信號處理技術上，我們將使用吾人發展的自身反應 Intrinsic Response 測試方法[1-8]，在不受其他模組的影響下，將每一模組的測試自身反應加以分離淬取出來，以分別判定每個模組的好壞優劣。就測試複雜度而言，它使用了內建的 AD/DA 與 DSP 模組，就測試成本而言是最為經濟的，他不必使用外掛的射頻測試機台。

三、研究方法及成果

在先行可行性探討上，我們先用簡單的例子，來測試我們的方法。在整個環境上，IF 的輸出與輸入可以由圖二所表示。



圖二、射頻測試概念圖

DAC 送出 x(t)與 ADC 接收 y(t)，在時域做 Convolution

$$y(t) = x(t)*h(t)$$

等於頻域相乘

$$Y(s) = X(s)\,H(s)$$

$$H(s) = Y(s) / X(s)$$

$$H(s)_{db} = Y(s)_{db} - X(s)_{db} \text{ (log scale)}$$

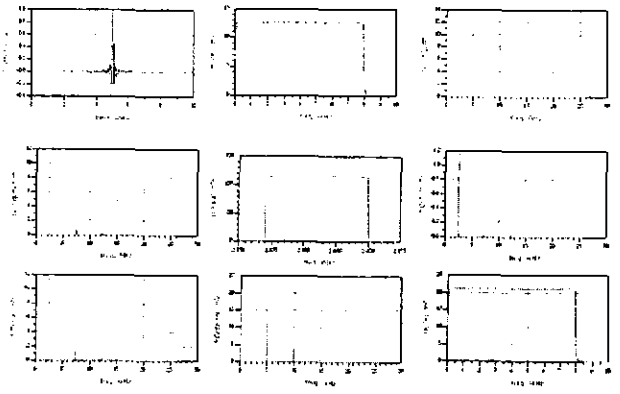
經由後端 DSP 的 FFT 運算，得到 X(s)與 Y(s)，而求得 RF 模組之轉移函數 H(s)，進而分析。

由於在 RF 系統中 DAC 與 ADC 均相當的高速。整體的系統也與前面獨立模組測試不盡相同。一般而言，測試信號可以有三種選擇，Single Tone、Multiple Tone、Chirp 與 Sinc 信號。我們先採用因此我們先採 Sinc 信號作為測試信號，取其平整之頻譜。Sinc 的時域信號與頻譜。會影響產生出來 IF 測試信號品質的因素有下列幾項：取樣頻率、解析度、線性度以及取樣的點數。為

射頻前端電路模擬

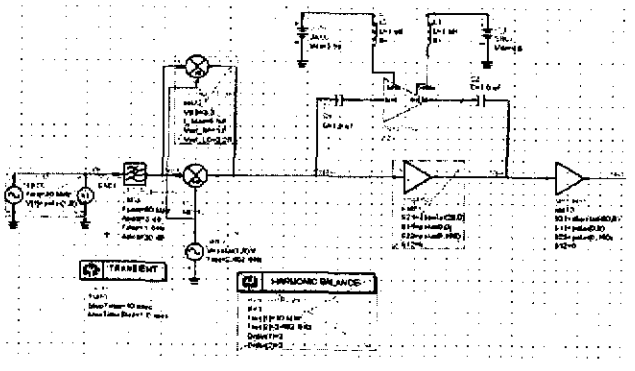
利用數位類比轉換器產生中頻訊號，接收到的訊號最後由類比睡位轉換器送交數位訊號處理器分析。圖七利用 Agilent ADS 模擬

的整個電路架構。電壓源 V_{in} 產生想要的類比中頻信號，然後再經射頻升頻 RF Up Convert 成為射頻信號，最後再以功率放大器 PA (Power Amplifier) 將之放大後送至天線傳送出去。在接收部分則走相反但卻相似的路徑。射頻信號由天線接收後，先經低雜訊放大器 LNA (Low Noise Amplifier) 加以放大，再經由射頻降頻 RF Down Convert 至中頻類比信號 V_{out}。圖八為理想的個節點電路模擬結果。



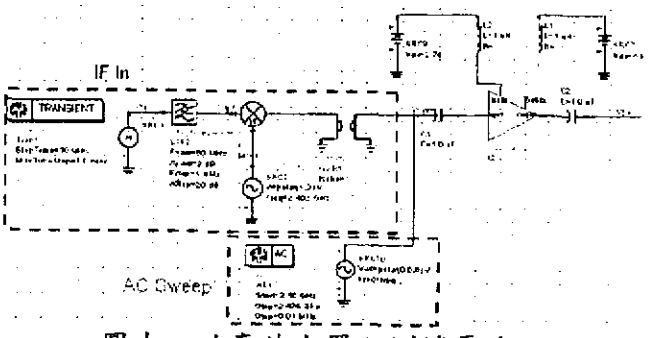
圖八、理想射頻前端電路模擬結果

當比較 V_{in} 與 V_{out} 可以確定出何項元件有不符規格或者是效能變差。為了驗證我們的方法，建構了理想的模型，行為式的模型，與真實電路的模型。圖九為理想模型與實際電路模型的電路途，可以利用這種架構測試與比較出異同。



圖九、理想與實際電路測試分析

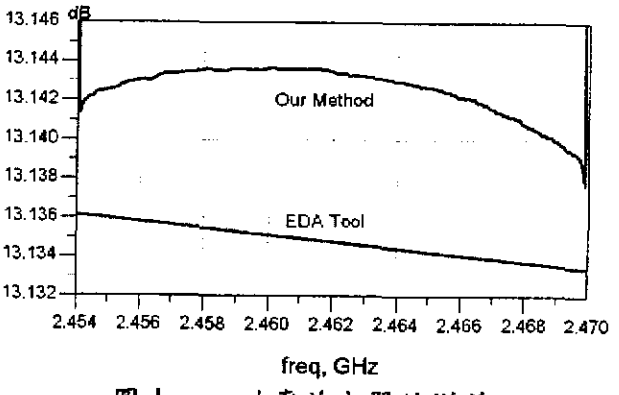
利用兩種不同的測試方法來產生訊號，EDA 軟體所提供的 AC 分析方式與暫態分析產生之 SINC 函數，然後轉換至頻域分析。測量功率放大器之電路圖如圖十所示。



圖十、功率放大器之測試電路

在射頻前端電路系統中有兩個或更多的放大器，在發射端，功率放大器是將訊號放大至天線送出的主要元件，在接收端，低雜訊放大器是將微弱的訊號放大至電路所能處理的範圍，雜訊指標與增益是重要的規格。

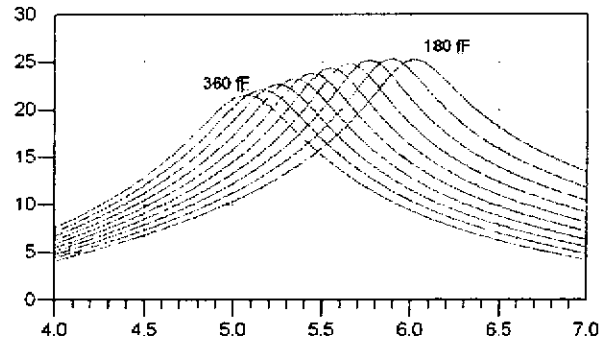
測試無線網路中非理想功率放大器的某些規格。首先，產生 8Mhz 頻寬的 sinc 波形然後利用理想的混波器升頻至 2.462 GHz。接著，量測功率放大器的輸出。最後接輸出與輸入的頻譜相減，得到功率放大器的增益。圖十一為比較此方法與利用 EDA 軟體頻域掃描的結果。兩者之間的誤差非常的微小，這是由於數值轉換的誤差所造成的。



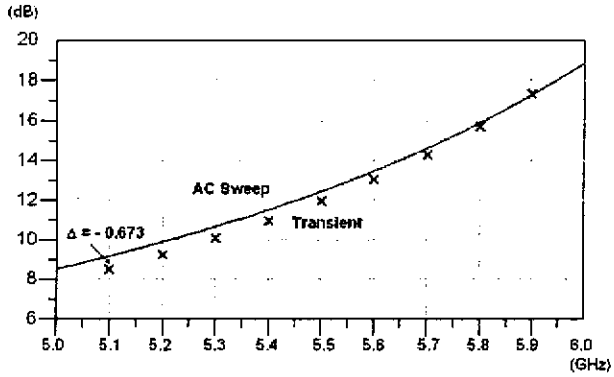
圖十一、功率放大器的增益

在接收端的低雜訊放大器的輸出阻抗會因為製成的飄移而造成高增益頻帶的範圍飄移，而產生出效能降低，甚至無法動作的結果，如圖十二所示。

利用我們的測試方法，可以得到如圖十三，於每個通道的頻率測試其增益，分析得到如頻譜的圖，用以確認其輸出等效阻抗飄動多少，是否符合規格。

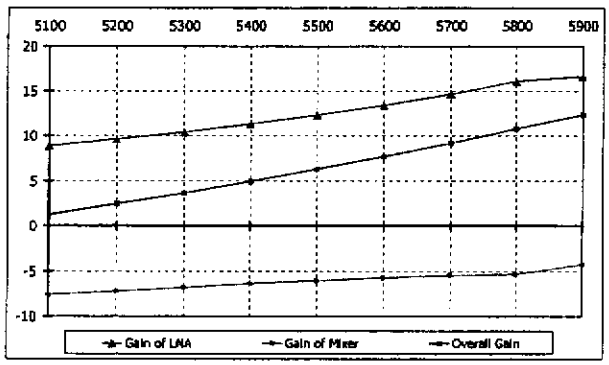


圖十二、低雜訊放大器輸出匹配造成之誤差



圖十三、低雜訊放大器的增益

整個接收器的測試包含了低雜訊放大器的增益與混波器的轉換增益，以及全部接收器的增益，如圖十四所示。



圖十四、整體接收端的增益分析

四、結論與討論

推導和實作一種利用產生以及接收中頻信號來測試與偵測混合信號電路中之中頻與射頻電路模組的方法。利用數位信號處理器產生測試訊號，經由數位類比轉換器產生較低頻的中頻測試訊號輸入至待測的發射器之中頻與射頻模組。在接收端，利用類比數位轉換器擷取波形，接著傳送到數位信號處理器作分析。

這個方法可以大幅節省測試成本，因為所需之各項元件均已內建在原本之電路中，且我們不需要額外的射頻自動化測試設備。

五、參考文獻

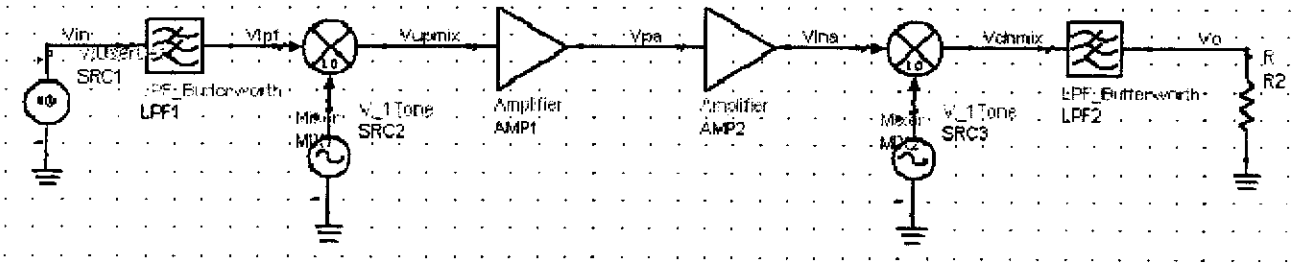
[1] Lupea, D.; Pursche, U.; Jentschel, H.-J., "RF-BIST: loopback spectral signature analysis," Design, Automation and Test in Europe Conference and Exhibition, 2003. pp. 478 – 483

[2] Heutmaker, M.S.; Le, D.K., „An architecture for self-test of a wireless communication system using sampled IQ modulation and boundary scan," Communications Magazine, IEEE , Volume: 37 , Issue: 6 , 1999 pp. 98 – 102

[3] Voorakaranam, R.; Cherubal, S.; Chatterjee, A., "A signature test framework for rapid production testing of RF circuits," Proc. Design, Automation and Test in Europe Conference and Exhibition, 2002. pp. 186 - 191

[4] Ferrario, J.; Wolf, R.; Moss, S., "Architecting millisecond test solutions for wireless phone RFICs, Proc. Test Conference, 2002. pp. 1151 - 1158"

[5] Kasten, J.S.; Kaminska, B., "An introduction to RF testing: device, method and system," Proc. VLSI Test Symposium, 1998. pp. 462 - 468



圖七、射頻前端電路架構圖

◆ 子計畫六：為 IEEE 802.11e WALNs 提供一服務品質模組 (廖婉君)

一. 簡介

由於IEEE 802.11已成為無線上網主流，而802.11e為提供各種應用程式不同服務品質之標準，但其支援的性質及程度有限，故本報告主要為802.11e WLANs提出一個新的提供服務品質系統模組，並介紹兩個不同方法以實作此系統模組。

二. 計畫緣由與目的：

802.11e[3]為提供各種應用程式不同服務品質之標準，在802.11e中，不同服務品質以一個新的存取方法，稱為HCF (Hybrid Coordination Function)來達成。在HCF中定義了兩種競爭存取方式，在這篇報告中，我們只針對其以競爭為基礎的媒介存取方式(EDCA)為主做討論。

802.11e EDCA主要針對不同存取種類(AC: Access Categories)提供不同優先順序的服務等級，一個無線網路用戶(station)可提供四個不同AC，分別為AC_VO, AC_VI, AC_BE, 及AC_BK。在802.11e裡，只有支援不同優先順序的服務，雖然可以根據不同的應用程式類別提供不同優先順序之服務，但因為其不是“絕對性”的優先順序，故此服務模組並不能對特定類別提供保證的服務品質；除此之外，原EDCA也不能提供以權重為比例分配無線網路頻寬，這對頻寬或資源共享是一個很重要且很實用的功能，因此對於802.11e EDCA而言，同時提供不同絕對性優先順序及依權重分配無線網路頻寬是非常重要的。

為了提供依權重分配頻寬，EDCA需要整合公平性的封包排程服務(fair service discipline)，在無線網路環境下提供此服務最大的挑戰在它完全是分散式環境，因此之前在有線網路環境下所提供的集中式排程法是不能直接於無線網路環境使用的。之前已有些文獻提出在802.11提供依權重分配網路頻寬的方法，包括Distributed Fair Queueing (DFS) [7], Priority-based fair Medium Access Control (P-MAC) [8] 及 Distributed Deficit Round Robin (DDRR) [9]。這些方法都是根據特定封包排程理論調整802.11 MAC的三個參數之一，分別為Backoff Interval (BI)，Inter-Frame Space (IFS) 及 Contention Window(CW)。

上述研究的目的都在提供不同無線用戶之間不同的服務品質，而非像EDCA針對不同AC間提供服務差別化。目前還沒有文獻志在提供AC間的服務差別化，Adaptive Fair Enhanced DCF (AFEDCF) [11]是唯一在同一AC間提供fairness的文獻，但它並非提供不同AC間的fairness，所以AFEDCF並不能提供解決的方法。

故這個計劃的目的在為IEEE 802.11e提供不同AC間不同的服務品質，在這個服務架構中，可以同時支援絕對性的優先順序及依權重分享頻寬。擁有最高優先順序AC的封包一定會最優先被傳送，接下來是次高優先順序AC的封包，而剩下的可用頻寬依AC的權重分配，也可用於只有絕對優先順序或依權重分享頻寬的網路政策上；故這是個非常彈性的資源分享架構，網路管理者可依其需要設定頻寬分享的策略。

三. BI-WF-SP 及 IDFQ-SP

在這章，我們提出兩個方法來支援此服務品質架構，這兩個方法皆可與原 802.11 標準相容。

A) 以 Backoff-interval 為基礎的排程服務

我們先描述 BIWF-SP (Backoff-Interval based Weighted Fair Scheduling with Strict Priority)。在 BIWF-SP 中，我們在每個 backoff entity 中使用 DFS 來提供依權重分配頻寬的服務，且透過一個準則來指定較小優先順序 AC(包括依權重分享頻寬的 AC)的參數來達到絕對的優先順序。如果使用 P-MAC 來決定每個 backoff entity 的 backoff interval 來達到

依權重分配頻寬的話，這個方法可以很簡單的延伸為以 Contention Window 為主的方法。

1) 依權重分享頻寬

為了達到這個服務，每個 AC 被指定一個正數，$\phi[AC]$，用來代表這個 AC 分享頻寬的比重，每個佇列的第一個封包的 backoff interval 由 DFS 的下列式子決定：

$$BI_i = \left\lfloor \left\lfloor Scaling\_Factor \times \frac{L_i}{\phi_i} \right\rfloor \times \rho \right\rfloor \quad (1)$$

，其中 L_i 是該封包的大小，ϕ_i 為其分配權重，ρ 為一個從 0.9 到 1.1 的平均分配變數，用以減少碰撞的機會，而 Scaling_Factor 用以選擇一個適當大小的 Backoff interval，以免降低整體運作的效率。

2) 絕對優先順序

為了提供絕對優先順序，優先順序較高的封包必須比優先順序較低的封包先傳送，雖然一般優先順序較高的封包的 backoff interval 較小，擁有較高的競爭優勢，但優先順序較小的封包，其 backoff interval 還是有減少到零的機會，這時就會違反絕對優先順多的情形，故我們必須利用 802.11e EDCA 能彈性設定每個 AC 的 AIFS 的特性，換句話說，較低優先順序的封包，其 AIFS 必須大於優先順序較高封包的 AIFS 加上其 AC 最大的 contention window，即：

$$AIFS[j] \geq AIFS[i] + CW_{\max[i]} \text{ if } j > i \quad (2)$$

總結來說，在 BIWF-SP 中，每個 AC 會搭配一個 backoff entity，還有一組 EDCA 的參數組，即 AIFS[AC]，CWmin[AC]，及 CWmax[AC]，但現在 AIFS[AC]是由(2)決定，不同優先順序 AC 的 backoff interval 的決定方式還是和 EDCA 相同，而依權重分享的 AC，其 backoff interval 則用(1)決定；另外，和 EDCA 不同的是，同一 station AC 的 backoff entity 都獨立運作，當同一 station 不同 AC 的 backoff interval 同時減少為 0 時，任何選擇的方法皆可。

我們必須注意的是：雖然 BIWF-SP 可以提供預期的服務模組，但它擁有較低效率的問題，即當較高優先順序的類別沒有封包要傳送時，較低優先順序仍需等待較大的 AIFS，如此會浪費無線網路頻寬。

B) 以 Inter-Frame Space 為基礎的排程服務

接下來，我們將提出一個以 IFS 為基礎的方法：IFS-based Distributed Fair Queueing with Strict Priority (IDFQ-SP)。IDFQ-SP 根據要傳送封包的 finish tag，選擇一個適當的 IFS，以致於在所有 backoff entity 中，擁有最小 finish tag 的封包將會被傳送。在這份報告中的模擬設定是根據 802.11b 的參數，但這個方法可以很容易的延伸到 802.11 其他版本，如 802.11 a/g。

1)依權重分享頻寬

為了提供依權重分享頻寬，我們以分散式的方法模仿 SCFQ：每個佇列的第一個封包都會附上一個 finish tag，然後將其 finish tag 轉換為 802.11 DCF 之 IFS，使得較小的 finish tag 被轉換為較小的 IFS；相對的，較大的 finish tag 將會有較大的 IFS，配合 IDFQ-SP 沒有 backoff process，故擁有最小 finish tag 的封包就會最先被傳送。除此之外，為了避免碰撞發生，我們將對應的 IFS 乘上一隨機變數。

在 IDFQ-SP 中，每個 backoff entity 會紀錄一個虛擬時間 $v_i(t)$，當 t=0 時，$v_i(0)=0$，每個要傳送出去的封包都會附上一個 finish tag，其決定的方式為：

$$S_i^k = \max\{v(A_i^k), F_i^{k-1}\}$$
$$F_i^k = S_i^k + \frac{L_i^k}{\phi_i} \quad (3)$$

故每個 backoff entity 在傳送或接收到封包時，可以根據此封包之 finish tag 來更新其虛擬變數：

$$v_i(t) = \max\{v_i(t), F\} \quad (4)$$

如此一來，我們就可以保證 backoff entity 內的虛擬時間將不會減少，即使因為在決定封包 IFS 時的隨機行為，使得 finish tag 較大的封包仍有可能被傳送。並且，因為所有的無線用戶都連接到同一個 AP，故系統內的所有 backoff entity 的虛擬時間應該都是一樣的。

接下來，我們介紹將 finish tag 轉換為 IFS 的方法。假設 F_i 為一 backoff entity 第一個封

包的 finish tag，L_{max} 及 ϕ_{min} 為系統內最大的封包長度及最小的分配權重，則一 backoff entity i 的 IFS 為：

$$IFS_i = \begin{cases} 40(\mu s) + \dfrac{random(0,1)}{\alpha} \times \beta, & F_i - v_i(t) = 0 \\ 40(\mu s) + \dfrac{F_i - v_i(t)}{\alpha} \times \beta, & otherwise \end{cases} \quad (5)$$

在這個式子中，$\alpha = L_{max} / \phi_{min}$ 為 $F_i - v_i(t)$ 的最大值，而 β 為大於 0 的 random number 用來減少 collision 的機率。

雖然擁有最小 finish tag 的封包通常會被傳送，但因為我們使用一個 random number 來避免 collision，使得一個擁有較大 finish tag 的封包有可能會被傳送，因此，$F_i - v_i(t)$ 的值有可能會小於 0；但，即使如此，$|F_i - v_i(t)|$ 還是會小於或等於 α，因此，$(F_i - v_i(t))/\alpha$ 的值還是會在-1 和 1 之間，故如果 β 平均分配在 0 到 10 之間的話，每個 backoff entity 的 IFS 就會介在 PIFS 和 DIFS 之間，如此一來就可保證 IDFQ-SP 可向後相容於 IEEE 802.11。

2)絕對優先順序

為了提供絕對優先順序，由 PIFS 到 DIFS 間，不重複且較小的 IFS 區段將被保留給較高優先順序的 AC。這些區段必須比依權重分配的 AC 小，必且不能和其重疊，為了達到這個目的，我們將(5)的 β 的範圍限制在[0,5]，並分配 AC[0] 及 AC[1] 的 IFS 分別為 $PIFS + uniform(0,2.5)$ 及 $PIFS + uniform(2.5,5)$。

總結來說，IDFQ-SP 沒有 backoff process，不同優先順序 AC 的 AIFS 為 PIFS 加上一不重疊的區段，而依權重分配的 AIFS 則由(5)決定；此外，系統的虛擬時間只依權重分配的 AC 需要使用，也就是說不同優先順序的 AC 不需要將其 finish tag 放在其封包中，而其他 backoff entity 聽到此類封包也不用更新其系統虛擬時間，以維持整個系統的正確性。

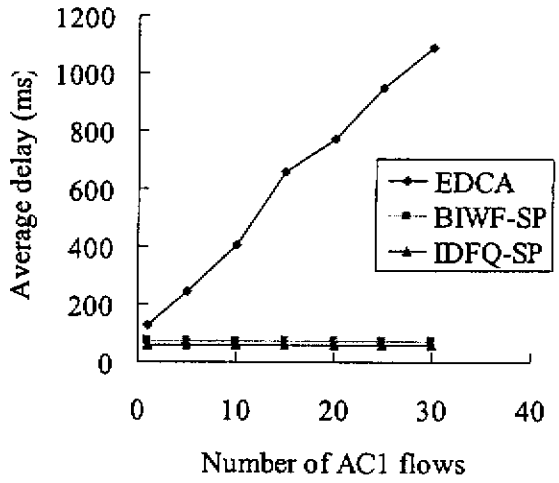
四. 系統表現評估

在這個章節，我們使用由 Chesson 及 Singla [13]所研發的 ns-2 EDCA 模組來評估我們所提出的兩個方法。在這個實驗環境中，無線網路的速度為 11Mbps，且每個 AC 傳送出來的網路流量為固定 8Mbps 的 CBR UDP 封包，其封包長度為 1500bytes。AC[0] 及 AC[1]為不同優先順序的 AC，且 AC[0]的優先順多大於 AC[1]，而 AC[2]及 AC[3]為依權重分配頻寬的 AC，且其權重比例為 2:1。使用 DFS 的參數 Scaling_Factor 為 0.01，且 EDCA 4 個 AC 的參數組列於表 1。

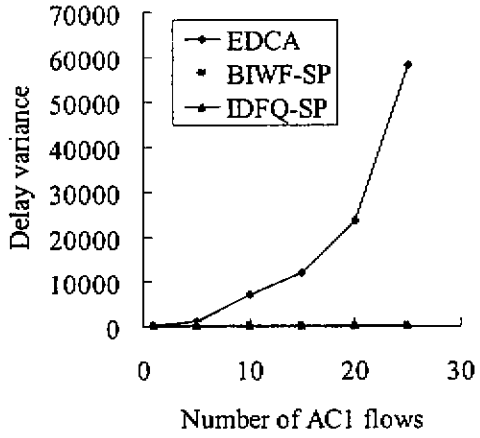
我們首先顯示 strict priority 的重要性及其效果。當屬於 AC[1]的連線越來越多時，圖一畫出了 AC[0]連線受到的平均延遲及延遲變異數。圖中顯示原來 EDCA 對較低優先順序的連線數非常敏感，相反的，我們提出的 BIWF-SP 及 IDFQ-SP 則維持一貫，不受較低優先順序連線多寡的影響，這顯示我們所提出的方法非常適合對延遲非常敏感的應用程式，如 VoIP。

Table 1. EDCA parameter sets used in the simulations

	AC_VO	AC_VI	AC_BE	AC_BK
CWmin	3	7	15	15
CWmax	7	15	1023	1023



(a) Average delay of AC0



(b) Delay variance of AC0

Figure 1. The delay and inter-arrival time of flows

接下來，我們比較 EDCA，BIWF-SP 及 IDFQ-SP 在同時提供絕對優先順序及依權重分享頻寬時，其 throughput 的表現。我們考慮以下三種情境：第一種為四個 AC 在同一 station 內；第二種情境為四個不同 AC 分散在不同無線用戶端中；第三種情境代表並非所有 AC 都在不同用戶端內，但一個用戶端內擁有多於一個 AC 在。在這個實驗中，每個 AC 的 sending rate 皆為 8Mbps，為了顯示不同優先順序的效果，我們在時間 15 秒，20 秒及 25 秒時，將 AC[0]的 sending rate 各改為 5Mbps，4Mbps 及 3Mbps；此外，將 AC[1] 的 sending rate 在 20 秒及 25 秒時，改為 2Mbps 及 1Mbps。因為三個情境的實驗結果都很相近，故我們只顯示第一個情境的結果，其結果如圖三。

圖三畫出原來 EDCA，BIWF-SP 及 IDFQ-SP 的 throughput。由圖所示，原來 EDCA 只提供不同優先順序的服務，但並不是絕對的優先順序，故 AC[1]的流量會降低 AC[0] 的表現，且 AC[2]和 AC[3]也沒有辦法依權重分配頻寬。相反的，BIWF-SP 及 IDFQ-SP 能同時提供絕對的優先順序及依權重比例分配頻寬，但若比較它們的結果，發現 IDFQ-SP 能提供較平穩的服務，且其 aggregate throughput 也較高。

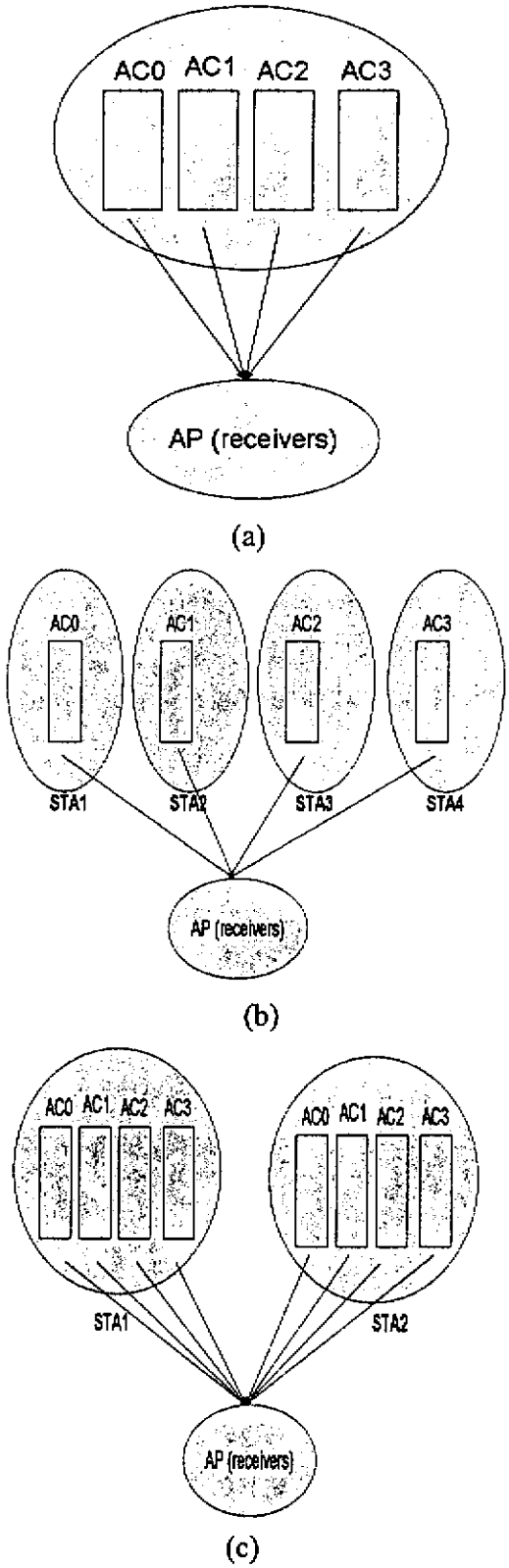
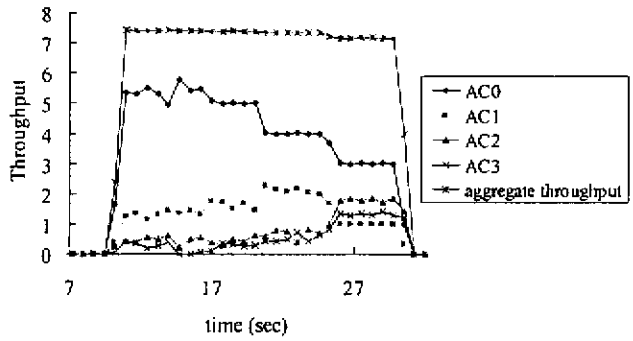
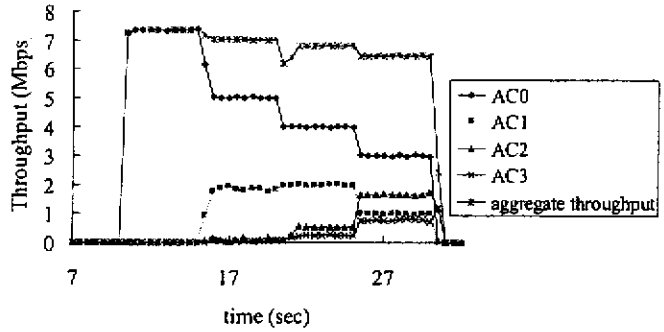


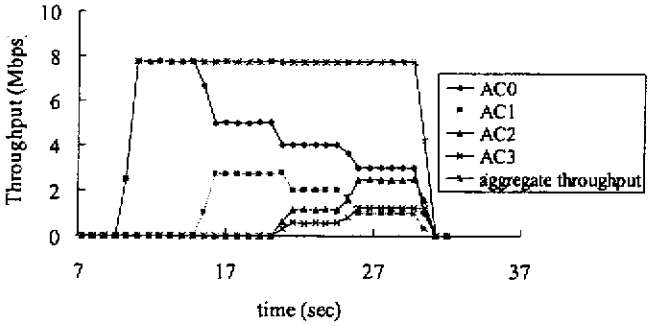
Figure 2. Simulation scenarios



(a) EDCA



(b) BIWF-SP



(c) IDFQ-SP

Figure 3. The throughputs of three schemes

五、結論：

本報告在 802.11e WLANs 提出兩個方法 BIWF-SP 及 IDFQ-SP，分別以 backoff interval 及 interframe space 為基礎以提供不同 AC 不同的服務品質，包括絕對的優先順序及依權重分配頻寬。在這份報告中，我們仔細的描述這兩個方法的作法，及透過模擬比較它們和原來 EDCA 的比較。模擬結果顯示 BIWF-SP 及 IDFQ-SP 在支援絕對的優先順序及依權重分享頻寬方面都表現的比 EDCA 好。和 IDFQ-SP 比較起來，BIWF-SP 在實際系統比較容易實作；但和 BIWF-SP 比較的話，IDFQ-SP 有較好的 aggregate throughput 及較穩定的優點。更重要的是，這兩個方法都和現有的 802.11e 標準相容，表示它們都很適合在 802.11 WLANs 裡對每個 AC 提供不同的服務品質。

六、 參考文獻

[1] IEEE 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications, *IEEE* Standard, Aug. 1999.

[2] I. Aad and C. Castelluccia, "Differentiation Mechanisms for IEEE," *IEEE INFOCOM 2001*.

[3] J. Deng and R.S. Chang, "A Priority Scheme for IEEE 802.11 DCF Access Method," *IEICE Trans. Commun.*, Vol. E82-B, no. 1, 1999, pp. 96-102

[4] V. Kanodia, C. Li, A. Sabharwal, B. Sadeghi and E. Knightly, "Distributed Multi-Hop Scheduling and Medium Access with Delay and Throughput Constraints," *ACM Mobicom 2001*

[5]S. Golestani, "A Self-Clocked Fair Queueing Scheme for Broadband Applications." *IEEE INFOCOM '94*, page 636-646, Toronto, CA, Jane 1994

[6] S. Lu, T. Nandagopal, and V. Bharghavan, "A Wireless Fair Service Algorithm for Packet Cellular Networks," *ACM Mobicom* 1998.

[7] N. H. Vaidya, P. Bahl, and S. Gupta, "Distributed Fair Scheduling in Wireless LAN," *ACM Mobicom* 2000.

[8] D. Qiao and K. G. Shin, "Achieving Efficient Channel Utilization and Weighted Fairness for Data Communications in IEEE 802.11 WALN under the DCF," *ACM IWQoS* 2002.

[9] W. Pattara-Atikom, S. Banerjee, and P. Krishnamurthy, "Starvation Prevention and Quality of Service in Wireless LANs." *IEEE WPMC* 2002.

[10] IEEE WG, Draft Supplement to Standard for Telecommunications and Information Exchange between Systems-LAN/MAN Specific Requirements- Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Medium Access Control (MAC), Enhancements for Quality of Service (QoS), *IEEE 802.11e Draft 4.1*, Feb. 2003.

[11] M. Malli, Q. Ni, T. Turletti and C. Barakat, "Adaptive Fair Channel Allocation for QoS Enhancement in IEEE 802.11 Wireless LANs." *IEEE ICC* 2004.

[12] M. Shreedhar and G. Varghese, "Efficient Fair Queuing Using Deficit Round-robin," *IEEE/ACM Trans. Net.*, vol. 4, no. 3, 1996, pp. 375-85.

[13] ns-2, available at http://www.isi.edu/nsnam/ns/

◆ 子計畫七：無線網路多點傳輸的壅塞控制以及應用於 MPEG-4 之調適性速率控制機制之設計（吳曉光）

一. 簡介

近年來無線網路迅速的發展，許多值得研究的議題也隨之產生。本報告包含的研究主題分別為應用於MPEG-4 即時視訊平台中，以線性模型為基礎的可調適性速率控制機制和無線網路環境下的可靠多點傳輸的壅塞控制。應用於MPEG-4 即時視訊平台中，以線性模型為基礎的可調適性速率控制機制一個可以幫助視訊串流服務適合於隨時在變動的可用頻寬現制下的可調適性速率控制機制必須要被發展設計出來。

另一主題為在無線網路環境下的可靠多點傳輸的壅塞控制，壅塞控制的機制使得網路傳輸，可以調整傳送速率符合網路上可用的頻寬，並且公平地與目前盛行的網路傳輸方式。

二. 計畫緣由與目的：

應用於MPEG-4之調適性速率控制機制設計

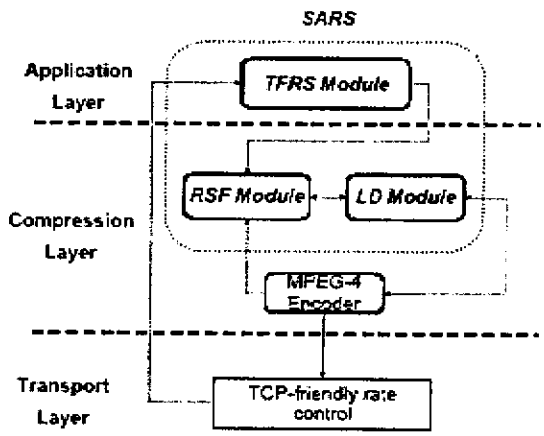
隨著在Internet 上或者無線網路中提供視訊串流服務的需求量大增，一個可以幫助視訊串流服務適合於隨時在變動的可用頻寬現制下的可調適性速率控制機制必須要被發展設計出來。隨時在變動的可用頻寬和視訊串流在網路上傳送的延遲時間讓即時視訊串流服務難以達到很高的頻寬使用率，並且也無法提供好的視訊品質給收看的使用者。MPEG-4 是目前最被廣為接受採用的一個視訊串流的標準格式，而一個採用MPEG-4 來作為視訊編碼解決方案的即時視訊串流系統中必須要提供一個能夠調整視訊壓縮位元率的控制機制來適應這個隨時隨地在改變的網路狀態。

無線網路多點傳輸的壅塞控制

可靠的多點傳輸協定提供有效率以及可靠的傳輸給有此需求的軟體，而且壅塞控制的機制使得網路傳輸，可以調整傳送速率符合網路上可用的頻寬，並且公平地與目前盛行的網路傳輸方式,TCP, 分享網路頻寬，有許多的服務建構在可靠的多點傳輸的壅塞控制傳輸方式上面。隨著無線網路環境的增加，多點傳輸的服務對於一直增加的可存取網路資源行動裝置日益重要，但是，原本的可靠多點傳輸的壅塞控制會被無線網路中，較高的封包遺失率所影響，造成效能低落。

三. 應用於 MPEG-4 之調適性速率控制機制設計與成果討論

在此研究中，我們分析了當一個即時視訊串流服務傳送端採用了TCP-friendly 的傳輸協定在Internet 中傳送MPEG-4 視訊時所產生的問題。而這篇論文主要著重在提供軟體層和壓縮層上問題的解決方案。根據這樣的目標，我們提供了一套解決方案叫做SARS。SARS是一個具有平滑而且可調適性的速率控制機制，而且它包含了三個主要的元件，分別為TFRS 模組、LD 模組和RSF 模組。TFRS 模組提供了視訊壓縮時比較平滑的速率，讓產生的視訊畫面間的畫質變動比較小，提供比較平滑的畫質。LD 模組可以幫助視訊壓縮時讓壓縮位元率逼近它的目標位元率來減少傳送的延遲時間。RSF 模組則提供了一個可調適性的速率控制機制，根據變動的頻寬來分配目標位元率，並且降低了被略過的畫面數目，圖一為整個系統的架構圖。



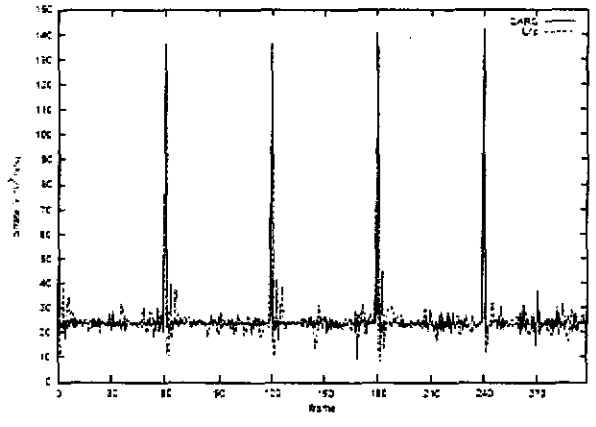
圖一、系統架構圖

在實驗部分我們評估了SARS 在做CBR和

VBR 壓縮時的效能，並且也針對不同TCP-friendly 傳輸協定進行模擬實驗。

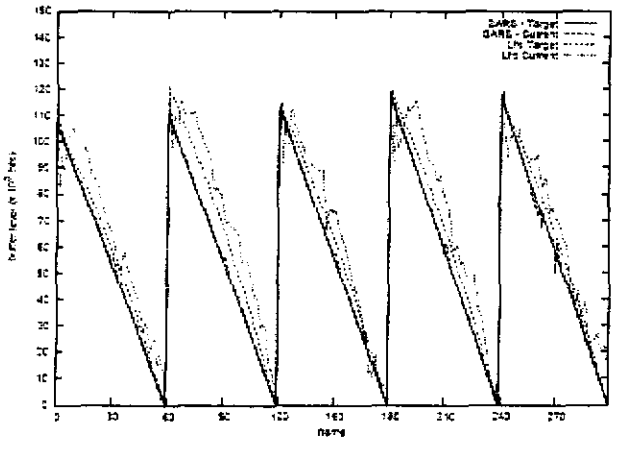
我們比較的對象為Li [1]所提出的rate control 機制。考慮兩種 video sequences，" Mobile" 和" Paris" ，video sequences大小為CIF(352x288)，frame rate為30fps，video sequences使用GOP做編碼，GOP的長度為60，考慮不同CBR和VBR，分別為CBR-768Kbps、CBR-512Kbps、CBR-256Kbps和VBR，以使用" Mobile" 在CBR-768Kbps的方案為例。

在圖二中Li[1]的方法有較大的coding bit rate的變動。圖三中Li的方法對於buffer有較大的變動。圖四中可以看出當P-frames較靠近I-frames時PSNR值稍微低於Li的方法，在後半部分高於Li的方法。圖五說明在每個GOP中平均的PSNR值，在此方案中SARS稍微增加了PSNR值。



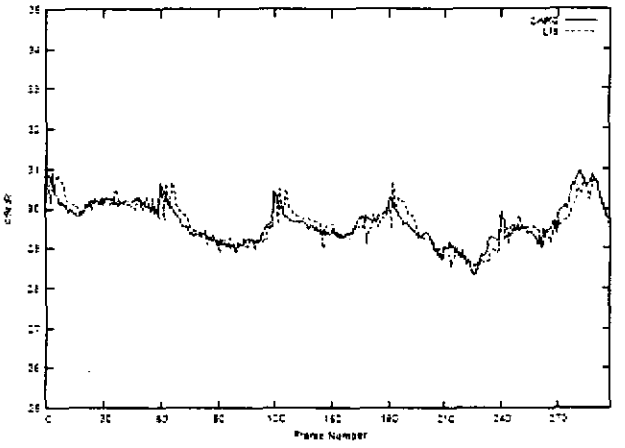
(a) actual coding bit rate of each frame

圖二、每個frame的實際coding bit rate



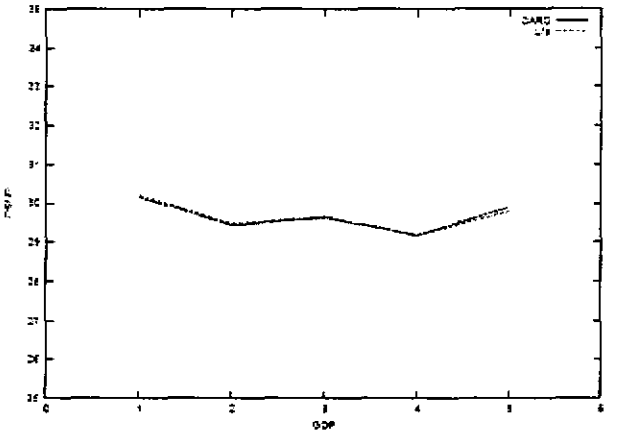
(b) buffer level variation at each frame

圖三、每個frame的buffer變動水準



(c) PSNR results of each frame

圖四、每個frame的PSNR結果



(d) average PSNR results of each GOP

圖五、每個GOP平均PSNR的結果

最後我們在rate based congestion control協定下評估SARS的表現。

(1) 網路拓普

圖六中R1和R2是routers，當traffic超過連結的容量則會發生congestion，Sn為傳送者，且會產生TCP、RAP、TMRC和TFRC flows，Dn為接收者。

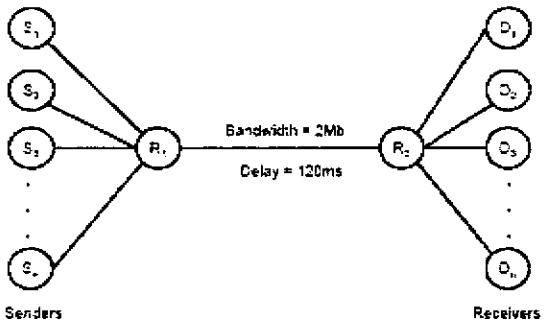
(2) 2TCP vs 2 Rate-Based Congestion Control Protocols

全部有三個方案分別為 2TCP vs 2RAP、2TCP vs 2TFRC 和 2TCP vs 2TMRC。圖七中，我們選擇RAP flows、TFRC flows和TMRC flows，描繪出它們的sending rate。

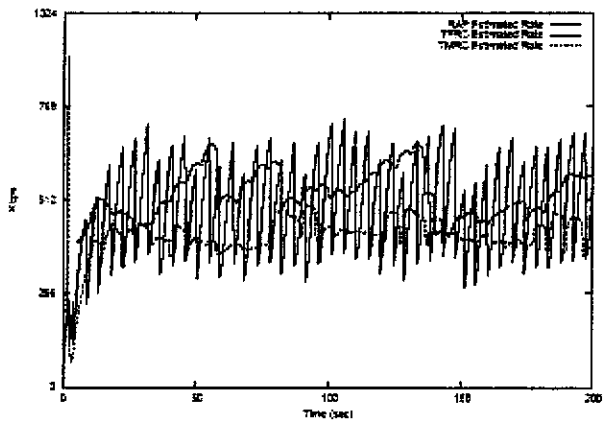
(3) 2TCP vs 2 Rate-Based Congestion Control Protocols with 4 TCP Competing during 20s and 80s

全部有三個方案分別為 2TCP vs 2RAP、2TCP vs 2TFRC 和 2TCP vs 2TMRC，將有4TCP開始於20s結束在80s。圖八中，我們選擇RAP flows、TFRC flows和TMRC flows，描繪出它們的sending rate。

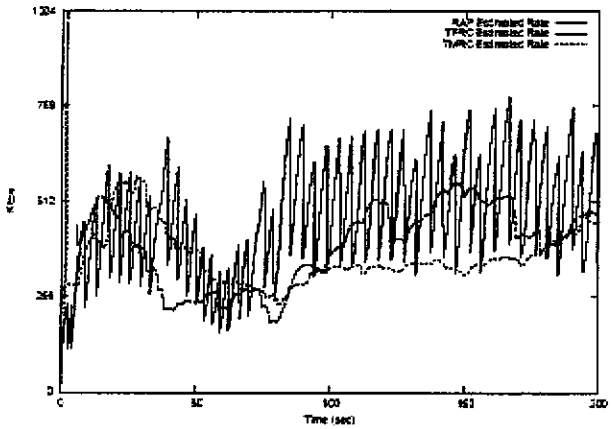
模擬結果顯示使用SARS方法，平均的PSNR在TFRC或TMRC低於Li[1]的方法。



圖六 網路拓普



圖七、估計的sending rate



圖八、估計的sending rate

四. 無線網路多點傳輸的壅塞控制設計與成果討論

本研究提出一個方法，叫做PGMCCW，用來減輕無線網路造成的封包遺失的影響；我們的方法可以分辨出在多點傳輸上無線網路所造成的封包遺失和壅塞所造成的封包遺失，並且在眾多接收端處於不同的網路環境的情況下，選擇適當的回饋資訊，正確地控制整個群組傳送的速率。

PGMCCW分為三個部分，分別為congestion estimation、wireless loss omission，representative switch determination。

(1) Congestion estimation

主要的目標就是要判斷是否傳輸路徑是壅塞的狀態，利用JTCP[2]算出jitter ratio，如果jitter ratio大於壅塞窗口(congestion window)的倒數(Jr>1/cwnd)則表示路徑為壅塞，反之則否。

(2) Wireless loss omission

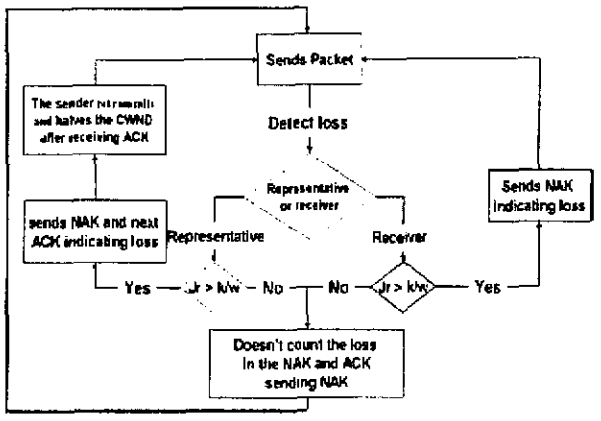
在PGMCC中，挑選一個接收者作為回應的代表，當傳送者收到ACK則增加壅塞窗口，但是當傳送者收到NACK則將壅塞窗口減半，這和TCP類似，但是在無線的環境之下則無法運作的很好，所以利用jitter ratio來判定是否為congestion loss，若Jr>1/cwnd則為congestion loss。

(3) Representative switch determination

接收者的代表由以下的公式來挑選

RTT x RTT x Loss Ratio

其中Loss Ratio利用jitter ratio，只考慮congestion loss，不包含wireless loss，圖九為PGMCCW的流程。



圖九、PGMCCW 流程圖

我們設計各種方案來驗證我們演算法的確是有效的，比較的對象為 PGMCC，分別針對 throughput vs. loss rate, fairness with TCP、receiver vs. throughput, wireless loss vs. bandwidth constrain, wireless loss vs. delay constraint。提出的方法模擬結果也驗證能有效地改進傳送的效能，在越來越普及的無線網路環境下，達到更佳的傳輸效率。

五、未來計畫：

本期末報告包含應用於 MPEG-4 之調適性速率控制機制設計和無線網路多點傳輸的壅塞控制，在 2005 年度將針對 IEEE802.16 相關技術做更進一步的研究。

六、 參考文獻

[1] Z. G. Li, C. Zhu, N. Ling, X. K. Yang, G. N. Feng, S. Wu, and F. Pan,"*A Unified Architecture for Real-Time Video-Coding Systems*", IEEETrans. Circuits Syst. Video Technol., vol 13, pp. 472-487, June 2003.

[2] Eric Hsiao-kuang Wu, Meizhen Chen, "*JTCP, Jitter-based for Wireless Networks*", accepted for IEEE Journal of Selected Area Communications.

[3] Wei-Ren Yang and Eric Hsiao-Kuang Wu,"Pragmatic General Multicast Congestion Control over Wireless Network"

[4] Chung-Yuan Knight Chang and Eric Hsiao-Kuang Wu,"SARS: A Linear Source Model Based Adaptive Rate-control Scheme for TCP-friendly Real-time MPEG-4 Video Streaming"

附錄一

A Dual-band IEEE 802.11a/b/g Receiver Front-end Using Half-IF and Dual-Conversion

Chun-Chih Hou, Ching-Chi Chang, and Chorng-Kuang Wang

Department of Electrical Engineering and Graduate Institute of Electronic Engineering,
National Taiwan University, Taipei, Taiwan 10617, R.O.C.

Abstract—This paper presents a dual-band receiver front-end architecture which combines the dual-conversion [4] and half-IF techniques for IEEE 802.11a/b/g. The proposed architecture receives dual-band signal with single receiver chain to reduce components count such as the 2nd down-conversion mixer and VCO. The required LO frequencies of the dual-band application can be synthesized by only one VCO in combination with a divide-by-four circuit. An LC tank aided mechanism of mixer is also proposed to deal with the flicker noise. The core portion of the front-end receiver has been fabricated in 1P6M 0.18μm CMOS technology. The delivered gain, noise figure, and IIP3 are 20 dB, 3.5 dB, and −13 dBm, respectively simulated. The chip occupies an area of 1.21×1.46mm^2 and the power consumption is 24mW under the supply voltage of 1.8V.

Index Terms—WLAN, LNA, mixer, half-IF, flicker noise, dual-conversion, dual-band.

I. Introduction

In the market of wireless local area network (WLAN), multi-standard transceivers have been the trend. The IEEE 802.11b standard at the 2.4 GHz ISM band provides data rates up to 11 Mbits/s [2]. The 802.11a standard at the 5GHz U-NII bands provides data rates up to 54 Mbits/s using OFDM modulation [1]. Released in 2003, the 802.11g standard, operated at the same band of 802.11b, uses the modulation method of 802.11a and provides data rates up to 54 Mbits/s [3]. Many products in the market can support each of the standards simultaneously, and the employed transceiver architectures are usually the direct-conversion ones.

As for the published dual-band architectures, most of them receive the two-band signal by two signal paths and the direct-conversion architecture is mostly adopted, too [5-7]. In this paper, a new architecture is proposed that can receive signal at two-band, 5.8GHz and 2.4GHz, with the same receiver chain. Due to the elaborate frequency conversion planning, the troublesome image problem is relaxed in this design. The organization of this paper is as follows. The receiver architecture is described in Section II. In Section III, the circuit topologies meeting the requirements of this architecture are presented. Finally, the simulation results and conclusions are given in Section IV and V, respectively.

II. Receiver Architecture

The proposed receiver front-end architecture is shown in Fig. 1. The components inside the dashed frame can be implemented into a single chip, and those outside the frame are the lumped components, which are switch, diplexer, and band-select band-pass filters. Two band-pass filters instead of one are used in order to provide sufficient rejection of the out-of-band interferences for each band. When the desired signal is at 5.8 GHz, the frequency conversion method, shown in Fig. 2., is used [4]. The 3.465GHz image is 2 GHz far away from the desired signal, and it can be filtered out by the band-pass nature of the on-chip amplifiers. The first local oscillator (LO) frequency, which is 4.62 GHz, is four times of the second LO, and thus they can be synthesized by a single voltage controlled oscillator (VCO) and a divide-by-four circuit. Because the VCO frequency is different from the signal frequency, the LO pulling problem will not occur.

This dual-band receiver architecture can also be switched to half-IF conversion to receive 2.4GHz RF signal, as shown in Fig. 3. The utilized 1.2 GHz LO frequency can be obtained by the same circuits composed of one VCO with 4.8-5 GHz tuning frequency and a divide-by-four circuit. As shown in Table I, the VCO frequency range of the two bands is close, so a VCO with wide tuning range can deliver the required LO frequencies for dual-band application. At the same time, the 90° phase shifter can be eliminated due to the utilization of dividers. Another benefit of this dual-band receiver is that the IF frequencies are close for this dual-band signal, so there is an ease to implement the circuits after the first down-conversion. As for the image signal around the zero frequency, the attenuation contributed by amplifiers and lumped components such as antenna and band-select filter is very large, and thus the image rejection ratio can be as high as 60dB [8]. Two problems that should be taken into account are the upconversion of flicker noise to IF frequency and the DC-offset that comes from LO-to-IF feedthrough. The circuits proposed in the next section will deal with the first problem. For the DC-offset problem, a high-pass filtering offset-cancellation circuit can solve it [8]. The SNR degradation due to the filtering is not significant because for IEEE 802.11a/g, the OFDM modulation leaves the center sub-carrier empty, and for IEEE 802.11b, spread-spectrum modulation is used.

This program is supported by the National Science Council (NSC), Taiwan, R.O.C.

III. Circuit Topologies

The only RF circuits that need to be carefully designed for our receiver architecture are the LNA and the first mixer. The design of other components is the same as the single band ones. For example, the VCO required is just a wide-tuning-range one. Besides, the second mixer and the amplifiers at baseband have single-band nature because the IF bandwidth is not very wide, which has been shown in the last section.

The LNA schematic is shown in Fig. 4. Because the LNA is designed to have the ability to receive and amplify the two-band signal, it has a load that contains two inductors and two capacitors [9]. The function of this LC tank can be comprehended as follows. When the frequency is low, the L and C in series are capacitive in effect. When the frequency is high, they are inductive in effect. Thus, the LC tank has two peaks in the frequency domain. The input matching of the LNA is achieved by off-chip microstrip lines, which are combinations of some open-stubs on the PCB board.

The mixer schematic is shown in Fig. 5. It is a gilbert-cell type mixer with source degeneration resistors to enhance the linearity. The mixer is a wide-band device in nature, so signal can drive the input ports and LO ports directly. Also, because signal frequency at the output of the mixer is about 1 GHz for both the two-band operation, the output port is single-band in nature and thus only one inductor is used, instead of the two-inductor case in LNA.

The mixer in [8] prevents the up-conversion of flicker noise of the transconductance stage to IF, but ignores the flicker noise of the switching pairs. The mixer proposed here uses an LC tank to achieve the filtering function. The resonant frequency of the LC tank is at the signal frequency such that it blocks the signal from passing through it. As for the flicker noise around zero frequency, it sees a short because the LC tank has an inductor in parallel. The circuit can thus be reduced to the half circuit as shown in Fig. 6. The topology of mixer is destroyed, and the flicker noise in neither the transconductance stage nor the switching stage will be up-converted to IF.

The linearity of amplifiers is an important consideration for wireless communication. When the signal received by the antenna is not very large, the third-order inter-modulation signal located at the signal frequency band because of the adjacent-channel interferences and the non-linearity of amplifiers must be low enough in order not to degrade the SNR. For WLAN, the IIP3 of the receiver chain should be at least –26dBm to meet this requirement. Meanwhile, when the input signal is large, which is at most –30dBm, –10dBm, and –20dBm for IEEE 802.11-a/b/g respectively, the amplifiers may be saturated. Moreover, for IEEE 802.11a/g that have 52 subcarriers, another add-on power back off of 10dB is needed to accommodate the large peak-to-average ratio (PAR). Thus, a gain-adjusting mechanism is necessary for the front-end amplifiers. A variable resistor can be placed in parallel with the load inductor, and in this case it is replaced by a digitally switched PMOS, as shown in Fig. 7. Discrete gain steps can also be achieved by many of such PMOS placed together.

IV. Simulation Results

The input return loss of the LNA is shown in Fig. 8., in which two bands matching is achieved, with bandwidth larger than 200 MHz in ISM band and larger than 300 MHz in the upper U-NII band. The S21 of LNA is shown in Fig. 9. The gain at each band is about 10 dB. The image rejection ratio (IRR) provided by this amplifier is larger than 60 dB for upper U-NII band, and 30 dB for ISM band. The noise figure of the LNA and mixer operated at 2.4 GHz when no LC tank is inserted in the LO switching pair is shown in Fig. 10. The up-converted flicker noise has a significant influence on the noise figure. After the LC tank is inserted, the influence from the flicker noise vanishes, as shown in Fig. 11.

Table II is the summary of the simulation results of the LNA and mixer. From the comparison between the two circuits with and without LC tanks to block the up converting of flicker noise, it can be observed that the noise figures for both bands improve. However, the conversion gain and IIP3 is not the case. This is because that the gain depends on the quality factor of the LC tanks and the linearity depends on the harmonic termination [10], which is not the same for the two bands.

The power consumption of the core circuit is 24mW. It has been fabricated by TSMC 1P6M 0.18μm CMOS. The chip area is 1.21×1.46 mm^2 and its microphotograph is shown in Fig. 12.

IV. Conclusion

A dual-band receiver front-end architecture is proposed in this paper. It combines the advantage of dual-conversion and half-IF receivers to provide IRR of 60dB and 30dB for U-NII and ISM band signal respectively. The IF frequency range of 1.14 to 1.24GHz relax the component requirement after the first downconversion. The dual-band LNA and mixer have been designed to deliver a conversion gain of 20dB and IIP3 of –14dBm with a power consumption of 24mW. The troublesome flicker noise problem in half-IF conversion has been solved by a circuit topology proposed in this paper.

Acknowledgement

The authors would like to thank to the technical support by Chip Implementation Center (CIC), Taiwan, R.O.C.

References

[1] IEEE Std 802.11, Wireless LAN Medium Access

Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band, Sept. 1999.

[2] IEEE Std 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer Extension in the 2.4 GHz Band, Sept. 1999.

[3] IEEE Std 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 4: Further Higher Data Rate Extension in the 2.4 GHz Band, June 2003.

[4] D. Su, M. Zargari, P. Yue, S. Rabii, D. Weber, B. Kaczynski, S. Mebta, K. Singh, S. Mendis and B. Wooley, "A 5GHz CMOS Transceiver for IEEE 802.11a Wireless LAN," in *IEEE Int. Solid-State Circuits Conf. Tech. Dig.*, Feb. 2002, pp. 70-405.

[5] M. Hotti, J. Kaukovuori, J. Ryynanen, K. Kivekas, J. Jussila, and K. Halonen, K, "A Direct Conversion RF Font-end for 2-GHz WCDMA and 5.8-GHz WLAN Applications," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 45-48, June 2003.

[6] B. U. Klepser, M. Punzenberger, T. Ruhlicke, and M. Zannoth, "5-GHz and 2.4-GHz Dual-band RF Transceiver for WLAN 802.11a/b/g Applications," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 37-40, June 2003.

[7] M.Y. Wang, R.R.-B. Sheen, D.T.-C. Chen, and R.Y.J. Tsen, "A Dual-band RF Front-end for WCDMA and GPS Applications," in *Proc. 2002 Int. Symp. Circuits and Systems*, vol.4, 2002, pp. 113-116.

[8] B. Razavi, "A 5.2-GHz CMOS Receiver with 62-dB Image Rejection," *IEEE J. Solid-State Circuits*, vol. 36, pp. 810-815, May 2001.

[9] H. Hashemi, H and A. Hajimiri, "Concurrent Multiband Low-noise Amplifiers - Theory, Design, and Applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, pp. 288-301, Jan. 2002.

[10] J. S. Fairbanks and L. E. Larson, "Analysis of optimized input and output harmonic termination on the linearity of 5 GHz CMOS radio frequency amplifiers," *Proceedings of Radio and Wireless Conference, 2003, RAWCON '03*, pp. 293-296, Aug. 2003.

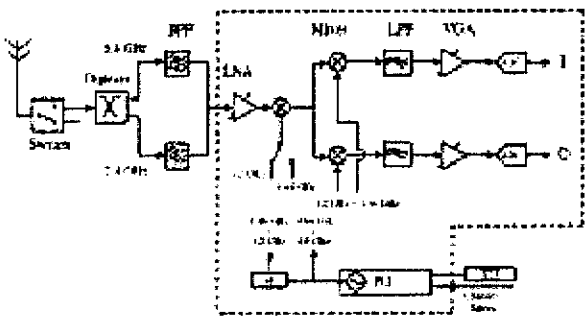


Fig. 1. Block diagram of the receiver.

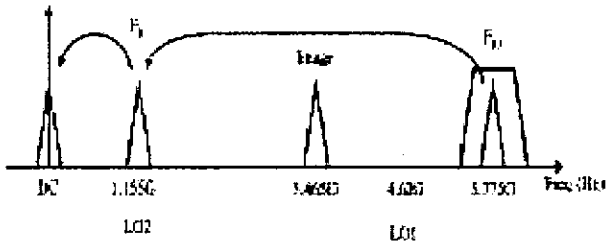


Fig. 2. Frequency planning of 5 GHz signal.

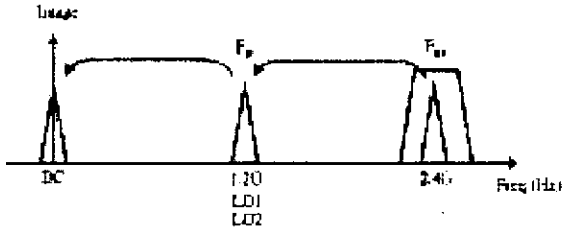


Fig. 3. Frequency planning of 2.4 GHz signal.

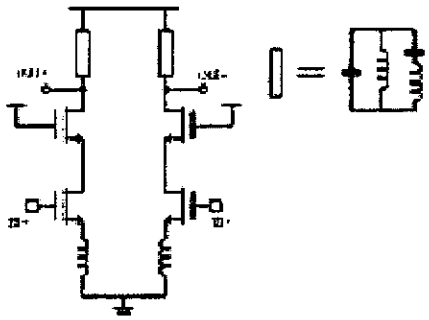


Fig. 4. Schematic of the LNA.

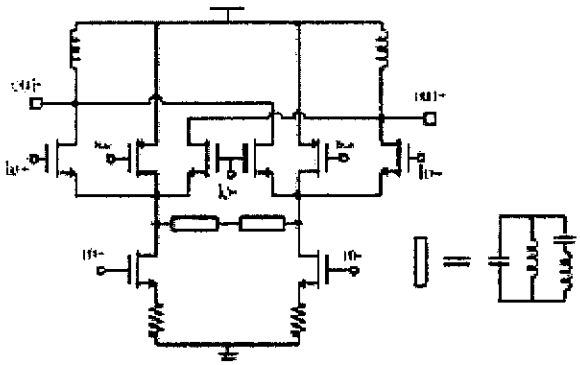


Fig. 5. Schematic of the first mixer.

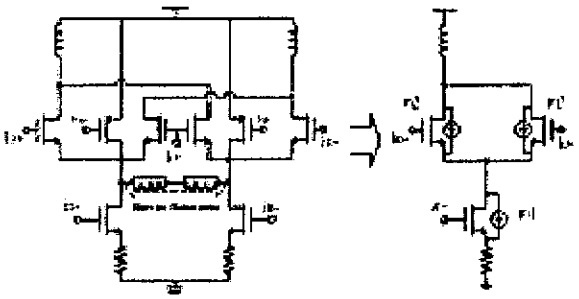


Fig. 6. Filtering mechanism for flicker noise.

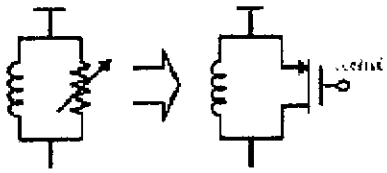


Fig. 7. Gain-adjusting circuit.

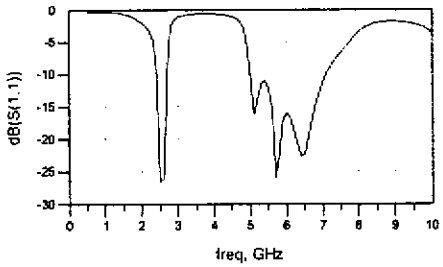


Fig. 8. Input return loss of the LNA.

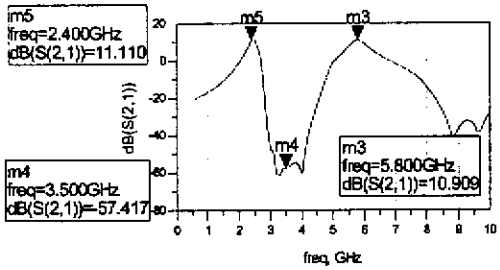


Fig. 9. S21 of the LNA.

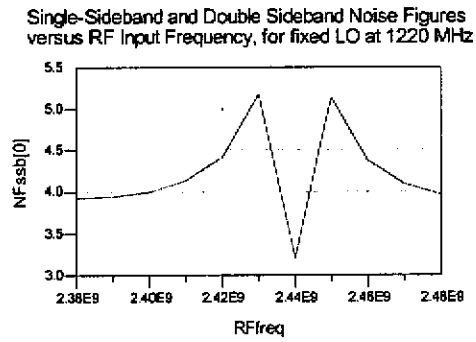


Fig. 10. Noise figure at the 2.4 GHz band, without LC tank inserted.

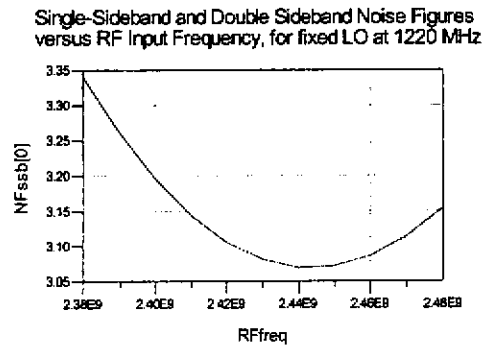


Fig. 11. Noise figure at the 2.4 GHz band, with LC tank inserted.

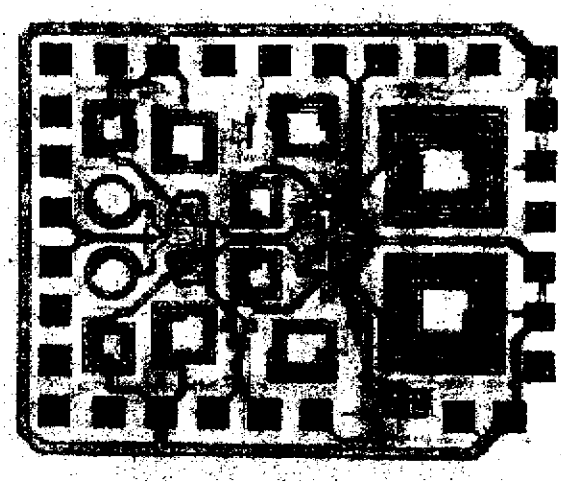


Fig. 12. Chip photograph.

TABLE I
FREQUENCY RANGE OF EACH STAGE

Frequency Band	Frequency Range	VCO Range	IF Range
Upper U-NII Band (GHz)	5.725-5.825	4.58-4.66	1.145-1.165
ISM (GHz)	2.40-2.4835	4.8-4.967	1.2-1.242

TABLE II
SIMULATED LNA AND MIXER PERFORMANCE.

Parameters		LNA+Mixer without flicker noise filtering circuitry	LNA+Mixer with flicker noise filtering circuitry
Supply Voltage		1.8 V	1.8 V
Power Consumption (without biasing and output driver circuits)		24 mW	24 mW
2.4 GHz	Gain	21.3 dB	20 dB
	NF	>4.5 dB	3.1 dB
	IIP3	-16.5 dBm	-13.4 dBm
	LO Power	-2 dBm	-2 dBm
5.8 GHz	Gain	16.4 dB	18.8 dB
	NF	4.25 dB	3.55 dB
	IIP3	-8.8 dBm	-11.4 dBm
	LO Power	-2 dBm	-2 dBm
Technology		TSMC 1P6M 0.18μm CMOS	
Chip Area			1.21×1.46 mm^2

A VARIABLE-LENGTH DHT-BASED FFT/IFFT PROCESSOR FOR VDSL/ADSL SYSTEMS

Tsung-Chieh Pao, Ching-Chi Chang, and Chorng-Kuang Wang*

Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University, Taipei, 106, Taiwan R.O.C.

ABSTRACT

In this paper, a matrix form of the radix-4 Discrete Hartly Transform (DHT) algorithm is derived. And two Process Element (PE) circuits are proposed for the variable-length application. One is for radix-2 with one real multiplier and three adders, and the other is for radix-4 with one real multiplier and seven adders. The radix-4 DHT processor of variable-length, namely, 8192/4096/2048/1024/512, meets the DMT-VDSL and ADSL requirements. Only six real multipliers, 2.01N (16467) registers and simple logic circuits are used to perform 8192-points FFT/IFFT in 231μs. Implemented in 0.25μm 1P5M typical CMOS technology, the core occupies an active area of 5mm×5mm. Power dissipation is 300mW using 2.5V supply voltage.

1. INTRODUCTION

Among standards with Multiple Carrier Modulation (MCM) technology using FFT/IFFT as the modulation technique, Very-high-bit-rate Digital Subscriber Line (VDSL) has potential for broadband communication of the next generation, owing to the popularity of twisted pairs [1]. The Discrete Fourier Transform (DFT) plays an important role in DMT/OFDM systems. Because of the computational complexity of N-points DFT is $O(N^2)$, it takes large operation time and power consumption to perform DFT directly, especially in large transform size (e.g. 8192 points in VDSL). Hence, various FFT algorithms have been proposed to reduce the computational complexity [2], such as Cooley-Turkey algorithm, Winograd Algorithm (WFTA), radar algorithm and Prime Factor Algorithm (PFA). Cooley-Turkey algorithm is the most popular one in VLSI implementation due to reducing the computational complexity from $O(N^2)$ to $O(N\log_2 N)$ and its regularity. Table 1 summaries several FFT architectures including Single-Path Delay Feedback (SDF) architecture [3], Multiple-Path Delay Commutator (MDC) [4], and DHT-based FFT/IFFT processor [5]. There are two key points shown in this table. One is that MDC architectures need more hardware than SDF architectures obviously. The other is that no matter what radix-r SDF architecture is used, 2(N-1) registers are essential. In order to reduce hardware in FFT/IFFT processor, the design goal should be to reduce the number of real multipliers without increasing the number of registers. Based on the advantage of Hermitian symmetric property in DMT-VDSL and ADSL systems, some DHT-based FFT algorithms and architectures have been presented to reduce the numbers of real multipliers [5-7]. But there are some drawbacks about these architectures such as more buffers for RAM-based designs and only radix-2 DHT-based FFT algorithm is delivered. If the higher radix FFT algorithm is adopted, the fewer multipliers can be employed with Cooley-Turkey algorithm. Our goal is to derive matrix form of higher radix DHT algorithm from previous work and then implement it using less hardware.

The organization of this paper is as follows. The matrix form of radix-4 DIF DHT algorithm is derived in Section 2. Section 3 presents the proposed circuit designs of the two PEs. Section 4 shows VLSI implementation and performance summary of the FFT processor for VDSL system, and the conclusion is given in section 5.

Table 1. Hardware comparisons of various FFT architectures.

	Real Multipliers	Memory	Through-put (sample/cycle)
Radix-2 SDF	$3(\log_2 N-2)$	$2(N-1)$	1
Radix-4 SDF	$3(\log_4 N-1)$	$2(N-1)$	1
Radix-8 SDF	$3(\log_8 N-1)$	$2(N-1)$	1
Radix-2 MDC	$3(\log_2 N-2)$	$2(1.5N-2)$	2
Radix-4 MDC	$3(3\log_4 N-3)$	$2(2.5N-4)$	4
Radix-8 MDC	$3(7\log_8 N-7)$	$2(4.5N-8)$	8
Radix-2 DHT-based [5]	$\log_2 N-1$	$2(N-1)$	1

2. RADIX-4 DHT ALGORITHM

Given N-points DHT definition as

$$y_k = \sum_{n=0}^{N-1} x_n(\cos\frac{2\pi kn}{N} + \sin\frac{2\pi kn}{N}) \quad (1)$$
$$= \sum_{n=0}^{N-1} x_n cas\frac{2\pi kn}{N},$$

This program is supported by the National Science Council (NSC), Taiwan R.O.C.

*Currently with Realtek Semiconductor Corp.

where k=0, 1, …, N-1. The frequency samples, y_{4k}, y_{4k+1}, y_{4k+2} and y_{4k+3} are derived as [8]:

$$y_{4k} = \sum_{n=0}^{N-1} x_n cas\frac{2\pi(4k)n}{N} = \sum_{n=0}^{N/4-1}[x_n + x_{n+\frac{N}{4}} + x_{n+\frac{N}{2}} + x_{n+\frac{3N}{4}}]cas\frac{2\pi(4k)n}{N}, \quad (2)$$

$$y_{4k+1} = \sum_{n=0}^{N-1} x_n cas\frac{2\pi(4k+1)n}{N} = \sum_{n=0}^{N/4-1}[(x_n - x_{n+\frac{N}{2}} + x_{\frac{N}{4}-n} - x_{\frac{3N}{4}-n})\cos\frac{2\pi n}{N} + (x_{n+\frac{3N}{4}} - x_{n+\frac{N}{4}} + x_{\frac{N}{2}-n} - x_{N-n})\sin\frac{2\pi n}{N}]cas\frac{2\pi(4k)n}{N}, \quad (3)$$

$$y_{4k+2} = \sum_{n=0}^{N-1} x_n cas\frac{2\pi(4k+2)n}{N} = \sum_{n=0}^{N/4-1}[(x_n + x_{n+\frac{N}{2}} - x_{n+\frac{N}{4}} - x_{n+\frac{3N}{4}})\cos\frac{2\pi 2n}{N} + (x_{\frac{N}{4}-n} + x_{\frac{3N}{4}-n} - x_{\frac{N}{2}-n} - x_{N-n})\sin\frac{2\pi 2n}{N}]cas\frac{2\pi(4k)n}{N}, \quad (4)$$

$$y_{4k+3} = \sum_{n=0}^{N-1} x_n cas\frac{2\pi(4k+3)n}{N} = \sum_{n=0}^{N/4-1}[(x_n - x_{n+\frac{N}{2}} - x_{\frac{N}{4}-n} + x_{\frac{3N}{4}-n})\cos\frac{2\pi 3n}{N} + (-x_{n+\frac{3N}{4}} + x_{n+\frac{N}{4}} + x_{\frac{N}{2}-n} - x_{N-n})\sin\frac{2\pi 3n}{N}]cas\frac{2\pi(4k)n}{N}, \quad (5)$$

where k=0, 1, …, N/4-1. From Eq. 2, 3, 4 and 5, the frequency samples, y_{4k}, y_{4k+1}, y_{4k+2} and y_{4k+3}, are respectively equal to the DHT of the four N/4-points sequences bracketed in these four equations. Then, four N/4×1 matrices, $\boldsymbol{Y}_{4k}$, $\boldsymbol{Y}_{4k+1}$, $\boldsymbol{Y}_{4k+2}$ and $\boldsymbol{Y}_{4k+3}$, and four N/4×N matrices, $\boldsymbol{X}_A$, $\boldsymbol{X}_B$, $\boldsymbol{X}_C$ and $\boldsymbol{X}_D$, are defined as follows.

$$\boldsymbol{Y}_{4k}=[y_0\ y_4\ y_8\ \cdots\ y_{4(N/4-1)}]^T,$$
$$\boldsymbol{Y}_{4k+1}=[y_1\ y_5\ y_9\ \cdots\ y_{4(N/4-1)+1}]^T,$$
$$\boldsymbol{Y}_{4k+2}=[y_2\ y_6\ y_{10}\ \cdots\ y_{4(N/4-1)+2}]^T,$$
$$\boldsymbol{Y}_{4k+3}=[y_3\ y_7\ y_{11}\ \cdots\ y_{4(N/4-1)+3}]^T,$$

$$\boldsymbol{X}_A=\begin{bmatrix} 1 & 1 & 1 & 1 & & & & \\ & & & & 1 & 1 & 1 & 1 \\ & & & & & & & \ddots \\ 1 & 1 & 1 & 1 & & & & \\ & & & & 1 & 1 & 1 & 1 \end{bmatrix}\boldsymbol{X}$$

$$\boldsymbol{X}_B=\begin{bmatrix} C_N^0 & C_N^0 & -C_N^0 & -C_N^0 & & & & \\ C_N^0 & C_N^0 & -S_N^0 & S_N^0 & & & & \\ & & & & -C_N^0 & -C_N^0 & S_N^0 & -S_N^0 \\ & & & & C_N^0 & C_N^0 & -S_N^0 & S_N^0 \\ & & & & & & & \ddots \\ C_N^{\frac{N}{4}-1} & C_N^{\frac{N}{4}-1} & -S_N^{\frac{N}{4}-1} & S_N^{\frac{N}{4}-1} & & & & \\ & & & & -C_N^{\frac{N}{4}-1} & -C_N^{\frac{N}{4}-1} & -S_N^{\frac{N}{4}-1} & S_N^{\frac{N}{4}-1} \end{bmatrix}\boldsymbol{X}$$

$$\boldsymbol{X}_C=\begin{bmatrix} C_{\frac{N}{2}}^0 & -C_{\frac{N}{2}}^0 & & & C_{\frac{N}{2}}^0 & -C_{\frac{N}{2}}^0 & & \\ C_{\frac{N}{2}}^0 & S_{\frac{N}{2}}^0 & & & -C_{\frac{N}{2}}^0 & -S_{\frac{N}{2}}^0 & & \\ & & \ddots & & & & \ddots & \\ C_{\frac{N}{2}}^{\frac{N}{4}-1} & S_{\frac{N}{2}}^{\frac{N}{4}-1} & & & -C_{\frac{N}{2}}^{\frac{N}{4}-1} & -S_{\frac{N}{2}}^{\frac{N}{4}-1} & & \\ & & & & C_{\frac{N}{2}}^{\frac{N}{4}-1} & S_{\frac{N}{2}}^{\frac{N}{4}-1} & & \\ & & & & & & \ddots & \\ & & & & C_{\frac{N}{2}}^{\frac{N}{4}-1} & -S_{\frac{N}{2}}^{\frac{N}{4}-1} & & \end{bmatrix}\boldsymbol{X}$$

$$\boldsymbol{X}_D=\begin{bmatrix} C_{\frac{N}{3}}^0 & & -C_{\frac{N}{3}}^0 & & -C_{\frac{N}{3}}^0 & & C_{\frac{N}{3}}^0 & \\ & C_{\frac{N}{3}}^0 & & -C_{\frac{N}{3}}^0 & & S_{\frac{N}{3}}^0 & & S_{\frac{N}{3}}^0 \\ & & \ddots & & \ddots & & \ddots & \\ & -C_{\frac{N}{3}}^{\frac{N}{4}-1} & & C_{\frac{N}{3}}^{\frac{N}{4}-1} & & S_{\frac{N}{3}}^{\frac{N}{4}-1} & & S_{\frac{N}{3}}^{\frac{N}{4}-1} \end{bmatrix}\boldsymbol{X}$$

where $\boldsymbol{X}=[x_0\ x_1\ x_2\ \cdots\ x_{N-1}]^T$, and $C_N^k=\cos(2\pi kn/N)$ and $S_N^k=\sin(2\pi kn/N)$. The matrix form of DHT is

$$\begin{bmatrix} Y_{4k} \\ Y_{4k+1} \\ Y_{4k+2} \\ Y_{4k+3} \end{bmatrix} = \begin{bmatrix} H(N/4) & 0 & 0 & 0 \\ 0 & H(N/4) & 0 & 0 \\ 0 & 0 & H(N/4) & 0 \\ 0 & 0 & 0 & H(N/4) \end{bmatrix}\begin{bmatrix} X_A \\ X_B \\ X_C \\ X_D \end{bmatrix} \quad (6)$$

Let $\boldsymbol{Q}(N)=[\boldsymbol{e}_0\ \boldsymbol{e}_4\ \boldsymbol{e}_8\ \cdots\ \boldsymbol{e}_{N-4}\ \boldsymbol{e}_1\ \boldsymbol{e}_5\ \boldsymbol{e}_9\ \cdots\ \boldsymbol{e}_{N-3}\ \boldsymbol{e}_2\ \boldsymbol{e}_6\ \boldsymbol{e}_{10}\ \cdots\ \boldsymbol{e}_{N-2}\ \boldsymbol{e}_3\ \boldsymbol{e}_7\ \boldsymbol{e}_{11}\ \cdots\ \boldsymbol{e}_{N-1}]^T$ be the permutation matrix, and $\boldsymbol{e}_i$ denotes an N×1 column matrix whose (i+1)th element is 1. Then Eq. 6 can be derived as a similar form of Radix-2 DHT [6], as shown in the following.

$$\boldsymbol{Q}(N)\boldsymbol{Y} = \begin{bmatrix} Y_{4k} \\ Y_{4k+1} \\ Y_{4k+2} \\ Y_{4k+3} \end{bmatrix} = \begin{bmatrix} H(N/4) & 0 & 0 & 0 \\ 0 & H(N/4) & 0 & 0 \\ 0 & 0 & H(N/4) & 0 \\ 0 & 0 & 0 & H(N/4) \end{bmatrix}\boldsymbol{P}(N)\boldsymbol{X}, \quad (7)$$

$\boldsymbol{H}_{N/M}(M)$ is the direct sum of N/M M×M matrices, $\boldsymbol{H}(M)$.

$$\boldsymbol{H}_{N/M}(M) = \boldsymbol{H}(M)\oplus\boldsymbol{H}(M)\oplus\cdots\oplus\boldsymbol{H}(M) = \begin{bmatrix} \boldsymbol{H}(M) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}(M) & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{H}(M) \end{bmatrix} \quad (8)$$

Eq. 7 can be simplified as

$$\boldsymbol{Q}_1(N)\boldsymbol{Y}=\boldsymbol{H}_4(N/4)\boldsymbol{P}_1(N)\boldsymbol{X}, \quad (9)$$

which represents that the N-points DHT of $\boldsymbol{X}$ becomes four parts of N/4-points DHT of $\boldsymbol{P}_1(N)\boldsymbol{X}$. And the order of the DHT result, $\boldsymbol{Y}$, is permuted to $\boldsymbol{Q}_1(N)\boldsymbol{Y}$. Repeating the similar decomposition, radix-4 DHT can be gotten as

$$\boldsymbol{Q}_{N/4}(4)\boldsymbol{Q}_{N/16}(16)\cdots\boldsymbol{Q}_1(N)\boldsymbol{Y}=\boldsymbol{P}_{N/4}(4)\boldsymbol{P}_{N/16}(16)\cdots\boldsymbol{P}_1(N)\boldsymbol{X} \quad (10)$$

It shows that the N-points DHT of $\boldsymbol{X}$ can be performed by several multiplications of matrix, $\boldsymbol{P}_{N/M}(M)$, which is corresponding to the ($\log_4 N/M$+1)th stage of the DHT algorithm. And the order of DHT result is bit-reversed of $\boldsymbol{Y}$.

3. CIRCUIT DESIGNS

To realize the 8192-points FFT for VDSL, one radix-2 DHT stage and six radix-4 DHT stages are needed as shown in Fig. 1. Besides, an external dual-port RAM and two adders are required to perform the rearrangement of bit-reversed DHT results and acquire the normal-order FFT results [7].

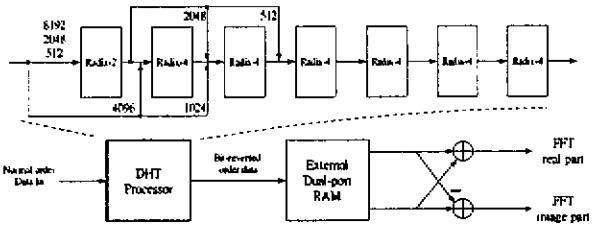


Fig. 1. Variable-length DHT-based FFT processor for 8192/4096/2048/1024/512-points VDSL and 512-points ADSL.

3.1. Radix-2 PE

Based on the matrix form of radix-2 DIF DHT algorithm [6],

$$Z = Q_{N/2}(2)Q_{N/4}(4)\cdots Q_1(N)Y = P_{N/2}(2)P_{N/4}(4)\cdots P_1(N)X, \quad (11)$$

the X column matrix denotes the sequential input data in time domain, and the Z column matrix indicates the sequential output data in bit-reversed order. The matrix, $P_{N/M}(M)$, is corresponding to the $(\log_2 N/M+1)$th stage PE. And it has the general form as Eq. 12 [6]. Fig. 2 shows the proposed circuit architecture of radix-2 DIF DHT PE, $P_{N/M}(M)$. Due to the utilization of the two-phase operating technique, only one real multiplier is employed to perform two multiplications.

$$P(M) = \begin{bmatrix} I_{M/2} & 0 \\ 0 & K(M/2) \end{bmatrix} \begin{bmatrix} I_{M/2} & I_{M/2} \\ I_{M/2} & -I_{M/2} \end{bmatrix} \quad (12)$$

$$K(M/2) = C(M/2) + S(M/2)$$

$$= \begin{bmatrix} C_M^0 & & & \\ & C_M^1 & & \\ & & \ddots & \\ & & & C_M^{M/2-1} \end{bmatrix} + \begin{bmatrix} S_M^0 & & & \\ & & & S_M^1 \\ & & \iddots & \\ & S_M^{M/2-1} & & \end{bmatrix}$$

3.2. Radix-4 PE

The matrix form of radix-4 DIF DHT algorithm is followed as Eq. 10 and the matrix, $P_{N/M}(M)$, is corresponding to the $(\log_4 N/M+1)$th stage PE. As the sequential input data, X, enter each PE stage, the input data are rearranged by the circuit as shown in Fig. 3. The 16-points radix-4 DHT is demonstrated in the following.

$$Z = Q_4(4)Q_1(16)Y = P_4(4)P_1(16)X, \quad \text{where} \quad (13)$$

$$P_1(16) = \begin{bmatrix} I & I & I & I \\ B & C & -B & -C \\ A & -A & A & -A \\ D & E & -D & -E \end{bmatrix}$$

$$I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} C_{16/2}^0 & 0 & 0 & 0 \\ 0 & C_{16/2}^1 & 0 & S_{16/2}^1 \\ 0 & 0 & C_{16/2}^2 + S_{16/2}^2 & 0 \\ 0 & S_{16/2}^3 & 0 & C_{16/2}^3 \end{bmatrix}$$

$$B = \begin{bmatrix} C_{16/1}^0 & 0 & 0 & 0 \\ 0 & C_{16/1}^1 & 0 & C_{16/1}^1 \\ 0 & 0 & C_{16/1}^2 + C_{16/1}^2 & 0 \\ 0 & C_{16/1}^3 & 0 & C_{16/1}^3 \end{bmatrix}, \quad C = \begin{bmatrix} C_{16/1}^0 & 0 & 0 & 0 \\ 0 & -S_{16/1}^1 & 0 & S_{16/1}^1 \\ 0 & 0 & S_{16/1}^2 - S_{16/1}^2 & 0 \\ 0 & S_{16/1}^3 & 0 & -S_{16/1}^3 \end{bmatrix}$$

$$D = \begin{bmatrix} C_{16/3}^0 & 0 & 0 & 0 \\ 0 & C_{16/3}^1 & 0 & -C_{16/3}^1 \\ 0 & 0 & C_{16/3}^2 - C_{16/3}^2 & 0 \\ 0 & -C_{16/3}^3 & 0 & C_{16/3}^3 \end{bmatrix}, \quad E = \begin{bmatrix} -C_{16/3}^0 & 0 & 0 & 0 \\ 0 & S_{16/3}^1 & 0 & S_{16/3}^1 \\ 0 & 0 & S_{16/3}^2 + S_{16/3}^2 & 0 \\ 0 & S_{16/3}^3 & 0 & S_{16/3}^3 \end{bmatrix}$$

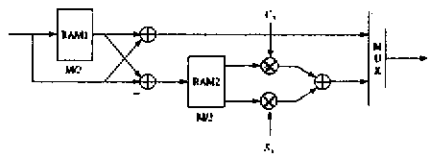


Fig. 2. Radix-2 DIF DHT PE.

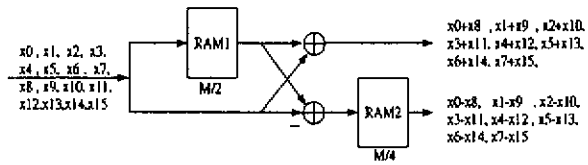


Fig. 3. Rearrangement of each stage input (e.g. M=16, for $P_1(16)$).

After the rearrangement circuit, the PE, $P_1(16)X$, becomes two simplified matrices shown as Eq. 14 and 15.

$$P_1(16)X = \begin{bmatrix} a_{4k} \\ a_{4k+1} \\ a_{4k+2} \\ a_{4k+3} \end{bmatrix}, \quad \text{where}$$

$$\begin{bmatrix} a_{4k} \\ a_{4k+2} \end{bmatrix} = \begin{bmatrix} I & I \\ A & -A \end{bmatrix} \begin{bmatrix} x_0 + x_8 & x_1 + x_9 & \cdots & x_7 + x_{15} \end{bmatrix}^T \quad (14)$$

$$\begin{bmatrix} a_{4k+1} \\ a_{4k+3} \end{bmatrix} = \begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} x_0 - x_8 & x_1 - x_9 & \cdots & x_7 - x_{15} \end{bmatrix}^T \quad (15)$$

In Eq. 14, this part of $P_1(16)X$ is the same as the general form of radix-2 DIF DHT PE derived in Eq. 12. The similar circuit architecture shown in Fig. 2 can be utilized. In Eq. 15, this part of $P_1(16)X$ is performed by the circuit architecture as shown in Fig. 4. The two-phase multiplication of single multiplier technique is also employed.

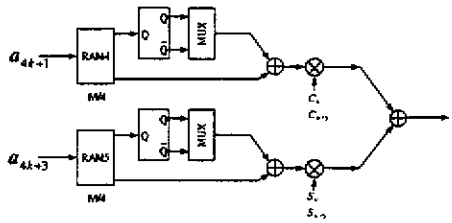


Fig. 4. Circuit architecture for computing a_{4k+1} and a_{4k+3}.

The circuits in Fig. 2 and 4 for radix-4 DHT use the same multiplier because of performing the multiplication at different cycles. By rearranging the computing order as a_{4k}, a_{4k+2}, a_{4k+1} and a_{4k+3}, the bit-reversed order in Z can be obtained.

As computing different points of radix-4 DHT, only the size of M in Fig. 2 and 4 is changed. The radix-4 DHT PE can be performed with latency of $3M/4$ cycles by only one real multiplier, seven real adders, $3M/4$ RAM as delay line, $3M/4$ RAM with double clock rate, two ROMs as sine and cosine coefficients, and some simple logic circuits.

4. VLSI IMPLEMENTATION AND PERFORMANCE SUMMARY

For a variable-length DHT-based FFT processor, several considerations in VLSI implementation such as variable-length, simplified ROM of sin/cos tables, cyclic extension removing and fixed-points evaluation for SNR requirements of VDSL system are all taken into account in this chip. The radix-4 DIF DHT processor is implemented with 0.25μm 1P5M typical CMOS technology. Chip design and simulation summary are described in Table 2, and chip layout is shown in Fig. 5. Compared in Table 1, it shows that the proposed radix-4 DHT-based FFT architecture has the fewest numbers of real multipliers among SDF architectures and radix-2 DHT-based FFT processor [5]. Otherwise, they use almost the same numbers of registers. Table 3 shows the comparisons with other 8192-points FFT chips.

Table 2. Chip design and simulation summary.

Process	CMOS 0.25μm 1P5M
Total Area RAM Area	34mm^2, 25mm^2 (core) 5.2mm^2
Maximum Speed	50MHz
Latency	8221@N=8192
Throughput	1 sample/cycle
Power Consumption	300mW@35MHz
Package	CQFP144
Supply Voltage	2.5V

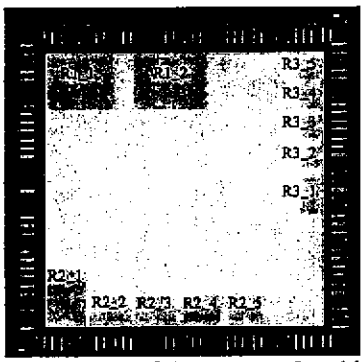


Fig. 5. Chip layout of the proposed architecture.

5. CONCLUSION

In this paper, the radix-4 DHT-based FFT algorithm is first derived according to the Hermitian symmetric property of the data in DMT-VDSL and ADSL systems, and the matrix form of the radix-2 DHT-based FFT designs [5][7]. Two PE circuits are proposed for variable-length application. One is composed of only one real multiplier and three adders for radix-2. The other is comprised with only one real multiplier and seven adders for radix-4. Finally, a variable-length 8192/4096/2048/1024/512 radix-4 DIF DHT processor is demonstrated. It meets the DMT-VDSL and ADSL requirements and uses only six real multipliers, 2.01N (16467) registers and some simple logic circuits. The chip is designed using 0.25μm 1P5M typical CMOS technology and occupies a core area of 5mm×5mm. Power dissipation is 300mW under 2.5V supply voltage while running at 35MHz.

Table 3. Comparison with other FFT chips.

	[9]	[10]	This Work
FFT Size	8192	8192	8192/4096/2048 /1024/512
Technology	0.5μm CMOS	0.6μm CMOS	0.25μm CMOS
Chip Area (core)	1cm^2	140mm^2 107mm^2	34mm^2 25mm^2
Normalized Area (core)	25mm^2	24.3mm^2 18.6mm^2	34mm^2 25mm^2
Clock	20MHz	20MHz	35MHz
Power Consumption	600mW	650mW	300mW
Word length: Input/Output sin, cos Internal results	 10/12 10 12 (bits)	 10/12 10 12 (bits)	 22/22 20 22 (bits)
Register	Delay- line	RAM	SRAM

REFERENCES

[1] D. J. Rauschmayer, *ADSL/VDSL Principles*, MTP, 1999.

[2] A. V. Oppenheim and R. W. Schafer, *Discrete Time Signal Processing*, Prentice Hall, 1999.

[3] S. He and M. Torkelson, "Design Pipeline FFT Processor for OFDM (de)Modulation," *Int. Symp. on Signal, System, and Electronics* , pp.257-262, 1998

[4] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Procession*, Prentice Hall, Englewood Cliffs, New Jersey.

[5] C. L. Wang and C. H. Chang, "A DHT-based FFT/IFFT Processor for VDSL Transceivers," *Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, 2001.

[6] C. L. Wang, C. T. Ho and Y. T. Chang, "A novel systolic design for fast computation of the discrete Hartley transform," *IEEE Thirtieth Asilomar Conf. on Signals, Systems & Computers*, pp. 1067-1071, Nov. 1996.

[7] C. L. Wang and C. H. Chang, "A novel DHT-based FFT/IFFT processor for ADSL transceivers," *Int. Symp. on Circuits and Systems* , Vol. 1, pp. 51 -54, 30 May, 1999.

[8] H. V. Sorensen, D. L. Jones, C. S. Burrus and M. T. Heideman, "On Computing the Discrete Hartley Transform", *IEEE Trans. On Acoustic, Speech and Signal Processing*, Vol. 33, pp. 1231-1238, Oct. 1985.

[9] E. Bidet, D. Castelain, C. Joanblanq, and P. Senn, "A Fast Single-Chip Implementation of 8192 Complex Point FFT", *IEEE Journal of Solid-State Circuits*, Vol. 30, No. 3, Mar. 1995

[10]L. Jia, "A New VLSI-Oriented FFT Algorithm and Implementation," *IEEE ASIC Conference*, pp. 337-341, 1998

IEEE 802.11a WLAN TRANSCEIVER USING SELF-CORRECTING DIGITAL TIMING RECOVERY DESIGN AND IMPLEMENTATION

Wei-Hsiang Tseng, Ching-Chi Chang and Chorng-Kuang Wang

Graduate Institute of Electronic Engineering, and Department of Electrical Engineering,
National Taiwan University, Taipei, Taiwan, R.O.C.
E-mail: ckwang@cc.ee.ntu.edu.tw

ABSTRACT

This paper presents an OFDM baseband transceiver for wireless LAN IEEE 802.11a implemented in 0.25μm CMOS technology. The architecture of the interpolator with the sliding window is proposed to realize the self-adjustment of sample skipping or duplicating that prevents the conventional digital timing recovery from fatal jitter problem caused by illegal indexing. In the receiver, the design of the symbol boundary detection, carrier recovery, timing recovery, FFT/IFFT, and equalizer are realized to deliver 10% packet error rate requirement under all specified SNRs. This chip occupies 3.5x3.5 mm^2 chip area and consumes 109 mW with 2.5 V power supply.

1. INTRODUCTION

Due to the demand for wide-band communication, it leads to a tremendous growth of wireless LAN. One of the main challenges in the wireless LAN is the indoor propagation characteristics. In 1997, the IEEE 802.11 standard was first established for the wireless LAN. It supports the data rates of 1Mbps and 2Mbps which are not sufficient for many multimedia applications. The IEEE 802.11a standard was proposed for high bandwidth efficiency, which selects orthogonal frequency division multiplexing (OFDM) as the basis for the physical layer and supports the data rates from 6 up to 54 Mbps in the 5 GHz band. OFDM is an excellent solution to combat burst noises and deal with multi-path fading robustly.

A conventional digital timing recovery using interpolator with long buffer can solve the problem of sample stuffed or rubbed when the timing offset occurs [5]. However, the buffer size may be infinite for non-stopping transmission and it is impractical. The proposed self-correcting digital timing recovery performs the sample skip/duplicate in the guard interval and can apply to both OFDM and Discrete-multi-tone (DMT) systems.

This paper is organized as follows. The baseband transceiver architecture is presented in Section 2. In Section 3, the circuits of the functional blocks are given. The simulation environment and the evaluation of the whole system are described in Section 4. Section 5 delivers the chip layout and the post-layout summary. Finally, the conclusion is drawn in Section 6.

2. TRANSCEIVER ARCHITECTURE

The proposed baseband transceiver architecture is shown in Figure 1. The upper part of the transceiver architecture is the transmitter, and the other is the receiver. The transmitter consists of the training sequence generator, QAM mapping, pilot insertion, IFFT, adding cyclic prefix, windowing and wave shaping filter. The wave shaping filter adopts raised cosine filter with roll-off factor 0.22. The design of the transmitter follows the IEEE 802.11a standard. It can send the regulated packet format and comply with the transmit spectrum mask requirement. The receiver design comprises the following functional blocks: interpolator, derotator, delay correlator, match filter, removing cyclic prefix, FFT, channel estimation, frequency-domain equalizer (FEQ), decision block, symbol boundary detection, carrier frequency offset (CFO) estimation, carrier recovery loop and timing recovery loop. The demodulator, FFT, is hardware-shared with the modulator, IFFT.

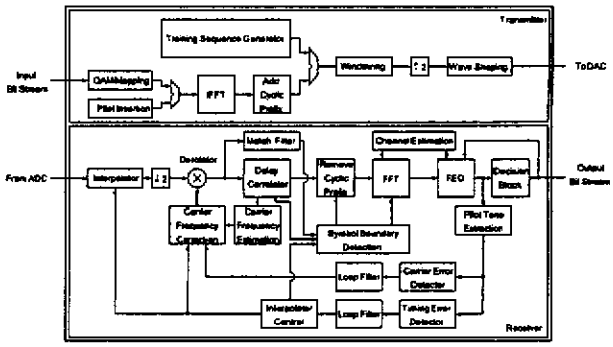


Fig. 1. Transceiver architecture

3. RECEIVER DESIGN

3.1. Training Structure

The training structure in the design of the receiver is shown in Figure 2 [1]. The preceding part of the short preamble, about 4.8 μs, is reserved for signal detection and AGC acquisition. At the end of the short preamble, the coarse symbol boundary detection and the CFO are finished. Then, the fine symbol boundary detection, the fine CFO estimation and the channel estimation are proceeding in the duration of the long preamble. The residual

This program is supported by the National Science Council (NSC), Taiwan, R.O.C.

CFO, the timing frequency offset (TFO) and the coefficients of the FEQ are kept tracking after the training sequences.

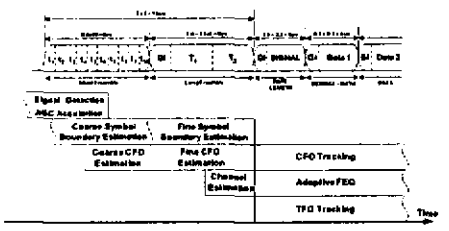


Fig. 2. Training structure of the receiver design

3.2. Symbol Boundary Detection

The symbol boundary detection consists of two steps. First, the delay correlator is adopted to detect the course symbol boundary since the characteristics of the short preamble are ten identical repetitions. A mechanism based on the binary search is applied to track the maximum of the correlation value. Second, the match filter is exploited to find the fine symbol boundary in the session of the long preamble.

3.3. Carrier Recovery

OFDM is very sensitive to the CFO, and there are two detrimental effects; one is the reduction of signal amplitude with phase shift and another is the inter-carrier inference (ICI) from the other carriers caused by loss of the orthogonal property [2]. Because the effects are greatly serious, it needs to be recovered before the demodulation. The carrier recovery is composed of three stages. Based on the maximum likelihood estimation [3], the coarse CFO estimation is achieved in the period of the short preamble, and the estimated CFO is compensated at the end of the short preamble by the aid of the 16-taps delay correlator. Similarly, in the long preamble, the fine CFO estimation and compensation are completed by the aid of the 64-taps delay correlator. A summary of the CFO estimation is listed in Table 1. By means of the pilot signal after the preambles, the remaining CFO is traced by a phase-locked loop. The PI filter is selected as the loop filter as shown in Figure 3. Table 2 lists the design parameters of the carrier recovery loop. To replace the multiplier with the shifter, the coefficients of PI filter are truncated to the power of 2. In our design, the effects due to the CFO can be suppressed within 0.1 dB SNR degradation.

Table 1. Summary of the CFO estimation

CFO Estimation Type	Coarse	Fine
Estimation Range	± 625 KHz	± 156 KHz
Maximum Estimation Error	± 32.6 KHz	± 6.22 KHz

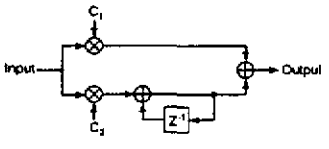


Fig. 3. PI filter

Table 2. Design parameters of the carrier recovery loop

Parameter	Value
Lock-in Range $\Delta\omega_L$	7.5 KHz
Damping Factor ξ	0.707
Natural Frequency ω_n	3.3×10^4 rad/s
Phase Detector Gain K_d	1 unit/rad
NCO Gain K_o	20 MHz/unit
PI Filter Coefficient C_1	2^{-11}
PI Filter Coefficient C_2	2^{-14}

3.4. Self-correcting Timing Recovery

The all-digital method is selected for the solution of the timing recovery. A phase-locked loop is also used to track the TFO, and the piecewise-parabolic interpolator in Farrow structure is chosen for compensating the sampling mismatch [4]. Due to the mismatch of the sampling clock, the interpolated phase drifts. Therefore, the skip or duplicate process in the cyclic prefix is required to reset the interpolating phase [5]. To combine the interpolator and the skip-and-duplicate process in view of VLSI, the architecture of the interpolator with a sliding window is proposed in Figure 4, and the hardware implementation is given in Figure 5. The PI filter is chosen as the loop filter in the timing recovery loop. Table 3 lists the parameters of the timing recovery loop.

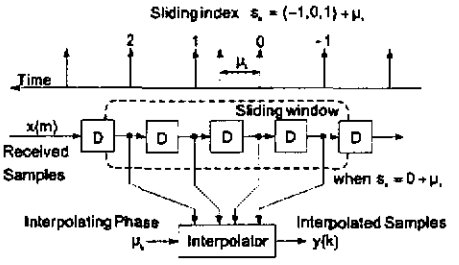


Fig. 4. Interpolator with the proposed sliding window

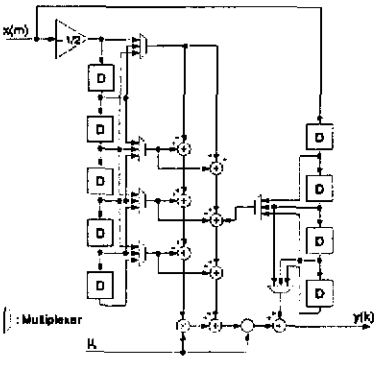


Fig. 5. Implementation of the interpolator with the proposed sliding window

Table 3. Design parameters of the timing recovery loop

Parameter	Value
Lock-in Range $\Delta\omega_L$	1.9 KHz
Damping Factor ξ	0.707
Natural Frequency ω_n	8.5×10^3 rad/s
Phase Detector Gain K_d	0.18 unit/rad
NCO Gain K_0	40 MHz/unit
PI Filter Coefficient C_1	2^{-12}
PI Filter Coefficient C_2	2^{-17}

3.5. FFT/IFFT

The 64-point pipeline FFT/ IFFT selects the single-path delay feedback (SDF) architecture for hardware implementation. The radix-2/4/8 algorithm is adopted due to the regularity [6]. For a 64-point FFT, it can be divided into two stages, and only one complex multiplier is required. Due to the property of FFT and IFFT, the operation of the IFFT can be derived by conjugating the input of the FFT, then computing the function of the FFT, and conjugating the output of the FFT which is scaled by 1/N. By such property, the FFT/ IFFT processor can be shared by the transmitter and the receiver for modulation and demodulation, respectively. The SQNR of FFT/IFFT is 46.8 dB.

3.6. Frequency-domain Equalizer

The frequency-domain equalizer (FEQ) exploits the Minimum Mean Square Error (MMSE) criterion to perform the channel estimation as [7]

$$C_K = H_K^* / \{|H_K|^2 + \sigma_n^2 / \sigma_s^2\} \quad (1)$$

where C_K is the coefficient of the equalizer over K-th carrier, the channel frequency response over the K-th carrier is H_K, σ_n^2 is the variance of the additive noise, and σ_s^2 is the variance of the transmitted symbols. The least-mean-square (LMS) algorithm for adaptive tracking that is given by [8]

$$C_K^{(j+1)} = C_K^{(j)} + \Delta \varepsilon_K Y_K^* \quad (2)$$

where $C_K^{(j)}$ is the j-th coefficient over K-th carrier, $C_K^{(j+1)}$ is the $(j+1)$-th one, Y_K^* is the conjugate of the demodulated output Y_K from FFT, ε_K is the decision error, and Δ is a scale factor that controls the rate of the adjustment. With LMS algorithm, it can further provide about 3 dB SNR improvement after the equalization.

4. SIMULATION RESULTS

Multipath fading, CFO, TFO, and AWGN are taken into consideration in the simulation environment as Figure 6 [9]. The indoor multipath channel model adopts Saleh's channel model [10].

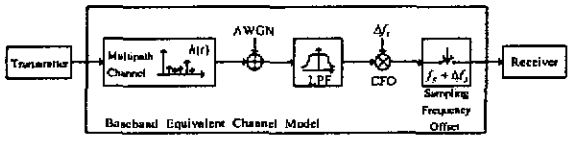


Fig. 6. Simulation environment [9]

The packet error rate (PER) is the indication to evaluate the system performance. In IEEE 802.11a standard, a PER less than 10% is required over the transmission at a PSDU length of 1000 bytes [1]. The simulation contains the cases of the multi-path delay spread equal to 0ns, 50ns, and 100 ns with the maximal tolerable CFO and TFO. The PER performance versus SNR for the delay spread 0 ns, 50ns, and 100 ns are shown in Figure 7, Figure 8 and Figure 9, respectively. Table 4 lists the SNR requirement for 10% PER.

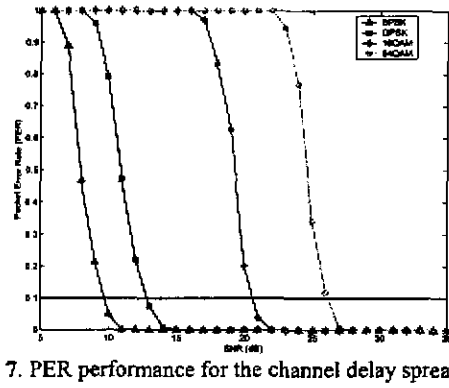


Fig. 7. PER performance for the channel delay spread 0 ns

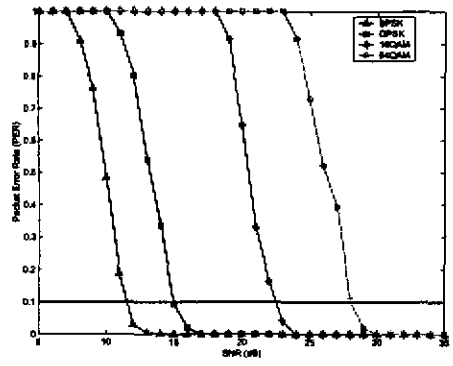


Fig. 8. PER performance for the channel delay spread 50 ns

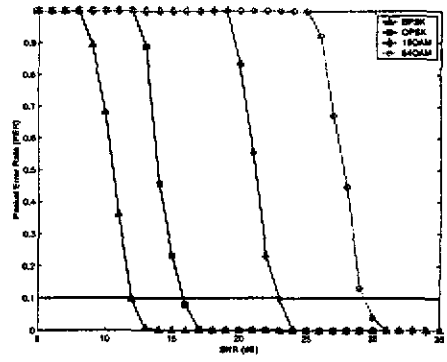


Fig. 9. PER performance for the channel delay spread 100 ns

Table 4. SNR requirement for delay spread 0 ns, 50 ns, and 100 ns

SNR Requirement for PER < 10% (without coding)			
Modulation \ Delay Spread	0 ns	50 ns	100 ns
BPSK	9.7 dB	11.6 dB	12.0 dB
QPSK	12.8 dB	15.0 dB	15.9 dB
16QAM	20.7 dB	22.6 dB	23.0 dB
64QAM	26.2 dB	28.0 dB	29.3 dB

5. CHIP LAYOUT AND POST-LAYOUT SUMMARY

Figure 10 shows the chip layout, and the major functional blocks are also marked. Total gate counts are about 302K. Using cell-based design methodology, this chip is implemented in 0.25μm CMOS technology and occupies 3.5x3.5 mm^2 chip area. With 2.5 V supply voltage, it consumes power dissipation 109 mW. The post-layout summary is listed in Table 5.

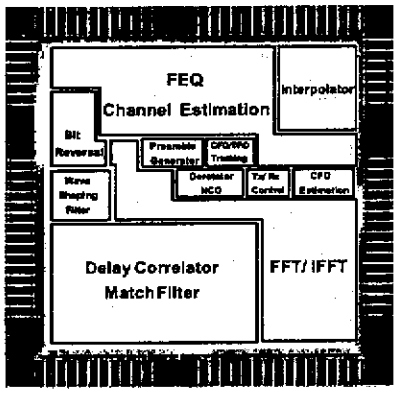


Fig. 10. Chip layout

Table 5. Chip post-layout performance

Data Rate	6, 9, 12, 18, 24, 36, 48, 54 Mbps
Modulation	OFDM (BPSK, QPSK, 16QAM, 64QAM)
Tolerance of CFO	+ 232.2 KHz
Tolerance of TFO	+ 1.6 KHz
Technology	0.25μm CMOS
Chip Area	3.5 x 3.5 mm^2
Gate Counts	302189
Clock Rate	40 MHz
Power Dissipation	109 mW
Supply Voltage	2.5 V
DA/AD Resolution	10 Bits

6. CONCLUSION

In this paper, an OFDM baseband transceiver architecture compliant with IEEE 802.11a standard is developed for wireless LAN communication. In the receiver design, the symbol synchronization has two steps for symbol boundary detection, and the carrier recovery has three stages for acquisition and tracking. The timing recovery utilizes the all-digital method, and an improved architecture of the interpolator with a sliding window is proposed for VLSI implementation in OFDM system. By MMSE and LMS algorithms, the FEQ is well-designed. 10% PER also can be delivered at a specified SNR. In 0.25μm CMOS process, this transceiver is implemented by 3.5x3.5 mm^2 chip area. The power consumption is 109 mW at 2.5 V power supply.

7. REFERENCES

[1] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications for High-speed Physical Layer in the 5 GHz Band. *IEEE std. 802.11a, Nov. 1999.*

[2] P. H. Moose, "A Technique for Orthogonal Frequency Division Multiplexing Frequency Offset Correction," *IEEE Trans. on Comm.*, Vol. 42, No. 10, pp. 2908-2914, Oct. 1994.

[3] J. J. van de Beek et al., "ML Estimation of Time and Frequency Offset in OFDM Systems," *IEEE Trans. on Signal Processing*, Vol. 45, No. 7, pp. 1800-1805, July 1997.

[4] L. Erup et al., "Interpolation in Digital Modems-Part II: Implementation and Performance," *IEEE Trans. on Comm.*, Vol. 41, No. 6, pp. 998-1008, June 1993.

[5] T. Pollet and M. Peeters, "Synchronization with DMT Modulation," *IEEE Comm. Mag.*, Vol. 37, No. 4, pp. 80-86, Apr. 1999.

[6] L. Jia et al., "A New VLSI-Oriented FFT Algorithm and Implementation," *IEEE ASIC Conf.*, pp. 337-341, Sept. 1998

[7] H. Sari et al., "Transmission Techniques for Digital Terrestrial TV Broadcasting," *IEEE Comm. Mag.*, Vol. 33, No. 2, pp. 100-109, Feb. 1995.

[8] J. G. Proakis, *Digital Communication*, 4th ed., McGraw Hill, 2001

[9] C-F Hsu, Y-H Huang and T-D Chiueh, "Design of an OFDM Receiver for High-speed Wireless LAN," *ISCAS*, Vol. 4, pp. 558-561, May 2001.

[10] A. A. M. Saleh and R. A. Valenzuela, "A Statistical Model for Indoor Multipath Propagation," *IEEE J. Selected Areas in Comm.*, Vol. SAC-5, No. 2, pp. 128-137, Feb. 1987.

Inductor-Aided Gain, Noise Figure, and IIP3 Enhancement Techniques for Active Mixer and the Application in WLAN

Chun-Chih Hou, Ching-Chi Chang, and Chorng-Kuang Wang

Department of Electrical Engineering and Graduate Institute of Electronics Engineering,
National Taiwan University, Taipei, Taiwan 10617, R.O.C.

Abstract—This paper presents the usage of spiral inductors in an active mixer. The influence of harmonic termination on the performance of mixer is discussed. Then, a technique of preventing the upconversion of flicker noise in a half-IF mixer is proposed. Simulation and measurement results are employed to verify this approach, and a dual-band WLAN receiver architecture that combines the features of half-IF and dual-conversion is proposed.

Index Terms—Flicker noise, half-IF, IIP3, inductor, mixer.

I. Introduction

There are three important parameters for the design of RF circuits. They are gain, noise figure, and IIP3. The gain of each component in a RF receiver chain should conform to the gain budget defined in the system design level. Higher gain is usually preferred for noise figure consideration, but too much high gain would distort the signal. For a wireless system, the received signal may contain in-band and out-of-band interferences. The nonlinearity of amplifiers would generate undesired harmonics from these interferences and may degrade the signal-to-noise ratio (SNR). The IIP3 is a parameter to evaluate the linearity of an amplifier, and should be as high as possible for good out-of-channel rejection.

For an active mixer, it has the advantage of higher conversion gain at the cost of lower linearity, when compared with passive mixers. A linearity enhancement technique based on harmonic tuning and multiple gating [1] is reviewed in section II. The commonly used folding structure also has good effects upon the linearity and noise figure performance, and is also discussed in section II.

In a half-IF receiver, the RF signal is first down-converted to half the RF frequency and then to zero frequency. It has the advantage of high image rejection ratio, but the flicker noise would be converted to the signal band if a normal mixer were used [2]. In section III, a circuit technique based on inductive termination at the load of the transconductance stage of fully differential mixers is used to prevent the flicker from worsening the noise figure. Simulation and measurement results are used to verify this circuit topology.

As for the published dual-band architectures, most of them receive the two-band signal by two signal paths and the direct-conversion architecture is mostly adopted, too [5-7]. In this paper, a new architecture, combining the features of dual-conversion and half-IF, is proposed that can receive signal at two-band, 5.8GHz and 2.4GHz, with the same receiver chain. Due to the elaborate frequency conversion planning, the troublesome image problem is relaxed in this design. The detail structure is described in section IV.

II. Folding Structure and Harmonic Tuning

For the design of active mixers, folding is a commonly used technique, as shown in Fig. 1. It has three advantages. First, the maximum number of cascoding MOSFETs is only two. Not only can it be a low-voltage design, but also the linearity performance is better than non-folding mixers. Second, the current flowing through M3 and M4 is usually small, so the required local oscillator (LO) swing is small. Finally, if the mixer is designed to down-convert signal to DC, the noise figure around zero frequency will be lower because of smaller flicker noise of low-current M3 and M5.

The LC tank composed of *L1* and *C1* resonates at frequency ω (RF frequency), and is short at 2nd harmonic frequency 2ω and sub-harmonic frequency $\Delta\omega$. It is known in [1] that 2ω and $\Delta\omega$ at common source noise of M3 and M4 worsen the linearity in conventional single balanced mixer. The LC tank can thus reduce the peaks at these harmonics and thus improve mixer's linearity.

III. Noise Figure Enhancement Technique for Half-IF Mixers

In [2], a mixer structure is proposed to prevent the up-conversion of flicker noise of the transconductance stage to IF. However, the consideration of the flicker noise in LO switching pair is also important. The proposed mixer topology shown in Fig. 2 uses an LC tank to achieve the filtering function. The resonant frequency of the LC tank is at the signal frequency such that it blocks the

This program is supported by the National Science Council (NSC), Taiwan, R.O.C.

signal from passing through it. If the quality factor of the tank is high enough, the conversion gain of the mixer will increase because the signal loss due to parasitic capacitance at the common-source node of the switching pair is reduced. As for the flicker noise around zero frequency, it sees a short because of the very low impedance around DC of the LC tank. The circuit can thus be reduced to the half circuit as shown in Fig. 3 when only flicker noise is considered. The topology of mixer is destroyed, and thus the flicker noise in neither the transconductance stage nor the switching stage will be up-converted to IF.

The circuit simulator of ADS [3] is used to simulate noise figure of circuits. A LNA in front of the mixer is concatenated in this simulation. When no LC tank is inserted in the LO switching pair, the simulation results is shown in Fig. 4. The up-converted flicker noise has a significant influence on the noise figure. After the LC tank is inserted, the influence from the flicker noise vanishes and the in-band noise figure reduces by more than 1 dB, as shown in Fig. 5. The circuit has also been realized in 0.18um CMOS 1P6M process and measured by ROHDE&SCHWAEZ FSQ 26 signal analyzer. The experimented result of noise figure is shown in Fig. 6. Though the noise figure is not as low as that of simulation due to some circuit loss, it shows that the proposed circuit can prevent the flicker noise to be up-converted. The simulated flicker noise floor is larger than 20 MHz in Fig. 4, and this phenomenon is not observed in Fig. 6. The occurrence of the peak when IF is equal to LO frequency is due to the LO-IF feedthrough of the mixer.

IV. Combined Dual-Conversion and Half-IF Receiver

Using the proposed mixer in section III in a half-IF receiver, and combing it with dual-conversion receiver chain, a dual-band receiver for WLAN is realized, as shown in Fig. 7. The components inside the dashed frame can be implemented into a single chip, and those outside the frame are the lumped components, which are switch, diplexer, and band-select band-pass filters. Two band-pass filters instead of one are used in order to provide sufficient rejection of the out-of-band interferences for each band. When the desired signal is at 5.8 GHz, the frequency conversion method, shown in Fig. 8., is used [4]. The 3.465GHz image is 2 GHz far away from the desired signal, and it can be filtered out by the band-pass nature of the on-chip amplifiers. The first LO frequency, which is 4.62 GHz, is four times of the second LO, and thus they can be synthesized by a single voltage controlled oscillator (VCO) and a divide-by-four circuit. Because the VCO frequency is different from the signal frequency, the LO pulling problem will not occur.

When the desired signal is at 2.4 GHz, the frequency conversion method shown in Fig. 9 is used. The utilized 1.2 GHz LO frequency can be obtained by the same circuits composed of one VCO with 4.8-5 GHz tuning frequency and a divide-by-four circuit. As shown in Table I, the VCO frequency range of the two bands is close, so a VCO with wide tuning range can deliver the required LO frequencies for dual-band application. At the same time, the 90° phase shifter can be eliminated due to the utilization of dividers. Another benefit of this dual-band receiver is that the IF frequencies are close for this dual-band signal, so there is an ease to implement the circuits after the first down-conversion. As for the image signal around the zero frequency, the attenuation contributed by amplifiers and lumped components such as antenna and band-select filter is very large, and thus the image rejection ratio can be as high as 60dB [2]. Two problems that should be taken into account are the upconversion of flicker noise to IF frequency and the DC-offset that comes from LO-to-IF feedthrough. The circuits proposed in the previous section deal with the first problem. For the DC-offset problem, a high-pass filtering offset-cancellation circuit can solve it [2]. The SNR degradation due to the filtering is not significant because for IEEE 802.11a/g, the OFDM modulation leaves the center sub-carrier empty, and for IEEE 802.11b, spread-spectrum modulation is used.

V. Conclusions

In this paper, the overview of the features provided by the inductor at the common source node of LO switches of a mixer are given. Inductors can help to peak the conversion gain and improve the linearity. The resulting folding structure can reduce the LO swing, and decrease the amount of flicker noise. For a half-IF receiver, a mixer topology is proposed to prevent the flicker noise to become in-band noise.

Acknowledgement

The authors would like to thank to the technical support by Chip Implementation Center (CIC), Taiwan, R.O.C.

References

[1] Tae Wook Kim, Bonkee Kim, and Kywro Lee, "Highly linear RF CMOS amplifier and mixer adopting MOSFET transconductance linearization by multiple gated transistors," in *IEEE*

Radio Frequency Integrated Circuits (RFIC) Symp. Dig., June 2003, pp. 107-110.

[2] B. Razavi, "A 5.2-GHz CMOS receiver with 62-dB image rejection," *IEEE J. Solid-State Circuits*, vol. 36, pp. 810-815, May 2001.

[3] Agilent Technologies, http://www.agilent.com.

[4] D. Su, M. Zargari, P. Yue, S. Rabii, D. Weber, B. Kaczynski, S. Mebta, K. Singh, S. Mendis and B. Wooley, "A 5GHz CMOS Transceiver for IEEE 802.11a Wireless LAN," in *IEEE Int. Solid-State Circuits Conf. Tech. Dig.*, Feb. 2002, pp. 70-405.

[5] M. Hotti, J. Kaukovuori, J. Ryynanen, K. Kivekas, J. Jussila, and K. Halonen, K, "A Direct Conversion RF Font-end for 2-GHz WCDMA and 5.8-GHz WLAN Applications," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 45-48, June 2003.

[6] B. U. Klepser, M. Punzenberger, T. Ruhlicke, and M. Zannoth, "5-GHz and 2.4-GHz Dual-band RF Transceiver for WLAN 802.11a/b/g Applications," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 37-40, June 2003.

[7] M.Y. Wang, R.R.-B. Sheen, D.T.-C. Chen, and R.Y.J. Tsen, "A Dual-band RF Front-end for WCDMA and GPS Applications," in *Proc. 2002 Int. Symp. Circuits and Systems*, vol.4, 2002, pp. 113-116.

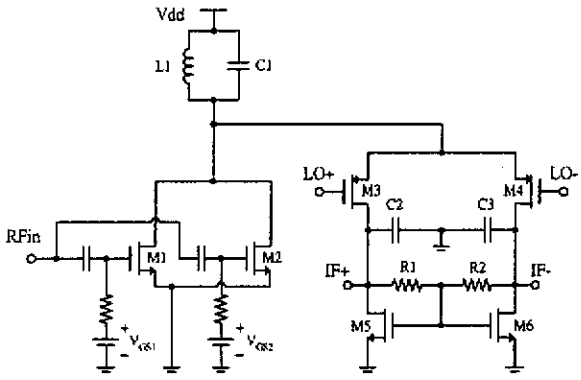


Fig. 1. Schematic of harmonic tuned MGTR mixer [1].

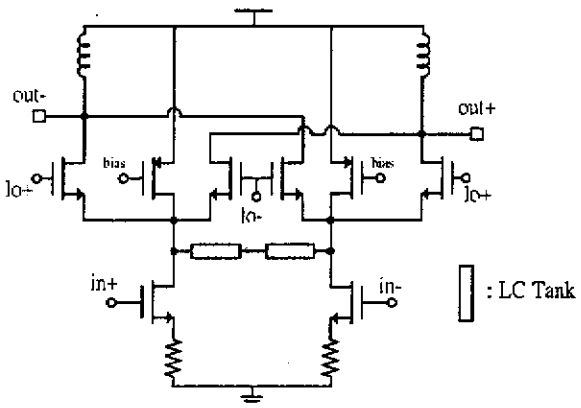


Fig. 2. Schematic of the first mixer in a half-IF receiver.

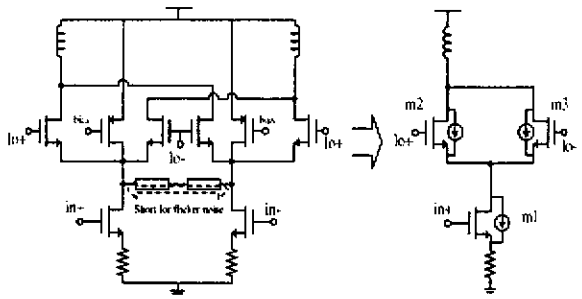


Fig. 3. Filtering mechanism for flicker noise.

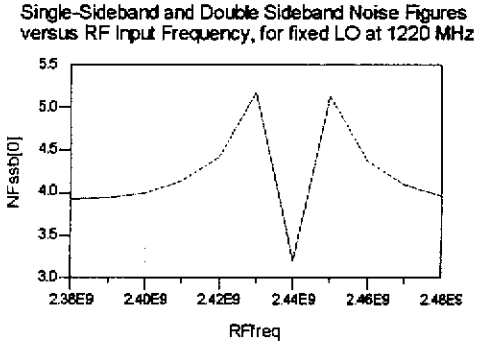


Fig. 4. Simulated noise figure of 2.4 GHz LNA+Mixer, without LC tank inserted.

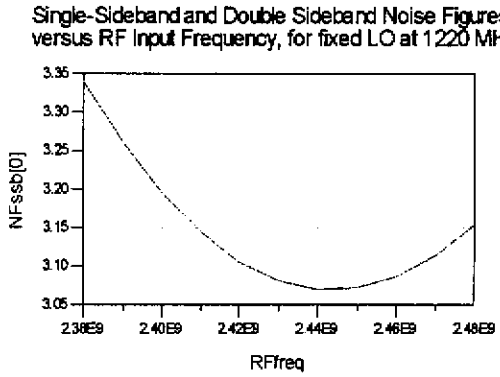


Fig. 5. Simulated noise figure of 2.4 GHz LNA+Mixer, with LC tank inserted.

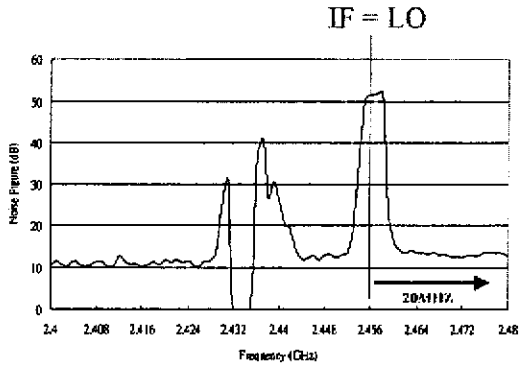


Fig. 6. Measured noise figure of 2.4 GHz LNA+Mixer, with LC tank inserted.

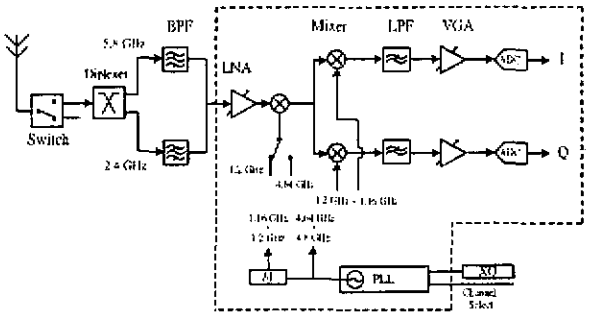


Fig. 7. Block diagram of the receiver.

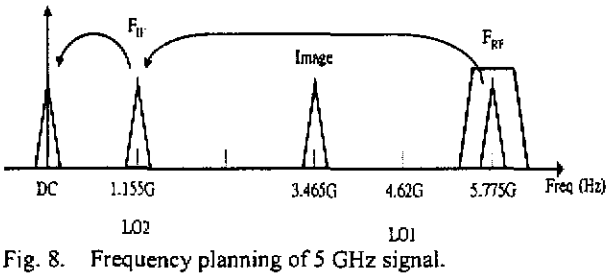


Fig. 8. Frequency planning of 5 GHz signal.

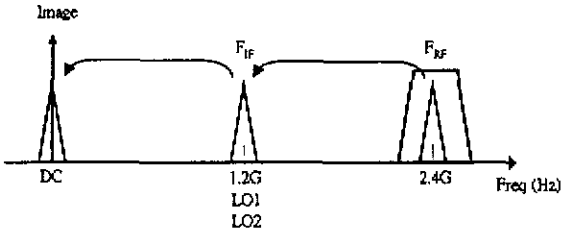


Fig. 9. Frequency planning of 2.4 GHz signal.

TABLE I
FREQUENCY RANGE OF EACH STAGE

Frequency Band	Frequency Range	VCO Range	IF Range
Upper U-NII Band (GHz)	5.725-5.825	4.58-4.66	1.145-1.165
ISM (GHz)	2.40-2.4835	4.8-4.967	1.2-1.242

An OFDM WLAN Baseband Receiver with Space Diversity

Chung-Jr Jan, Pin-Hsun Lin, and Tzi-Dar Chiueh
Graduate Institute of Electronics Engineering and
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan 10617

Abstract— **In the paper, we present the design and implementation of a baseband receiver with space diversity for IEEE 802.11a OFDM wireless local area network (WLAN). The single-antenna OFDM receiver has been implemented in a single chip. To improve the system performance, we adopt two antennae and apply maximal ratio combining (MRC) for attaining space diversity. We then implement an MRC-based IEEE 802.11a WLAN baseband receiver using two receiver ICs and an FPGA. Actual measurements demonstrate that under several different channel conditions the space diversity receiver outperforms the original single-antenna receiver by up to 8 dB in SNR.**

I. Introduction

Recently, WLAN has received more and more attention in both academia and industry. Basically WLAN systems are designed to provide high-speed indoor links between portable devices and fixed networks. One major effect that deteriorates the performance of wireless communication is multipath fading. Conventional solution to the multipath fadin involves time-domain adaptive equalizers, which consum huge amount of hardware. The orthogonal frequency-divisio multiplexing (OFDM) technique overcomes the fading effec with low-complexity frequency-domain equalizers (FEQ). I addition, OFDM utilizes the bandwidth more efficiently tha conventional single-carrier communication systems. Therefore OFDM has been used in the IEEE 802.11a WLAN standard

To further increase the transmission rate, multiple-antenn techniques can be combined with the OFDM technology. I this paper, we design and implement an OFDM baseban receiver that achieves space diversity by using two baseban receiver chips (one for each antenna) and a signal processin module built in field-programmable gate array (FPGA). Th contents of this paper are as follows. Section 2 briefly de scribes the operation of the baseband receiver IC, includin its timing synchronization, carrier frequency synchronizatior and equalization. In Section 3, we introduce the technique o space diversity, which includes transmit diversity and receive diversity. We discuss only the receiver diversity and show the comparison between two combining methods: selection combining (SC) and maximum ratio combining (MRC). In Section 4, we will give a description of the design of the complete OFDM baseband receiver with space diversity and the measurement results. Finally, a conclusion is given in section 5.

II. Baseband OFDM Receiver Chip Design

Fig. 1 depicts the block diagram of the OFDM baseband receiver IC that has been designed and implemented. The receiver IC, as well as the A/D converter, operates at 20MHz. In the beginning, the receiver finds the symbol boundary by the delay correlator and the correlator bank. At the same time, it also estimates the coarse carrier frequency offset (CFO) according to the phase in accumulated delay correlation output values. Once the coarse CFO is estimated it is immediately compensated in the time domain by an oscillator generating a sinusoidal wave with the same frequency as the coarse CFO but in the opposite direction. The remaining CFO, however, is tracked and compensated in the frequency domain. The baseband receiver IC also estimates the channel gain and correct the channel effect in the frequency domain using a frequency-domain equalizer. In the following we will briefly describe these operations.

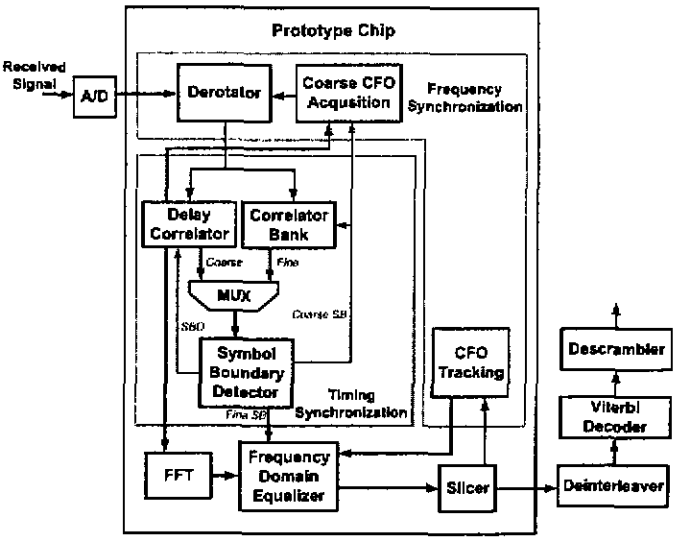


Fig. 1. The block diagram of the OFDM baseband chip.

A. Timing Synchronization

An 802.11a packet starts with ten identical short preamble symbols and then they are followed by two and a half long preamble symbols [1]. As soon as the automatic gain control

(AGC) block has settled, the symbol boundary detector in the baseband receiver starts to determine the correct FFT window. This task is divided into coarse symbol boundary detection and fine symbol boundary detection. Because of the repetitive characteristic of the short preamble segment, the receiver estimates the coarse symbol boundary by means of delay correlation with a delay matching to the length of one or several short preamble symbol. The coarse symbol boundary is detected at the falling edge of the delay correlation (loss of periodicity). In [2], the fine symbol boundary is detected by finding the first path of the channel impulse response (CIR) with an additional FFT operation. In this chip we estimate the CIR in the time domain by a bank of correlators matched to the ensuing long preamble symbol. With the first path of the CIR, the receiver sets the FFT window accordingly to minimize inter-symbol interference (ISI).

B. Frequency Synchronization

As carrier frequency offset (CFO) between the transmitter and the receiver always, the orthogonality between different subcarriers can be destroyed and inter-carrier interference (ICI) introduced. In addition, the phase of the received signal will be rotated by a quantity proportion to the CFO and the time index. Within one OFDM symbol, the phase rotation is 1.44 degree with 1-KHz CFO, and according to [1], the maximum CFO between the transmitter and the receiver can be as large as 232.2 KHz. Hence, CFO is a significant issue that needs be tackled.

In the proposed chip, frequency synchronization is accomplished in two steps: coarse CFO acquisition and CFO tracking. We exploit the fact that CFO is proportional to the phase of the delay correlation to estimate the coarse CFO. The residual frequency offset in the compensated signal can still incurs phase rotation in the frequency-domain data. In the ASIC, another PLL is used to track the phase error by compensating in the time domain.

C. FFT and FEQ

We select the radix-2^2 algorithm in order to reduce the number of the complex multiplications per output sample. The reason is that it has the same multiplicative complexity as the radix-4 algorithm and still retains the regular butterfly structure of radix-2 algorithm. Another operation related to frequency-domain signal processing is FEQ. In the indoor environments, we assume that the channel fades slowly. Therefore, we can apply the channel gain estimated from the long preamble signal to the following OFDM symbols in the same packet without adaptation. The least squares method is adopted in the estimation of channel gains from the long preamble symbols.

D. Channel Decoder and Design Summary

Since we implement only the inner receiver in this chip, the error-correcting code decoder is not implemented. In the following simulation, we collect the output bit streams of the slicer and decode these data using a software decoder. The Viterbi algorithm is chosen as the channel decoder because its complexity will not grow exponentially as the number of data bits increases.

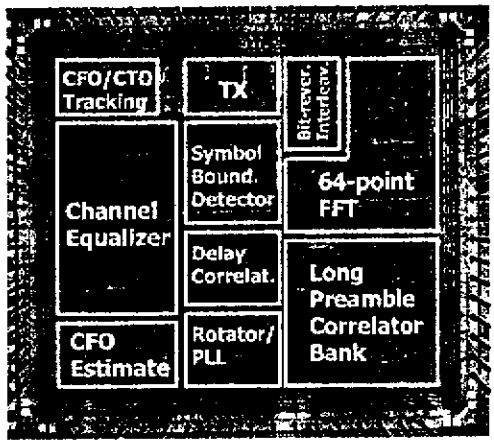


Fig. 2. Die photograph of the OFDM receiver chip ($4102 \times 4600 \mu m^2$).

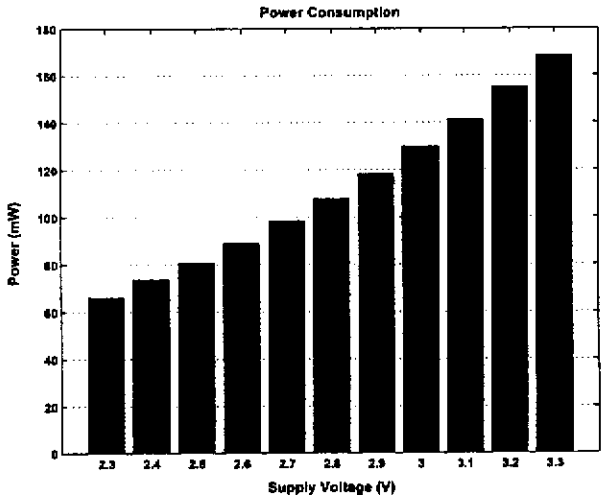


Fig. 3. Power consumption of the OFDM receiver chip.

The OFDM receiver ASIC is fabricated in a $0.35 \mu m$ CMOS process and is packaged in a 68-pin CLCC package. It is a cell-based design with several customized memory blocks (mostly delay buffers) to reduce not only the chip power consumption but also the active area. The transistor count is 429659, which is approximately 100K gates and the layout size of the chip including the pad ring is $4102 \mu m \times 4600 \mu m$. Fig. 2 shows the photograph of the chip and Fig. 3 shows the power consumption at different supply voltages.

III. Space-Diversity Receiver Design

Space diversity is used to increased the overall received SNR by utilizing multiple antennae. The antennae are arranged with sufficient spacing such that the channels are assumed independent. Because there are more independent paths between the transmitter and the receiver through proper diversity algorithms, the probability of occurrence of deep fades in all the diversity channels is significantly decreased. So the

performance can be highly improved as the number of diversity increase. Various space diversity techniques have been researched recently [3][4][5][6][7][8][9]. Generally speaking, there are two broad categories of space diversity:

- Transmit diversity: such as delay diversity, space-time trellis codes, space-time block codes, and so on.
- Receive diversity: such as selection combining (SC), maximum ratio combining (MRC), equal gain combining (EGC), and so on.

We focus only on receive diversity because it is much easier to implement with the existing receiver chip. In the following we discuss properties of some most popular combination technique for receive diversity such as SC, EGC, and MRC and make a comparison among these methods:

A. Selection Combining

SC is the simplest technique which just chooses the largest SNR in each symbol interval. However, the SNR is usually difficult to estimate, so most systems use the total power of signal and noise instead of SNR. In other words, we choose the signal with the highest received power. Since the other insignificant signals are completely ignored, we need only one receiver chains. It is an attractive characteristic because we the hardware complexity will not be increased by much. Although SC is quite easy to implement, it is not the optimal combining technique because it discards the information from all but one antenna.

B. Maximum Ratio Combining

In MRC, the received signal from different antennae are rotated so that they are co-phased, weighted proportionally by the magnitude of the channel gain, and finally summed. Conceptually speaking, if the magnitude of the channel gain is larger, the corresponding signal suffers less channel-fading effect and becomes more reliable. In one word, MRC can perform much better than SC because it collect the information from all the antennae and weight the signal by the optimal coefficients. However, the performance somewhat degrades if there exists channel estimation errors and frequency offset [6].

C. Equal Gain Combining

In EGC, the received signal from different antennae are rotated to be co-phased, and then combined. In other words, all the signal branches are weighted equally. EGC is usually used in case the magnitude of channel gain is not convenient to estimate. In general, EGC performs slightly worse than MRC because it lacks the information about the magnitude of channel gain, which contains the reliability in the subcarrier.

We compare the performance of different combining schemes through computer simulation. For simplicity, we only consider the case of 2-antenna diversity receiver. We use the same architecture of synchronization blocks and the FFT block. QPSK modulation with 1/2 rate convolutional code is used and the number of OFDM symbols is set to 50 per packet. We also take the average of the packet error rate over several static channels with the same RMS delay spread. Fig. 4 shows the simulation results of packet error rates using different diversity techniques under multipath fading channel. It is clear that the MRC receiver with 2-antenna diversity outperforms 1-antenna receiver by about 8 to 9 dB. The performance of EGC receiver is slightly worse than MRC receiver. In addition, the SC receiver with 2-antenna diversity only outperforms the single-antenna receiver by less than 2 dB. No matter what kind of diversity technique is used, the curve with two antennae is steeper than the one with one antenna. It is because the probability of signals at the same subcarrier from two antennae are both deep faded is much less. Since the magnitude of the channel gain can be estimated by the receiver chip, it is not difficult to realize the MRC technique. As a result, we determine to adopt MRC technique for the implementation of the baseband space diversity receiver.

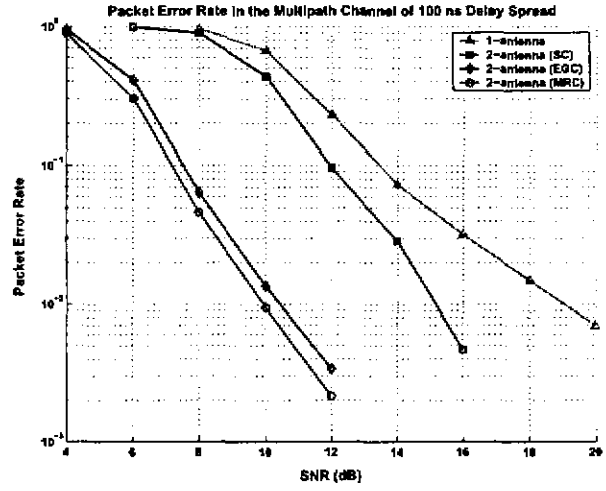


Fig. 4. The packet error rates with 2-antenna diversity of different combining schemes.

IV. Hardware Implementation and Measurement Results

For the sake of simplicity, we implement the space diversity receiver with only two antennae. Fig. 5 illustrates the architecture of the space diversity receiver. We reuse the block of the coarse symbol boundary detector, the fine symbol boundary detector, the coarse frequency estimation loop, and FFT in the chip, and implement the other circuits, including the CFO tracking loops, the frequency domain equalizer and combination circuit, and the slicer, using an FPGA.

Fig. 6 shows the photograph of the space diversity receiver hardware. We have also implemented the diversity receiver in FPGA simulator environment to verify the functionality of the design. From the simulation results, we can see that although the signal from one of the antennae is in deep fade, the diversity receiver can still get the correct data by combining it with signal from the other antenna. Fig. 7 shows the packet error rate of the space diversity receiver in different Rayleigh fading channels. The practical measurement result

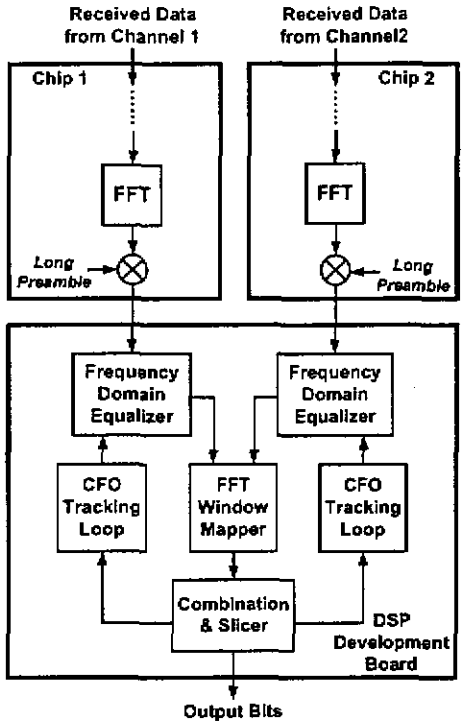


Fig. 5. Block diagram of the OFDM baseband receiver with space diversity.

degrades about 1 dB compared from the simulation. Although the hardware requirement of the space diversity receiver is approximately twice as large as that of single-antenna receiver, it outperforms the single-antenna receiver by about 8 dB in SNR.

Fig. 6. Photograph of the OFDM baseband receiver with space diversity.

V. Conclusion

In the paper, we present the design and implementation of a baseband receiver with space diversity for IEEE 802.11a OFDM WLAN. This system includes two single-antenna OFDM receiver chips and an FPGA board that executes MRC as an approach to attain space diversity. Actual measurements demonstrate that under several different channel conditions

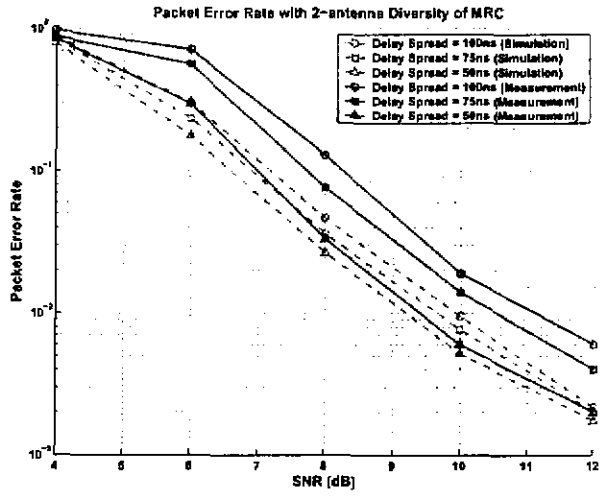


Fig. 7. Packet error rates of the 2-antenna diversity receiver using MRC under channels with different delay spread.

the space-diversity receiver outperforms the original single-antenna receiver by up to 8 dB in SNR.

References

[1] IEEE, *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-speed Physical Layer in the 5 GHz Band, Std 802.11a-1999.*

[2] B. Yang, K. B. Letaief, R. S. Cheng, and Z. Cao, "Burst frame synchronization for OFDM transmission in multipath fading links," in *Proc. of IEEE Vehicular Technology Conference*, 1999, vol. 1, pp. 300V304.

[3] D. Lee, G. J. Saulnier, Z. Ye, and M. J. Medley, "Antenna diversity for an OFDM system in a fading channel," in *Proc. of IEEE Military Communications Conference*, Nov. 1999, vol. 2, pp. 1104V1109.

[4] T. Eng, N. KOng, and L. B. Milstein, "Comparison of diversity combining techniques for rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 44, no. 9, pp. 1117V1129, Sept. 1996.

[5] X. Ouyang, M. Ghosh, and J. P. Meehan, "Optimal antenna diversity combining for IEEE 802.11a system," *IEEE Trans. Consumer Electron.*, vol. 48, no. 3, pp. 738V742, Aug. 2002.

[6] D. Bartolome and A. I. Perez-Neira, "MMSE techniques for space diversity receivers in OFDM-based wireless LANs," *IEEE J. Select. Areas Commun.*, vol. 21, no. 2, pp. 151V160, Feb. 2003.

[7] R. Narasimhan, "Performance of diversity schemes for OFDM systems with frequency offset, phase noise and channel estimation errors," *IEEE Trans. Commun.*, vol. 50, no. 10, pp. 1561V1565, Oct. 2002.

[8] D. Gesbert, M. Shafi, D. S. Shiu, P. J. Smith, and A. Naguib, "From theory to practice: An overview of mimo space-time coded wireless systems," *IEEE J. Select. Areas Commun.*, vol. 21, no. 3, pp. 281V301, Apr. 2003.

[9] A. N. Barreto, "Antenna transmit diversity for wireless OFDM systems," in *Proc. of IEEE Vehicular Technology Conference*, May 2002, vol. 2, pp. 757V761.

DESIGN AND SIMULATION OF A BASEBAND TRANSCEIVER FOR IEEE 802.16a OFDM-MODE SUBSCRIBER STATIONS

Sang-Jung Yang, Yi-Ching Lei, and Tzi-Dar Chiueh

Graduate Institute of Electronics Engineering
and Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan 10617

ABSTRACT

In this paper, we propose a subscriber station (SS) physical layer (PHY) inner transceiver architecture for IEEE 802.16a OFDM mode. Algorithms for symbol boundary detection, carrier frequency offset (CFO), timing offset (TO) and frequency-domain equalization (FEQ) as well as channel interpolation are embedded in this receiver. Since IEEE 802.16a standard contains various mandatory requirements, such as channelization type, baseband equivalent channel model, carrier frequency band, data throughput, etc., the main consideration of the proposed architecture is then flexibility and robustness to conform to various kinds of channel combinations. Simulation results show that the proposed architecture can fulfill the bit-error-rate (BER) requirements under most situations.

1. INTRODUCTION

The OFDM modulation technique offers an attractive solution to high-rate data access for its robustness against frequency-selective multipath fading and its simple equalization scheme. Moreover, it is also very efficient in spectrum utilization since the spectra of the adjacent subcarriers overlap. OFDM has therefore been adopted in several standards such as DVB-T, VDSL, IEEE 802.11a and IEEE 802.16a [1].

IEEE 802.16a is a broadband wireless metropolitan area network (WirelessMAN™) standard from the Institute of Electrical and Electronics Engineers (IEEE) provides for fixed broadband wireless access (BWA) between 2 and 11 GHz, and is an extension of the global IEEE 802.16™ WirelessMAN standard for 10 to 66 GHz. In the base IEEE 802.16 standard, the advanced technology it defines is designed from first principles to support multimedia services such as videoconferencing, voice and gaming. New features, including an optional mesh architecture, are also included. As for IEEE 802.16a, it facilitates the widespread deployment of 2 to 11 GHz wireless MANs as an economical alternative to wireline connections to the internet.

The 802.16a standard supports licensed and license-exempt spectra between 2 and 11 GHz. These frequencies are well suited for residential and small business applications using non-line-of-sight links. Major telephone/data carriers and wireless internet services providers will be the major customers for equipments developed under this standard. Since the technology integrates well with the IEEE 802.11 wireless LANs, IEEE 802.16a base stations are excellent candidates for wireless links between 802.11 hot spots and the Internet. Moreover, the standard can also play a pivotal role in underdeveloped regions where only sporadic wired infrastructures are available.

Table 1. Parameters for 802.16a OFDM mode.

RF frequency	2-11GHz
FFT Size	256
Effective subcarriers	192
Bandwidth (MHz)	BW
Guard time (us)	Tg
Data time (us)	Tb
Symbol time (us)	Tg + Tb
Subcarrier spacing (kHz)	Δf
Sampling rate (MHz)	Fs = BW x 8/7
Maximum data rate (Mbps) (BW=28MHz, 64QAM, code rate 3/4)	104.73

Table 2. ETSI Channelization schemes.

BW (MHz)	OFDM(N_{FFT}=256)					
	Δf(kHz)	Tb(us)	Tg(us)			
			Tb/32	Tb/16	Tb/8	Tb/4
1.75	$7\frac{13}{16}$	128	4	8	16	32
3.5	$15\frac{5}{8}$	64	2	4	8	16
7.0	$31\frac{1}{4}$	32	1	2	4	8
14.0	$62\frac{1}{2}$	16	$\frac{1}{2}$	1	2	4
28.0	125	8	$\frac{1}{4}$	$\frac{1}{2}$	1	2

There are three non-interoperable modes defined in the IEEE 802.16a standard: WirelessMAN-SCa (single carrier), WirelessMAN-OFDM, and WirelessMAN-OFDMA (multiple access). We choose to implement the OFDM mode receiver to exploit the advantages of OFDM modulation.

To provide further insight into what IEEE 802.16a OFDM mode is, we list several parameters defined in the standard [1] in Table 1. Note that the data time Tb, the guard time Tg and the subcarrier spacing Δf all vary with the bandwidth. Table 2 lists the channelization schemes for IEEE 802.16a licensed spectra as defined [1]. Based on the parameters in the above two tables, we design and implement a baseband receiver for the IEEE 802.16a

OFDM-mode subscriber station. Moreover, simulation results of the implemented software receiver are shown and accordingly the validity of the design is assured.

This paper is organized as follows. In section 2 we will introduce the symbol description and packet format defined in the standard. Next, a receiver architecture is presented in Section 3. Then, Section 4 provides simulation results of this system under the Stanford University Interim (SUI) [2] channel model. Finally, Section 5 concludes this paper.

2. SYMBOL DESCRIPTION AND PACKET FORMAT

The OFDM symbol structure is illustrated in Fig. 1 and Fig. 2. Fig. 1 illustrates the OFDM symbol time-domain structure and Fig. 2 shows the frequency-domain description. Note that 8 pilot subcarriers are evenly inserted in every OFDM symbol. These pilots are power boosted by 3dB in magnitude and shall be transmitted using binary phase shift keying (BPSK). In the proposed receiver, we utilize these pilots for tracking residual CFO/TO. Details of the tracking loop will be described in a later section.

The IEEE 802.16a OFDM packet contains two parts, preamble symbols and data symbols. Each of The data symbols in a packet contains a 256-sample OFDM symbol with a variable-length cyclic prefix. The preamble structure is shown in Fig. 3. Note that the short preamble symbol uses only subcarriers whose indices are multiples of four. As a result, the time-domain short preamble waveform consists of four periods of a 64-sample fragment, which are preceded by a cyclic prefix. The long preamble utilizes only even subcarriers, resulting in a time-domain waveform composed of two repetitions of a 128-sample fragment, which are again preceded by a cyclic prefix.

The short preamble is used for coarse symbol boundary detection and fractional CFO estimation, whereas the long preamble is for fine symbol boundary detection, integer CFO estimation and channel estimation.

3. RECEIVER FUNCTION AND ARCHITECTURE

The overall block diagram of the proposed SS physical layer baseband receiver is depicted in Fig. 4. For clarity, the receiver is roughly partitioned to three parts. *Part I* involves mostly time-domain processing, such as coarse and fine symbol boundary detection and compensating the CFO by phase rotation. *Part II* is intended for transformation of the signals from time domain to frequency domain and the related signal processing across two domains. These include an FFT block, tracking loop for the residual CFO, phase rotation for timing offset (TO) compensation and FFT window control. At last, the blocks in *Part III* are used for frequency-domain signal processing such as channel estimation, equalization, and slicing.

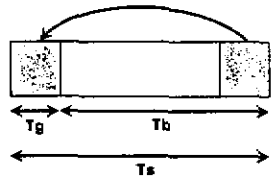


Fig. 1. OFDM symbol time domain structure.

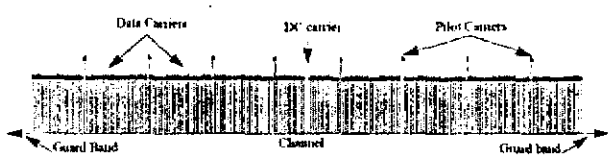


Fig. 2. OFDM Symbol frequency domain description.

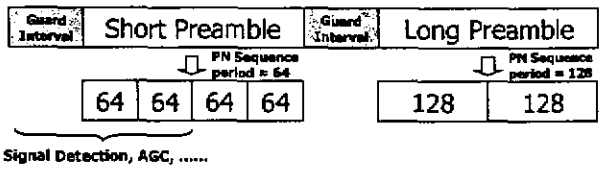


Fig. 3. Preamble structure of 802.16a OFDM mode.

3.1. Synchronization for Acquisition

In the proposed receiver, several tasks are completed as soon as the receiver starts up. These include finding the boundary of FFT window and the CFO, which is crucial for OFDM receivers. Since the short preamble is the very first symbol of each packet, we assume some part of the short preamble is used for tasks such as signal detection, automatic gain control, and others. Therefore we assume in the worst case only two 64-sample segments of the short preamble are available for the synchronization task.

The procedures for the symbol boundary detection and the CFO estimation are summarized as follows :

(i) Compute the normalized delay correlation M(d) and its moving average with length equals to guard interval [3] .

$$P(d) = \sum_{m=0}^{L-1} (r_{d+m}^{*} r_{d+m+L}) \tag{1}$$

$$R(d) = \sum_{m=0}^{L-1} \left| r_{d+m+L} \right|^2 \tag{2}$$

$$M(d) = \frac{|P(d)|^2}{((R(d))^2} \tag{3}$$

where d is the time index, r_m denotes the received m-th signal sample, L is the period of short preamble and is set to 64 in our case. $P(d)$ is the 64-point delay correlation. $R(d)$ is the received energy of 64 consecutive samples, and $M(d)$ is the normalized 64-point delay correlation.

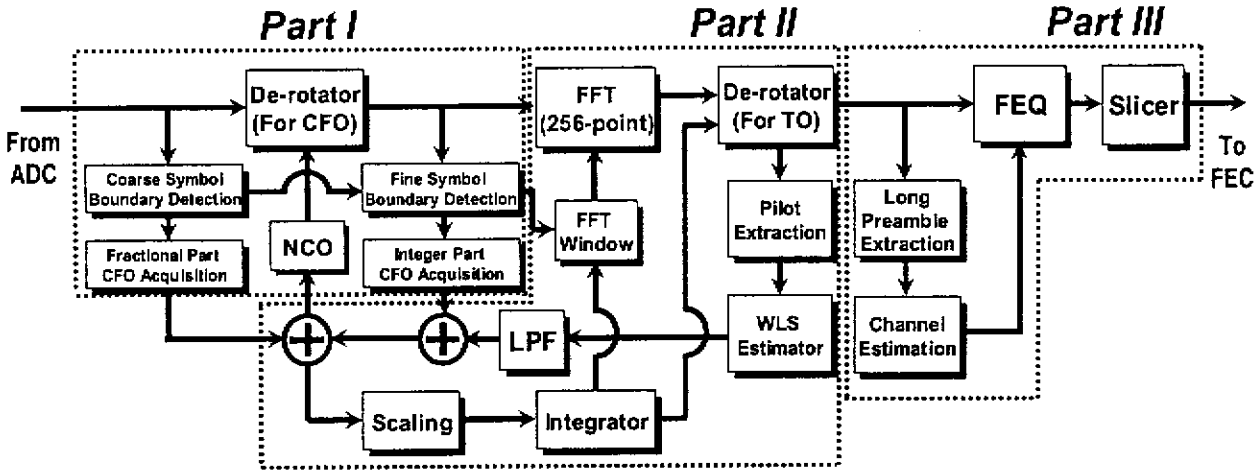


Fig. 4. Block diagram of the proposed physical layer inner receiver.

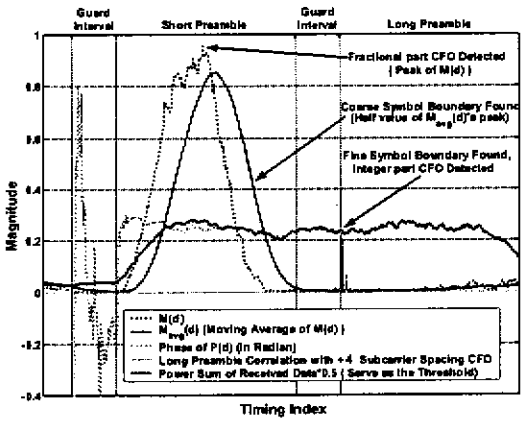


Fig. 5. Simulation of symbol boundary detection and CFO estimation. (SUI-4, SNR=5dB, CFO=+4 subcarrier spacing)

Note that the moving average of $M(d)$ with length equals to the guard interval N_g is defined by

$$M_{avg}(d) = \frac{1}{N_g}\sum_{m=0}^{N_g} M(d-m) \qquad (4)$$

Moving average is used for smoothing $M(d)$ and thus can yield more accurate result of symbol boundary detection.

(ii) Find the peak of the moving average, and compute the fractional CFO from the phase of its corresponding delay correlation.
(iii) When the current moving average drops to half of its peak value, we set the coarse symbol boundary at 128 samples to the right of this point.
(iv) Find the fine symbol boundary by searching in a interval that spans 16 samples in both directions from the previous coarse symbol boundary.
(v) Use the long preamble correlator bank at the searching window. The set with peak occurs indicates the correct integer part CFO, and the peak position is then the fine symbol boundary.
(vi) To handle the situation that the first path is not the strongest path, we use a threshold to find the peak. The threshold is set to be half of R(d), which is half of the power sum of 64 consecutive received samples.

3.2. CFO/TO Tracking Loop

The task of CFO/TO tracking loop is to estimate the residual CFO/TO and track it by a carefully-designed loop. In this paper, we adopt the joint weighted least squares algorithm proposed in [4], which is a pilot-based estimation scheme and is suitable for 802.16a system.

In the tracking loop, pilots are first extracted from each OFDM symbol and are passed to the WLS estimator. Since we assume there is only one oscillator in the receiver, we can simply estimate the CFO and then scale it to get TO. Moreover, since the CFO introduces much more ICI then the TO, we compensate the CFO in the time domain while compensating the TO in the frequency domain.

Since ICI and noise tend to deteriorate the estimation accuracy and induce jitters, we include a first-order low-pass proportion-integral (PI) filter in the loop. By adjusting its two coefficients C1 and C2, we can make a trade-off between convergence speed and jitter level, as depicted in Fig. 6. To reduce the hardware complexity, both C1 and C2 are chosen to be powers of two. With (C1,C2) = (0.25, 0.125), the loop filter output can have fairly well performance.

In the case when the timing error overflows or underflows, we can simply adjust the FFT window for an OFDM symbol right after detecting an overflow or an underflow. For example, if the overflow happens in i-th OFDM symbol, i.e. the timing clock in the receiver is too fast that one sample needs to be dropped, then the receiver move the FFT window one sample ahead during the Fourier transformation of the (i+1)-th OFDM symbol and vice versa. With this mechanism, the receiver works smoothly without any up-sampling.

3.3. Channel Estimation and FEQ

As mentioned before, the long preamble is utilized for channel estimation. However, since the long preamble utilizes only even subcarriers, interpolation is needed to estimate the channel responses of the odd subcarriers. We adopt a raised-cosine filter with "shift

term"[5] to achieve this task. Fig. 7 illustrates the improvement by adding the "shift term" in the interpolator. Note that the peaks near the subcarrier 0 in Fig. 7(b) and 7(d) are due to no DC subcarrier. From these figures, we can see that with shift term added, the estimation error is greatly reduced.

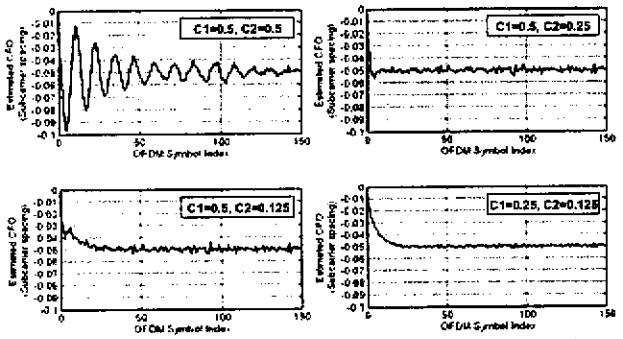


Fig. 6. Simulation results of different loop filter coefficients. (SUI-3, SNR=15dB, residual CFO = -0.05 Δf, BW=28MHz)

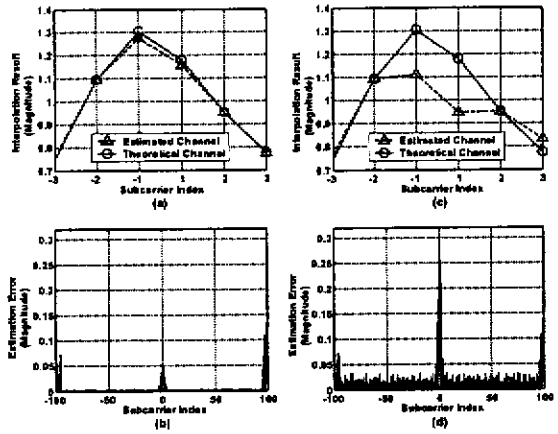


Fig. 7. Estimated channel response and estimation error with/without shift term. (SUI-6, BW=1.75MHz). (a) estimated channel response with shift term, (b) estimation error with shift term, (c) estimated channel response without shift term, and (d) estimation error without shift term.

4. SIMULATION RESULTS

Before illustrating the simulation results, the channel model used in the simulation is introduced. As mentioned before, SUI channel model 1~6 are adopted. In brief, these six models differ from one another in delay spread, multipath fading, and Doppler frequency. Since 802.16a is for fixed broadband application, Doppler frequency is quite small. In fact, the maximum Doppler frequency is only 2.5Hz.

In addition to multipath fading, bandwidth is another important factor in our simulation. Since different bandwidth yields different sampling frequency and hence different amount of timing error, we combine each SUI model with the widest bandwidth for the highest possible data rate. The criterion of choosing the bandwidth for each SUI model is the system being free of inter-symbol

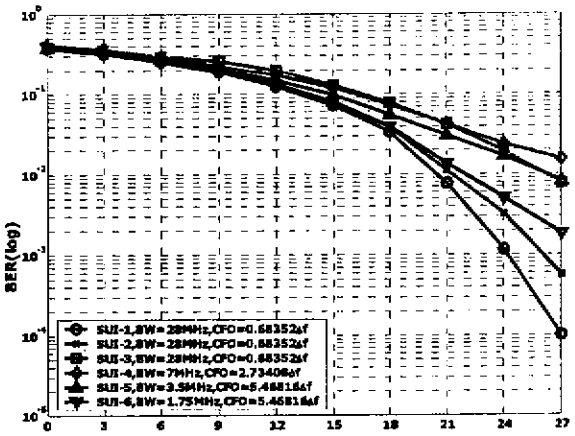


Fig. 8. Simulation results of uncoded BER v.s. SNR using 64-QAM modulation.

interference. Moreover, we set both CFO and TO to the maximum level specified in the standard. The simulated uncoded bit error rate (BER) performance of the proposed receiver using the highest 64-QAM constellation is illustrated in Fig. 8. If we further implement forward error correcting (FEC) codes, the case with uncoded BER of less than 10^{-2} is equivalent to a coded BER of about 10^{-5}~10^{-6} and hence satisfies the BER requirement listed in the standard.

5. CONCLUSION

In this paper, a physical layer inner receiver architecture for IEEE 802.16a OFDM mode subscriber station is proposed. It is designed to meet various kinds of combinations of SUI channel models and bandwidths. Algorithms of symbol boundary detection, CFO acquisition and CFO/TO tracking loop as well as channel estimation are designed. Uncoded BER performances under SUI-1~6 and different bandwidths show that the proposed receiver is capable of reliable transmission in most scenarios.

6. REFERENCES

[1] "Air Interface for Fixed Broadband Wireless Access Systems - Medium Access Control Modifications and Additional Physical Layer Specifications for 2-11GHz," *IEEE Std 802.16a-2003*, Jan. 2003.

[2] V. Erceg, K.V.S. Hari, M.S. Smith, D.S. Baum, K.P. Sheikh, C. Tappenden, J.M. Costa, C. Bushue, A. Sarajedini, R. Schwartz, D. Branlund, S. Kaitz, D. Trinkwon, "Channel Models for Fixed Wireless Applications," *802.16a-03/01*. Jun. 2003.

[3] T. M. Schmidl, D. C. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Trans. Commun.* vol. 45, pp. 1613-1621, Dec. 1997.

[4] Pei-Yun Tsai, Hsin-Yu Kang, Tzi-Dar Chiueh, "Joint weighted least squares estimation of frequency and timing offset for OFDM systems over fading channels," *IEEE Vehicular Technology Conference*, 2003, Volume 4, pp. 2543-2547 ,April 22-25, 2003

[5] Pei-Yun Tsai, Tzi-Dar Chiueh, "Frequency-domain interpolation-based channel estimation in pilot-aided OFDM systems," *IEEE VTC*, Spring, May 2004.

DESIGN AND SIMULATION OF A MIMO OFDM BASEBAND TRANSCEIVER FOR HIGH THROUGHPUT WIRELESS LAN

Chi-Yeh Yu, Zih-Yin Ding, and Tzi-Dar Chiueh

Graduate Institute of Electronics Engineering
and Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan 10617

ABSTRACT

Multiple-input multiple-output (MIMO) signal processing techniques combined with orthogonal frequency division multiplexing (OFDM) can provide a high-throughput link as well as an increased diversity gain in frequency-selective fading channels. This paper proposes a MIMO OFDM baseband inner transceiver design for the next-generation high-throughput wireless LAN using two transmit antennas and two receive antennas. A MIMO OFDM receiver with algorithms for timing and frequency synchronization, tracking, channel estimation, and MIMO detection is designed and implemented in software. Simulation results show that the proposed receiver is capable of transmission with a data rate that is twice that of the current IEEE 802.11a wireless LAN standard.

1. INTRODUCTION

One of the main objectives of current wireless communication research is to increase the link throughput in a spectrally-efficient and reliable manner. Multiple-input multiple-output (MIMO) techniques have the potential of increasing the transmission rate by several folds [1]. In general, there are two types of MIMO techniques: (a) rate maximization schemes such as vertical Bell Laboratories Layered Spare-Time code (V-BLAST) [2] and (b) diversity maximization schemes such as space-time block code (STBC) [3]. The former can achieves a higher data rate, while the latter gains more diversity. Both types of MIMO techniques were developed originally for flat fading channels, and lately have been applied to orthogonal frequency division multiplexing (OFDM) technology used in frequency-selective fading channels.

In this paper, we propose a baseband inner transceiver capable of doubling the data rate of IEEE 802.11a standard with two transmit and two receive antennas. The uncoded data rate ranges from 12 × 2 Mbps to 72 × 2 Mbps. A new packet format based on the original format is designed for accommodating MIMO channel estimation. Moreover, several synchronization algorithms for MIMO OFDM receiving are implemented. Functional simulation results of the proposed receiver establish the validity of it effectiveness in high-throughput wireless LAN applications.

The organization of the paper is as follows. Section 2 describes the system as well as the packet format. Section 3 gives the receiver architecture and explains respective algorithms. In Section 4 the simulation results are provided. Finally, conclusions are drawn in Section 5.

2. SYSTEM DESCRIPTION AND PACKET FORMAT

The proposed MIMO OFDM system is illustrated in Fig. 1. At the transmitter, data symbols are first processed by two MIMO techniques namely spreading and MIMO coding. Data spreading is used to enhance the diversity, and is done by a Walsh-code spreader [5]. MIMO coding will arrange symbols in 2 × 2 VBLAST or in 2 × 2 STBC according to the adopted data rate. The processed data symbols are then sent to an IFFT operation and a cyclic prefix (CP) is inserted. At the receiver, CP is removed, and the data symbols are recovered after an FFT operation. Finally, data symbols are decoded, despread, and then detected.

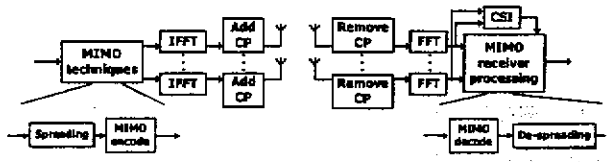


Fig. 1. MIMO OFDM system model.

Physical layer key parameters used in the proposed MIMO OFDM transceiver are listed in Table 1. Note that all are compatible with the IEEE 802.11a standard. Additional double data rates newly provided by the MIMO techniques are listed in Table 2. Note that VBLAST-OFDM can achieve a higher throughput due to parallel transmission, while STBC-OFDM gains more diversity by coding over different branches.

The new MIMO OFDM packet format is shown in Fig. 2. The preamble consists of 10 short symbols, denoted as t_1 to t_{10}, and four long symbols, denoted as T_1 to T_4. The preamble is followed by the SIGNAL field and DATA symbols. Identical to the IEEE 802.11a standard,

Table 1. Physical layer key parameters.

Modulation type	BPSK, QPSK, 16QAM, 64QAM
FFT size	64
Subcarrier in use	52 (4 for pilot, 48 for data)
Bandwidth	20 MHz (Occupied 16.6 MHz)
Subcarrier spacing	312.5 KHz (=20 MHz/64)
FFT period	3.2 us (=1/312.5 KHz)
Guard interval duration	0.8 us
OFDM symbol duration	4.0us (=3.2+0.8us)
FFT period	3.2 us (=1/312.5 KHz)
Guard interval duration	0.8 us
OFDM symbol duration	4.0us (=3.2+0.8us)

Table 2. Achievable data rates.

Uncoded data rate (Mbps)	Modulation	MIMO coding	Spectral Efficiency (bps/Hz)
12×2	BPSK	2 × 2 STBC	2
24×2	QPSK	2 × 2 VBLAST	4
48×2	16QAM	2 × 2 VBLAST	8
72×2	64QAM	2 × 2 VBLAST	12

the short preamble utilizes only 12 out of 52 subcarriers, whose indices are a multiple of 4, and hence results in a periodicity of $T_{FFT}/4 = 0.8$ us. The long preamble also reuses the IEEE 802.11a design, with 52 subcarriers (excluding a zero value at DC) modulated by a fixed ± 1 sequence. The period of long preamble is $T_{FFT} = 3.2$ us. Two periods are preceded by a guard interval (GI) of 1.6 us. For MIMO channel estimation, there are four periods and the last two periods and the associated GI in one of the transmitter are inverted.

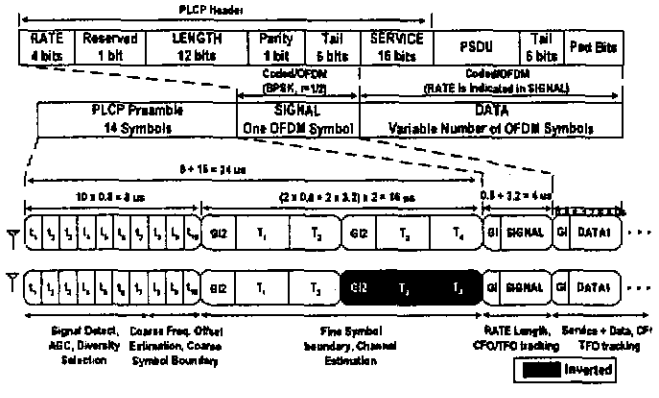


Fig. 2. Proposed MIMO OFDM packet format.

3. RECEIVER FUNCTION AND ARCHITECTURE

The overall receiver functional block diagram is shown in Fig. 3. We divide the receiver into three parts: Part I is responsible for timing and frequency synchronization in the acquisition mode; Part II is intended for residual carrier frequency offset (CFO) tracking and residual timing offset (TO) tracking; and Part III deals with channel estimation and data recovery.

3.1. Synchronization in the acquisition mode

Timing and frequency synchronization are performed in the acquisition mode. These tasks includes signal detection, fractional CFO estimation, coarse symbol boundary detection [4], and fine symbol boundary detection. To allow time for automatic gain control (AGC) and energy detection, we assume only the last five short preamble periods are available for synchronization.

Step 1) Signal detection and coarse symbol boundary detection — Due to the periodicity of the short preamble, we use the normalized delay correlation $M(k)$ to detect the existence of short preambles, and averaging the results from the two receive branches,

$$P(k) = \sum_{q=1}^{2} \sum_{n=k-L+1}^{k} r_q(n) r_q^*(n-d), \qquad (1)$$

$$R(k) = \sum_{q=1}^{2} \sum_{n=k-L+1}^{k} |r_q(n)|^2, \qquad (2)$$

$$M(k) = \frac{|P(k)|^2}{(R(k))^2}, \qquad (3)$$

where k is the time index, q is the receive antenna index, L is the summation length of delay correlation, d is the delay, and $r_q(n)$ denotes the received n-th sample from the q-th antenna. $P(k)$, $R(k)$, and $M(k)$ represent the delay correlation, signal energy, and normalized delay correlation, respectively.

After applying moving average to smooth the delay correlation, the coarse symbol boundary is determined at the falling edge of the moving average $P_{ave}(k)$,

$$P_{ave}(k) = \frac{1}{w} \sum_{n=k-w+1}^{k} P(n), \qquad (4)$$

where w is window size for moving average.

Step 2) Fractional CFO estimation — Carrier frequency offset between the transmitter and the receiver will reflect on the phase of the received time-domain signal. We can average the delay correlation and calculate the corresponding phase shift to estimate the CFO.

$$\Delta \hat{f} = \frac{\hat{\phi}}{2\pi d T_s}, \qquad (5)$$

where ϕ is the phase of $P(k)$ and T_s is the sampling interval.

Step 3) Fine symbol boundary detection — The long preamble has a PN-code like time-domain characteristic, so we match the received signal to several time-shifted

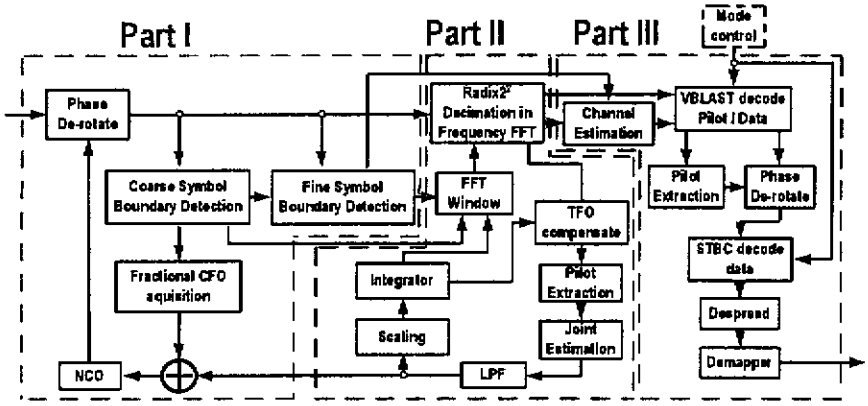


Fig. 3. Block diagram of the proposed MIMO OFDM receiver.

versions of the long preamble waveform by a bank of correlators. The outputs of the correlator bank thus approximate the channel impulse response, from which the first-arrival path can be determined.

The synchronization process simulation is illustrated in Fig. 4

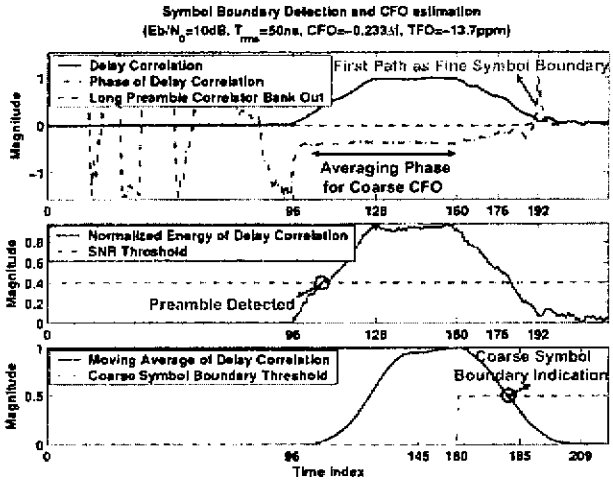


Fig. 4. Synchronization process when receiving the preambles of a packet.

3.2. CFO/TO tracking

Though residual CFO/TO are small after fractional CFO compensation, the accumulated phase rotation can still deteriorate system performance as time goes by. Consequently, we need a CFO/TO tracking loop to dynamically compensate such signal phase drift.

To estimate the residual CFO/TO, we adopt the joint weighted least squares (WLS) algorithm [6]. Because we have double the number of pilots from two branches, the accuracy can be enhanced when compared to the case with single receiving antenna. Assume there is only one oscillator in the receiver, WLS algorithm can be simplified, and we can estimate the CFO first and scaling the CFO by a constant to get the TO. For ICI level and simplicity, we choose compensate the CFO in the time domain and compensate the TO in the frequency domain. To limit the noise, we introduce a first-order low-pass proportional-integral (PI) filter in the tracking loop. Finally, as the sampling timing overflows or underflows, we shift the FFT window forward or backward by one sample so as to avoid inter-symbol interference.

3.3. Channel estimation and data recovery

The long preamble is especially designed for MIMO channel estimation. By inverting the last two periods of the long preamble in one of the transmitter, we can adopt the STBC combining method to estimate the channel as in [4]. The two extra periods of long preamble can also improve the channel estimation accuracy.

For a certain subcarrier i, the received symbol matrix $\mathbf{R}$ can be expressed in terms of the transmit symbol matrix $\mathbf{S}$, the channel matrix $\mathbf{H}$, and the noise matrix $\mathbf{N}$ as

$$\mathbf{R}_{i,2\times 2} = \mathbf{S}_{i,2\times 2}\mathbf{H}_{i,2\times 2} + \mathbf{N}_{i,2\times 2}. \quad (6)$$

Channels can then be estimated by

$$\mathbf{H}_{i,2\times 2} = \mathbf{S}_{i,2\times 2}^{-1}\mathbf{R}_{i,2\times 2}. \quad (7)$$

To further enhance the channel estimation performance, we add a frequency-domain filter to smooth the estimation result. Such filter is designed in such a way that its time-domain waveform is a window of proper shape and length to extract the channel impulse response and reject as much noise as possible [7]. Mean square errors of the estimated channel responses using the filtered and the unfiltered schemes are shown in Fig. 5. The improvements of the filtered channel estimation method is quite apparent. Using the estimated channel responses, we can detect the VBLAST OFDM signal and STBC OFDM signal as described in [8] and [9], respectively.

4. SIMULATION RESULTS

The whole MIMO OFDM receiver is implemented in software and the simulated uncoded bit error rates (BER) of

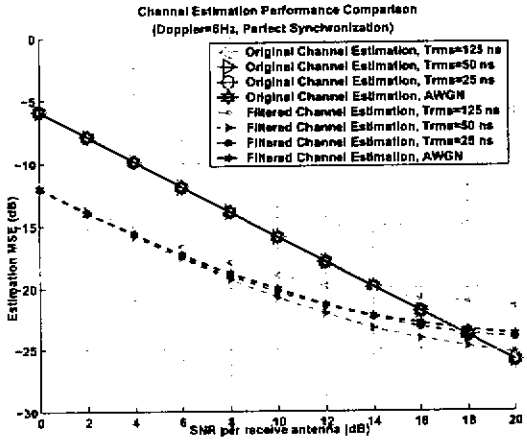


Fig. 5. Channel estimation MSE performance.

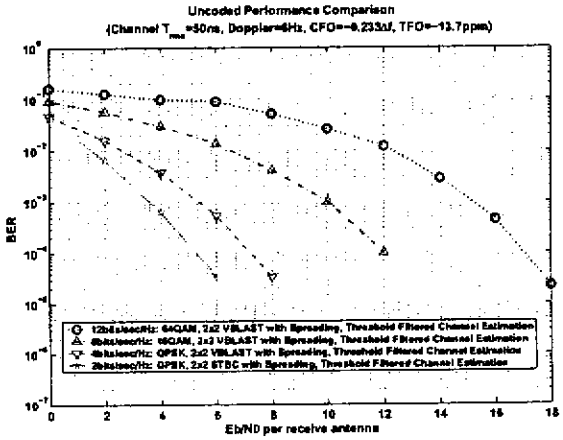


Fig. 6. Uncoded BER performance of the proposed MIMO OFDM receiver.

the demodulated signals are depicted in Fig. 6. In this simulation, independent scalar channels in typical office environments are used for simplicity. The center frequency is set to be 5320 MHz, which is the highest in the middle UNII bands. The CFO and TO are both at -13.7 ppm according to TGn comparison criteria [10], and the Doppler frequency is 6 Hz, or about 1.2 km/hr.

From the simulation results, the uncoded BER performance in the highest data rate case reaches 10^{-2} at $E_b/N_0 = 12$ dB, or equivalently at SNR=20 dB. With the help of outer coding, we can reasonably achieve a coded BER of 10^{-5}, or a packet error rate (PER) of 8% for 1000-byte packets, and thus satisfying the IEEE 802.11 standard requirement on PER.

5. CONCLUSION

In this paper, we propose a MIMO OFDM baseband inner transceiver design for the next-generation high-throughput wireless LAN using two transmitting antennas and two receiving antennas. Algorithms for timing and frequency synchronization, tracking, channel estimation, and MIMO detection are designed and implemented. Finally, the BER performance of the receiver is presented and the results indicate that the proposed receiver is capable of delivering reliable transmission with double the data rate achieved by the current IEEE 802.11a standard.

6. REFERENCES

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment using multiple antenna," *Wireless Personal Commun.*, vol. 6, pp. 311–335, Mar. 1998.

[2] P.W. Wolniansky, G.J. Foschini, G.D. Golden, and R.A. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *URSI International Symposium on Signals, Systems, and Electronics*, Oct. 1998, pp. 295–300.

[3] Siavash M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1451–1458, Aug. 1998.

[4] A. N. Mody and G. L. Stuber, "Receiver implementation for a MIMO OFDM system," in *Globecom*, 2002, pp. 716–720.

[5] Timothy A. Thomas, Kevin L. Baum, and Frederick W. Vook, "Modulation and coding rate selection to improve successive cancellation reception in ofdm and spread OFDM MIMO systems," in *ICC*, May 2003, pp. 2842–2846.

[6] Pei-Yun Tsai, Hsin-Yu Kang, and Tzi-Dar Chiueh, "Joint weighted least squares estimation of frequency and timing offset for OFDM systems over fading channels," in *Vehicular Technology Conference*, April 2003, vol. 4, pp. 2543–2547.

[7] Pei-Yun Tsai and Tzi-Dar Chiueh, "Frequency-domain interpolation-based channel estimation in pilot-aided OFDM systems," *IEEE VTC Spring*, May 2004.

[8] R. J. Piechocki, P. N. Fletcher, A. R. Nix, C. N. Canagarajah, and J. P. McGeehan, "Performance evaluation of BLAST-OFDM enhanced hiperlan/2 using simulated and measured channel data," in *Electron. Lett.*, Aug. 2001, vol. 37, pp. 1137–1139.

[9] K. Lee and D. B. Williams, "A space-time coded transmitter diversity technique for frequency selective fading channels," *IEEE Sensor Array and Multichannel Signal Processing Workshop*, pp. 149–152, Mar. 2000.

[10] IEEE 802.11 TGn, "Comparison criteria," DCN: 11-03-0814-r16.

A SCALABLE REED-SOLOMON DECODING PROCESSOR BASED ON UNIFIED FINITE-FIELD PROCESSING ELEMENT DESIGN

Jih-Chiang Yeo, Huai-Yi Hsu, An-Yeu (Andy) Wu

Graduate Institute of Electronic Engineering
and Department of Electrical Engineering,
National Taiwan University, Taipei 106, Taiwan

ABSTRACT

Reed-Solomon (*RS*) codes play an important role in providing error protection and data integrity. Many researches of RS decoder are implemented in parallel architecture, which can perform the highest data throughput rate in all RS decoder architectures. However, for some specifications such as *Digital Video Broadcasting system* (*DVB*), the requirement of data throughput rate is very low and a low-power folding architecture is enough to achieve the data throughput rate requirement. The hardware cost of folding architecture is also comparably much smaller than the hardware cost of parallel architecture. In this paper we present a design of a scalable RS codec processor design with a reconfigurable architecture. The reconfigurable architecture has good flexibility for tradeoff between the data throughput rate and power consumption. Compared with the DSP type architecture, the proposed reconfigurable architecture can perform higher data throughput rate with shorter latency. Besides, with combination of two RS processor engines, we can easily double the performance with the scalable design. This good scalability is another advantage of our proposed reconfigurable architecture.

1. INTRODUCTION

Reed-Solomon codec is a widely used *Forward Error Correcting* (*FEC*) technique [1]. RS codec has good error correction capability for burst errors. Many applications have employ RS codec, such as digital audio and video, computer memory storage, magnetic and optical recording, wireless mobile and satellite communications, cable modem and xDSL [2].

The block diagram of RS decoder is shown in Fig. 1. Many ASIC designs of hardware implementation use a parallel architecture with pipelined data paths to achieve a very high data throughput rate [3][4]. However, the hardware cost is very large for a parallel architecture. For some specifications, the requirement of data throughput rate is very low, for example, the DVB system [5]. Recently, there are some researches of RS decoder design employing a DSP core combined with a hardware array of fine-grain processing elements [6][7][8]. The DSP type architecture provides a lower data throughput rate with a smaller hardware area, but it needs to use a DSP to maintain the control of procedures of data processing.

In this paper, we investigate in details the methodologies of the RS decoding procedure, and design a coarse-grain unified finite-field processing element. Then, based on this coarse-grained processing element, we can simplify the complexity of control and replace the DSP core circuit with only a very small control circuit.

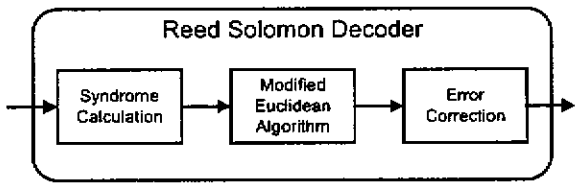


Fig. 1. Block diagram of a Reed-Solomon decoder.

2. UNIFIED FINITE-FIELD PROCESSING ELEMENT

There are three major DSP operations in the RS decoder, syndrome calculation, key equation solving, and error correction. In this section, we will analyze the common features of those DSP operations, and define the unified finite-field processing element.

2.1. Syndrome Calculation

The syndrome calculation module receives codes from channel, denoted as $r(x)$, and calculates syndrome polynomial $S(x)$.

$$S(x) = \sum_{i=0}^{2t-1} S_i x^i, \tag{1}$$

where t is the error correcting capability. Then, the coefficients of syndrome polynomial can be calculated by

$$S_i = r(\alpha^i) = \sum_{j=0}^{n-1} r_j \alpha^{ij},\ i = 0,1,\ldots,(2t-1), \quad (2)$$

where n is the code length and α^0, α^1,..., α^{2t-1} are the roots of the generation polynomial in encoding. The Eq.(2) can be represented in a recursive format as

$$S_i = r(\alpha^i) = (\ldots((r_{n-1}\alpha^i + r_{n-2})\alpha^i + r_{n-3})\ldots)\alpha^i + r_0, \quad (3)$$

and each iteration only needs a finite-field multiplication and a finite-field addition. The processing unit of syndrome calculation module is shown below.

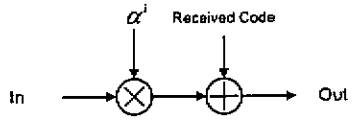


Fig. 2. Processing unit of syndrome calculation.

2.2. Modified Euclidean Algorithm

The key equation-solving module is the critical part of the RS decoder. We use the modified Euclidean algorithm to solve the key equation [9]. The key equation is

$$S(x)\sigma(x) = \omega(x) \bmod (x^{2t}). \quad (4)$$

where $\sigma(x)$ is the error location polynomial and $\omega(x)$ is the error magnitude polynomial. The modified Euclidean algorithm can be explained by following iteration descriptions.

A. *Initial Conditions*

$$\begin{cases} R_0(x) = x^{2t} \\ Q_0(x) = S(x) \end{cases}, \begin{cases} L_0(x) = 0 \\ U_0(x) = 1 \end{cases} \quad (5)$$

B. *Update Equations in Each Iteration*

$$\begin{cases} R_i(x) = s_{i-1}\left(\begin{bmatrix} bR_{i-1}(x) \\ aQ_{i-1}(x) \end{bmatrix} + x^1 \begin{bmatrix} aQ_{i-1}(x) \\ bR_{i-1}(x) \end{bmatrix}\right) \\ Q_i(x) = s_{i-1}\left(\begin{bmatrix} Q_{i-1}(x) \\ R_{i-1}(x) \end{bmatrix}\right) \end{cases} \quad (6)$$

$$\begin{cases} L_i(x) = s_{i-1}\left(\begin{bmatrix} bL_{i-1}(x) \\ aU_{i-1}(x) \end{bmatrix} + x^1 \begin{bmatrix} aU_{i-1}(x) \\ bL_{i-1}(x) \end{bmatrix}\right) \\ U_i(x) = s_{i-1}\left(\begin{bmatrix} U_{i-1}(x) \\ L_{i-1}(x) \end{bmatrix}\right) \end{cases} \quad (7)$$

$$\begin{cases} l_{i-1} = \deg(R_{i-1}(x)) - \deg(Q_{i-1}(x)) \\ s_{i-1} = \begin{cases} \begin{bmatrix} 1 & 0 \end{bmatrix} & if \quad l_{i-1} \geq 0. \\ \begin{bmatrix} 0 & 1 \end{bmatrix} & if \quad l_{i-1} < 0. \end{cases} \end{cases} \quad (8)$$

C. *Stop Condition*

$$\deg(R_{i-1}(x)) < t \quad (9)$$

D. *Output Assignments*

$$\omega(x) = Q_{i-1}(x), \quad \sigma(x) = U_{i-1}(x) \quad (10)$$

where "a" is the highest degree coefficient of $R(x)$, and "b" is the highest degree coefficient of $Q(x)$. Each iteration needs four finite-field multiplication operations and two finite-field addition operations. The processing unit of key equation-solving module is shown in Fig. 3.

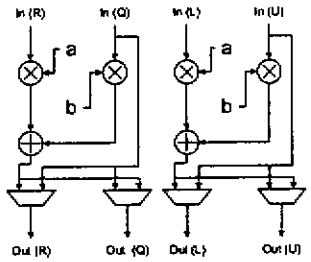


Fig. 3. Processing unit of solving key equation.

2.3. Error Correction

The error correction module finds the locations of errors by check if $\sigma(\alpha^{-k}) = 0$, from $k = 0$ to $k = (n-1)$. This is called the Chien's search. Chien's search evaluates

$$\sigma(\alpha^{-k}) = \sum_{i=0}^{v} \sigma_i \cdot (\alpha^{-i})^k, k = 0,1,\ldots,(n-1). \quad (11)$$

Forney algorithm is used for calculating the error values. If there are some errors at the coefficient r_k in the received code polynomial $r(x)$, the error value at r_k can be calculated by

$$error = \frac{\omega_0 + \omega_1(\alpha^{-1})^k + \omega_2(\alpha^{-2})^k + \ldots + \omega_{v-1}(\alpha^{-v+1})^k}{\sigma_1(\alpha^{-1})^k + \sigma_3(\alpha^{-3})^k + \ldots\sigma_v(\alpha^{-v})^k} \quad (12)$$

Eq.(11) and Eq.(12) need a finite-field multiplication and a finite-field addition for each iteration. The processing unit of error correction module is shown below.

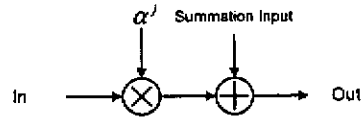


Fig. 4. Processing unit of error correction.

2.4. Unified Finite-Field Processing Element

Based on the results above, we introduce a unified finite-field *processing element* (*PE*) as shown in Fig. 5, which takes the advantage that all of the three procedures can be done at the same time. Oppositely, for a DSP type RS decoder, there is only a single procedure can be done each time, and the DSP core is needed to maintain the complex control.

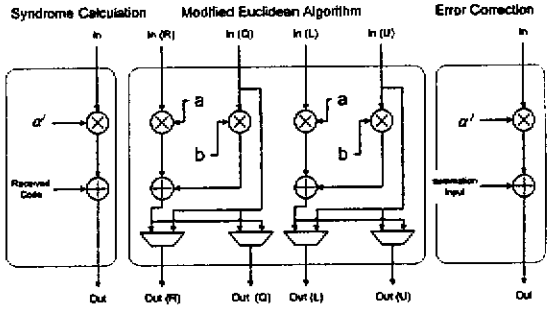


Fig. 5. Hardware architecture of the unified finite-field processing element

3. SCALABLE DESIGN OF RECONFIGURABLE RS DEODING PROCESSOR

3.1. Proposed Reconfigurable Architecture

Since a parallel architecture takes large area and dynamic power consumption, it is not suitable for low data throughput rate specifications. And a DSP type architecture needs a DSP core circuit to maintain the complex control. A reconfigurable architecture may be a good solution because it can be reconfigured to operate at a low power mode, and doesn't need a complex controller.

Assume that the number of processing elements is m. Fig. 6 shows our proposed architecture of reconfigurable m-PE RS decoder design. We split all the data registers into m segments, and each of them consists of $2t/m$ data registers. For each data register segment, we assign one PE to be responsible for all the operations performed on this data register segment.

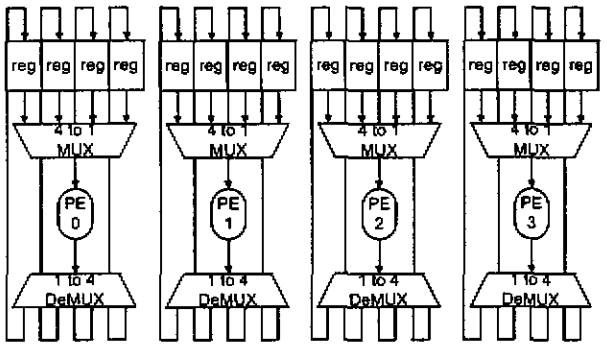


Fig. 6. Proposed scalable architecture of reconfigurable m-PE RS decoder design. ($t = 8$, $m = 4$ example)

When the data throughput rate requirement of the system specification is very low, our proposed scalable reconfigurable m-PE RS decoder can operates in single-PE mode that takes low power consumption and performs the same data throughput rate with the $2t$-folding RS decoder. When the data throughput rate requirement becomes higher, our proposed scalable m-PE RS decoder can switch to a more PEs mode to satisfy the system specification.

While in single-PE mode, as shown in Fig. 7, each time there is only one PE on operating when other PEs idle without power dissipation. However, every PE needs to operate $2t/m$ times, for processing its own $2t/m$ data registers. In details, PE0 processes reg0, reg1, reg2 and reg3 separately at time slot *0, 1, 2* and *3*, then the PE0 idles at time slot *4* to *15*. Similarly, PE1 processes reg4, reg5, reg6 and reg7 at time slot *4* to *7*, and idles at other time slots. PE2 uses time slot *8* to *11* to work, and PE3 uses time slot *12* to *15* to work. The decoder spends $2t$ time slots (clock cycles) for completing each iteration.

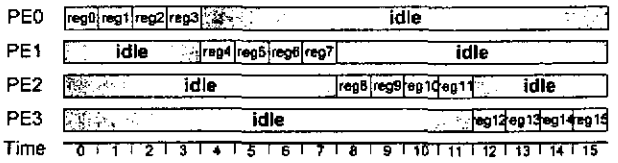


Fig. 7. Timing diagram of m-PE RS decoder in *1*PE mode for low power consumption ($t = 8$, $m = 4$ example).

While in *2*-PE mode, each time there are only two PEs on operating. And every PE still needs to operate $2t/m$ times for processing its own $2t/m$ data registers. The decoder spends t time slots for completing each iteration, which are half of *1*PE mode, as shown in Fig. 8. Other cases are similar to these. No matter in what mode, every PE always operates $2t/m$ times for processing its own $2t/m$ data registers.

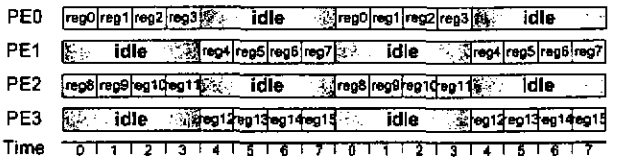


Fig. 8. Timing diagram of m-PE RS decoder in *2*PE mode ($t = 8$, $m = 4$ example).

As the result, each PE needs a $(2t/m)$-to-*1* symbol-wise multiplexer to choose from the $2t/m$ register data inputs, and outputs through a *1*-to-$(2t/m)$ de-multiplexer to its own $2t/m$ data registers. Besides, the control circuit, which generates data path control signals is just a counter, and no DSP cores are needed. The total delay on data path control is $(\log_2 t - \log_2 m + 2)$ 2-to-*1* multiplexers delay. The total data path control area is $(24t - 6m)$ symbol-wise multiplexers.

3.2. Scalability

When the data throughput rate requirement is too high, and the m-PE RS decoder cannot achieve the system specification even with m-PE mode, we can extend the m-PE RS decoder by combining two or more the m-PE RS decoder to form a *2* x m-PE or higher mode RS decoder. Fig. 9 is an example of two *4*-PE RS decoder combine together forming a *2* x *4*-PE RS decoder. Fig. 10 shows an example of four *4*-PE RS decoder combine together forming a *4* x *4*-PE RS decoder, which performs the same data throughput rate with $2t$-parallel RS decoder when $t = 8$.

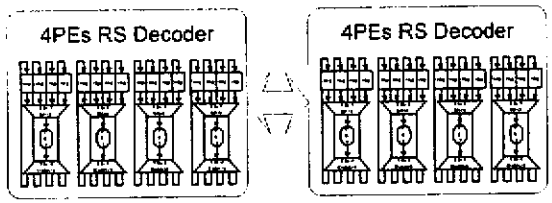


Fig. 9. Extended *2* x *4*-PE RS decoder

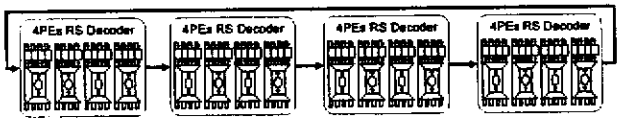


Fig. 10. Extended *4* x *4*-PE RS decoder

4. PERFORMANCE COMPARISON

By using a *16*-parallel architecture with our proposed PEs, we can use less gate counts to achieve 1 symbol/clock data throughput rate, which is the same with Lee's *16*-parallel architecture design [3]. Besides, with a scalable reconfigurable architecture for four PEs design, our *4*-PE scalable RS decoder still has good performance, with *0.25* symbol/clock data throughput rate, compared with the DSP type architectures [6][7][8]. The comparison table is shown in Table 1.

Table 1. Comparison with other architectures ($t = 8$).

	Area (gate)	Latency (clock)	Data throughput rate (symbol/clock)
Lee's *16*-parallel [3]	*55210*	$n + 66$	*1*
16-parallel with proposed PEs	*35001*	$n + 16$	*1*
Proposed scalable in *4*-PE mode	*21526*	$4n + 64$	*0.25*
Lee's DSP [6]	*N/A*	$4n + 114$	*0.22*
Ji's DSP [7]	*N/A*	$4n + 236$	*0.19*
Song's DSP [8]	*N/A*	$6n + 37$	*0.17*

For applications of the proposed *m*-PE scalable RS decoder, when the data throughput rate requirement is low, the *m*-PE scalable RS decoder can be reconfigured to single-PE mode and operate with very low power consumption similar with a *2t*-folding RS decoder design. When the data throughput rate requirement becomes higher, the *m*-PE scalable RS decoder can be easily tuned to get the higher data throughput rate requirement. When the data throughput rate becomes higher than a *m*-PE scalable RS decoder can provide, we can even extend the *m*-PE scalable RS decoder to a *2* x *m*-PE or more PEs RS decoder. We list some applications of our proposed *m*-PE scalable RS decoder in Table 2, compared with *2t*-parallel and *2t*-folding architectures.

Table 2. Comparison table based on the same PE.

	Data throughput rate (symbol/clock)	Hardware used each clock cycle			
		reg	mult	add	*2*-to-*1* mux
2t-parallel	*1*	$12t$	$12t$	$8t$	$8t$
Extended $(2t/m)$x *m*-PE	*1*	$12t$	$12t$	$8t$	$(48t^2/m) - 4t$
Extended *2* x *m*-PE	m/t	$12m$	$12m$	$8m$	$48t - 4m$
m-PE mode	$m/2t$	$6m$	$6m$	$4m$	$24t - 2m$
2-PE mode	$1/t$	*12*	*12*	*8*	$(48t/m) - 4$
*1*PE mode	$1/2t$	*6*	*6*	*4*	$(24t/m) - 2$
2t-folding	$1/2t$	*6*	*6*	*4*	*4*

5. CONCLUSIONS

In this paper, we have developed a scalable VLSI architecture of reconfigurable Reed-Solomon decoding processor based on unified finite-field processing elements. Our proposed *m*-PE scalable RS decoder architecture has very good flexibility performing different data throughput rate from very low to very high, with acceptable power consumption on data path control. It also has an advantage on a good scalability.

6. REFERENCES

[1] S. Lin and D. J. Costello. Jr., *Error Control Coding: Fundamentals and Applications*, Englewood Cliffs, NJ: Prentice-Hall, 1983.

[2] Dennis J. Rauschmayer, *ADSL/VDSL Principles*, Macmillan Technical Publishing, Indianapolis, 1999.

[3] Hanho Lee, Meng-Lin Yu, and Leilei Song, "VLSI Design of Reed-Solomon Decoder Architectures," *IEEE ISCAS*, pp. 705-708, May 2000.

[4] T. K. Matsushima, T. Matsushima, and S. Hirasawa, "Parallel Architecture for High-Speed Reed-Solomon Codec," *Telecommunications Symposium, 1998. ITS '98 Proceedings. SBT/IEEE International*, vol. 2 , pp. 468-473, Aug. 1998.

[5] "Digital Video Broadcasting (DVB); Framing Structure, Channel Coding and Modulation for Digital Terrestrial Television," *ETSI Standard EN 300 744 V1.4.1*, Jan. 2001.

[6] Jung H. Lee, Jaesung Lee, and Myung H. Sunwoo, "Design of Application-Specific Instructions and Hardware Accelerator for Reed-Solomon Codes," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 13, pp.1346-1354, Dec. 2003.

[7] H. Michael Ji, "An Optimized Processor for Fast Reed-Solomon Encoding and Decoding," *IEEE ICASSP'02*, vol. 3, pp.3097-3100, May 2002.

[8] Leilei Song, Keshab K. Parhi, Ichiro Kuroda, and Takao Nishitani, "Hardware/Software Codesgin of Finite-Field Datapath for Low-Energy Reed-Solomon Codecs," *IEEE Transactions on VLSI Systems*, vol. 8, no. 2, April 2000.

[9] Huai-Yi Hsu and An-Yeu Wu, "VLSI Design of A Reconfigurable Multi-Mode Reed-Solomon Codec for High-Speed Communication systems," *IEEE Asia-Pacific Conference on ASIC, 2002, Proceedings*, pp.359-362, Aug. 2002.

Buffering Hardware Nested Loop of Parameterized and Embedded DSP Core

Ya-Lan Tsao and Wei-Hao Chen

Electrical Engineering Department
National Central University,
No.300, Jung-Da Rd., Jung-Li City, Taiwan 32054, R.O.C.
Tel : +886-3-4227151ext34576
Fax : +886-3-4255830
e-mail : alan@ee.ncu.edu.tw

Shyh-Jye Jou

Department Of Electronics Engineering
National Chiao Tung University
No. 1001 Ta-Hsueh Rd., Hsinchu City, Taiwan 30050, R.O.C
Tel : +886-3-5131475
Fax : +886-3-572-4361
e-mail : jerryjou@cc.nctu.edu.tw

Abstract - In this paper, a hardware nested looping structure with an instruction buffer memory is proposed for the parameterized and embedded DSP core. An optional buffer memory for the instructions in the loop is used to save much of the memory accessing power consumption during the transaction of the program memory fetching. The size of the instruction buffer memory and nested loop depth are parameterized parameters. Design examples show that the hardware overhead for the nested hardware looping scheme is only 4.20% and saves 12% power consumption.

I Introduction

Combining high performance DSP processor core with customized circuits for better performance is a trend in an application specific product. We had proposed a module generator of a parameterized and embedded DSP core [1]. Fig. 1 is block diagram of NCU_DSP. The DSP core includes basic function blocks, optional multi-function blocks, optional special blocks, parameterized data bus and data/program memory. The design parameters are selected by the designer and then a synthesizable Verilog code is generated. For many DSP functions of communication systems, our DSP core can be easily customized to meet the requirement of performance [1].

Table 1 Comparison of Software & Hardware looping

	8-point FFT (Software Looping)	8-point FFT (Hardware Looping)
Program length	161	142
Cycle time	927	736

Efficient looping operation is critical to digital signal processing because most of the DSP algorithms are repetitive. Consequently, a good DSP core shall support zero-overhead looping with dedicated internal hardware [1,2]. Table 1 shows the efficiency of hardware looping comparing with software looping for a 3-level nested loop. Beside hardware looping, using a small buffered memory for the instructions in a loop is a good idea to save memory access power consumption [3,4,5]. However, the design of instruction buffer for only the inner most loop are not efficient in executing nested loop [3,4]. Also, the use of instructions cache has some instruction cycle penalty if the cache is missed [4,5]. In our design, the instruction buffer only stores the instructions in loop. The size of the instruction buffer is much smaller than the program memory. So a small instruction buffer can be placed closer to the instruction decoder to save power consumed not only cause by large program memory but also by longer bus. The operation of instruction buffer involves no miss or instruction cycle penalties. Moreover, in our parameterized DSP core, the designer can decide which level of the nested loops will be stored into instruction buffer. Also, a key difference of our design is that our embedded DSP core is parameterized and the buffer size can be selected by designer according to the length of loops in the applications.

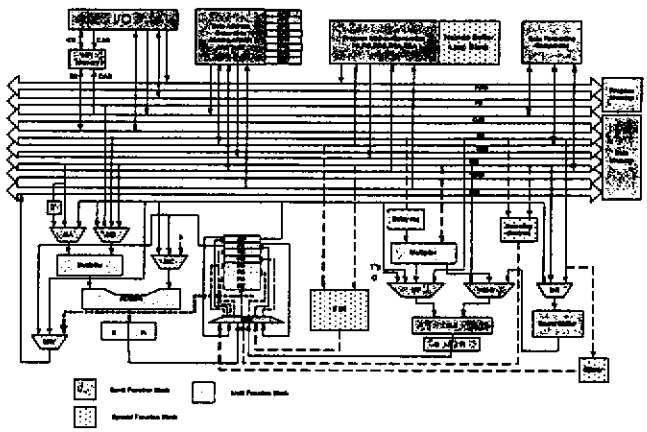


Fig. 1 Block diagram of NCU_DSP

II. Design of Buffered Hardware Looping

To achieve the function of hardware looping, program address generating unit (PAGU) shall includes Block Repeat Counter (BRC), Repeat Start Address register (RSA) and Repeat End Address register (REA). Also two comparators shall be used to check if BRC equal to zero and if Program Counter (PC) is equal to REA. The loops of a DSP program can be partitioned into 3 categories: loops with single instruction, loops with a block of instruction (non-nested) and loops with nested looping.

For repeating a single instruction, the pointer of the program memory address is program counter (PC). Once the instruction is fetched and stored in the instruction register of instruction decoder, this instruction will not fetch again for the rest of looping.

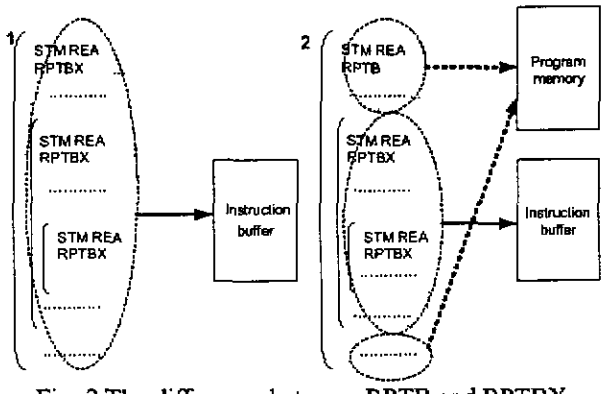


Fig. 2 The difference between RPTB and RPTBX

The circuit for loops with a block of instructions needs another pointer like PC when using instruction buffer because the instructions are now in the buffer. The addresses of instructions in buffer are different from those in program memory. We use the buffered program counter (BPC) as a pointer to indicate the addresses of instructions in the buffer and control the fetch of instruction buffer.

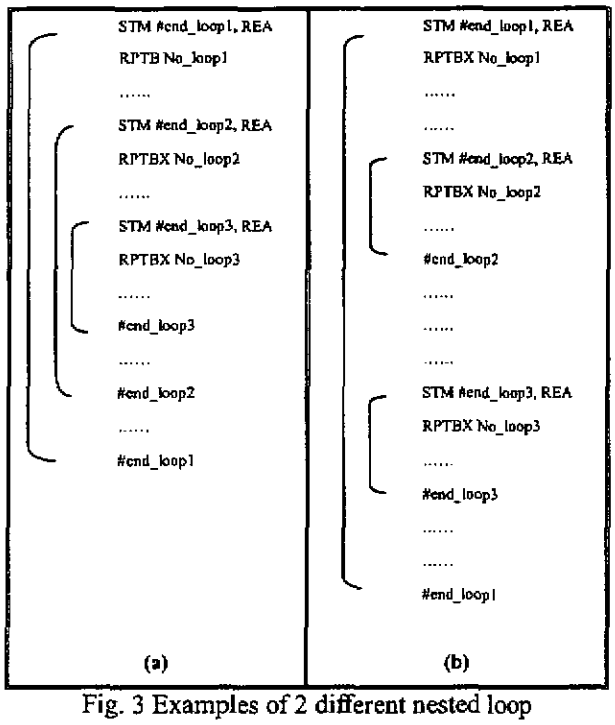


Fig. 3 Examples of 2 different nested loop

Basically, the loops with nested looping use the same concept of loops with non-nested looping. However, in the beginning of execution the lower level loop, system has to store the information in the upper level loop. The control circuits will store the information in a memory by first in last out sequence like a stack. When instructions in the lower level loop inside a nested loop are executed, the RSA, REA and BRC of the upper level loop will be pushed to a stack. Furthermore, the circuit needs a different set of RSA and REA when using instruction buffer. The RSA and REA in program memory have to transfer to the address of the instruction buffer when nested loops fetching from instruction buffer.

The instruction of repeating block instructions is RPTB and RPTBX when using instruction buffer. Fig. 2 illustrates the difference of instruction fetch between RPTB and RPTBX. On the left of Fig. 2, entire nested loop is intended to store in to the instruction buffer. All loops are using RPTBX instruction. On the right of Fig. 2, loop of level 1 is not intended to store into the instruction buffer. Only level 2 and level 3 use RPTBX instruction. It is note that the beginning of the loop is a STM instruction following by a RPTBX or RPTB. This is because there are 3 values has to be stored, RSA, REA and BRC. The machine code of the proposed DSP core instructions are 24 bits and can only pass 2 sets of operands. Therefore, looping operation has to be partitioned into 2 instructions to accomplish 3 sets of operands. The RPTB and RPTBX can be used in a nested loop at the same time. Fig. 3 shows two examples of nested looping in different configurations. In Fig. 3(a), the top level of the loop is not stored in the instruction buffer. The instruction block of "loop2" and "loop3" will be stored into the instruction buffer. Fig. 3(b) illustrates another complicated nested loop can be executed in this design. Thus, for a long loop with small loop count, RPTB used to avoid the requirement of using larger instruction buffer. However, for a short loop with large loop count, like most DSP operation, RPTBX should be used. The instruction buffer size is the maximum number of the instructions in the RPTBX nested loop.

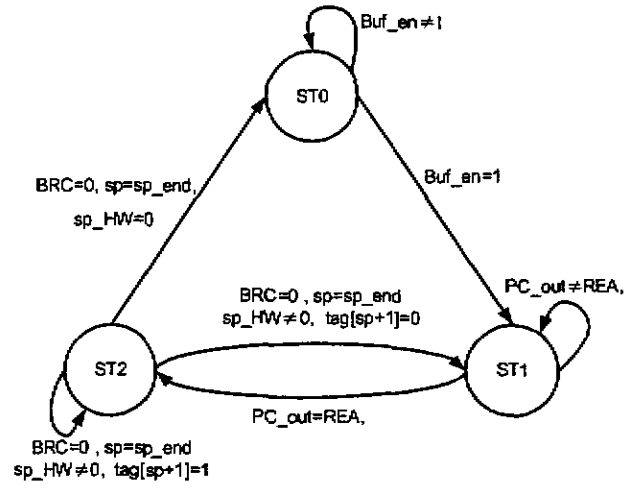


Fig. 4 The state diagram of RPTBX

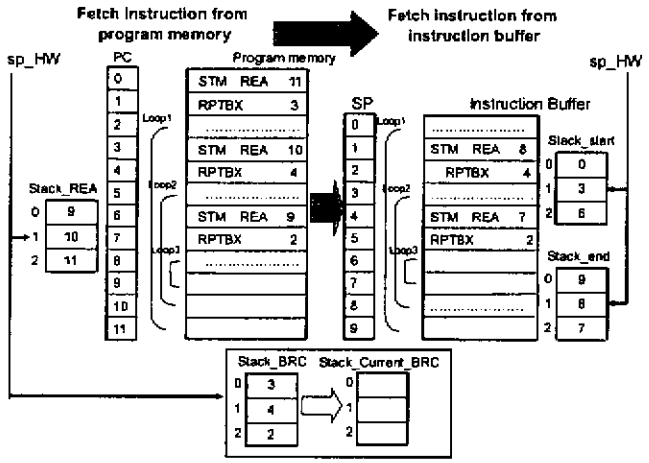


Fig. 5 An example of RPTBX instruction

Fig. 4 shows the control sequence of buffered hardware nested looping. The ST0 state is the default setting to fetch instructions from the program memory. It will stay in this state unless the RPTBX is being executed. The ST1 state is to fetch instructions from the program memory and in the same time store the instruction to instruction buffer. This state is the first looping time of RPTBX instruction. In this state, the PC and BPC are increased in the same time. In the successive looping, ST2 state entered, instructions are fetched from the instruction buffer. The PC and program memory are disabled in this state. Although the overhead circuit of instruction buffer scheme will consummate extra power, but some circuits including program memory are zero switching at this state. Fig. 5 is an example to show the different between BPC, REA and RSA. For example, REA of loop1 is 11 in program memory; 9 in the instruction buffer.

III. Design analysis and Results

When using a DSP core with proposed instruction buffer design, user must aware the size of instruction. We suggest the user can generate DSP core according the maximum loop which needed to store in the instruction buffer. But keeping the instruction buffer size smaller is better. The design example in Table 2 shows 4.20% of overhead with the instruction buffer. Table 2 shows the chip summary and the hardware overhead of instruction buffer we used in the design. It is note that NCU_DSP core is a parameterized DSP core. Designer can select parameters according application demanded.

We use a 8-point FFT program to verify power consumption saving in buffering hardware nested loop scheme. This program executes the FFT algorithm in three-level nested buffered loops. Total clock cycle time is 736 clock cycles. The loop body has 22 instructions. The loop had been repeat 14 times.

On the other hand, the same program using three-level non-buffered loops (RPTB) will also be simulated for comparison.

The DSP core example is implemented by TSMC 0.25um standard cell process. The length of instruction buffer is 32 words. The total power is 12% saving in post-layout simulation by NanoSim®[6]. Table 3 shows the comparison of RPTB and RPTBX. By further observation, the simulation indicates the first loop of the buffered loop (RPTBX) is only consumed 1.63% extra power because of fetch program memory and store to the instruction buffer. It also shows this small instruction buffer and its control circuit consumed only 1/7 power of the program memory.

Table 2 Chip summary

Process	TSMC 0.25um 1P5M
Cell library	Artisan 0.25 Cell Library
Supply voltage	2.5V
Maximum operating frequency	125MHz
DSP core area with instruction buffer(32*24bit)	123172 gate counts
DSP core asrea without instruction buffer	118196 gate counts
Memory	512×24-bit (Program memory), 512×16-bit(Data memory), 64×16-bit (HPI memory)

Table 3 Power consumption of RPTB and RPTBX

8-bit FFT nested loop	Using RPTB	Using RPTBX
Power@100MHz	275.96 mw	247.08 mw

IV. Conclusions

The buffered hardware nested looping in our design is not only provides a zero-overhead looping scheme. It will also save the power consumption in program memory fetch. Our design has optional instruction for programmer to decide which level of nested loop to be stored to the buffer according the usage of instruction buffer size. The programmer can decide which loop or which level of the loop to be buffered by specified instructions. The design result shows 4.20% of hardware overhead saving 12% of power in each instruction execution. The power of instruction buffer and its control circuit are 1/7 of the program memory.

Acknowledgements

This work was supported in part by the National Science Council, Taiwan.

References

[1] Y. L. Tsao, W. H. Chen, M. H. Tan, M. C. Lin and S. J. Jou, "Low Power Embedded DSP Core for Communication Systems," *EURASIP Journal on Applied Signal Processing*, pp. 1355-1370, vol. 2003, no. 13, Dec 2003.

[2] P. Lapsley, J. Bier, A. Shoham, E. A. Lee, *DSP Processor Fundamentals, IEEE Press*, 1997.

[3] M. Lewis, L. Brackenbury, "An instruction buffer for a low-power DSP," *Proceeding of IEEE Symposium on Asynchronous Circuit and System*, pp.176 -186, 2000

[4] R.S. Bajwa, M. Hiraki, H. Kojima, D.J. Gorny, K. Nitta, A. Shridhar, K. Seki, K. Sasaki, " Instruction buffering to reduce power in processors for signal processing," *IEEE Transactions on VLSI Systems*, Vol.5, Issue 4, pp.417-424 1997

[5] C. T. Wu, T.T. Hwang, "Instruction buffering for nested loops in low power design," *Proceedings of International Symposium on Circuit and System*, vol.4, pp.IV-81 -IV-84, May 2002..

[6] NanoSim®, Synopsys, Inc. Version U-2003.03-SP2.

LOW POWER CORRELATOR OF DSP CORE FOR COMMUNICATION SYSTEM

Ya-Lan Tsao, Jun-Xian Teng, Maw-Ching Lin
Electrical Engineering Department of National Central University, Taiwan, R.O.C. alan@ee.ncu.edu.tw

Shyh-Jye Jou
Department Of Electronics Engineering National Chiao Tung University, Taiwan, R.O.C
jerryjou@cc.nctu.edu.tw

ABSTRACT

Correlator is the most demanded block in communication systems. In this paper, low power correlator block is implemented into an embedded DSP core as a special block. This special block of correlator can employ data bus bandwidth of DSP efficiently. The design results show that it use only 3235 gate counts and can reduce the correlation operations from 20 instructions (without correlator) to only 4 instructions.

1. INTRODUCTION

An important feature of the WCDMA system, one of the candidates of 3G standard, is the requirement to spread data. Fig. 1 shows a basic transceiver structure of WCDMA system. In order to transmit data with different rate in the same bandwidth, all the base-band data is spread by a set of codes in the transmitter into a unity code rate. The spreading code generation definition can be found in 3GPP specification [1]. Under this spreading code scheme, the receiver needs a match filter block to recover data. This operation is defined as "dispreading data". The dispreading can be done by using correlation block with smaller chip area [2]. Although the primary function of the correlator is data dispreading, it also can be used in cell search procedures. Thus, in different sections of the system, different correlation operations are required.

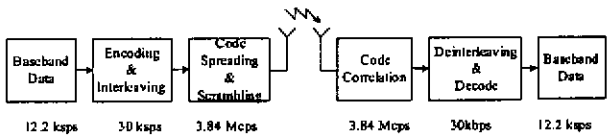


Fig. 1 Block diagram of a WCDMA communication system

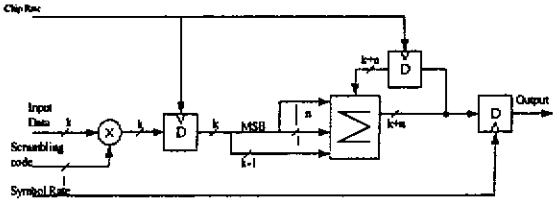
There are three steps in the WCDMA cell search procedure [3]. In each step, there are different correlation requirements. This implies we need different data flow and hardware to complete all three steps. The ASIC approaches will never meet all the requirements encountered without designing all data-path [4]. In other aspect, a DSP core based approach can be easily changed to fit the different requirement. There are some approaches that combine the correlator into a DSP core. In [5], it is much like a bank of correlator which is controlled by DSP core and works as a co-processor.

There are also now several DSPs on the market that have been designed with wireless applications in mind, for instance, the TI C54x series, the Lucent 16000 series and the ADI21xx series, etc.. In the consideration of WCDMA application, a correlator coprocessor is built up of TI C54x as a loosely coupled coprocessor (LCC). The concepts of LCC indicating the coprocessor in C54x operated in symbol rate which occupies many clock cycles. The function of coprocessor can easily be changed by changing instructions. However, this design needs timing synchronization and data I/O between different coprocessor. In order to accelerate the processing ability of DSP in WCDMA, we propose a correlator as a special block in our NCU_DSP core [6]. The proposed correlator special block works in the same clock rate as the DSP core. Thus, the correlator can benefit from the full bandwidth of DSP core.

In the following of this paper, Section 2 will demonstrate a low power correlator design. Section 3 will show the design of the Correlator in DSP Core. The design results will be shown in Section 4. Finally, a conclusion will be made in Section 5.

2. LOW POWER CORRELATOR DESIGN

There are several different correlator structures have been proposed, two's complement, sign magnitude and biased implementation [2]. Fig.2. shows the three different structures. The implementation in two complement is shown in Fig. 2(a). The incoming data are first multiplied with scrambling code and the results are accumulated. The most power consuming operation in correlation is accumulation. Thus, a low power accumulator is needed in the correlator. There are two main sources that cause power consumption: high switching probability and sign extension.



(a)

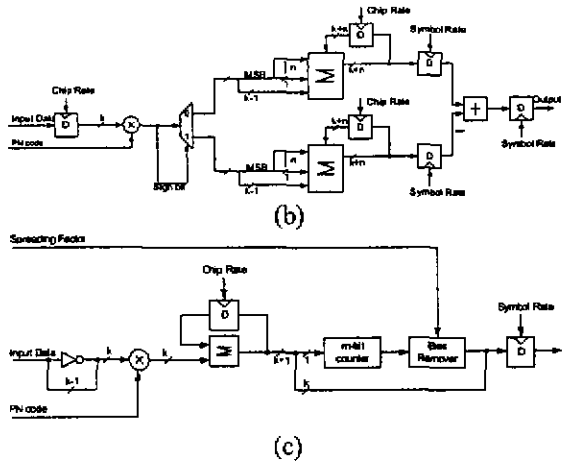


Fig.2 Accumulator of Correlators by using (a) two complement, (b) sign magnitude and (c) biased number implementation.

As the code characteristics shown in Table 1, the outputs of correlator are suspend in the range from $(2^{(L+1)/2}-1)$ to $-(2^{(L+1)/2}+1)$ (with L Odd) or from $(2^{(L+2)/2}-1)$ to $-(2^{(L+2)/2}+1)$ (with L even) most of times, where L is the length of the PN Code generator. Thus, we can say that the range of correlator result is around zero. Two complement number system has a large switching probability in output when the value of data is around zero (from positive to negative or vice verse). In CMOS circuits, the dynamic power consumption equation $Pd= Pt(C*f*V_{dd}^2)$ in CMOS circuits, where Pt indicates switching probability, C for switched capacitance in the circuits, f for operation frequency and V_{dd} for supply voltage. Thus, large switching probability consumes large power consumption. The code multiplier output is k-bits long while the other operand from D-FF is (k+n) bits long. (n is the number of accumulation and it depends on the spreading factor. k is the number of ADC output bits and it depends on the system precision requirement). Therefore, the output of the multiplier has to be sign extended to (k+n) bits. Thus, it needs extra n bit full adder which means more switching activity in the accumulator.

Table 1 The orthogonal characteristics of correlation

User Pattern	Received signal			
	1,1,1,1	1,1,-1,-1	1,-1,1,-1	-1,1,1,-1
1,1,1,1	4	0	0	0
1,1,-1,-1	0	4	0	0
1,-1,1,-1	0	0	4	0
1,-1,-1,1	0	0	0	4
	Correlated Results			

In order to solve the power consumption problem, there are two different structures being proposed: Sign Magnitude with a Positive and Negative Accumulator [7] and Biased Number System [8].

The main feature of the sign magnitude system is to accumulate positive values and negative values separately. Its block diagram of sign magnitude is shown in Fig.2(b). During the operation of correlator, the input data are correlated with PN code at first and then separated into two arms according to its sign bits. The positive values and negative values will be accumulated separately by two independent accumulators in chip rate. Every symbol period, a subtraction operation between positive accumulator outputs and negative outputs is performed to get the final value. Due to the characteristics of sign magnitude system, we can lower the switching probability significantly when the accumulator results are around zero. Another proposed correlator structure is the biased number system. The basic idea of this implementation is based on Eqn.(1):

$$A_{comp} = -(a_{n-1}) \cdot 2^{n-1} + \sum_{i=0}^{n-2} a_i 2_i$$
$$A_{biased} = (\overline{a_{n-1}}) \cdot 2^{n-1} + \sum_{i=0}^{n-2} a_i 2_i = 2^{n-1} - (a_{n-1}) \cdot 2^{n-1} + \sum_{i=0}^{n-2} a_i 2_i = 2^{n-1} + A_{comp} \quad (1)$$

Every operand in biased number system is added with a MSB value. In a k-bit system, all the incoming data is biased with a 2^{k-1} amount. The bias amount addition can be accomplished by simply reverse the MSB of incoming data. The most important characteristic of biased number system is that it shifts the represented numbers to the range 0 to 2^{k-1} (the two complement number can express numbers from -2^{k-1} to $2^{k-1}-1$.) Since all the input bits are positive value, outputs of the correlator in biased number system are always positive value. As a result, the high switching probability caused by zero around output value can be avoided. However, we need to subtract the bias from the result of the accumulator. Due to the characteristic of binary number system, this can be achieved by simply reverse the bit corresponding to the accumulated biased value. With this comprehension, we can design a correlator with different spreading factors. The block diagram of the correlator in biased number system is shown in Fig.2(c).

The sign extension problem can be avoided by using an m-bit counter to handle upper bits of correlator result. Because incoming data is fixed in k bits long, which means adder only need to handle the lower k bits addition, and the counter is responsible for carry out counting. Combining the counter value and k bits of the accumulator is the correlator output. With this methodology, the higher switching activity of clock signal in higher bits is eliminated. The design results show that by using the biased (sign-magnitude) number implementation in the accumulator of correlator, the power consumption saving is 25% (16%) as compared with two complement implementation. Obviously, the biased number based correlator structure is the one consumes the lowest power.

3. SPECIAL BLOCK OF CORRELATOR IN DSP CORE

The modulation and demodulation procedures of WCDMA are divided into several different steps. In the following, we will put our focus on the correlator requirement of each steps, more detail information about cell search procedure can be found in [1] [3].

Step 1: Slot synchronization

In the WCDMA system, messages are transmitted in the format shown in Fig.4. Thus, receiver has to accomplish two-step synchronization, slot and frame synchronization. The first step is to do slot synchronization with information in a specialized channel, P-SCH (Primary Synchronization CHannel). P-SCH is composed of PSC

(Primary Synchronization Code), as shown in Fig.5. The same PSC is broadcasted in every cell in the system [9]. Thus, correlation between received data and a fixed PSC is needed in the slot synchronization step.

Step2: Frame synchronization and code-group identification

In the WCDMA system, Secondary SCH is designed for frame synchronization. Distinct from step 1, there are 64 different S-SCH code group defined in [1]. Each code group defines a set of fifteen codes to compose a frame, or fifteen slots. In the CDMA system, PN code group is identified by SSC [10]. Thus, not only the frame synchronization is done in step 2, the PN code-group identification is also accomplished in step 2. In short, we need to correlate the received data with 64 different S-SCH codes for two purposes, one for frame synchronization and one for code-group identification.

Step3: Scrambling-code identification

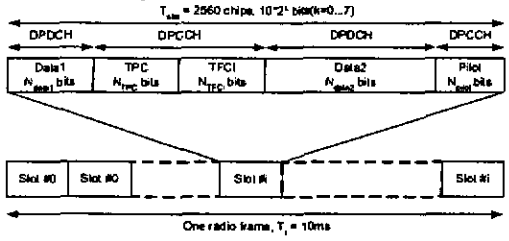


Fig. 4 Downlink dedicated physical channel frame structure

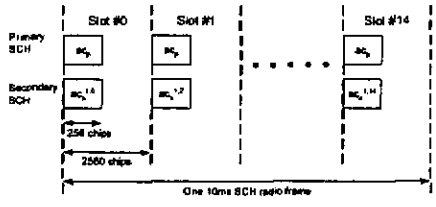


Fig.5 Synchronization channel structure

After step 2, we need to identify which code group the received signal is. Thus, the receiver narrow the spreading code searching range from 512 primary PN codes to 8 primary PN codes. It means the correlation requirement in the step 3 is to correlate received data with eight different primary PN codes.

Besides the cell search procedure, correlator also plays an important role in the tracking phase. The mobile station has to synchronize frequency and track phase. In summary, there are three different correlation conditions, a data sequence are correlated with a certain code, a data sequence are correlated with different numbers of codes, and a permuted data sequence are correlated with a certain code.

There are tree possible input data arrangement conditions of correlator:

1. Batch signal data correlation
2. Serial input data for real-time correlation
3. Different data sources for synchronization

Based on the analysis of correlation conditions and DSP cooperation consideration, we propose a correlator special block as shown in Fig.6. The main design principles are trying to design a low power, hardware sharing DSP special block, which is specialized for receiver. Basically, it can correlate the data in 4 different correlation conditions. It also can correlate 4 data in parallel within 1 correlation operation.

As the system requires different kinds of PN code [1], so the PN codes need lot of memory space. We try to use code generator to do PN code generation, and D-FF are going to store PSC and SSC read from memory. The four "C_gen/DFF" blocks are designed to be responsible for reference code generation and code shifting operation for proper operands alignment. The hardware of code generator is designed with specification in [1], and the four D Flip-Flops in four "C_gen/DFF" blocks are actually built in the form of shift registers.

Besides code source, the data source is properly designed for different correlation conditions. The data bus (word length) in NCU_DSP core is N bits long, where N depends on user configuration. In here N is assumed to be 16, and the receiver system supports 4-bit precision in ADC. If DSP read a 4-bit received data by a 16-bit data bus each time, it obviously causes a great waste of bandwidth. Therefore, we rearrange the incoming 4 data as a word. Thus, each memory access can read four received data each time. Under this data arrangement, a three level carry save adder (CSA) is built for multi-operands accumulation. There are two main benefits from using this data arrangement:

1. Higher data bus utility rate: Data bus utility rate is improved from twenty-five percents to nearly one hundred percents.

Lower power consumption: Because this proposed correlator accumulating four data in one cycle, we adopt CSA architecture for multi-operand accumulation. This implies that the proposed correlator only needs one-fourth cycle of traditional correlator, and eliminates unnecessary carry propagation. Thus, we can expect for a further power consumption reduction.

As described previously, there are three different correlation conditions. We illustrate the data flow for different conditions.

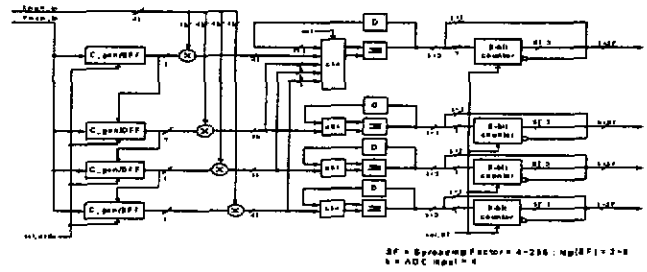


Fig.6 Architecture of correlator

Condition 1: As the data arrangement illustrated in Fig. 7(a), four successive incoming data are read in one cycle. In this example the data word (p_{ij}) is 4 bits. The word length of our NCU_DSP core is 16. It can correlate 4 data in one clock cycle. The four incoming data also can be multiplied with four different reference codes depending on the application. Results from four multipliers are accumulated in four corresponding accumulators independently.

Condition 2: As shown in Fig. 7(b), one received data is correlated with one dedicated code in an instruction

cycle. Each incoming data are first multiplied with its corresponding code. The results from four code multipliers are then put together and accumulated in the accumulator at the uppermost arm.

Condition 3: During tracking phase, received data are permuted in the sequence of early, late and data. In the mean time, correlator has to compute the CPICH correlation for additional channel information [11]. Thus, there are totally four different correlations required. The data flow is shown in Fig. 7(c).

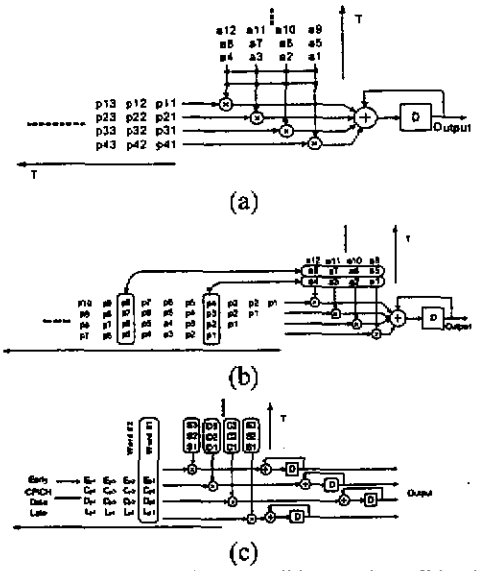


Fig.7 Data flow of correlator in (a) condition 1, (b) condition 2 and (c) condition 3.

4. DESIGN RESULT

NCU_DSP core is a configurable DSP core, which means it is flexible in many applications. We propose the correlator as a special functional block which is suitable for communication system especially for WCDMA. With proper design, a minor change in arithmetic scheme and data format can help a lot in reducing instruction cycle. As shown in Table 4 the correlator special block can help DSP to have five times higher efficiency during correlation operation. It not only reduces four times less memory access, but also reduces eight times less instruction cycle consumed in data-path operation. Table 5 shows the synthesis results of the correlator special function block. As a special functional block of NCU_DSP core, the correlator is only 6.04% of area increasing. Table 6 shows the summary of proposed design in 0.35um and 0.25 um.

Table 4 Number of instructions used with and without correlator block

	Data-path without Correlator	Data-path with Correlator
Input Data Writing	4 ins. cycle	1 ins. cycle
Operands Reading	4 ins. cycle	1 ins. cycle
Correlation Executing	4 ins. cycle in XOR 4 ins. cycle in accumulating	1 ins. cycle
Results Writing	4 ins. cycle	1 ins. cycle
Total	20 ins. cycle	4 ins. cycle

5. CONCLUSION

In this paper, we have proposed a low power correlator for better performance as a key processing element in 3G communication applications. Fig. 8 shows the block diagram and the special block of correlator in the NCU_DSP core. It wouldn cause extra delay in the critical path of NCU_DSP core and can finish the execution of correlation operation in one clock cycle. As mention previously, the NCU_DSP core can perform better because the correlator that has high data-bus utility rate and low-power design. The example of 16-bit word-length NCU_DSP core and 4 bits data word-length of correlation operand is shown in Table 4 and 5. The delays of this correlator at critical path are 5.96ns (0.25um) and 7.9ns (0.35um) and consume about 3235 gate counts.

Table 5 Synthesis results of correlator block

	0.25um	0.35um
Area	253um x 253um	3235 gate counts
Timing	5.96ns (5.99ns)[1]	7.9ns (8.96ns)[1]
Area Overhead	6.04%[2]	6.48%[2]

1. Critical path of NCU_DSP
2. The area of DSP core excluding memory in 0.35 um is estimated as 49895 gate count and area in 0.25um process is estimated as 1,067,057um².

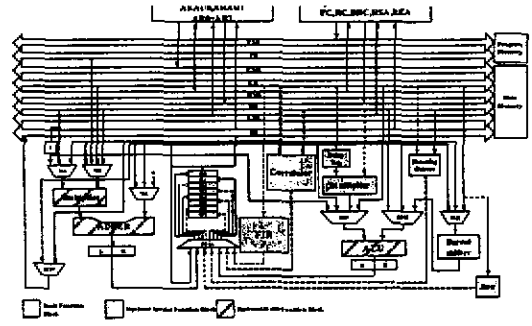


Fig.8 The special block of correlator in NCU_DSP core

11. REFERENCES

[1] 3G TS 25.213 "Technical Specification Group Radio Access Network." 3rd Generation Partnership Project, 3GPP Technical Specification, Version 5.0.0, Mar, 2002.

[2] S. Kim, B.A Daneshrad, "100 uW, 20 Meps versatile correlator chip for third generation WCDMA systems," The Thirty-Third Asilomar Conference on Signals, Systems, and Computers, Vol.1 pp.130 -134, 1999.

[3] K. Higuchi, M. Sawahashi and F. Adachi, "Fast cell search algorithm in inter-cell asynchronous DS-CDMA mobile radio," IEICE Transaction on communication, E81-B, pp1527-1534, 1998.

[4] C. Li, W. Sheen, Ho, J., Y. Chu, "ASIC design for cell search in 3GPP W-CDMA," Vehicular Technology Conference, VTC 2001 Fall. IEEE VTS 54th, Vol. 3, pp1383-1387, 2001.

[5] C. Chen, P. Tseng, Y. Chang, L. Chen, "A digital signal processor with programmable correlator array architecture for third generation wireless communication system," IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, , Vol.48 Issue 12 pp.1110 -1120, Dec 2001.

[6] Y. Tsao, S. Jou, H. Lee, Y. Chen, M. Tan "An Embedded DSP Core for Wireless Communication," Proceedings of International Symposium on Circuit and System, Vol. 4, pp.524-527, May 2002.

[7] A. Chandrakasan and R. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," Proc. of the IEEE, Vol.83, No.4, Apr.1995.

[8] M. Ercegovacc and T. Lang, "Low-Power Accumulator (Correlator)," IEEE Symp. on Low Power Electronics, Digest of Technical Papers, pp.30-31, Oct. 1995.

[9] G. D. Mandyam and J. Lai, Academic Press, "Third Generation CDMA Systems for Enhanced Data Services," chapter6, WCDMA Overview, 2002.

[10] 3G TS 25.214 "Technical Specification Group Radio Access Network; Physical layer procedures (FDD)." 3rd Generation Partnership Project, 3GPP Technical Specification, Version 5.0.0, Mar, 2002.

[11] A. Gatherer, T. Stetzler, M. McMahan, and E. Auslander, "DSP-Based Architectures for Mobile Communications: Past, Present and Future," IEEE Communications Magazine, Vol.38 Issue 1, pp.84 -90, Jan 2000.

A MODULE GENERATOR FOR PARAMETERIZED DSP CORE

Ya-Lan Tsao, Yu-Chun Lin, Wei-Hao Chen, Bo-Shiang Huang
Electrical Engineering Department of National Central University, Taiwan, R.O.C. alan@ee.ncu.edu.tw

Shyh-Jye Jou
Department Of Electronics Engineering National Chiao Tung University, Taiwan, R.O.C
jerryjou@cc.nctu.edu.tw

ABSTRACT

In this paper, a parameterized module generator of DSP core for embedded application is proposed. Some special functions for the communication applications are included in this DSP core. Moreover, key parameters of the DSP core are identified and analyzed to show their effects on performance index such as execution cycle, hardware gate count and power consumption. The parameters of the DSP core and the special functions required can be specified by user which is required by the application. The DSP core which generated by the generator is synthesizable and flexible RTL code for system integration. The turn around time of design process can be dramatically decreased. Furthermore, the generator can automatically optimize the structure of DSP core for different parameters. The design examples show that the variety of selection in implementation by using this parameterized generator and its flow provides the system designer to obtain better result in area, speed and power trade-off.

1. INTRODUCTION

Some DSPs can achieve high throughput by exploiting parallelism with specialized data paths at moderate clock frequency. For example, VLIW (Very Long Instruction Word) and SIMD (Single Instruction Multiple Data) approaches can be used to further enhance processor performance[1][2][3]. However, these approaches are not economical for dedicated application in area and power terms. Therefore, combining a dedicated, high performance DSP core with some special blocks yields a highly integrated system will be more suitable for embedded core[4][5]. To achieve higher performance on the communication algorithms, special functions like the dual MAC unit, the Hamming distance calculator, the dedicated CSD code based FIR filter, the multi-level slicer unit, the SWP (sub-word parallel) structure, and the multi-level hardware loop blocks etc. are provided as optional blocks in the core to be used as an embedded core. Also, both the peripheral and memory architecture should be designed for embedded considerations[5].

We had proposed a DSP core for embedded system applications, the NCU_DSP core [5]. The NCU_DSP core itself is parameterized with several independent parameters. In this paper, we propose a parameterized DSP core generator and its parameterized flow to automatically generate the embedded DSP core. In this way, the parameters of the DSP core and optional special blocks can be configured and the whole DSP cores can be generated for the dedicated applications.

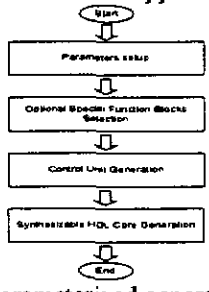


Fig. 1 Flow of parameterized generator in theoretical

To be used in the automatic IC and FPGA design flow, the RTL design of the whole DSP core can be generated by the proposed generator flow as shown in Fig.1. First, according to the specified application, all parameters can be determined by user. Then the user select the optional special functions of this DSP core which are required for computational intensity application and the control unit will be set up according to the function blocks selected. Finally, the synthesizable Verilog code will be generated by the module generator.

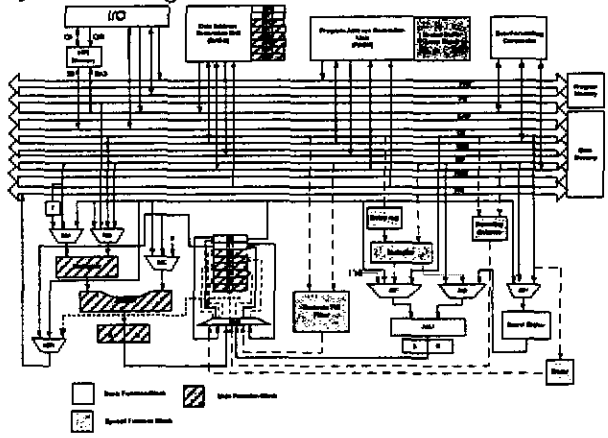


Fig. 2 Block diagram of NCU_DSP core.

In the following of this paper, section 2 overviews the architecture of the NCU_DSP core. In section 3, we will show the parameters and special functions. Section 4 will show the implementation of module generator and give some example of the implementations of the DSP core. Finally, a conclusion will be made in Section 5.

2. THE ARCHITECTURE OF DSP CORE

Fig.2 illustrates the overall architecture of NCU_DSP core. NCU_DSP is a fixed-point DSP core. The grey blocks in Fig.2 are the special functional blocks and are optional. The dashed blocks are multifunction blocks with optional functions. The DSP core itself is parameterized with several independent parameters. Users can set the parameters to fit the requirements of application. However, bus, memory architecture, I/O interface, pipeline stage sequence and the blank blocks in Fig. 2 are the same for all the parameter settings.

2.1. Bus and Memory Architecture

NCU_DSP core uses the modified Harvard architecture in bus architecture. The modified architecture contains one program memory bank and one data memory bank with separate program and data bus. The program and data memory are single port and dual port RAM, respectively. The dual port RAM indicates that the DSP core simultaneously can make two accesses to RAM. This arrangement provides a maximum of one program access and two data accesses per instruction cycle to enhance memory access capacity.

NCU_DSP core supports three addressing modes, namely the indirect addressing, register-direct addressing and immediate addressing modes. Moreover, NCU_DSP core supports circular addressing and bit-reversed addressing in the indirect addressing mode as optional functions. The program memory address is coded by of Gray code instead of straight binary code.

2.2. I/O Interface

The I/O interface of DSP core contains three operating modes, the direct data access (DMA) mode, the hand shaking mode and the merge mode. The DMA mode is provided for transferring data quickly and conveniently in batch transaction. The transfer rate is the same with the clock in the DSP core. The hand shaking mode is for transferring real-time data but the data rate is not regular. The data rate in merge mode is synchronous and regular but slower than the internal clock of the DSP core. The data transfer in the handshaking and merge modes occurs between the data outside the DSP core and the host programmable interface (HPI) memory. The HPI memory resembles a buffer of data memory.

2.3. Pipeline Stage

The NCU-DSP core contains six pipeline stages, namely instruction fetch stage (IF), instruction decode stage (ID), operand fetch stage (OP), execution one stage (EX1), execution two stage (EX2), and write back stage (WB).

3. PARAMETERIZED DSP CORE

3.1 Parameters

In this section, we will discuss the parameters of DSP core. The parameters are listed in Table 1.

Data_length, Dmem_size, Pmem_size, HPImem_size, PC_stack_size:

The data word length of DSP processor is the key parameter. Most of the data operation blocks of the DSP core will vary as the data word changed. Different data word length will have different size of chip area. For example, the number of pipeline registers, the data memory and the area of data-path blocks will be proportional to the data word length. Thus, for a fixed-point DSP algorithm, maximum required data word length shall be determined carefully to optimize the core area. The data memory size (Dmem_size), program memory size (Pmem_size) and the host programming interface memory size (HPImem_size) can be selected according to application demand. The stack memory size of program counter (PC_stack_size) will determine the number of the looping calls. From the area point of view, the memory usually occupies a huge part of the total area of DSP core. Therefore, the parameters, Dmem_size, Pmem_size, HPImem_size and PC_stack_size should be carefully determined according to applications for optimal performance.

Table 1 Parameters of DSP core for module generator

Parameter	Range	Relation	Description
Data_length	8 ~ 48 bits	m	Data word length
Dmem_size	8 ~ 65536 words	S_D	The size of Data memory It determine the data memory address length—DAL
Pmem_size	8 ~ 65536 words	S_P	The size of Program memory It determine the program memory address length--PAL
HPImem_size	8 ~ 65536 words	S_{HPI}	The size of Host port interface memory and it determine the HPI memory address length--HPIAL
IO_width	14 ~ 32 bits	max (16,m,DAL, PAL,HPIAL)	Input/Output operand width
G_width	2 ~ 16 bits	G	Guard bit length
A_width	18 ~ 112 bits	(2 * m + G)	The word length of accumulator
R-num	2 ~ 8	N_R	Number of accumulator register
AR-num	2 ~ 8	N_{AR}	Number of auxiliary register
Loop-num	2 ~ 8	N_{HL}	Hardware nested loop number
PC_stack_size	4 ~ 128 words	S_{PCS}	The size of PC stack. It determine the overall loop levels and branches.

IO_width

The I/O bus is used to transmit data, data memory address, program instruction, program memory address, HPI data, HPI memory address, and I/O control signal---HPIC (16 bits). So the IO_width uses the maximum value between these parameters---Data length, data address length (DAL), program address length (PAL), HPI address length (HPIAL) and HPIC.

G_width, A_width:

The data-path contains two major blocks, arithmetical logic unit (ALU) and multiplier and accumulator (MAC). When ALU and MAC are used within a loop, "overflow" condition shall be avoided. The guard-bit parameter—G_width is a scheme to prevent overflow during accumulation. The overall bit length of the

accumulator (A_width) is the sum of multiplier output length and the guard-bit length.

R-num:

In the DSP core, we also use some RISC concepts. The main difference between RISC processor and DSP processor is their source of operands; the RISC processor is register oriented (load-store processor) but the DSP processor is memory oriented. Obviously, the register oriented has advantages on power consumption and flexibility. Therefore, R-num, which determines the number of accumulator register, is also regarded as a parameter.

AR_num:

The DSP core provides many kinds of indirect addressing mode and the DSP needs auxiliary register to do the operation required to generate address. The parameter---AR-num can modify the auxiliary register number according to the operation requirement.

Loop_num:

The DSP core provides instructions to execute nested loops. We use the stacks to store the nested loop information. But the more nested loop level number, the more the stack size needed. Thus, user can determine Loop_num according to how many levels to be execute in the application for optimal performance.

3.2 Special Function Blocks

We integrate some application-specific block into the DSP core. These blocks are regarded as optional modules and could be chosen by the user. The optional modules we provide are listed in Table 2.

Table 2 Optional special function blocks and their overhead

	clock cycle (with/without special block)	Process	Overhead (gate count)	Overhead* (Percentage)
Hamming distance calculator (16 bits)	1/33	0.35um	174	0.35%
		0.25um	238	0.38%
Multi-Level Slicer (16 bits)	1/2N;N =log(Symbol Level)	0.35um	65	0.13%
		0.25um	175	0.28%
Sub-word MAC unit (16 bits)	1/4	0.35um	1,396	2.79%
		0.25um	1,669	2.70%
Dedicated FIR filter with N taps	1/N; N/2(Dual MAC)	Depends on spec.		
Advanced Hardwart:Looping	1/5	0.35um	850	1.70%
		0.25um	988	1.60%
Dual MAC (Additional Multiplier)	1/2	0.35um	2,488	4.98%
		0.25um	1,337	2.17%

* Area of whole DSP exclude memory in 0.35 um is estimated as 49,895 gates and area in 0.25um process is estimated as 61,751 gates.

Table 2 also shows the design information of the special functional blocks. The data of the designs were implemented by standard cell design flow of TSMC 0.25um 2.5v and 0.35um 3.3v CMOS process. Gate count of the whole DSP core excluding memory in 0.35 um and 0.25um are estimated as 49,895 gates (operating at 100Mhz) and 61,751 gates (operating at 140Mhz) respectively. The overhead are relatively small. The special blocks are carefully integrated so that they will not cause extra delay of the critical path in the DSP core.

4. MODULE GENERATOR &DESIGN EXAMPLE

The generation flow of the proposed parameterized DSP core generator has been shown in Fig. 1. The parameterized module generator is developed in C language. The design flow and its generating functions are integrated in a graphic user interface (GUI) program. The program can be executed in any PCs with Microsoft Windows operation system. The module generator is partitioned into four major parts as shown in Fig. 3: (1) Parameter Setup, (2) Data path and special blocks generation, (3) RAM Generation, (4) Control blocks generation and final DSP core.

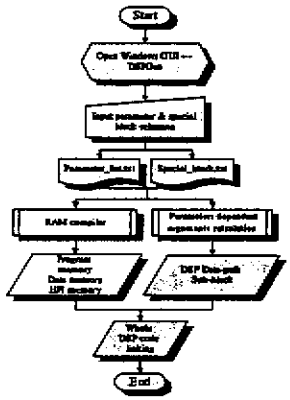


Fig. 3 Program flow of proposed DSP core generator

Although the user specifies the parameters and selects the special blocks in different steps of the user interface. Most of the parameters are also applied to the special blocks. The special blocks are generated along with the data-path modules. The memories of the DSP core are generated by a RAM compiler. The module generator passes the required parameters and synthesis for output script according to the process technology file to the RAM compiler. After all the special blocks and data-path blocks were generated, the module generator then links all the blocks and produces the control blocks. Finally, synthesizable RTL code of the DSP core and scripts for the synthesis of the core are generated in text file format under specified directory path.

4.1 Parameterized data-path design and control scheme

In the generation of data-path blocks, special functional blocks and control blocks for different parameter settings, the module generator need to consider two things: control dependency and structure dependency.

4.1.1 Structure design of parameterized data-path and special functional blocks:

In the generation of these blocks, we use two main methodologies to generate the structure for optimal performance. The first method is to describe hardware in RTL code by proper coding style, which contains the parameter information. These blocks are coded in behavioural level. The generator applies the parameters to the RTL code directly without changing the structure among the whole range of the parameters. For example, the structure of multilevel slicer is the same in all data word length. The second method is to construct different structures for different range of the selected parameters for (in) optimal performance. The structure consists of some sub-blocks. The sub-blocks are coded by RTL

code and the parameters can be directly applied in the whole range or certain rang like the first method. In this method, the structure of the blocks may different for optimal performance in different selected parameters. For example, the word length will change the levels of adder tree in the accumulator, ALU and multiplier. The barrel shifter has to balance the shift amount in two levels if the word length changes. The sub-blocks of RTL code are organized by the generator written in C language and then the generator will export a RTL code to fit for certain parameter configuration.

4.1.2 Control schemes:

Different structures may need different arrangement of control signals. We have to collect design parameters in the configurations and to generate the control signals in RTL code according to different structures.

For example, the register number of accumulation registers can range from 2 to 8. It needs different kinds of decoder for the register units. In the mean time, the existence of dual MAC path determines the number of output signals that will be fetched from the accumulation registers.

Another example is the data rearrangement in the input of sub-word multiplier. The data arrangement needs a much more complicated generator unit to keep the data in correct alignment. Different configuration needs different control signal. All these modifications have to reflect in final generated RTL blocks. This issue is handled by C sub-routine in the parameterized generation flow.

4.2 Examples

In the following, we will show some design examples and applications of the parameterized design flow. All the design examples are generated by the proposed module generator and then are synthesized by synthesis tool. The standard cell library used is in TSMC 0.25um CMOS process.

Table 3 Area and timing of Single/Sub-word MAC

	16x16 Single/Sub-word MAC	32x32 Single/Sub-word MAC
Area (gate count)	7000 / 8669	16406 / 20293
Timing (ns)	3.35 / 3.61	4.00 / 4.12

The results in Table 3 are timing and area performance of two specific MAC of 16*16 to 40 and 32*32 to 72. The area of 32*32 is about 2.5 times larger than the area of 16*16. The delay time of 32*32 design is not increased much as compared to 16*16 design because the generator optimizes the architecture for better performance and area trade-off.

Table 4 shows the DSP core generated by different parameters. The most important parameter of the DSP core is data length. It will affect every data-path blocks of the DSP core. So we use the generator to generate three cases---DSP_8 (8 bits), DSP_16 (16 bits) and DSP_32 (32 bit). The three cases are different only in data length (Data_length) and guard bit (G_width). The other conditions are the same as shown in Table 4. Also, the gate count percentage of every major blocks are also shown in Table 4. As you can see, as the data word length is increased, the gate count of the DSP is increased and the maximum operation frequency is decreased. Moreover, the percentage of the data path to the DSP core increases from 25% to 58% when the data word length increases from 8 to 32. This example shows that the generator has the capability to optimize the structure for better performance.

Table 4 Comparison of different design cases in 8, 16, 32 bits data word length

	DSP_8	DSP_16	DSP_32
MAC	Dual MAC		
Slicer	No		
Hamming distance	No		
Advanced Nested loop	Yes		
Indirect addressing modes	Circular mode and Bit reversal mode		
Dedicate FIR filter	No		
Performance			
Area (gate count)	48374	53107	92449
Clock rate(Max clock in MHz)	194	170	153
Power (mw@100MHz)	699	902	1,369
Data-path units			
MAC	8%	17%	31%
ALU	5%	7%	8%
Shifter	2%	3%	3%
Accumulation register	10%	15%	16%
Memory related & other units			
PAGU	35%	24%	16%
DAGU	13%	9%	5%
Instruction decoder	2%	2%	1%
Others	25%	23%	20%

5. CONCLUSION

The structure of NCU_DSP core is parameterized which can easily meet the requirement as demanded. Various methods of saving power have been proposed for use in the design of DSP core. These methods include bus segmentation, data access reduction, program memory access reduction, Gray code addressing and pipeline register reduction. The special functional blocks of this DSP core can achieve improved performance and flexibility with minimum area overhead. In this paper, we had proposed a parameterized module generator which can generate a NCU_DSP core by user specified parameters. It provides the designer to reduce the turn-around-time in the applications. The generator has capability to optimize the data-path structure among different parameters.

6. REFERENCES

[1]. H. Yagi, R. E. Owen, "Architectural considerations in a configurable DSP core for consumer electronics," Workshop on IEEE Signal Processing, pp.70-81, Sep 1995

[2]. C. K. Chen, P. C. Tseng, Y. C. Chang and L. G. Chen; "A digital signal processor with programmable correlator array architecture for third generation wireless communication system" *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 48, no 12, pp. 1110-1120, Dec 2001

[3]. I. Verbauwhede, M. Touriguian, "A Low Lower DSP Engine for Wireless Communications," *Journal of VLSI Signal Processing Systems*, vol. 18, no. 2, pp.177-186, Feb 1998.

[4]. B. W. Kim, J. H. Yang and C. S. Hwang, "MDSP-II: A 16-Bit DSP with Mobile Communication Accelerator," *IEEE Journal of Solid-State Circuits*, vol. 34, no.3, pp. 397-404, Mar 1999.

[5]. Y. L. Tsao, W. H. Chen, M. H. Tan, M. C. Lin and S. J. Jou, "Low Power Embedded DSP Core for Communication Systems," *EURASIP Journal on Applied Signal Processing*, pp. 1355-1370, vol. 2003, no. 13, Dec 2003.

Multiplierless Multirate Decimator / Interpolator Module Generator

Shyh-Jye Jou, Kai-Yuan Jheng*, Hsiao-Yun Chen and An-Yeu Wu*

Department of Electrical Engineering, National Central University, Chung-Li, 320, Taiwan, R.O.C
*Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, 106, Taiwan, R.O.C

ABSTRACT

A module generator, which can automate the process of designing high-speed low-complexity multistage multirate decimator / interpolator, is presented. The generator exploit architectural symmetries in linear phase filters and multistage multirate interpolated FIR filter design methodology for low complexity. In addition, the polyphase representation is used to decompose the filter into subfilters. The resulting filters utilize canonic signed digit (CSD) multipliers, a transposed direct form structure, and carry-save addition for high speed. A filter design example with TSMC 0.25μm standard cell for 64-QAM baseband demodulator shows that the area is reduced by 39% for low-complexity applications. Moreover, for high-speed application, the chip can operate at 714MHz. Finally, a designed decimator which is used in the CDMA cellular shows that the area is reduced by 70% as compared with conventional approach.

I. INTRODUCTION

The applications of digital FIR filter and up / down sampling (DSP) techniques are found everywhere in modern electronic products. For every electronic product, lower circuit complexity is always an important design target since it reduces the cost. In this paper, a general-purpose multiplierless multirate decimator / interpolator module generator, which is based on canonic signed digit (CSD) code representation and multistage multirate interpolated FIR (IFIR) filter design methodology [1], are proposed. Several design methodologies were adopted to reduce the hardware complexity at architectural level. Therefore, by using this module generator, an efficient design of a digital FIR filter / decimator / interpolator can successfully completed in a few minutes.

The rest of this paper is organized as follows. In Section 2, the design flow of the module generator and several hardware reduction methods are presented. Some experimental results of the decimator and the interpolator design examples synthesized with our module generator are then shown in Section 3. Finally, some conclusions will be given in Section 4.

II. MODULE GENERATOR IMPLEMENTATION

The system configuration and dataflow of the module generator are shown in Fig. 1. The module generator consists of many subprograms. The main subprograms are the multistage architecture analysis and synthesis, the coefficient optimization, the word length estimation and the synthesizable Verilog code generation. All programs are written in C++ language.

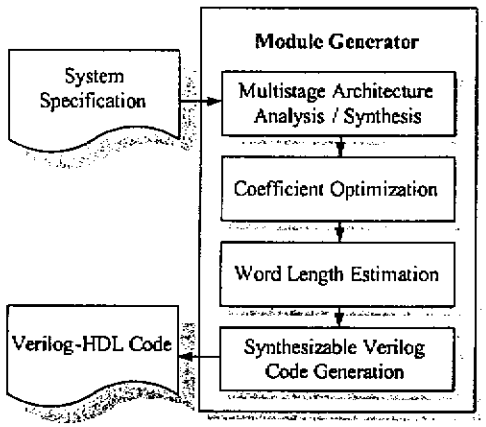


Fig. 1 Design flow of the module generator.

Table 1 Input data of system specification.

Filter Type (LP, HP, BP, BS, Decimator, Interpolator)	
Normalized Passband and Stopband Edge Frequencies	ω_P , ω_S
Passband Ripple in Magnitude or dB	δ_P / A_P
Stopband Attenuation in Magnitude or dB	δ_S / A_S
Input and/or Output Data Word Length (bit)	W_{in} / W_{out}
Signal to Noise Ratio (dB)	SNR
Up Conversion Ratio (for Interpolator)	L
Down Conversion Ratio (for Decimator)	M

In this system, the input is the system-level specification, which listed in Table 1. Moreover, user can also directly input the coefficient of the filter.

A. Multistage Architecture Analysis

In many applications, it is usually necessary to design a decimator / interpolator with a large decimation / interpolation ratio. Although this can be done by designing a filter directly and using the polyphase structure to save the arithmetic operations, it is more efficient to design in multiple stages, and the IFIR technique [1] is still applicable. The choice of the number of stages K and the sampling rate conversion ratio M / L is not a trivial problem. However, in practice, the number of stages K is rarely larger than four. Furthermore, take decimator as an example, for a given value of M, there are only a limited set of possible integer factors. Thus, a feasible approach is to determine all the possible factors of M. We use multistage IFIR technique and follow the relationship of the sampling rate conversion ratio (SRCR):

$$M1 \geq M2 \geq \ldots \geq Mk$$

to decompose the decimator into all the possible multistage sets. For a K-stage decimator, the filter specification for each stage to ensure that the overall filter requirements are

met are shown in Fig. 2, where the passband ripple is δ_P/K, and the stopband ripple is δ_S.

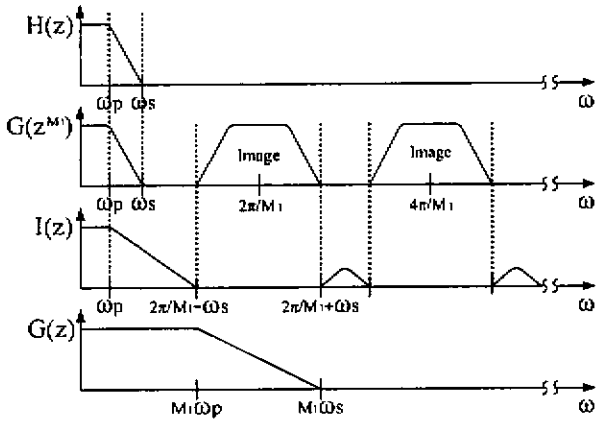


Fig. 2 Specifications of multistage IFIR filter design.

Moreover, the polyphase decomposition [3] is used to decompose the filter into subfilters. In order to take both high-speed and low-speed applications into consideration, the transposed direct form structure is chosen. However, the decomposition of the linear phase filter into M subfilters will usually result in nonlinear phase subfilters. Therefore, when the SRCR is even, the filter with odd tap length N can be redesigned to be N+1 to have two more number of linear phase subfilters to reduce the hardware complexity [4].

B. Coefficient Optimization

In this subprogram, we integrate the MATLAB engine into our module generator. The floating-point filter coefficient set is calculated by the generalized Reméz method as given in the MATLAB gremez.m function.

Numerous search algorithms for the design of multiplierless filters with canonic signed digit (CSD) or signed powers-of-two (SPT) coefficients have been proposed. However, they did not explore the possibility of further reduction of nonzero digits by taking the filter order as a variable parameter. In general, the tap length of a filter will be increased when only the passband ripple is decreased or the stopband attenuation is increased without changing the frequency specifications. Thus, we can allow more margins for coefficient quantization error by increasing the filter tap length. Besides, an observation shows that one can start with a filter, which exceeds the given criteria that may involve acceptable level of increase in the filter order, but with much lesser total nonzero digits than the initial design. Therefore, we adopt a two-step search methodology to first exploit the filter order and then use a two-step local search algorithm proposed by Samueli [5] to search for a optimal coefficients of a filter with CSD codes.

C. Word Length Estimation

In digital signal processing, the finite word length has a strong effect on the system performance since it dominates the precision of the output signals. The increment of internal word length will lead to a better signal-to-noise ratio (SNR), but it will also increase the hardware complexity, consume

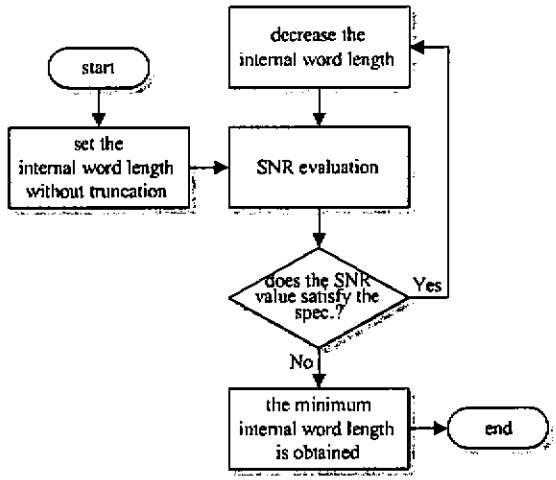


Fig. 3 Internal word length reduction flowchart.

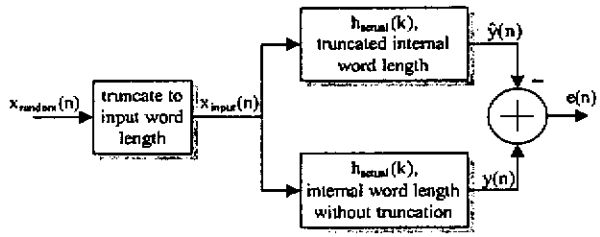


Fig. 4 SNR evaluation block.

more power, and slow down the system operation frequency. Therefore, it is a trade-off that the designer should be take care. In this subprogram, the SNR is defined as

$$SNR = 10\log\left[\frac{E\left(y^2(n)\right)}{E\left(e^2(n)\right)}\right] = 10\log\left[\frac{E\left(y^2(n)\right)}{E\left(y(n)-\hat{y}(n)\right)^2}\right]$$

The internal word length reduction flowchart and SNR evaluation block are shown in Fig. 3 and Fig. 4 respectively. The initial internal word length will be evaluated for the result that does not introduce any error first. Then the internal word length will be decreased to the value that its SNR value still fits the specification. Finally, the minimum internal word length, which fulfills the specification, will be obtained.

D. Synthesizable Verilog Code Generation

a. Multirate Decimator / Interpolator

Considering a decimator shown in Fig. 5(a), the lowpass filter H(z) will be a narrow band case as the decimation ratio M becomes large. The IFIR technique can be used to reduce the hardware complexity of H(z). If we carefully design the interpolation factor L of the periodic model filter $G(z^L)$ to be M_1, as shown in Fig. 5(b), the structure of the decimator can be reconstructed into Fig. 5(c) from noble identity. By this structure, the decimator is divided into two sections, and both of them can be implemented by polyphase representation with less filter coefficients resulting from image suppressor I(z) and model filter G(z), as shown in Fig. 5(d). In addition, the interpolator can be designed in the same way, as shown in Fig. 6.

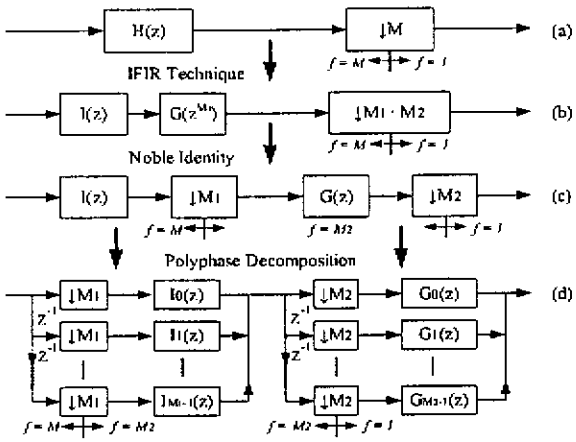


Fig. 5 Multistage IFIR decimator design.

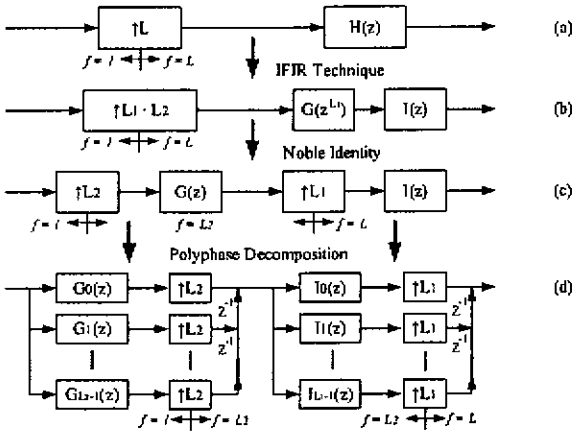


Fig. 6 Multistage IFIR interpolator design.

Both decimator and interpolator can have two structures in direct form or transposed direct form. For high-speed application, transposed direct form is selected. For the transposed direct form of decimators, the registers will be shared between the subfilters, as shown in Fig. 7 for the example of M=3, N=9. This is the so-called memory-saving technique. The transposed direct form of interpolators allow multipliers to be shared between the subfilter in each mirror symmetric pair, as shown in Fig. 8. This is the so-called mirror symmetric filter pairs technique.

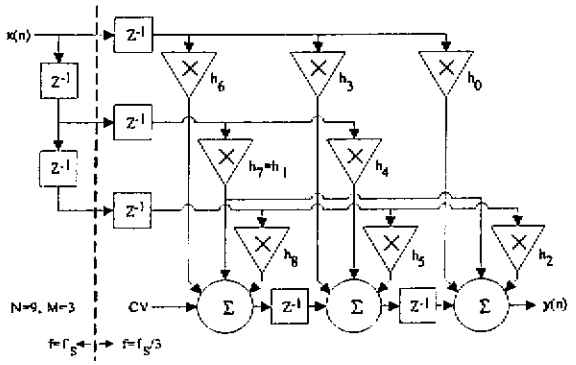


Fig. 7 Transposed direct form decimator with memory -saving technique.

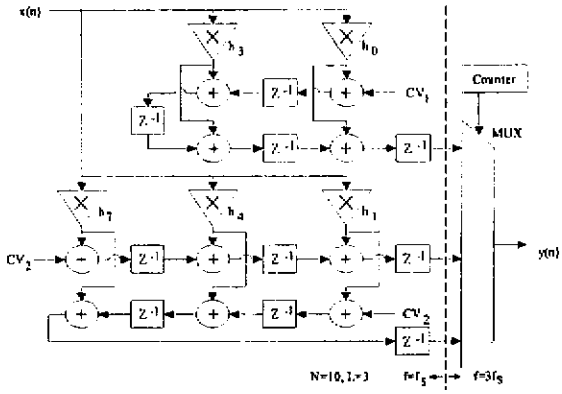


Fig. 8 Transposed direct form interpolator with mirror symmetric filter pairs.

b. FIR filter

The module generator generates three types of the symmetric transposed direct form for each stage FIR filters. Structure A: The transposed direct form filter structure is adopted and written in behavior level synthesizable Verilog-HDL code, which allows the synthesis tool to select the appropriate architecture for user's constraints. Structure B (Fig. 9(a)): The transposed direct form filter structure utilizes with carry save adders (CSA). Structure C (Fig. 9(b)): We exploit structure B with pipelining to achieve at most two CSA delay critical path for the maximum allowed number of SPT terms per coefficient is three.

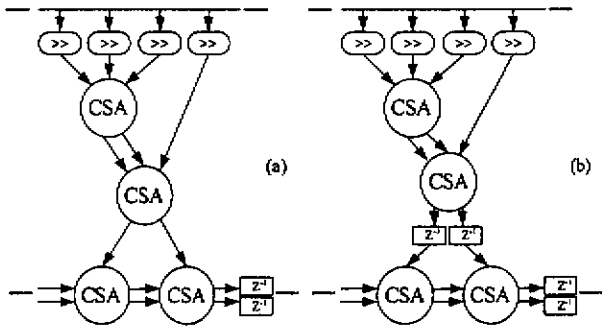


Fig. 9 Transposed direct form filter strcture utilizes with (a) carry save adders (CSA) and, (b) carry save adders (CSA) with pipelining.

III. DESIGN EXAMPLE

A linear-phase low-pass FIR filter is designed using our proposed method, the mixed integer linear programming (MILP) algorithm [6], and Samueli's local search algorithm [5]. The pass-band and stop-band edge frequencies are 0.3π and 0.5π, respectively. The normalized peak ripple (NPR) δ_{NPR}=-50dB. The word length of the input signal is assumed to be 14 bits. The minimum number of SPT terms required by the various methods mentioned above is summarized in Table 2. When the maximum allowed number of SPT terms per coefficient is limited to four, the filter designed by our methods saves 22%(21%~24%) SPT terms and costs 5%(4%~7%) additional tap length. If the application requires

us to limit the maximum number of SPT terms per coefficient to three, for a higher throughput rate, the filter designed using Samueli's algorithm failed to reach -50 dB NPR. However, using our proposed method can save 16% SPT terms and costs 4% additional tap length.

Table 2 Minimum number of SPT terms required to attain -50dB NPR.

Algorithm	#SPT	N
Max. SPT per coeff. = 4		
MILP [6]	68	28
Samueli [5]	66	28
Our Work #1	54	29
Our Work #2	52	30
Max. SPT per coeff. = 3		
MILP [6]	68	28
Samueli [5]	cannot reach -50 dB	
Our Work #3	57	29

A filter design example for 64-QAM baseband demodulator is given [7]. The chip is deigned with TSMC 0.25μm process and shows that the area is reduced by 39% for low-complexity applications. Moreover, for high-speed application, the chip can operate at 714MHz. The synthesis results are summarized in Table 3. The area is measured in equivalents of 2-input NAND gates.

The designed multirate decimators, which specification is the first version of the CDMA cellular proposed by Qualcomme [8] are summarized in Table 4. The frequency responses of the conventional decimator that is single-stage decimator using the polyphase structure to save the arithmetic operations and the multistage multirate decimators designed by our proposed method as shown in Fig. 10. When the timing constraint of the conventional decimator and the multistage multirate decimators are equals, the area of the multistage multirate decimator is reduced by 70% as compared with conventional decimator. If use same specification to design the multistage multirate interpolator that can also saves 70% hardware complexity as compared with conventional decimator.

IV. CONCLUSION

We have implemented a module generator for multiplierless multistage multirate decimator / interpolator. In the architecture design, the module generator uses IFIR (multistage) to reduce the tap number of the filter. Then, multirate and polyphase methods are used to decompose the filter and down / up sampling block into subfilters with different operation rate to save power dissipation and further reduce the hardware complexity. In each subfilters, have also shown that the local search algorithm together with changing filter order can further reduce the number of total nonzero digits. All the analysis, design process and trade-off considerations are coded in the module generator to automate the process. Finally, synthesizable Verilog-HDL code of the design is generated. Design examples show that it can complete the design in a few minutes and have area reduction of about 70% as compared with design using conventional design method.

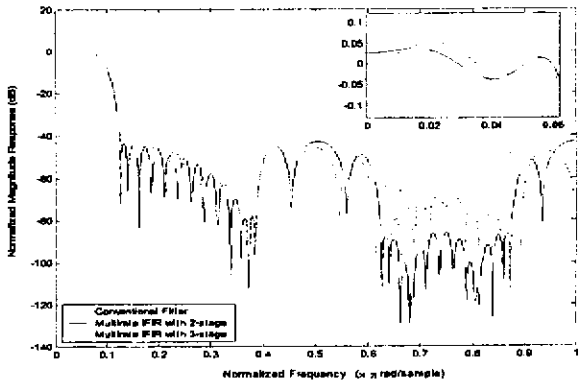


Fig. 10 The frequency responses of the conventional FIR and multirate IFIR filters.

Table 3 Synthesis results of the example for 64-QAM baseband demodulator.

	Work#1	Work#2	[7]
Technology(um)	0.25	0.25	0.25
Max. Operating Frequency	714 MHz	146 MHz	72 MHz
Total Gate Count	11117	5155	8477
Combination Area	5011	2496	5938
Noncombination Area	6106	2659	2539

Work#1 Specifications under high-speed constraint.
Work#2 Specifications under low-complexity constraint.

Table 4 The comparison with the conventional decimator and the multirate decimator [8].(Decimation Ratio, M=8)

Technology: TSMC 0.25um Input Frequency: 200 MHz	Conventional Decimator (M=8)	Multirate Decimator with two-stage	
		Stage#1 (M_1=4)	Stage#2 (M_2=2)
Total Gate Count	13742	1580	2554
Combination Area	12567	1162	1792
Noncombination Area	1174	418	761

Technology: TSMC 0.25um Input Frequency: 200 MHz	Multirate Decimator with three-stage		
	Stage#1 (M_1=2)	Stage#2 (M_2=2)	Stage#3 (M_3=2)
Total Gate Count	638	808	2417
Combination Area	336	496	1706
Noncombination Area	301	312	710

V. REFERENCES

[1] Y. Neuvo, C. Y. Dong, and S. K. Mitra, "Interpolated finite impulse response filters," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-32, pp. 563-570, June 1984.
[2] B.C. Wong and H. Samueli, "A 200-MHz all-digital QAM modulator and demodulator in 1.2μm CMOS for digital radio applications," IEEE JSSC, pp. 1970-1979, Dec. 1991.
[3] M. Bellanger, G. Bonnerot, and M. Coudreuse, "Digital filtering by polyphase network: application to sample rate alteration and filter banks," IEEE Trans. ASSAP, vol. ASSP-24, pp. 109-114, April 1976.
[4] R. A. Hawley, B. C. Wong, T.-J. Lin, J. Laskowski, and H. Samueli, "Design techniques for silicon compiler implementations of high-speed FIR digital filters," IEEE JSSC, vol. 31, pp. 656-667, May 1996.
[5] H. Samueli, "An improved search algorithm for the design of multiplierless FIR filters with powers-of-two coefficients," IEEE Trans. Circuits Syst., pp. 1044-1047, Jul. 1989.
[6] Y. C. Lim and S. R. Parker, "FIR filter design over a discrete powers-of-two coefficient space," IEEE Trans. ASSP, Jun. 1983.
[7] S. J. Jou, C. H. Kuo, M. T. Shiau, J. Y. Heh and C. K. Wang, "VLSI implementation of timing recovery and carrier recovery for QAM/VSB dual mode," International Symp. on VLSI Technology, Systems and Applications, Taipei, R. O. C. June 1999, pp.159-162.
[8] "The CDMA network engineering handbook, volume 1: concepts in CDMA," Qualcomm Inc., March 1993.
[9] K.Y. Jeng, S.J. Jou and A.Y. Wu, "A Design Flow for Multiplierless Linear-Phase FIR Filters: From System Specification to Verilog Code," IEEE Circuits and Systems, 2004, ISCAS '04.

Convolution Based RF Tranceiver Testing Methodology

Hung-kai Chen and Chauchin Su
Department of Electrical and Control Engineering
National Chiao Tung University, Hsinchu 300, Taiwan

Abstract

To derive and implement a methodology to test and debug IF and RF modules using the IF signal generated and received at the mixed signal modules. Low IF test signal is generated by DAC from DSP module and fed to the IF/RF module at transmitting end. At the receiving end, the response waveform is captured by the ADC and transported to DSP for analysis. By this method, the test cost is minimal because the AD/DA converters and DSP modules are built in already. There is no external RF ATE needed.

1. Introduction

After years of development, wireless communication has become essential for telephony and data communication. In mobile communication, there are many different standards and systems in use simultaneously, such as GSM, DECT, PHS, GPRS, WCDMA, and CDMA2000. In data communication, Bluetooth and 802.11 are competing for market acceptance. Therefore, RF circuit design and applications are wide and extensive. Furthermore, in wired communication, fiber and high speed copper wired serial links are all fall in RF frequency range. Due to its high added value, design houses, foundry, and system houses are focus on this sector of semiconductor industry in order to make high profit margin. However, the testing of RF circuits and systems are not well developed as design and manufacturing.

Traditionally, RF circuit and system test are achieved in two ways. The first one is to use specialized instrument to send and receive high quality RF signals to determine the circuit parameters and go-no-go of the circuits. The disadvantage is the cost overhead. The instrument in RF range is not only expensive but also very difficult to operate and maintain. It seems that it is difficult to meet the requirement of low cost mobile phone handset or wireless local area network. The second type uses the loop back mechanism for the test. The digital data is sent by the transmitter and received and recovered by the receivers. It checks the bit-error rate (BER) of the received data to determine the function of the CUT. It has the properties opposite to the first method. The test equipment cost is low. But, the test time is very long.

2. RF Test Methodology

We would like to propose an RF test methodology which takes the advantage of the RF system architectures nowadays. We utile the built-in analog/digital (AD/DA) converters and digital signal processors (DSP) as the test resources for the test.

2.1 RF Test Architecture

The DSP is responsible for sending digital IF signal to the DAC. Then DAC coverts the discrete signal into analog signal. After that, it is up converted into RF band by the RF transmitter module. The RF receiver module receives the RF signal and down converts it into IF band. The ADC converts it into digital form. Finally, the DSP collects the received digital signal and use DSP techniques to extract the circuit parameters. The DSP techniques will be based on the "intrinsic response extraction". The simple architecture of wireless system is shown in Figure 1.

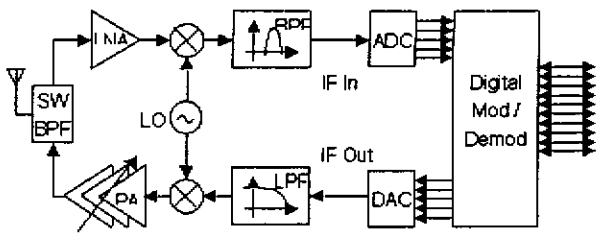


Figure 1. Wireless RF/IF and mixed signal circuit

2.2 RF Test Method

We use convolution technique to develop our method. When convolution in time domain,

$$y(t) = x(t)*h(t)$$

It takes multiplication in frequency domain

$$Y(s) = X(s)\,H(s)$$

$$H(s) = Y(s) / X(s)$$

$$H(s)_{db} = Y(s)_{db} - X(s)_{db} \text{ (log scale)}$$

The RF block in wireless transmission system is shown in Figure 2. Use build-in DAC in the transmitter to generate x(t), and receive y(t) by ADC in the receiver. Then use FFT to convert time domain signal, x(t) and y(t), to frequency domain, X(s) and Y(s). Finally we can easily find system transfer function H(s) = Y(s) / X(s).

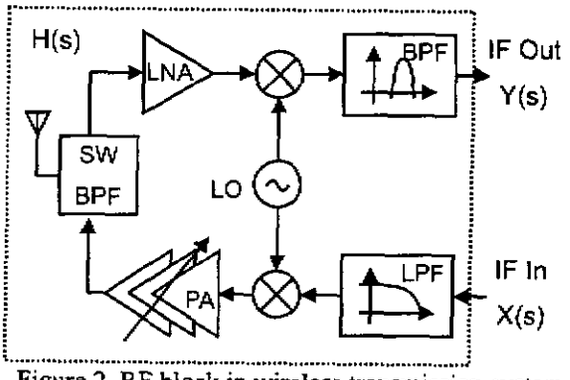


Figure 2. RF block in wireless transmission system

3. Emulation and Measurement Results

To verify the proposed methodology, use EDA Tools to simulate and measurement by bench equipment like a waveform generator and a spectrum analyzer.

3.1 Filter Simulation and Measurement

We have use discrete components to build a simple 1st order filter environment. The circuit diagram is as the one shown in Figure 3.

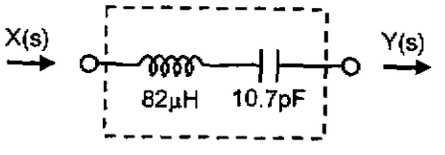


Figure 3.1st-order filter, H(s)

Use EDA-tool to simulate this filter, the simulation result is shown in Figure 4.

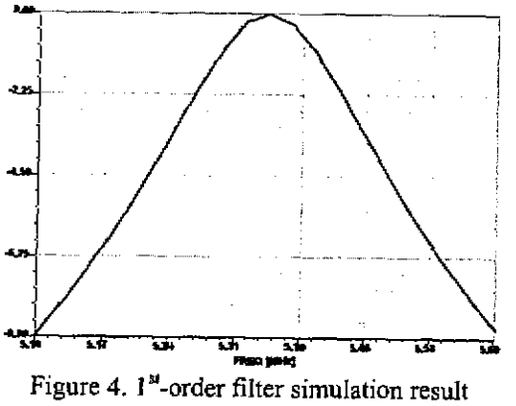


Figure 4. 1st-order filter simulation result

Generate the input signal x(t) by using Agilent 33250A Function / Arbitrary Waveform Generator. The test signal is sinc function which bandwidth is 500 kHz and center frequency is 5.45 MHz. Then feed test signal to the emulation circuit, 1st-order filter. Finally, measure input Signal, X(s), and filter output, Y(s), by signal analyzer. The measurement result is shown in Figure 5.

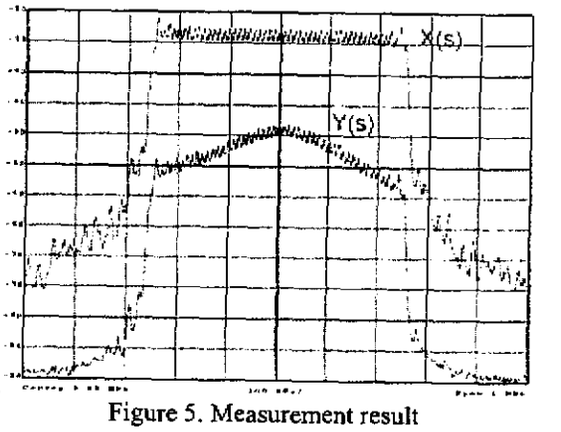


Figure 5. Measurement result

Use Matlab to calculate H(s) = Y(s) – X(s), H(s) is system response shown in Figure 6.

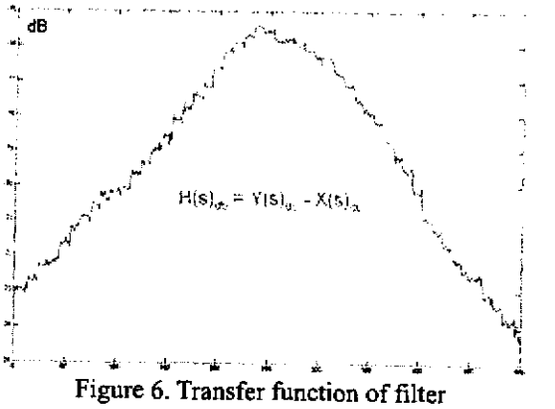


Figure 6. Transfer function of filter

Also take measurement by HP Network Analyzer. Compare two methods in spectrum.

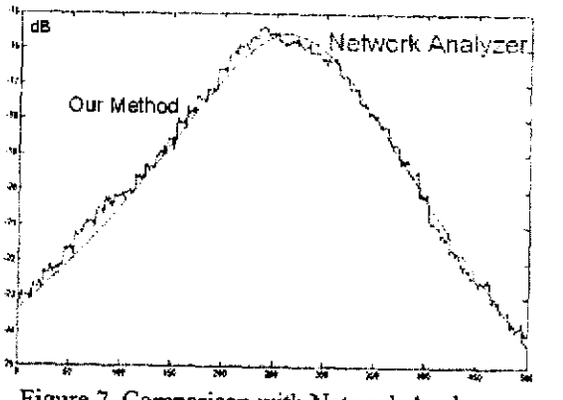


Figure 7. Comparison with Network Analyzer

The difference between our method and a network analyzer measurement result is small. It is reduce the cost of network analyzer and extra time for calibrate the instrument.

3.2 RF Front-end Simulation

This section we build whole RF front-end system from the DAC generates IF a signal to the ADC receives IF signals. And use Agilent ADS tool to simulate our circuits. The schematic is shown in Figure 8. The voltage source, V_{in}, generate desire IF signals and through a filter, a up-converter mixer, and a power amplifier to the channel. Then RF signal fed in a low noise amplifier, a down-convert mixer, and a filter to the IF output, V_{out}. The simulation for ideal components is shown in Figure 9. It cans analysis results at each node.

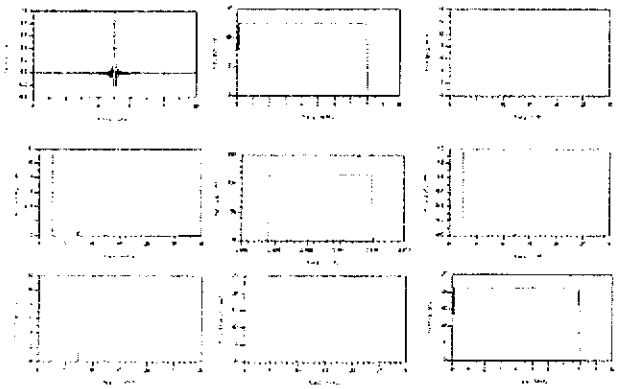


Figure 9. Ideal RF Front-End simulation results

With compare V_{in} and V_{out}, can determinate which component and which specification has errors or degrade. The ideal mixer spectrum is shown in Figure 10 and the mixer with 1 dBm third interception point shown in Figure 11.

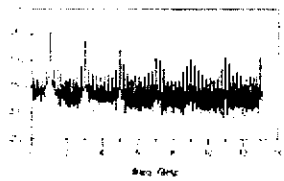


Figure 10. Ideal Mixer

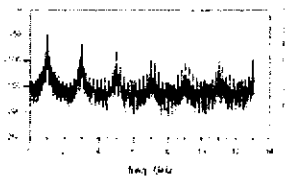


Figure 11. IP3=1 dBm

4. Conclusions

In this paper, we focus on the IF and RF testing. Low IF test signal is generated by DAC from DSP module and fed to the IF/RF module at transmitting end. At the receiving end, the response waveform is captured by the ADC and transported to DSP for analysis. The DSP is used as the engine for test generation and response analysis digitally.

To derive and implement a methodology to test and debug IF and RF modules using the IF signal generated and received at the mixed signal modules. The sinusoidal single tone or multi-tone signal in conjunction with Fourier transform can be used to test amplifiers, mixers, and oscillators. While, sinc function based test waveforms can be used to test filters effectively.

By this method, the test cost is minimal because the AD/DA converters and DSP modules are built in already. There is no external RF ATE needed.

References

[1] Lupea, D.; Pursche, U.; Jentschel, H.-J., "RF-BIST: loopback spectral signature analysis," Design, Automation and Test in Europe Conference and Exhibition, 2003. pp. 478 – 483

[2] Heutmaker, M.S.; Le, D.K., „An architecture for self-test of a wireless communication system using sampled IQ modulation and boundary scan," Communications Magazine, IEEE , Volume: 37 , Issue: 6 , 1999 pp. 98 – 102

[3] Voorakaranam, R.; Cherubal, S.; Chatterjee, A., "A signature test framework for rapid production testing of RF circuits," Proc. Design, Automation and Test in Europe Conference and Exhibition, 2002. pp. 186 - 191

[4] Ferrario, J.; Wolf, R.; Moss, S., "Architecting millisecond test solutions for wireless phone RFICs, Proc. Test Conference, 2002. pp. 1151 - 1158"

[5] Kasten, J.S.; Kaminska, B., "An introduction to RF testing: device, method and system," Proc. VLSI Test Symposium, 1998. pp. 462 - 468

[6] IEEE Std. 1057-1994, *IEEE Standard for Digitizing Waform Recorders*, IEEE 1994.

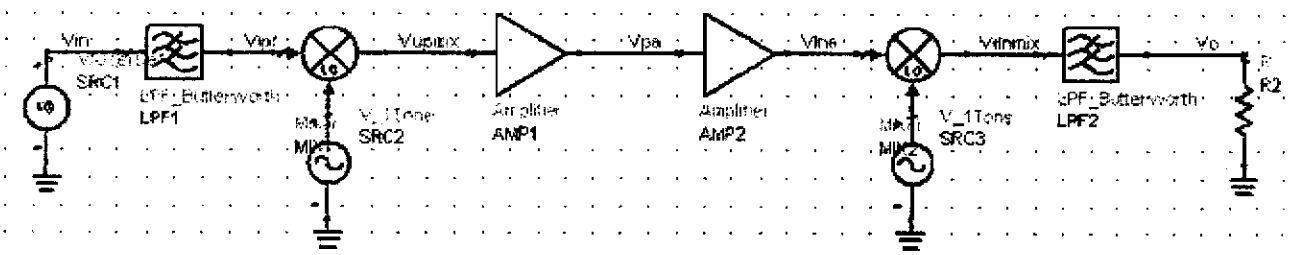


Figure 8. Schematic of RF Front-end

CONVOLUTION BASED RF TRANSCEIVER TESTING METHODOLOGY

Hung-kai Chen and Chauchin Su

Department of Electrical and Control Engineering
National Chiao Tung University, Hsinchu 300, Taiwan

ABSTRACT

To derive and implement a methodology to test and debug IF and RF modules using the IF signal generated and received at the mixed signal modules. Low IF test signal is generated by DAC from DSP module and fed to the IF/RF module at transmitting end. At the receiving end, the response waveform is captured by the ADC and transported to DSP for analysis. By this method, the test cost is minimal because the AD/DA converters and DSP modules are built in already. There is no external RF ATE needed.

1. INTRODUCTION

After years of development, wireless communication has become essential for telephony and data communication. In mobile communication, there are many different standards and systems in use simultaneously, such as GSM, DECT, PHS, GPRS, WCDMA, and CDMA2000. In data communication, Bluetooth and 802.11 are competing for market acceptance. Therefore, RF circuit design and applications are wide and extensive. Furthermore, in wired communication, fiber and high speed copper wired serial links are all fall in RF frequency range. Due to its high added value, design houses, foundry, and system houses are focus on this sector of semiconductor industry in order to make high profit margin. However, the testing of RF circuits and systems are not well developed as design and manufacturing.

Traditionally, RF circuit and system test are achieved in two ways. The first one is to use specialized instrument to send and receive high quality RF signals to determine the circuit parameters [3, 4, 5] and go-no-go of the circuits. The disadvantage is the cost overhead. The instrument in RF range is not only expensive but also very difficult to operate and maintain. It seems that it is difficult to meet the requirement of low cost mobile phone handset or wireless local area network. The second type uses the loop back mechanism [1, 2] for the test. The digital data is sent by the transmitter and received and recovered by the receivers. It checks the bit-error rate (BER) of the received data to determine the function of the CUT. It has the properties opposite to the first method. The test equipment cost is low. But, the test time is very long..

2. RF TEST METHODOLOGY

We would like to propose an RF test methodology which takes the advantage of the RF system architectures nowadays. We utile the built-in analog/digital (AD/DA) converters and digital signal processors (DSP) as the test resources for the test.

2.1 RF Test Architecture

The DSP is responsible for sending digital IF signal to the DAC. Then DAC coverts the discrete signal into analog signal. After that, it is up converted into RF band by the RF transmitter module. The RF receiver module receives the RF signal and down converts it into IF band. The ADC converts it into digital form. Finally, the DSP collects the received digital signal and use DSP techniques to extract the circuit parameters. The DSP techniques will be based on the "intrinsic response extraction". The simple architecture of wireless system is shown in Figure 1.

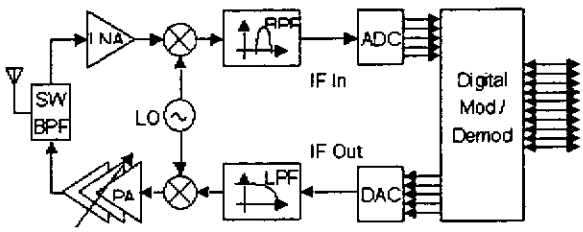


Figure 1. Wireless RF/IF and mixed signal circuit

2.2 RF Test Method

We use convolution technique to develop our method. When convolution in time domain,

$y(t) = x(t)*h(t)$

It takes multiplication in frequency domain

$Y(s) = X(s)\,H(s)$

$H(s) = Y(s) / X(s)$

$H(s)_{db} = Y(s)_{db} - X(s)_{db}$ (log scale)

The RF block in wireless transmission system is shown in Figure 2.

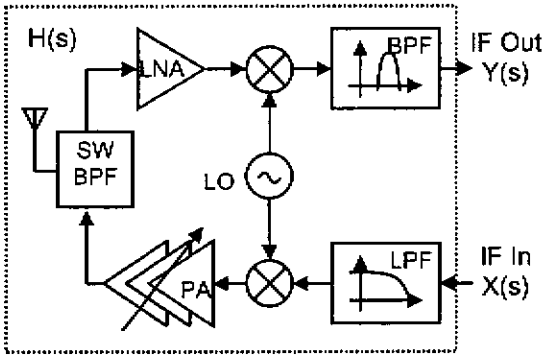


Figure 2. RF block in wireless transmission system

Use build-in DAC in the transmitter to generate x(t), and receive y(t) by ADC in the receiver. Then use FFT to convert time domain signal, x(t) and y(t), to frequency domain, X(s) and Y(s). Finally we can easily find system transfer function $H(s) = Y(s) / X(s)$.

There are many king signals can be used in our testing method. The single tone sinusoidal waveform is easy to test some frequency response. The two tones test can test harmonics and intermodulation distortion. The sinc function is a band limited signal. It has a flat spectrum in some band. However, the power density of sinc waveform is too small, so we can use multitone signal to test our circuits more efficiency.

3. EMULATION AND MEASUREMENT RESULTS

To verify the proposed methodology, use EDA Tools to simulate and measurement by bench equipment like a waveform generator and a spectrum analyzer.

3.1 Filter Simulation and Measurement

We have use discrete components to build a simple 1st order filter environment. The circuit diagram is as the one shown in Figure 3.

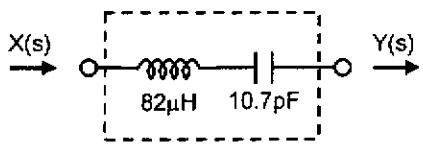


Figure 3. 1st-order filter, H(s)

Use EDA-tool to simulate this filter, the simulation result is shown in Figure 4.

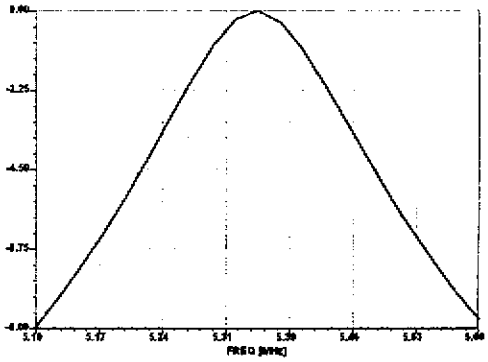


Figure 4. 1st-order filter simulation result

Generate the input signal x(t) by using Agilent 33250A Function / Arbitrary Waveform Generator. The test signal is sinc function which bandwidth is 500 kHz and center frequency is 5.45 MHz. Then feed test signal to the emulation circuit, 1st-order filter. Finally, measure input Signal, X(s), and filter output, Y(s), by signal analyzer. The measurement result is shown in Figure 5.

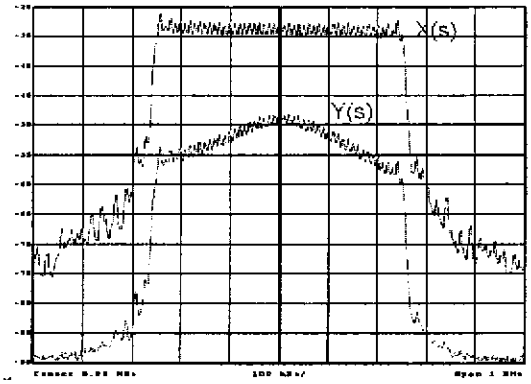


Figure 5. Measurement result

Use Matlab to calculate $H(s) = Y(s) - X(s)$, H(s) is system response shown in Figure 6.

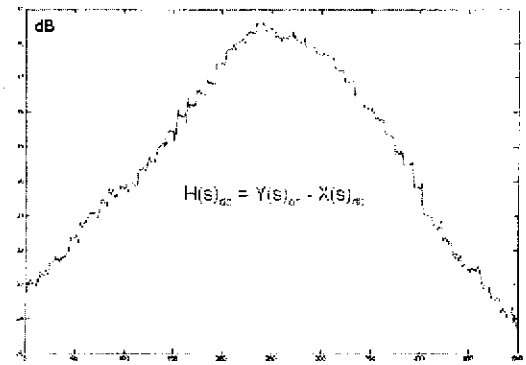


Figure 6. Transfer function of filter

Also take measurement by the HP Network Analyzer. And compare two methods in the frequency domain shown in Figure 7.

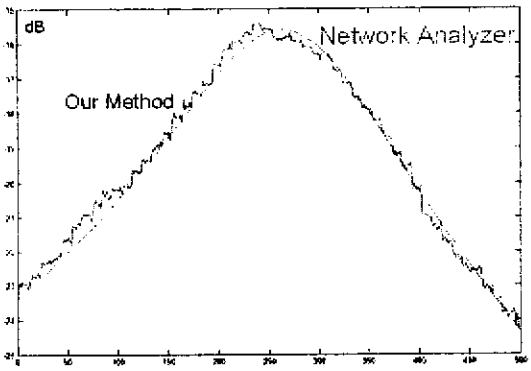


Figure 7. Comparison with a Network Analyzer

The difference between our method and a network analyzer measurement result is small. It is reduce the cost of a network analyzer and extra time for calibrate the instrument.

3.2 RF Front-end Simulation

This section we build whole RF front-end system from the DAC generates IF a signal to the ADC receives IF signals. And use Agilent ADS tool to simulate our circuits. The schematic is shown in Figure 8. The voltage source, V_{in}, generate desire IF signals and through a filter, a up-converter mixer, and a power amplifier to the channel. Then RF signal fed in a low noise amplifier, a down-convert mixer, and a filter to the IF output, V_{out}. The simulation for ideal components is shown in Figure 9. It cans analysis results at each node.

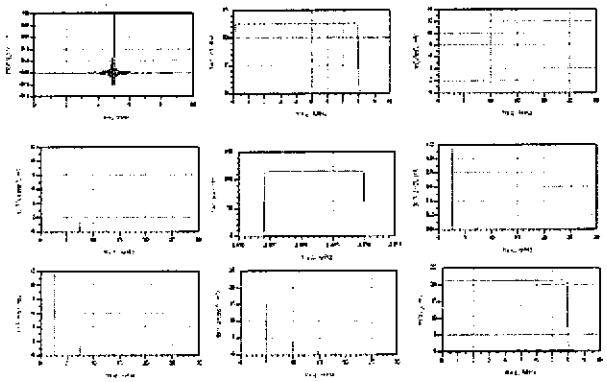


Figure 9. Ideal RF Front-End simulation results

With compare V_{in} and V_{out}, can determinate which component and which specification has errors or degrade. The ideal mixer spectrum is shown in Figure 10 and the mixer with 1 dBm third interception point shown in Figure 11.

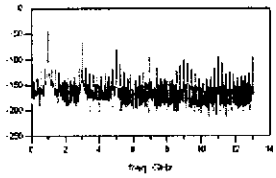


Figure 10. Ideal Mixer

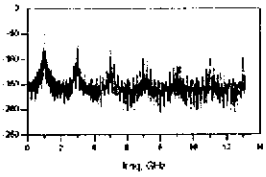


Figure 11. IP3=1 dBm

There are two or more amplifiers in the RF front-end. In transmitter end, the power amplifier is key component to amplify signal to antenna. In receiver end, the small signal is amplified by low noise amplifier. The noise figure and gain is the important specifications for amplifier.

In order to correlate our methodology, build the ideal system, behavior model, and real model, circuit model. There are two power amplifiers, an ideal one and non-ideal model shown in Figure 12. Can use this architecture to test and compare each other.

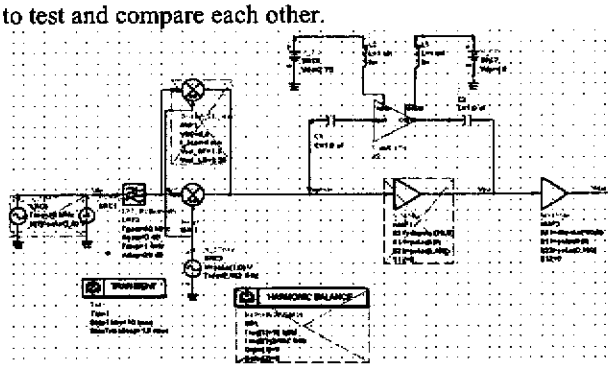


Figure 12. Build an ideal model vs. real circuits

Generate testing signal by two methods. One is AC sweep by EDA Tools, another is sinc function generated by voltage source in transient analysis then convert to frequency domain. The schematic of testing a power amplifier is shown in Figure 13.

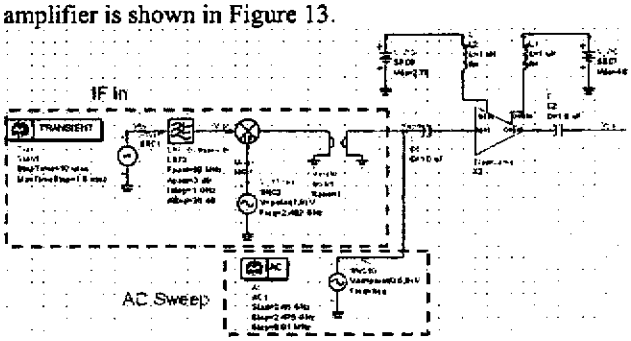


Figure 13. The schematic of testing a power amplifier

Test some specifications of a non-ideal power amplifier in a wireless LAN transceiver. First, generate the sinc function with 8 MHz bandwidth and up-converter to 2.462 GHz by using an ideal mixer. Second, measure the output of the power amplifier. Finally, calculate the

gain of power amplifier by subtracting the output from the input.

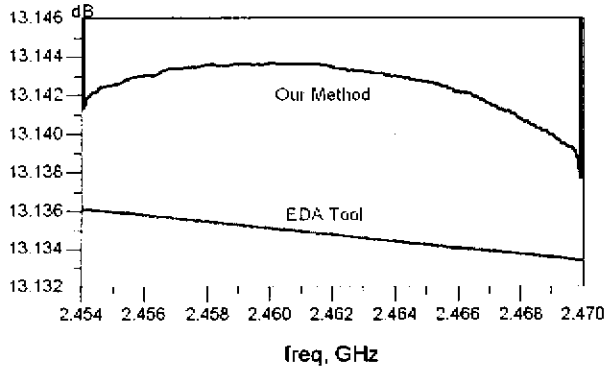


Figure 14. The gain of a power amplifier

Compare with the simulation result of frequency sweep in an EDA Tool is shown in Figure 14. The error between two methods is very small which is about 0.06 dB. It is caused by precision of numerical transformation.

Also can generate multitone signals to test circuits is shown in Figure 15. Generate the large number of tones at the corner frequency or some interested frequency. And in flat band, it is not necessary to generate many tones. This way can increase the power efficiency and gain the larger signal to noise ratio.

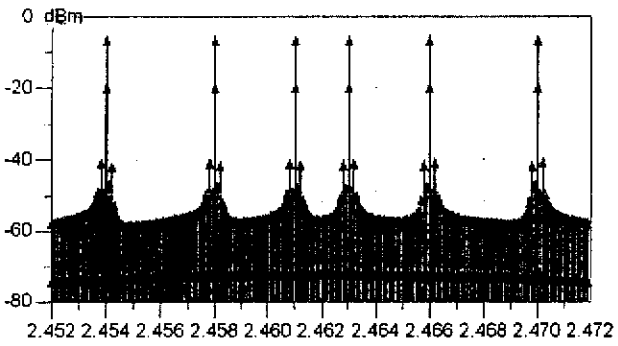


Figure 15. The multitone test of a power amplifier

4. CONCLUSIONS

In this paper, we focus on the IF and RF testing. Low IF test signal is generated by DAC from DSP module and fed to the IF/RF module at transmitting end. At the receiving end, the response waveform is captured by the ADC and transported to DSP for analysis. The DSP is used as the engine for test generation and response analysis digitally.

To derive and implement a methodology to test and debug IF and RF modules using the IF signal generated and received at the mixed signal modules. The sinusoidal single tone or multi-tone signal in conjunction with Fourier transform can be used to test amplifiers, mixers, and oscillators. While, sinc function based test waveforms can be used to test filters effectively.

By this method, the test cost is minimal because the AD/DA converters and DSP modules are built in already. There is no external RF ATE needed.

5. REFERENCES

[1] Lupea, D.; Pursche, U.; Jentschel, H.-J., "RF-BIST: loopback spectral signature analysis," Design, Automation and Test in Europe Conference and Exhibition, 2003. pp. 478 – 483

[2] Heutmaker, M.S.; Le, D.K., „An architecture for self-test of a wireless communication system using sampled IQ modulation and boundary scan," Communications Magazine, IEEE , Volume: 37 , Issue: 6 , 1999 pp. 98 – 102

[3] Voorakaranam, R.; Cherubal, S.; Chatterjee, A., "A signature test framework for rapid production testing of RF circuits," Proc. Design, Automation and Test in Europe Conference and Exhibition, 2002. pp. 186 – 191

[4] Ferrario, J.; Wolf, R.; Moss, S., "Architecting millisecond test solutions for wireless phone RFICs, Proc. Test Conference, 2002. pp. 1151 - 1158

[5] Kasten, J.S.; Kaminska, B., "An introduction to RF testing: device, method and system," Proc. VLSI Test Symposium, 1998. pp. 462 – 468.

[6] IEEE Std. 1057-1994, IEEE Standard for Digitizing Waveform Recorders, IEEE 1994.

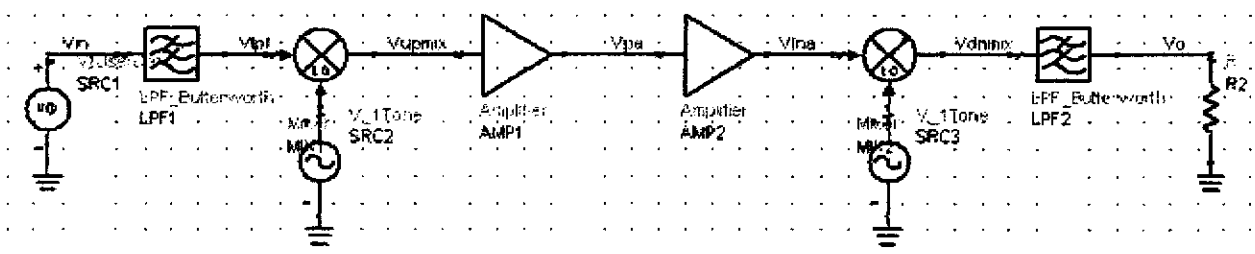


Figure 8. Schematic of RF Front-end

Fair Scheduling with QoS Support in Ad Hoc Wireless Networks

Hsi-Lu Chao, *Student Member, IEEE*, and Wanjiun Liao, *Member, IEEE*

***Abstract*—In this paper, we study fair scheduling with Quality of Service (QoS) support for wireless ad hoc networks. two types of flows are considered: best effort and guaranteed flows. The goal is to satisfy the minimum bandwidth requirements of guaranteed flows and to provide fair share of residual bandwidth to all flows. We compare two types of mechanisms: timestamp-based and credit-based mechanisms, and evaluate the feasibility of all existing timestamp-based and credit-based fair scheduling schemes for multimedia wireless multihop networks. We suggest a flow weight calculation scheme for existing timestamp-based mechanisms to support both best-effort and guaranteed flows, and propose a credit-based mechanism called Credit-based Slot Allocation Protocol (CSAP). For comparison purposes, several metrics are defined to evaluate the performances of these two kinds of mechanisms. The simulation results show that CSAP outperforms the other approaches in terms of meeting the minimum requirements of guaranteed flows, fairly sharing the residual bandwidth among all flows, and improving overall system throughput.**

***Index Terms*—QoS, fair scheduling, ad hoc networks**

I. Introduction

An ad hoc network is a self-organizing wireless network comprised only of mobile nodes without the support of any pre-existing wired infrastructure. In such a network, each node plays both roles of a terminal and a router. Due to the fully distributed characteristic of the network, collisions may occur and the system throughput may be low without proper coordination among the transmitting nodes.

Providing Quality of Service (QoS) in ad hoc networks is a new area of research. Existing work focuses mainly on QoS routing [1,2], which finds a path to meet the desired service levels of flows. This paper studies a different QoS issue in multihop, multimedia wireless networks. We consider a mix of guaranteed and best-effort flows and investigate fair scheduling with QoS support for the network. The goal is to guarantee the minimum bandwidth requirements of guaranteed flows, and ensure fair share of residual bandwidth to all flows. In the following, we first define "fairness" in the context of ad hoc networks, and then examine related issues of fair scheduling with QoS support for multimedia ad hoc networks.

Manuscript received August 8, 2002; revised Feb 2003. This work was supported in part by MOE program for Promoting Academic Excellence of Universities under grant number 89E-FA06-2-4-7, and in part by the National Science Council, Taiwan, under grant number NSC91-2213-E-002-057.

Both authors are with the Department of Electrical Engineering and Graduate Institute of Communication Engineering, National Taiwan University, Taipei 106, Taiwan.

A. Fairness

Fairness is an important criterion of resource sharing in the best-effort Internet, especially when there is competition for their share due to unsatisfied demands [3]. Fairness is defined as follows [4,5,6]. Each flow, say f, is allowed to share a certain percentage of link capacity based on its flow weight indicated as w_f. Let $W_f(t_1,t_2)$ denote the aggregate resource received by flow f in the time interval $[t_1,t_2]$. The allocation is called ideal fair if (1) is satisfied for all intervals $[t_1,t_2]$ in which both flows f and g are backlogged.

$$\left|\frac{W_f(t_1,t_2)}{w_f}-\frac{W_g(t_1,t_2)}{w_g}\right|=0 \qquad (1)$$

Scheduling is an effective way to ensure fair sharing of resource among flows [7]. It determines the service order of packets according to such metrics as packet timestamp, delay bound, and cumulative credit. Of these metrics, timestamp is the most common one used by existing scheduling schemes [8,9,10,11]. For these timestamp-based scheduling protocols, the system maintains a virtual time as in [4]. Each packet is associated with two timestamps, i.e., start tag and finish tag, based on the system-defined virtual time. Let S_f^j and F_f^j denote the start tag and the finish tag for the j^{th} packet of flow f, respectively.

$$\begin{cases} S_f^j = max\left\{ v\left[A\left(p_f^j\right)\right],\ F_f^{j-1}\right\},\ j \ge 1 \\ F_f^j = S_f^j + \dfrac{l_f^j}{w_f}, j \ge 1 \end{cases}, \qquad (2)$$

where p_f^j, l_f^j and $A(p_f^j)$ are the j^{th} packet of flow f, the length of p_f^j, and the arrival time of p_f^j, respectively. $v[A(p_f^j)]$ denotes the virtual arrival time of p_f^j.

The main objective of existing fair scheduling approaches is to ensure E, defined in (3), as close to zero as possible. The closer E is to zero, the fairer the approach.

$$E=\left|\frac{W_f(t_1,t_2)}{w_f}-\frac{W_g(t_1,t_2)}{w_g}\right| \qquad (3)$$

B. Design Issues of Fair Scheduling in Ad Hoc Networks

Due to the characteristics of wireless ad hoc networks, three challenges must be addressed when a fair scheduling protocol is designed [9].

1) Location-dependent contention: Wireless transmission is broadcast locally to the network. Uncoordinated simultaneous transmissions may result in packet collisions. Those flows that incur packet collisions when transmitted simultaneously are called "*contending flows.*" For example, in Fig. 1, there are nine nodes (denoted N1, N2, ..., N9) and four flows (denoted F1, F2, F3, F4). The line connecting two nodes indicates each is within the same transmission range of the other. Flows F1 and F2 cannot be transmitted simultaneously, or a collision occurs at N2. Node N4 is interfered by flow F3, because node N4 is also within the transmission range of node N7. Likewise, nodes N3 and N5 are interfered by flow F4.

2) Spatial channel reuse: Multiple flows can be transmitted simultaneously without collisions if these flows do not interfere with each other. This is called "spatial channel reuse." For example, flows F2 and F3 in Fig. 1 can be transmitted simultaneously. Spatial channel reuse is an important technique to improve system throughput. A good design of fair scheduling should provide a simple and cost-effective way of finding the maximum number of flows that can be transmitted simultaneously.

3) Conflict between global fairness and maximal channel utilization for wireless systems: Existing work on fair scheduling for ad hoc networks focuses on global fairness, because they assume all flows are in a single cluster. Typically, a maximum window size is defined to restrict the maximum bandwidth each flow can use in a time interval. In reality, wireless networks, such as IEEE 802.11 and Bluetooth, are cluster-based and make use of channel reuse to maximize system throughput. A good mechanism should balance the trade-off between fairness and maximum channel reuse when fair scheduling is provided.

C. Related Work

1) Fair Scheduling in Ad Hoc Networks: Most existing works for fair scheduling in ad hoc networks are timestamp-based [8,9,10]. For timestamp-based protocols, each newly arrived packet is assigned two timestamps, i.e., start and finish tags, based on (2). In [10], both tags are used to calculate a backoff value. The backoff value determines when the packet will be sent. It works as follows. To send a packet, a node monitors the channel until it is free. The node then calculates a backoff value based on the two tags assigned to the packet, and decrements the backoff timer by one at each time slot until it reaches zero. If the node with a zero timer finds the channel is free, the packet is transmitted. Nodes with zero-timer packets do not coordinate before the packets are transmitted. Thus, collisions may occur.

In [8,9], a flow graph must first be generated. Each vertex in the flow graph represents a flow. When two flows contend for resource, an edge is added between them. Depending on the scheduling discipline, one of the two tags is chosen as the service tag. The packet with the smallest service tag is sent first. The mechanisms in [8] and [9] differ in how they achieve spatial channel reuse. In [8], the graph coloring theory is used to mark the contending flows of a channel with different colors. The flows with the same color can be transmitted simultaneously. In [9], each packet of a flow is associated with a backoff value. This backoff value is equal to the number of flows contending for a channel but with smaller service tags. Once this backoff value reaches zero, the packet is transmitted.

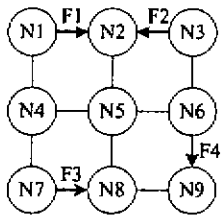


Figure 1. Contending flows

The timestamp-based algorithms suffer two problems. One is the need to sort all packets in the queue in ascending order of their service tags. The other is that the virtual clock cannot be reinitialized with zero unless the system is empty [6].

In [12], a credit-based fair scheduling algorithm is proposed for ad hoc networks. Each flow simply maintains a counter to record the transmission credit, instead of using two tags as in timestamp-based mechanisms. The basic scheduling concept is "the less excess in usage value, the higher the transmission priority." This approach makes use of clustering to implement spatial channel reuse.

2) Fair Scheduling with QoS Support : The mechanisms in [8] and [9] are two major timestamp-based solutions to provide fair scheduling for best effort flows in ad hoc networks. To extend a timestamp-based mechanism to support QoS, an intuitive approach is to assign each flow with a different flow weight according to the respective service requirement. The higher the priority, the larger the flow weight. However, both mechanisms in [8] and [9] do not describe how flow weights are assigned. In the following, we extend [9] to enable QoS support by a distributed flow weight assignment scheme, and call the enhanced version of [9] as Q_EMLM-FQ in the rest of the paper.

The Bottleneck Considered Fair Scheduling (BCFS) mechanism [11] is extended from [8]. It assumes that a flow weight assignment scheme is given, and introduces a new parameter called "contending power" to provide QoS service on top of the fair scheduling mechanism described in [8].

The mechanism in [12] is a credit-based mechanism for fair scheduling in ad hoc networks. In Sec. II, we extend [12] to provide QoS, and call the enhanced version of [12] as CSAP in the rest of the paper. We then compare the performance of BCFS, Q_EMLM-FQ, and CSAP via simulation, and discuss the pros and cons of these three approaches.

a) Q_EMLM-FQ: Suppose that each QoS flow will specify its minimum bandwidth requirement to ensure certain quality.

For a flow, say f, passing through node N, the flow weight w_f is defined as

$$w_f = Resv_f + \frac{1 - \sum_{i \in B} Resv_i}{Num}, \quad (4)$$

where $Resv_f$ and Num are the minimum bandwidth requirement of flow f and the total number of flows passing through node N, respectively; B is the set of backlogged flows. If flow f is a best effort flow, its $Resv_f$ value is zero. As such, guaranteed flows are assured to have flow weights larger than best-effort flows. Consequently, the service tags of guaranteed flows increase more slowly than best-effort ones, allowing their service requirements to be satisfied.

b) BCFS: To support QoS service, a new parameter called "contending power" is defined in [11] for each flow as the scheduling metric. The contending power of a flow, say f, is defined as follows. For each fully-connected graph in which the flow is involved, the flow weights of all the flows are first summarized. The contending power of flow f (denoted as $P(f)$) is equal to the maximum over the summation of the flow weights in each graph, say G_f, involved by flow f, i.e., $P(f) = \max_{\forall G_f} \left\{ \sum_{\forall f_i \in G_f} w_{f_i} \right\}$, where f_i denotes a flow in graph G_f.

Fig. 2 illustrates how the contending power of a flow is calculated. Each vertex indicates a flow. (Fx, y) denotes the flow ID and its predefined flow weight, respectively. F1 is involved in two fully connected subgraphs, i.e., (F1, F2, F3) and (F1, F4, F5). Thus, F1 contends with F2 and F3, and also with F4 and F5. The total flow weights of these two subgraphs are 2+2+3=7 and 2+4+4=10, respectively. Therefore, the contending power of F1 is equal to max {7,10}=10.

The scheduling rules of fair scheduling with QoS support in [11] are described as follows.

- The flow with the highest contending power has the priority to transmit packets.
- If there are multiple flows with the same contending power, the one with the least service tag has the priority to transmit.
- BCFS adopts the same graph-coloring approach as [8] to implement spatial channel reuse. Each vertex is marked a different color from its direct neighbors. The set of flows in the same color can be transmitted simultaneously. The set of flows which has the maximum number of flows in the same color is chosen to transmit simultaneously with the flow selected based on rules (1) or (2).
- If there are several sets with the same maximum number of flows which can be transmitted simultaneously, the set with the highest aggregate contending power will be selected.
- BCFS sets a parameter called "*scheduling window*" to avoid starving flows which have low contending powers. A flow must stop transmission temporarily when a "scheduling window" worth of packets has been transmitted in a pre-defined time interval. In other words, based on a sliding window, of size η, BCFS allows a flow to send up to η packets in a pre-defined time interval.

BCFS suffers from the following problems:

- Flow weight assignment: as in [9], BCFS assigns each flow a positive integer as the flow weight, which indicates how the flow shares the resource. For example, if flows 0, 1, and 2 are assigned flow weights of 2, 3, and 1, respectively, the proportion of resource used by these three flows is 2:3:1. However, there is no word in [11] to mention how flow weights are assigned and learned, and to discuss the impact of different flow weight assignments on the performance of BCFS.
- The impact of flow topology: In BCFS, the contending power of each flow is calculated for scheduling. The flow with the highest contending power has the priority to transmit. However, a flow with a higher contending power does not imply a higher flow weight. Recall that the flow weight is used to indicate how the flow shares system resource. The higher the flow weight, the more time slots the flow can use. For BCFS, the contending power of a flow is determined based on its neighbors in the flow graph. Both the node degree and neighbor flows' weights affect the contending power of the flow. In addition, the spatial reuse mechanism adopted may further reverse the service priority. As a result, a flow with a lower weight may be allocated more slots than one with a higher flow weight. Fig. 3 shows an example to explain why the performance of BCFS is affected by flow topology. Each vertex represents a flow. Let x, y, and z in x(y, z) denote the flow ID, flow weight, and contending power, respectively. In this example, flow 5 has both the highest flow weight and the highest contending power. Thus, it has the highest priority to transmit packets. However, flow 4, which has the second largest flow weight, may not have the second highest priority to transmit packets. The reason is as follows. Considering spatial channel reuse, flow 5 can be transmitted simultaneously with some other flows. For example, it can be transmitted with flows 1, 6, and 7, or with flows 1 and 4. Flow 5 will choose the first set (i.e., flows 1, 6, 7) because the first set has more flows (i.e., 1, 6, and 7) to be transmitted simultaneously. As a result, flow 4 may be allocated fewer slots than flows 1 or 7, even though flow 4 has a higher weight than either flows. The example above shows that the flow topology significantly affects the performance of BCFS. In [11], however, there is no discussion about this issue.
- A centralized scheduling mechanism: In [11], a globally centralized coordinator is assumed available to schedule packet transmissions, but there is no description of how such centralized mechanism is implemented in a fully distributed ad hoc network.

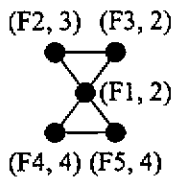


Figure 2. An example of contending power

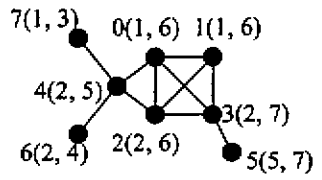


Figure 3. An illustrative example for BCFS

D. Problem Specification for CSAP

In this paper, we propose a new mechanism called the credit-based slot allocation protocol (CSAP), which provides fair scheduling with QoS support for multimedia ad hoc networks. We consider a mix of best effort packets and guaranteed flows, and use "bandwidth" as the QoS metric due to its importance for real-time applications. In [6], bandwidth requirement in time-slotted network is measured in terms of fraction of a time slot per unit time, where unit time is a time slot. As in [6], we use a fraction of one slot time to denote the bandwidth requirement of each guaranteed flow.

The objectives of CSAP are described as follows:

1) Avoid collisions caused by simultaneous, uncoordinated transmissions, thus improving system throughput.

2) Maximize spatial channel reuse, thus maximizing global bandwidth utilization.

3) Guarantee minimum bandwidth for guaranteed flows, and provide fair share of residual bandwidth among all flows (i.e., both best effort and guaranteed flows). Let F_b and F_g denote the sets of best-effort and guaranteed flows in a cluster, respectively. The objective of CSAP is to make E_1 and E_2 in (5) and (6) as close to zero as possible. Both flows i and j are in F_g, and have minimum bandwidth requirements r_i and r_j, respectively. Flow k is in F_b. $W_i(t_1,t_2)$, $W_j(t_1,t_2)$, and $W_k(t_1,t_2)$ denote the aggregate resource received by flows i, j, and k, respectively, in the time interval $[t_1,t_2]$. $C_i(t_1,t_2)$ is the cumulated credit of flow i in $[t_1,t_2]$. $S_i(t_1,t_2)$ and $S_j(t_1,t_2)$ are the residual bandwidth shared by flows i and j in $[t_1,t_2]$, respectively.

$$E_1 = \frac{W_i(t_1,t_2) - S_i(t_1,t_2)}{r_i} - \frac{W_j(t_1,t_2) - S_j(t_1,t_2)}{r_j} \quad (5)$$

$$E_2 = \left| W_i(t_1,t_2) - C_i(t_1,t_2) - W_k(t_1,t_2) \right| \quad (6)$$

The rest of this paper is organized as follows. Section II describes the credit-based slot allocation protocol for fair scheduling with QoS support. Section III shows the simulation results. Finally, the conclusion is drawn in Section IV.

II. Credit-based Slot Allocation Protocol (CSAP)

This section describes the proposed credit-based slot allocation protocol (CSAP) in detail. The characteristics of CSAP are described as follows.

(1) A two-tier hierarchy is used to avoid uncoordinated transmissions.

(2) A cluster-based mechanism is used to maximize spatial channel reuse.

(3) A credit-based mechanism is used to provide QoS service for guaranteed flows while ensuring fair scheduling.

Note that our mechanism works with different residual bandwidth allocation schemes, including fair scheduling for best effort flows only, for a mix of best effort and guaranteed flows, or for a mix of flows but with an upper limit on the resource share for guaranteed flows. In this paper, we just demonstrate the second case, i.e., for a mix of flows without an upper bound on each flow.

A. Assumptions

1) Wireless media may exhibit time-varying errors, which affect network users in a number of ways, such as fading, interference, link errors, and collisions. In this paper, we only consider the errors caused by collisions.

2) We consider packet-switched multihop wireless networks, but do not consider host mobility as in all related work in ad hoc networks [8,9,11,12].

3) We assume a TDMA-based system on a single channel shared by all hosts. To avoid collisions, CDMA can be overlaid on the top of the TDMA system. We assume a code assignment algorithm is running in the lower layer of our system as in [1].

4) There are always some flows backlogged.

B. Service Registration

1) Cluster-based Mechanism: CSAP is a cluster-based mechanism. The network is logically partitioned into clusters. Each cluster has exactly one scheduler. Clusters may be overlapped, each with a designated code to avoid interferences. CSAP can cooperate with any existing clustering algorithm (e.g., [13,14]) to form clusters. For example, a cluster could be formed as follows. Each mobile node periodically broadcasts beacons to nodes located within its transmission range. Based on the received beacons, mobile nodes learn of their neighbors and related information, such as node ID, node stability, etc., for operation. According to the selected criterion, one node is elected as the scheduler per cluster. Each scheduler periodically advertises itself to its neighbors, from which newly arriving mobile nodes learn where to register. Those nodes that can hear the same scheduler form a logical cluster. Each node can hear several schedulers, but can only be registered with one scheduler at a time.

2) Table Structure: CSAP defines two scheduling tables.

a) Node Allocation Table (NAT): maintained by the scheduler; each entry of the NAT is for a (node, service type) pair and is of six fields: a node ID, a service type, a *Resv*, a *Credit*, a *Usage*, and an *Excess*. For a guaranteed service entry, the value in the *Resv* field indicates the total reservation level required to serve all guaranteed flows for the corresponding node. For a best effort entry, the *Resv* value represents the total number of best effort flows scheduled by the node.

b) Flow Allocation Table (FAT): maintained by the node; each entry of the FAT is for a flow and is of seven fields: a flow ID, a scheduler ID, a service type, a *Resv*, a *Credit*, a *Usage*, and an *Excess*.

3) Path Construction: before a transmission starts, the sender of the flow must first determine a flow path, irrespective of whether it is a best effort or guaranteed (i.e., QoS) flow. Best

effort flows do not place any resource requirement on the nodes along the path. However, for a QoS flow, each node along the path should provide resource for the flow to meet its minimum requirement. CSAP can be easily integrated with any of the existing work on ad hoc routing for path construction (e.g., [1,2,15]). For best effort flows, ad hoc routing is used; for guaranteed flows, QoS routing is required. Below we just simply describe an existing path construction mechanism, and will use it as an example to illustrate how CSAP works in the rest of the paper.

To construct a path, a sender broadcasts an RREQ message to its neighbors. Each node receiving this RREQ will execute the service registration phase described in the next section. If the registration fails, this RREQ message will be discarded. Duplicated RREQ messages are discarded when received. When an RREQ message reaches the destination, the destination will send an RREP message in reply along the reverse path back to the sender. Unlike the RREQ message, the RREP message is delivered via unicast. Upon receiving an RREP message at a node, the reservation is confirmed. This process continues until the RREP message reaches the sender, when the path construction phase terminates.

4) Service Registration: Upon receipt of an RREQ message, a node checks if the following condition holds, namely, the summation of the total reserved usage (denoted as *Resv*) of all flows scheduled by the node, including the newly arriving one, is less than or equal to the target link utilization. If the condition holds, the node further sends a resource Allocation ReQuest (ARQ) to its scheduler to check if the condition holds at the scheduler as well. If the condition fails to hold at the node or the scheduler, the request is denied and the RREQ packet is discarded; otherwise, the request is tentatively accepted. The scheduler creates a temporary entry for the node in the NAT, and responds to the node with a temporary acceptance. The node in turn creates a temporary entry for the flow in the FAT, and then rebroadcasts this RREQ to its neighbors. Note that temporary entries in the scheduling table (i.e., in both FAT and NAT) are not involved in the scheduling process.

Once the corresponding RREP message is received, the node will change the entry in the FAT from temporary to active, and will inform its scheduler to make the entry in the NAT active. Those temporary entries not confirmed by the corresponding RREP within a predefined period of time will be cleared in the scheduling table.

C. *Two-Tier Slot Allocation*

1) Scheduling Parameters: There are four parameters defined in CSAP:

a) Resv: a fraction of one time slot, used to indicate the minimum requirement of a flow. For a guaranteed flow, the *Resv* value is a positive real number between zero and one. For a best effort flow, its *Resv* value is always zero.

b) Credit: the cumulative time slots reserved for a guaranteed flow to meet the QoS demand, or the remaining slot quota for a best effort flow per iteration[1]. For the guaranteed service type, the value of an increase at each time slot is equal to a *Resv* value. For the best effort service type, the *Credit* value is initialized with the total number of best effort flows currently scheduled by the node, and is decreased by one after a slot is assigned. Note that for a best effort flow, the *Credit* field is only valid in the NAT and is always set to zero in FAT.

c) Usage: the cumulative time slots assigned to a request. For a scheduler, it records a *Usage* for each node. For a node, it records a *Usage* for a flow. The *Usage* value can be a zero or any positive integer.

d) Excess: this value is used to determine to whom the next time slot is assigned. The next slot is assigned to the request with the smallest *Excess* value. For guaranteed-service entries, the *Excess* values are equal to "*Usage* minus *Credit*," and are updated at each time slot. For best effort-service entries, the *Excess* values are initialized with zero, and incremented by one at each "update." The *Excess* value of a best effort service entry is only updated when the corresponding *Credit* value is count down to zero. Once the *Excess* value has been updated, the *Credit* value is reset to a *Resv* value, and a new iteration starts.

2) Slot Allocation: The slot allocation is based on a two-tier mechanism. The scheduler assigns time slots based on the first tier slot allocation mechanism. The node, say N, to whom the time slot is assigned then in turn assigns the time slot to a flow based on the second tier slot allocation mechanism.

Suppose that there are m and n entries in NAT and FAT, respectively. The m entries in NAT can be divided into two sets: S_g and S_b, which indicate guaranteed and best effort service types, respectively. The first tier slot allocation mechanism works as follows.

a) The scheduler assigns the next time slot to the node with the smallest *Excess* value in the NAT at the decision time. Suppose that node A is the node scheduled to send at the next time slot. The scheduler then updates the values of *Credit*, *Usage*, and *Excess* in all entries of the NAT, as follows.

b) Increment the *Usage* value of the entry of node A and leave all the others intact, i.e., *Usage* $(A) \leftarrow$ *Usage* $(A) + 1$, where *Excess* $(A) = \min \{Excess\ (x) | x=1,2,\ldots,m\}$.

c) Increase *Credit* by *Resv* for each guaranteed-service entry in the NAT, i.e., *Credit* $(x) \leftarrow$ *Credit* (x) + *Resv* (x), $\forall x \in S_g$. If the time slot is assigned to a best effort service entry, the Credit value of node A is decremented by one, i.e., *Credit* $(A) \leftarrow$ *Credit* $(A) - 1$.

d) Update the *Excess* values of all guaranteed-service entries in the NAT, i.e., *Excess* $(x) \leftarrow$ *Usage* (x) – *Credit* (x), $\forall x \in S_g$. If the time slot is assigned to a best effort service entry, and the *Credit* value of node A equals zero, the Excess value is incremented by one and the *Credit* value is reset to the *Resv* value, i.e., if *Credit* $(A) = 0$, *Excess* $(A) \leftarrow$ *Excess* $(A) + 1$, and *Credit* $(A) \leftarrow$ *Resv* (A); the Excess values of other best effort entries stay unchanged.

The second tier slot allocation algorithm works as follows.

[1] Each iteration indicates a cycle in which the slot quota is counted down from the original value to zero.

a) The node cannot schedule any flow transmission unless a time slot is assigned to it. Once a time slot is assigned, the node assigns the time slot to the flow which satisfies the following two conditions: (i) the flow has the smallest Excess value, and (ii) the flow's service type matches the service type of the node granted by the scheduler.

Suppose that flow F is scheduled to transmit at the next time slot. The node updates the values of *Credit*, *Usage*, and *Excess* in all entries of the FAT, as follows:

b) Increment the *Usage* value of the entry of flow F, and leave the others intact. $Usage\ (F) \leftarrow Usage\ (F) + 1$, where $Excess\ (F) = \min\{Excess\ (y) | y = 1,2,\ldots,n\}$.

c) Update the *Excess* values of all entries in the FAT, i.e., $Excess\ (y) \leftarrow Usage\ (y) - Credit\ (y), y = 1,2,\ldots,n$.

d) $Credit\ (y) = Credit\ (y) + Resv\ (y), y = 1,2,\ldots,n$.

D. Global Fairness in CSAP

So far, we have described how to ensure per-cluster fairness while providing QoS service to guaranteed flows. In [8,9,11], the mechanisms focus on global fairness. That is, no matter where the flow is, the resource usage is proportion to the flow weight. To achieve this goal, they set a parameter called "*scheduling window*" to ensure global fairness. CSAP can be easily extended to ensure global fairness when a scheduling window mechanism similar to that described in [8,9,11] is implemented. The window mechanism in [8,9,11], however, has some problem when applied directly to CSAP. For example, the service guarantee fails if the timer expires before the time slots for the minimum requirement of a flow are allocated.

Let η be the scheduling window for CSAP. The scheduling rules for global fairness in CSAP are revised as follows.

1) When a flow, say, i has sent up to η additional packets in a period of time after its minimum QoS requirement has been satisfied, the node, say, n will stop scheduling further time slots to flow i, and will send an update message to the scheduler to temporarily degrade its Resv level. For example, node N manages three flows F0, F1, and F2, with a Resv of 0.2, 0.2, and 0.3, respectively. All these three flows are registered with the same scheduler S. Suppose F1 has sent up to η packets in a period of time. Thus, node N will send an update message to S to degrade its total Resv level from 0.7 to 0.5.

2) For a scheduler, the rules to make scheduling decisions are intact as described in Sec. II.C.(2).

3) When node N has been assigned the next time slot, it will in turn assign the slot to the flow satisfying the following three conditions: (i) the flow is not frozen to be scheduled, (ii) the flow's service type matches the service type of the node granted by the scheduler, and (iii) the flow has the smallest *Excess* value.

After the expiry of the predefined period of time, node N resumes the scheduling for frozen flow and informs the scheduler to resume its original *Resv* level.

III. Simulation

In this section, the performance of CSAP, Q_EMLM-FQ, and BCFS is compared by simulation. There are 20 mobile nodes randomly distributed in a 670-by-670 area. We randomly select nodes, some as flow senders and some as flow receivers. Each flow may be of best effort or guaranteed. For best effort flows, their Resv values are set to zeros, and for guaranteed flows, the values are randomly generated between zero and one. The packet length is fixed, and is assumed to occupy one time slot. The time period and the window size used for the scheduling window of the global fairness mechanism are set to 200 time slots and 150 packets, respectively.

A. Performance Metrics

1) Share degree (ϕ): the percentage of residual bandwidth shared by each flow. Here the residual bandwidth means the bandwidth left after the minimum requirements requested by all QoS flows have been allocated. The share degree of flow F is $Usage(F)$, divided by total number of time slots, minus $Resv(F)$, i.e., $\varphi = \frac{Usage(F)}{total\ slots} - Resv(F)$. The share degree of a flow is value between [−1,1]. The larger the value, the more residual bandwidth the flow has used. A flow with a negative ϕ value must be a QoS flow whose minimum bandwidth requirement is not satisfied.

2) Satisfaction index (ρ): this parameter is used to indicate how well all QoS flows are satisfied by each approach. ρ is defined in a way similar to the definition of the fairness index in [16]. The parameter x_i indicates if a QoS flow has been satisfied. If the time slot usage of a QoS flow is larger than or equal to its minimum requirement, x_i=1; otherwise x_i indicates the insufficient portion. Thus $\rho = \frac{1}{n}\sum_{i=1}^{n} x_i / x_f$, where $x_i = \begin{cases} 1 & , if\ \varphi(i) \geq 0 \\ 1 - \alpha \times \frac{|\varphi(i)|}{Resv(i)} & , if\ \varphi(i) < 0 \end{cases}$, whose $\alpha \geq 1$ is a weighting factor, and $x_f = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i}$. The satisfaction index of an approach is a value between zero and one. The larger the value, the more flows are satisfied with the desired service.

3) Fairness index (κ): this parameter indicates how fair the residual bandwidth is shared by all flows for each approach. The parameter is defined as $\kappa = \frac{1}{m}\left\{\sum_{j=1}^{m}\left(\frac{1}{n_j}\sum_{k=1}^{n_j}\frac{x_k}{x_f(j)}\right)\right\}$, where m is the number of clusters, n_j is the number of flows in cluster j; x_k indicates the share degree of flow k, and $x_k = \begin{cases} \varphi(k), if\ \varphi(k) \geq 0 \\ 0 \quad , if\ \varphi(k) < 0 \end{cases}$; the x_f value of cluster j is $x_f(j) = \frac{\sum_{k=1}^{n_j} x_k^2}{\sum_{k=1}^{n_j} x_k}$. Fairness index κ is a value between zero and one. The larger the value, the fairer the residual bandwidth is

shared among all flows. Note that this index can be used for global fairness when m is set to one.

4) Network throughput: defined as $\sum_{i \in N} \frac{Usage(i)}{total\ slots}$, where N is the set of all flows in the ad hoc network. The larger the network throughput, the more efficient the spatial channel reuse.

A. Simulation Results

In the following, simulation results are presented. We first examine the single cluster case, and then to multiple clusters. We also demonstrate that CSAP can still ensure minimum bandwidth requirements for QoS flows and provide global fairness for all flows in the last simulation. Finally, we show the impact of different clustering schemes and cluster size on the performance of each mechanism.

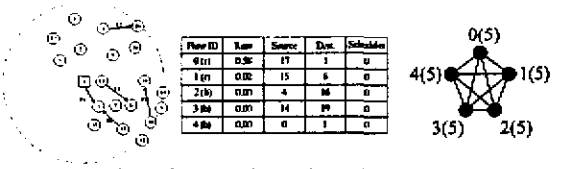
1) Single Cluster

We first present the simulation results for single cluster scenarios. The transmission range of each node is set to 670 units. In the first experiment, 5 pairs of nodes in a single cluster are randomly selected as flow senders and receivers, and five flows are randomly generated: two guaranteed flows and three best effort flows. Flows 0 and 1 are guaranteed flows with a *Resv* value of 0.58 and 0.02, respectively. Flows 2, 3, and 4 are best-effort flows. Ideally, the share of the residual bandwidth for each flow should be (1–0.58–0.02)/5=0.08.

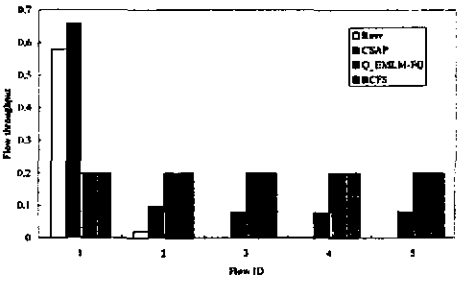
Fig. 4(a) shows the network topology and flow distribution. In this figure, the square node indicates the cluster scheduler, the circles represent mobile nodes, an arrow shows the direction of a flow, from the source to the destination, the solid lines indicate best effort flows, and the dotted lines indicate guaranteed flows. Fig. 4(b) shows the corresponding flow graph as described in Sec. I.C.(2).(b). x and y in x(y) means the flow ID and the contending power of the flow, respectively. The flow graph is a 4-degree fully connected graph, in which each flow has four neighbors and each has the same contending power of 5. Since each flow has the same value of contending power, the scheduling decision made by BCFS is based on the least service tag. Thus, it has the same result as Q_EMLM-FQ. Fig. 4(c) shows the flow throughput to the five flows for each mechanism. We see that CSAP meets the minimum requirements of all flows (see flows 1 and 2 in Fig. 4(c)), and thus has a satisfaction index of one (Fig. 4(d)). In addition, it has equal share of residual bandwidth among five flows, each with a share degree of 0.08, the same as the ideal sharing (Fig. 4(e)). Thus, CSAP has a fairness index of one (Fig. 4(f)).

Both Q_EMLM-FQ and BCFS allocate resource equally to the five flows (Fig. 4(c)), despite different types of service requested by the flows. As a result, they do not guarantee the minimum bandwidth requirement of the guaranteed flow (i.e., flow 0 in Fig. 4(c)), and has a lower satisfaction index. Meanwhile, both Q_EMLM-FQ and BCFS have a flow with a negative share degree. Thus, they have a lower fairness index, which is 0.7984.

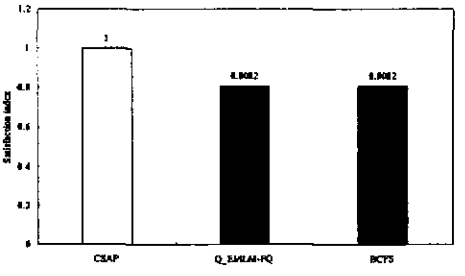
We then randomly generate 10 guaranteed flows in a single cluster ad hoc network. Fig. 5(a) shows the network topology,



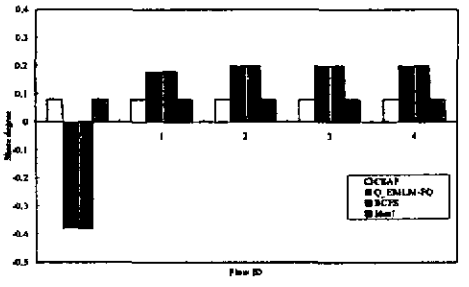
(a) Network topology and flow information (b) Flow graph



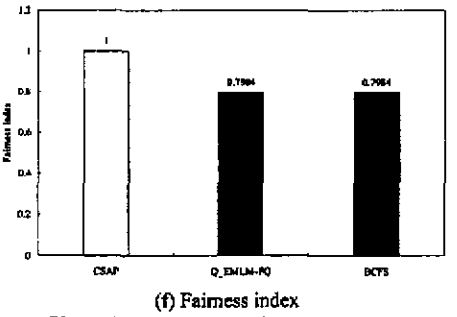
(c) Flow throughput



(d) Satisfaction index



(e) Share degree



(f) Fairness index

Figure 4. Single cluster with a mix of flows

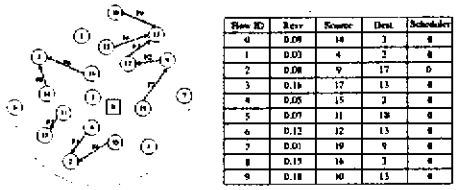
and Fig. 5(b) shows the flow information. The flow graph of this experiment is, again, the same as in Fig. 5(b). Fig. 5(c) shows the flow throughput of the 10 flows for each mechanism. Both Q_EMLM-FQ and BCFS allocate slots relatively equally to each flow, irrespective of whether it is a best effort or guaranteed flow. Three guaranteed flows are not satisfied (i.e., flows 3, 6, 8 in Fig. 5(c)). Thus, both have a satisfaction index of 0.8934 (Fig. 5(d)). In addition, they do not share the residual

bandwidth fairly (see Fig. 5(e)), and have a fairness index of 0.4314 (Fig. 5(f)). CSAP satisfies the requirements of all flows except flow 6, and shares residual bandwidth more fairly. Thus, it has a higher satisfaction index and a higher fairness index.

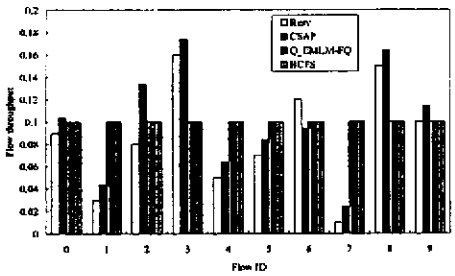
2) Multiple Cluster

We then extend the experiment from one cluster to multiple clusters. We randomly generate a mix of best effort and guaranteed flows. This time the network is logically partitioned into four clusters. Flows 0, 1, 3, 6, and 7 are scheduled by node 0. Ideally, each should share a fraction (1–0.1–0.27–0.05–0.26)/5=0.064 of the residual bandwidth. Flows 2, 5, and 8 are managed by node 4, and each should have an equal share of (1–0.20–0.16)/3=0.2133. Both flows 4 and 9 are managed by node 3, and each shares a fraction of residual bandwidth equal to (1–0.29)/2=0.355. Fig. 6(a) shows the network topology, the flow distribution and the flow graph. The detailed flow information is shown in Fig. 6(b). Figs. 6(c) and 6(e) show the flow throughput and the share degree of each flow. Figs. 6(d) and 6(f) show the satisfaction and fairness indices of each approach.

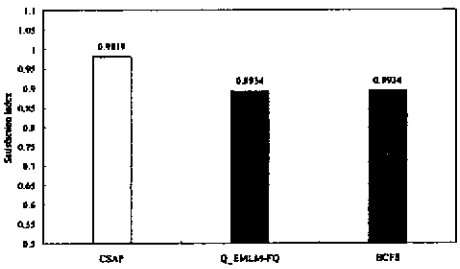
Again, CSAP provides the minimum bandwidth requirements for all flows and ensures fair share for per-cluster residual bandwidth. As a consequence, both the satisfaction and fairness indices are equal to one. Q_EMLM-FQ, on the



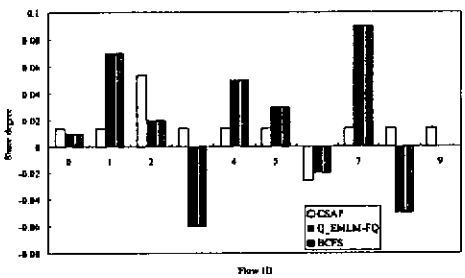
(a) Network topology (b) Flow information



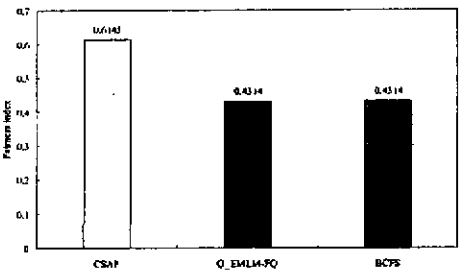
(c) Flow throughput



(d) Satisfaction index

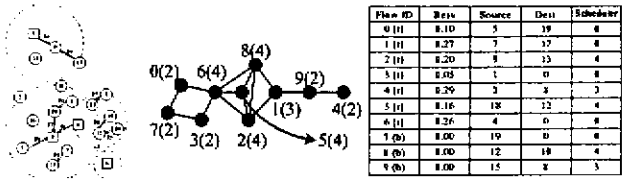


(e) Share degree

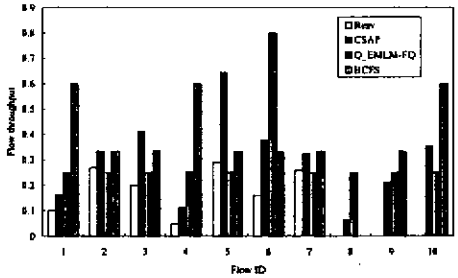


(f) Fairness index

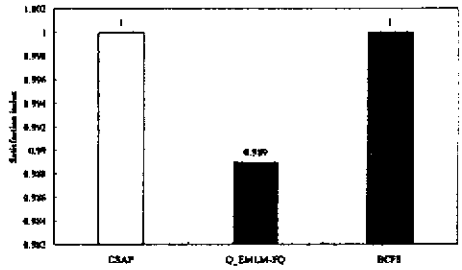
Figure 5. Single cluster with guaranteed flows only



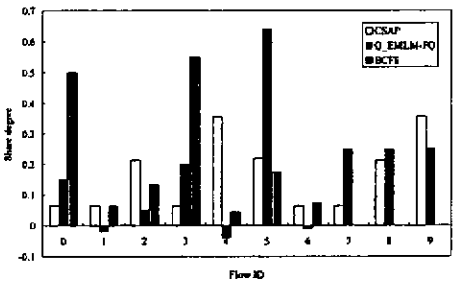
(a) Network topology and flow graph (b) Flow information



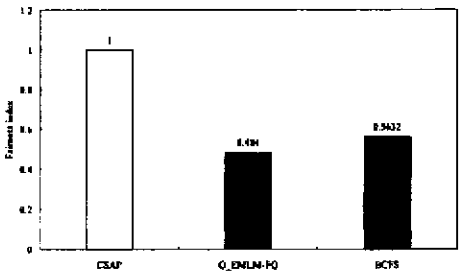
(c) Flow throughput



(d) Satisfaction index



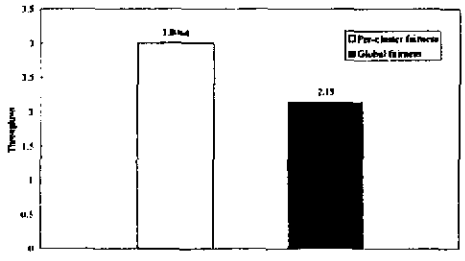
(e) Share degree



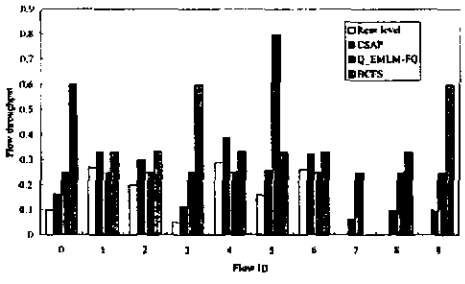
(f) Fairness index

Figure 6. Multiple clusters with a mix of flows

other hand, cannot always satisfy the requirements of all flows. BCFS can meet the minimum requirements of all QoS flows,



(a) Per-cluster vs. global fairness



(b) Flow throughput with global fairness
Figure 7. Global fairness

but fails to fairly share the residual bandwidth.

The last experiment is conducted to observe the impact of global fairness on the performance of CSAP. Fig. 7(a) shows the average system throughputs to compare per-cluster and global fairness in the multi-cluster scenario shown in Fig. 6. We see that the throughput of CSAP degrades when the global fairness mechanism described in Sec. II.B.(5) is implemented. Fig. 7(b) shows that CSAP with the global fairness mechanism still ensures the minimum requirements of QoS flows and provides global fairness to all flows in sharing the residual bandwidth.

IV. Conclusion

In this paper, we have proposed a credit-based mechanism called CSAP to provide fair scheduling with QoS supports for multimedia wireless d hoc networks. In addition, we have proposed Q_EMLM-FQ, which introduces a distributed flow weight calculation scheme to [9], to make it capable of QoS support. We have also pointed out some problems of BCFS, which is extended from [8] for QoS service.

To compare CSAP with Q_EMLM-FQ and BCFS quantitatively, some performance indices are defined. We have also conducted simulations to observe the fairness and satisfaction degrees of these three approaches. The results show that of these three mechanisms, CSAP performs the best because it can guarantees QoS flows with the minimum bandwidth requirements, and share the residual bandwidth more fairly with all flows in the ad hoc network. Both Q_EMLM-FQ and BCFS usually allocate bandwidth equally, despite different types of service requested by the flows. As a result, they may not function very well.

References

[1] C. R. Lin and J.-S. Liu, "QoS routing in ad hoc wireless networks," *IEEE Journal on Selected Areas in Communication*, Vol. 17, No. 8, Aug. 1999, pp.1426-1438.

[2] S. Chen and K. Nahrstedt, "Distributed quality-of-service routing in de hoc networks," *IEEE Journal on Selected Areas in Communication*, vol. 17, no. 8, Aug. 1999, pp.1488-1505.

[3] P. Gevros, J. Crowcroft, P. Kirstein, and S. Bhatti, "Congestion control mechanisms and the best effort service model," *IEEE Network*, May/June 2001, pp. 16-26.

[4] P. Goyal, H. M. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Transaction on Networking*, vol. 5, Oct. 1997, pp. 690-704.

[5] J. M. Jaffe, "Bottleneck flow control," *IEEE Transaction on Communication*, vol. 29, no. 7, Jul. 1981, pp. 954-962.

[6] E. Hossain, and V. K. Bhargava, "A centralized TDMA-based scheme for fair bandwidth allocation in wireless IP networks," *IEEE Journal on Selected Areas in Communication*, vol. 19, no. 11, Nov. 2001, pp. 2201-2214.

[7] Cao, Y. and Li, V.O.K., "Scheduling algorithms for broadband wireless networks," *Proc. of the IEEE*, Vol. 89, No. 1, Jan 2001, pp. 76 - 87.

[8] H. Luo and S. Lu, "A topology-independent fair queueing model in ad hoc wireless networks," *Proceedings of IEEE International Conference on Network Protocols*, Nov. 2000, pp. 325-335.

[9] H. Luo and S. Lu, "A self-coordinating approach to distributed fair queueing in ad hoc wireless networks," *Proceedings of IEEE INFOCOM*, Apr. 2001, pp. 1370-1379.

[10] N. H. Vaidya and P. Bahl, "Fair scheduling in broadcast environments," *Microsoft Research Technical Report MSR-TR-99-61*.

[11] X. Wu, Y. Gao, H. Wu, and B. Li, " Fair scheduling with bottleneck consideration in wireless ad-hoc networks," *Proceedings of IEEE International Conference on Computer Communication and Networks*, Oct. 2001, pp. 568-572.

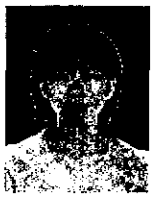
[12] H. L. Chao, and W. Liao, "Credit-based fair scheduling in wireless ad hoc networks," *Proceedings of IEEE Vehicular Technology Conference*, Sep. 2002.

[13] H.-C. Lin and Y.-H. Chu, "A clustering technique for large multihop mobile wireless networks," *Proceedings of IEEE Vehicular Technology Conference*, Sep. 2000, pp. 1545-1549.

[14] D. J. Baker and A. Ephremides, "The architecture organization of a mobile radio netework via distributed algorithm," IEEE Transactions on Communications, Nov. 1981, pp. 1694-1701.

[15] C. E. Perkins and E. M. Royer, "Ad-hoc on demand distance vector routing," *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, Feb. 1999, pp. 90-100.

[16] R. K. Jain, D. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," *DEC Technical Report DEC-TR-301*, 1984.



Hsi-Lu Chao received the B.S. degree in Electronic Engineering from Fengchia University in 1992, and the M.S. degree in Electrical Engineering from University of Southern California in 1996. She is currently a Ph.D. candidate in the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan. Her research interests include wireless communication and QoS issues in ad hoc networks.

Wanjiun Liao received the BS and MS degrees from National Chiao Tung University, Taiwan, in 1990 and 1992, respectively, and the Ph.D. degree in Electrical Engineering from the University of Southern California, Los Angeles, California, USA, in 1997. She joined the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, as an Assistant Professor in 1997. Since August 2000, she has been an Associate Professor. Her research interests include wireless networks, optical networks, and broadband Internet.

Dr. Liao is actively involved in the international research community. She is an Associate Editor for IEEE Transactions on Wireless Communications. Dr. Liao has received many research awards. She was a recipient of the Outstanding

Research Paper Award in Electrical Engineering at the University of Southern California in 1997. Two papers she co-authored with her students received the Best Student Paper Award in the First IEEE International Conferences on Multimedia and Expo (ICME) in 2000, and the Best Paper Award in the First IEEE International Conference on Communication, Circuits and Systems in 2002. Dr. Liao was elected as one of Ten Distinguished Young Women in Taiwan in 2000 and is listed in the Marquis Who's Who in 2001-2003, and the Contemporary Who's Who in 2003.

Fair Scheduling with Error Compensation for Mobile Ad Hoc Networks

Hsi-Lu Chao and Wanjiun Liao
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
Email: wjliao@cc.ee.ntu.edu.tw

Abstract- **In this paper, we study fair scheduling in ad hoc networks, accounting for channel errors. Since wireless channels are susceptible to failures, to ensure fairness, it may be necessary to compensate those flows with error-prone channels. Existing compensation mechanisms need the support of base stations and only work for one-hop wireless channels. Therefore, they are not suitable for multihop wireless networks. Existing fair scheduling protocols for ad hoc networks can be classified into timestamp-based and credit-based approaches. None of them takes channel errors into account. In this paper, we investigate the compensation issue of fair scheduling and propose a mechanism called Timestamp-Based Compensation Protocol (TBCP) for multimedia mobile ad hoc networks. We evaluate the performance of TBCP by simulation and analyze its long-term throughput. The results show that our analytical result provides accurate estimation for TBCP.**

Keywords- **fair scheduling, compensation, ad hoc networks**

I. INTRODUCTION

An ad hoc network is a self-organizing wireless network comprised only of mobile nodes without the support of any pre-existing wired infrastructure. According to different decision metrics, existing work for fair scheduling in ad hoc networks can be classified into two categories: *timestamp based* [1,2] and *credit based* [3]. The timestamp mechanisms in [1] and [2] work similarly. Using [2] as an example, each arriving packet is locally assigned two timestamps: a start tag and a finish tag. Either timestamp can be chosen as the service tag. The packet with the smallest service tag will be sent first. In [3], scheduling is based on the credit value. The unused credit can be accumulated for future use. Each flow is associated with three parameters: a credit, a usage, and an excess, where excess = usage – credit. The one with the smallest excess value has the priority to transmit.

All existing work of fair scheduling in ad hoc networks [1-3] simply assumes channels are error-free. This assumption, however, is unrealistic for wireless networks, which usually suffer from high error rates, such as location-dependent errors or bursty errors. In this paper, we investigate the impact of channel errors on fair scheduling and discuss the channel compensation issue of fair scheduling in ad hoc networks.

A. *Related Work on Fair Scheduling for Wireless Networks with Channel Errors*

Existing work proposed to achieve fairness with error compensation for wireless networks [4-8] are all based on the support of base stations and work for one-hop wireless channels.

In [5], the Channel-condition Independent packet Fair Queueing (CIF-Q) is proposed. Each session maintains a parameter called *lag* to indicate if it should be compensated. If a session is not leading and the channel is error-free at its scheduled time, its head-of-line packet is transmitted; otherwise, the time slot is released to other sessions. The problem with CIF-Q is that leading sessions are not allowed to terminate unless all their leads have been paid back, regardless of whether such terminations are caused by broken routes. This property makes CIF-Q an infeasible solution for ad hoc networks, because a connection may be broken due to node movements.

In [6], the Idealized Wireless Fair-Queueing (IWFQ), and the Wireless Packet Scheduling protocol (WPS) are proposed. According to the weight of each session, the base station calculates the number of slots per frame each session can use. If a session experiences channel errors at its scheduled time, the base station first determines if any other sessions can exchange their slots with this session within the same frame; otherwise, the base station will compensate the error slots of this session in a later frame. Again, this mechanism is not suitable for ad hoc networks, due to the need of support from base stations.

In [7], a virtual compensation session is introduced into the scheduling process. The virtual session is a session, which always experiences an error-free channel at its scheduled time, but it does not really generate packets. The slots received by this virtual session are used for compensation, i.e., they will be reassigned to lagging sessions. Since this mechanism only works for single-hop wireless networks, again, it is not suitable for multihop wireless ad hoc networks.

In [8], Bandwidth-Guaranteed Fair Scheduling with Effective excess Bandwidth Allocation (BGFS-EBA) is proposed. In BGFS-EBA, each flow can be a lagging,

leading, or in-synch flow as compared to its bandwidth allocation in an error-free environment. Each flow is associated with a parameter called deadline. Bandwidth resource is first scheduled according to such deadlines, and then further allocated based on flow status (i.e., lagging, leading, or in-synch) as in CIF-Q. Unlike other mechanisms which allow a flow to transmit only in a good state, a flow in BGFS-EBA is still allowed to transmit, albeit at lower rates, when its channel is error-prone. Again, this mechanism also needs the support of base stations.

B. *Problem Statement*

In this paper, we study fair scheduling in ad hoc networks in which channel errors are considered. In particular, we focus on the timestamp-based fair scheduling mechanism. In credit-based mechanisms (e.g., CSAP [3]), unused credits can be accumulated for future use and error compensation is naturally available. Note that we call the mechanisms considering channel errors in their scheduling as channel error models, and those assuming that channels are error-free in their scheduling as error-free models. In addition, we call the packets experiencing channel errors at their scheduled time as channel-error packets, and the ones without errors as error-free packets.

There are two timestamp-based error-free mechanisms in the literature proposed for ad hoc networks (i.e., [1,2]). However, these two protocols only work for single-hop flows, a special case in ad hoc networks. In the next section, we will show how to extend these single-hop timestamp mechanisms to serve multihop wireless networks.

The rest of this paper is organized as follows. Sec. II describes the proposed timestamp-based protocol. Sec. III shows the simulation results. Sec. IV provides the long-term throughput of TBCP. Finally, the paper is concluded in Sec. V.

II. TIMESTAMP-BASED COMPENSATION PROTOCOL (TBCP)

In this section, a timestamp-based fair scheduling mechanism, called Timestamp-Based Compensation Protocol (TBCP), is proposed for multihop wireless networks, under which channel errors are considered. TCP adopts Start-time Fair Queueing (SFQ) [9] as its scheduling discipline and selects the start tag as its service tag.

A. *System Model*

We assume a TDMA-based system operates over a single channel shared by all hosts, and the bandwidth is represented as frames of time slots. Each frame has several slots: some are control slots and some are data slots. A valid route should be constructed, using whatever ad hoc routing mechanisms, before data packets are transmitted. Note that the spatial channel reuse mechanisms in [1,2] are still applicable to our mechanism.

Each node periodically measures the Signal-to-Noise Ratio (*SNR*) on the channel to determine the channel state. When the *SNR* value falls below a predefined threshold, the channel status is deemed error-prone and the node is prevented from transmitting packets temporarily. If the *SNR* value exceeds the predefined threshold again, the node is allowed to transmit again.

B. *Multihop Flows*

In ad hoc networks, each mobile node acts as both a router and a host. Thus, a multihop flow can be modeled as multiple single-hop flows, and each node schedules the single-hop flows passing through it independently. It is insufficient for a node to schedule for multihop flows relying only on the selected scheduling parameter (e.g., timestamp or credit values). To solve this problem, a new parameter called Q-size is defined. The Q-size of a flow is used to indicate the number of packets received from its previous hop and waiting to be sent to the next hop. In other words, those single-hop flows belonging to the same multi-hop flow are correlated with the Q-size parameter. Therefore, among all nodes with nonzero Q-size values, the node with the least scheduling parameter value can use the next time slot to transmit.

C. *Normalized Flow Weight*

Assume that there are n flows passing through node N. Let x_i^N denote the bandwidth share of flow i at node N (called the flow weight of flow i at node N in this paper), and is expressed as

$$x_i^N = \frac{1 - \sum_{j=1}^{n} Resv_j}{n} + Resv_i, \qquad (1)$$

where $Resv_i$ is the minimum bandwidth requirement of flow i, represented as a fraction of the channel bandwidth. Flow weight x_i^N indicates the fraction of channel bandwidth used by flow i at node N, so as to fairly share the residual bandwidth of node N and meet the minimum bandwidth requirement of flow i. For example, assume that there are three flows, say, f_1, f_2, and f_3 passing through node N. Flows f_1 and f_2 are guaranteed flows, each with a *Resv* value of 0.2; flow f_3 is a best effort flow. Based on (1), the flow weights of the three flows at node N are $(x_1^N, x_2^N, x_3^N)=(0.4, 0.4, 0.2)$.

In TBCP, the actual bandwidth share for each flow at node N is determined by the flows at all nodes in the interference range (i.e., nodes located within two hops), not just solely depending on those flow passing through node N. Due to lack of coordination for transmissions, nodes located in the interference range contend the channel, and should be considered in the calculation of bandwidth share for each flow. To reflect this fact, the normalized flow weight of each flow is defined as follows.

Let F_N and $F_{N'}$ be the set of flow segments passing through node N and N', respectively, where N' is within two-hop range of node N. Besides, let S_N be the set of node including node N and all other nodes in its interference range (i.e., two hops). For each flow segment f at node N, its normalized flow weight is expressed as.

$$\omega_f^N = \frac{x_f^N}{\sum_{i \in (F_N \cup F_{N'}), j \in S_N} x_i^j} \quad (2)$$

This normalized flow weight ω_f^N is then used for the calculation of service tag of flow f at node N. The number of slots per frame that flow f can obtain at node N is the multiplication of ω_f^N and the number of slots per frame. Note that since the calculations of (1) and (2) are performed independently at each node, different single-hop flow segments of a multihop flow may have different bandwidth shares and normalized flow weights.

For example, considering two nodes, say N_1 and N_2, are in an ad hoc network. These two nodes are within the transmission range of each other. Let $Resv_i^j$ indicate the $Resv$ value of the i^{th} flow at node j. The flow information of N_1 is $(Resv_1^{N_1}, Resv_2^{N_1}, Resv_3^{N_1}) = (0.2,0.2,0)$, and that of N_2 is $(Resv_1^{N_2}, Resv_2^{N_2}, Resv_3^{N_2}) = (0.4,0,0)$. Thus the flow weight calculated by N_1 and N_2 are (0.4,0.4,0.2) and (0.6,0.2,0.2), respectively, and the normalized flow weights of all flows are $(\omega_1^{N_1}, \omega_2^{N_1}, \omega_3^{N_1}, \omega_1^{N_2}, \omega_2^{N_2}, \omega_3^{N_2})$ = (0.2,0.2,0.1,0.3,0.1,0.1).

D. TBCP Operations

C.1 Transmission Order Determination at a Node

The transmission order of a packet is determined with three factors: the service tag of the packet, the number of slots per frame a flow can use, and flow's Q-size. For example, the numbers in Fig. 1(a) show the packets' service tags of three flows f_1, f_2, and f_3 at node N. Assume that each frame is comprised of five slots, and f_i^j indicates the j^{th} packet of flow i. Besides, the number of slots per frame each flow can use is (2,2,1). The transmission order of packets at node N in Fig. 1(a) is then $< f_1^1, f_3^1, f_1^2, f_2^1, f_2^2 >$. Note that since f_1 has already used two slots, the last slot of this frame is assigned to f_2 even though the service tag of f_1^3 is smaller than f_2^2.

C.2 Message Exchange

Each node exchanges the information about the transmission order of packets determined at step 1) with its neighbors. Thus each node knows the service tags of other nodes, and also learns when it will transmit packets. For example, assume that there are two nodes in an ad hoc network, and the corresponding transmission order with the service tag is shown in Fig. 1(b). Let N_i^j indicates the j^{th} slot of node i. The transmission sequence is $< N_1^1, N_2^1, N_2^2, N_2^3, N_1^2 >$, $< N_1^3, N_2^4, N_1^4, N_1^5, N_2^5 >$. Note that if multiple nodes have packets with the same service tag, the tie will be broken based on their node IDs.

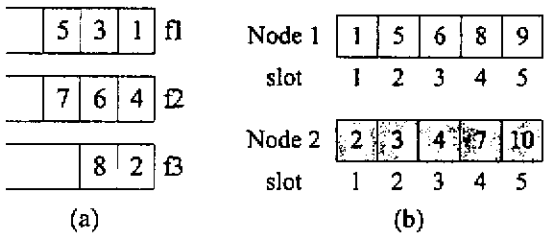


Figure 1. An example to explain TBCP.

C.3 Channel Error Handling

Each node keeps monitoring its channel state. When the channel is error-prone, the node stops exchanging transmission messages with its neighbors. Once the channel recovers, the error-prone node resumes the exchanges. Consequently, if this node has packets with service tags smaller than its neighbors after recovery, these packets still have higher priority to be transmitted. If node 2 experiences an error-prone state at slot 4, the modified transmission sequence is $< N_1^1, N_2^1, N_2^2, X, N_1^2 >$, $< N_2^3, N_1^3, N_2^4, N_1^4, N_1^5 >$.

III. Simulation

In this section, we provide simulation results to evaluate the performance of the fair scheduling mechanisms with and without considering channel errors. The simulation environment is described as follows. 20 mobile nodes are randomly distributed in a 1000-meter by 1000-meter area. The transmission range of each node is set to 250 meters. We randomly select nodes, some as flow sources and some as flow destinations. Each flow may either be a best effort or guaranteed flow and is continuously with backlogged packets. Each packet is assumed to occupy one time slot, and has fixed packet length. The mobility pattern of each node follows the modified random waypoint model [10].

The wireless channel is modeled as a two-state discrete Markov chain in our simulation, as in [11]. Let p_g be the probability that the next time slot is in the good state given that the current time slot is error-prone, and let p_e be the probability that the next time slot is error-prone given that the current slot is in the good state. Then the steady state probabilities P_G and P_E in the good and error-prone states, respectively, are given by

$$P_G = \frac{p_g}{p_g + p_e} \quad \text{and} \quad P_E = \frac{p_e}{p_g + p_e}.$$

In this simulation, we compare TBCP with TBP (the error-free model of TBCP, i.e., assuming that channels are

error-free in the scheduling). Besides, we implement the spatial channel reuse scheme as in [2]. We also show CBCP and CSAP [4] (denoted as CBP in this paper) in the figures for comparison, where CBP and CBCP are the credit-based mechanisms for channel error-free and channel error models, respectively. CBCP differs from CBP in that when a node detects an error-prone state at its scheduled time, the node will inform its scheduler to stop assigning further slots to it. The Credit value of this node at the scheduler continues to be accumulated. Thus, once the channel has recovered, the node has a small Excess value as compared to other error-free nodes, which allows this node to have priority to obtain slots. This node then in turn assigns this slot to the flow with the least Excess value among all nonzero Q-size flows.

The performance metrics measured in the simulation include the network throughput, relative network throughput, satisfaction index, and fairness index.

(1) Network throughput (ρ): this parameter is used to compare the performance of each approach for both error-free and error-prone channel models. The better the compensation scheme, the less throughput degradation. Network throughput is the summation of all flows' throughputs.

(2) Relative network throughput (ψ): this parameter is defined as

$$\frac{throughput_{CBCP}}{thgoughput_{CBP}} \text{ and } \frac{throughput_{TBCP}}{thgoughput_{TBP}},$$

for credit-based mechanism and for timestamp-based mechanism, respectively.

(3) Satisfaction index (ξ): this parameter shows if guaranteed flows are satisfied for each approach, considering channel conditions. Thus we only count QoS flows and define ρ as

$$\frac{\left(\sum_i x_i\right)^2}{m\sum_i \left(x_i^2\right)},$$

where x_i is the residual bandwidth share (i.e., flow throughput divided by simulation time, and then minuses the *Resv* value), and is set to one if the corresponding x_i is equal or larger than zero, otherwise x_i is one minuses the shortage divided by its *Resv* value; m is the number of guaranteed flows.

(4) The fairness index (κ): this parameter indicates how fair the residual bandwidth is shared by all flows for each approach, and is defined as in [12], i.e.,

$$\frac{\left(\sum_i x_i\right)^2}{n\sum_i \left(x_i^2\right)},$$

where x_i is the residual bandwidth share if x_i is equal or larger than zero, otherwise x_i is zero; n is the number of flows in an ad hoc network. Note that we do not use the proportional fairness defined in [5,6], i.e.,

$$\left|\frac{W_i(t_1,t_2)}{\omega_i} - \frac{W_j(t_1,t_2)}{\omega_j}\right|,$$

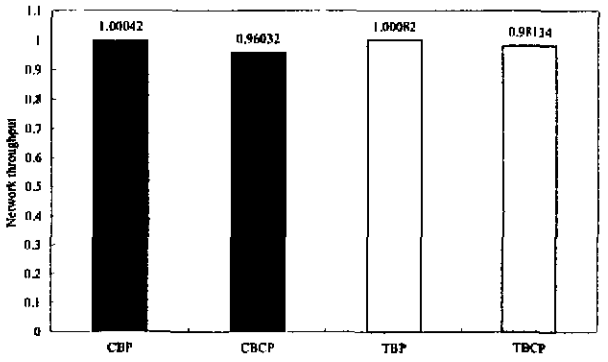
as the fairness index. The reason is for a multihop flow, the single-hop flow segments may not have the identical flow weight.

We generate five multihop flows: three guaranteed flows and two best-effort flows. The steady state probabilities P_G and P_E are 0.8 and 0.2, respectively. The Min and Max speeds are set to be 10 meters per sec and 30 meters per sec, respectively. The network throughput of each approach is shown in Fig. 2(a). The satisfaction and fairness indices of each approach are shown in Figs. 2(b) and (c), respectively. We find that mobility causes broken routes more frequently, thus degrades network throughput. However, node mobility may increase the probability for a flow to change the area in which it is located (i.e., cluster for the credit-based schemes and interference range for time-stamp based mechanisms). This helps improve global fairness.

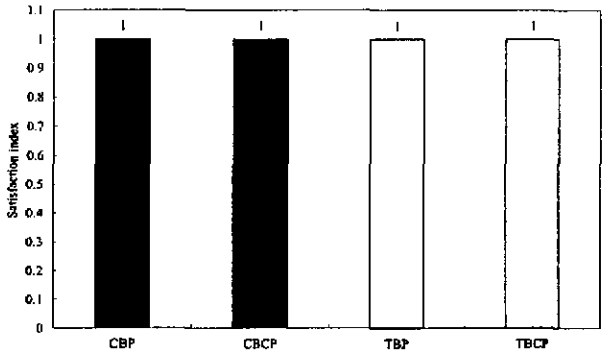
Finally, we study the impact of channel errors on the proposed approaches. We vary the value of P_G but fix all other parameters. The relative network throughput of each approach is shown in Fig. 2(d). As the value of P_G increases, the relative network throughput increases. The reason is that a larger P_G value means more good slots, leading to more successful transmissions and higher network throughput. The satisfaction and fairness indices of each setting are shown in Figs. 2(e) and 2(f), respectively. Since a large P_G value makes more good slots, the possibility to satisfy QoS flows is higher. Thus, it gives a higher satisfaction index. A large P_G value also increases the opportunity for a node to be compensated after its channel recovers, and thus has a higher fairness index.

IV. PERFORMANCE ANALYSIS

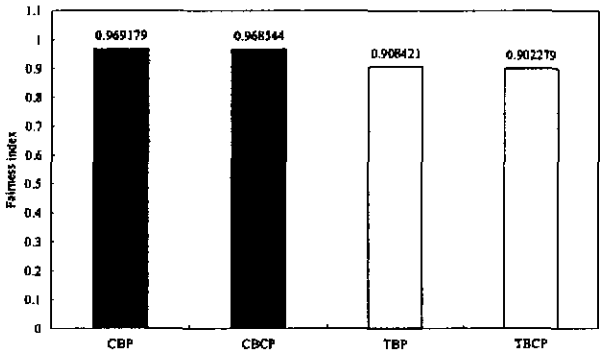
In this section, we analyze the long-term throughput of TBCP. We do not consider node mobility, and ignore spatial channel reuse in this analysis. TBCP is analyzed with two-state channel errors, which is based on the two-state Markov chain in [10].



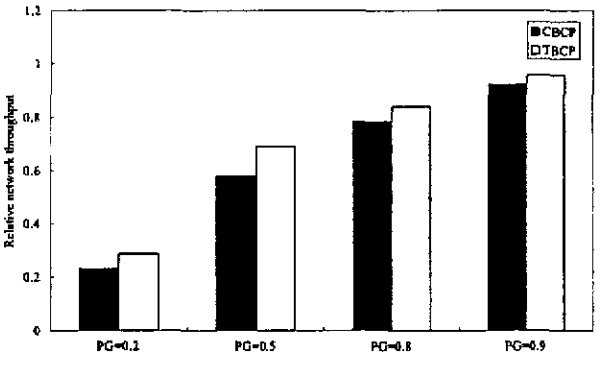
(a) Network throughput



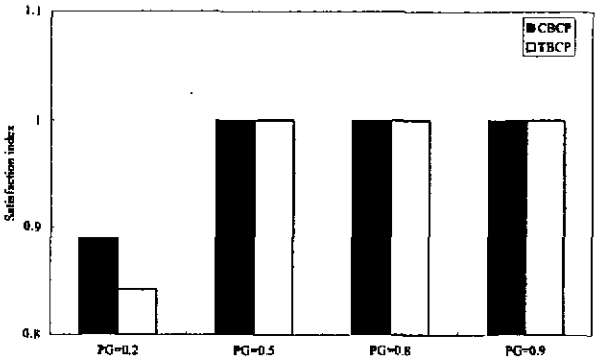
(b) Satisfaction index



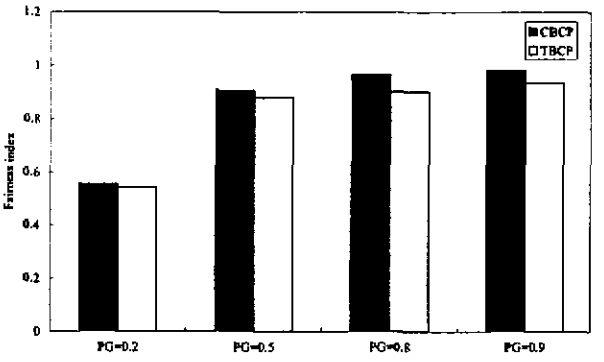
(c) Fairness index



(d) Relative network throughput



(e) Satisfaction index for different P_G values



(f) Fairness index for different P_G values

Figure 2. Two-state Markov-chain error model

A. Analysis

For each node in the ad hoc network, the average number of error-free slots a node incurs is $T \times P_G$, where T is the measurement window, and P_G is the steady state probability of the good state. Let F_N and $F_{N'}$ indicate the sets of flow segments passing through node N, and segments within node N's two-hop range, respectively. For each flow segment f passing through node N, its bandwidth share rate is $x_f^N = Resv_f + \frac{1 - \sum_{i=1}^{n} Resv_i}{n}$, where n is the number of flow segments managed by node N. Furthermore, the normalized flow weight is $\omega_f^N = \frac{x_f^N}{\sum_{i \in (F_N \cup F_{N'}), j \in \{N \cup N'\}} x_i^j}$. For the multihop flow that flow segment f belongs to, its normalized flow weight is $\omega_f = min\left\{ \omega_f^i \middle| i \in P \right\}$, where i is a path node and P is

flow's corresponding path. Therefore, the corresponding flow throughput for flow f is in (3).

$$W_f = \frac{T \times P_G \times \omega_f}{T} = P_G \times \omega_f, \quad f \in F_N \qquad (3)$$

B. Analytical vs. Simulation Results

We randomly generate five flows in the simulation: four best effort flows and one guaranteed flow. The detailed flow and topology information is in Fig. 3(a). Let F_{ij} indicate the j^{th} segment of flow i. Using node 7 as an example, $x_{F_{02}}$ =0.331+(1-0.331)/2=0.6655, and x_{F_2} =(1-0.331)/2=0.3345. Besides, nodes within node 7's two-hop ranges include nodes 6, 11, 12, 14, and 19, thus normalized flow weight will cover F_{01}, F_{02}, F_1, F_2, and F_3. Therefore, the normalized flow weight of F_{02} is $\omega_{F_{02}}$ =0.6655/(0.6655+0.3345+1+1+1)=0.1664. Similarly, $\omega_{F_{01}} = 0.333$. Thus ω_{F_0} =0.1664. Applying the same calculation to other nodes, we obtain $(\omega_{F_0}, \omega_{F_1}, \omega_{F_2}, \omega_{F_3}, \omega_{F_4})$ =(0.1664,0.25,0.083625,0.25,0.33).

We set P_G to be 0.8. Consequently, for the two-state model, each flow's estimated bandwidth share is (0.8×0.1664, 0.8×0.25, 0.8×0.083625, 0.8×0.25, 0.8×0.33) = (0.13312, 0.2, 0.0669, 0.2, 0.264). The comparison with simulation result is in Fig. 3(b). We see that our analytical model provides accurate estimation for the performance of TBCP.

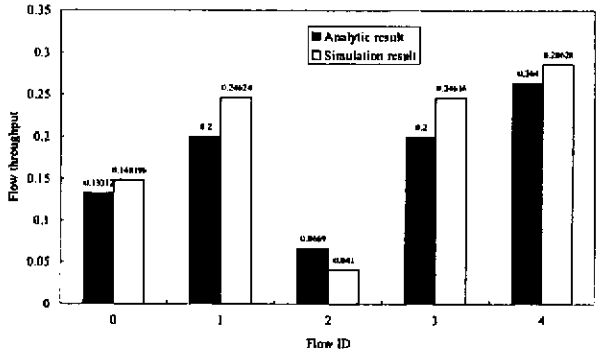
V. CONCLUSION

In this paper, we discuss the channel compensation issue of fair scheduling and propose a mechanism called TBCP for mobile multihop networks. TBCP supports multihop flows, and performs well when node mobility is supported. We describe the detailed operation of TBCP, and conduct simulations to evaluate its performance. From the simulation results, we demonstrate that TBCP satisfies QoS flow demands and provides global fairness for best effort flows. We also show that mobility helps break the locality problem of both protocols for multihop flows. Finally, we also analyze the flow throughput of TBCP, and verify the analytical result with simulation. The results show that our analytical results provide accurate performance estimation for TBCP.

FID	S	D	Type	Resv	Route
0	14	12	G	0.331	14-7-12
1	19	12	B	0	19-12
2	7	6	B	0	7-6
3	6	11	B	0	6-11
4	16	19	B	0	16-19

NID	Neighbors
0	2, 9, 13, 15
1	8, 14
2	0, 9, 13, 15, 16
3	5
4	10, 11, 14, 16, 19
5	3
6	7, 10, 11, 12, 17
7	6, 10, 11, 12, 14, 17
8	1, 14
9	0, 2, 13, 15, 16
10	4, 6, 7, 11, 12, 14, 16, 19
11	4, 6, 7, 10, 12, 14, 19
12	6, 7, 10, 11, 17, 19
13	0, 2, 9, 15
14	1, 4, 7, 8, 10, 11
15	0, 2, 9, 13
16	2, 4, 9, 10, 19
17	6, 7, 12, 18
18	17
19	4, 10, 11, 12, 16

(a) Flow and topology information



(b) Flow throughput comparison

Figure 3 Analytical vs. simulation results for TBCP

REFERENCES

[1] Haiyun Luo and Songwu Lu, "A topology-independent fair queueing model in ad hoc wireless networks," *IEEE ICNP'00*, pp. 325-335.

[2] H. Luo and S. Lu, "A self-coordinating approach to distributed fair queueing in ad hoc wireless networks,". *IEEE INFOCOM'01*, pp. 1370-1379.

[3] H. L. Chao, and W. Liao, "Fair scheduling with QoS support in ad hoc networks," *IEEE LCN*, '02.

[4] Y. Cao and V. O. K. Li, "Scheduling algorithms in broadband wireless networks," *Proc. of the IEEE*, vol. 89, no. 1, Jan. 2001

[5] T. S. Eugene Ng, I. Stoica and H. Zhang, "Packet fair queueing algorithm for wireless networks with location-dependent errors," *IEEE INFOCOM*, pp. 1130-1111, 1998.

[6] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. on Networking*, Vol. 7, No. 4, Aug 1999.

[7] S. Lee, K. Kim, and A. Ahmad, "Channel error and handoff compensation scheme for fair queueing algorithms in wireless networks," *IEEE ICC*, pp. 3128-3132, 2002.

[8] Y. Cao and V. O. K. Li, "Wireless packet scheduling for two-state link models," *IEEE Globecom*, Taipei, Taiwan Nov. 2002.

[9] P. Goyal, H. M. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Trans. on Networking*, vol. 5, no. 5, pp. 690-704, 1997.

[10] J. Yoon, M. Liu, and B. Noble, "Random waypoint considered harmful," *IEEE INFOCOM*, Apr. 2003.

[11] E. O. Elliot, "Estimates of error rates for codes on burst-noise channels," *Bell Systems Technical Journal* 42 (Sep. 1963) 1977-1997.

[12] R. K. Jain, D. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," *DEC Tech. Rpt. DEC-TR-301*, 1984.

Improving TCP Performance in Mobile Networks

Wanjiun Liao, Chang-Jung Kao, and Chin-Hei Chien
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
Email: wjliao@cc.ee.ntu.edu.tw

Abstract- **In this paper, a new transport layer mechanism is proposed to improve the performance of TCP in mobile networks. The proposed mechanism is comprised of two parts: a Loss Classifier (LC) and a Congestion Window Extrapolator (CWE). Based on LC, the cause of packet loss during roaming is determined. If the loss is considered to be caused by congestion in the wireline, the congestion window is halved; otherwise, the packet is considered to be lost in the last hop, the wireless portion, and the sender adjusts the size of the congestion window based on CWE. We conduct simulations to evaluate the performance of the proposed mechanism. The results show that our mechanism significantly improves TCP performance as compared to existing solutions for mobile networks.**

Keywords- **wireless TCP, mobile networks**

I. INTRODUCTION

TCP is a transport layer protocol providing reliable and ordered data service in the Internet. While TCP performs well in wired networks, when used in mobile wireless networks, TCP performance may degrade due to high bit error rates on the wireless link or temporary disconnections caused by handoffs. Much research effort has been expended to enable wireless or mobile TCP. Existing work on this subject can be classified into two categories: (1) approaches based on base station assistance, such as I-TCP [1], MTCP [2], Snoop [3], M-TCP [4], and WTCP [5], and (2) end-to-end approaches, such as Fast retransmission [6], Explicit Bad State Notification [7], and Freeze TCP [8].

In this paper, we focus on the performance problem with temporary disconnections caused by handoffs, and propose an end-to-end mechanism to improve the performance of TCP for roaming users in mobile networks. Such disconnections may further result in the following problems: (1) data loss during the handoff period, and (2) throughput degradation due to handoffs. To solve these problems, the proposed mechanism is comprised of two parts: a Loss Classifier (LC) and a Congestion Window Extrapolator (CWE). The LC targets the first issue and determines the reason of a packet loss during the handoff period; the CWE focuses on the second issue and improves the throughput of the connection after handoffs. Note that most of the existing work focuses on the enhancement of TCP performance over wireless networks at the receiver side, i.e., the operation is performed at the receiver side. In this paper, we discuss this problem from the perspective of the sender.

The rest of the paper is organized as follows. Sec. II describes the proposed mechanism. Sec. III shows the simulation results. Finally, the paper is concluded in Sec. IV.

II. PROPOSED MECHANISM

In this section, the proposed mechanism, comprised of a Loss Classifier (LC) and a Congestion Window Extrapolator (CWE), is described. Our mechanism is implemented at the mobile sender only. The receiver still runs TCP Reno.

A. Loss Classifier (LC)

The Loss Classifier is used to determine why packets are not received after handoffs. The loss of a packet during a handoff may be caused by congestion in the wireline, or by handoffs, or by transmission errors on the channel. In LC, transmission errors on the channels are treated in the same way as errors due to network congestion. In both cases, the congestion window of the TCP sender is shrunk to half of that of the normal TCP operation.

To determine the cause of a packet loss, two loss probabilities are calculated: P_H and P_C. P_H is the probability that a loss is caused by the handoff, and P_C the loss is due to congestion. If $P_C > P_H$, the mechanism regards the loss due to congestion, and the normal congestion control mechanism is initiated; otherwise, the loss is due to handoffs, and the sender can continue sending unsent packets in the usable window size based on the Congestion Window Extrapolator (CWE) which will be described later.

We briefly describe how P_H and P_C are calculated. When a timeout occurs, a period called Possible ACK returning Period (PARP) is calculated according to eqs. (1) and (2).

$$\begin{aligned} T_{Pb} &= \text{the beginning time of the PARP} \\ &= \text{Transmit\_time} + \text{SRTT} - \text{RTTVAR} \qquad (1) \\ T_{Pe} &= \text{the end time of the PARP} \end{aligned}$$

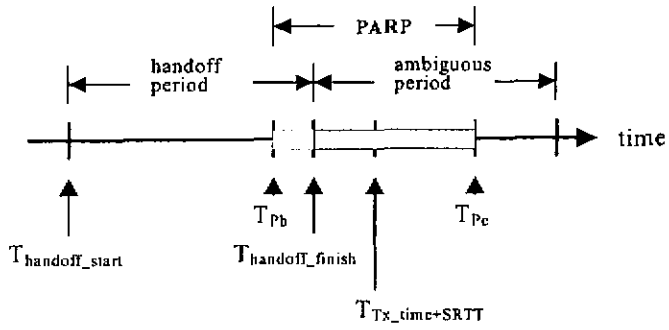


Figure 1. Illustration of PARP

$$= \quad \text{Transmit\_time} + \text{SRTT} + \text{RTTVAR} \qquad (2)$$

Here SRTT means smoothed RTT and RTTVAR means RTT variation. Fig. 1 illustrates PARP, which is the period from T_{Pb} to T_{Pe}. Note that after a handoff, there is a short period called the ambiguous period (e.g., 1xSRTT). During this period, the reason of packet losses is ambiguous.

To determine P_H, the sender checks if the PARP of a lost packet is within the handoff period. If the PARP falls before $T_{handoff\_finish}$, P_H is set to *Max* (i.e. 100%). If the PARP is completely beyond the handoff period, P_H is set to *Min* (i.e. 0%). If the PARP is in the ambiguous period, P_H is calculated as

$$P_H = \frac{T_{handoff\_finish} - T_{P_b}}{T_{P_e} - T_{P_b}}.$$

P_C is calculated based on the *congestion history* α, which is the ratio of RTT to SRTT (i.e. RTT = αSRTT). Whenever segment losses are detected, *excluding* the losses occurring in the ambiguous period, the congestion history, α, is calculated as $\alpha_{Sample} = \frac{RTT\_value}{SRTT}$, and $\alpha_{i+1} = w_c \times \alpha_i + (1 - w_c) \times \alpha_{Sample}$, where w_c is the weighting factor, e.g. 0.2. The RTT_value is set to an RTO value if the reason of segment loss is retransmission timeout; otherwise, the RTT_value is set to the "last available" sampled RTT.

If RTT is less than SRTT, P_C is set to *Min* (i.e. 0%); if RTT is greater than αSRTT, P_C is set to *Max* (i.e. 100%). Otherwise, P_C will be given by

$$P_C = \frac{RTT}{\alpha \times SRTT}.$$

B. Congestion Window Extrapolator (CWE)

The Congestion Window Extrapolator is used to determine how the congestion window changes after a handoff. If no packet is lost due to congestion and the sender has an empty usable window in the ambiguous period, the window size increases by the amount which should have occurred during the handoff. The detailed operation of CWE is described in Table I. Here an *acknowledgement interval* T_{ack} is calculated as $T_{ack,i+1} = w_a \times T_{ack,i} + (1 - w_a) \times T_{ack}$, where w_a is the weighting factor, $0 \le w_a \le 1$.

```
if ( (T_handoff_finish − T_handoff_start) > estimated_RTT )
        T_increase_interval = estimated_RTT ;
else
        T_increase_interval = T_handoff_finish − T_handoff_start ;

for ( i = 0; i < (int)( T_increase_interval / T_ack ); i++){
     if( cwnd_ < ssthresh_ ) /* slow start */
        cwnd_ += 1;
     else                    /* congestion avoidance */
        cwnd  += 1 / cwnd  ;  }
```

Table I. CWE

III. PERFORMANCE EVALUATION

In this section, we describe the simulations conducted to evaluate the performance of the proposed mechanism in a wireless network. The simulation environment is shown in Fig. 2. There are *n* routers between a base station (BS) and a wired receiver. The delay on each wired link is 50 ms. The proposed mechanism is implemented at the mobile sender, and the receiver runs TCP Reno. The mobile sender moves between BS1 and BS2 every 20 seconds. The data rates of the wireless links to BS1 and BS2 are set to B_{w1} and B_{w2}, respectively. The wireless delay D_w is variable. There is an FTP connection between the TCP sender and the receiver. The FTP connection lasts for the duration of the simulation. The value of each parameter is listed as follows: n=3, B_{w1} = 10Mbps, B_{w2} = 10Mbps, and D_w = 100 ms. The simulation results are

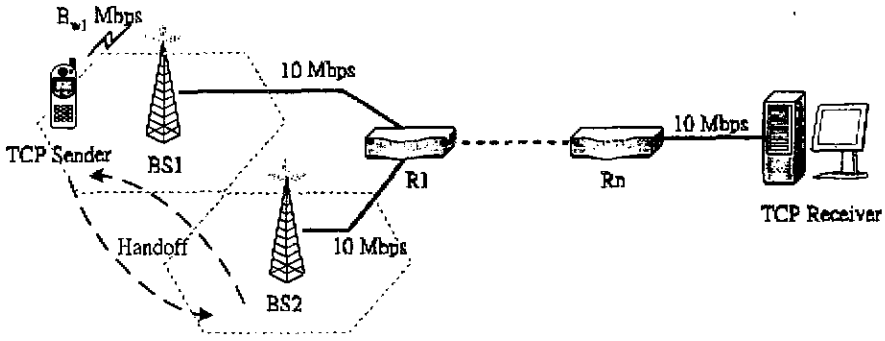


Figure 2. Simulation environment

generated using simulation tool ns-2. We compare the following mechanisms in the simulation: TCP Reno and Fast Retransmit with Freeze timer.

In Fig. 3, the proposed mechanism is compared with TCP Reno. Fig. 3 (a) plots the congestion windows of the proposed mechanism and TCP Reno as a function of time. The curves of Reno and the proposed mechanism are plotted after a handoff (at simulation time 206 sec). The curve marked "no-handoff" is only for comparison, showing the congestion window of a Reno connection when no handoff occurs. The proposed mechanism performs as if the handoff has never occurred, while Reno degrades when a handoff occurs. We then compare the throughputs of the proposed mechanism and TCP Reno. To better see the performance improvement of the proposed mechanism over TCP Reno, we define the performance "gain" as

$$Gain = \frac{Throughput_{proposed\_TCP} - Throughput_{Reno\_TCP}}{Throughput_{Reno\_TCP}} \times 100\%$$

and plot the gain curve of each component over TCP Reno with different handoff periods in Fig. 3 (b). The one with diamonds corresponds to LC only (the middle one), the curve with rectangles is based on CWE only (the bottom one), and the curve with triangles accounts for both (the top one). We see that LC is more important than CWE in the proposed mechanism in terms of the throughput improvement for mobile TCP connections. The three curves of performance gains all exceed one, indicating that our mechanism outperforms TCP Reno.

Fig. 4 shows the hit ratio of LC, i.e., the percentage of "right guess" for packet losses by LC. Here T_{bad} indicates the duration in which the wireless channel is in the bad state in every 10 sec. When it is in the bad state, the wireless delay is set to twice the original delay of the link. The figure depicts that LC gives the right classification of packet losses over 95% of the time on the average. The wrong "guess" of packet losses being classified as congestion is mainly due to our treating transmission errors on the links in the same way as congestions in LC. Fig. 5 compares the performance of three mechanisms, including TCP Reno, the proposed mechanism, and a TCP variant having both freeze timer and fast-transmit mechanism. Since Freeze TCP and existing work focus on the receiver side, we implement the combined mechanism of Freeze TCP [8] with Fast Transmit [6] at the sender and compare it with our mechanism. The curve with squares corresponds to the proposed mechanism, the one with diamonds is the combined freeze timer and fast transmit, and the one with triangles is TCP Reno. Again, varying the handoff period from 0 to 3 sec, we see that our mechanism outperforms both TCP Reno and Fast Retransmit with Freeze Timer.

IV. Concluding Remarks

In this paper, we have proposed an end-to-end mechanism which improves the performance of TCP for roaming users in mobile networks. The proposed mechanism is comprised of a Loss Classifier (LC) and a Congestion Window Extrapolator (CWE). LC determines why packets are not received after handoffs, and CWE determines how to adjust the congestion window after handoffs. We have conducted simulations to evaluate the performance of our mechanism. The results show that the proposed mechanism significantly improves TCP performance in mobile networks.

References

[1] A. Bakre and B. R. Badrinath, "I-TCP: Indirect TCP for Mobile Hosts," Proc. of the 15th International Conference on Distributed Computing Systems, June 1995.

[2] I. Rhee, N. Balaguru, and G.N. Rouskas, "MTCP: Scalable TCP-like Congestion Control for Reliable Multicast," IEEE INFOCOM '99, 1999.

[3] H. Balakrishnan, S. Seshan, and R.H. Katz, "Improving Reliable Transport and Handoff

Performance in Cellular Wireless Networks," ACM Wireless Networks, Dec. 1995.

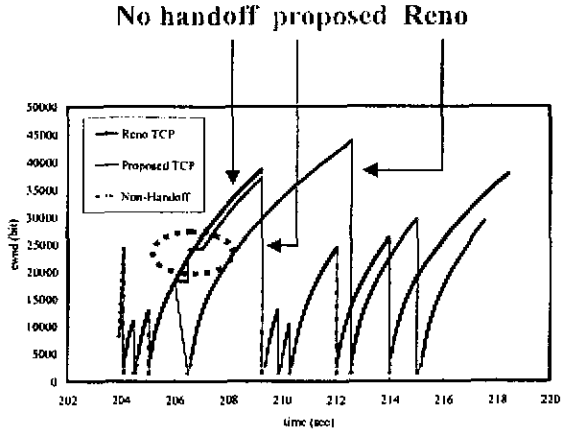
[4] K. Brown and S. Singh, "M-TCP: TCP for Mobile Cellular Networks," ACM Computer Communications Review, vol. 27, 1997.

[5] P. Sinha, et al., "WTCP: A Reliable Transport Protocol for Wireless Wide-Area Networks", Mobicom'99, 1999.

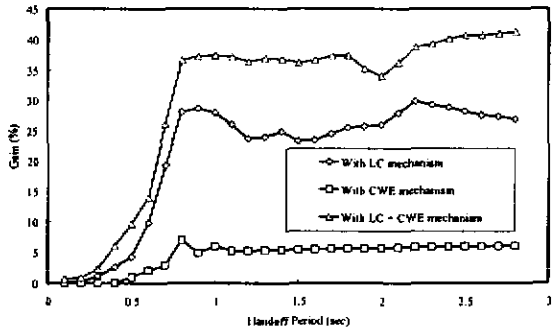
[6] R. Caceres and L. Iftode, "Improving the Performance of Reliable Transport Protocols in Mobile Computing Environments," IEEE Journal on Selected Areas in Communications, June 1995.

[7] B. S. Bakshi, P. Krishna, N. H. Vaidya, and D. K. Pradhan, "Improving Performance of TCP over Wireless Networks," International Conference on Distributed Computing Systems, 1997.

[8] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta, "Freeze-TCP: A True End-to-End TCP Enhancement Mechanism for Mobile Environments," IEEE INFOCOM '00, 2000.



(a) Congestion window



(b) Throughput gain

Figure 3. TCP Reno vs. proposed TCP

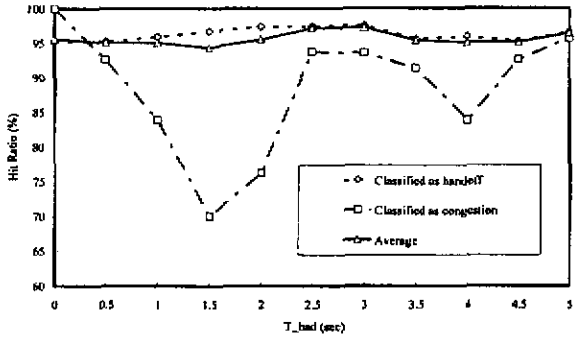


Figure 4. The hit ratio of LC

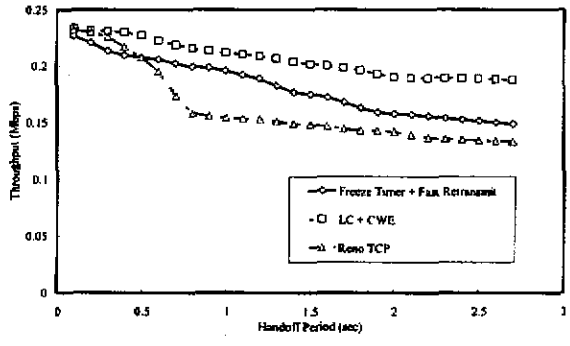


Figure 5. Comparison of three mechanisms

The Behavior of TCP over DOCSIS-based CATV Networks

Wanjiun Liao

Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
Email: wjliao@cc.ee.ntu.edu.tw

Abstract This paper studies the impact of the DOCSIS MAC mechanism on the performance of TCP in Hybrid Fiber Coax (HFC) networks. We define an asymmetry ratio for DOCSIS systems, taking into account the media access control mechanism in the upstream direction. Based on the new asymmetry ratio, the round trip delay of sending a data packet is derived for both one-way and two-way TCP transfers. We then verify the analytical results of the behavior of TCP by simulation. The results show a good match between our analysis and the simulation.

Keywords: TCP, HFC, DOCSIS, CATV, asymmetry ratio

1. INTRODUCTION

This paper studies the behavior of TCP over DOCSIS-based Hybrid Fiber Coax (HFC) Networks, a popular choice for broadband access networks. An HFC network is a tree-and-branch network. Each branch is comprised of a downstream channel and an upstream channel. The downstream channel is a point-to-multi-point, broadcast channel, and the upstream channel is a multipoint-to-point, shared channel. The only transmitter on the downstream channel is the Cable Modem Termination System (CMTS) at the Head End, and all the Cable Modems (CMs) connected to the channel are receivers. For the upstream channel, all the CMs are transmitters and the CMTS is the only receiver. To arbitrate random access to the upstream channel, a media access control mechanism is required. The media access control (MAC) mechanism in HFC may follow the Data-Over-Cable Service Interface Specifications (DOCSIS) of the Multimedia Cable Network System Partners (MCNS) [1] or IEEE 802.14 [2]. Since DOCSIS is the *de facto* standard in the cable industry, we will focus on DOCSIS's operation in this paper.

In DOCSIS, an upstream channel is modeled as frames of mini-slots. Each frame is comprised of contention minislots and data minislots. The details of each frame are specified via control messages,

called MAP, periodically transmitted on the downstream channel by the CMTS. In a MAP, a variable number of information elements (IEs) are specified, each of which defines a range of mini-slots to be used. The number of IEs, and the corresponding mini-slots, varies from MAP to MAP. Typically, each MAP contains a Request IE, some Data Grant IEs, a Null IE, and some management information. The Request IE specifies the contention interval of the next frame; a Data Grant IE describes the transmission interval for a CM to transmit packets, and the Null IE terminates the assignment list. Each MAP specifies the details of the next frame, and thus it must be received by all CMs before the next frame starts.

Each time a CM has a packet to transmit, it requests the assignment of a Data Grant IE in a later MAP. If it is a backlogged CM (i.e., a CM with a non-empty packet queue) and is assigned a Data Grant IE in the newly received MAP, the CM piggybacks the request on an outgoing packet using the assigned Data Grant IE in the data minislots; otherwise, the CM contends for the use of the upstream channel via contention minislots. Each contention request occupies one mini-slot. If several CMs compete for one mini-slot in the contention interval, all the requests are corrupted due to collision. After sending a request, each CM waits for a Data Grant IE or a Data Pending IE in the next MAP. Once either is received, the contention resolution is complete. Each CM can use the Data Grant IE to transmit a data packet if assigned, or keep waiting for a Data Grant IE and send no further request if a Data Pending IE is received. The CM regards the contention request as lost if neither a Data Grant IE nor a Data Pending IE is received. The CM then increases the back-off window (i.e., contention window) by a factor of two, as long as the window size is less than the maximum back-off window. The CM randomly selects a number within its new back-off window, pauses until the timer expires, and then repeats the contention. This retry process continues until the maximum number of retries has been reached, at which time the packet must be discarded.

The design issues of HFC networks have been widely studied for years. An efficient MAC layer scheduling and allocation algorithm for the upstream channel of HFC networks is proposed in [3-6].

The performance of HFC networks is evaluated in [7-9]. The flow control mechanism of TCP is tuned in [10-13] to improve the performance of TCP over asymmetric networks. The performance of TCP in HFC networks is investigated in [14-17]. The impact of DOCSIS on the performance of TCP is studied in [18-19].

Previous studies [10-13] have shown that the performance degradation of TCP operating in asymmetric networks is due to bandwidth asymmetry of downstream and upstream channels. This paper further investigates the degradation problem but focuses on DOCSIS-based HFC networks, an asymmetric network which needs media access control mechanisms to arbitrate random access among multiple hosts. We discuss the impact of the DOCSIS MAC mechanism on the performance of TCP, and analyze the behavior of TCP operating over DOCSIS. Our analysis is based on TCP Reno, plus delayed ACK, for congestion control.

The rest of the paper is organized as follows. Section 2 identifies the performance problem of TCP in DOCSIS-based HFC networks and defines an asymmetry ratio for DOCSIS systems. Section 3 analyzes the behavior of one-way TCP transfers over the MAC layer of DOCSIS. Section 4 analyzes two-way TCP transfers over the DOCSIS MAC layer. Section 5 shows simulation results to verify the analysis. Finally, concluding remarks is included in Section 6.

2. PROBLEM STATEMENT

In HFC networks, a CMTS controls multiple branches, each of which is typically comprised of one downstream channel and one upstream channel and connects 200 to 500 CMs. Each branch works identically but independently. Without loss of generality, we consider one branch only in the analysis.

2.1 Notations and Assumptions

To better explanation, some terms are defined in this paper.

(1) The delivery offset of a MAP (denoted as D_{MAP}) is defined as the one-way propagation delay

between the CMTS and each CM plus some extra delay, as shown in Fig. 1. This time is spent for the CMTS to inform all CMs of the slot allocation details for the next frame. Thus, each MAP should be sent D_{MAP} seconds earlier before the next frame starts.

(2) Upstream user service time (denoted as T_{usv}) is defined as the delay between sending two consecutive packets in the upstream buffer at a CM, as shown in Fig. 1.

(3) Access delay (denoted as T_{ad}) is defined as the time between when a packet is received by a CM and when the CM starts sending this packet out to the channel.

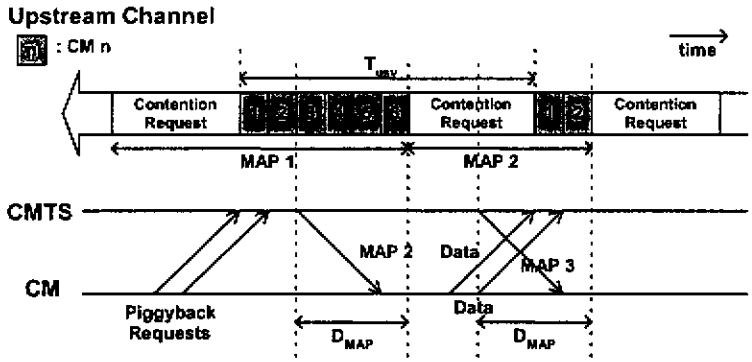


Figure 1. The MAC mechanism of DOCSIS

The notations used in the analysis are summarized as follows.

C_d	Capacity of the downstream channel, in bps
C_u	Capacity of the upstream channel, in bps
L_{data}	Data packet length, in bits
L_{ack}	ACK packet length, in bits
B_{CM}	Buffer size (i.e., backlog) at a CM, in terms of number of packets
T	One-way propagation delay between the CMTS and CMs, in seconds
N_{dCM}	Number of CMs performing TCP transfers in the downstream direction
N_{uCM}	Number of CMs performing TCP transfers in the upstream direction
t_{ms}	One mini-slot time , in seconds
N_{frame}	Total number of mini-slots granted in a MAP for a frame
N_c	Number of contention mini-slots described in a MAP
N_{u_ack}	Number of mini-slots for one ACK packet

N_{u_data} Number of mini-slots for one TCP data packet

T_{ad} Access delay, in seconds

T_{usv} Upstream user service time, in seconds

D_{MAP} Delivery offset of a MAP, in seconds

N_{p_REQ} Maximum number of pending requests received during a D_{MAP}

We make the following assumptions:

1. We only consider TCP traffic in the analysis. The result can serve as the lower bound of the performance when UDP and TCP traffic co-exists on the channel.
2. The propagation delay between the CMTS and all CMs are assumed the same.
3. In DOCSIS, multiple outstanding MAPs are allowed. Here we restrict the number of outstanding MAPs to one to simplify the analysis. Nevertheless, the analytical approach is still applicable to the case with multiple MAPs outstanding.
4. The number of contention minislots in a DOCSIS frame, i.e., *Nc*, is a constant.
5. The contention minislots described in a MAP is assumed larger than the delivery offset of the MAP, i.e., $N_c \times t_{ms} > D_{MAP}$.

2.2 The Effect of Bandwidth Asymmetry on the Performance of TCP

Asymmetric networks, such as HFC and xDSL, are networks with different channel bandwidths in the downstream and upstream directions. The main effect of bandwidth asymmetry on the performance of TCP is that the ACK clocking may be disrupted [10-13]. To better understand the behavior of TCP in asymmetric networks, an asymmetry ratio k [10] is defined as follows.

$$k = \frac{\text{forward channel bandwidth}}{\text{reverse channel bandwidth}} \times \frac{\text{ACK packet length}}{\text{data packet length}} = \frac{C_d}{C_u} \times \frac{L_{ack}}{L_{data}}. \quad (1)$$

The physical meaning of this ratio is explained as follows. TCP behaves normally when k is less than or equal to one. When this ratio exceeds one, the network is asymmetric. Operating TCP over an asymmetric network makes ACK packets arrive at the bottleneck link in the reverse direction at a rate

faster than the bottleneck link can support, and leads to two consequences. First, the time spacing between consecutive ACKs received by the sender is larger than the symmetric case. As a result, the sender clocks out data at a slower rate, and slows down the growth of the congestion window. This in turn degrades the throughput in the downstream direction. Second, the buffer in the reverse bottleneck link is filled up with ACK packets rapidly. A full buffer may cause serious ACK drops. This may lead to a bursty sender because one ACK may acknowledge several TCP data packets in an unpredictable way. In addition, fewer ACK packets received may slow down the growth of the sender's congestion window because TCP increases its congestion window by counting the number of ACKs received.

2.3 The Asymmetry Ratio for DOCSIS Systems

To examine the adequacy of the expression k in (1) to describe the behavior of TCP over DOCSIS-based HFC networks, we estimate the performance of TCP using the parameters in Table 1. The bandwidths of the downstream and upstream channels are set to 26.97 Mbps and 2.56 Mbps, respectively. Considering the delayed ACK policy of TCP Reno (i.e., sending one ACK packet to acknowledge the receipt of d data packets), the expression of k is modified as $k = \frac{C_d}{C_u} \times \frac{L_{ack}}{L_{data}} \times \frac{1}{d}$. Using the packet lengths of ACK = 64 bytes and Data = 1024 bytes, and d = 2, we obtain $k = \frac{26.97}{2.56} \times \frac{64}{1024} \times \frac{1}{2} \cong 0.33$. Since $k \leqq 1$, this network should be symmetric for TCP, i.e., TCP is supposed to behave normally.

We then verify the behavior of TCP in such a network by simulation using ns-2 [18], again based on the parameters listed in Table 1. We observe the influence of different number of downloading CMs (i.e., N_{dCM}) on the aggregate downstream throughput and the backlog in the upstream buffer (which buffers ACK packets) at the CM. Fig. 2 plots the aggregate downstream throughput of the network, varying N_{dCM} from 1 to 50, and Fig. 3 shows the respective buffer size (i.e., backlog). Interestingly, we find that when N_{dCM} is small (e.g., less than 13 in this example), the network behaves like an asymmetric network, i.e., low downstream throughput and full upstream ACK queues. However, as N_{dCM} becomes

large (e.g., exceeds 13 in this example), TCP starts behaving normally. This implies that the network experiences bandwidth asymmetry (i.e., the asymmetry ratio is greater than one) when N_{dCM} is small, and the performance degrades in this region. However, according to the estimation based on k in (1), the network should be symmetric. Thus, we learn that while k in (1) captures well the behavior of TCP in point-to-point asymmetric links, it cannot adequately explain the behavior of TCP operating over DOCSIS-based HFC networks, an asymmetric network which needs a media access control mechanism to arbitrate random access among end hosts. To better capture the abnormal behavior of TCP over DOCSIS-based HFC networks, we need to consider the mini-slot reservation-based random access in the upstream direction of DOCSIS systems.

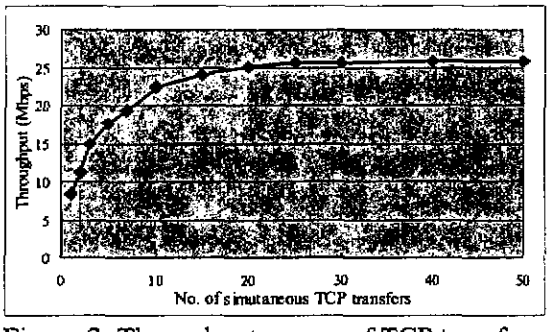


Figure 2. Throughput vs. no. of TCP transfers

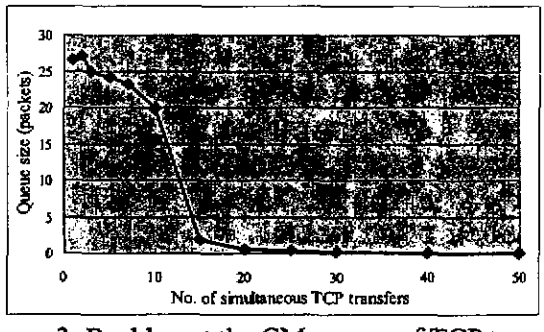


Figure 3. Backlog at the CM vs. no. of TCP transfers

We revise the asymmetry ratio k in (1) as η in (2) for DOCSIS networks, considering the MAC layer operation of DOCSIS.

$$\eta = \frac{C_d \times T_{usv}}{d \times L_{data} \times N_{dCM}} \tag{2}$$

This new ratio is obtained as follows. Considering the upstream channel in HFC networks being a TDMA-like channel, $\frac{L_{ack}}{C_u}$ in (1) is replaced with T_{usv}, where T_{usv} is the duration that a CM should wait for its scheduled time. For downstream transmissions, $\frac{C_d}{L_{data}}$ in (1) is shared by all simultaneous TCP transfers, and should be replaced with $\frac{C_d}{L_{data} \times N_{dCM}}$. We also consider the delayed ACK policy. Thus,

we multiply $1/d$ by $\frac{C_d \times T_{usv}}{L_{data} \times N_{dCM}}$ to obtain η.

In the following, the impact of the DOCSIS MAC mechanism on the performance of TCP is studied in terms of the round trip delay (RTT) of each TCP transfer. This RTT in turn determines the throughput of the TCP connection in DOCSIS networks.

3. ONE-WAY TCP TRANSFER

We first examine one-way TCP transfers, i.e., all CMs download TCP traffic on the downstream channel. Thus, all data packets go on the downstream channel and all ACK packets go on the upstream channel.

3.1 The Effect of the DOCSIS MAC Mechanism on Bandwidth Asymmetry

In DOCSIS, the MAP for a frame should be received by all CMs before the frame starts. This may cause that some requests, especially piggybacked ones, in the current frame have arrived and been processed at the CMTS after the MAP for the next frame is sent. The late requests will be pending and become backlogs for the next second frame. These pending requests plus new requests which arrive at the CMTS during the next frame will be waiting to be granted in the next MAP. Each MAP can describe at most 2048 minslots. Compared to the MAP limitation of 2048 mini-slots, the ACK packet length is relatively small. To simplify the analysis, we assume that the mini-slot limitation of a MAP will not be exceeded when the CMTS grants all the waiting requests. In fact, even if the limitation is exceeded, the analytical approach is still applicable because the global ordering of data grants placed in MAPs remains intact. Recall that contention minislots is assumed larger than the delivery offset of a MAP. Thus, the pending requests will be granted in the next MAP, not in later MAPs. Since the packet length of an ACK is fixed, the maximum number of pending requests, N_{p_REQ}, (i.e., those arriving at the CMTS during a D_{MAP}) in terms of number of minislots can be expressed as $N_{p\_REQ} = \left\lfloor \frac{D_{MAP}}{t_{ms}} \times \frac{1}{N_{u\_ack}} \right\rfloor$.

The value of T_{usv}, can be calculated in two mutually exclusive, collectively exhaustive cases.

Case 1: $N_{dCM} \leqq 2N_{p\_REQ}$

When $N_{dCM} \leqq N_{p\_REQ}$, no piggybacked requests are granted in the next MAP. On the average, each CM receives a data grant in every other frame, i.e. there are N_{dCM} requests granted every two frames. Thus, T_{usv} can be expressed as

$$T_{usv} = \left(2N_c + N_{dCM} N_{u\_ack}\right) t_{ms}, \tag{3}$$

where $N_{u\_ack} = \left\lceil \frac{L_{ack}}{C_u} \times \frac{1}{t_{ms}} \right\rceil$.

Case 2: $N_{dCM} > 2N_{p\_REQ}$

If $N_{dCM} > 2N_{p\_REQ}$, some CMs may obtain their data grants immediately in the next frame. When piggybacked requests arrive at the CMTS before the next MAP is sent, these requests will be scheduled after those pending requests arriving in the current frame. These "lucky" piggybacked requests are scheduled after the pending requests in the next frame. Those requests placed in the last N_{p_REQ} data grants of the next frame become the pending requests of the next second frame. As such, all CMs take turn to be the "lucky" ones in each frame, and the average data grants each CM can obtain in every other frame is more than one. Thus, T_{usv} can be expressed as

$$T_{usv} = \frac{\left[N_c + \left(N_{dCM} - N_{p\_REQ}\right) N_{u\_ack}\right] t_{ms} N_{dCM}}{N_{dCM} - N_{p\_REQ}}. \tag{4}$$

Substituting (3) and (4) into (2), we can derive the asymmetry ratio η as follows:

1. $N_{dCM} \leqq 2N_{p\_REQ}$

$$\eta = \frac{C_d}{d \times L_{data}} \times \frac{(2N_c + N_{u\_ack} N_{dCM}) t_{ms}}{N_{dCM}} \tag{5}$$

2. $N_{dCM} > 2N_{p\_REQ}$

$$\eta = \frac{C_d}{d \times L_{data}} \times \frac{\left[N_c + (N_{dCM} - N_{p\_REQ}) N_{u\_ack}\right] t_{ms}}{N_{dCM} - N_{p\_REQ}} \tag{6}$$

Note that since only ACK packets are transmitted on the upstream channel, all pstream packets have the same length. As a consequence, the number of mini-slots for each data grant (i.e., N_{u_ack}) is fixed, and the number of pending requests (i.e., N_{p_REQ}) is also a fixed value. In addition, we fix the number of contention mini-slots (i.e., N_c). Given the values of the downstream and upstream channel capacities and TCP packet length, the asymmetry ratio η varies only with the number of CMs transferring simultaneously (i.e., N_{dCM}). Furthermore, from (5) and (6), η decreases as N_{dCM} increases. This implies that as the number of simultaneous transfer increases, the network asymmetry decreases, matching the observation we have in Figs. 2 and 3. From (6), when N_{dCM} becomes very large, η will be approximated as $\frac{C_d}{d \times L_{data}} \times t_{ms} N_{u\_ack}$. Substituting $N_{u\_ack} = \left\lceil \frac{L_{ack}}{C_u} \times \frac{1}{t_{ms}} \right\rceil$ into this approximation, we obtain η as

$$\eta = \frac{C_d}{d \times L_{data}} \times \frac{L_{ack}}{C_u}. \tag{7}$$

Interestingly, η in (7) is equal to k in (1)[1], just as if the DOCSIS MAC mechanism has not been taken into account. Therefore, we can conclude that when the number of simultaneous transfers becomes very large, the effect caused by the DOCSIS MAC mechanism on TCP becomes negligible. At this point, the value of η is determined by the asymmetry ratio of the pair of channels and the packet size ratio of TCP and ACK packets. However, when there is only one CM performing TCP transfers in the downstream direction, the asymmetry ratio reaches its maximum value. This maximum value of η may be much larger than one, rendering the bandwidths rather asymmetric.

[1] Here we mean (1) with the consideration of delayed ACK, i.e., $k = \frac{C_d}{C_u} \times \frac{L_{ack}}{L_{data}} \times \frac{1}{d}$.

3.2 The Effect of DOCSIS MAC Layer on TCP Round Trip Delay

We now analyze the round trip delay of a TCP packet in DOCSIS systems. To send an ACK packet, the CM needs to send a request first and wait for a data grant allocated by the CMTS. The average round trip time is then given by

$$RTT = 2T + \frac{L_{data}}{C_d} + \frac{L_{ack}}{C_u} + T_{ad}. \tag{8}$$

Access delay, T_{ad}, of a packet is defined as the time between when the packet is received by the CM from the upper layer and when the CM starts sending it out. The packet may experience a queueing delay if the buffer at the CM is non-empty. Every time the CM wants to send an ACK packet, it will first issue a request. Then, it will wait for the MAP that contains the data grant specifying the period of time it is allowed to send. This needs two times more than the propagation delay between the CMTS and the CM (i.e., one propagation delay for the request, one for the MAP, plus the processing delay at the CMTS). Later at the scheduled time, it will send the head-of-line packet from the buffer.

The asymmetry ratio η greatly affects the value of T_{ad}. If $\eta > 1$, the upstream ACK buffer will always stay full. All requests experience long queuing delay and are sent piggybacked. If $\eta \leq 1$, a request will often find an empty queue, and will be sent via contention. In the following, we calculate T_{ad} in these two cases.

Case 1: asymmetry ratio $\eta > 1$

When η is larger than one, the upstream buffer will build up rapidly with ACK packets and stay full constantly. Due to the piggyback mechanism, the order of data grants placed in the consecutive MAPs remains intact until at least one CM has no more packets to send. The distribution of the upstream user service time for each packet will almost always remain unchanged, and each packet in the upstream buffer experiences a delay of ($B_{CM} \times T_{usv}$). Thus, the access delay is expressed as

$$T_{ad} = B_{CM} \times T_{usv}, \tag{9}$$

where B_{CM} is the upstream buffer size (i.e., backlog) at the CM in terms of the number of ACK packets.

Substituting (9) to (8), we obtain the TCP RTT for the asymmetric case.

Case 2: asymmetry ratio $\eta \leq 1$

If $\eta \leq 1$, the upstream channel is not the bottleneck anymore and the upstream buffer will not stay full. The buffer size may drop to zero rapidly, and the requests for sending ACK packets may not be piggybacked to the CMTS. Assume that there is no collision in the contention minislot. The mean waiting time of a request is about half a frame time. When a request arrives at the CMTS, it should wait until the next MAP is issued, at which time the data grant will be assigned. Later the CM will receive the MAP with its data grant assigned. It will keep waiting until its turn to send. On the average, this waiting time is about another half a frame time. Thus, the average access delay on the upstream channel can be approximately given by two frame time, i.e.,

$$T_{ad} = 2N_{frame}t_{ms}. \tag{10}$$

The total number of mini-slots in a frame is determined as follows. ACK packets use only a portion of bandwidth on the upstream channel: $\frac{L_{ack}}{d \times L_{data}} C_d$. The rest of the bandwidth is used by the contention mini-slots. The average number of mini-slots used by ACK packets is then given by $\frac{N_c \times \frac{L_{ack}}{d \times L_{data}} C_d}{C_u - \frac{L_{ack}}{d \times L_{data}} C_d}$.

Thus, on the average, the total number of mini-slots in a frame is given by

$$N_{frame} = N_c + \frac{N_c \times \frac{L_{ack}}{d \times L_{data}} C_d}{C_u - \frac{L_{ack}}{d \times L_{data}} C_d}. \tag{11}$$

Substituting (11) to (10) and then to (8), we obtain the TCP RTT for the symmetric case.

□

To reflect the impact of these two cases of η on RTT, the expression in (8) should be revised as

$$RTT = w \times RTT_a + (1-w) \times RTT_s, \tag{12}$$

where RTT_a is the round trip delay calculated based on $\eta > 1$, RTT_s is based on $\eta \le 1$, and w is a weighting factor with $w = \begin{cases} 1, & \eta > 2 \\ 1 - e^{\frac{1-\eta}{\tau}}, & 1 < \eta \le 2 \\ 0, & \eta < 1 \end{cases}$. The interpretation of this setting will be described later in the numerical example section.

4 TWO-WAY TCP TRANSFERS

This section studies two-way TCP transfers over DOCSIS-based HFC networks. With two-way transfers, some clients download TCP traffic (i.e., do TCP transfers in the downstream direction), and some upload TCP traffic (i.e., transfer TCP traffic in the upstream direction). Thus, both data and ACK packets can go in either direction.

4.1 The Effects of DOCSIS MAC Layer on Bandwidth Asymmetry

Since the network has asymmetric bandwidths and different operations for both channels, downloading CMs experience network asymmetry differently from uploading CMs. In the following, we discuss the asymmetry ratios for clients doing TCP transfers in these two directions.

A) Downstream TCP transfers

With two-way TCP transfers, both data and ACK packets can be transmitted on the upstream channel. These two kinds of packets are very different in size. ACK packets are fixed and small, and TCP data packets are variable and large. Typically, the data packet is ten times larger than the ACK packet. Thus, the data packet transmission time, $\frac{L_{data}}{C_u}$, may be larger than the delivery offset of a MAP, D_{MAP}. When a MAP is ended with a data grant for a long data packet, it is very likely that no pending requests will arrive at the CMTS in the frame described by the MAP. However, in practice, data grants

may be ordered in a MAP rather randomly. Consequently, it is hard to accurately determine the number of pending requests in each frame. We can only conclude that if there are pending requests, they are usually from ACK packets.

Moreover, both data and ACK packets can be transmitted in the downstream direction. Only a portion of the downstream capacity is used to transmit data packets. Considering the "delayed ACK" policy, the asymmetry ratio of downstream TCP transfers can be expressed as

$$\eta = \frac{N_{dCM} L_{data}}{N_{dCM} L_{data} + \frac{N_{uCM} L_{ack}}{d}} \times \frac{C_d}{d \times L_{data}} \times \frac{T_{usv}}{N_{dCM}}. \quad (13)$$

The first term in (13) (i.e., $\frac{N_{dCM} L_{data}}{N_{dCM} L_{data} + \frac{N_{uCM} L_{ack}}{d}}$) is close to one, because (i) the ACK packet length is relatively small and (ii) the "delayed ACK" policy is used to reduce the number of ACK packets to be sent. Thus, transmitting ACK packets in the downstream direction has very minor impact on η due to only a small portion of the channel capacity being used for the ACK packets. The MAP limitation of 2048 mini-slots, however, may easily be exceeded in the two-way transfer case. Fortunately, the MAP limitation has very minor impact on η, thanks to the global ordering of data grants placed in MAPs remaining intact. The only difference is that the data grants unable to fit in one MAP should be put in the next MAP. To save space, we assume this limitation is not exceeded although the analytical approach is still applicable to handle the case when it is exceeded.

Based on the analysis in Sec.3.1 (i.e., (5) and (6)), η is highly dependent on the relationship between N_{dCM} and $2N_{p\_REQ}$. Besides, the backlog in the upstream buffer is usually non-zero in the two-way transfer case. Due to the uncertain number of pending requests, we can only determine the upper and lower bounds of η instead of the deterministic value as in the one-way scenario. Using the same approach as in Sec. 3.1, we can derive the two bounds of T_{usv} for downstream transfers in two cases.

$$T_{usv} = (N_c + N_{uCM} N_{u\_data} + N_{dCM} N_{u\_ack})\, t_{ms}. \quad (17)$$

Substituting (16) and (17) into (13), the two bounds of η can be obtained accordingly.

Data grants placed in a MAP may be rather random. Suppose that no pending requests are caused by the MAP limitation. Once a long data grant (i.e., satisfying $N_{u\_data} \times t_{ms} > D_{MAP}$) is placed in the end of a MAP, all requests which arrive in the frame described by the MAP will be able to obtain their data grants in the next MAP. Thus, we conclude that while the worst case may happen, the network tends to operate in the best case in the steady state. We will demonstrate such an effect in the simulation section.

When the numbers of downloading and uploading CMs are equal and very large, the lower bound of η can be approximated as

$$\eta \cong \frac{N_{dCM} L_{data}}{N_{dCM} L_{data} + \frac{N_{dCM} L_{ack}}{d}} \times \frac{C_d}{d \times L_{data}} \times \frac{(N_{dCM} N_{u\_data} + N_{dCM} N_{u\_ack})\, t_{ms}}{N_{dCM}} = \frac{C_d}{d \times C_u}. \quad (18)$$

In this case, η just reflects the inherent bandwidth asymmetry in HFC networks.

B) Uploading TCP transfers

From the viewpoint of uploading CMs, the upstream user service rate (i.e., the inverse of the upstream user service time) is their packet transmission rate. In addition, the upstream service rates (in terms of packets per sec) of uploading CMs are equal to those of downloading CMs. Thus, η of uploading CMs is

$$\eta = \frac{1}{d \times T_{usv}} \times \frac{N_{uCM} \times L_{ack}}{C_d} \times \frac{N_{dCM} L_{data} + \frac{N_{uCM} L_{ack}}{d}}{N_{uCM} L_{ack}} \quad (19)$$

where T_{usv} is the upstream user service time of the uploading CM.

Multiplying η in (19) with η in (13) yields a constant of $1/d^2$. When the asymmetry ratio of downloading CMs increases, the asymmetry ratio of upstream CMs decreases. Due to the nature of

Case 1: $N_{dCM} \leqq 2N_{p\_REQ}$

First, we determine the asymmetry ratio η for the case of $N_{dCM} \leqq 2N_{p\_REQ}$, regardless of the number of uploading CMs. The worst case T_{usv} of downloading CMs occurs when all data grants for ACK packets are placed in the end of a MAP. In this case, the piggybacked requests of downloading CMs are all pending. Thus, the worse case T_{usv} of downloading CMs is given by

$$T_{usv} = (2N_c + 2N_{uCM}N_{u\_data} + N_{dCM}N_{u\_ack}) \times t_{ms}, \quad (14)$$

where $N_{u\_data} = \left\lceil \frac{L_{data}}{C_u} \times \frac{1}{t_{ms}} \right\rceil$ and $N_{u\_ack} = \left\lceil \frac{L_{ack}}{C_u} \times \frac{1}{t_{ms}} \right\rceil$.

The best case T_{usv} of downstream TCP transfers occurs when the data grant of any long TCP packet is put in the end of a MAP, where the packet length is long enough to satisfy the condition of $N_{u\_data} \times t_{ms} > D_{MAP}$. In this case, all piggybacked requests can obtain their data grants in the next MAP. Thus, the best-case T_{usv} is given by

$$T_{usv} = (N_c + N_{uCM}N_{u\_data} + N_{dCM}N_{u\_ack}) \times t_{ms}. \quad (15)$$

Substituting (14) and (15) into (13), the two bounds of η can be obtained accordingly.

Case 2: $N_{dCM} > 2N_{p\_REQ}$

We determine η for the case of $N_{dCM} > 2N_{p\_REQ}$, regardless of the number of uploading CMs. The worst case T_{usv} of downstream TCP transfers occurs when more than N_{p_REQ} data grants for ACK packets are put in the end of a MAP. Thus, the worst case T_{usv} of downloading CMs is

$$T_{usv} = \frac{[N_c + N_{uCM}N_{u\_data} + (N_{dCM} - N_{p\_REQ})N_{u\_ack}]\, t_{ms}N_{dCM}}{(N_{dCM} - N_{p\_REQ})}. \quad (16)$$

As in case 1, the best case T_{usv} occurs when the data grant of a long data packet is put in the end of a MAP and the packet size is long enough to satisfy the condition of $N_{u\_data} \times t_{ms} > D_{MAP}$. Thus, the best case T_{usv} is

bandwidth asymmetry in HFC networks, downstream TCP transfers usually have much larger asymmetry ratios.

4.2 The Effects of Bandwidth Asymmetry on TCP Round Trip Delay

The relationship between the TCP round trip delay and the asymmetry ratio for two-way transfers is the same as the one-way case (see eqs. 8-11), except that the downloading CMs mostly operate as in asymmetric networks. The average round trip delay of sending a packet in the two-way case is

$$RTT = 2T + T_{trans} + T_{od}, \text{ where } T_{trans} = \begin{cases} \frac{L_{data}}{C_d} + \frac{L_{ack}}{C_u}, \text{for downstream traffic} \\ \frac{L_{data}}{C_u} + \frac{L_{ack}}{C_d}, \text{for upstream traffic.} \end{cases} \quad (20)$$

The long RTT experienced by downloading CMs in the two-way case is due to long TCP packets going upstream. Long round trip delay in turn makes the TCP congestion window grow very slowly. In other words, the longer the round trip delay, the lower the downstream throughput. Thus, the two-way transfer case has lower downstream throughput as compared to the one-way transfer.

5. PERFORMANCE EVALUATION

Parameter	Value
C_d	26.97 Mbps
C_u	2.56 Mbps
T	0.5 ms
D_{MAP}	2 ms
t_{ms}	50 μs
N_c	50 (mini-slots)
MAP limit	2048 mini-slots and 240 IEs
L_{data}	1024 bytes
L_{ack}	64 bytes
B_u	20 (packets)
d	2

Table 1. Parameters used for performance evaluation

In this section, we provide numerical examples and simulation results using *ns*-2 to verify the analytical results presented in the previous two sections. The parameters we use in this section are listed in Table 1.

5.2 Numerical Examples

A. One-way TCP Transfers

We first examine the analysis results for one-way TCP transfers based on Table 1. Fig. 4 plots η as a function of N_{dCM}. When N_{dCM} is less than 13, η is larger than one; otherwise, η is less than or equal to one. This implies TCP has worse performance when N_{dCM} is small. In the extreme case, when there is only one TCP transfer, η reaches its maximum value and the TCP transfer results in the worst performance. Fig. 5 shows the relationship between the TCP round trip delay and the number of simultaneous transfers. The solid curve is plotted based on RTT in (8), and the dashed curve is from RTT in (12). It is observed that the asymmetry ratio η has significant impact on TCP round trip delay. When η >1, the round trip delay is in the order of 100 ms, and when $\eta \le 1$, the round trip delay is in the order of 10 ms. The large RTT for η >1 is due to the long waiting time in the upstream buffer at CMs. The linear growth of the curve as η >1 is due to the increasing total number of mini-slots granted by the MAP. The curve stays flat as $\eta \le 1$, invariant to the value of N_{dCM}.

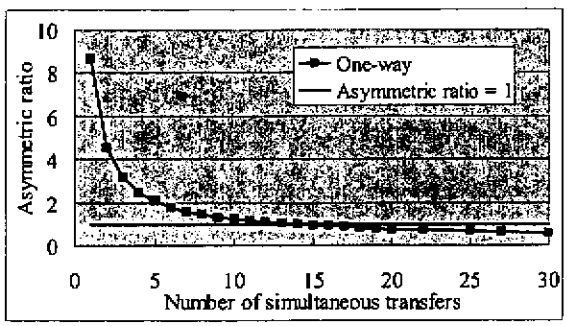


Figure 4. Asymmetry ratio η

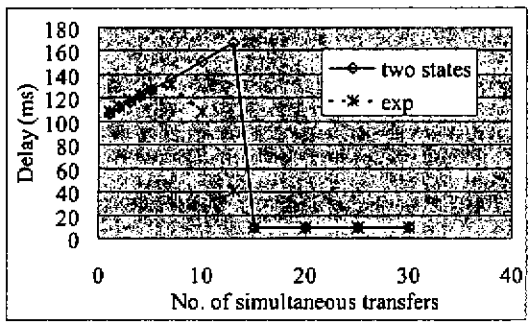


Figure 5. TCP round trip delay

We see that the delays of the solid curve are strictly classified into two categories, i.e., an increasing curve (as $\eta > 1$) and a flat curve (as $\eta \le 1$), due to the result of the following setting: $w=1$ when $\eta > 1$, and $w=0$ when $\eta \le 1$. In other words, there is a sharp drop in the curve at N_{dCM}=13 (i.e.,

$\eta = 1$). In reality, the behavior of TCP should be similar on either side of the point of $\eta = 1$ (e.g., 1.000001 and 0.999999). This phenomenon can also be explained with Fig. 4. When Fig. 4 is rotated counter-clock-wise by 90 degrees, the curve of the asymmetry ratio demonstrates an exponential-like growth. This implies that the RTT value should gradually decrease rather than increase when η approaches one and $1 < \eta \leq 2$. With the exponential setting, the RTT curve decreases smoothly as η approaches to one, better matching our expectation.

B. Two-way TCP Transfers

We fix the total number of simultaneous transfers to 30. Let N_{uCM} of 30 clients transfer TCP in the upstream direction and the rest transfer downstream. The other parameters are based on Table 1.

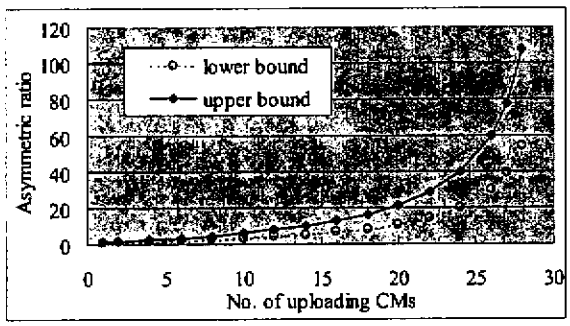


Figure 6. Asymmetry ratio η

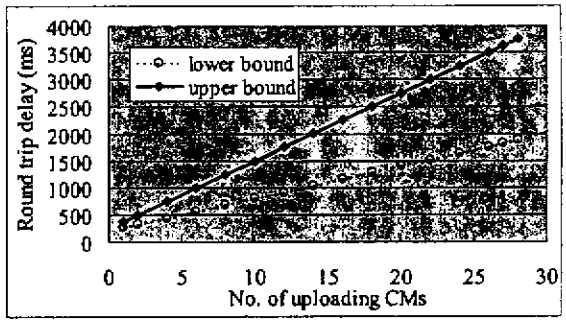


Figure 7. TCP Round trip delay

The asymmetry ratio η and the TCP round trip delay of downloading CMs are plotted in Figs. 6 and 7, respectively. Fig. 6 shows that η of downloading CMs increases sharply as the number of simultaneous uploading CMs increases. η is too high to tolerate when the number of simultaneous uploading CMs is beyond 20, i.e. two-third of clients are uploading. Interestingly, η is larger than one so long as the number of simultaneous transfers is larger than one. This means even with just a few clients uploading, all downloading clients in this network suffer from bandwidth asymmetry. Fig. 7 shows that the TCP round trip delay of downloading CMs grows linearly as the number of simultaneous uploading CMs increases. The large round trip delay is due to a very large η, which is much larger than one and increases exponentially as the number of uploading CMs increases. Compared with one-way

transfers (see Fig. 5), Fig. 7 has much longer delay. This is because long packets (TCP packets) are transferred on the bottleneck channel, causing long waiting times for short packets (ACK packets).

5.2 Analysis vs. Simulation

In this section, the analytical results are compared with the simulation results. Figs. 8 and 9 show the access delays of downloading users for one-way and two-way transfers, respectively. The solid curves are from the simulation, and the dashed curves are from the analysis (U is for the upper bound and L is for the lower bound). We can see that in the one-way case, the two curves stay very close; in the two-way case, the simulation curve is very close to the lower bound of the analytical curve, matching our argument in Sec. 4.1.

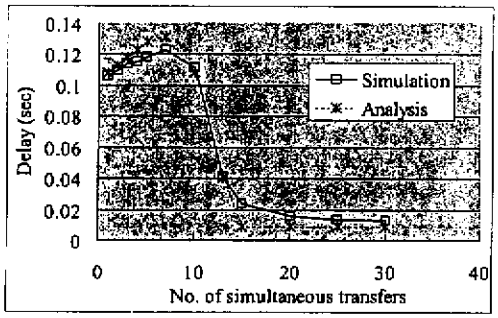


Figure 8. One-way transfers

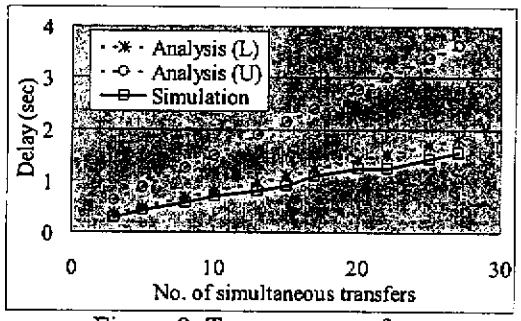


Figure 9. Two-way transfers

6. CONCLUSION

We have studied the impact of DOCSIS MAC layer operation on the performance of TCP over HFC networks. We learn that the asymmetry ratio k expressed in [10] cannot adequately explain the behavior of TCP in DOCSIS-based HFC networks. To better capture the effect of DOCSIS, the asymmetry ratio is expressed in another way, denoted as η, considering the TDMA-like MAC layer operation of DOCSIS. When $\eta > 1$, TCP behaves as in a symmetric network; when $\eta \leq 1$, the system acts as in an asymmetric network, and the performance of TCP degrades. We find that the number of simultaneous transfers significantly affects the asymmetry ratio. If the number of active CMs is below two times the maximum number of pending requests in a transmission period, the value of η is larger than one,

regardless of the value of k. However, as the number of active CMs becomes very large, the effect of DOCSIS on TCP becomes negligible, and the asymmetry ratio is determined by the bandwidth ratio of channels times the length ratio of data and ACK packets.

In our analysis, both one-way and two-way transfers are considered. With two-way transfers, the downloading CMs mostly experience network asymmetry, but operate in the lower bound of η in the steady state. Based on η, the round trip delay of sending a data packet is derived for both one-way and two-way transfers, and the buffer requirement at the CMTS is discussed. We have also conducted simulation to verify the analytical results. The results show a good match between the simulation and the analysis.

This paper does not consider collisions in the contention period. In the future, we will further investigate the impact of collisions on the performance of TCP in DOCSIS networks, and the impact of DOCSIS's MAC layer operations on QoS scheduling mechanisms for multimedia traffic over HFC networks.

REFERENCES

[1]. CableLabs, Data-Over-Cable Service Interface Specifications - Radio Frequency Interface Specification, MCNS Consortium, 2000, SP-RFIv1.1-106-001215

[2]. IEEE Project 802.14/a Draft 3 Revision 1, Apr. 1998.

[3]. M. Droubi, N. Idirene, and C. Chen. "Dynamic bandwidth allocation for the HFC DOCSIS MAC protocol," Proc. IEEE IC3N 2000.

[4]. D. Sala, J. O. Limb, and S. U. Khaunte, "Adaptive Control Mechanism for Cable Modem MAC Protocols," Proc. IEEE INFOCOM '98.

[5]. Y-D Lin, C-Y Huang and W-M Yin, "Allocation and Scheduling Algorithms for IEEE 802.14 and MCNS in Hybrid Fiber Coaxial Networks," IEEE Transaction on Broadcasting, Vol. 44, No. 4 Dec. 1998

[6]. Dolors Sala and John O. Limb, "Comparison of Contention Resolution Algorithms for A Cable Modem MAC Protocol," Proc. Broadband Communications, 1998.

[7]. V. Sdralia, C. Smythe, P.Tzerefos and S. Cvetkovic, "Performance Characterisation of the MCNS DOCSIS 1.0 CATV Protocol with Prioritised First Come First Served Scheduling," IEEE Transactions on Broadcasting, Vol. 45, No. 2, June 1999.

[8]. Th. Orfanoudakis, N. Leligou, E. Meciu and A. Harsanyi, "Evaluation of IP oriented HFC access protocols," Proc. Broadband Communications, 2000.

[9]. M. T. Ali, R. Grover, G. Stamatelos, and David D. Falconer, "Performance Evaluation of Candidate MAC Protocols for LMCS/LMDS Networks," IEEE Journal on Selected Areas in Communications, Vol. 18, No. 7,

July 2000.

[10]. H. Balakrishnan, V. N. Padmanabhan and Randy H. Katz, "The effects of asymmetry on TCP performance," Mobile Networks and Applications 4, p219-p241, 1999.

[11]. Y.- D. Li and W. Liao, "Improving TCP Performance over Asymmetric Networks," Proc. IEEE ICC 2001, Helsinki, Finland, June 2001.

[12]. S. Varma, "Performance and Buffering Requirements of TCP Applications in Asymmetric Networks, " Proc. IEEE INFOCOM '99, New York 1999.

[13]. T. V. Lakshman, U. Madhow and Bernhard Suter, "Window-based Error Recovery and Flow Control with A Alow Acknowledgement Channel: A Study of TCP/IP Performance," Proc. IEEE INFOCOM '97, 1997.

[14]. R. Cohen and S. Ramanathan, "TCP for High Performance in Hybrid Fiber Coaxial Broad-Band Access Networks," IEEE/ACM Transactions on Networking, Vol. 6 No. 1, Feb. 1998

[15]. O. Elloumi, N. Golmie, H.Afifi and D.Su, "A Study of TCP Dynamics over HFC Networks," Proc. IEEE GLOBECOM '98.

[16]. R. Cohen and S. Ramanathan, "Using Proxies to Enhance TCP Performance over Hybrid Fiber Coaxial Networks," Computer Communications, vol. 20, 1998.

[17]. I. M. C. Tam, M. K. Patnam, K. T. Chan and F. S. Fook, "Performance Evaluation of Web browsing over Hybrid Fiber Coaxial Broad-/band Networks," IEEE, 1999

[18]. UCB/LBNL/ISI/VINT Network Simulator - ns, Version 2, http://www.isi.edu/nsnam/ns/

MODLEING USER MOBILITY FOR RELIABLE PACKET DELIVERY IN MOBILE IP NETWORKS

Jiunn Ru Lai and **Wanjiun Liao**
Department of Electrical Engineering and
Graduate Institute of Communication Engineering
National Taiwan University
Taipei, Taiwan
Email:wjliao@ntu.edu.tw

Abstract **In this paper, we analyze the handoff behavior for reliable packet delivery in Mobile IP networks. In Mobile IP, packets destined to roaming hosts are intercepted by their home agents and delivered via tunneling to its Care of Address (CoA). A mobile node may roam across multiple subnets during receiving data reliably. Upon each boundary crossing, a handoff is initiated in which the CoA is updated and a new tunnel is established. We find that reliable packet delivery in Mobile IP networks can be modeled as a renewal process. We derive the probability distribution of boundary crossings for each successfully transmitted packet. We also provide numerical examples to demonstrate how to use our model to calculate the probability distribution of boundary crossings, given the distributions of residence time and local retransmission attempts.**

I. INTRODUCTION

Mobile IP [1] is the dominant standard for host mobility in the Internet. In Mobile IP, mobility service is enabled with the cooperation of mobility agents, i.e., the home agent and foreign agents, based on the operations of "binding" and "tunneling." A mobile node (MN) in the home network does not need support from its HA. When the MN moves away from the home network, the MN registers with its home agent a care-of-address (CoA) temporarily allocated in the foreign network. The home agent will then bind the received CoA with the home IP address of the roaming node. It intercepts packets for the mobile node to the home network and forwards the packets to the roaming node via tunneling, i.e., the original packets are encapsulated in new packets destined to the CoA of the node and sent to the roaming nodes in their new locations using normal IP delivery. Such address binding is also performed upon each handoff. As such, mobile nodes can go anywhere while retaining their home IP addresses to receive packets.

In this paper, we study the impact of node movements on reliable packet delivery. As in IP, Mobile IP provides unreliable datagram service for mobile nodes. Packets may also be lost due to node mobility. Reliable packet delivery here refers to that packets being correctly delivered to the roaming nodes. Such a study is important because providing reliable transport service over Mobile IP, such as TCP or reliable multicast [2-4], will play an important role for current or emerging wireless applications. To the

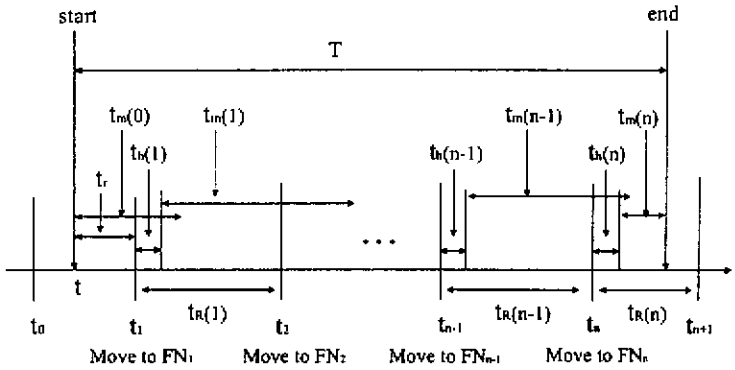


Figure 1. Timing diagram for node movement during a packet delivery time

best of our knowledge, there is no existing work on the analysis of node movements for Mobile IP.

The rest of the paper is organized as follows. In Sec. II, the analytical model is derived. In Sec. III, numerical examples are shown. Finally, the paper is concluded in Sec. V.

II. ANALYTICAL MODEL

In this section, the analytical model of node movements for reliable packet delivery in Mobile IP networks is derived. Let $t_R(i)$ denote a random variable representing for the residence time of an MN in subnet i, i=1,2,3,.. and $t_m(i)$ be a random variable denoting the time needed for successful packet retransmission to MN in subnet i, i=1,2,3, ..We make the following assumptions:

1. We assume that the arrival process of packets destined to an MN is a renewal process.
2. The handoff time (i.e., $t_h(i)$ for moving to subnet i) is assumed the same for all handoff events.
3. We make no assumptions on the distributions of random variables $t_R(i)$ and $t_m(i)$, where i =1,2,...n, except that they are independent of each other and that $\{t_R(i)\}$ and $\{t_m(i)\}$ are sequences of identically distributed random variables.

In this paper, the *packet delivery time* is defined as the total time between when a packet is relayed by the HA of a roaming node and when a packet is received successfully by the node. The packet may be received correctly during the period when the node stays in a subnet (called residence time in this paper), or spanning over multiple subnets. In the former case, the packet delivery time is smaller than the residence time of the node in the subnet. In the latter case, a handoff occurs whenever the node moves across a subnet boundary. On each handoff, the node needs to re-register with the home agent fits new CoA in the new foreign network. This requires a new address binding between the home agent and the roaming node. The time spent in acquiring a new CoA and rebinding is defined as the handoff time in this paper. Once a handoff occurs before the packet is successfully delivered, all the retransmission attempts made in previous subnets become useless, i.e., the packet should be sent as a new copy in the new tunnel, irrespective of how many retries have been attempted in the previous subnets. Since this operation repeats after each handoff and behaves independently from the previous history, the behavior of reliable packet delivery can be modeled as a renewal process with identically and independently distributed (*iid*) processing times.

Fig. 1 shows the timing diagram for N handoffs in a packet delivery time. Suppose that a mobile node MN is roaming to a foreign network, say subnet FN_0,

and has been registered with its HA for address binding. The process begins when a packet destined for the MN is intercepted by its HA. The HA then sends the packet via tunneling to MN. Assume that the packet received by MN is in subnet FN_0 at time t. According to the random observer probability from the renewal theory [5], the distribution of the residual time t_r for MN staying in subnet FN_0 can be derived as follows. Let $f_r(t)$ denote the probability density function of t_r, and $f_r^*(s)$ denote the Laplace transform of $f_r(t)$. Thus, we have

$$f_r(t) = \lambda_r \int_t^\infty f_X(\tau)d\tau,\ t \geq 0, \text{ and}$$

$$f_r^*(s) = \lambda_r \times \frac{\left[1 - f_X^*(s)\right]}{s},$$

where random variable X is the residence time of the MN receiving a new packet in subnet FN_0, $f_X(\tau)$ is the probability density function of X, and λ_r is the reverse of the mean residual time in subnet FN_0.

Let N denote a random variable representing the number of handoffs during the packet delivery time, with the following probability distribution:

$$\Pr[N=0] = \Pr[t_m(0) \leq t_r],$$

$$\Pr[N=n] = \Pr[t_m(0) > t_r] \times \prod_{i=1}^{n-1} \Pr\{t_m(i) + t_h(i) > t_R(i)\} \times \Pr[t_m(n) + t_h(n) \leq t_R(n)]$$

Note that we make no assumptions on the distributions of random variables $t_R(i)$, $t_h(i)$, and $t_m(i)$, where $i = 1,2,\ldots n$, except that they are independent of each other and that $\{t_R(i)\}$ and $\{t_m(i)\}$ are sequences of identically distributed random variables. We also assume that the handoff time is identical for all handoff events. Thus, we omit the argument i.

Define $P_N(n) = \Pr(N=n)$. Thus,

$$P_N(0) = \Pr[t_m \leq t_r]$$

$$= \int_0^\infty \Pr(t_m \leq t) \times f_r(t)dt = \int_0^\infty \int_0^t f_m(\tau)d\tau f_r(t)dt$$

$$= \int_0^\infty \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_m^*(s)}{s} e^{st} ds f_r(t)dt$$

$$= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_m^*(s)}{s} \int_0^\infty f_r(t)e^{st}dtds \qquad (1)$$

$$= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \frac{f_m^*(s)}{s} f_r^*(-s)ds$$

$$= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[\frac{f_m^*(s)}{s}\right]\left[\frac{\lambda_R\left(1 - f_R^*(-s)\right)}{(-s)}\right]ds$$

$$= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \lambda_R \left(\frac{f_m^*(s)}{s^2}\right)\left[f_R^*(-s) - 1\right]ds$$

where $f_m(t)$ is the probability density function of t_m and $f_m^*(s)$ is the Laplace transform of $f_m(t)$; $f_R(t)$ is the probability density function of t_R and $f_R^*(s)$ is the Laplace transform of $f_R(t)$; σ is a sufficiently small positive number appropriately chosen for the inverse Laplace transform and is chosen to be less than the smallest of the real parts of the poles of $f_R^*(-s)$.

Let $g_X(t) = \text{Prob}(X < t) = \int_0^t f_X(\tau)d\tau$, and $g_X^*(s) = \frac{f_X^*(s)}{s}$. Thus, $g_X(t - T_h) = \int_0^{t-T_h} f_X(\tau)d\tau$, and its Laplace transform $= e^{-T_h \times s} \times g_X^*(s) = e^{-T_h \times s} \times \frac{f_X^*(s)}{s}$.

We then derive $P_N(i) = \Pr(N=i)$, $i = 1,2,\cdots,n$. Let $\theta = \Pr[t_m + T_h > t_R]$. Thus, $1-\theta$ can be obtained as follows.

$$\Pr[t_m + T_h \leq t_R]$$

$$= \int_0^\infty \Pr(r_R \leq t - T_h) \times f_R(t)dt = \int_0^\infty \left\{ \int_0^{t-T_h} f_m(\tau)d\tau \right\} f_R(t)dt$$

$$= \int_0^\infty \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[e^{-T_h \times s} \times \frac{f_m^*(s)}{s}\right] e^{st} ds f_R(t)dt$$

$$= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[e^{-T_h \times s} \times \frac{f_m^*(s)}{s}\right] \int_0^\infty f_R(t)e^{st}dtds \qquad (2)$$

$$= \frac{1}{2\pi j} \int_{\sigma-j\infty}^{\sigma+j\infty} \left[e^{-T_h \times s} \times \frac{f_m^*(s)}{s}\right] f_R^*(-s)ds$$

Therefore, $P_N(n) = (1 - P_N(0)) \times \theta^{n-1} \times (1-\theta)$.

We see that the right hand side of the integrand in equations (1) or (2) (i.e., $\lambda_R\left(\frac{f_m^*(s)}{s^2}\right)$ and $\left[e^{-T_h\times s}\times\frac{f_m^*(s)}{s}\right]$, respectively) are analytic at the right half open complex plane. If $f_R^*(-s)$ has no branch point and has only finite possible isolated poles at the right half plane, we can apply the Residue Theorem to (1) and (2) using a semi-circular contour in the right half plane. Let Ω denote the set of the poles of the function $f_R^*(-s)$ in the right half complex plane. Applying the Residue Theorem to (1) and (2), we obtain the following results.

$$P_N(0)=(-1)\times\sum_{n\in\Omega}\operatorname*{Res}_{s=n}\left[\frac{\lambda_R\times f_m^*(s)}{s^2}\times\left[f_R^*(-s)-1\right]\right]$$

$$P_N(n)=(1-P_N(0))\times\theta^{n-1}\times(1-\theta),\quad n=1,2,\cdots$$

where

$$\theta=1+\sum_{n\in\Omega}\operatorname*{Res}_{s=n}\left[e^{-T_h\times s}\times\frac{f_m^*(s)}{s}\times f_R^*(-s)\right].$$

Interestingly, θ can be treated as the "moving-out" probability from a subnet for a packet delivery time in Mobile IP networks, and can be used to express $P_N(n)$. Later in the numerical example section, we will further investigate the impact of θ on $P_N(n)$.

III. NUMERICAL EXAMPLE

In this section, we use the following example to show how to calculate the probability distribution of boundary crossings for a packet delivery time in Mobile IP networks.

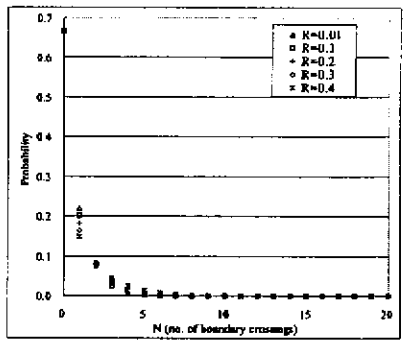
Let t_R be exponentially distributed with probability density function $f_R(r_0)=\lambda_R e^{-\lambda_R r_0}$, $r_0\geq 0$. The Laplace transform of $f_R(r_0)$ is then $f_R^*(s)=\frac{\lambda_R}{(s+\lambda_R)}$. Substituting s in $f_R^*(s)$ with $-s$, we have $f_R^*(-s)=\frac{\lambda_R}{(-s+\lambda_R)}$. Therefore,

$$P_N(0)=f_m^*(\lambda_R),$$

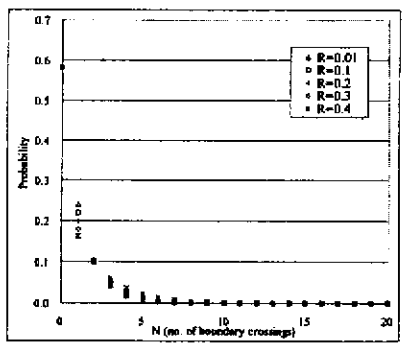
and $$P_N(n)=(1-f_m^*(\lambda_R))\times\theta^{n-1}\times(1-\theta),$$

$n=1,2,\cdots$, where $\theta=1-e^{-T_h\times\lambda_R}\times f_m^*(\lambda_R)$.

Next, we consider $f_m(m_0)$ is exponentially distributed with rate λ_m. Its Laplace transform $f_m^*(s)$ can be expressed as $f_m^*(s)=\frac{\lambda_m}{(s+\lambda_m)}$. Substituting s in $f_m^*(s)$ with λ_R, we yield $f_m^*(\lambda_R)=\frac{\lambda_m}{(\lambda_R+\lambda_m)}$, from which $P_N(0)$ and $P_N(n)$, $n=1,2,\cdots$, can both be determined accordingly.



(a) $\frac{E[t_m]}{E[t_R]}$=0.5, and $\lambda_R=0.01$



(b) $\frac{E[t_m]}{E[t_R]}$=0.7, and $\lambda_R=0.01$

Figure 2. $P_N(n)$

In Fig. 2, $P_N(n)$ is plotted as a function of n for exponentially distributed t_R and exponentially distributed t_m. To observe the impact of other parameters on the probability distribution of the

number of boundary crossings for successful packet delivery, we vary the following ratios: (1) $\frac{E[t_m]}{E[t_R]}$ and (2) $\frac{E[t_h]}{E[t_R]}$, where $\frac{E[t_m]}{E[t_R]} = \frac{\lambda_R}{\lambda_m}$ and $\frac{E[t_h]}{E[t_R]} = T_h \times \lambda_R$.

In Fig. 2, R is defined as $\frac{E[t_h]}{E[t_R]}$. We have the following observations.

(1) The curves of $P_N(n)$ in all cases drop rapidly at small n. In other words, it is more likely that a packet will be delivered successfully after a few handoffs.

(2) The curves with the same value of $\frac{E[t_m]}{E[t_R]}$ have the same value at n=0, i.e., $P_N(0)$. The smaller the value of $\frac{E[t_m]}{E[t_R]}$, the larger the value of $P_N(0)$.

(3) The $P_N(n)$ curve drops more sharply for smaller "moving-out" probability (i.e., θ) from a subnet. Such moving-out speed is determined by $\frac{E[t_h]}{E[t_R]}$ and $\frac{E[t_m]}{E[t_R]}$. A smaller value of $\frac{E[t_h]}{E[t_R]}$ will have a smaller moving-out speed. This is because the value of $\frac{E[t_h]}{E[t_R]}$ will become smaller when $E[t_h]$ decreases and $E[t_R]$ is fixed, or when $E[t_h]$ is fixed and $E[t_R]$ increases. Thus, a smaller $\frac{E[t_h]}{E[t_R]}$ incurs less boundary crossings, leading to a higher probability to successfully deliver a packet within a subnet. Thus, the curve drops more sharply. Similarly, a smaller $\frac{E[t_m]}{E[t_R]}$ will have a smaller θ, and the curve also drops more sharply.

IV. CONCLUSION

In this paper, we have modeled the node movement behavior for reliable packet delivery in Mobile IP networks. The probability distribution of the number of boundary crossings for reliable packet delivery is derived. We also provide numerical examples to demonstrate how to use our model to calculate the probability distribution of boundary crossings, given the distributions of residence time and local retransmission attempts.

ACKNOWLEDGEMENT

This work was supported by the National Science Council, Taiwan, under Grant Number NSC 92-2213-E-002-013.

REFERENCE

[1] C. Perkins, "IP Mobility Support," IETF RFC 2002, Oct. 1996.

[2] H. Balakrishnan, S. Seshan, and R.H. Katz., "Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks," Proc. ACM Wireless Networks, Dec. 1995.

[3] W. J. Liao, J.-A Ke, and J.-R. Lai, "Reliable Multicast with Host Mobility," in *Proc. IEEE Globecom 2000.*

[4] J. R. Lai and W. J. Liao, "Reliable Multicast for Mobile IP Networks: Analytical Study," Proc. IEEE Globecom 2001.

[5] D. C. Cox, Renewal Theory, New York: Wiley, 1962.

Fair Scheduling in Mobile Ad Hoc Networks with Channel Errors

Hsi-Lu Chao, *Student Member, IEEE*, and Wanjiun Liao, *Member, IEEE*

***Abstract*—In this paper, we study fair scheduling in ad hoc networks, accounting for channel errors. Since wireless channels are susceptible to failures, to ensure fairness, it may be necessary to compensate those flows with error-prone channels. Existing compensation mechanisms need the support of base stations and only work for one-hop wireless channels. Therefore, they are not suitable for multihop wireless networks. Existing fair scheduling protocols for ad hoc networks can be classified into timestamp-based and credit-based approaches. None of them takes channel errors into account. In this paper, we investigate the compensation issue of fair scheduling and propose a mechanism called Timestamp-Based Compensation Protocol (TBCP) for mobile ad hoc networks. We evaluate the performance of TBCP by simulation and analyze its long-term throughput. The results show that our analytical result provides accurate performance estimation for TBCP.**

***Index Terms*—fair scheduling, wireless ad hoc networks, fairness, channel compensation.**

I. Introduction

An ad hoc network is a self-organizing wireless network comprised only of mobile nodes. In such a network, nodes are connected via a multihop path and without the support of any pre-existing wired infrastructure. In wireless networks, radio spectrum (i.e., bandwidth) is a precious and limited resource. Such resource should be shared by nodes in a fair way, especially when there is competition for this resource due to unsatisfied demands.

Existing fair scheduling mechanisms in ad hoc networks can be classified into two categories: timestamp-based [1,2] and credit-based [3], according to different decision metrics. The timestamp-based mechanisms in [1] and [2] work similarly. In [2], a flow graph must first be generated. Each vertex in the flow graph represents a flow. When two flows contend for the same resource, an edge is added between them. The scheduling discipline in [2] is based on Start-time Fair Queueing (SFQ) [4], and each newly arriving packet is locally assigned a start and a finish tag according to (1).

$$\begin{cases} S_f^j = max\left\{v\left[A\left(p_f^j\right)\right], F_f^{j-1}\right\}, j \geq 1 \\ F_f^j = S_f^j + \dfrac{l_f^j}{\omega_f}, j \geq 1 \end{cases}, \quad (1)$$

where S_f^j is the start tag and F_f^j is the finish tag of the j^{th} packet of flow f; p_f^j, l_f^j, $A(p_f^j)$, and $v\left[A(p_f^j)\right]$ are the j^{th} packet of flow f, the length of p_f^j, the arrival time of p_f^j, and the virtual arrival time of p_f^j, respectively; ω_f is the flow weight of flow f, indicating the fraction of link capacity to be shared by f. Either timestamps can be chosen as the service tag. The packet with the smallest service tag is transmitted first. Spatial channel reuse is implemented by means of backoff parameters. Each packet of a flow is associated with a backoff value. This backoff value is equal to the number of flows contending for the channel but with smaller service tags, and is decremented by one at each time slot. Once this backoff value reaches zero, the packet is transmitted.

In [3], a two-tier and cluster-based mechanism is proposed. The network is logically partitioned into clusters, each with a scheduler. Each flow is associated with three parameters: *Credit*, *Usage*, and *Excess*, where *Excess* = *Usage* − *Credit*. The scheduler assigns time slots to mobiles in the respective cluster based on the first tier algorithm. The mobiles scheduled to send at the next time slot then in turn assign the time slot to the flows determined by the second tier algorithm. Each cluster scheduler operates independently. The scheduling disciplines in both tiers are all based on the *Credit* value, i.e., the one with the smallest *Excess* value has the priority to transmit. The unused credit can be accumulated for future use.

The problem in [1-3] is that the transmission channels are all assumed error-free. This assumption is, however, unrealistic for wireless networks [5] due to their high channel error rates, such as location-dependent errors or bursty errors. In what follows, we first survey the related work on error compensation for wireless networks, and then describe our main contribution to the addressed problem for ad hoc networks.

Manuscript received Oct. 15, 2003. Manuscript revised on Jan. 4, 2004, and accepted on Feb. 22, 2004. This work was supported in part by MOE program for Promoting Academic Excellence of Universities under grant number NSC93-2752-E-002-006-PAE, and in part by the National Science Council, Taiwan, under grant number NSC93-2213-E-002-001.

H. L. Chao is with the Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan.

W. J. Liao is with the Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan

A. Related Work on Error Compensation in Wireless Networks

Existing works proposed to achieve fairness for wireless networks with channel errors are all based on the support of base stations and work for one-hop wireless channels [6-9].

In [6], the Channel-condition Independent packet Fair Queueing (CIF-Q) is proposed. Each session is associated with a parameter called *lag* to indicate if the session should be compensated. If a session is not leading and the channel is error-free at its scheduled time, its head-of-line packet is transmitted; otherwise, the time slot is released to other sessions. The problem with CIF-Q is that leading sessions are not allowed to terminate unless all their leads have been paid back, regardless of whether such terminations are caused by broken routes. This property makes CIF-Q an inadequate solution for ad hoc networks, because a connection may be broken due to node movements.

In [7], the Idealized Wireless Fair-Queueing (IWFQ) and the Wireless Packet Scheduling (WPS) protocols are proposed. If a session experiences channel errors at its scheduled time, the base station first attempts to swap the slots assigned to this session with other sessions within the same frame. If such attempt fails, the base station will compensate the error slots of this session in a later frame. Again, this mechanism is not suitable for ad hoc networks, due to the need of support from base stations.

In [8], a virtual compensation session is introduced in the scheduling process. The virtual session is a session, which always experiences an error-free channel at its scheduled time but does not really generate packets. The slots received by this virtual session will be used for compensation, i.e., they will be reassigned to lagging or handoff sessions. This mechanism, however, does not consider multihop flows. Thus it is not suitable for ad hoc networks.

In [9], Bandwidth-Guaranteed Fair Scheduling with Effective excess Bandwidth Allocation (BGFS-EBA) is proposed. In BGFS-EBA, each flow can be a lagging, leading, or in-synch flow as compared to its bandwidth allocation in an error-free environment. Each flow is associated with a parameter called deadline. Bandwidth resource is first scheduled according to such deadlines, and then further allocated based on flow status (i.e., lagging, leading, or in-synch) as in CIF-Q. Unlike other mechanisms, which allow a flow to transmit only in a good state, a flow in BGFS-EBA is still allowed to transmit, albeit at lower rates, when its channel is error-prone. Again, this mechanism needs the support of base stations. Thus it is not suitable for ad hoc networks.

B. Problem Specification

It is a challenge to provide channel compensation while sharing resource fairly for ad hoc networks. The challenge is mainly caused by the fully distributed nature of mutli-hop wireless networks. In ad hoc networks, there is no support from base stations, and multip-hop flows passing through different hops are processed in an uncoordinated way. Existing channel compensation mechanisms for wireless networks, however, are mainly designed for single-hop networks with the support of base stations. This renders them inadequate for ad hoc networks.

In this paper, we study fair scheduling, accounting for channel errors for ad hoc networks. In particular, we focus on timestamp-based mechanisms. Existing timestamp-based mechanisms for ad hoc networks (i.e., [1], [2]) only work for single-hop flows. Implementing fully distributed scheduling based on timestamps for multihop flows is much more challenge. In this paper, we propose a new mechanism called Timestamp-Based Compensation Protocol (TBCP), which is a multihop timestamp-based fair scheduling mechanism providing error compensation for flows experiencing channel errors. We consider a mix of best effort and QoS flows in the network, where QoS flows are flows whose minimum bandwidth requirements should be guaranteed. Our objective is to guarantee the minimum bandwidth requirements of QoS flows, while ensuring global fairness for best effort flows. Note that for credit-based mechanisms (e.g., CSAP [3]), unused credits can be accumulated for future use and error compensation is naturally available. As a result, they can be easily extended to handle channel errors.

The rest of this paper is organized as follows. TBCP is described in Sec. II. The performance of TBCP is evaluated in Sec. III and IV. Finally, the paper is concluded in Sec V.

II. Timestamp-Based Compensation Protocol (TBCP)

In TBCP, the scheduling discipline is also based on SFQ [4]. The start tag in (1) is selected as the service tag for TBCP. When there is no packet to be sent, the service tag is set to ∞.

A. System Model

The general system model for TBCP is described as follows.

(1) A multihop wireless network is considered, in which a TDMA-based system is assumed to operate over a single channel shared by all hosts. The bandwidth is represented as frames of time slots and the slot allocation to packets waiting to be transmitted is based on their service tags. Each TDMA frame contains a fixed number of time slots. The network is synchronized on a frame and slot basis, using existing synchronization protocols [10,11]. Consider the tiered adaptive timing synchronization procedure (TATSP) [10] as an example. In TATSP, a small set of fast nodes is quickly identified. These nodes are given more chances to send out beacons. Each node should adopt the timing received from any beacon with a timestamp later than its timestamp.

(2) We support two types of flows: best-effort and QoS flows. For QoS flows, the minimum bandwidth requirement is guaranteed, and is represented as a fraction of one time slot, denoted as the *Resv* value in this paper. Thus, for a QoS flow, the *Resv* value is a positive real number between zero and one; for a best effort flow, the *Resv* value is always zero.

(3) A valid route should be constructed before data packets are transmitted, using whatever ad hoc routing mechanisms. For best effort flows, an ad hoc routing protocol (such as [12], [13]) is used; for QoS flows, QoS routing (such as [14], [15]) is used. The routing mechanism should also re-route packets on the broken path caused by node movements. The rerouting may be either locally constructed or fully reconstructed from the source to all nodes.

(4) Simple admission control is performed at each node during path construction. A flow is admitted only when the sum of the reservation levels (i.e., *Resv*) of all flows, including the newly arriving one, is less than or equal to the target link utilization.

(5) We consider only short-term channel errors. To avoid resumed flows hogging the channel resource after long-term error conditions, the flows without message exchanges longer than a threshold due to channel errors will be removed from the scheduling. As such, the resumed flows need to compete for the resource as newly arriving flows.

(6) The channel state is detected as follows. Each node periodically measures the Signal-to-Noise Ratio (*SNR*) of the channel. If the *SNR* value falls below a predefined threshold, the channel is deemed error-prone and the node stops from transmitting packets temporarily. If the *SNR* value exceeds the predefined threshold again, the transmission resumes.

Note that to better distinguish between the mechanisms with and without implementing compensation, those considering channel errors in their scheduling are referred to as compensation models, and those assuming that channels are always error-free in their scheduling as error-free models. In addition, the nodes experiencing error states are called channel-error nodes, and those without errors are error-free nodes. Similarly, packets transmitted on error-prone channels are called channel-error packets and those on error-free channels are error-free packets.

B. Multihop Flows

Each node in mobile ad hoc networks acts as both a router and a host. In TBCP, a multihop flow is modeled as multiple single-hop flow segments, and each node schedules the single-hop flows passing through it independently. In TBCP, each end-to-end flow is referred to as a "flow," and the single-hop flows belonging to a multi-hop flow are called "flow segments." For example, flow F in Fig. 1 is comprised of five single-hop flow segments, i.e., F_{12}, F_{23}, F_{34}, F_{45}, and F_{56}. The source and the destination of flow F are nodes 1 and 6, respectively. Note that each single-hop flow has one and only one flow segment.

Each flow segment of a multi-hop flow may be scheduled by different nodes. As a result, even though a downstream flow segment has been assigned a slot, it may not have packets queued to be transmitted unless the upstream flow segments have all been allocated a slot. To solve this problem, a new parameter called *Q-size* is defined to correlate multiple flow segments belonging to the same multihop flow. Each flow is associated with a *Q-size* parameter at each node on the flow path. This parameter is initialized to zero at all nodes except for the sender, which has a non-zero *Q-size* depending on the number of packets to be transmitted. The *Q-size* of a flow segment indicates the number of packets received from its previous hop and waiting to be sent to the next hop. A node with nonzero *Q-size* means at least one of its flow segments has packets waiting to be sent, and this node will choose the least service tag of queued packets as its service tag. Among all nodes with nonzero *Q-size* values, the node with the least service tag is scheduled for the next time slot.

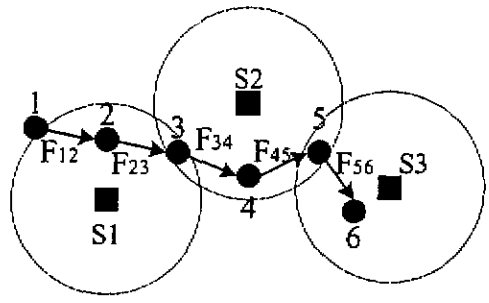


Figure 1. An example to explain *Q-size*

C. Normalized Flow Weight

Assume that there are n flows passing through node N. Let x_i^N denote the bandwidth share of flow i at node N (called the flow weight of flow i at node N in this paper), expressed as

$$x_i^N = \frac{1 - \sum_{j=1}^{n} Resv_j}{n} + Resv_i, \qquad (2)$$

where $Resv_i$ is the minimum bandwidth requirement of flow i, represented as a fraction of the channel bandwidth. Bandwidth share x_i^N indicates the fraction of channel bandwidth used by flow i at node N, so as to fairly share the residual bandwidth of node N and meet the minimum bandwidth requirement of flow i. For example, assume that there are three flows, say, f_1, f_2, and f_3 passing through node N. Flows f_1 and f_2 are guaranteed flows, each with a *Resv* value of 0.2; flow f_3 is a best effort flow. Based on (2), the bandwidth share of the three flows at node N are (x_1^N, x_2^N, x_3^N)=(0.4, 0.4, 0.2).

In TBCP, the actual bandwidth shared by each flow at node N is collaboratively determined by all flows at all nodes in the interference range (i.e., nodes located within two hops) of node N, not just solely depending on those flows passing through node N. Due to lack of coordination for transmissions, nodes located in the interference range may contend for the channel, and such contention should be considered in the calculation of bandwidth share for each flow. To reflect this fact, the normalized flow weight of each flow at node N is defined.

Let F_N and $F_{N'}$ denote the two sets of flow segments passing through node N and N', respectively, where N' is within two hops of node N. Let S_N be the set of nodes, including N and all other nodes in its interference range (i.e., two hops). The normalized flow weight of flow segment f at node N is

$$\omega_f^N = \frac{x_f^N}{\sum_{i \in (F_N \cup F_{N'}), j \in S_N} x_i^j} \quad (3)$$

This normalized flow weight ω_f^N is then used to calculate the service tag of flow f at node N. The number of slots per frame obtained by flow f at node N is equal to ω_f^N times the number of slots per frame. Note that since the calculations of (2) and (3) are performed independently at each node, different single-hop flow segments of a multihop flow may have different bandwidth shares and normalized flow weights.

For example, consider two nodes, say N_1 and N_2, in an ad hoc network. These two nodes are within the transmission range of each other. Let $Resv_i^j$ indicate the $Resv$ value of the i^{th} flow at node j. The flow information of N_1 is $(Resv_1^{N_1}, Resv_2^{N_1}, Resv_3^{N_1})$ = (0.2,0.2,0), and that of N_2 is $(Resv_1^{N_2}, Resv_2^{N_2}, Resv_3^{N_2})$ = (0.4,0,0). Thus the bandwidth shares calculated by N_1 and N_2 are (0.4,0.4,0.2) and (0.6,0.2,0.2), respectively, and the normalized flow weights of all flows are $(\omega_1^{N_1}, \omega_2^{N_1}, \omega_3^{N_1}, \omega_1^{N_2}, \omega_2^{N_2}, \omega_3^{N_2})$ = (0.2,0.2,0.1,0.3,0.1,0.1).

So far we have not mentioned how the signaling protocol works for the message exchange among nodes in the interference range. Recall that each TDMA frame is comprised of a fixed number of time slots. The network is synchronized on a frame and slot basis. A frame is comprised of a control phase and a data phase. The signaling messages are exchanged in the control phase. The messages to be exchanged are flooded to a region within two hops (with TTL=2) during the control phase. Based on the received messages, the normalized flow weight of each flow at a node can be calculated.

D. TBCP Operations

1) Determination of Packet Transmission Order at a Node: The transmission order of a packet at a node is affected by three factors: the number of slots per frame the flow can use, the flow's *Q-size* at that node, and the service tag of the packet. The normalized flow weight ω_f^N of flow f at node N is first calculated. Based on ω_f^N, flow f at node N is allocated with $\omega_f^N \times S$ slots per frame, where S is the number of slots per frame. The allocated slots may exceed the requirement if the number of packets waiting to be transmitted, i.e., the *Q-size* of flow f, is less than $\omega_f^N \times S$. In this case, the unused slots are released to other flows, and flow f will not be compensated with these give-away slots. The transmission order is then determined by the service tag. For example, the numbers in Fig. 2(a) show the packets' service tags of three flows f_1, f_2, and f_3 at node N. Assume that each frame is comprised of five slots, and f_i^j indicates the j^{th} packet of flow i. The number of slots per frame each flow can use is (2,2,1). For flow f_1, the first two least service-tag packets are transmitted, i.e., f_1^1 and f_1^2. After the comparison with other two flows, the transmission order of packets at node N in Fig. 2(a) is $< f_1^1, f_3^1, f_1^2, f_2^1, f_2^2 >$. However, if f_2 has only one packet waiting to be sent as in Fig. 2(a), its unused slot will be released to other flows. In this case, the transmission order arranged by node N becomes $< f_1^1, f_3^1, f_1^2, f_2^1, f_1^3 >$.

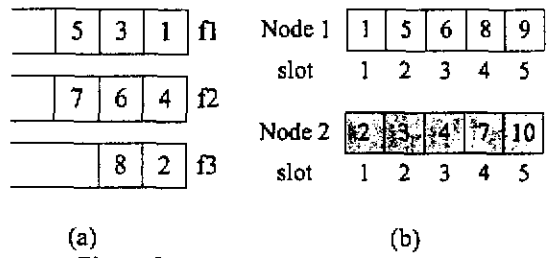


Figure 2. An example to explain TBCP

2) Message Exchange: Each node exchanges the information about the transmission order of packets determined at step 1) with its neighbors on a per-frame basis. Thus each node knows the service tags of other nodes, and also learns when it will transmit packets. For example, assume that there are two nodes in an ad hoc network, and the corresponding transmission order with the service tag is shown in Fig. 2(b). Let N_i^j indicates the j^{th} slot of node i. The transmission sequence is $< N_1^1, N_2^1, N_2^2, N_2^3, N_1^2 >, < N_1^3, N_2^4, N_1^4, N_1^5, N_2^5 >$. Note that if multiple nodes have packets with the same service tag, the tie will be broken based on their node IDs.

Note that the messages are exchanged by flooding with TTL=2. As such, the exchanged messages are restricted to those neighbors within two hops. The communication overhead of TBCP is mainly caused by such message exchanges within two hops

3) Channel Error Handling: Each node keeps monitoring its channel state. When the channel is error-prone, the node stops exchanging transmission messages with its neighbors. Once the channel recovers, the error-prone node resumes the exchanges. Consequently, if this node has packets with service tags smaller than its neighbors after recovery, these packets still have higher priority to be transmitted. For example, as shown in Fig. 2(b), node 2 experiences an error-prone state at slot 4 during the transmission of frame 1, but the channel recovers before the frame 2 is sent. The modified sequence then becomes $< N_1^1,$

N_2^1, N_2^2, X, N_1^2 >, < N_2^3, N_1^3, N_2^4, N_1^4, N_1^5 >. Note that one slot is wasted in this example, because messages are exchanged on a per-frame basis.

III. Simulation Results

In this section, TBCP is evaluated via simulation. We compare TBCP with TBP, the error-free model of TBCP (i.e., assuming that channels are error-free). We also show CBCP and CSAP [3] (denoted as CBP in this paper) in the figures for comparison, where CBP and CBCP are the credit-based mechanisms for channel error-free and compensation models, respectively. CBCP differs from CBP in that when a node detects an error state at its scheduled time, the node will inform its scheduler to stop assigning further slots to it. The *Credit* value of this node at the scheduler continues to be accumulated. Thus, once the channel has recovered, the node has a small *Excess* value as compared to other error-free nodes, which allows this node to have priority to obtain slots. This node then in turn assigns this slot to the flow with the least *Excess* value among all nonzero *Q-size* flows

A. Simulation Environment and Performance Metrics

In the simulation, 20 nodes are randomly distributed in a 1000-meter by 1000-meter area. The transmission range of each node is 250 meters. Some nodes are randomly selected as flow sources and some as flow destinations. Each flow may either be a best effort or a guaranteed flow and is continuously backlogged. Each packet is assumed to occupy one time slot, and has fixed packet length.

We implement spatial channel reuse for both TBP and TBCP in the simulation, using a mechanism similar to that in [2]. Each flow maintains a parameter called *Backoff*, which decides if it can use the next slot. The *Backoff* value of a flow, say flow i, is the number of flows which are located in flow i's interference range and have smaller service tags. This *Backoff* value counts down at each time slot. When the *Backoff* value becomes zero, the flow can use the next slot.

The proposed mechanism is evaluated with two scenarios: multihop flows without node mobility, and multihop flows with mobility support. The mobility pattern of each node follows the modified random waypoint model [16]. Each node randomly selects a target position and speed to move. The speed a mobile node selects is within (Min, Max) meters per second. After arriving at the target position, the mobile node stays there for a predefined period of time, called the pause time. The mobile host then randomly selects a new target position and a speed, and moves again.

The performance metrics measured in the simulation include the relative network throughput, satisfaction index, and fairness index.

(1) Network throughput (ρ): this parameter is defined as the summation of all flows' throughput achieved by an approach.

(2) Relative network throughput (ψ): this parameter is defined as $\frac{throughput_{CBCP}}{throughput_{CBP}}$ for the credit-based mechanism and $\frac{throughput_{TBCP}}{throughput_{TBP}}$ for the timestamp-based mechanism.

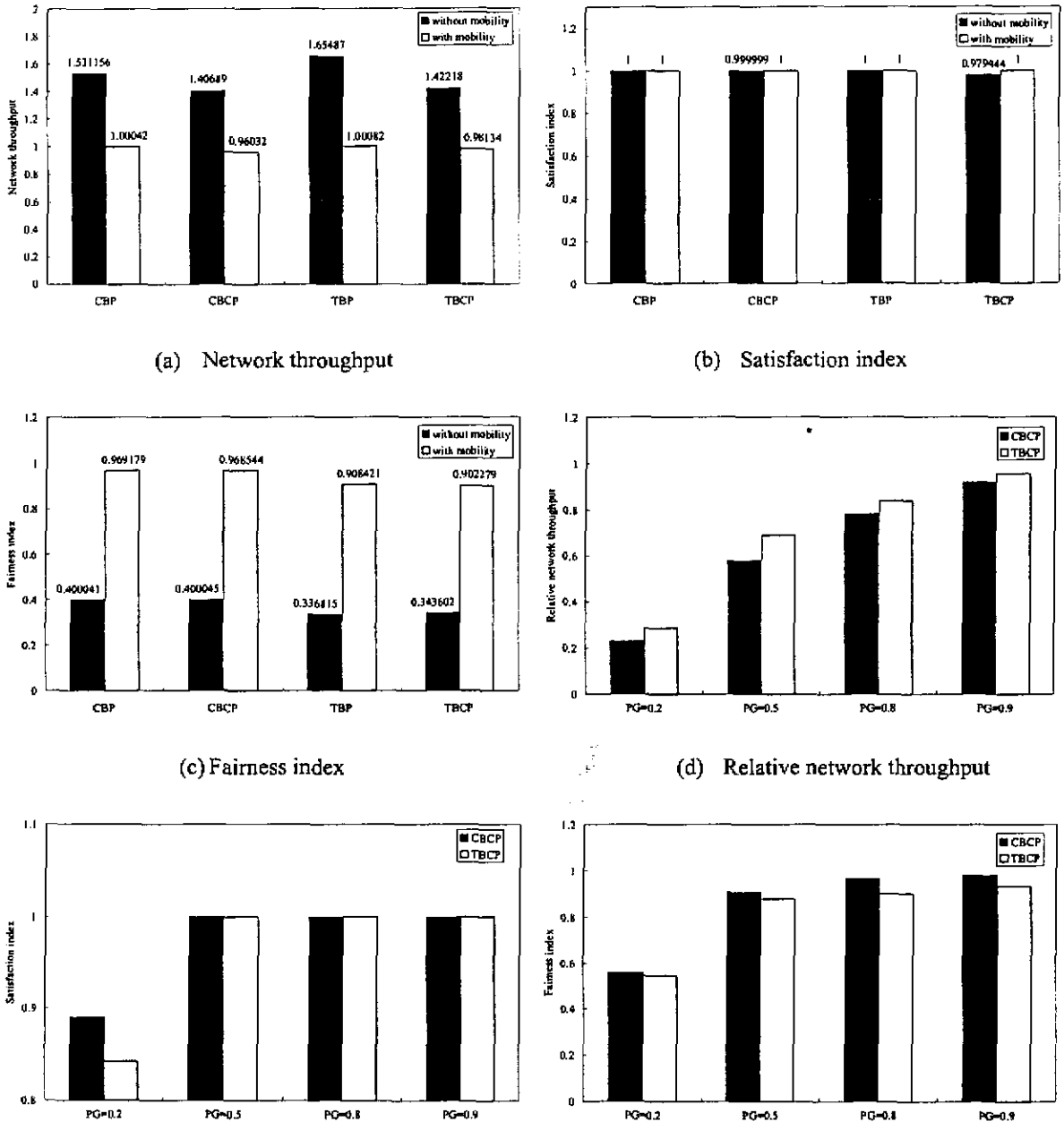
(3) Satisfaction index (ξ): this parameter indicates how well all QoS flows are satisfied for each approach, considering channel conditions. ξ is defined in a way similar to the definition of the fairness index in [17]. Thus we only count QoS flows and define it as $\xi = \frac{\left(\sum_i x_i\right)^2}{m\sum_i \left(x_i^2\right)}$, where x_i is the residual bandwidth share (i.e., flow's slot usage divided by simulation time, less the *Resv* value), and is set to one if the corresponding x_i is equal or larger than zero, otherwise x_i is one minused the shortage divided by its *Resv* value; m is the number of guaranteed flows.

(4) The fairness index (κ): this parameter indicates how fair the residual bandwidth is shared by all flows for each approach, and is defined as in [17], i.e., $\frac{\left(\sum_i x_i\right)^2}{n\sum_i \left(x_i^2\right)}$, where x_i is the residual bandwidth share if x_i is equal or larger than zero, otherwise x_i is zero; n is the number of flows in an ad hoc network. Note that we do not use the proportional fairness defined in [6,7,8], i.e., $\left|\frac{W_i(t_1,t_2)}{\omega_i} - \frac{W_j(t_1,t_2)}{\omega_j}\right|$, as the fairness index. The reason is for a multihop flow, the single-hop flow segments may not have identical flow weights.

B. Simulation Results

We adopt a two-state discrete Markov chain as in [18] to model wireless channel errors. Let p_g denote the probability that the next time slot is in the good state given that the current time slot is bad, and let p_b denote the probability that the next time slot is bad given that the current slot is in the good state. The steady state probabilities P_G and P_B in the good and error states, respectively, are given by $P_G = \frac{p_g}{p_g + p_b}$ and $P_B = \frac{p_b}{p_g + p_b}$.

We first generate ten multihop flows: six guaranteed flows and four best-effort flows. p_g and p_b in the error model are set to 0.08 and 0.02, respectively. Thus, the steady state probabilities P_G and P_B in the good and bad states are 0.8 and 0.2, respectively



(a) Network throughput

(b) Satisfaction index

(c) Fairness index

(d) Relative network throughput

(e) Satisfaction index for different P_G values

(f) Fairness index for different P_G values

Figure 3. Simulation results

Fig. 3(a) (see black bars) shows the network throughput. We find both CBCP and TBCP have similar network throughput. This is because the throughput of each multihop flow is determined by the area with the largest congestion degree[1] it passes through (here we call the area the "bottleneck cluster[2]" in the case of credit-based schemes, and "bottleneck interference range[3]" in the timestamp-based schemes), and not solely determined by each cluster (or interference range). Another interesting finding is that TBP has higher network throughput than CBP. This is because the proposed timestamp-based scheme only allows nonzero *Q-size* nodes to exchange transmission order messages, and thus it is more

[1] The congestion degree of a cluster is defined as the total reservation level recorded by the scheduler.

[2] The bottleneck cluster of a network is the cluster that is the most congested, i.e., with the highest *Resv* value.

flexible in slot allocation. Figs. 3(b) and 3(c) (see black bars) show their satisfaction and fairness indices, respectively. We see that both mechanisms still satisfy the demands of most QoS flows in this scenario, but have low fairness indices. This is because each multihop flow's throughput is constrained by the bottleneck cluster. Flows passing through the less congested area will have larger bandwidths, and vice versa. Fortunately, this problem can further be improved when node mobility is supported, as shown in the next case.

Next, we make the five multihop flows with mobility. The Min and Max speeds are set to be 10 meters per sec and 30 meters per sec, respectively. The network throughput of each approach is shown in Fig. 3(a), and the satisfaction and fairness indices of each approach are shown in Figs. 3(b) and 3(c), respectively (see white bars). We find that mobility causes broken routes more frequently, thus degrading network throughput. However, node mobility may increase the probability for a flow to change the area in which it is located (i.e., cluster for the credit-based schemes and interference range for time-stamp based mechanisms). This helps improve global fairness.

Finally, we study the impact of channel errors on the proposed approaches. We vary the value of P_G but fix all other parameters. The relative network throughput of each approach is shown in Fig. 3(d). As the value of P_G increases, the relative network throughput increases. The reason is that a larger P_G value means more good slots, leading to more successful transmissions and higher network throughput. The satisfaction and fairness indices of each setting are shown in Figs. 3(e) and 3(f), respectively. Since a larger P_G value makes more good slots, the probability to satisfy QoS flows is higher. Thus, it gives a higher satisfaction index. A larger P_G value also increases the opportunity for a node to be compensated after its channel recovers, and thus has a higher fairness index.

IV. Performance Analysis

In this section, we analyze the long-term throughputs of CBCP and TBCP. In this analysis, we do not consider node mobility, and ignore spatial channel reuse. Each mechanism is analyzed with the two-state channel errors.

A. Credit-Based Compensation Protocol (CBCP)

We calculate the throughput of CBCP. Suppose that the network is partitioned into m clusters. Without loss of generality, the m clusters are sorted in a descending order of the congestion degree, i.e., $Resv(1) > Resv(2) > \ldots > Resv(m)$. Let N_i denote the number of flow segments in cluster i, and $N_{i,j}$ denote the number of flow segments in both clusters i and j. Let S_i denote the set of flows in cluster i, and $S_{i,j}$ denote the set of flows in both clusters i and j. Assume that the scheduler of cluster i manage N_i flow segments. The residual bandwidth shared by flow f in cluster i is calculated as $y_f^i = \frac{1 - \sum_{j=1}^{N_i} Resv_j}{N_i}$, and the total bandwidth used by flow f in cluster i is given by $z_f^i = Resv_f + y_f^i$. The throughput of flow f is defined as $W_f = \frac{P}{T}$, where P is the number of packets received at the destination node, and T is the measurement window size.

1) Estimated Throughput: We consider the two-state error model. The average number of error-free slots obtained by each node is $T \times P_G$, where P_G is the steady state probability of the good state. For the most congested cluster, i.e., cluster 1, each flow segment's throughput is proportional to its z_f^1, as shown in (4).

$$W_f = \frac{T \times P_G \times \left(Resv_f + \frac{1 - \sum_{j=1}^{N_1} Resv_j}{N_1} \right)}{T} = P_G \times z_f^1, \quad f \in S_1 \qquad (4)$$

Similarly, the throughputs of those flows in cluster 2 but not in cluster 1 can be calculated as follows.

$$W_g = P_G \times \left(Resv_g + \frac{1 - \sum_{j \in S_{1,2}} z_j^1 - \sum_{k \in (S_2 - S_{1,2})} Resv_k}{N_2 - N_{1,2}} \right), \quad g \in (S_2 - S_{1,2}) \qquad (5)$$

Assume the bottleneck of flow segment h is in cluster r. Let Q denote the set of flows in cluster r but their bottlenecks are in more congested clusters than cluster r (i.e., clusters $1, 2, \ldots,$ or $r-1$), and R be the set of flows in less congested clusters (i.e., $r+1, r+2, \ldots, m$). For all flows in Q, z_f^b indicates the bandwidth share of flow segment f in its bottleneck cluster b. Let N_R denote the number of flow segments in set R. Therefore, the throughput of flow segment h is expressed as

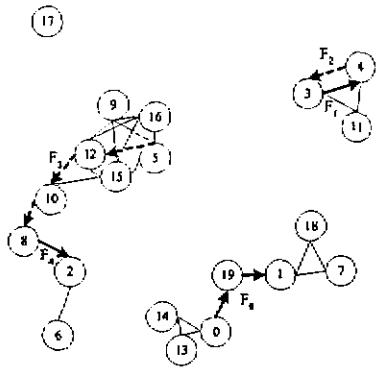
$$W_h = P_G \times \left(Resv_h + \frac{1 - \sum_{i \in Q} z_i^b - \sum_{j \in R} Resv_j}{N_R} \right), \quad h \in R \qquad (6)$$

For those flows passing through both clusters i and j, and $i<j$, their shares of residual bandwidth are constrained by the more congested cluster, i.e., cluster i. Thus, they will release some resources to the flows in cluster j but not in cluster i. Based on this concept, to calculate the throughputs of flows bottlenecked

[3] The bottleneck interference range of a network is the 2-hop area with the highest total flow weight.

at cluster j, we should first recursively derive the flow segments' throughputs bottlenecked at cluster *1, 2, ..., (j-1)*.

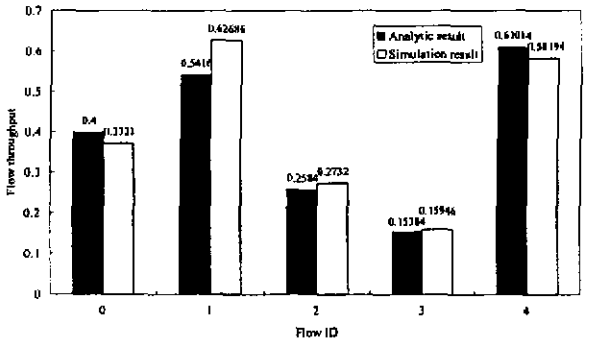
Therefore, for a multihop flow, its estimated flow throughput is $min\{ W_{i \in P} \}$, where i is a path node and P is the flow path.



(a) Network topology

FID	S	D	Type	Resv	Route	Single-hop flows		
						S	D	Sch
0	0	1	G	0.495	0-19-1	0	19	0
						19	1	0
1	3	4	G	0.354	3-4	3	4	3
2	4	3	B	0	4-3	4	3	3
3	5	8	B	0	5-12-10-8	5	12	5
						12	10	2
						10	8	2
4	8	2	G	0.423	8-2	8	2	2

(b) Flow information



(c) Result comparisons

Figure 4. Analytical result vs. simulation result for CBCP

2) Analytical vs. Simulation Results: We verify our analysis via simulation. Five flows are randomly generated in the same network as in Sec. III: three guaranteed flows and two best effort flows. The network topology and flow information are shown in Figs. 4(a) and 4(b). The comparison of the analytical result with the simulation result is shown in Fig. 4(c). We see that our analytical model provides very good estimation for the performance of CBCP.

B. *Timestamp-Based Compensation Protocol (TBCP)*

In an ad hoc network, a transmitting node, say node n, will interfere with another node, say n', due to the following two reasons:

(1) Node n' is within the transmission range of node n, and is the receiver of another flow segment, simultaneously.

(2) Node n' is within the transmission range of node n's corresponding receiver, and is the sender of another flow segment.

Unlike CBCP, TBCP has no schedulers to implement coordination. When a node is transmitting, those nodes within its two-hop are interfered. Therefore, the following analysis deals with this issue by calculating a flow's normalized weight.

1) Estimated Throughput: We consider the two-state error model. Assume there are n nodes in the ad hoc network. The average number of error-free slots for a node is $T \times P_G$, where T is the measurement window, and P_G is the steady state probability of the good state. Let S_m indicate the set of flow segments passing through node m, and $S_{m'}$ represents the set of flow segments within node m's interference range. For each flow segment f passing through node m, its bandwidth share is $x_f = Resv_f + \frac{1 - \sum_{j=1}^{M} Resv_j}{M}$, where M is the number of flow segments managed by node m. The normalized flow weight of f is calculated as $\omega_f = \frac{x_f}{\sum_{j \in \{S_m \cup S_{m'}\}} x_j}$. The corresponding flow throughput of flow f is given by

$$W_f = \frac{T \times P_G \times \omega_f}{T} = P_G \times \omega_f, \ f \in S_m \quad (7)$$

For a multihop flow, its estimated flow throughput is $min\{ W_{i \in P} \}$, where i is a path node and P is the flow path.

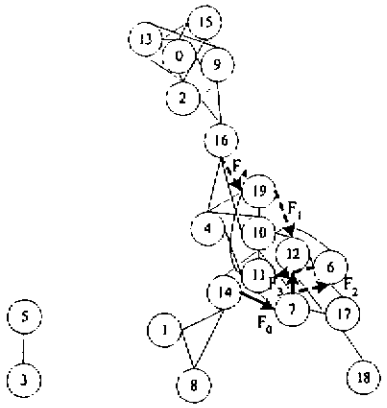
2) Analytical vs. Simulation Results

We randomly generate five flows in the simulation: four best effort flows and one guaranteed flow. The node graph is in Fig. 5(a). Furthermore, the information of flows and the corresponding interfering nodes is shown in Fig. 5(b). The comparison with simulation results is shown in Fig. 5(c). Again, we see that our analytical model provides accurate estimation for the performance of TBCP.

V. Conclusion

In this paper, we discuss the channel compensation issue of fair scheduling and propose one such mechanism called TBCP for mobile multihop networks. TBCP supports multihop flows, and performs well when node mobility is supported. We describe the detailed operation of TBCP, and conduct simulations to evaluate its performance. From the simulation results, we demonstrate that TBCP satisfies QoS flow demands and provides global fairness for best effort flows. We also show

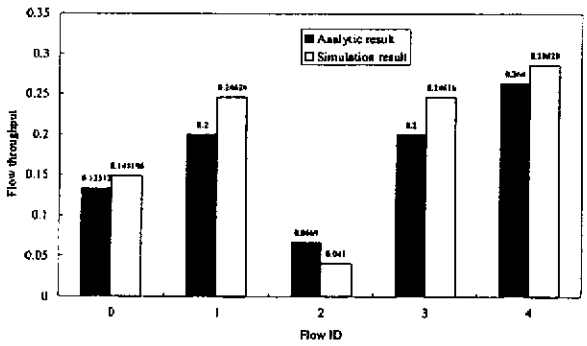
that mobility helps break the locality problem for multihop



(a) Network topology

FID	S	D	Type	Resv	Route	Single-hop flow	Interfering node
0	14	12	G	0.331	14-7-12	14-7	1,4,6,8,10,11,12,16,17,18,19
						7-12	1,4,6,8,10,11,14,16,17,19
1	19	12	B	0	19-12	19-12	2,4,6,7,9,10,11,14,16,17,18
2	7	6	B	0	7-6	7-6	1,4,8,10,11,12,14,16,17,18,19
3	6	11	B	0	6-11	6-11	1,4,7,10,12,14,16,17,18,19
4	16	19	B	0	16-19	16-19	0,2,4,6,7,8,9,10,11,12,13,14,15,17

(b) Information of flows and the corresponding interfering nodes



(c) Result comparisons

Figure 5. Analytical result vs. simulation result for TBCP

flows. Finally, we analyze the flow throughputs of CBCP and TBCP, and verify the analytical result with simulation. The results show that our analytical results provide accurate performance estimations for the proposed mechanisms.

This paper addresses fair scheduling for multimedia mobile ad hoc networks with channel errors. The QoS demands of flows basically are described via bandwidth. In the future, we will investigate fair scheduling with other QoS metrics such as delay or delay jitter for ad hoc networks.

REFERENCES

[1] H. Luo and S. Lu, "A topology-independent fair queueing model in ad hoc wireless networks," *IEEE International Conference on Network Protocol*, Nov. 2000.

[2] H. Luo and S. Lu, "A self-coordinating approach to distributed fair queueing in ad hoc wireless networks," *IEEE INFOCOM*, Apr. 2001.

[3] H. L. Chao, and W. Liao, "Fair scheduling with QoS support in ad hoc networks," *IEEE Conference on Local Computer Networks*, Nov. 2002.

[4] P. Goyal, H. M. Vin, and H. Cheng, "Start-time fair queueing: a scheduling algorithm for integrated services packet switching networks," *IEEE/ACM Transaction on Networking*, vol. 5, no. 5, 1997.

[5] Y. Cao and V. O.K. Li, "Scheduling algorithms for broadband wireless networks," *Proceedings of the IEEE*, vol. 89, no. 1, Jan. 2001, pp. 76-87.

[6] T. S. Eugene Ng, I. Stoica and H. Zhang, "Packet fair queueing algorithm for wireless networks with location-dependent errors," *IEEE INFOCOM*, 1998.

[7] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Transaction on Networking*, Vol. 7, No. 4, Aug 1999.

[8] S. Lee, K. Kim, and A. Ahmad, "Channel error and handoff compensation scheme for fair queueing algorithms in wireless networks," *IEEE ICC* 2002.

[9] Y. Cao and V. O.K. Li, "Wireless packet scheduling for two-state link models," *IEEE Globecom*, 2002.

[10] L. Huang and T.-H. Lai, "On the scalability of IEEE 802.11 ad hoc networks," *Proceedings of MobiHoc*, 2002.

[11] T.-H. Lai and D. Zhou, "Efficient and scalable IEEE 802.11 ad-hoc-mode timing synchronization function," *Proceedings of the 17th IEEE International Conference on Advanced Information Networking and Applications*, 2003.

[12] C. E. Perkins and E. M. Royer, "Ad-hoc on demand distance vector routing," *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, Feb. 1999, pp. 90-100.

[13] C. E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers," *ACM SIGCOMM* '94, pp. 234-244.

[14] S. Chen and K. Nahrstedt, "Distributed quality-of-service routing in ad hoc networks," *IEEE J. Select. Areas Comm.*, vol. 17, no. 8, Aug. 1999, pp.1488-1505.

[15] C. R. Lin and J.-S. Liu, "QoS routing in ad hoc wireless networks," *IEEE J. Select. Areas Commun.*, Vol. 17, No. 8, Aug. 1999, pp.1426-1438.

[16] J. Yoon, M. Liu, and B. Noble, "Random waypoint considered harmful," *IEEE INFOCOM*, Apr. 2003.

[17] R. K. Jain, D. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer system," *DEC Technical Report DEC-TR-301*, 1984.

[18] E. O. Elliot, "Estimates of error rates for codes on burst-noise channels," *Bell Systems Technical Journal 42*, Sep. 1963

Hsi-Lu Chao was born in Taiwan in 1970. She received the B.S. degree in Electronic Engineering from Fengchia University, Taiwan, in 1992, the M.S. degree in Electrical Engineering from the University of Southern California, Los Angeles, in 1996, and the Ph.D. degree in electrical Engineering from the National Taiwan University, Taiwan, in 2004. Since August 2004, she has been an Assistant Professor, Department of Computer and Information Science, National Chiao Tung University. Her research interests include wireless communication and QoS issues in ad hoc networks.

Wanjiun Liao received the BS and MS degrees from National Chiao Tung University, Taiwan, in 1990 and 1992, respectively, and the Ph.D. degree in Electrical Engineering from the University of Southern California, Los Angeles, California, USA, in 1997. She joined the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, as an Assistant Professor in 1997. Since August 2000, she has been an Associate Professor. Her research interests include wireless networks, overlay networks, and broadband Internet.

Dr. Liao is actively involved in the international research community. She is currently an Associate Editor of IEEE Transactions on Wireless Communications. Dr. Liao has received many research awards. She was a recipient of the Outstanding Research Paper Award in Electrical Engineering at the University of Southern California in 1997. Two papers she co-authored with her students received the Best Student Paper Award of the First IEEE International Conferences on Multimedia and Expo (ICME) in 2000, and the Best Paper Award of the First International Conference on Communication, Circuits and Systems (IEEE Communication Press) in 2002. Dr. Liao was elected as one of Ten Distinguished Young Women in Taiwan in 2000 and is

listed in the Marquis Who's Who in 2001-2003, and the Contemporary Who's Who in 2003

Modeling the Behavior of Flooding on Target Location Discovery in Mobile Ad Hoc Networks

Jia-Chun Kuo and Wanjiun Liao
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
Email: wjliao@ntu.edu.tw

Abstract- In this paper, we model the behavior of packet forwarding via a multihop path in mobile ad hoc networks. In our analysis, we consider a densely populated network and assume a flooding protocol for packet forwarding. We find that the behavior of packet forwarding in such an environment can be regarded as a dropping a stone into a lake, and then counting the number of ripples moving out from the source (i.e., the center) to the destination. What makes the problem complicated is the node mobility during the packet forwarding. As a result, we cannot solely count the number of ripples in between as the number of hops in the path on which the packet traverses. We then derive the probability distribution function of hop counts for packet forwarding, accounting for node movements. Based on the analytical model, we then evaluate several different types of flooding mechanisms commonly adopted for target searching in ad hoc networks. Compared with existing work, which assumes a snapshot of the network and all nodes are static in the analysis, our analytical framework provides more insights for the study of efficient flooding in mobile ad hoc networks.

Keywords: mobile ad hoc networks, multihop packet forwarding, modeling

I. INTRODUCTION

A mobile ad hoc network is a multihop wireless network. In such a network, each node in the network plays both roles of a host and a router. Data packets are then relayed via multiple hops, without the support of fixed infrastructure. Nodes may be roaming around during packet forwarding. Whenever a node on the routing path has moved away, the path must be rerouted.

In this paper, we study the behavior of packet forwarding on a multihop path in mobile ad hoc networks. The behavior of packet forwarding in mobile ad hoc networks may be affected by many different factors, such as node mobility, spatial distribution of nodes, the routing algorithm, and network topology. The impacts of all factors on the performance of packet forwarding have been widely studied in the literature [1-6]. In [1], many mobility models are surveyed. In [2], the impact of packet routing protocols on the end-to-end transmission delay is discussed. The authors also claim that node density affects the delay. In [3], the authors find that node degree and connectivity both have impact on the delay, and thus network topology is also an important factor to consider. In addition, they claim that so long as the transmission range of nodes is properly chosen, the network connectivity is ensured.

In [4], both the length and duration of a movement epoch and the direction of travel according to the random waypoint mobility model are proven to be nonuniform and affected not only by the spatial position of the node but also the simulation region. In [5], the node spatial distribution based on the random waypoint model is analyzed. In [6], a forwarding technique, called GeRaF, is proposed based on the geographic location of nodes. The circular intersection area between the transmission ranges of adjacent nodes and the distance of the destination node are used to derive the analytical bounds for the multihop performance of GeRaF. However, the assumption that the geographic location of all nodes is known *a priori* is not realistic in general.

In this paper, we model the behavior of packet forwarding on a multihop path in mobile ad hoc networks. In particular, we derive the probability distribution function of the number of hops in a multihop path for packet forwarding, allowing node movements. Based on the analytical model, we further evaluate several different types of flooding mechanisms commonly adopted for target searching in ad hoc networks, including blind flooding (e.g., flooding mentioned in [7]), two-tier flooding (e.g., DSR [8]), and expansion-ring flooding (e.g., AODV [9]), with respect to their cost and target searching latency. Compared with [10], which assumes a snapshot of the network and all nodes are static in the analysis, our analytical framework provides more insights for the study of efficient flooding in mobile ad hoc networks, because we consider node mobility and many network parameters (including distance between nodes, transmission radius and the processing time of nodes) together.

The rest of the paper is organized as follows. In Sec. II, the system model and assumptions are described, and an analytical model is derived. In Sec. III, different types of flooding schemes are evaluated based on the analytical model. Finally, the paper is concluded in Sec. IV.

II. ANALYTICAL MODEL

A. System Model and Assumptions

The system we consider is a homogeneous mobile ad hoc network, in which all nodes are equipped with the same transmission capacity and data are forwarded via a multihop path. Each node is continuously roaming. We do not assume a certain mobility model, but capture the moving behavior by means of the circles centered at the initial location of roaming nodes. In other words, the

circles for those fast-moving nodes grow faster than those slow-moving ones. We consider a densely populated network. Thus, the isolated node problem will not happen. We also don't consider the interference and handoff problems.

In this paper, we model the distance between source and destination in terms of the number of hops for packets to travel, taking node movements into account. Data packets are forwarded via flooding, considering "flooding" being widely used for target searching in the route discovery phase of most on-demand ad hoc routing protocols. The notation used in the analysis is summarized in Table I.

a	The radius of the transmission range of each node.
p	Packet size (bits)
c	Link speed (bps)
t_x	Packet transmission time, $t_x = \frac{p}{c}$
t_i	Processing delay at the ith node counted from the source on a multihop path
L	The random variable representing the initial distance between the source and the destination, with probability density function $f_L(l)$.
r_c	The growing rate of the circle centered at the destination
H	The random variable representing the number of hops to be traversed for packet forwarding
$F_H(h_0)$	The cumulative distribution function of H
$P_H(k)$	The probability density function of H

TABLE I. NOTATION IN THE ANALYSIS

B. The Hop Counts for Multihop Packet Forwarding

In our model, we consider a mobile ad hoc network with high node density. In other words, no matter which direction the next hop is, there is always a node there to forward the packet via flooding. Fig. 1 illustrates an example to forward a packet over a multihop path in our system, where R denotes the radius of the transmission range of each node. Interestingly, such phenomena makes the behavior of packet forwarding along a multihop path for a source-destination pair just like dropping a stone into a lake and generating ripples moving out from the source. What interests us is the number of ripples between the source and the destination, which is equal to the number of hops for the packet to traverse in between.

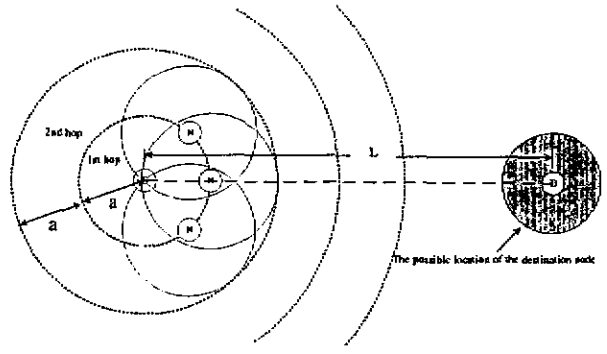
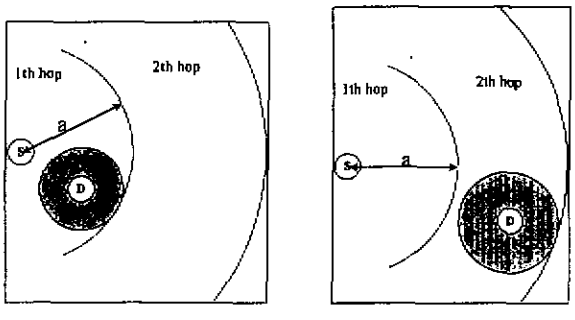


Figure 1. The behavior of multihop packet forwarding

However, this hop count can only be regarded as the initial distance between the source and the destination, because the destination node may be moving around during the packet forwarding. If the destination node is moving towards the source, the final hop count will be smaller than the initial one, and vice verse. What makes this problem complicated is that it is hard to predict which direction the destination node is moving at any moment. The only thing we can be sure is that if the movement speed is higher, the possible area in which the destination node may be located become wider. To capture such an effect, we adopt a circle centered at the destination to represent the possible area in which the destination node may be located at any moment. The higher node mobility then results in a larger circle. As a result, once the moving speed is given, we can predict how the circle grows.



(a) $H \le 1$ (b) $H \le 2$

Figure 2. Two examples of packet forwarding

Let H denote the random variable of the hop count for the source-destination pair of interest, and $F_H(h_0)$ be the cumulative probability distribution of H. Given the initial distance L, the number of hops the packet traverses is at most one only if the following condition holds (as shown in Fig. 2 (a)):

$$a \ge L + r_c \cdot t_x, \text{ i.e., } L \le \frac{ac - r_c p}{c},$$

where a is the node transmission radius, r_c is the circle growing rate, and t_x is the packet transmission delay in one hop (the propagation delay is ignored). Thus, the probability with H at most one is

$$F_H(1) = \Pr(H \le 1) = \Pr(L \le \frac{ac - r_c p}{c}).$$

Now we calculate the probability of the hop count greater than one. Using a two-hop path as an example (Fig. 2 (b)), we have $2a \ge L + r_c \cdot (2t_x + t_1)$, where t_1 is the processing delay of the first relaying node from the source. Since $t_x = \frac{p}{c}$, we obtain $t_1 \le \frac{2ac - Lc - 2r_c p}{r_c c}$. In addition, $0 \le t_1$, thus we have $\frac{2ac - Lc - 2r_c p}{r_c c} \ge 0$. Due to $r_c c \ge 0$, $2ac - Lc - 2r_c p \ge 0$ must hold, i.e., $L \le \frac{2ac - 2r_c p}{c}$. Thus, the probability that the packet traverses at most two hops is expressed as

$$F_H(2) = \int_0^{\frac{2ac-2r_c p}{c}} P(t_1 \le \frac{2ac - l_0 c - 2r_c p}{r_c c} \mid L = l_0) P(L = l_0) dl_0$$

Similarly, the packet traversed at most N hops (i.e., N-1 intermediate relaying nodes) only when the following condition holds: $aN \ge L + r_c \cdot \left(N \cdot t_x + \sum_{i=1}^{N-1} t_i \right)$, where t_i is the processing delay of the ith relaying node counting from the source on the path, $i = 1, 2, \cdots, N-1$. Substituting $t_x = \frac{p}{c}$ into the inequality, we obtain $\sum_{i=1}^{N-1} t_i \le \frac{Nac - Lc - Nr_c p}{r_c c}$. Since $0 \le \sum_{i=1}^{N-1} t_i$, we get $L \le \frac{Nac - Nr_c p}{c}$. Thus, the probability of at most N-hop traveling is expressed as

$$F_H(N) = \int_0^{\frac{Nac-Nr_c p}{c}} P\left(\sum_{i=1}^{N-1} t_i \le \frac{Nac - l_0 c - Nr_c p}{r_c c} \mid L = l_0 \right) P(L = l_0) dl_0.$$

Assume that $\{ t_i : i = 1, 2, \cdots \}$ is a set of independent, identically distributed random variables with probability density function $f(t)$. Let $f_L(l_0)$ be the probability density function of L. The probability distribution function that the packet can travel at most N hops $F_H(N)$ can be further expressed as

$$\frac{1}{2\pi j} \int_{\sigma - j\infty}^{\sigma + j\infty} \frac{[f^*(s)]^{N-1}}{s} \cdot \left[e^{\frac{s(Nac - Nr_c p)}{r_c c}} \cdot \int_0^{\frac{Nac - Nr_c p}{c}} f_L(l_0) \cdot e^{-\frac{s l_0}{r_c}} dl_0 - \int_0^{\frac{Nac - Nr_c p}{c}} f_L(l_0) dl_0 \right] ds$$

where $f^*(s)$ is the Laplace transform of $f(t)$.

Let $P_H(N)$, $N = 1, 2, \cdots$, denote the probability of exactly N-hop traversal for packet forwarding, i.e., $P_H(N) = \Pr(H = N)$. Therefore,

$$P_H(N) = \begin{cases} F_H(N), & N = 1 \\ F_H(N) - F_H(N-1), & N \ge 2 \\ 0, & \text{otherwise} \end{cases}.$$

III. THE MODEL VERIFICATION

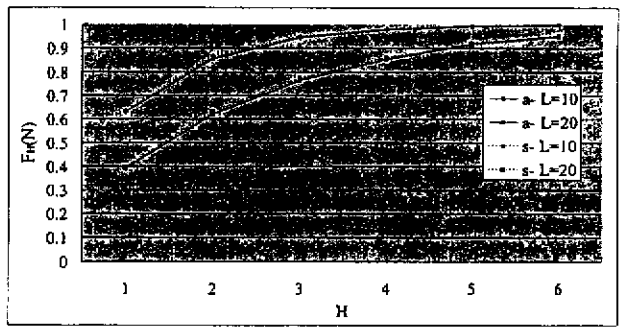


Figure 3. The analytical and simulation results for $F_H(N)$

To validate the analytical model by simulations, we plot the cumulative distribution function $F_H(N)$ with two different settings of L. The settings of other parameters are: t_i=1sec, a=10 units and r_c=0.5 units per sec. As shown in Fig.3, the analytical and simulation results fit very well.

IV. THE PERFORMANCE OF DIFFERENT FLOODING SCHEMES ON TARGET DISCOVERY

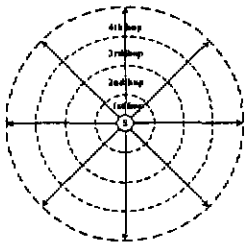
We have modeled the behavior of a packet being flooded from source to destination in a dense mobile ad hoc network. Based on the analytical model, we can evaluate the cost and latency of different flooding schemes for searching a target destination under various system parameters. In what follows, three different types of flooding schemes for searching a target destination are considered.

(1) Blind flooding: the entire network is flooded (e.g., [7]).
(2) Two-tier flooding: the finite-hop neighbors are searched first. If the target is not found, the entire network is flooded. The searching packet for target location in DSR [8] is an example of two-tier

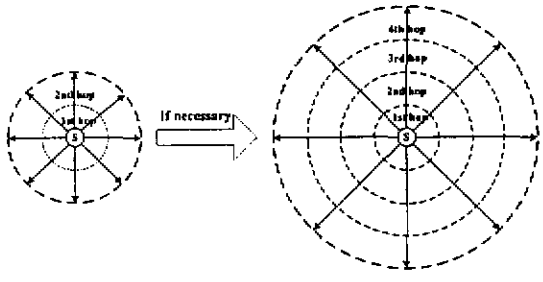
flooding.

(3) Expansion-ring flooding: the source incrementally enlarges the searching range from an initial value to a predefined threshold. If the target is still not found, the entire network is flooded. The searching packet for target location in AODV [9] is an example of expansion-ring flooding.

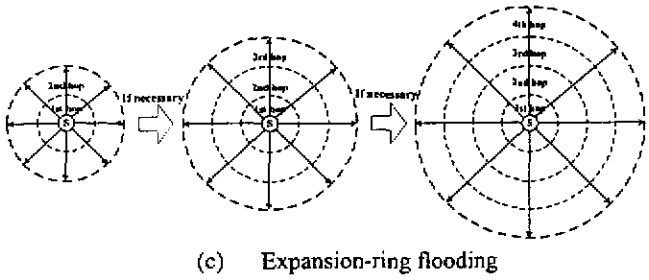
Fig. 4 illustrates the use of the three flooding schemes to search a target in a 4-hop ad hoc network centered at node *S*. The searching range of the network is controlled by the "Time-to-Live" (TTL) parameters (in terms of the number of hops) in the packet header. Fig. 4 (a) illustrates blind flooding, in which a searching packet is flooded to the entire network. Fig. 4 (b) shows two-hop two-tier flooding, in which a searching packet is first broadcast with TTL=2. If the target destination is not found, another searching packet is flooded to the entire network. Fig. 4 (c) depicts expansion-ring flooding, in which a searching packet with initial TTL, say, 2, is first broadcast. In case the target is not found, the TTL is incremented by one and a searching packet is broadcast with the updated TTL. This is repeated until the predefined threshold (say 3) is reached. If the search with the threshold TTL still fails, a final searching packet is flooded to the entire network.



(a) Blind flooding



(b) Two-tier flooding



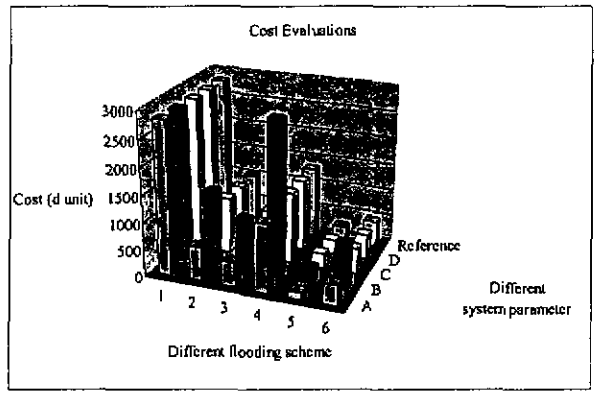
(c) Expansion-ring flooding

Figure 4. Three different flooding schemes

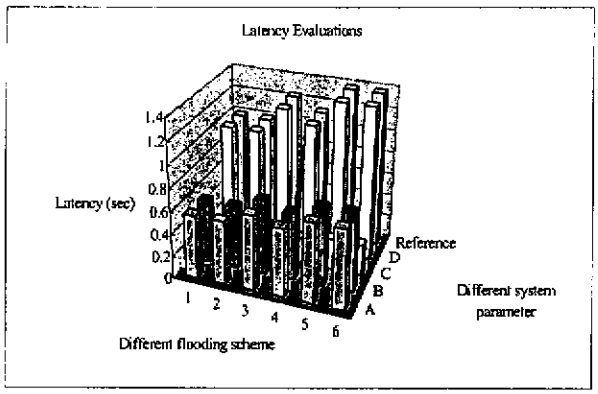
We first evaluate the cost of sending a packet for each flooding scheme, where the cost is defined as the number of times the packet is broadcast by each scheme. A packet is referred to as k-hop limited if the packet is propagated up to k hops from the source. The eligible nodes for a k-hop limited packet are those located in the area with transmission radius $(k-1)a$, $k=1,2,\cdots$, but not including those exactly k hops away, from the source. For example, the set of nodes eligible for a 1-hop limited packet include the source node only; the set of nodes eligible for a 2-hop limited packet include the nodes located in the radius of one hop, i.e., $\pi a^2 d$, where a is the transmission radius and d is the node density. Similarly, the set of nodes eligible for a k-hop limited packet includes the nodes located in the area of $(k-1)^2 \pi a^2 d$, $k=1,2,\cdots$. Note that since we consider a dense ad hoc network, $d >> 1$.

We consider six different schemes. Scheme 1 represents blind flooding. Schemes 2-4 indicate two-tier flooding with different settings to limit the flooding of the initial packet: for Scheme 2, the first searching packet is one-hop limited, for Scheme 3, the first packet is two-hop limited, and for Scheme 3, the first packet is 3-hop limited. Schemes 5-6 are for expansion-ring flooding with different settings of the initial packet: for Scheme 5, the first packet is one-hop limited, and for Scheme 6, the first packet is two-hop limited. There is no predefined threshold for Schemes 5 and 6.

Fig. 5 (a) shows the cost of each flooding with different network topologies as shown in Table II, in which each column indicates a different network parameter, and each row shows a different topology setting. The reference row (i.e., Ref) indicates the baseline of comparison for different flooding schemes. Compared to the baseline, row A achieves 33% reduction in L (i.e., initial distance between source and destination), row B 50% increase in a (transmission radius of a mobile node), row C 80% reduction in t_i (i.e., processing delay of an intermediate node), and row D 80% reduction in r_c (i.e., node mobility). Fig. 5 (b) shows the searching latency of each scheme. From Fig.5, we have some observations.



(a) Cost



(b) Latency

Figure 5. Performance evaluation

	L (unit)	t_i (sec)	a (unit)	r_c (unit/sec)
A	6.67	1	10	0.5
B	10	1	15	0.5
C	10	0.2	10	0.5
D	10	1	10	0.1
Ref	10	1	10	0.5

TABLE II. NETWORK PARAMETERS

1) Trade-off between the flooding cost and the searching latency

Fig. 5 shows the trade-off between the flooding cost and the searching latency. For blind flooding, the packet is broadcast to the entire network but never retransmitted. Thus, it has the highest flooding cost but the lowest latency. For two-tier flooding, the packet is hop-limited in its first probing, followed by blind flooding if the first trial fails. As a result, it significantly reduces the cost, without dramatically increasing the latency. For expansion-ring flooding, the flooding is done in an incrementally increasing fashion. Thus, it has the lowest cost, but at the expense of high latency due to multiple re-flooding.

2) The impact of different parameters on each flooding scheme

From Fig.5, we learn that only the initial distance between source and destination and the transmission radius of each node are of interest to us. The other parameters affect the metrics of all schemes in a similar way. For example, a reduction in L causes a decrease in the latency for all schemes; similarly, an increase in a decreases the latency of all schemes.

We examine the impacts of a change in L (i.e., row A) and a change in a (i.e., row B) on the flooding cost of all schemes. Blind flooding is immune to the changes in the values of L and a, because it always floods to the entire network no matter what happens. For the two-tier and the expansion-ring schemes, the observation is that (a) a reduction in L can lower (and an increase in a can raise) the flooding cost of both types of schemes, and (b) a smaller initial search range saves more flooding cost (see columns 2 and 5 for example, which correspond to the cases of one-hop limited). Such observations match our intuition.

V. CONCLUSION

In this paper, we model the behavior of packet forwarding on a multihop path for mobile ad hoc networks with high node density. The node mobility is captured by a circle centered at the initial location of the destination node. Data packets are forwarded via flooding, considering "flooding" being widely used for target searching in the route discovery phase of most on-demand ad hoc routing protocols.

We find that the behavior of multihop packet forwarding can be regarded as dropping a stone into a lake and then counting the number of ripples between the source and the destination as the initial hop count. We then derive the probability distribution functions of hop counts in a multihop path, considering that the destination node may be roaming during packet forwarding. Based on the analytical model, we can then evaluate the cost and latency of different flooding schemes for target discovery.

ACKNOLWDGEMENT

This project was supported in part by National Science Council (NSC), Taiwan, under a Center Excellence Grant NSC93-2752-E-002-006-PAE, and in part by NSC under Grant Number NSC93-2213-E-002-132.

REFERENCES

[1] T.Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research," *Wireless Comm. and Mobile Computing (WCMC)*, vol. 2, no. 5, pp.483-502, 2002.

[2] C.E. Perkins, E.M. Royer, S.R. Das, and M.K. Marina, "Performance Comparison of Two On-Demand Routing Protocols for Ad Hoc Networks," *IEEE Personal Comm.*, Vol.8, Issue 1, pp.16-28, Feb. 2001.

[3] C. Bettstetter, "On the Minimum Node Degree and Connectivity of a Wireless Multihop Network," *Proc. ACM MobiHoc*, June, 2002.

[4] C. Bettstetter, H. Hartenstein, and Xavier Perez-Costa, "Stochastic Properties of the Random Waypoint Mobility Model: Epoch Length, Direction Distribution, and Cell Change Rate," *Proc. ACM MSWiM*, Sep. 2002

[5] C. Bettstetter, G. Resta, and P. Santi, "The Node Distribution of the Random Waypoint Mobility Model for Wireless Ad Hoc Networks," *IEEE Trans. on Mobile Computing*, vol. 2, no. 3, Jul-Sep, 2003.

[6] M. Zorzi and R. R. Rao, "Geographic Random Forwarding (GeRaF) for Ad Hoc and Sensor Networks: Multihop Performance," *IEEE Trans. on Mobile Computing*, vol. 2, no. 4, Oct-Dec, 2003.

[7] C. Ho, K. Obraczka, G. Tsudik, and K. Viswanath, "Flooding for reliable multicast in multi-hop ad hoc networks," *Proc. the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communication (DIALM)*, pp. 64-71, 1999.

[8] D. Johnson and D. Maltz, "Dynamic Source Routing in Ad Hoc Wireless Networks," T. Imielinski and H. Korth, Eds. *Mobile Computing*, Ch. 5, Kluwer, 1996.

[9] C. E. Perkins, E. M. Royer, and S. R. Das, "Ad Hoc on Demand Distance Vector (AODV) Routing," IETF Internet Draft.

[10] Z. Cheng, W. B. Heinzelman, "Ad Hoc and Sensor Networks: Flooding Strategy for Target Discovery in Wireless Networks," *Proc. ACM Modeling Analysis and Simulation of Wireless and Mobile Systems*, Sep. 2003.

Performance Analysis of Reliable MAC-Layer Multicast for IEEE 802.11 Wireless LANs

Chong-Wei Bao and Wanjiun Liao
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
Email: wjliao@ ntu.edu.tw

Abstract- In this paper, we study reliable multicast at the MAC layer for IEEE 802.11 Wireless LANs. In IEEE 802.11, multicast transmissions are unreliable in the sense that multicast frames are transmitted from the Access Point (AP) without an ACK being returned from each receiver as in unicast transmissions. As a result, transmitted multicast frames may be lost due to collisions or errors. There are two types of reliable multicast at the MAC layer proposed for 802.11 in the literature: one is ACK-based and the other is a leader-based mechanism. For the ACK-based mechanism, the AP monitors the frame reception progress of all receivers and retransmits frames whenever no ACK is received from any receiver; for leader-based mechanism, the AP retransmits frames only when no ACK frame is received from the selected leader. We analyze their performance in terms of frame holding time at the AP, and conduct simulation to validate our analytical model. We also propose a new channel acquisition and multicast notification mechanism called RTS/CTS/SEQ to improve the performance of existing work for reliable multicast in IEEE 802.11 WLANs.

Keywords: reliable multicast, IEEE 802.11 WLANs

I. INTRODUCTION

In this paper, we analyze the performance of two reliable MAC-layer multicast mechanisms for IEEE 802.11 Wireless LANs. In IEEE 802.11 [1], multicast transmissions are unreliable in the sense that multicast frames are transmitted from the Access Point without an ACK being returned from each receiver as in unicast transmissions. As a result, transmitted multicast frames may be lost due to collisions or errors. While reliable multicast in the transport layer has been an active research topic, this issue in the MAC layer for IEEE 802.11, has drawn much less attention. Currently, very few protocols, such as Leader-Based Protocol (LBP) [2], have been proposed for IEEE 802.11-based Wireless LANs.

LBP works as follows. A receiver is selected as the leader for the multicast group. The AP then sends a multicast RTS (denoted as m-RTS) frame to all receivers, and only the leader transmits a multicast CTS (denoted as m-CTS) frame in reply to the AP. The AP is then assured that the channel is granted and starts the transmission of a multicast data frame. The leader sends an ACK in reply if the data is received correctly, or sends a NAK otherwise. If any other receiver detects a transmission error, a NAK is sent. This NAK frame will collide with the ACK, if any, sent by the leader. This leads to the AP not hearing any ACK, and thus retransmitting the lost frame.

Applying LBP to IEEE 802.11 suffers from two problems. One is the hidden terminal problem due to the m-RTS in LBP serving both roles of channel acquisition and multicast notification. The other is the poor performance when the channel error rate is high. This is because the receivers in LBP cannot access the frame sequence number before the frame is received, as there is no such field in the structure of RTS/CTS frames for multicast. As a result, the performance of the error recovery scheme for LBP degrades, particularly for lossy channels.

In this paper, we analyze two types of reliable multicast mechanisms at the MAC layer for IEEE 802.11 WLANs: ACK-based (e.g., [3]) and leader-based (e.g., LBP in [2]). For the ACK-based mechanism, the AP monitors the frame reception progress of all receivers and retransmits frames whenever no ACK is received from any receiver; for leader-based mechanism, the AP retransmits frames only when no ACK frame is received from the selected leader. Thus, the ACK-based mechanism can be regarded as a centralized scheme, and the leader-based mechanism, a distributed one. Since LBP suffers from the hidden terminal problem, and has poor performance when the channel error rate is high, we will enhance LBP to solve these problems in this paper and called the enhanced version of LBP as ELBP. Note that since the work in [3] is designed for ad hoc networks, we will also modify it for IEEE 802.11 WLAN with APs, and called the modified version as "ACK-based Multicast Protocol (AMP)." We will then analyze both types of reliable multicast mechanisms based on AMP and ELBP, and verify the analytical model via simulations.

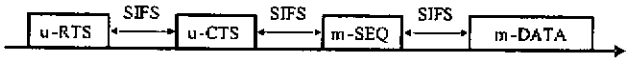
The rest of the paper is organized as follows. In Sec. II, the two reliable multicast mechanisms to be analyzed are described. In Sec. III, the performance of AMP and ELBP with respect to frame holding time is analyzed. In Sec. IV, the analytical model is verified via simulation. Finally, the paper is concluded in Sec. V.

II. RELIABLE MAC-LAYER MULTICAST

The issues of reliable MAC-layer multicast can be decomposed into two problems: (1) channel acquisition, and (2) error-recovery strategy. In what follows, we propose a channel acquisition mechanism, and enhance the two types of error-recovery schemes (i.e., ACK-based and LBP), which together will provide reliable multicast transmissions for IEEE 802.11 WLANs.

A. *Channel Acquisition and Multicast Notification*

In this paper, we propose a channel acquisition mechanism called RTS/CTS/SEQ for IEEE 802.11. In RTS/CTS/SEQ, the structures of both RTS and CTS frames remain intact. The SEQ is a new control frame defined as in Fig. 1, where TA represents the Transmitter Address, and Sequence Control shows the sequence number and the fragment number of the frame to be transmitted.



u-RTS / u-CTS : unicast RTS / CTS
m-SEQ / m-DATA : multicast SEQ / DATA

Figure 1. The format of an SEQ frame

The RTS/CTS/SEQ exchange works as follows. A receiver, say host X, is randomly selected to respond to the RTS/CTS exchange. The AP sends a unicast RTS frame to host X, and host X sends a unicast CTS frame in reply after an SIFS period. The AP multicasts an SEQ frame after an SIFS if the CTS frame is received correctly. No one needs to respond to this multicast SEQ. If the channel has been acquired after the previous u-RTS/u-CTS exchange, the m-SEQ will not collide and will thus achieve multicast notification. After another SIFS, a data frame is multicast by the AP. Error recovery described in Sec II. B then follows.

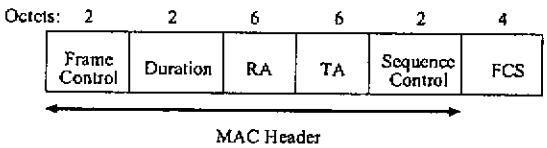


Figure 2. The RTS/CTS/SEQ exchange

B. *Error Recovery*

Two protocols are proposed for error recovery: (1) an ACK-based scheme called ACK-based Multicast Protocol (AMP), and (2) an enhanced LBP scheme called ELBP. Both protocols are based on the RTS/CTS/SEQ exchange for channel acquisition and multicast notification.

1) AMP

After the RTS/CTS/SEQ exchange and data multicast, the RAK/ACK exchange follows. The AP polls each multicast receiver in turn via a Request for ACK frame (denoted as RAK) and each receiver sends an ACK in reply after an SIFS. If no ACK is heard for a RAK, the AP records the unsuccessful receiver, say host Y, and contends the channel for frame retransmission. After the frame is retransmitted in multicast, the AP resumes the polling, starting from host Y. This process repeats until all multicast receivers have completed a RAK/ACK exchange.

2) ELBP

After the RTS/CTS/SEQ exchange and data multicast, the AP waits for an ACK from the leader selected by a leader election procedure as in [2]. As in LBP, if no ACK is heard from the leader within a period of time, the data is retransmitted in multicast.

However, the operation of non-leader hosts is different from LBP, in which non-leader hosts send a NAK to collide with the leader's ACK whenever the received frame is in error, regardless of whether this erroneous frame has been received successfully before or not, due to the unknown sequence number. In ELBP, because of the announcement from SEQ frames, non-leader hosts send a NAK only when the following multicast frame has not been received correctly yet, and stay quiet for any retransmitted frame which has been received successfully before. As a result, ELBP can effectively eliminate the redundant retransmissions caused by LBP.

Note that together with the RTS/CTS/SEQ exchange, ELBP can further eliminate the hidden terminal problem suffered by LBP. For LBP, a u-RTS from a hidden terminal of the leader, say H1, may collide with the AP's m-RTS, which may destroy the functionality of multicast notification between the area covered by the AP and H1. However, the multicast notification in ELBP is provided by SEQ frames, instead of m-RTS frames. Therefore ELBP and AMP are free from the hidden terminal problem thanks to the RTS/CTS/SEQ exchange.

III. FRAME HOLDING TIME AT AP

In this section, the performance of both AMP and ELBP with respect to frame holding time at the AP is analyzed. The frame holding time is defined as "*the time period from the start of an AP sending a certain multicast frame to the end of the successful reception by all members.*"

A. System Model and Assumptions

R	The number of multicast receivers in the group
N	The total number of stations in the WLAN, including the AP
M	The number of transmissions for all receivers to receive a frame correctly
T_D	The time to transmit an RTS, a CTS, an SEQ and a multicast data frame, i.e., $T_D = \text{RTS} + \text{CTS} + \text{SEQ} + \text{DATA} + 3\,\text{SIFS}$
T_{R+A}	The time to transmit a RAK and an ACK, i.e., $T_{R+A} = \text{RAK} + \text{ACK} + 2\,\text{SIFS}$
T_{ACK}	The time to transmit an ACK, i.e., $T_{ACK} = \text{ACK} + \text{SIFS}$
τ	The probability that a station transmits a frame in a slot
T_S	The time of a successful transmission slot
T_C	The time of a collision slot
T_{idle_slot}	The duration of an idle slot
T_{av_slot}	The average slot time observed by the AP in the contention phase for AMP and ELBP

Table I. Notations used in the analysis

Consider a system consisting of N stations, including the AP, forming a single subnet. There are R receivers belonging to the multicast group, i.e., $R<N$, and the multicast flow is transmitted from the AP to each multicast receiver. In our model, all stations are located in the transmission range of the AP. The error probability of each link from the AP to a receiver is assumed independent and identical, and denoted as p_e. Thus, a station experiencing an error state does not imply the links of any other stations are also in error.

We assume that all the stations operate synchronously. The propagation time is assumed negligible in the analysis. To simplify the analysis, we further assume that all hosts transmit data based on the RTS/CTS exchange, and RTS, CTS, SEQ, ACK and NAK frames will not be lost or received in errors. Thus, the performance results may be thought of as the upper bounds of the real-world performance.

All stations are assumed backlogged. In other words, the outgoing queue of each station always has at least one frame to send, and each station always tries to access the channel to send out the frame. In this analysis, we do not consider the hidden terminal problem and the capture effect of the CSMA/CA protocol. The notations used in the analysis are summarized in Table I.

B. Number of Transmissions

Let M denote the number of transmissions for an m-data (i.e., multicast data frame) to be received correctly by all R receivers. In [4], the *pdf* and the expectation of transmissions for R receivers to receive a frame correctly are expressed as follows.

$$P(M = m) = \sum_{i=0}^{R} \binom{R}{i} (-1)^i p_e^{i(m-1)} (p_e^i - 1) \quad , m = 1, 2, \cdots$$

$$E[M] = \sum_{m=1}^{\infty} m \cdot P(M = m) = \sum_{i=1}^{R} \binom{R}{i} (-1)^{i+1} \frac{1}{(1 - p_e^i)}$$

C. DCF's Backoff Model

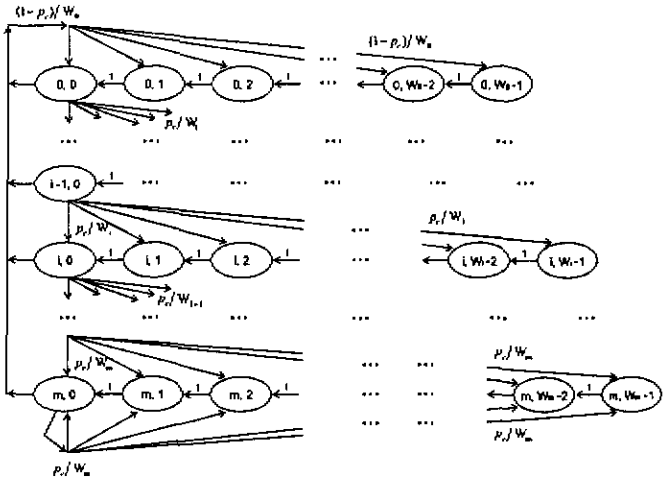


Figure 3. The two-state Markov chain for IEEE 802.11

In [5], the exponential backoff mechanism of an IEEE 802.11 station can be modeled as a two-state Markov chain shown in Fig. 3. The state of the Markov chain is defined as {s(t), b(t)}, where s(t) is the backoff stage of the station at time t, and b(t) is the backoff count for a given station at time t.

Let P_c denote the probability of a collision for a frame being transmitted on the channel. Thus,

$$P_c = 1 - (1 - \tau)^{N-1} \qquad (1)$$

$$\tau = \frac{2(1 - 2P_c)}{(1 - 2P_c)(W_0 + 1) + P_c W_0 (1 - (2P_c)^m)}. \qquad (2)$$

Inverting (1), we obtain $\tau = 1 - (1 - P_c)^{\frac{1}{N-1}}$ (3)

Together with (2) and (3), P_c and τ can be solved by using numerical techniques.

D. Analysis of Frame Holding Time

1) AMP

Let X_{AMP} denote the frame holding time of AMP. Thus,

$$E[X_{AMP}] = E[M] \cdot T_D + (R + E[M] - 1) \cdot T_{R+A} + (E[M] - 1) \cdot [DIFS + (\frac{1-q}{q}) \cdot T_{av\_slot}] \quad (4)$$

where $q = \tau(1-\tau)^{N-1}$, and

$$T_{av\_slot} = (1-\tau)^N \cdot T_{idle\_slot} + (N-1)\tau(1-\tau)^{N-1} \cdot T_S + [1-(1-\tau)^N - N\tau(1-\tau)^{N-1}] \cdot T_C.$$

The first term in (4) corresponds to the total time from when the AP starts sending a multicast frame until it is successfully received by all receivers. This is equal to the mean number of multicast frame transmissions, each with $T_D = RTS + CTS + SEQ + DATA + 3\,SIFS$. The second term is the total time for the AP inquiring the receivers until an ACK is received from all receivers. This is equal to the mean number of RAK/ACK exchanges for a data frame being received successfully by all receivers. Since the AP should at least inquire R receivers, and poll the failed receiver again after each retransmission, it comes to a total of $R + E[M] - 1$ RAK/ACK exchanges, each with

$$T_{R+A} = RAK + ACK + 2\,SIFS.$$

The third term in (4) corresponds to the total time of the AP contending for channel access. Each m-data experiences E[M]-1 retransmissions. For each retransmission, the AP contends the channel after waiting a DIFS. Since the successful transmission probability of a frame sent by the AP is $q = \tau(1-\tau)^{N-1}$, the mean number of time slots between the successful transmissions of two frames, which is a geometric distribution with parameter q, is (1-q)/q, and the mean time of each slot in this period is T_{av_slot}. Thus, the mean contention time for each retransmission is $((1-q)/q) \cdot T_{av\_slot}$, yielding the total time as

$$(E[M]-1) \cdot [DIFS + (\frac{1-q}{q}) \cdot T_{av\_slot}].$$

T_{av_slot} is calculated as follows. The probability that the AP sees an idle slot during the contention phase is $(1-\tau)^N$, which takes time T_{idle_slot}. The probability that the AP observes a success in transmission caused by any other host is $(N-1)\tau(1-\tau)^{N-1}$, which takes time T_S. The AP detects a collision with a probability of $1-(1-\tau)^N - N\tau(1-\tau)^{N-1}$, and a collision slot takes time T_C. Thus,

$$T_{av\_slot} = (1-\tau)^N \cdot T_{idle\_slot} + (N-1)\tau(1-\tau)^{N-1} \cdot T_S + [1-(1-\tau)^N - N\tau(1-\tau)^{N-1}] \cdot T_C.$$

2) ELBP

Let X_{ELBP} denote the frame holding time of ELBP. Thus,

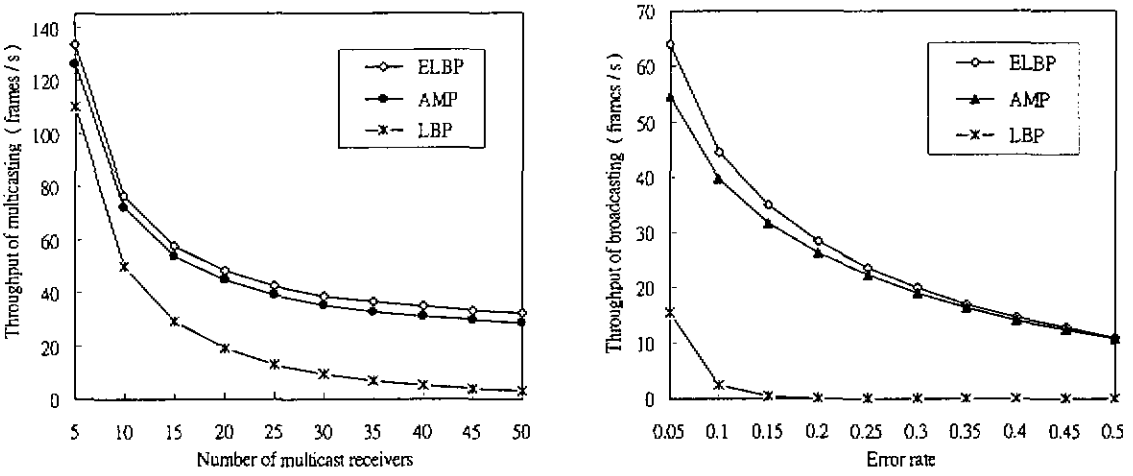
$$E[X_{ELBP}] = E[M] \cdot (T_D + T_{ACK}) + (E[M]-1) \cdot [DIFS + (\frac{1-q}{q}) \cdot T_{av\_slot}] \quad (5)$$

The first term in (5) corresponds to the total time from when a multicast frame is sent until the frame is correctly received by all receivers. Again, the mean number of times each frame is transmitted is $E[M]$, and each transmission takes $T_D + T_{ACK}$. Thus, it spends $E[M] \cdot (T_D + T_{ACK})$ in total for an m-data. The second term is the time for the AP contending the channel for successfully sending an m-data.

IV. PERFORMANCE EVALUATION

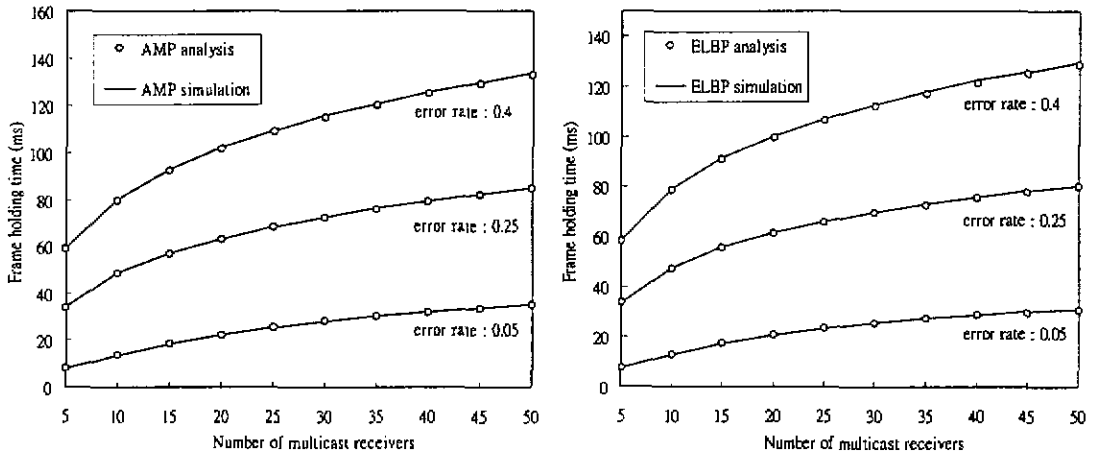
In this section, we provide simulation results to validate our analytical models. The physical parameters used in the simulation follows the IEEE 802.11a standard: data rate: 54Mbps; an idle slot: 9μs; an SIFS: 16μs; a DIFS: 34μs; an RTS: 20 octets; a CTS: 14 octets; data frames: 2304 octets; an ACK: 14 octets; MAC overhead: 28 octets; *CWmin*: 16; *CWmax*: 1024.

Fig. 4 shows the performance of LBP, AMP, and ELBP. We see that both AMP and ELBP outperform LBP in terms of the AP's throughput, defined as the inverse of the mean frame holding time at the AP. Note that LBP's throughput is very poor at high error rates. The performance advantage of our proposed protocols is thanks to the proposed channel acquisition mechanism. Figs. 5 (a) and 5 (b) compare the analytical models of AMP and ELBP with the simulation results. The three pairs of curves in each figure correspond to three different sets of channel error probabilities. The lower the error rates, the lower the frame holding time, matching our expectation. In all cases, all pairs of curves in both figures show that both simulation and analytical results match very well.



(a) Number of receivers, with error rate = 0.05 (b) Error rates, with no. of receivers = 30

Figure 4. The throughputs of the three protocols



(a) AMP (b) ELBP

Figure 5. Analytical and simulation results

V. CONCLUSION

In this paper, we have analyzed the two types of reliable multicast protocols at the MAC layer for IEEE 802.11 WLANs: AMP and ELBP. Both protocols are based on the proposed RTS/CTS/SEQ exchange for IEEE 802.11 WLANs, and are derived from existing works. From the simulation, the performance of AMP and ELBP are both superior to that of LBP, thanks to the channel acquisition mechanism RTS/CTS/SEQ. We have conducted simulations to validate the proposed analytical model. The results show that the simulation and analytical curves match very well, showing that our analytical model is excellent in describing reliable multicast mechanisms for IEEE 802.11 WLANs.

REFERENCES

[1] IEEE 802.11 Working Group, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, *ANSI/IEEE Std 802.11*, 1999.

[2] J. Kuri and S. K. Kasera, "Reliable Multicast in Multi-Access Wireless LANs," *ACM/Kluwer Wireless Networks Journal*, vol.7, no.4, pp. 359-369, August 2001.

[3] M. T. Sum, L. Huang, A. Arora, and T. H. Lai, "Reliable MAC Layer Multicast in IEEE 802.11 Wireless Networks," *Proc. of ICPP*, pp. 527-536, Aug. 2002.

[4] D. F. Towsley, J. Kurose and S. Pingali, "A Comparison of Sender-Initiated and Receiver-Initiated Reliable Multicast Protocols," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No.3, April 1997.

[5] G. Bianchi, "Performance Analysis of the IEEE802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, Vol. 18, No. 3, March 2000.

SARS : A Linear Source Model Based Adaptive Rate-control Scheme for TCP-friendly Real-time MPEG-4 Video Streaming

Chung-Yuan Knight Chang*, Eric Hsiao-Kuang Wu†
Department of Computer Science and Information Engineering
National Central University, Chung-Li, Taiwan, R.O.C.
*tnnd@wmlab.csie.ncu.edu.tw †hsiao@csie.ncu.edu.tw

***Abstract*— In this paper we analyze the problem of transmitting MPEG-4 video over IP when senders use a TCP-friendly rate control protocol. This paper is focused on solving the challenging issues in application layer, compression layer. With this aim, we provide a solution called "*SARS*" (Smoothed and Adaptive Rate-control Scheme) in compression layer and application layer for an MPEG-4 video transmission system and discuss the characteristics of each its elements. The *SARS* contains three main components, including *TFRS module*, *LD module*, and *RSF module*. The *TFRS module* provides a smoothed rate and helps the video encoder to generate smoothed video quality. The LD module helps the video encoder to meet its target closer and lowers the transmitting delay. The *RSF module* provides an adaptive rate control scheme, reduces the number of skipped frames, and makes a smoother presentation. The overall performance of *SARS* will be evaluated in the CBR scenarios or VBR scenarios. We also evaluate the performance of *SARS* over several rate based congestion control protocols such as RAP, TFRC and TMRC.**

***Index Terms*— real-time MPEG-4, adaptive rate control, linear source model**

I. Introduction

With the ever-growing amount of the transmission bandwidth on the Internet, the range of applications being deployed today is getting wider than it was a few years ago. In addition to traditional applications such as E-mail, FTP, and the World Wide Web, the Internet enables people to use it as the transmission medium for various kinds of applications. Applications that rely on the delivery of real-time video, such as Internet television, unicast video-conferencing, IP telephony, distance learning, and video-on-demand, are gaining a lot of attentions [4] and are able to change the way people watch video and to compete with traditional distribution methods of video content, such as broadcasting video over airwaves or cable networks.

As streaming video over the Internet, there exists a major problem that Internet only offers best effort services. There are no guarantees on bandwidth, packet-loss ratio, and end-to-end delay while the delivery of real-time video has these requirements. Specifically, these characteristics are unknown and dynamic. Therefore, designing the protocols and mechanisms in a real-time video streaming system has three key challenging issues.

The first issue is the bandwidth problem. In order to provide an acceptable perceptual quality, a minimum bandwidth is required for real-time video streaming applications. Actually, available bandwidth in Internet today is dynamic. If excessive flows transmit data faster than available bandwidth, congestion collapse occurs. The throughput of other flows will be reduced and the transmission of the ream-time video streams is affected due to congestion collapse [5]. The second issue is the packet-loss problem. In worst case, receiver will unable to have a completely reconstructed video frame due to the lost of video packets. High packet-loss ratio is also caused by congestion collapse. The third issue is end-to-end delay problem. Real-time video streaming is delay constrained (e.g. 150ms for interactive applications and 500ms for video on demand) and the playing of video at receiver is continuous. If a video frame is not received and played out in time, the video frame is considered useless and the effects can propagate. Excessive delay is also introduced due to network congestion.

Thus, in order to solve the problems addressed above, there are two typical approaches can be adopted [6]. The first approach is to enhance the the current Internet architecture with support for QoS flows to guarantee bandwidth, end-to-end latency, and packet-loss. IETF (Internet Engineering Task Force) has defined two models: IntServ (Integrated Services) [7] and DiffServ (Differentiated Services) [8]. However, there are challenges and drawbacks, such as the needs of provisioning for DiffServ, billing and monitoring, and loss of granularity. One of the biggest drawbacks of both the IntServ and DiffServ models comes from the fact that signaling/provisioning happens separate from the routing process. Thus, there may exist a path in the network that has the required resources, even when DiffServ fails to find the resources. Furthermore, the deployment of QoS-support nodes for IntServ/DiffServ in the current network is still difficult and premature due to the huge cost.

In contrast to the first approach, the second one is to introduce control techniques at the end systems without any support from the network. The installation of QoS-support devices in the Internet which changes the Internet architecture to guarantee QoS. The end systems with control techniques can then achieve the requirements which described above and adapt to the changing network condition. It is the feasible and required approach at present and for the future. Therefore, an general end-to-end real-time video streaming system which employs control techniques can then be illustrated in Fig. 1 as

a layer-structure system [30]

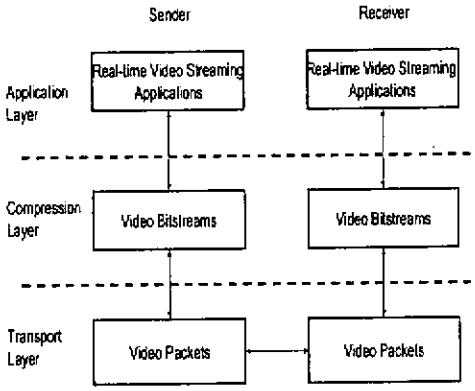


Fig. 1. Layer structure of an end-to-end real-time video streaming system.

The system consists of application layer on the top, compression layer and transport layer. They are described as follows:

Application Layer: It defines the maximum end-to-end latency and the acceptable visual quality in this layer. The real-time video streaming applications are subject to different end-to-end latency constraints.

Compression Layer: It is the video encoder that provides a wide range of coding bit rate and encodes the live video based on adaptive rate control algorithm.

Transport Layer: It is required to estimate the bandwidth and to control the streaming flows' throughput in order not to exceed the available bandwidth or congestion will occur.

The challenging issues of bandwidth, end-to-end latency, and packet-loss consist in *Compression Layer* and *Transport Layer*. We discuss the approaches for *Transport Layer* and *Compression Layer* to solve the challenging issues.

a) Transport Layer Approach: A transport protocol which employs a control technique provides a good solution to solve the challenging issues in this layer. The transport protocols, Transmission Control Protocol (TCP) and User Data Protocol (UDP), are the most dominant transport protocols in the packet-switched network,. TCP is designed for the reliable transmission, either on the congested or light-loaded network. On the TCP connection, slow start and congestion avoidance mechanisms will be applied to avoid congestion. However, the TCP transport protocol is not suitable for these real-time video streaming applications because of its complex retransmission mechanism, which introduces delays, and the excessive burstiness of the traffic transmitted. For this reason, UDP, which makes no attempts to retransmit lost packets is more suitable for video streaming applications as the transmission protocol. The design methodology for UDP is arriving the destination as quickly as possible. Since the real-time video streaming applications are usually built on the top of UDP, UDP does not embed any congestion control mechanism and results in negative impacts on the Internet. These excessive streaming flows using UDP as the transport protocol affect other competing flows and reduce the network utilization. Therefore, an adaptive, end-to-end congestion control protocol for real-time video streaming applications is expected to be introduced in the end-systems: 1) to react to network congestion; 2) to adjust sending rate to avoid congestion collapse; and 3) to keep high network utilization and low packet-loss ratio. There are two types of such control schemes, window-based and rate-based, are adopted to control the traffic flows. There exists a long propagation delay in window based method. The rate based method is more attractive for real-time multimedia traffic then the window based method.

In reaction to packet-loss due to network congestion, there are generally two methods to control errors caused by packet-loss in Internet video transmission, namely packet retransmission [9] and Forward Error Correction (FEC) [10], [11]. The idea of the FEC is to generate redundant packets at the sender, which can be used at the receiver to recover lost video data packets as long as the number of lost packets does not exceed the error correction capability of the FEC code. The most popular FEC code is the Reed-Solomon codes. As for retransmission-based error control method, ARQ (Automatic Repeat reQuest), the sender attempts to retransmit the lost packets upon receiving the packet retransmission requests. It is only feasible to recover burst loss with minimum cost of network bandwidth if round trip delay is low [12].

b) Compression Layer Approaches: State-of-the-art video coding standards, such as H.263, H.263+, H.264, and MPEG-4, are designed for transmission of video on the Internet. H.263 is a provisional ITU-T standard coder for video-conferencing. It was designed for low bitrate communication, early draft specified coding bit rate less than 64Kbits/s, however the limitation has now been removed. H.263 was later extended to H.263+ (version 2) and H.263++ (version 3).

MPEG-4 video coding standard, unlike previous standards from the Motion Pictures Experts Group Consortium (MPEG-1 and MPEG-2), is gaining increasing acceptance and offers the digital video community an open standard in the face of the dominant, yet proprietary, digital video formats from RealNetworks and Microsoft [13]–[16]. It provides a wide range of coding bit rate and resolution to perform adaptive encoding.

In addition, there are generally two method to control errors due to packet-loss, namely encoder error resilient tools and decoder error concealment techniques. MPEG-4 video coding standard provides an error resilient tool to add periodic resynchronization points into the encoded bitstreams. Such a tool limits the impact of the incorrect data within a local area because the decoder the option of advancing to the next resynchronization point to continue decoding when it encounters an error. The decoder error concealment technique is to employ some algorithms to reconstruct the lost information from the already received data. If the reconstruction errors caused by using error concealment techniques are small, they are tolerable to human eyes.

However, an video encoder with error control techniques is still not enough for solving the challenging issues in this layer.

Therefore, an video encoder which embeds the adaptive rate control schemes or supports scalable coding mode is needed to combat these these problems. The adaptive rate control schemes allocate the target bit rate to each frame for video encoding, control the bit rate generated by video encoder, and adapts to the time-varying available bandwidth. If the coding bit rate of a frame is controlled well, the frame can be transmitted, received, played out at receiver on time if no packet-loss occurs.

The remaining of this paper is organized as follows. In Section II, we briefly review recent researches on the end-to-end congestion control mechanisms and rate control schemes in the video coder. The proposed scheme *SARS* is detailed in Section III. The performance comparison of *SARS* and the rate control scheme in [1] under CBR and VBR and over several rate based congestion control protocols are evaluated in Section IV. At the end of this paper, we summarize our study and point out the future works in Section V.

II. Related Work

A. MPEG-4 Video Rate Control

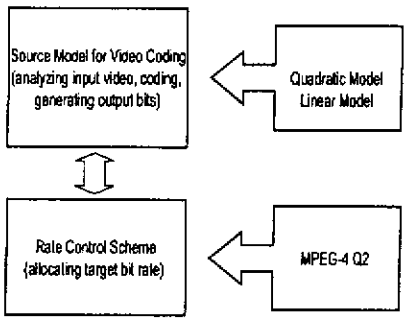


Fig. 2. The main components of a common rate control scheme in a video encoder.

A conventional rate control scheme in a video encoder is a technique used to determine the target bit rate R of each encoded frame and maintain a good visual quality and minimize the distortion measure D under a limited bandwidth constraint [23]. The distortion D means that the difference between the original source picture and the reconstructed picture after it has been decoded. In typical transform coding, both the bit rate R and the distortion D are controlled by the quantization parameter. The rate control scheme allocates the bit budget to each group of pictures (GOP), individual picture and/or macroblock in a video sequence and derive the quantization parameter from a wide range of quantization parameter sets. The conventional rate control scheme embedded in the video encoder generally has two components as shown in Fig.2.

- Rate Control : It consists of the rate allocation and the skipping frame mechanisms. In rate allocation, it estimates the number of bits available to encode the frames in the video sequence. Video delivery over the Internet meet the constraints imposed by the changing network conditions. The allocated bit rate has to adapt to the time-varying available bandwidth. In order to keep the below the pre-defined threshold of end-to-end latency, the skipping frame mechanism is initiated.
- Source Model : The source model is used to help the video encoder to produce the output bit rate closer to its target.

We briefly review current researches related to these two components in the following sections.

1) Researches on Source Model for Video Coding: A source model, aka rate-quantization model, for video coding is the most important part of a rate control scheme. The rate-quantization model defines the relationship between the coding bit rate and quantization parameter. A suitable quantization parameter used for encoding a frame or a macroblock to achieve the target bit rate is derived from the rate-quantization model.

Many conventional rate-quantization model adjust the quantization parameter based on the rate-distortion (R-D) theory. Ding et al. [24] proposed a rate-quantization model based on an exponential equation coupled with several control parameters. A feedback re-encoding method is developed with the rate-quantization model. It is claimed that the model is accurate and the computation is modest. However, the computational complexity is actually very high and the re-encoding method is not suitable for real-time video streaming.

In [25], [26], the logarithmic rate-quantization model is proposed. This source model which is derived from classic R-D theory is used to predict the R-D characteristics and the coding bit rate of a frame is adjusted to ensure minimum perceptual quality. A quadratic/second-order rate-quantization model, which was proposed by Chiang and Zhang, was used to predict the frame bit rate from distortion measure and quantization parameter [27] and it was adopted as a rate control tool in MPEG-4 verification model. Lei et al. [28], [29] proposed an efficient and low-delay macroblock layer rate control algorithm based on a Lagrange optimization technique and this rate control scheme was adopted as a rate control tool in test model TMN8 and a rate control tool with some modifications in MPEG4 verification model. Lee et al. [30] proposed a scalable rate control (SRC) scheme for different coding granularity including the frame level, the object level and the MB level. The proposed rate control algorithm was also based on the quadratic rate-distortion model. It was also adopted as part of the rate control tool in MPEG-4 verification model.

However, the rate-distortion theory used by these source models is the branch of information theory and it makes two assumptions. First, it assumes that the source statistics are Laplacian or Gaussian distributed and uncorrelated. Second, it assumes that the coding bit rate is approximated by the empirical entropy. Actually, these two assumptions are not entirely correct [31]. There will be a large mismatch between the target bit rate and the actual coding bit rate, especially at a lower bit rate, using the rate-distortion theory [32].

He et al. [2], [3] proposed a linear source model which defines the linear relationship between the percentage of non-zero among quantized DCT coefficients and the coding bit rate. The coding bit rate is strongly proved to be a function of the percentage of non-zero quantized coefficients. The linear source model is used to predict the coding bit rate and determine the quantization parameter for encoding macroblocks in a video frame. In contrast to the theoretical rate-distortion model, new concepts are introduced in this source model. The number of zeros among DCT coefficients play a key role in transform coding and the coding bit rate is also a function of percentage of zeros among the quantized DCT coefficients. However, the process of encoding macroblocks is macroblock by macroblock and the use of percentage of non-zeros among the quantized coefficients means that the non-zero coefficients are evenly allocated to each macroblock. A macroblock will be assigned a larger quantization parameter if it contains more non-zero DCT coefficients. But the macroblock may need more bits to encode it's texture to provide a better quality.

2) Researches on Rate Control for Video Coding: A rate control algorithm generally includes three steps. A fixed size of buffer based on the maximum acceptable latency is determined in the first step. The second step is to estimate the target bit rate according to the buffer occupancy, remaining bits in this video sequence and available bandwidth. The allocated bit budget can then be applied on the source models to derive the quantization parameter. The final step is to control the buffer occupancy and start the frame skipping mechanism if buffer overflow occurs.

Many researches focus on the MPEG-4 Q2 rate control algorithm which is based on the quadratic rate-distortion model [27], [30], [33]–[35].

Pan et al. [36] proposed an improved MPEG-4 Q2 rate control scheme by modifying the calculation of target bit rate for P-frames. It introduces a weighted bit budget allocation algorithm for P-frames that the P-frame nearer to I-frame is allocated more bits than other P-frames. The perceptual quality of the reconstructed video is improved in this scheme.

However, the MPEG-4 Q2 rate control schemes [27], [30], [36] are heuristic and are designed for CBR applications which are not suitable for real-time streaming video on Internet. In [1], Li adopted the fluid-flow traffic model and the linear tracking theory to calculate the target bit rate and the quantization parameter for P-frames corresponding to time-varying available bandwidth. Li's rate control scheme divides the allocation of target frame bit rate into two steps. The first target bit rate is determined based on the target buffer level and current buffer level. The final target bit rate is determined based on the first target bit rate and remaining bits of a GOP. There is a problem with such an adjustment. If target buffer level is low and current buffer level is high, the first target bit rate is possible to be negative. But the final adjustment with remaining bits will make the target bit rate positive. The buffer level thus remains increasing and is unable to meet the target buffer level. If the rate decreases suddenly, the buffer overflow will occur. Furthermore, there is also a problem with the skipping frame control. It is possible to cause frames to be skipped continuously even some frames had already been skipped.

B. Congestion Control

Much attention has focused on developing congestion control algorithms for streaming media applications in recent years. A congestion control protocol have to provide the following characteristics:

- **TCP-friendly** Non-TCP flows long-term throughput does not exceed the throughput of a conformant TCP connection under the same conditions [5]. Thus, a congestion control protocol should provide a sending rate that is close to competing TCP flows to achieve TCP-friendly.
- **Smoothness** A congestion control protocol should provide a smooth sending rate that the variations are small over time even with accidentally packet loss.
- **Aggressiveness** If the available bandwidth is getting much more, a congestion control protocol should increase sending rate as fast as possible to improve the network utilization.
- **Responsiveness** With the change of network condition that the occurrence of congestion becomes more and more, a congestion control protocol should decrease the sending rate as fast as possible.

There are tradeoffs among smoothness, aggressiveness and responsiveness. A smooth sending rate, which is less sensitive to the change of network condition, is less aggressive and responsive.

Existing congestion control protocols can be roughly divided into two categories: *window-based* and *rate-based*. Window-based congestion control protocols use a AIMD-like method to increase or decrease the congestion window [37], [38]. But the window based congestion control is not suitable for streaming applications. The rate-based congestion control, including RAP [39], TEAR [40], LIMD/HD [36], TFRC [41], and TMRC [42], provide a more accurate formula for the sending rate to make flows perform TCP-friendly. We briefly describe AIMD-based and equation-based rate based congestion control protocols in the following sections.

1) AIMD Based: [39] presents the Rate Adaptation Protocol (RAP), which employs AIMD algorithm. Its primary goal is to be fair and TCP-friendly while separating AIMD network congestion control from application-level reliability. RAP, in the absence of packet loss, increase the transmission rate in a step-like fashion. The transmission rate is controlled by adjusting the inter-packet-gap(IPG). The transmission rate is computed by dividing the packet size by inter-packet-gap. Basic RAP behaves in a TCP-friendly in a specific range of likely conditions, but the author also devised a fine-grain rate adaptation mechanism to extend this range further.

LIMD/H (Increase Multiplicative Decrease With History) [43], similar to RAP [39], uses AIMD for rate control instead of window control and it is based upon the RTP/RTCP. Additionally, it uses the history of losses across measurement periods to adapt its backoff strategy. LIMD/HD [36] is an augmentation of the LIMD/H. It defines five cases to adjust the sending rate: 1) packet-loss ratio is reduced abruptly; 2) packet-loss ratio is reduced gradually; 3) packet-loss ratio is increased gradually; 4) packet-loss ratio is increased abruptly;

and 5) no RTCP packet is received within 5s. The rate increasing/decreasing are different in each case, especially the rate is halved in case 5. However their increase rule is still additive. These two protocols result in lacking aggressiveness and convergence-to-fairness [44].

2) Equation Based: TFRC [41] is proposed to be a TCP-friendly, rate-based and end-to-end congestion control protocol for unicast traffic. The TFRC sender explicitly adjusts its sending rate as a function the measured rate of loss events by receiver and ensures fair behavior against competing flows. Instead of reacting to single congestion event (e.g. packet loss), TFRC changes its sending rate in response to the loss rate, sampled over a single round-trip time. It is desirable to feature of delivering maximally smooth, TCP-friendly transmission rate. However, it is characterized by very slow responsiveness to changing network conditions. It also result in lacking the aggressiveness.

TMRC [42] is proposed based on a new approach using time-based model for TCP-friendly rate estimation. The average TCP sending rate in the time-based model, instead of collecting packet loss rate, is measured as a function of the packet size, round-trip time and the congestion event period. As the results shown in [42], TMRC exhibits good aggressiveness and responsiveness and provides smoother sending rate.

Motivated by the above discussion, the problems dealt with in this paper arise from the fact that although the many rate based congestion control algorithms, such as RAP [39], TFRC [41] and TMRC [42], were proposed to support real-time video streaming applications, they focused on achieving the friendliness, aggressiveness, and responsiveness to be a suitable congestion control on Internet. Li et al [1] proposed a real-time video coding system . However, as mentioned above, the source model which this system adopted can not it's target bit rate and it's computational complexity is a little bit high. Additionally, the allocation of target bit rate in Li's scheme will cause the increasing of buffer occupancy and the frame skipping mechanism in Li's rate control scheme is not correct so that the number of skipped frames is still high. A MPEG-4 real-time video streaming applications using these rate based congestion control protocols as the transport layer protocol is still an attractive topic. However, the fast changing of the estimated sending rate makes an uncontrolled real-time video streaming applications unable to meet its delay constraints and also causes the fluctuating video quality. Therefore, this paper starts from considering a real-time MPEG-4 video streaming system concerning the use of an rate based congestion control protocol as the transport layer protocol and the *SARS* (Smoothed and Adaptive Rate-control Scheme) which cooperates with the estimated rate reported from transport protocol. Note that we do not consider the error control techniques in this paper.

The rate based congestion control protocol allows the real-time streaming applications in the sender to know the available bandwidth and helps them to reduce the packet loss ratio while competing with other flows on Internet. The *SARS*, including *TFRS module*, *LD module*, and *RSF module*, is thus introduced in the real-time streaming applications. The *TFRS module* is a TCP-friendly rate smoother with memory wiper which periodically average the rate reported from transport layer. The *LD module* is a linear source model for video coding which helps the video encoder to meet its target closer and lower the delay. Considering the number of non-zero among DCT coefficients are discarded after quantization, there must exists another relationships between the number of non-zero quantized coefficients and the coding bit rate. We need to model these relationships to met the target bit rate more accurately for video encoding and provide better perceptual quality. For the *RSF module*, it can determine the target rate for video encoding based on the rate reported from the rate based congestion control protocols and reduce the number of skipped frames. In order to adapt to dratical changing network conditions, further adjustment is also needed in the rate control scheme. We target to provide the solutions for solving the issues of delivering real-time MPEG-4 video over the Internet when end systems adopt the rate based congestion control protocols and to evaluate the performance of our solutions.

III. SARS : Smoothed and Adaptive Rate-control Scheme

A. Overview of The System Architecture

In this section we describe the whole system architecture proposed in this paper to transmit ream-time MPEG-4 video over rate based congestion control protocols. There are two objectives in this system. The first objective is that the system has to regulate the output bit rate that adapts to the time-varying available bandwidth. The second objective is coding bit rate allocation and coding bit rate adjustment for the MPEG-4 video encoder. The bit budget will be allocated before encoding and controlled during the encoding process.

The whole system architecture is depicted in Fig. 3, where four main components of the system are presented: the *TCP-friendly Rate Control*, the *TFRS Module*, the *LD Module*, and the *RSF Module*, whose functions are described in the following sections.

B. Bandwidth Adaptation: TCP-friendly Rate Control

Internet conditions change with time; as a result, real-time application should tailor their transmission rate in a manner that achieves high utilization while sharing bandwidth appropriately with competing traffic (e.g., Email, web traffic, etc.). Additionally, the quality of the received video should react to the available bandwidth, so that users receive the highest possible quality video for their available bandwidth.

In order to adapt to time varying channel bandwidth appropriately, real-time streaming applications must:

- Adapt to transient changes in available bandwidth by applying congestion control scheme in the transport layer, and
- Adjust the coding bit rate adapt to the estimated available rate

A rate based congestion control protocol provides an estimated tcp-friendly sending rate calculated either from receiver or at sender. Every time an estimated rate report is received from receiver or calculated at sender, the sender reacts with the report of rate to adjust its sending rate with the new

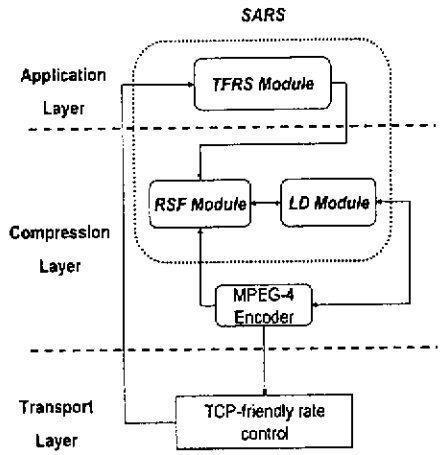


Fig. 3. The system architecture of real-time applications.

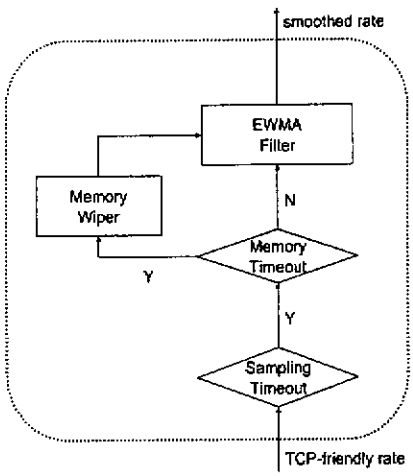


Fig. 4. TFRS Module.

one. We denote the $T_{rate}(n)$ as the n^{th} estimated tcp-friendly sending rate. The event of changing rate and the information of $T_{rate}(n)$ are then passed to the *Rate Smoother* component. The transmission of packets will be scheduled using an transmitting packet interval.

C. TFRS Module : A TCP-friendly Rate Smoother with Memory Wiper

In multimedia streaming applications, the video encoder in the compression layer plays an important role in available bandwidth adaptation. However, to provide graceful quality degradation, the rate for video encoding may not be varied so frequently and a rate smoothing techniques has to be utilized. The *TFRS Module* is to smooth the rate for video encoding and It has three main functions. We depict it in Fig. 4.

The first function is to sample the estimated tcp-friendly sending rate. A sampling timer which is represented by frames is placed in the *TFRS Module*. If an event of timeout occurs, the *TFRS Module* will obtain the newly estimated tcp-friendly sending rate T_{rate} to replace the old one. We define $S_{interval}$ as the sampling time which is represented by number of frames.

The second function is to wipe the memory of the rate. A memory timer which is represented by frames is placed in the *TFRS Module*. The memory wiper will wipe the history in the EWMA filter if a timeout occurs. The new rate will be passed through the EWMA filter and applied directly to give the proposed rate control scheme *SARS* more responsiveness to the TCP-friendly rate control protocols.

The third function is to apply the T_{rate} on the EWMA filter to get a smoother rate after getting the newer tcp-friendly sending rate T_{rate}. The Exponentially Weighted Moving Average (EWMA) is a statistic for monitoring the process that averages the data in a way that gives less and less weight to data as they are further removed in time. The function of EWMA filter is to adapt the bandwidth conservatively. This EWMA filter is expressed as follows:

$$T_{rate_{smoothed}}(n) = (1-\gamma) \times T_{rate_{smoothed}}(n-1) + \gamma \times T_{rate}(n) \quad (1)$$

where N_c is current number of coded frames, γ is the weighting parameter whose default value is 0.5, *n* and *n-1* are the time that n^{th} and $n-1^{th}$ rate-change events occurred, respectively.

The main reason behind this process is to help the rate-adaptive video encoding to provide a consistent perceptual quality. Thus, the weighting parameter of EWMA filter, the sampling timer, and the memory timer play important roles on the tradeoff between the bandwidth and quality adaptation. If we shorten memory timer, sampling timer, and increase weighting parameter, it help video encoder follow the behavior required by tcp-friendly rate control but sacrificing the consistency of video quality. On the contrary, keeping memory timer long, sampling timer long, and weighting parameter low will enable the video encoder to provide more consistent quality, but the multimedia system becomes unstable.

D. LD Module : A Linear Source Model for MPEG-4 Video Coding

1) Statistical Properties of Video Source: In digital video compression, all significant coders employ the DCT and quantization. The current video coding standards are based on motion estimation, motion compensation and DCT transformation, quantization and coding. A MB(macroblck) in a frame is further divided into six smaller blocks and DCT is then applied to each of the six blocks. After transforming, the DCT coefficients are quantized. The process of quantization is to do a division by the quantization parameter followed by a rounding operation. Finally, the MB header, motion vector, and DCT quantized coefficients are variable-length encoded and multiplexed into the compressed bit stream. The compressed bit stream rate are regulated by determining the quantization parameter of the DCT blocks and are composed from three main parts:

- Header Bits

- Texture Bits
- Motion Vector Bits

The nontexture bits are generally considered as invariant number that could be predicted and excluded from the target bit rate estimation. In order to estimate the texture bits, we have to find out the relationship between the bit rate and statistical properties of video source. Only the quantization parameter is not enough to represent the statistical properties and model the characteristic rate curve.

As mentioned in Section II, all the source models are to find the best expression for the coding bit rate R in terms of the quantization parameter Q_p. In this section, we try to derive another relationship between the coding bit rate of a picture and the ratio of the non-zero DCT coefficients and the non-zero quantized coefficients. The non-zero DCT coefficients NZ_{DCT} are the amount of non-zero coefficients in each MB after applying discrete cosine transformation method. The quantized non-zero DCT coefficients NZ_{DCT_Q} are the amount of non-zero coefficients in each MB after applying a specified quantization parameter. Twelve video sequences, "Akiyo", "Coastguard", "Container", "Dancer", "Flower", "Foreman", "Mobile", "Mother", "News", "Paris", "Silent", "Singer", "Stefan" and "Table tennis", are chosen for this study. We use the MPEG-4 ISO Reference coder [45] and run it on the video sequence "Akiyo" using particular values of quantization parameter ranging from 1 and 31. The video coder outputs two types of frames. One is I-frame, and the other is P-frame. I-frame is self-coded without motion compensation and it always produces more bit count than P-frame does. In order to look at the relation of bit count and the ratio of non-zero DCT coefficients and non-zero quantized coefficients, the curves shown in Fig. 5 are plotted using the bit count and the non-zero ratio for both the horizontal and vertical axes.

The ratio of non-zero DCT and quantized coefficients NZ_{ratio} is defined as follow:

$$NZ_{ratio} = NZ_{DCT_Q}/NZ_{DCT} \tag{2}$$

I-frame interval size between the selected frames is 30, which is determined to represent a general GOP length of 30. A noticeable point in Fig. 5(a) is that the rate-NZ_{ratio} characteristics exhibited a linear relationship over a wide range of Q_p. To examine the above characteristics for the same sequence, a similar simulation is performed on several P-frames sampled from the "Akiyo" sequence and other video sequence, including "Flower", "Mobile", and "Singer". They all show the same linear relationship.

2) Construction of Linear Source Model: As mentioned above, we observe that there is a linear relationship between the coding bit rate R and non-zero ratio of transformed and quantized DCT coefficients NZ_{ratio}. We can see that the coding bit rate R could be a linear function of NZ_{ratio}. Therefore, based on these observations, the linear relationship between R and NZ_{ratio} can be characterized by the following expression:

$$R = \kappa \times NZ_{ratio} \tag{3}$$

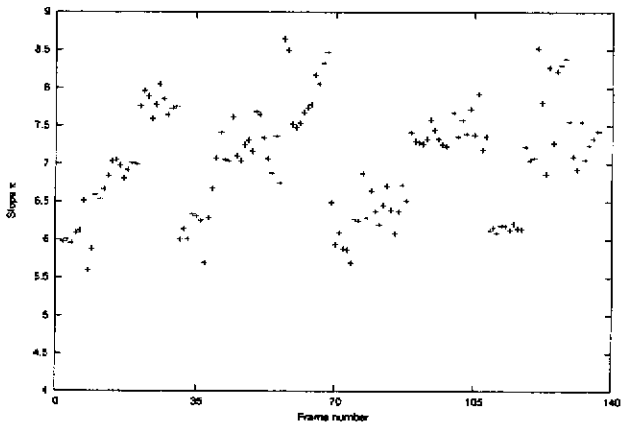


Fig. 9. The variation of model parameter κ.

where κ is the model parameter and it is the slope of the approximation line and R is the coding bit rate represented by bits per pixel.

The linear source model is an augmentation of the source model proposed in [2], [3] and there is still only one model parameter. The number of non-zero quantized coefficients can be easily derived from (3). But the slope κ is also a key role in the linear source model, we provide an adaptive estimation method for slope κ.

3) Adaptive Estimation of The Model Parameter κ: Model parameter κ determines the slope of the R-NZ_{ratio} curve and it is the only parameter of the proposed model. The values of slope κ for the sampled frames from those test video sequences used by this paper is plotted in Fig. 9 for sampled frames. It can be seen that the variation of each slope κ is large.

Therefore, the model parameter κ have to be adaptively estimated to achieve the target bit rate accurately. The estimation method is similar to the method in [2], [3].

Let NZ_{DCT_m} and NZ_{Q_m} be the number of non-zero DCT coefficients and non-zero quantized coefficients in m^{th} macroblock, respectively. Let R_m be the actual coding bit rate in m^{th} macroblock. Let M be the number of encoded macroblocks. Note that in a macroblock, there are total 384 coefficients. The model parameter κ can be estimated using the following expression:

$$\kappa = \frac{\sum_{m=1}^{M} R_m}{384 \times M} \times \frac{\sum_{m=1}^{M} NZ_{DCT_m}}{\sum_{m=1}^{M} NZ_{Q_m}} \tag{4}$$

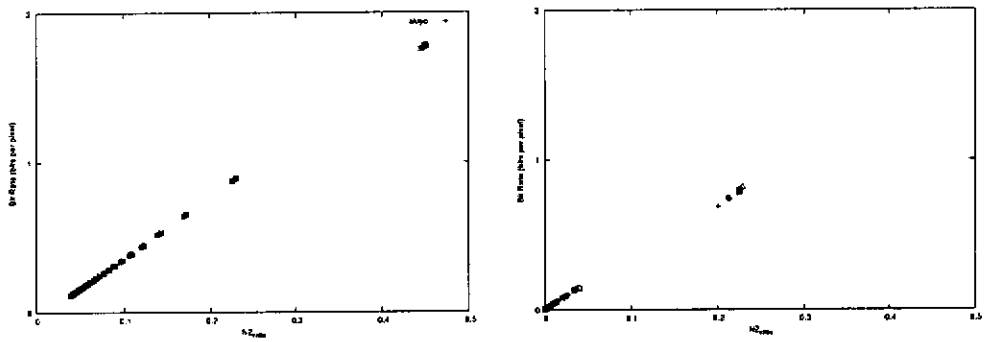
The initial value of κ is set to 7.6438.

Based on (3) and (4), the linear source model can be adopted in the rate control scheme to provide lower delay. We will discuss the linear source model based rate control scheme in the next section.

E. RSF Module : A Reduced Skipped Frames and Adaptive Rate-control Scheme

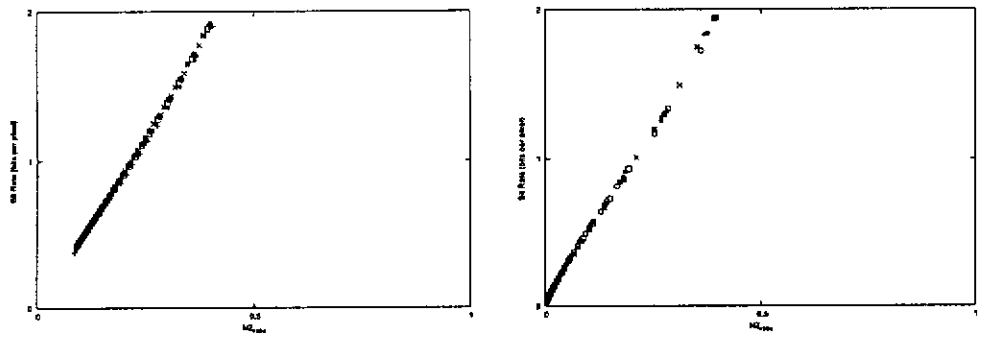
The *RSF Module* is summarized and depicted in Fig. 10.

In the following sections, we describe the proposed adaptive rate control scheme at different levels including GOP layer rate control, frame layer rate control and macroblock layer rate



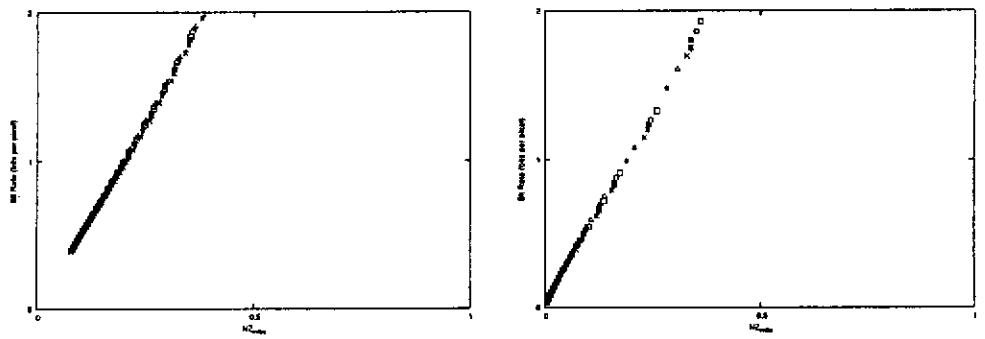
(a) I-frames (b) P-frames

Fig. 5. Plots of bit rate and NZ_{ratio} for the different types of frames in "Akiyo" video sequence.



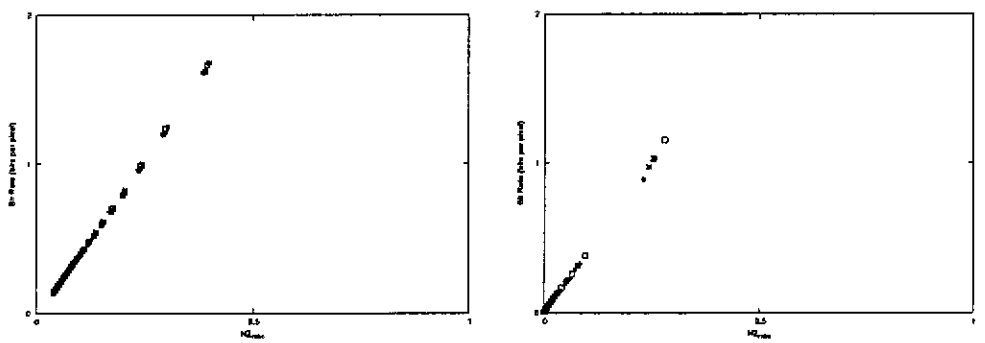
(a) I-frames (b) P-frames

Fig. 6. Plots of bit rate and NZ_{ratio} for the different types of frames in "Flower" video sequence.



(a) I-frames (b) P-frames

Fig. 7. Plots of bit rate and NZ_{ratio} for the different types of frames in "Mobile" video sequence.



(a) I-frames (b) P-frames

Fig. 8. Plots of bit rate and NZ_{ratio} for the different types of frames in "Singer" video sequence.

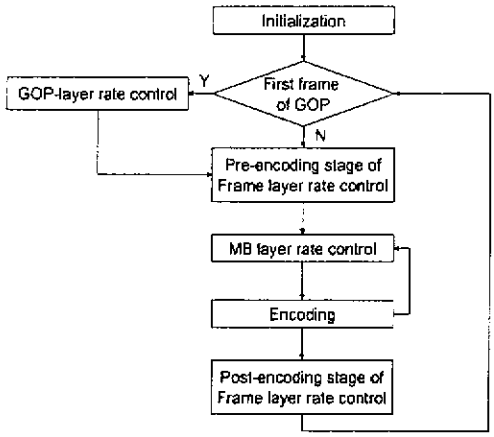


Fig. 10. RSF Module.

control. The GOP layer rate control estimates the bit budget for each GOP. The frame layer rate control estimates the bit budget for each frame. The more accurate rate control, macroblock layer rate control, estimates the bit budget for each macroblock and derive the quantization parameter for encoding. In our rate control scheme, we only consider the GOP structure as IPPP...P.

1) Initialization: As described in Appendix A, an fluid-flow traffic model containing a virtual buffer is introduced. The size of the virtual buffer is decided according to the time-varying bandwidth and the pre-defined maximum end-to-end latency. Let $D_{latency}$ denote the maximum acceptable latency during the transmission. For a given tcp-friendly rate $T_{rate_{smoothed}}$, the size of the virtual buffer model B_s is computed as follows:

$$B_s(0,0) = T_{rate_{smoothed}} \times D_{latency} \quad (5)$$

The size of the virtual buffer model is not fixed during the real-time video streaming, it will be adjusted if the rate $T_{rate_{smoothed}}$ is changed. The use of such a virtual buffer model is to assure that the video data in the buffer could be transmitted out under the acceptable delay constraints.

2) GOP-Layer Rate Control: Before allocating the target bit rate for each frame, we first estimate a rough bit budget for an entire GOP at the beginning of each GOP. The calculation of bit budget for each GOP is just because the structure of a GOP. A GOP always starts with an I-frame and the I-frame is always encoded using a pre-defined quantization parameter. We can not know the actual bit rate until the I-frame is coded. Therefore, we allocate the rough bit budget for each GOP and the remaining bits for remaining frames can be calculated after I-frame is coded. The target bit rate for each remaining frames is then determined by the actual bit rate of I-frame and remaining bits of a GOP.

We thus estimate it from the predicted available bandwidth $T_{rate_{smoothed}}$, frame rate and the total number of frames in a GOP. Let N_{GOP} denote the number of frames in the GOP and there are one I-frame and N_{GOP} - *1* P-frames.

The rough bit budget for i^{th} GOP is calculated as follows:

$$R_{GOP}(i,1) = \frac{T_{rate_{smoothed}}(n)}{F_r} \times N_{GOP} - B_l(i,1) \quad (6)$$

where $R_{GOP}(i,1)$ is the total number of bits for i^{th} GOP, F_r is the frame rate, and $B_l(i,1)$ is buffer level at the beginning of i^{th} GOP.

It can be seen that the rough bit budget for each GOP is determined by the draining rate, the value of N_{GOP} and the current virtual buffer level. The draining rate is the number of bits drained at each frame interval. In our rate control scheme, the virtual buffer level after encoding each GOP should be kept at 0 to assure that the frames are transmitted on time. If each GOP can not meet the allocated bit budget after encoding, there will be video data left in the virtual buffer that is unable to be transmitted at previous GOP interval. These remaining bits of i^{th} GOP in the virtual buffer will affect the target bit allocation of $i + 1^{th}$ GOP. The rough bit budget allocated to $i + 1^{th}$ GOP will be less.

3) Pre-Encoding Stage of Frame-Layer Rate Control: There are two stages in our frame layer rate control scheme. The first stage is the pre-encoding stage and the second stage is the post-encoding stage. We first describe the pre-encoding stage of frame layer rate control in this section. The post-encoding stage will be detailed in Section 4.3.5 due to the macroblock layer rate control.

The most important processes in the pre-encoding stage are to calculate the target bit rate for a frame.

Considering a frame is encoded and filled into the virtual buffer while the transport layer drains the video data from the virtual buffer to packetize video packets for transmitting. We could treat such a process as fluid-flow traffic model as described in Appendix A. Therefore, we can calculate the target bit rate for a frame based on the available bandwidth and the encoding frame rate in the video encoder. The draining rate *d* at each frame interval is also the same as the target bit rate for a frame. The calculation is expressed as follows:

$$R_f^{'}(i,j) = d(i,j) = \frac{T_{rate_{smoothed}}(n)}{F_r} \quad (7)$$

where $R_f^{'}(i,j)$ and *d(i,j)* are the target bit rate and draining rate of j^{th} frame in i^{th} GOP respectively, $T_{rate_{smoothed}}(n)$ is the newest estimated rate, and F_r is frame rate.

However, the I-frame of i^{th} GOP in our rate control scheme is encoded with a pre-defined quantization parameter. It is similar to the existing rate control schemes in [1], [30], [33]. The actual coding bit rate of an I-frame is unknown till it is encoded and it is usually larger than the coding bit rate of P-frames. After an I-frame is encoded and filled into the virtual buffer, it can't be drained out completely at the interval and part of the I-frame is left in the virtual buffer. The delay is introduced if we still use (7) to allocate target bit rate to the P-frames in the same GOP. Therefore, the excess bit rate used by the I-frame must be paid by reducing the target bit rate of other P-frames. A target virtual buffer level is used to cooperate with the allocation of target bit rate for P-frames. We define the target virtual buffer level as a function of the

frame position in a GOP and compute it for each P-frame in a GOP as follows:

$$TB_l(i,j) = TB_l(i,j-1) - \Delta_p \quad (8)$$

$$TB_l(i,1) = B_l(i,1) \quad (9)$$

$$\Delta_p = \frac{TB_l(i,1)}{N_{GOP}-1} \quad (10)$$

It can be shown from (9) that we reset the target virtual buffer level based on current virtual buffer level after coding I-frame. We then gradually decrease the target virtual buffer level by Δ_p using (8) and set the target virtual buffer level for j^{th} frame. It can be shown from (8) that the target buffer level for the last P-frame is 0 to assure all the frames in the GOP is transmitted out in time.

We now have the target virtual buffer level for each P-frame in i^{th} GOP. The target bit rate for j^{th} frame in i^{th} GOP can be calculated based on the target virtual buffer level $TB_l(i,j)$, current virtual buffer level $B_l(i,j)$, encoding frame rate F_r, and the available bandwidth $T_{rate_{smoothed}}(n)$. We divide the allocation of target bit rate for j^{th} frame in i^{th} GOP into two steps :

Step 1: Allocation of target bit rate for j^{th} frame

$$R_f(i,j) = \frac{T_{rate_{smoothed}}(n)}{F_r} + (TB_l(i,j) - B_l(i,j)) \quad (11)$$

where $R_f(i,j)$ is the target bit rate for j^{th} frame in i^{th} GOP, $TB_l(i,j)$ is the target virtual buffer level, and $B_l(i,j)$ is the current virtual buffer level.

It can be shown from (11) that the target bit rate for the j^{th} frame is calculated based on the difference of target virtual buffer level and current virtual buffer level. It is because that the actual coding rate is impossible to be exactly controlled as the same as its target.

4) MB-Layer Rate Control: The flow of macroblock layer rate control is depicted in Fig. 11.

We divide the MB layer rate control into the following steps:

Step 1: Calculating the target bit rate T_{MB}

Before choosing the quantization parameter, we need to allocate target bit rate to each macroblock. The sum of absolute difference (SAD) for motion estimation provides important information about the activities in the macroblocks. We thus calculate the target bit rate for m^{th} MB using the SAD information. The calculation is expressed as follows:

$$R_{MB}(i,j,m) = \frac{SAD_{MB}(i,j,m)}{SAD_{frame}(i,j) - \sum_{v=0}^{m-1} SAD_{MB}(i,j,v)} \times (R_f(i,j) - \sum_{v=0}^{m-1} A_{MB}(i,j,v)) \quad (12)$$

where $R_{MB}(i,j,m)$ is the target bit rate for m^{th} MB, $SAD_{MB}(i,j,m)$ is the sum of absolute difference in m^{th} MB of j^{th} frame in i^{th} GOP, $SAD_{frame}(i,j)$ is the amount of SAD in all macroblocks of i^{th} frame, and $A_{MB}(i,j,v)$ is the actual coding bit rate of v^{th} MB.

It can be shown from (12) that the more SAD a MB has, the more bit rate it is allocated.

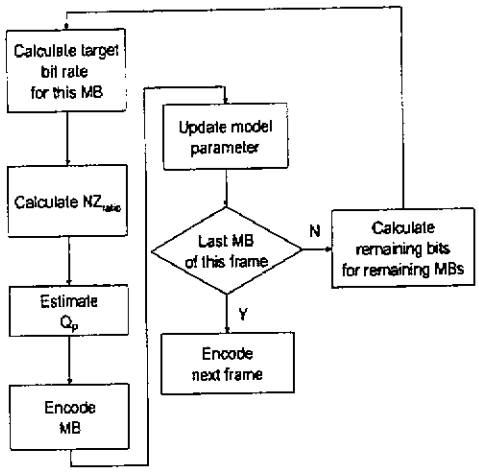


Fig. 11. The flow of MB layer rate control.

Step 2: Calculating NZ_{ratio}

We apply the target bit rate for m^{th} MB on (3) to get the non-zero ratio of quantized coefficients NZ_{ratio}.

Step 3: Estimating quantization parameter

We train huge amount of data and store the mapping of non-zero ratio NZ_{ratio} and quantization parameter in a table. The quantization parameter QP_{MB} is determined by the pre-trained lookup table.

Step 4: Encoding m^{th} MB

The quantization parameter QP_{MB} is applied for encoding m^{th} MB.

Step 5: Update model parameter

The model parameter κ is updated adaptively by using (4).

Step 6: Loop steps 1 to 5 for the next MB until all the MBs in the current frame is encoded.

5) Post-Encoding Stage of Frame Layer Rate Control: If the sampling timeout occurs, the $T_{rate_{smoothed}}$ is re-calculated by using (1). The remaining bit budget of i^{th} GOP is adjusted by the new $T_{rate_{smoothed}}$. The adjustment of remaining bit budget of i^{th} GOP is expressed as follows:

$$R_{GOP}(i,j) = R_{GOP}(i,j-1) + \Delta_{GOP} - A(i,j-1) \quad (13)$$

$$\Delta_{GOP} = \frac{T_{rate_{smoothed}}(n) - T_{rate_{smoothed}}(n-1)}{F_r} \times N_{rGOP}(i) \quad (14)$$

where $N_{rGOP}(i)$ is the remaining frames to be encoded in the i^{th} GOP and $A(i,j-1)$ is the actual coding bit rate of $j-1^{th}$ frame.

The virtual buffer level is increased by adding the actual bits produced by the j^{th} frame in the i^{th} GOP and decreased by draining the bits from the virtual buffer using (??). If the virtual buffer level is too high, we start the frame-skipping control mechanism. The frames will be skipped until the virtual buffer level is under a safe level. The buffer condition is expressed as follows:

$$B_l(i,j) < \rho \times B_s(i,j) \tag{15}$$

where ρ is a constant and its typical value is 0.8.

Also, if buffer overflow occurs, the target buffer level $TB_l(i,j)$ will be recalculated using the following equation:

$$TB_l(i,j) = B_l(i,j - N_{skipped}) - \frac{T_{rate_{smoothed}}}{F_r} \times N_{skipped} \tag{16}$$

where $N_{skipped}$ is the number of skipped frames after the frame-skipping control is started.

It can be seen from (16) that we reset the value of target virtual buffer level as current virtual buffer level after skipping frames. Therefore, Δp will also be recalculated using (10).

We provide a solution for streaming real-time MPEG-4 video over rate based congestion control protocols. We thus demonstrate the performance of the proposed rate control scheme in next section.

IV. Simulation Results

Numerical experiments have been conducted to evaluate the performance of our proposed rate control algorithm in this section. We implement the proposed rate control algorithm in the MPEG-4 video encoder [45]. The purpose of this section is to demonstrate the following.

- Compared to rate control scheme proposed in [1], the *SARS* controls the coding bit rate more accurately and keep our buffer level closer to its target. Also, the perceptual quality is improved using *SARS*.
- Overall performance of *SARS*. The rate control scheme is experimented over three rate based congestion control protocols. The coding bit rate is adapted to the estimated rate. The output bits from MPEG-4 video encoder can meet the throughput of transport protocols. All the frames can be transmitted out under the delay-constraints. The rate smoother makes the quality deviation smaller and improves the overall perceptual quality.

A. Performance Evaluation of SARS under CBR and VBR

We consider two video sequences "Mobile" and "Paris" in this scenario. The size of the video sequences are CIF(352x288). The frame rate is 30 fps. These video sequences are coded by the unit of GOP. The length of a GOP is 60. The pre-defined quantization parameter for I-frame is 15. The experimental results are discussed in the following sections.

1) CBR-768Kbps: The experimental results are plotted from Fig. 12(a) to Fig. 13(d)

It can be seen from Fig. 12(a) that the variation of coding bit rate in Li's scheme [1] is large due to the source model used in it can not help the video encoder to meet its target. Fig. 12(b) shows the large fluctuation of buffer level by using Li's scheme, while the buffer level can meet its target in *SARS*. The initial increase of the target buffer level at 59^{th} frame in Li's scheme is due to a large amount of bits remained in last GOP. Fig. 12(c) shows the PSNR of each frame by using Li's scheme and *SARS*. It can be shown that, for *SARS*, the PSNR of P-frames nearer to I-frames are slightly lower than Li's scheme, while the PSNR of P-frames at late part of a GOP are higher than Li's scheme. Fig. 12(d) shows the average PSNR of each GOP. It can be shown that, for *SARS*, the average PSNR has been improved slightly in this scenario. The results in Fig. 13(a) - Fig. 13(d) also show the improved performance by using *SARS*.

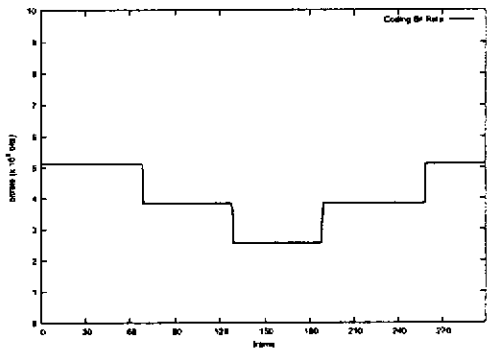


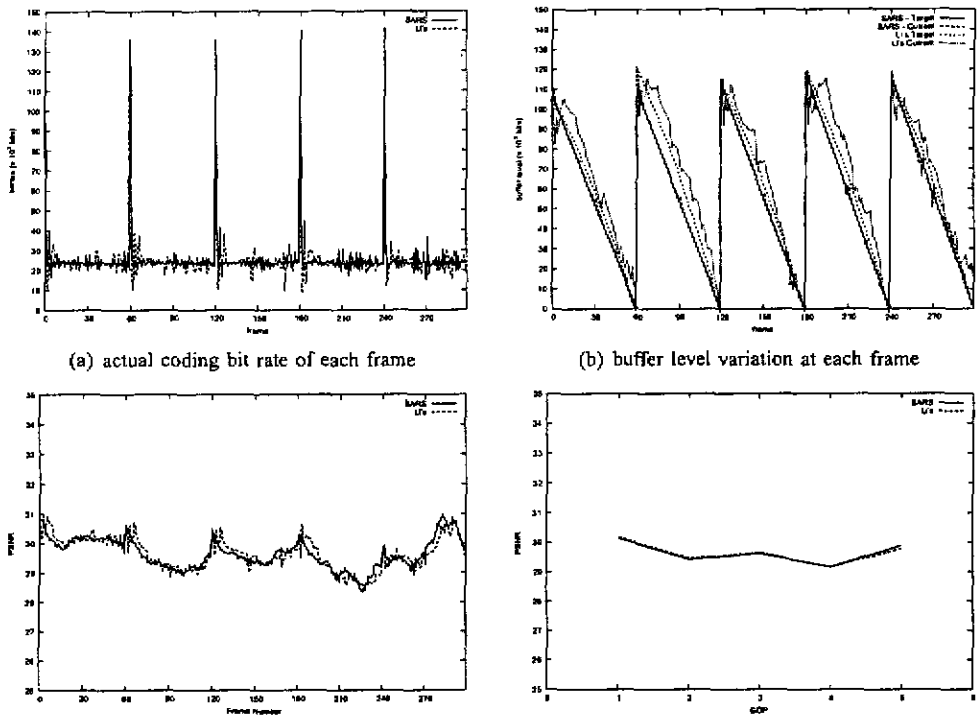
Fig. 18. Target bit rate for encoding.

2) CBR-512Kbps: Fig. 14(a) - 15(d) show the similar results that we have tested in CBR-768Kbps scenario. Fig. 15(d) shows the average PSNR of each GOP of the sequence "paris" by using Li's scheme and *SARS*. It can be shown that the *SARS* produces much higher PSNR values than that of Li's scheme.

3) CBR-256Kbps: Fig. 16(a) shows the actual coding bit rate of each frame by using Li's scheme and *SARS*. It can be shown that, for *SARS*, the variation of coding bit rate is still smaller than that in Li's scheme. Fig. 16(b) shows the buffer level variations of the test sequence "mobile" by using Li's scheme and *SARS*. The initial increase of buffer level in Li's scheme is due to the control error. It can be seen clearly that the frames are started to be skipped in the first GOP and the skipping frame can't be stopped by using Li's scheme. It is noted that by using Li's scheme, there are 38 frames skipped, while there are 15 frames skipped by using *SARS*. In Fig. 17(a) - Fig. 17(d), the bits used by the I-frames in "Paris" video sequence are smaller than that in "Mobile" video sequence and there are enough remaining bits for allocating bit rate to other frames. The overall performance of *SARS* is still better than Li's scheme.

4) VBR: The VBR scenario is a pre-test before we experiment the *SARS* over tcp-friendly congestion control protocols. The target coding bit rate of these two test video sequences are given in Fig. 18. The frame rate is 30 and the GOP length is 60. The experimental results are shown in Fig. 19(a) - Fig. 20(d).

Fig. 19(a) and Fig. 19(b) show the coding bit rate of each frame and the variation of buffer level in test video sequence "Mobile", respectively. It can be shown that, for *SARS* in VBR scenario, the fluctuation of coding bit rate is small and the buffer level can meet its target, while the variation of buffer level in Li's scheme is large. Fig. 19(c) and Fig. 19(d) show the PSNR of each frame and each GOP in test video sequence



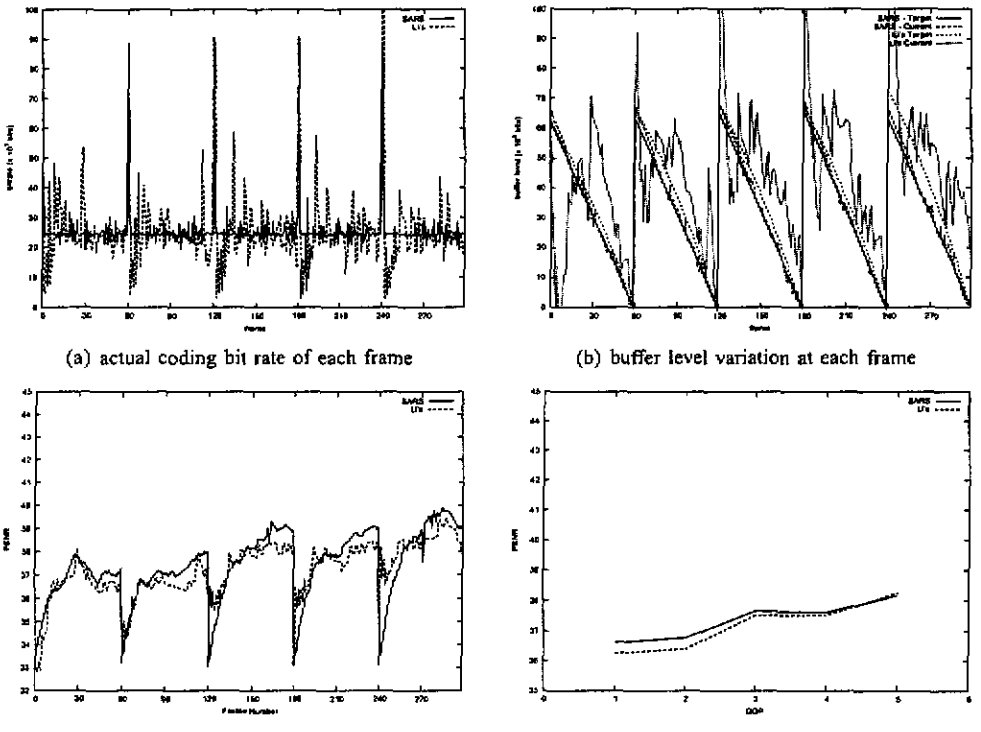
(a) actual coding bit rate of each frame

(b) buffer level variation at each frame

(c) PSNR results of each frame

(d) PSNR results of each GOP

Fig. 12. Performance comparison of video sequence "Mobile" in CBR-768Kbps scenario when the proposed rate control scheme and Li's adaptive rate control scheme [1] are applied to the MPEG-4 coding.



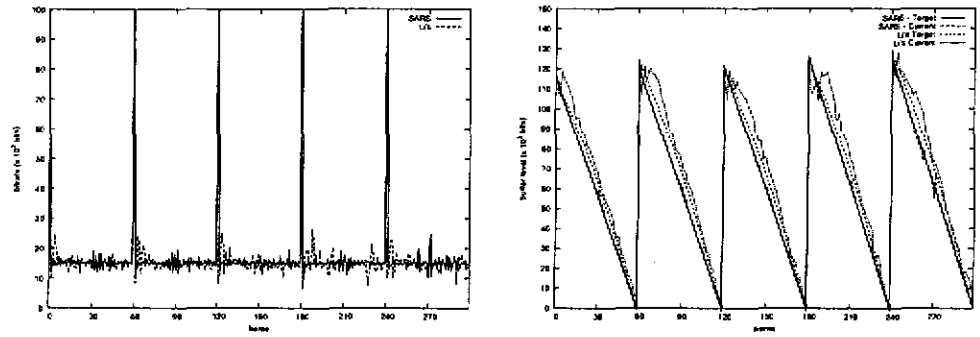
(a) actual coding bit rate of each frame

(b) buffer level variation at each frame

(c) PSNR results of each frame

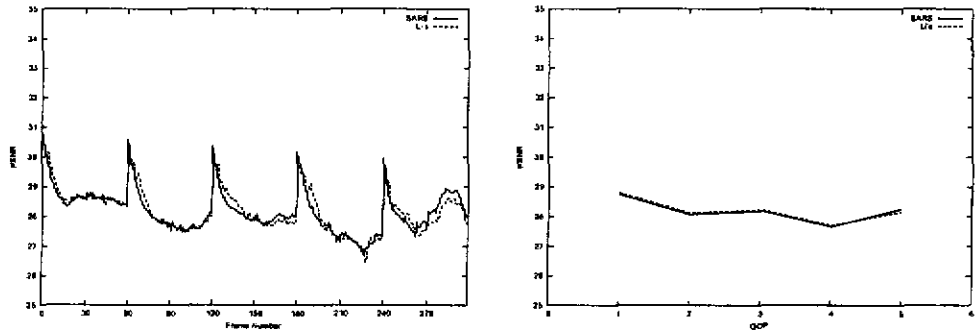
(d) PSNR results of each GOP

Fig. 13. Performance comparison of video sequence "Paris" in CBR-768Kbps scenario when *SARS* and Li's adaptive rate control scheme [1] are applied to the MPEG-4 coding.



(a) actual coding bit rate of each frame

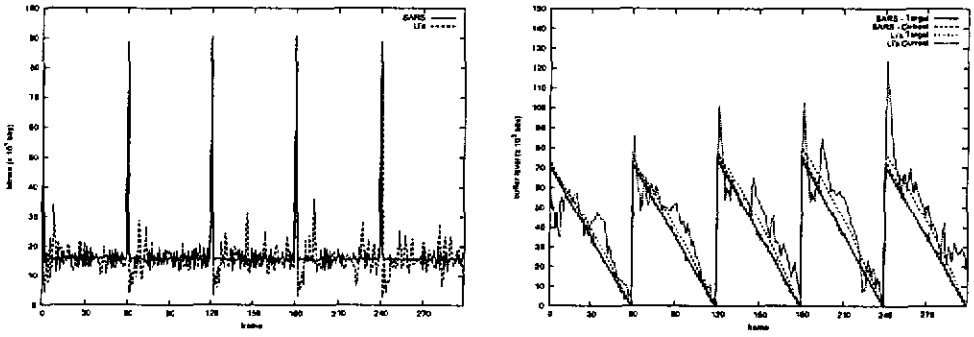
(b) buffer level variation at each frame



(c) PSNR results of each frame

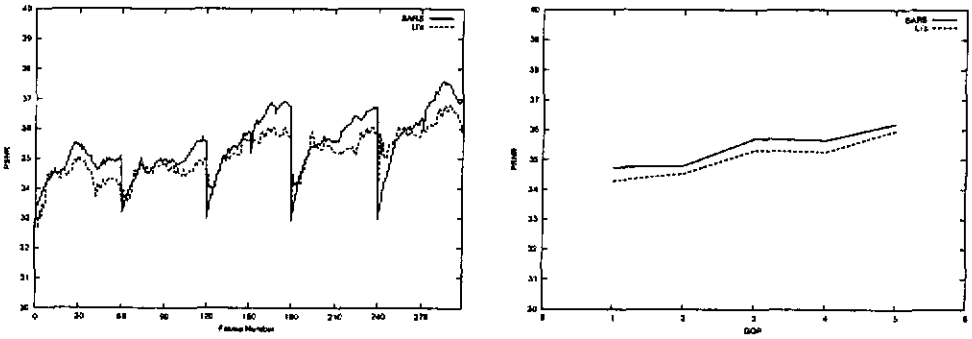
(d) PSNR results of each GOP

Fig. 14. Performance comparison of video sequence "Mobile" in CBR-512Kbps scenario when *SARS* and Li's scheme [1] are applied to the MPEG-4 coding.



(a) actual coding bit rate of each frame

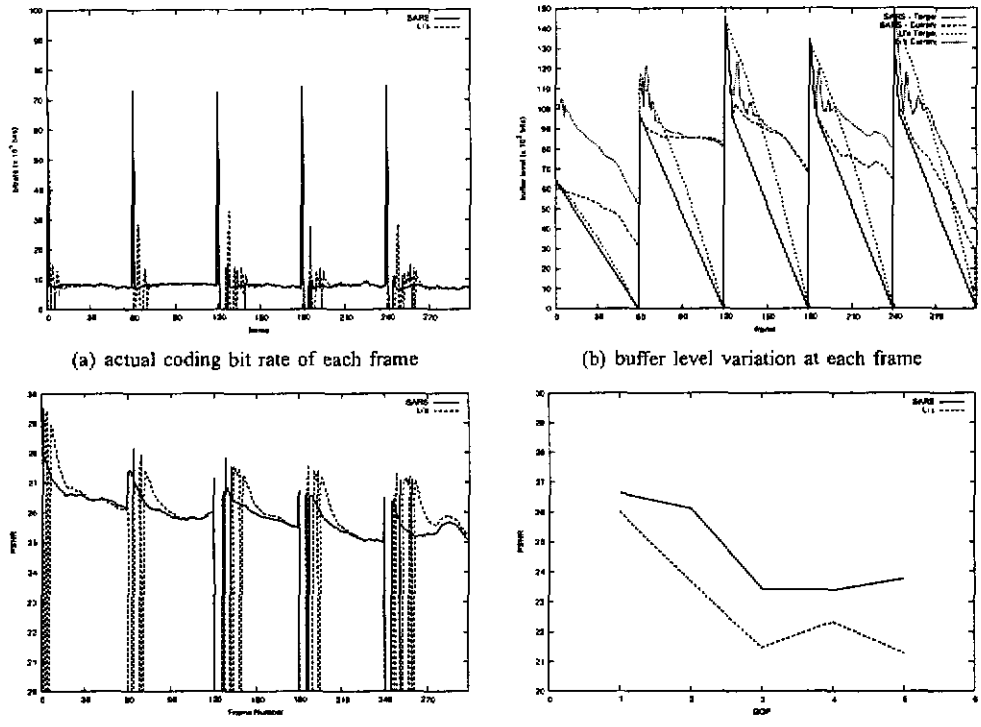
(b) buffer level variation at each frame



(c) PSNR results of each frame

(d) PSNR results of each GOP

Fig. 15. Performance comparison of video sequence "Paris" in CBR-512Kbps scenario when *SARS* and Li's scheme [1] are applied to the MPEG-4 coding.



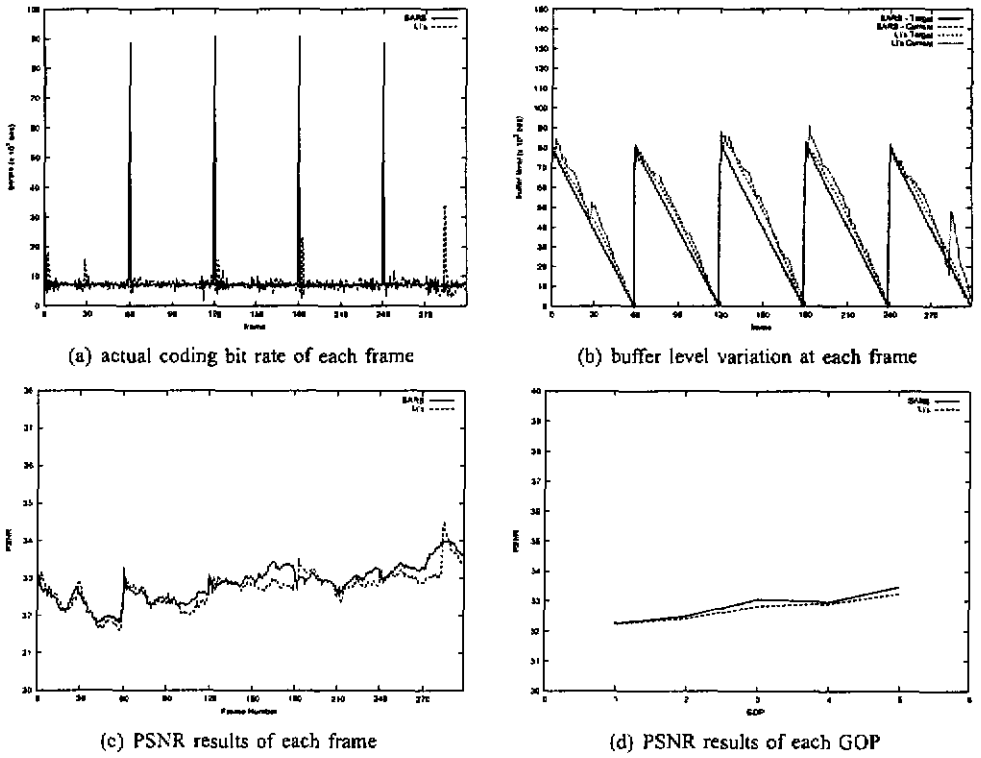
(a) actual coding bit rate of each frame

(b) buffer level variation at each frame

(c) PSNR results of each frame

(d) PSNR results of each GOP

Fig. 16. Performance comparison of video sequence "Mobile" in CBR-256Kbps scenario when *SARS* and Li's adaptive rate control scheme [1] are applied to the MPEG-4 coding.



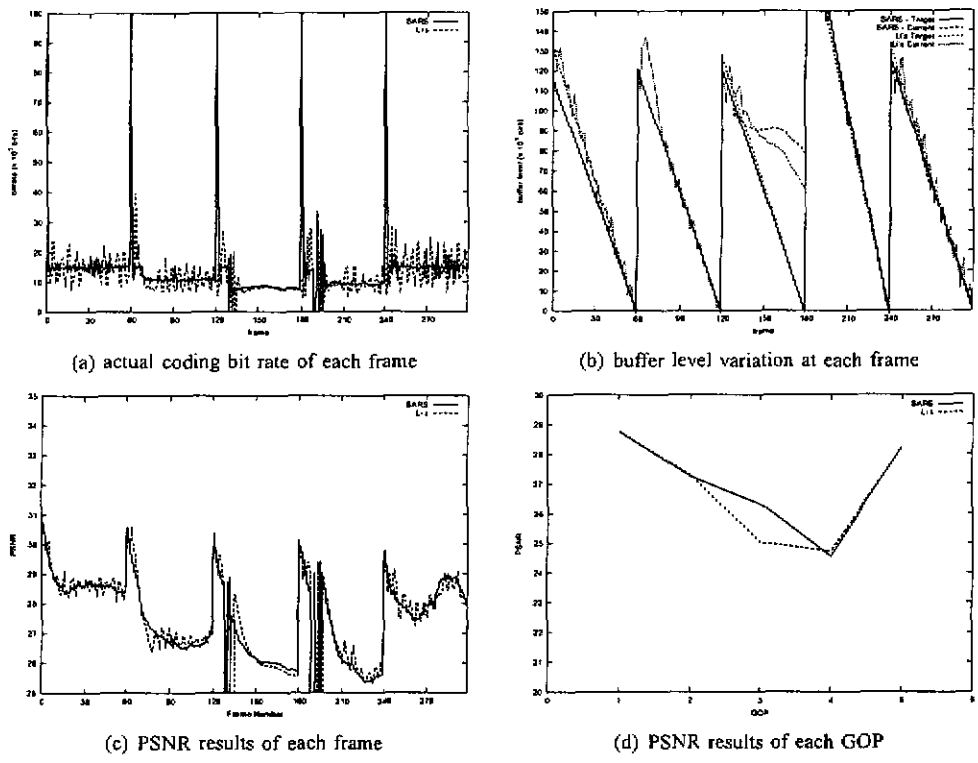
(a) actual coding bit rate of each frame

(b) buffer level variation at each frame

(c) PSNR results of each frame

(d) PSNR results of each GOP

Fig. 17. Performance comparison of video sequence "Paris" in CBR-256Kbps scenario when *SARS* and Li's adaptive rate control scheme [1] are applied to the MPEG-4 coding.



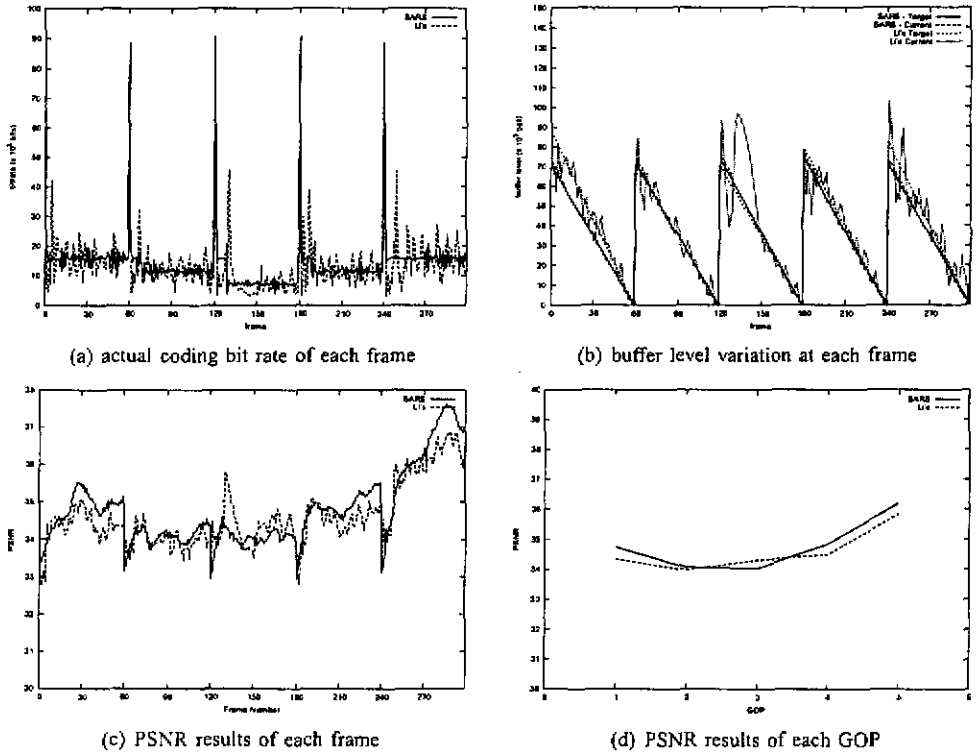
(a) actual coding bit rate of each frame

(b) buffer level variation at each frame

(c) PSNR results of each frame

(d) PSNR results of each GOP

Fig. 19. Performance comparison of video sequence "Mobile" in VBR scenario when *SARS* and Li's adaptive rate control scheme [1] are applied to the MPEG-4 coding.



(a) actual coding bit rate of each frame

(b) buffer level variation at each frame

(c) PSNR results of each frame

(d) PSNR results of each GOP

Fig. 20. Performance comparison of video sequence "Paris" in VBR scenario when *SARS* and Li's adaptive rate control scheme [1] are applied to the MPEG-4 coding.

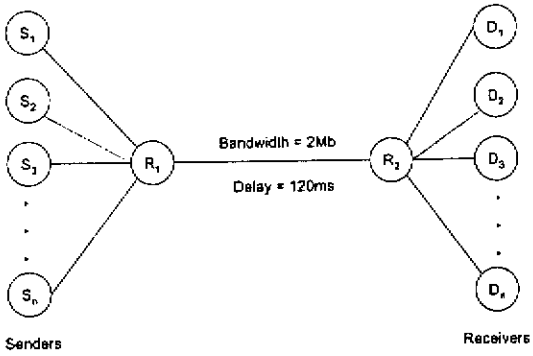


Fig. 21. Network Topology

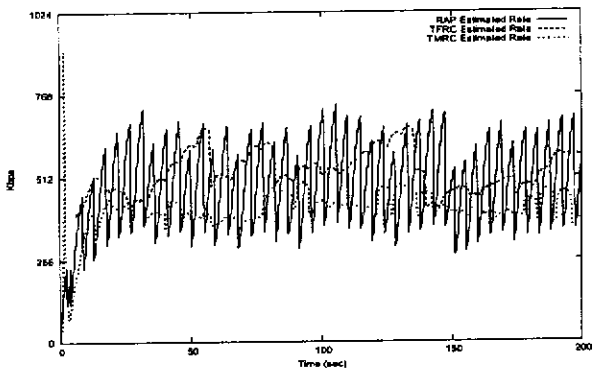


Fig. 22. Estimated sending rate of different rate based congestion control protocols. The scenario is 2 TCP versus 2 rate based congestion control protocols.

"Mobile", respectively. It can be seen clearly that, for Li's scheme, the average PSNR of GOP is much lower than that in *SARS* due to more skipped frames in Li's scheme. It is noted that by using Li's scheme, there are 17 frames skipped, while there are 9 frames skipped by using *SARS*. In Fig. 20(a) - Fig. 20(d) show the experimental results of the test video sequence "Paris". It can be seen clearly that, for Li's scheme, the average PSNR of third GOP is higher than that in *SARS* due to the high control error of coding bit rate at 130^{th} frame. However, the overall performance of *SARS* is still better than Li's scheme.

5) Summary of The Experimental Results: The experimental results are summarized in Table I. It can been shown that the performance of *SARS* is better than the rate control scheme in [1]. The average PSNR is improved up to 1.7dB and the number of skipped frames is reduced up to 23.

B. Performance Evaluation of SARS over rate based congestion control protocols

In this section we will analyze the performance of the MPEG-4 video transmission system described in Section III. In order to analyze the performance of *SARS* over time-varying available bandwidth, we implement the streaming function in the MPEG-4 encoder with proposed rate control scheme for network simulator (ns2) [46]. We first generate the traffic trace from the network simulator (ns2) and then simulate the MPEG-4 video streaming over rate based congestion control protocols by using the traffic trace and our proposed rate control scheme.

1) Network Topology: As a case study the experiments are performed using a relatively simple network topology shown in Fig. 21 with a bottleneck link. R_1 and R_2 are routers at the both ends of the bottleneck link that congestion may occur while traffic exceed to link capacity. S_n denotes senders which generate TCP, RAP, TMRC and TFRC flows and D_n denotes the corresponding receivers.

2) 2 TCP vs 2 Rate-Based Congestion Control Protocols: The experiment is run with the following parameters:

- Bottleneck bandwidth : 2Mb
- Number of competing TCP flows : 2
- Number of competing RAP or TFRC or TMRC flows : 2
- Link latency : 120ms

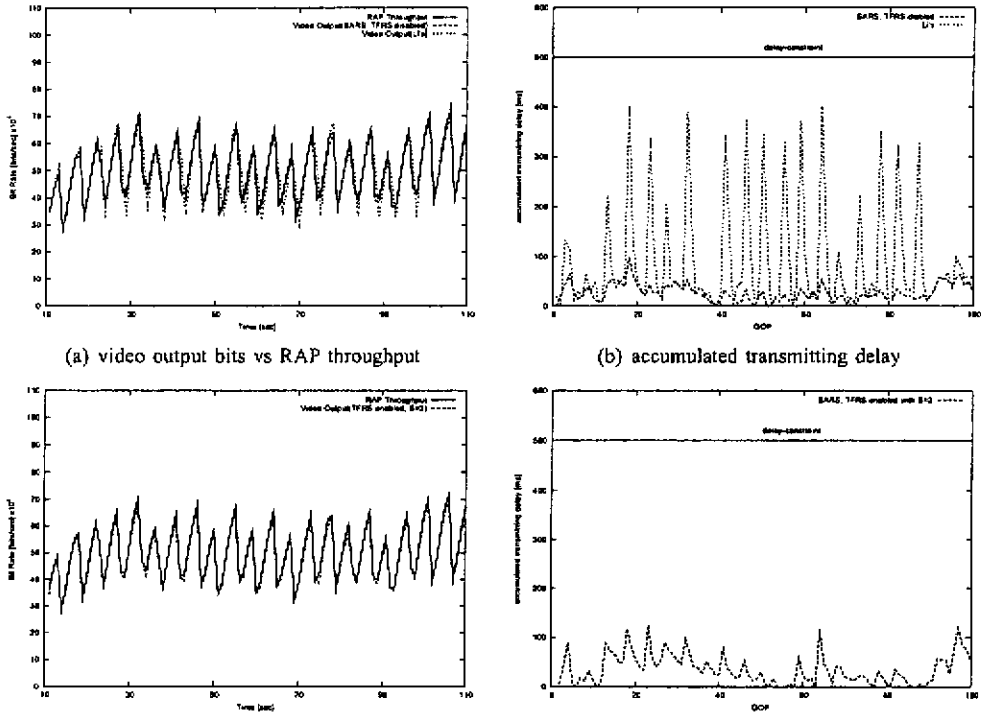
There are total three scenarios in this section, including 2 TCP vs 2 RAP, 2 TCP vs 2 TFRC, and 2 TCP vs 2 TMRC. We choose one of RAP flows, TFRC flows and TMRC flows and plot their estimated sending rate in Fig. 22

The video streaming is initialized at time 10s and stopped at time 110s. Total simulation time is 100s. The maximum delay-constraint is 500ms. The video sequence "Coastguard" with CIF size is tested. The simulation results, including video output versus throughput, accumulated transmission delay over different rate based congestion control protocols are plotted in Fig. 23 - Fig. 25, respectively. The transmission delay is the time taken to transmit all the frames in a GOP out at each GOP interval. Since the length of GOP is 30 and the frame rate is 30, the GOP interval is 1 second. The end-to-end delay of a frame is defined as the time taken to transmit a frame from sender to receiver.

Fig. 23(a) shows the amount of bits produced by video encoder and the throughput of RAP. It can be shown that, for Li's scheme, the video output does not meet the throughput of RAP, while *SARS* works well with RAP. Fig. 23(b) shows the accumulated transmitting delay of each GOP. It can be seen clearly that, for *SARS*, the transmitting delay is smoother than that in Li's scheme. It is noted that by using Li's scheme, there are 74 frames skipped, while there is no frame skipped by using *SARS*. The overall performance of *SARS* over TFRC or TMRC also show the better results than that of Li's scheme.

Table II shows the summarized results by using Li's scheme and *SARS*. The transmission delay of *SARS* is under the maximum delay-constraint and it is also smoother than that of Li's rate control scheme. It can be shown that, for *SARS*, the average PSNR is improved up to 0.018 dB with TFRS module enabled and the deviation of PSNR is also reduced. However, TFRS module can not enhance the average PSNR while streaming video over TFRC or TMRC much due to the estimated rate is already smoother than RAP. The lost frames using RAP are much more than they using TFRC or TMRC although the delay is smaller using RAP.

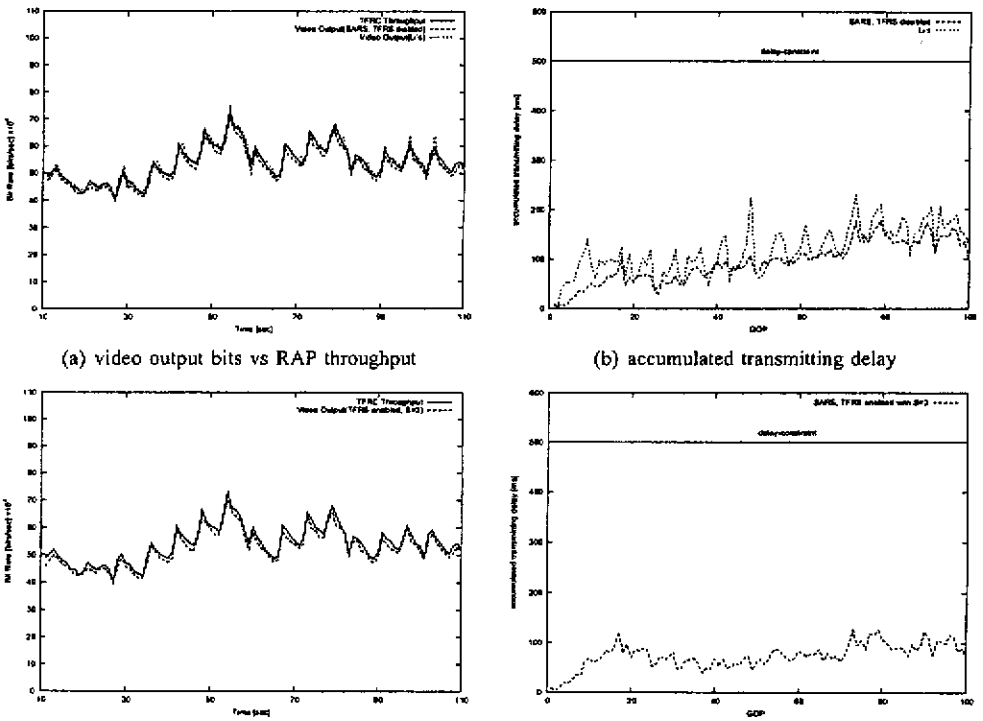
3) 2 TCP vs 2 Rate-Based Congestion Control Protocols with 4 TCP Competing during 20s and 80s: The experiment



(a) video output bits vs RAP throughput (b) accumulated transmitting delay

(c) video output bits vs RAP throughput (d) accumulated transmitting delay

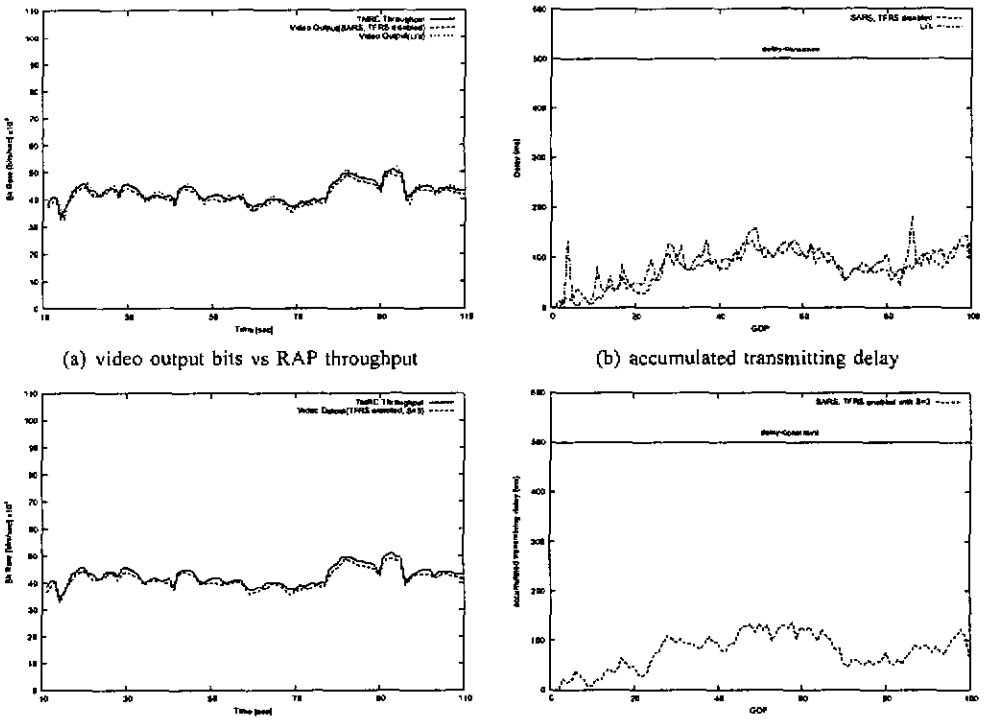
Fig. 23. Performance comparison of video sequence "Coastguard" using RAP as the transport protocol. (c) and (d) are the results that the TFRS in *SARS* is enabled with sampling timer S=3.



(a) video output bits vs RAP throughput (b) accumulated transmitting delay

(c) video output bits vs RAP throughput (d) accumulated transmitting delay

Fig. 24. Performance comparison of video sequence "Coastguard" using TFRC as the transport protocol. (c) and (d) are the results that the TFRS in *SARS* is enabled with sampling timer S=3.



(a) video output bits vs RAP throughput

(b) accumulated transmitting delay

(c) video output bits vs RAP throughput

(d) accumulated transmitting delay

Fig. 25. Performance comparison of video sequence "Coastguard" using TMRC as the transport protocol. (c) and (d) are the results that the TFRS in *SARS* is enabled with sampling timer S=3.

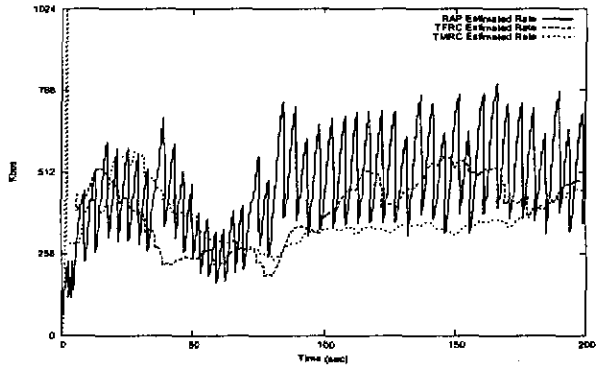


Fig. 26. Estimated sending rate of different rate based congestion control protocols. The scenario is 2 TCP versus 2 rate based congestion control protocols with 4 TCP starting at 20s and stopping at 80s

is run with the following parameters:

- Bottleneck bandwidth : 2Mb
- Number of competing TCP flows : 6(4 TCP start at 20s and stop at 80s)
- Number of competing RAP or TFRC or TMRC flows : 2
- Link latency : 120ms

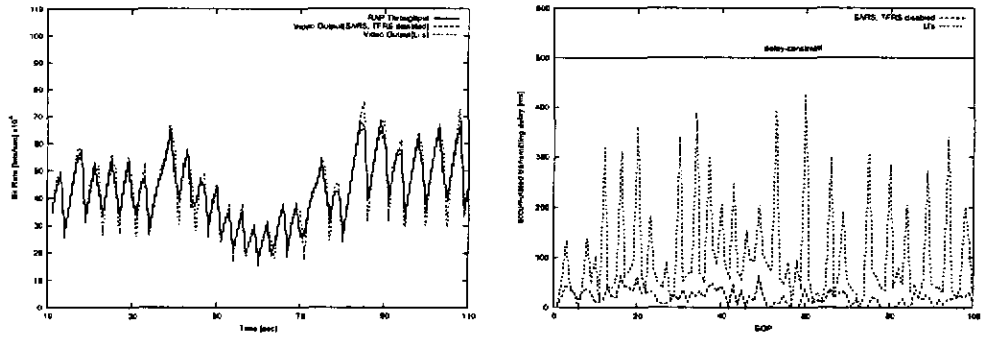
There are total three scenarios in this section, including 2 TCP vs 2 RAP, 2 TCP vs 2 TFRC, and 2 TCP vs 2 TMRC. There will be 4 TCP started at 20s and stopped at 80s in these scenarios. We choose one of RAP flows, TFRC flows and TMRC flows and plot their estimated sending rate in Fig. 26.

The configuration of the simulations in this section is the same as it in previous section while the video sequence "Container" with CIF size is tested. The simulation results, video output versus throughput, over different rate based congestion control protocols are shown in Fig. 27 - Fig. 29, respectively.

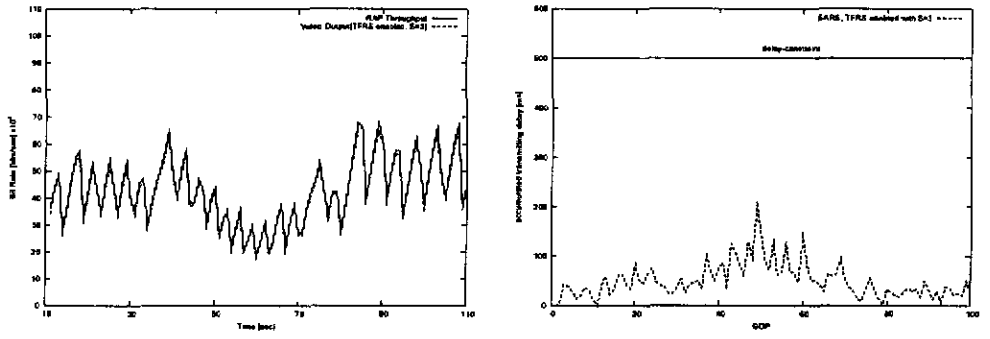
Fig. 27(a) shows the bits of video output and the throughput of RAP protocol. It can be shown that, for Li's scheme, video output still can not meet RAP's throughput, while *SARS* works well as the same as it does in previous scenario. Fig. 27(b) shows the accumulated transmitting delay by using Li's scheme and *SARS*. It can be seen clearly that, for Li's scheme, the accumulated transmitting delay is still large and worse than that in previous scenario. It is noted that by using Li's scheme, there are 88 frames skipped, while there are still no frame skipped by using *SARS*. Table III shows the summarized results in this scenario. However, the average PSNR by using *SARS* over TFRC or TMRC is lower than that by using Li's scheme.

V. Conclusion

In this paper we analyze the problem of transmitting MPEG-4 video over IP when senders use a TCP-friendly algorithm to be friendly to TCP traffic. This paper is focused on solving the challenging issues in application layer, compression layer. With this aim, we provide a solution called "*SARS*" (Smoothed and Adaptive Rate-control Scheme) in compression layer and application layer for an MPEG-4 video transmission system and discuss the characteristics of each its elements. The *SARS* contains three main components, including *TFRS module*, *LD module*, and *RSF module*. We study the statistical properties of video source and propose an improved linear source model

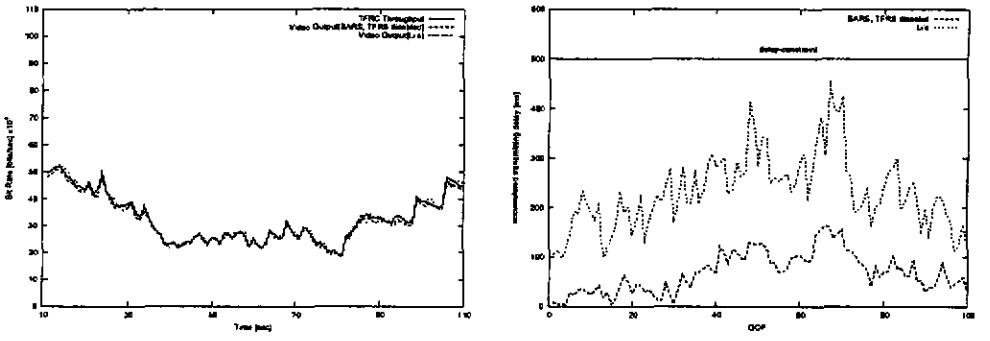


(a) video output bits vs RAP throughput (b) accumulated transmitting delay

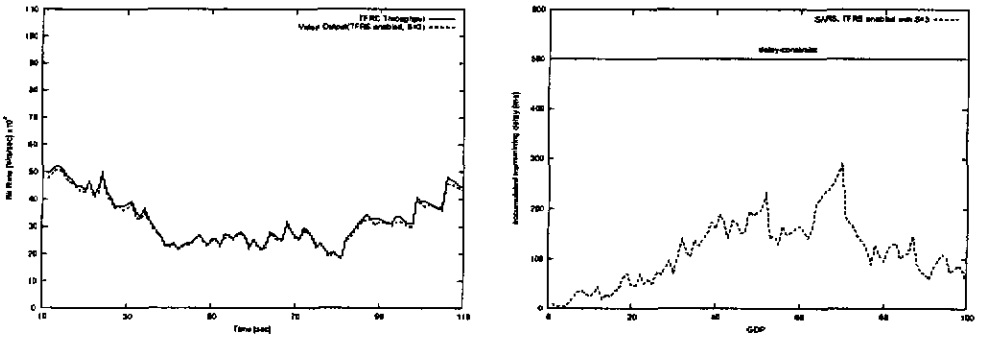


(c) video output bits vs RAP throughput (d) accumulated transmitting delay

Fig. 27. Performance comparison of video sequence "Container" using RAP as the transport protocol. (c) and (d) are the results that the TFRS in *SARS* is enabled with sampling timer S=3.

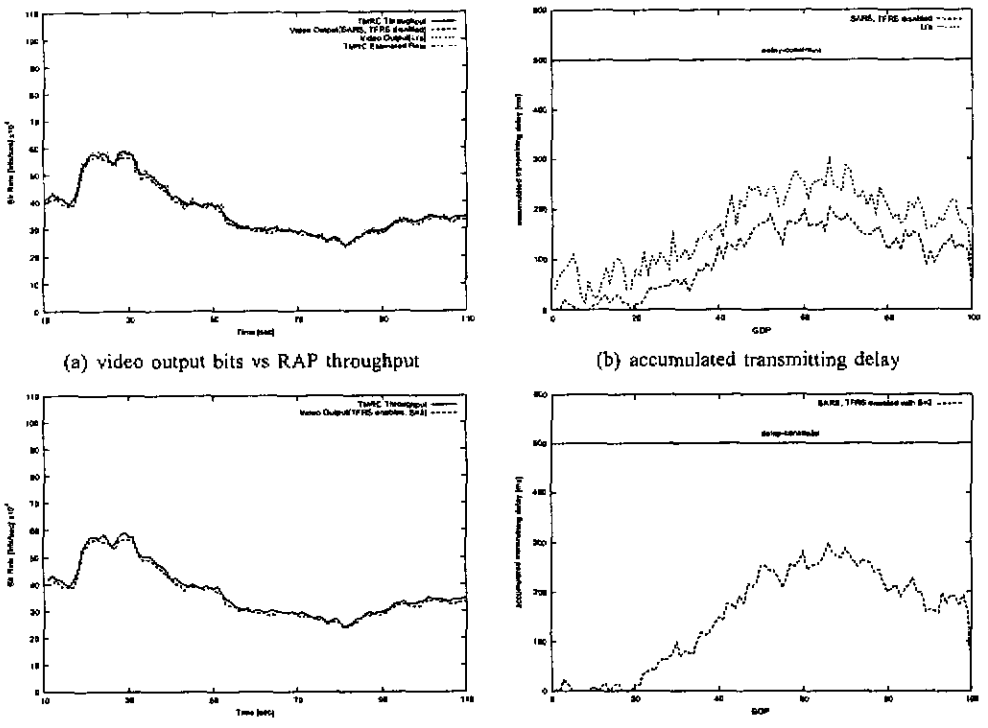


(a) video output bits vs RAP throughput (b) accumulated transmitting delay



(c) video output bits vs RAP throughput (d) accumulated transmitting delay

Fig. 28. Performance comparison of video sequence "Container" using TFRC as the transport protocol. (c) and (d) are the results that the TFRS in *SARS* is enabled with sampling timer S=3.



(a) video output bits vs RAP throughput

(b) accumulated transmitting delay

(c) video output bits vs RAP throughput

(d) accumulated transmitting delay

Fig. 29. Performance comparison of video sequence "Container" using TMRC as the transport protocol. (c) and (d) are the results that the TFRS in *SARS* is enabled with sampling timer S=3.

named "*LD module*" using the non-zero ratio of quantized coefficients. The proposed linear source model can provide better prediction of coding bit rate and better perceptual video quality. The delay is lowered with this linear source model. Based on the linear source model, an accurate and adaptive rate control scheme named "*RSF module*" is proposed. The coding bit rate and buffer level are much closer to their target with the proposed rate control scheme. The number of skipped frames are also reduced with the proposed adaptive rate control scheme. The overall performance of the *SARS* is increased even in the CBR scenarios or VBR scenarios. We also evaluate the performance of *SARS* over several rate based congestion control protocols such as RAP, TFRC and TMRC. A tcp-friendly rate smoother with memory wiper named "*TFRS module*" is further designed to prevent the video quality from fluctuating with the time-varying available bandwidth. The average PSNR is increased and the PSNR deviation is decreased on AIMD-model rate based congestion control protocols such as RAP. The average PSNR is also slightly increased using equation-model rate based congestion control protocols such as TFRC and TMRC. The transmission delay is controlled under the maximum delay constraint and it is shorter and smoother than Li's rate control scheme. The number of lost frames is a little bit high using RAP as transport protocol although RAP has lower transmission delay. TFRC or TMRC has higher transmission delay but lower number of lost frames. There are tradeoffs for a real-time video streaming applications using these different tcp-friendly protocols.

Even though we provided the solutions for the challenging issues of bandwidth and end-to-end delay, there is still a salient characteristic of delivering real-time video streams on Internet and it is the vulnerability of packet loss. As mentioned in Section I and Section II, the error control techniques is needed to be introduced to recover the packet loss and then to reduce the presentation displeasing to human eyes, especially for delivering real-time video over wireless link. The error control techniques which demand additional bandwidth further interact with *SARS* and the proposed layer-structure real-time video streaming system could be thus conducted in the end-to-end QoS video streaming solutions.

REFERENCES

[1] Z. G. Li, C. Zhu, N. Ling, X. K. Yang, G. N. Feng, S. Wu, and F. Pan, "A Unified Architecture for Real-Time Video-Coding Systems", IEEE Trans. Circuits Syst. Video Technol., vol 13, pp. 472-487, June 2003.

[2] Z. He, Y.-K. Kim, and Sanjit K. Mitra, "Low-Delay Rate Control for DCT Video Coding via ρ-Doamin Source Modeling", IEEE Trans. Circuits Syst. Video Technol., vol 11, pp. 928-940, Aug. 2001.

[3] Z. He and Sanjit K. Mitra, "A Linear Source Model and a Unified Rate Control Algorithm for DCT Video Coding", IEEE Trans. Circuits Syst. Video Technol., vol 12, pp. 970-982, Nov. 2002.

[4] H. Shojania and B. Li, "Experiences with MPEG-4 Multimedia Streaming", ACM Multimedia, pp. 492-494, 2001.

[5] B. Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J. Wroclawski, and L. Zhang, "Recommendations on Queue Management and Congestion Avoidance in the Internet", RFC 2309, Informational, Apr. 1998.

[6] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Transporting Real-Time Video over the Internet: Challenges and Approaches", Proc. IEEE, vol. 88, pp. 1855-1877, Dec. 2000.

[7] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, June 1994.

[8] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, IETF, Dec. 1998.

[9] G. Carle and E. W. Biersack, "Survey of Error Recovery Techniques for IP-based Audio-Visual Multicast Applications", IEEE Networks, vol. 11, no. 6, pp. 24-36, Nov. 1997.

[10] U. Horn, K. Stuhlmüller, M. Link, and B. Girod, "Robust Internet Video Transmission Based on Scalable Coding And Unequal Error Protection", Signal Processing: Image Communication, vol.15, no. 1-2, pp. 77-94, Sept. 1999.

[11] Q. Zhang, W. Zhu, and Y.-Q. Zhang, "Resource Allocation for Multimedia Streaming over The Internet", IEEE Trans. Multimedia, vol. 3, no. 3, pp. 339-355, Sept. 2001.

[12] C.-H. Wang, R.-I Chang, J.-M. Ho, S.-C. Hsu, "Rate-Sensitive ARQ for Real-time Video Streaming", in Proc. GLOBECOM'03, San Francisco, USA, Dec. 2003.

[13] L. Chiariglione, "MPEG-4 FAQs", Technical report, ISO/IEQ JTC1/SC29/WG11, July 1997.

[14] ISO/IEC JTC1/SC29/WG11, "Overview of the MPEG-4 Standard", N4668, Mar. 2002.

[15] ISO/IEC JTC1/SC29/WG11, "Information Technology - Coding of Audiovisual Objects - Part 2: Visual", N2502, FDIS 14496-2.

[16] MPEG Homepage, http://www.cselt.it/mpeg/.

[17] H. Radha, Y. Chen, K. Parthasarathy, and R. Cohen, "Scalable Interenet Video Using MPEG-4", Signal Processing: Image Commun., no. 15, pp. 95-126, Sept. 1999.

[18] H. Radha and Y. Chen, "Fine-Granular-Scalable Video for Packet Networks", Packet Video, Arp. 1999.

[19] S. Nelakuditi, R. R Harinath, E. Kusmierek, and Z.-L. Zhang, "Providing Smoother Quality Layered Video Stream", in Proc. of NOSSDAV'00, Chapel Hill, North Carolina, June 2000.

[20] P. de Cuetos and K. Ross, "Adaptive Rate Control for Streaming Stored Fine-Grained Scalable Video", in Proc. of NOSSDAV'02, pp. 3-12, Miami, Florida, May 2002.

[21] P. de Cuetos, P. Guillotel, K. Ross, and D. Thoreau, "Implementation of Adaptive Streaming of Stored MPEG-4 FGS Video Over TCP", in Proc. of ICME'02, pp. 405-408, Lausanne, Switzerland, Aug. 2002.

[22] L. Zhao, J. Kim, and C.-C Jay Kuo, "MPEG-4 FGS Video Streaming with Constant-Quality Rate Control and Differentiated Forwarding", in Proc. VCIP'02, San Jose, CA, Jan. 2002.

[23] D. T. Hoang and J. S. Vitter, "Efficient Algorithm for MPEG Video Compression", Wiley-Interscience Inc. Sep. 2001.

[24] W. Ding and B. Liu, "Rate Control of MPEG Video Coding and Recording by Rate-Quantization Modeling", IEEE Trans. Circuit Syst. Video Technol., vol 6, pp. 12-20, Feb. 1996.

[25] H.-M. Hang and J.-J. Chen, "Source Model for Tranform Video Coder and Its Application - Part I: Fundamental Theory", IEEE Trans. Circuits Syst. Video Technol., vol 7, pp. 287-298, Apr. 1997.

[26] Bo Tao, B.W. Dickson, and H.A. Peterson, "Adaptive Model-Driven Bit Allocation for MPEG Video Coding", IEEE Trans. Circuits Syst. Video Technol., vol. 10, pp. 147-157, Feb. 2000.

[27] T. Chiang and Y.-Q. Zhang, "A New Rate Control Scheme Using Quadratic Rate Distortion Model", IEEE Trans. Circuits Syst. Video Technol., vol 7, pp. 246-250, Feb. 1997.

[28] Jordi Ribas-Corbera and Shawmin Lei, "Rate Control in DCT Video Coding for Low-Delay Communications", IEEE Trans. Circuits Syst. Video Technol., vol. 9, pp. 172-185, Feb. 1999.

[29] Jordi Ribas-Corbera and Shawmin Lei, "A Frame-Layer Bit Allocation for H.263+", IEEE Trans. Circuits Syst. Video Technol., vol. 10, pp. 1154-1158, Oct. 2000.

[30] H.-J. Lee, T. Chiang, and Y.-Q. Zhang, "Scalable Rate Control for MPEG-4 Video", IEEE Trans. Circuits Syst. Video Technol., vol. 10, pp. 878-894, Sept. 2000.

[31] Z. Lei and N. D. Georganas, "Rate Adaptaion Transcoding for Precoded Video Streams", in Proc. ACM Multimedia'02, Juan-les-Pins, France, Dec. 2002.

[32] Z.G. Li, N. Ling, G. N. Feng, F. Pan, K. P. Lim, and S. Wu, "Adaptive Rate Control For Real Time Video Coding Process", in Proc. DCV'02, Clearwater, Florida, Nov. 2002.

[33] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 Rate Control for Multiple Video Objects", IEEE Trans. Circuits Syst. Video Technol., vol. 9, pp. 186-199, Feb. 1999.

[34] ISO/IEC JTC1/SC29/WG11, "MPEG-4 Video Verification Model Version 18.0", N4350, July 2001.

[35] F. Pan, Z. Li, K. Lim, and G. Feng, "Reducing Frame Skipping in MPEG-4 Rate Control Scheme", in Proc. ICASSP 2002, vol. 4, pp.3409-3412, Orlando, FL, May 13-17 2002.

[36] F. Pan, Z. Li, K. Lim, and G. Feng, "A Study of MPEG-4 Rate Control Scheme and Its Improvements", IEEE Trans. Circuits Syst. Video Technol., vol 13, pp. 440-446, May 2003.

[37] Y. R. Yang and S. S. Lam, "General AIMD Congestion Control", in Proc. ICNP 2000, Osaka, Japan, Nov. 2000.

[38] D. Bansal and H. Balakrishnan, "Binomial Congestion Control Algorithms", in Proc. INFOCOM'01, Apr. 2001.

[39] R. Rejaie, M. Handley, and D. Estrin, "RAP: An End-to-end Rate-based Congestion Control Mechanism for Realtime Streams in the Internet", in Proc. INFOCOM'99, New York, Mar. 1999.

[40] I. Rhee, V. Ozdemir, and Y. Yi, "TEAR: TCP Emulation at Receiver - Flow Control for Multimedia Streaming", NCSU Technical Report, Apr. 2000.

[41] S. Floyd, M. Handley, J. Padhye, and J. Widmer, "Equation-Based Congestion Control for Unicast Applications", in Proc. ACM SIGCOMM Symposium on Communications Architecture and Protocols, Aug. 2000.

[42] W. Tseng, E. Wu and K. Chang, "TMRC: A Load-Adaptive TCP-Friendly Rate Control Protocol for Real-Time Multimedia Applications", Appear on Proc. on IEEE ICC 2004.

[43] R. Puri, K.-W. Lee, K. Ramchandran, and V. Bharghavan, "An Integrated Source Transcoding and Congestion Control Paradigm for Video Streaming in the Internet", IEEE Trans. Multimedia, vol. 3, pp. 18-32, Mar. 2001.

[44] S. Jin, L. Guo, I. Matta, and A. Bestavros, "A Spectrum of TCP-Friendly Window-Based Congestion Control Algorithms", IEEE/ACM Trans. Networking, vol. 11, pp. 341-355, June 2003.

[45] Simon A.J. Winder, "ISO/IEC 14496 (MPEG-4) Video Reference Software", Microsoft-FDAM1-2.4-021205.

[46] http://www.isi.edu/nsnam/ns/

Video Sequence	Target Bit Rate Kbps	Actual Bit Rate Kbps ([1])	Actual Bit Rate Kbps (SARS)	Skipped Frames ([1])	Skipped Frames (SARS)
mobile	768	769.05	768.05	0	0
paris	768	768.85	767.88	0	0
mobile	512	512.54	511.96	0	0
paris	512	513.52	511.94	0	0
mobile	256	260.30	258.81	38	15
paris	256	256.31	256.02	0	0
mobile	VBR	404.26	405.85	17	9
paris	VBR	421.28	424.90	0	0

Video Sequence	Target Bit Rate Kbps	Average PSNR dB ([1])	Average PSNR dB (SARS)	Gain in PSNR dB
mobile	768	29.655	29.656	+0.001
paris	768	37.191	37.376	+0.185
mobile	512	28.209	28.193	-0.016
paris	512	35.077	35.427	+0.350
mobile	256	22.945	24.666	+1.721
paris	256	32.727	32.853	+0.126
mobile	VBR	26.006	26.618	+0.612
paris	VBR	34.592	34.781	+0.189

TABLE I

THE SUMMARY OF PERFORMANCE COMPARISON AMONG DIFFERENT RATE CONTROL SCHEMES.

Protocols	Average PSNR dB	PSNR deviation	Received Frames	Lost Frames
RAP (Li's)	36.9684	2.8286	2810	116
RAP (SARS, TFRS disabled)	37.9781	0.6786	2885	115
RAP (SARS, S=0)	37.9940	0.6760	2887	113
RAP (SARS, S=3)	37.9962	0.6675	2887	113
TFRC (Li's)	38.0401	0.4654	2985	15
TFRC (SARS, TFRS disabled)	38.2012	0.5256	2984	16
TFRC (SARS, S=0)	38.2057	0.5256	2985	15
TFRC (SARS, S=3)	38.2073	0.5198	2986	14
TMRC (Li's)	37.5717	0.4225	2983	17
TMRC (SARS, TFRS disabled)	37.6535	0.4940	2982	18
TMRC (SARS, S=0)	37.6603	0.4940	2984	16
TMRC (SARS, S=3)	37.6637	0.4900	2981	19

Protocols	Skipped Frames	Average Transmitting Delay ms	Average End-to-End Delay ms
RAP (Li's)	74	90.06	338.70
RAP (SARS, TFRS disabled)	0	26.86	280.06
RAP (SARS, S=0)	0	31.31	285.49
RAP (SARS, S=3)	0	35.49	290.16
TFRC (Li's)	0	117.15	399.09
TFRC (SARS, TFRS disabled)	0	93.01	376.64
TFRC (SARS, S=0)	0	91.43	374.45
TFRC (SARS, S=3)	0	74.28	356.61
TMRC (Li's)	0	87.86	388.48
TMRC (SARS, TFRS disabled)	0	77.35	370.88
TMRC (SARS, S=0)	0	71.88	366.66
TMRC (SARS, S=3)	0	77.49	371.05

TABLE II

THE OVERALL PERFORMANCE OF OUR RATE CONTROL SCHEME OVER DIFFERENT RATE BASED CONGESTION CONTROL PROTOCOLS WHICH COMPETE WITH 2 TCP FLOWS. THE TEST VIDEO SEQUENCE "COASTGUARD" IS USED.

Protocols	Average PSNR dB	PSNR deviation	Received Frames	Lost Frames
RAP (Li's)	38.0286	3.2823	2795	117
RAP (SARS, TFRS disabled)	38.7930	0.9280	2878	122
RAP (SARS, S=0)	38.8043	0.9216	2880	120
RAP (SARS, S=3)	38.8182	0.9138	2876	124
TFRC (Li's)	38.2312	0.8983	2922	78
TFRC (SARS, TFRS disabled)	38.0183	0.7284	2923	77
TFRC (SARS, S=0)	38.0195	0.7269	2921	79
TFRC (SARS, S=3)	38.0197	0.7262	2926	74
TMRC (Li's)	38.6289	0.8557	2908	92
TMRC (SARS, TFRS disabled)	38.4215	0.7256	2908	92
TMRC (SARS, S=0)	38.4218	0.7218	2908	92
TMRC (SARS, S=3)	38.4235	0.7218	2905	95

Protocols	Skipped Frames	Average Transmitting Delay ms	Average End-to-End Delay ms
RAP (Li's)	88	102.25	367.52
RAP (SARS, TFRS disabled)	0	20.33	289.24
RAP (SARS, S=0)	0	37.04	305.52
RAP (SARS, S=3)	0	50.28	319.15
TFRC (Li's)	0	231.77	538.04
TFRC (SARS, TFRS disabled)	0	68.58	367.49
TFRC (SARS, S=0)	0	81.41	377.84
TFRC (SARS, S=3)	0	113.68	411.43
TMRC (Li's)	0	165.01	473.09
TMRC (SARS, TFRS disabled)	0	102.16	406.38
TMRC (SARS, S=0)	0	116.40	420.64
TMRC (SARS, S=3)	0	148.55	452.58

TABLE III
THE OVERALL PERFORMANCE OF OUR RATE CONTROL SCHEME OVER DIFFERENT RATE BASED CONGESTION CONTROL PROTOCOLS WHICH COMPETE WITH 2 TCP FLOWS. THERE ARE 4 TCP FLOWS STARTING AT 20S AND STOPPING AT 80S. THE TEST VIDEO SEQUENCE "CONTAINER" IS USED.

Low Latency and Efficient Packet Scheduling for Streaming Applications

Eirc Hsiao-Kuang Wu*, Hsu-Te Lai*, Meng-feng Tsai *, Cheng-Fu Chou†

*Department of Computer Science and Information Engineering
National Central University, Chung-Li, Taiwan, R.O.C.
†Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.
hsiao@csie.ncu.edu.tw

Abstract— **Adequate bandwidth allocations and strict delay requirements are critical for real time applications. Packet scheduling algorithms like Class Based Queue (CBQ), Nested Deficit Round Robin (Nested-DRR) are designed to ensure the bandwidth reservation function. However, they might cause unsteady packet latencies and introduce extra application handling overhead, such as allocating a large buffer for playing the media stream. High and unstable latency of packets might jeopardize the corresponding Quality of Service since real-time applications prefer low playback latency. Existing scheduling algorithms which keep latency of packets stable require knowing the details of individual flows. GPS (General Processor Sharing)-like algorithms does not consider the real behavior of a stream. A real stream is not perfectly smooth after forwarded by routers. GPS-like algorithms will introduce extra delay on the stream which is not perfectly smooth. This thesis presents an algorithm which provides low latency and efficient packet scheduling service for streaming applications called LLEPS.**

I. Introduction

The number of new real-time applications is growing dramatically as more and more broadband services are available. The current best-effort service for Internet delivery of data cannot be sufficient for new multimedia applications. A variety of new multimedia applications such as voice over IP, video on demand and teleconferencing rely on network traffic scheduling algorithms in switches and routers to assure performance bounds to satisfy different QoS requirements. Typical VoIP (Voice over IP) applications require low latency and less bandwidth; VoD (Video on Demand) applications require a huge volume of bandwidth, but they can tolerate higher latency than VoIP applications. Unlike bursting data traffic applications, real-time applications usually do not try to consume all available bandwidth. In most cases, the required maximum bandwidth for a real-time application can be expected, such as a typical streaming application.

Since most of existing routers only provide best-effort services, the traffic of streaming applications will compete with other traffic of data services like FTP, HTTP and so on. The packets of streaming applications may experience higher latency or be dropped because of the competition among different kinds of traffic. Therefore, the quality of streaming application can not be guaranteed and new technologies are necessary to offer quality of service (QoS) for these real time applications.

A number of research solutions propose to enforce the fairness among Internet flows. The proposed fairness schemes makes that each flow consumes almost the similar bandwidth. They prevent the competition between flows and assure no starvation of any flow. The above researches assume that all flows have end to end congestion control mechanisms to adapt to the same compromised bandwidth assignments. The real-time application can only survive with enough bandwidth and different real-time applications may demand different bandwidths. Therefore, enforcing each flow getting the same bandwidth is not applicable.

Differentiated services (DiffServ) [1][2] and integrated services (IntServ) [3] are proposed to make the network to be QoS aware in IP networks in the mid 1990s. RSVP [4] is proposed for reservation for the IntServ approach and it defines how to allocate bandwidth hop by hop. Bandwidth allocation makes that each flow can get different bandwidth. Real-time applications can allocate bandwidth as they need, then they don't need to compete with other flows anymore.

Bandwidth reservation can ensure the delay of packets of real-time applications bounded into some value. Class Based Queue (CBQ) [5] has proposed a link sharing scheme which can be extended for bandwidth reservation implementation. However, CBQ does not consider the latency of packets and its service could be bursty for delay-bounded service constraints.

In addition to bandwidth guarantee, the latency of packets is required to be low and stable for some particular real-time applications. Bandwidth guarantee ensures no packets of streaming applications being dropped. Stable and low latency makes the playback latency of streaming-applications low. Therefore, *bandwidth guarantee* and *stable* and *low latency* are the two major requirements for stream applications. The bursting service providing by CBQ can not keep the latency of packets stable. Besides CBQ, other packet scheduling algorithms providing bandwidth reservation have been developed. They can be separated into three categories: GPS (Generalized Processor Sharing) like, frame based and regulative.

GPS is a concept of how multiple tasks (sessions) sharing a single processor in an OS. The sessions can get different

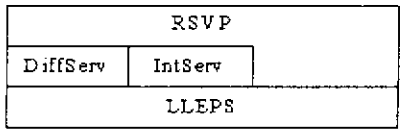


Fig. 1. DiffServ, IntServ, RSVP and LLEPS

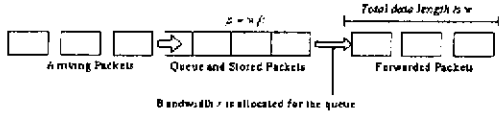


Fig. 2. Parameters of a queue

service rates at the same time, and they are served simultaneously. For applying GPS to the network technologies, it should be not noted that the service of GPS is bit by bit. However, the network service is packet by packet. Due to the different service units of GPS (bit-based) and Internet (packet-based), the algorithms WFQ [6] and WF^2Q [7] have been developed. Since they are all simulating GPS, they are classified as GPS like algorithms. The GPS like algorithms can extend the bandwidth sharing into bandwidth reservation and assure the packet behaviors smooth. In other words, GPS like packet scheduling algorithms mostly consider the latency of packets and bandwidth sharing ratio.

Frame based packet scheduling algorithms like DRR [8], NDRR [9] are different from GPS. They adopt different approaches to provide bandwidth sharing (reservation) services. They service all queues round by round and have no complex control mechanisms. The major advantage of frame based packet scheduling algorithms is efficiency. The disadvantage is that they only consider the queues which are always backlogged. That makes the latency of packets of the queues not always backlogged unstable.

The algorithms [10][11] are regulative packet scheduling algorithms. In regulative packet scheduling algorithms, each queue has a regulator which controls forwarding of packets. They can keep the latency of packets stable and the provide bandwidth reservation, but they are not scalable and hard to implement. Since the regulator demands to know the details of the flow, each regulator is only designed or configured for a particular flow. The regulative packet scheduling algorithm suffers the same problem of Integrated Service Framework [3].

II. Overview of LLEEPS

LLEPS is a packet scheduling algorithm. It can be the foundation of DiffServ, IntServ and RSVP. The relationship of them is represented in Figure 1. LLEPS can provide underlining packet scheduling service for them.

Whenever the bandwidth of bottleneck is not overwhelmed for on-going distinct multimedia applications, the consequent queuing delays of packets will be small and stable. Since the propagation time and transmission time of a packet on a link are constant values, the major cause of jitter is the queuing delay. In other words, as the queuing delay is small, the jitter will also be small. Streaming applications (ex: Mpeg Video Streaming) are not like bursty data oriented applications, such as TCP. They will not try to consume the bandwidth as much as possible. They usually consume the bandwidth as much as they need. LLEPS assumes streaming applications will not utilize more bandwidth than they need. The latency of packets can be bounded into a L/R value if the flow never uses more bandwidth than reserved bandwidth. L denotes the size of the packet and R denotes the reserved bandwidth.

LLEPS maintains multiple queues and each queue is used only for a session. Upon receiving a packet of a session, LLEPS puts it into a specific queue. The scheduler chooses the queue with the highest priority and forwards packets for the queue.

$[t_s,t_f]$ denotes a time interval from time t_s to t_f. The reserved bandwidth for the session i is r_i. The bandwidth of the link is B. The link can forward W bytes in the time interval $[t_s,t_f]$. The number of sessions in the system is n. The data length consumed by the session i from t_s to t is w_i^t. Under the condition that the inequality (3) holds, LLEPS can work. If the session i does not consume more bandwidth than r_i, the inequality (4) always holds.

$$W = B \times (t_f - t_s) \tag{1}$$

$$w_i = r_i \times (t_f - t_s) \tag{2}$$

$$\sum_{i=1}^{n} w_i \leq W \tag{3}$$

$$w_i^t \leq w_i \tag{4}$$

LLEPS gives the session i a higher priority when equation (4) is true. It denotes that LLEPS can guarantee bandwidth for each session even when other sessions consume bandwidth more than the bandwidth of the link.

$$w_i^t \leq r_i \times (t - t_s) \tag{5}$$

$$p_i = w_i^t / r_i \tag{6}$$

If the session i temporarily consumes more bandwidth than the reservation, the inequality (5) will not be true. LLEPS determines the priority of each session from this concept. Let p_i be the priority of the session i. The equation (6) shows the way the value of p_i is determined. The lower p_i denotes the higher priority in LLEPS. Figure 2 shows the relationship of p, w, r and the queue.

The value of p_i can be an indicator to reflect the greedy status for a particular session. At any moment, p_i of each queue should be similar. As a session is more aggressive than others, its p_i value will be larger than others'. According to the property of p_i, LLEPS only needs to forward packets of

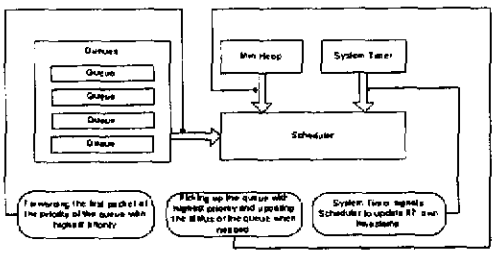


Fig. 3. Components of LLEPS

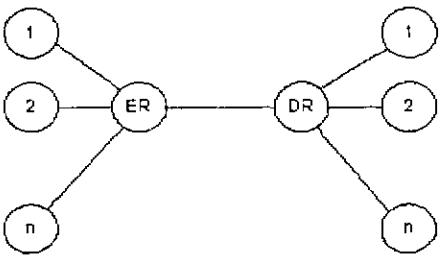


Fig. 4. Topology of Simulation

the queue with a lower p_i value than others. The behavior of LLEPS is able to keep the p_i value of each queue similar.

LLEPS provides the packet forwarding service based on the concept of time slice. LLEPS defines a time interval $[t_s,t_f]$ and tries to reserve bandwidth for each session. It can guarantee the bandwidth in the current time interval, the next time interval and so on. Let T be the length of $[t_s,t_f]$ and LLEPS must update the usage status of each queue every T seconds. It ensures that each session can get the bandwidth reserved for it in each time slice.

A. Components of LLEEPS

LLEPS contains these components: a system timer, a scheduler, queues and a min heap. Figure 3 introduces the basic architecture of LLEPS and the relationships between the three components.

1) Queues:

There are numbers of queues in the system. The relationship between queues and sessions is one to one. A session may contain multiple flows. When a packet arrives, LLEPS classifies it by the IP address or the port number of UDP or TCP and puts it into the corresponding queue. The way to classify packets could be flexible for distinct requirements.

There are several properties for a queue: Q, w and ts. Q value is the data length LLEPS will reserve for the session (queue) in a time slice. The value of w is the number of bytes that the queue has been served in the time slice. The value of ts is the time stamp of a queue and it denotes the time that usage status of the queue has been updated. These properties can be utilized to reflect the usage status of the queue.

LLEPS uses the value of Q and the value of w to determine the priority of the queue and uses ts to determine that the usage status of the queue should be updated or not.

2) Min Heap:

Min Heap is used to store the indexes of queues. It makes the scheduler able to pick the queue with the highest priority quickly.

3) System Timer:

System Timer is used to update the time stamp of the scheduler. After the scheduler picks a queue, it learns that the usage status of the queue should be updated accordingly as the time stamp of the queue is older than the time stamp of the scheduler.

System Timer updates the time stamp of the scheduler periodically. The period is defined as tick.

4) Scheduler:

The scheduler picks the queue with the highest priority to serve. After forwarding a packet from the queue, the scheduler will update the usage status w of the queue and put it into the Min Heap. The scheduler will update the ts value of the queue if the time stamp of the queue is smaller than the time stamp of the scheduler before putting the queue into the Min Heap.

There is a trick between ts and the Min Heap. While comparing the priority values of two queues, the queue with a smaller ts value will get a higher priority. This ensures that the scheduler can pick the queue which needs to be updated the w and ts values without any packet forwarding.

III. Simulation Experiments

A. Simulation Environment

The simulations are implemented using Network Simulator 2. The topology used in the simulation is presented by Figure 23. The circle with a number is a node and it denotes a sender or a receiver. The number n on it denotes that it's nth pair of a sender and a receiver. The circle with *ER* or *DR* is a router. The bottleneck of the simulations is the link between the two routers in Figure 4. The bandwidth of the links between nodes and routers is 100Mbps and the propagation delay is 1 millisecond. The link between two routers is 1Mbps and 15 milliseconds. LLEPS and Nested-Deficit Round Robin are applied on the simplex link from *ER* to *DR*.

B. Bandwidth Sharing

The bandwidth sharing experiment is to show the ability of LLEPS to provide bandwidth sharing. In the bandwidth sharing experiment, the same scenario is applied for LLEPS and NDRR. Three UDP flows are created: 500Kbps for UDP 1, 300Kbps for UDP 2 and 200Kbps for UDP 3. Each UDP flow runs with 10Mbps. UDP 1 starts at 0th second, stops at 15th second and starts at 25th second again. UDP 2 starts at 5th second. UDP 3 starts at 10th second. In NDRR, Q_min is set to 1500 bytes. The packet size of the test is set to 500 bytes.

The receiving rate of the three UDP flows should be 5:3:2 at any time in the idea case. Figure 5 shows the result of LLEPS and Figure 6 shows the result of NDRR. The ratio of

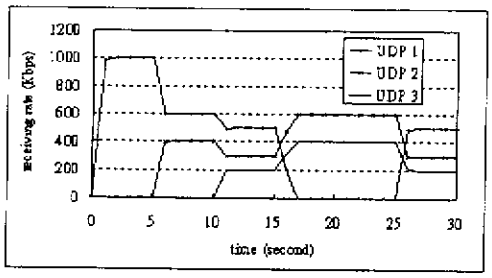


Fig. 5. Bandwidth Sharing Test/LLEPS

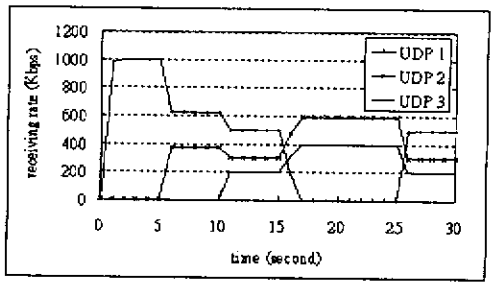


Fig. 6. Bandwidth Sharing Test/NDRR

reserved bandwidth for the three UDP flows is 5:3:2. Figure 6 demonstrates that NDRR always keeps the sharing ratio of all flows the same as the ratio of reserved bandwidth. However, LLEPS does not always keep the sharing ratio but it ensures that all flows can get reserved bandwidth.

In Figure 5 from 5th second to 10th second, only UDP flow 1 and UDP flow 2 are running. The sharing ratio between the two UDP flows should be 5:3. But the real sharing ratio in LLEPS is 3:2. In the experiment, after LLEPS updating the queue for UDP flow 1 and UDP flow 2, the value w is always equal to the value Q because the two UDP flows using more bandwidth than the reserved bandwidth all the time. The updating of all queues happens at the same time and UDP flow 2 may sent one or two more packets before the updating. The unfairness is never memorized by LLEPS. That's the reason why the sharing ratio between the two flows is not 5:3.

C. Bandwidth Reservation

The goal of the experiment is to test the ability of LLEPS to provide bandwidth guarantee. LLEPS allocates 200Kbps for UDP flow 1 and 300Kbps for UDP flow 2. The bit rate of UDP flow 1 and UDP flow 2 are 200Kbps and 300Kbps. UDP flow 1 starts at 5th second and UDP flow 2 starts at 10th second. There are four background TCP flows competing with the two UDP flows.

Figure 7 presents the result of the experiment. Before the starting of the two UDP flows, four TCPs take all the bandwidth of the bottleneck. After the first UDP flow starts, it gets the bandwidth reserved to it. The second UDP flow also gets the reserved bandwidth. Since LLEPS allocates bandwidth

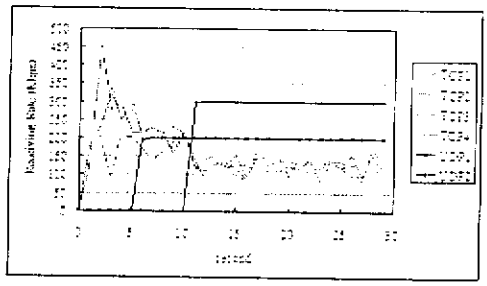


Fig. 7. Bandwidth Reservation Test of LLEPS

for the two UDP flows, the bandwidth for the two flows will not be used by the four TCPs. The four TCPs share the left bandwidth 500Kbps. The result shows that LLEPS is able to provide bandwidth guarantee.

D. Buffer Under Run Problem

As described, NDRR and other frame-based packet scheduling algorithms do not perform well when some queues are not always active. The delay of packets is not stable.

In the simulation, ten flows are running and both LLEPS and NDRR allocate 100Kbps for each of the ten flows. The bit rate of the first nine UDP flows is 1Mbps and the bit rate of the tenth UDP flow is only 80Kbps. The queue belonging to the tenth UDP flow will be empty at most of time and other queues will be always active. The first 9 UDP flows are background traffic and keep the bottleneck link busy.

The 10th UDP flow is used to simulate a streaming application. Most streaming applications like VoD and VoIP do not have special control mechanism to control the behavior (bit rate or something else) of the stream. A CBR (Constant Bit Rate) MP3 (MPEG Audio Layer 3) broadcasting is just like a CBR UDP flow. This is why an UDP flow is able to simulate a streaming application here.

This paper uses *Delay Jitter* to represent the stability of latency of packets. The definition of *Delay Jitter* is the same as RTP. R_i is the receiving time of packet i. S_i is the sending time of packet i. $D_{i-1,i}$ denotes the delay variation between two consecutive packets. Ji is the smoothed function of $D_{i-1,i}$.

$$D_{i-1,i} = (R_i - R_{i-1}) - (S_i - S_{i-1}) \tag{7}$$

$$J_i = 15/16 \times J_{i-1} + 1/16 \times |D_{i-1,i}| \tag{8}$$

In the experiment, only the J_i of the 10th UDP flow is plotted. The packet size is set to 100 bytes, 500 bytes and 1400 bytes. Qmin in NDRR is set to 1500 bytes. In LLEPS, *tick* is set to 250 milliseconds.

Figure 8, Figure 9 and Figure 10 are the results of the simulations. The solid line is the result of LLEPS and the dashed line is the result of NDRR. In these simulations, no packet of the 10th UDP flow is dropped. In Figure 8, the transmission time of a packet on the bottleneck link is 0.8 milliseconds. In Figure 9, it's 4 milliseconds. In Figure 10,

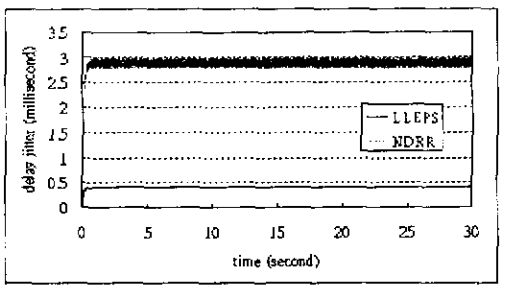


Fig. 8. Delay Jitter with 100 bytes packet size

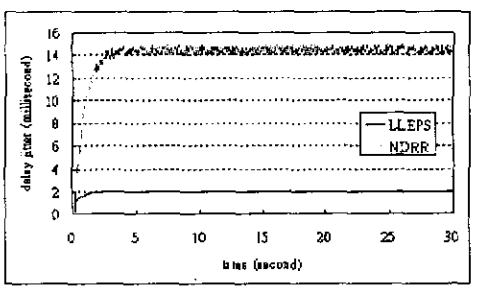


Fig. 9. Delay Jitter with 500 bytes packet size

it's 11.2 milliseconds. If the transmission time is longer, the delay jitter will also become longer. In LLEPS, the delay jitter is always approximate to 50 percent of the transmission time of a packet. The experiment shows that LLEPS can handle the problem much better than NDRR.

IV. Conclusion And Future Work

This paper introduces a new scheduling protocol (LLEPS) to provide bandwidth guarantee service efficiently. In addition to ensuring bandwidth guarantee services, LLEPS does not introduce much delay jitter effects for packets. NDRR does not address *"Buffer Under Run Problem"* but LLEPS does. Since LLEPS always serves the queue with the highest priority, it provides preemption mechanism for packets. However, in NDRR, the service sequence of each queue is constant. The continuous work of LLEPS is to design a better way to determine the priority of each session. LLEPS determines the priority of each session based on the corresponding transmission rate. LLEPS estimates the transmission rate of a session in a time interval. When the time interval is too small, the estimation will become very unstable and no meaning for LLEPS. The continuous efforts will address the issue.

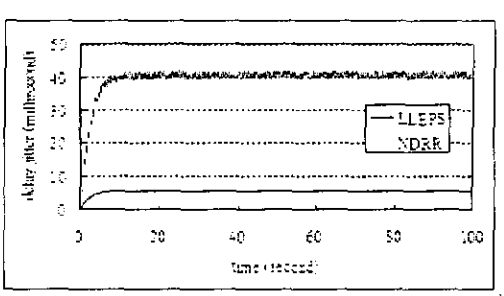


Fig. 10. Delay Jitter with 1400 bytes packet size

References

[1] S. Blake, D. Black, M.Carlson, E. Davies, Z. Wang and W. Weiss, "An architecture for differentiated services", RFC 2475, December 1998.

[2] K. Nichols, V. Jacobson and L. Zhang, "Two-bit differentiated services architecture for the Internet", IETF RFC 2638, July 1999.

[3] R. Braden and D. Clark, "Integrated Services in the Internet Architecture: An Overview", RFC 1633, July 1994.

[4] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation protocol (RSVP) - Version 1 Functional Specification", RFC 2205, September 1997.

[5] S. Floyd and V. Jacobson, "Link-Sharing and resource management models for packet networks", IEEE/ACM Transactions on Networking, August 1995.

[6] A. Demers, S.Keshav and S.Shenker, "Design and Analysis of a fair queueing algorithm", Proceeding of ACM SIGCOMM, September 1989. algorithms", Proc. SIGCOMM'96, August 1996.

[7] Jon C.R. Bennett and Hui Zhang, "WF2Q: Worst-case Fair Weighted Fair Queueing", INFOCOM '96, Proceedings IEEE, Mar 1996

[8] M. Shreedhar and George Varghese, "Efficient fair queueing Using deficit round-robin", IEEE Transactions on Networking, Jane 1996.

[9] Salil S. Kanhere and Harish Sethu, "Fair, Efficient and Low-Latency Packet Scheduling using Nested Deficit Round Robin", Proceedings of the IEEE Workshop on High-Performance Switching and Routing (HSPR), May 2001

[10] H. Zhang and D. Ferrari, "Rate-Controlled static-priority queueing", Proc. IEEE INFOCOM '93, September 1993

[11] S. Iatrou and I. Starvrakakis, "A Dynamic Regulation and Scheduling Scheme for Real-Time Traffic management", IEEE/ACM Transactions on Networking, February 2000.

TMRC: A Load-Adaptive TCP-Friendly Rate Control Protocol For Real-time Multimedia Environment

Eric Hsiao-kuang Wu*, Chien-Shao Wesley Tseng*, Chung-Yuan Knight Chang*, Meng-feng Tsai*
Chunhung Richard Lin†

*Department of Computer Science and Information Engineering
National Central University, Chung-Li, Taiwan, R.O.C.
†Department of Computer Science and Information Engineering
National Sun Yat-Sen University, Kaosiung, Taiwan, R.O.C.
hsiao@csie.ncu.edu.tw

Abstract—**In this paper, we introduce and analyze a new congestion control approach for multimedia streaming, called TMRC (Time-based Model TCP-Friendly Rate Control). TMRC is a time-based end-to-end congestion control mechanism, which uses a model to estimate sending rate and assures the fairness against competing flows. With our mechanism, the sender can explicitly adjust its sending rate as a function of the measured rate reported by the correspond receiver. In long term, TMRC acts as a TCP-friendly flow with the same throughput as other flows while in short term it performs more stable in its sending rate. This makes our mechanism suitable for real-time multimedia streaming. At last, we compare TMRC to other TCP-friendly rate control mechanism, such as TFRC.**

I. Introduction

Real-time multimedia streaming transmission is now world-wide used over the Internet and has become a striking application, such as Internet television, video conference, IP telephony, distance learning, video-on-demand. Compared to traditional applications, real-time multimedia applications demand that the sufficient quality of service (QoS) (e.g. a bounded delay, jitter, or guaranteed bandwidth) should be provided. However, the current Internet environment provides a best-effort service in general, that is data transmissions on such a non-QoS guaranteed environment may suffer from bandwidth variation, end-to-end delay variation, and packet loss. Therefore, an adaptive, best-effort, end-to-end congestion control protocol is required to react with congestion to avoid congestion collapse and keep network utilization as high as possible.

This paper presents the design and evaluation of the TMRC protocol extended from [1] in which a new TCP-friendly rate estimation approach is proposed.

The rest of this paper is organized as follows. We review some of the related work in section II. In section III, we introduce the TMRC protocol. Detail description of our simulation results are presented in section IV and the comparisons with other TCP-friendly rate control protocols are shown in section V. Finally, section VI concludes the paper and discusses some of our future work.

II. Related Works

Existing congestion control protocols can be roughly categorited into two types: *window-based* and *rate-based*. The window-based approaches perform additive increase and multiplicative decrease (AIMD) rate control at a sender as in TCP. Typically, they probe for the availabel network bandwidth by slowly increasing a congestion window, and reduce the congestion window greatly when congestion is detected (indicated by one or more packets loss) [2], [3]. The rapid rate reduction in response to congestion is essential to avoid network callapse. Because of AIMD property, the window-based protocols are provably stable and fair under steady state.

On the other hand, the rate-based approaches estimate the throughput of a TCP sender as a function (or formula) of packet loss rates and round-trip-time (RTT). Since the window-based protocols typically involving with retransmission which can introduce intolerable delays, the rate-based protocols are usually employed for transporting real-time multimedia streaming data. However, these protocols may not converge to the fair bandwidth share of network paths (owing to either over-allocation or under-allocation of bandwidth) [4], [5], [6], [7], [8], [9]. This happens because of inaccuracy in estimating loss rate and in the formula itself. Furthermore, the TCP formula may be not reliable when packet losses or RTT is correlated to the transmission rate of the flow being controlled.

As for rate-based congestion control, it is significant to investigate a more accurate formula for the TCP sending rate to make flows perform *TCP-compatible* or *TCP-friendly*, use no more bandwidth than conformant TCP running under comparable conditions in steady-state, [5]. [7] observed and

introduced the TCP response function as following:

$$T = \frac{s}{R\sqrt{\frac{2p}{3}} + t_{RTO}(3\sqrt{\frac{3p}{8}})p(1+32p^2)} \quad (1)$$

This gives an upper bound on the sending rate T (bytes/sec), as a function of the packet size s (bytes), round-trip time R (sec), steady-state loss event rate p, and the TCP retransmit timeout value t_{RTO} (sec).

TFRC [4] is proposed to be a *TCP-friendly* rate-based end-to-end congestion control protocol for unicast traffic. The TFRC sender explicitly adjusts its sending rate as a function of the measured rate of loss events by receiver and ensures fair behavior against competing flows. Instead of reacting to single congestion event (i.e. packet loss), TFRC changes its sending rate in response to the loss rate, sampled over a single round-trip time.

Rhee et al. [6] developed a new flow control approach for multimedia streaming, called *TCP emulation at receiver (TEAR)*, which shifts most of flow control mechanism to receivers. In TEAR, a receiver does not send the congestion signals (such as packet arrivals, packet losses, and timeout) to the sender but rather processes them immediately to determinate its own appropriate receiving rate.

Both TFRC and TEAR flows provide smoother transmission rate than TCP, however these protocols may result in lacking the aggressiveness and responsiveness of TCP [8], [9].

III. TMRC Protocol

Chen et al. [1] proposed a new approach using time-based model for TCP-friendly rate estimation:

$$T = s \times \frac{\left(\frac{P}{R} - 1\right) \times \frac{3}{2}}{R} \quad (2)$$

Instead of collecting packet loss rate, in the time-based model average TCP sending rate T (bytes/sec) is measured as a function of the packet size s (bytes), round-trip time R (sec), and the *congestion event period* P (sec), time interval of two adjacent loss/congestion events.

Simulation results show that this new approach provides rate estimation more accurate than TFRC does. With the successful rate estimate approach, we design a new rate control scheme, named *time-based model rate control (TMRC) protocol*. The primary goal of time-base model rate control is to provide a TCP-friendly congestion control mechanism to maintains a relatively steady sending rate for media streaming flows.

A. Protocol Model

TMRC protocol adpots the receiver-driven model, shifts most functionarities to the receiver end. While the sender concentrates itself on delivering data in a specified bitrat, the receiver keeps track of essential information along with network changes, The information is then applied to equation (2) to calculate a up-to-date sending rate. Necessarily or periodically, the receiver reports the sending rate adjustment to the sender.

```
Period := Now - PreviousEventTime;

if (loss/congestion event detected) {
    if (Period > SustainedTime) {
        PreviousEventTime = Now;
        Add Period to PeriodHistories;
        EventPeriod :=
            Average(PeriodHistories);
    }
}
else {
    NewEventPeriod :=
        Average(Period, EventPeriod);
    if (NewEventPeriod > EventPeriod)
        EventPeriod = NewEventPeriod;
}
```

Fig. 1. Algorithm for calculating the event period when a packet arrived

B. Sender Functionality

The sender plays quite a simple role in TMRC protocol. Every time a rate adjustment report is received, the sender reacts with the report to adjust its sending rate. In additional, the sender cooperates with the receiver to calculate the round-trip time R (described later).

C. Receiver Functionality

In order to apply equation (2) as the control function for congestion control, several parameters, the round-trip time R and the congestion period P, are required by the receiver. (The packet size s used by the rate estimate equation should be either provided by application or probed by other protocol, such as path MTU discovery [10].)

1) Round-Trip Time Estimation: The receiver and sender collaborate to measure the round-trip time R. Periodically the receiver sends the sender a echo request consists of the sending time, and then the sender feeds a corresponding echo response along with the timestamp back to the receiver. Once a feedback is received, the receiver can determinate the round-trip time.

2) Event Period Estimation: To calculate the congestion event period is one of the most critical part of TMRC. For the purpost of illustration, we briefly describe the algorithm in Fig. 1. The estmated event period should increase only in response to a new event period that is longer than perviously-calculated average, or be updated while packet arriving with a loss/congestion event. There is a clear trade-off between updating the event period over a short period of time to respond rapidly to degradation in the available bandwidth, and updating over a longer period of time to get a smoother adjustment that is much less noisy. Let the *sustained time* be defined as the quiescent period of time to prevent the event period from updating excessively. In general, the sustained time is given with twice RTT.

The method of calculating the average of event period has been the subject of much discussion and testing. Obvious methods we looked at include the EWMA method and the LIWA method which is the method we chose. The EWMA

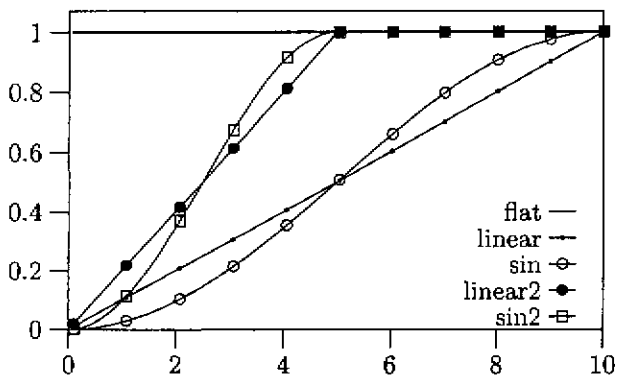


Fig. 2. LIWA weight curve

method uses an exponentially weighted moving average of the event period. However, it is difficult to choose an sufficient EWMA weight that responds to the actually network changes due to the EWMA method keeps memory of untimely event period. The LIWA method computes a weighted average of the event period over the last n periods with the weight curve. Fig. (2) is an example for $n = 10$. In our experiments, we use *linear*2 curve which gives equal weights to the most recent periods and fades the past period.

3) Rate Estimation: The receiver calculates the event period when a packet arrived. if necessary (i.e. a loss/congestion event is detected), a rate adjustment report is send to the sender. In order to aggressively find and use the available bandwidth, the receiver periodcally send a rate adjustment report to the sender as well.

4) Slow Start: For lack of the congestion window used by window-based protocol, TMRC needs a initiate mechanism similar to the window-based slow-start procedure. Roughly, the receiver reports the sender with a base sending rate T_0, and doubles the rate every each round-trip time, while the sender starts with a low sending rate (i.e. 16 Kbps), and adjusts its sending rate as the receiver reported. When a loss/congestion event is detected, slow-start procedure goes terminated.

$$\begin{cases} T_0 = 2 \times \frac{\text{packet size}}{\text{round-trip time}} \\ T_{t+1} \leftarrow 2 \times T_t \end{cases} \quad (3)$$

As Fig. 4 and Fig. 5 shown, TMRC slow-start behavior is more similar to TCP.

D. Improvement

In some cases, such as only few flows share the same link, the delay in reporing rate adjustment information to the sender may cause sending rate oscillations. Owing to the adjustment information does not take effect immediately to the network, the receiver may under-allocate or over-allocate sending rate. Moreover, one oscillatory flow may effect to competing flows when the shared link is loaded. In order to avert from rate oscillations, TMRC applies *anti-overload* mechanism in rate estimation. An anti-overload value P_{ao} in time t is calculated by the following expression:

$$P_{ao}(t) = \sqrt{1 - (Lr(t) + Jr(t))} \quad (4)$$

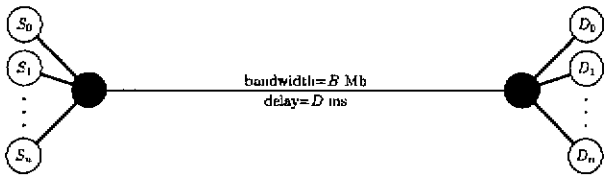


Fig. 3. Simulation topology (delays of links for which no delay is specified are all equal to 5 ms and bandwidth of links for which no bandwidth is specified are all equal to 100 Mb).

TABLE I
SIMULATION CONFIGURATION

Description	Value
Packet Size	1000 bytes
TCP maximum windows size	65535 packets
TCP version	New Reno
TCP timer granularity	0.2 seconds
New Reno TCP parameters	newreno_changes on newreno_changes1 on
Queue type	Droptail

Lr is the packet loss ratio in time scale from $t - UT$ to t:

$$Lr(t) = \frac{N_{loss}}{N_{loss} + N_{arrival}} \quad (5)$$

and Jr is the jitter ratio correspondingly:

$$Jr(t) = \frac{(t_R(t) - t_R(t - UT)) - (t_S(t) - t_S(t - UT))}{t_R(t) - t_R(t - UT)} \quad (6)$$

Generally, loss ratio is equal to zero if no packet dropped, and jitter ratio is equal to 1 while the shared link sitting on steady state. Once the bottleneck link suffers from a congestion, the estimated loss ratio and jitter ratio get increasing. On the contrary, when the bottleneck link suffers through a congestion, both of loss ration and jitter ratio decrease. With these atributes, the anti-overload value may indicate network variations. Therefore, we modified the *period* by $period \times P_{ao}$ in the algorithm shown as Fig. 1. Obviously, to decide the value of the unit time UT is a significant factor.

IV. Simulation And Results

To assess the performance of TMRC, we use the *ns* simulator to study the TCP-friendliness, TCP-compatibility, aggressiveness, and smoothness of our proposed algorithm. Each experiment is run with the following parameters: the bottleneck bandwidth, the number of competing TCP flows, and the number of competing TMRC (or TFRC) flows. The bottleneck queue configuration and other simulation parameters are listed in Table I, and simulation topology is shown as Fig. 3. The simulation duration is set to 200 seconds, and each network flow is started at zero.

We hightlight some subset of the results to illustrate the performance comparison between TMRC and TFRC.

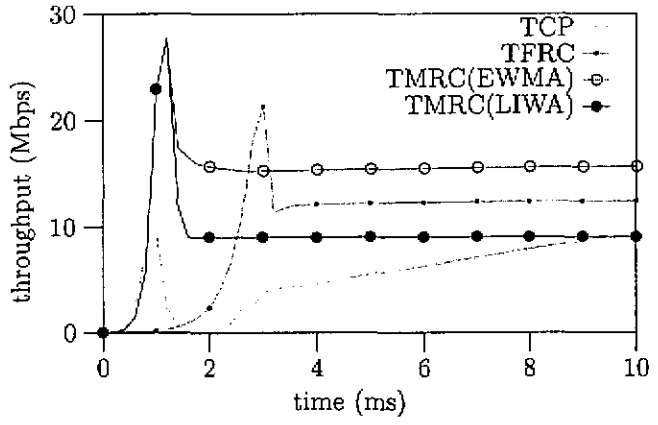


Fig. 4. Slow-start behavior

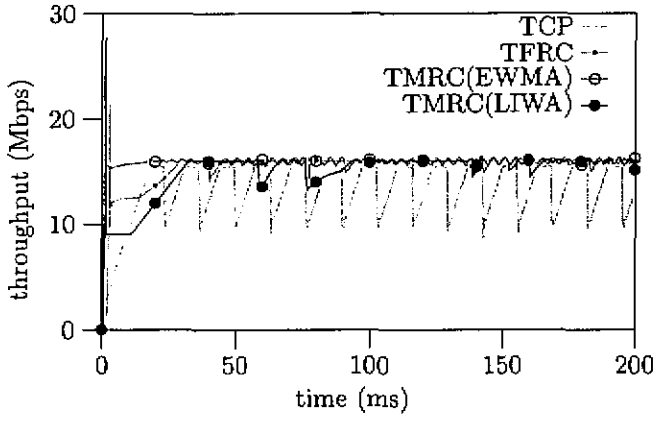


Fig. 5. Overall performance

1) TCP-friendliness: We conduct the following experiment to test TCP-compatibility of our TMRC protocol. A single flow under investigation in travelling a flat link with bottleneck bandwidth $B = 16$ Mbps, and link delay $D = 40$ ms ($RTT =$ 100 ms). Fig. 4 and Fig. 5 shows the slow-start behaviors of TCP, TFRC, and TMRC (with EWMA and LIWA). Obviously in Fig. 4, TMRC slow-start procedure is much similar to TCP. In particular, TMRC with EWMA method acquires the available bandwidth instantly. However, this is not suitable for all case, EWMA method makes TMRC flows smoother but less aggressive while competing with other flows. Both TFRC and TMRC performs TCP-friendly in long term throughput.

2) Intra-Protocol Fairness: Fig. 6 shows the case that flows with single type protocol compete to each other. We observed that TMRC flows vibrate a little in the steady state while TFRC flows keeping more flat. This happens becuase of that in our experiment we use the *simple model*[1] for event detecting. With the simple model, a loss event may be ignored and then TMRC estimate a relatively higher sending rate. This condition will be solved when the *complex model*[2] is involved our proposed TMRC protocol.

3) TCP-compatibility: This section we exam the TCP-compatibility of our TMRC protocol. Fig. 7 shows 4 TMRC flows and 4 TCP flows compete for bandwidth over a shared bottleneck link. Fig. 8 shows the other case with 1 TMRC flows and 7 TCP flows. In each case, results are shown for a bottleneck link bandwidth of 16 Mbps and round-trip time of 100 ms. As can be observed form the results, TMRC achives a slightly lower average throughtput than TCP when TFRC gaining more. Moreover, TMRC flows performs smoother and more steady than TFRc does. In additional, TMRC flows converge on the fair share bandwidth much closely to each other.

V. Conclusion

We proposed a new rate-based congestion control algorithm called TMRC (Time-based Model Rate Control), which uses time-based model TCP rate estimation approach. Instead of gathering packet loss rate, TMRC calculates loss/congestion event period. We have shown that TMRC performs TCP-friendly as well as TCP-compatible. We have also shown that TMRC is smoother and more steady while competing with other flows.

Our future focus will be to investigate the complex model that uses congestion events instead of loss events for event period measurement, and to implement TMRC algorithm under the real network.

References

[1] Shi-Yang Chen, Eric Hsiao-kuang Wu, and Yi-Shuang Chen, "A New Approach Using Time-Based Model for TCP-Friendly Rate Estimation", Proceedings of IEEE ICC2003.

[2] Y. Richard Yang and Simon S. Lam, "General AIMD Congestion Control", Technical Report TR-2000-09, May 9, 2000; Proceedings ICNP 2000, Osaka, Japan, November 2000.

[3] Yang Richard Yang, Min Sik Kim, and Simon S. Lam, "Two Problems of TCP AIMD Congestion Control", TR2000-13, June, 2000.

[4] Sally Floyd, Mark Handley, Jitendra Padhye, and Joerg Widmer, "Equation-Based Congestion Control for Unicast Application", February 2000.

[5] Sally Floyd and Kevin Fall, "Promoting the Use of End-to-End Congestion Control in the Internet", IEEE/ACM Transactions on Networking, August 1999.

[6] Injong Rhee, Volkan Ozdemir, and Yung Yi, "TEAR: TCP Emulation at Receiver — Flow Control for Multimedia Streaming", NCSU Technical Report, April 2000.

[7] J. Padhye, V. Firoio, D. Towsley, and J. Kurose, "Modelling TCP Throughput: A Simple Model and Its Empirical Validation", Proceedings of SIGCOMM 98.

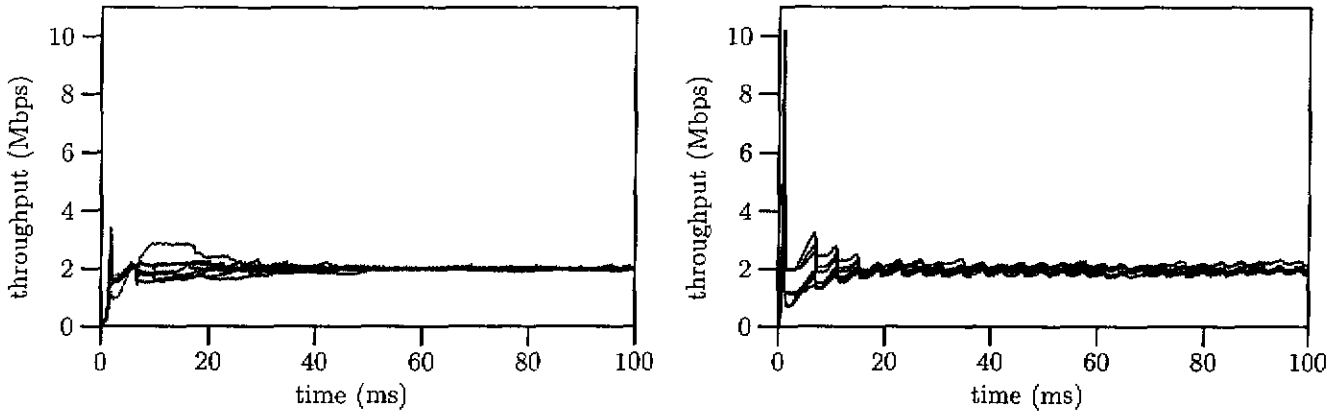
[8] Sally Floyd, Mark Handley, and Jitendra Padhye, "A Comparison of Equation-based and AIMD Congestion Control", May 2000.

[9] Y. Richard Yang, Min S. Kim, and Simon S. Lam, "Transient Behaviors of TCP-Friendly Congestion Control Protocols", Technical Report TR-2000-14, July 2000

[10] J. Mogul and S. Deering, "Path MTU Discovery", RFC 1191, Draft Standard, November 1990.

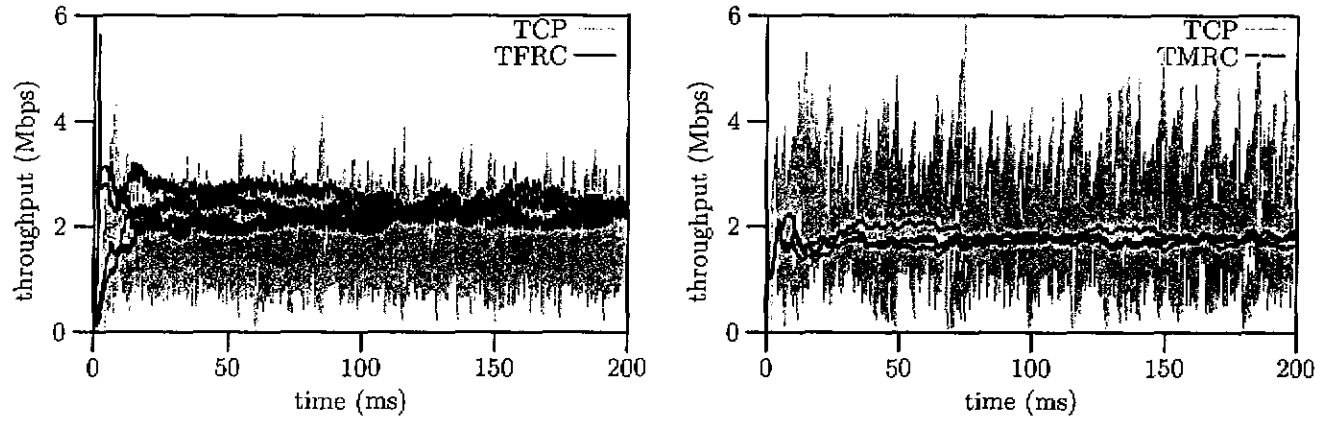
[1] A simple model of TMRC determinates events by packet loss.

[2] In complex model, we detect a event by congestion events depends on the cooperation of jitter ratio and loss ratio described in section III-D. The complex model is not be well-defined nowadays, so that this will be a future work for our study.



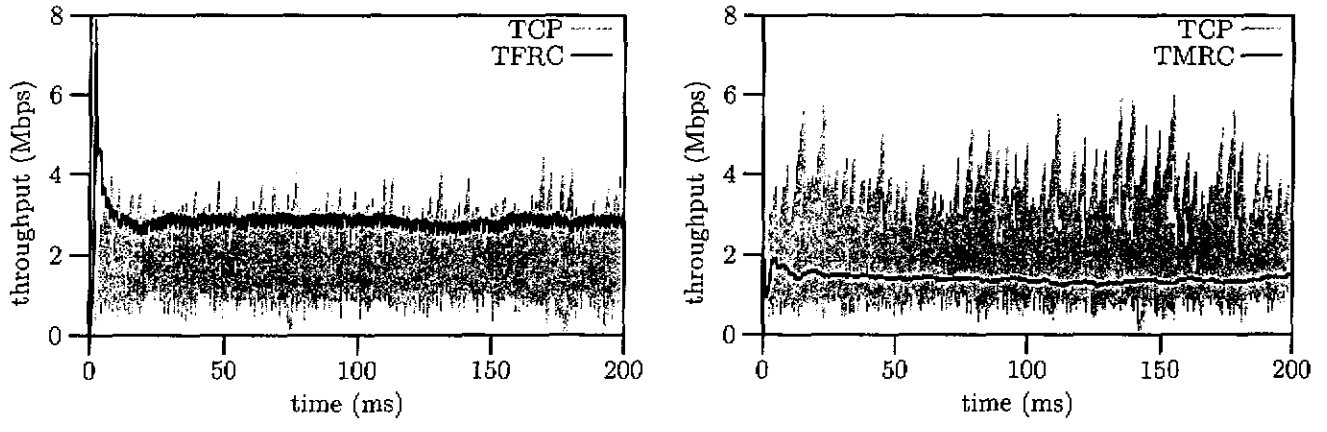
(a) 8 TFRCs (b) 8 TMRCs

Fig. 6. Bottleneck: 16 Mbps, RTT=100ms



(a) 4 TCPs vs 4 TFRCs (b) 4 TCPs vs 4 TMRCs

Fig. 7. Bottleneck: 16 Mbps, RTT=100ms



(a) 7 TCPs vs 1 TFRC (b) 7 TCPs vs 1 TMRC

Fig. 8. Bottleneck: 16 Mbps, RTT=100ms

Noncooperative Admission Control for Differentiated Services in IEEE 802.11 WLANs

Yu-Liang Kuo*, Eric Hsiao-Kuang Wu†, and Gen-Huey Chen*
* Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.
†Dept. of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan, R.O.C.

Abstract— **Recent developments in wireless LAN (WLAN) standardizations have been successful to offer high-speed data services. Hence, a variety of types of traffic must be accommodated in future WLAN environments. IEEE 802.11 working group has been developing a new distributed MAC, called enhanced distributed function (EDCF), to support service differentiation in the IEEE 802.11 MAC protocol. Besides, high-speed WLAN environments are also expected to provide wireless Internet services in hot spots such as airports, parks, and etc. Since there are usually multiple service providers competing for providing wireless network access in hot spots, mobile users are free to choose their own service providers. For a specific service provider, more flows are admitted to transmit in its coverage range, more revenue (e.g., utility or profit) is gained. However, admitting many flows may make the wireless medium overloaded and degrade the QoS satisfactions of ongoing flows since the wireless medium in IEEE 802.11 is share and collision-based. Hence some mobile users may leave the current service provider and subscribe to wireless network access with another service provider. In this paper, we analyze the resource management problem in the competitive environment, in which EDCF protocol is implemented in all access points (APs) and mobile stations. We formulate admission control as a game and prove the game owns a Nash equilibrium solution. Based on the game, a service provider not only fulfills most part QoS satisfactions of ongoing flows but also increases its own revenue. We evaluate the performance by means of throughput, packet delay and bandwidth violation ratio.**

I. Introduction

With the provisioning of high-speed WLAN environments, data services (e.g., VoIP or video-conference) with different QoS requirements will be available in future WLANs. Hence, data services will be categorized into multiple traffic classes and different priorities are used to access the wireless medium. However, in the current access mechanism of IEEE 802.11, all mobile stations have the same priority to access the wireless medium. In order to satisfy multiple traffic classes with different QoS requirements, it is desired to provide service differentiation in the IEEE 802.11 standard [5].

Providing service differentiation requires a MAC protocol to support different access priorities among different traffic classes. Recently, the IEEE 802.11 working group has been developing a new distributed protocol, called EDCF, to support service differentiation in the MAC layer [3]. EDCF is an extension of the existing distributed coordination function (DCF) to support service differentiation. More details about DCF and EDCF will be presented in Section II.

Apart from the MAC service differentiation, high-speed WLAN environments are also expected to provide public wireless Internet services in "hot spots" such as airports, parks and etc., where a high density of mobile users will create demand for wireless network access [10]. Hence multiple wireless service providers dispose their infrastructures, e.g., APs, in hot spots to provide wireless network access. Mobile users that intend to access wireless Internet services are charged by prepaid cards, online purchasing, or other charging mechanisms. Since there are usually multiple service providers in a local region, mobile users are free to choose their own service providers. Admitting more flows to transmit in the coverage range of an AP (belonging to a certain service provider) will gain more revenue for the service provider. However, admitting many flows may make the wireless medium overloaded and degrade the QoS satisfactions of ongoing flows since the wireless medium in IEEE 802.11 is share and collision-based. Hence some mobile users may leave the current service provider and subscribe to wireless network access with another service provider.

In this paper, we analyze the resource management problem in the competitive environments, in which EDCF is implemented in access points (APs) and mobile stations. We formulate admission control as a game, which is nonzero-sum and noncooperative. Based on the game, a service provider not only fulfills most part of QoS satisfactions of ongoing flows in its radio coverage but also increases its own revenue.

The rest of this paper is organized as follows. Section II reviews the DCF and EDCF protocols. Section III suggests an admission control game formulation under which it owns a Nash equilibrium solution. Section IV shows the performance evaluation for the admission control. Section V concludes this paper.

II. DCF and EDCF

DCF operates based on carrier sense multiple access with collision avoidance (CSMA/CA). A mobile station that intends to transmit a packet first senses the channel. If the channel is idle for a time period of DCF interframe space (DIFS), it can immediately start transmission. Otherwise, it generates a backoff counter. The counter starts decrement if the channel is sensed idle for a time period of DIFS. Then the counter continues to decrease until the channel is busy or the counter counts down to zero. If the channel is busy, the decrement will pause and resume after another idle time period of DIFS. When the counter counts down to zero, the mobile

0-7803-8794-5/04/$20.00

TABLE I
FOUR ACS SPECIFIED IN THE IEEE 802.11E

	AC_0	AC_1	AC_2	AC_3
Values of AIFSN	2	1	1	1
Values of CW_{min}	32	32	16	8
Values of CW_{max}	1024	1024	32	16

station starts transmission. In order to avoid channel capture, a mobile station has to wait a random backoff time between two consecutive packet transmissions, even if the channel is idle for a time period of DIFS.

The backoff counter is randomly assigned a value from the range [0, CW−1], where CW is the contention window. Initially, let CW=CW_{min}, the minimum contention window. When the transmission (or retransmission) fails, the value of CW is doubled until it reaches the maximum $CW_{max}=2^m CW_{min}$, where m is called *maximum backoff stage*.

DCF employs two access mechanisms for packet transmission. One is two-way handshaking and the other is four-way handshaking. For the former, an ACK (acknowledgement) message is used to indicate that the transmitted packet has been correctly received by the destination station. For the later, an RTS (request-to-send) message is first sent by the source station. When the destination station receives the RTS, it replies a CTS (clear-to-send) message. After receiving the CTS message, the source station is allowed to transmit a packet. Finally, the destination station informs the source station of a successful transmission by replying an ACK message.

RTS and CTS messages carry information about the identifiers of the source and destination stations and the duration for transmitting the packet. Once hearing the RTS or CTS message, any other station will update its NAV (network allocation vector), which records the duration when the channel is busy, and defer its access to the channel.

In the IEEE 802.11 standard [1], the four-way handshaking is used to reduce the collision time when the size of the transmitted data packet is greater than a predefined length, i.e., *RTSThreshold*. Otherwise, the two-way handshaking is used. The four- way handshaking can also avoid the hidden terminal problem.

EDCF, which is an enhanced version of DCF, can provide a distributed access mechanism to support service differentiation in IEEE 802.11. EDCF introduces the concept of access categories (ACs). Traffic classes belonging to different ACs use different values of CW_{min}, CW_{max}, and arbitration interframe spacing number (AIFSN) to contend the channel. There are four ACs specified in IEEE 802.11e as shown in Table I, where the 802.11b physical layer [2] is used.

EDCF requires that a mobile station has to wait a time period of AIFS before transmitting a packet or generating a backoff counter. Let T_{AIFS} and T_{SIFS} denote the lengths of AIFS and short IFS (SIFS), respectively. T_{AIFS} is computed as follows: $T_{AIFS} = T_{SIFS} + \text{AIFSN} \times \delta$, where AIFSN ≥ 1 and δ is the length of a time slot. A traffic class with smaller AIFSN has smaller T_{AIFS} and hence has a higher probability of seizing the channel.

EDCF is backward compatible with DCF. A mobile station without EDCF has to wait at least a time period of DIFS before it can transmit a packet. Recall that IEEE 802.11 set T_{DIFS}, the length of DIFS, to be $T_{SIFS} + 2\delta$. A mobile station with EDCF has a higher probability of seizing the channel than those without EDCF, if it sets AIFSN=1.

III. ADMISSION CONTROL GAME FORMULATION

In this section, we formulate admission control as a two-player game. The admission control game is formed when a user sends a flow request to an AP. In the following subsections, we first present the system model. Second, we review the basic concepts of two-player game. Third, we present the admission control game formulation. Fourth, an important ratio − QoS satisfaction ratio − used in the game is presented. Finally, we will prove that the game owns a Nash equilibrium solution.

A. System Model

The system environment we consider is within a local region in which multiple APs compete for providing wireless network access. We assume that different APs in the local region are belonging to different service providers. In the competitive environment, mobile users are free to choose an AP for wireless Internet services. A mobile user may leave the current AP and subscribe to wireless network access with another service provider according to the degree of QoS satisfaction for the current quality.

In order to support service differentiation, EDCF protocol is implemented in all APs and mobile stations. Suppose that there are K traffic classes with distinct QoS requirements in the system, where $K \geq 1$. We assume that each mobile user has packets of the same traffic class. For convenience, we refer to a mobile user that has traffic class k packets as a class-k user. We use n_k to denote the number of class-k users currently in the system, where $0 \leq k \leq K-1$. The term r_k is referred to as the data rate for traffic class k.

B. Preliminaries on Two-Player Game

In a two-player game [8], each player has a set of *strategies*. A *configuration* is a form of a pair of strategies, in which one strategy comes from one player and the other strategy comes from the other player. Associated with each configuration, i.e., a pair of two strategies, there are two payoffs, one for each player. A *payoff* is a number that reflects the desirability of a configuration to a player. For a noncooperative game, each player makes a decision independently and in a manner maximizes its own payoff without knowing the decisions of other players.

Specifically, a two-player non-cooperative game has two players p and q. We assume there are m and n strategies for p and q, denoted as $\{u_i : 1 \leq i \leq m\}$ and $\{v_j : 1 \leq j \leq n\}$, respectively. Hence the configurations of the game are denoted by $\{u_i, v_j\}$, where $1 \leq i \leq m$, $1 \leq j \leq n$. The payoffs for p

0-7803-8794-5/04/$20.00 © 2004 IEEE

and q are defined as two payoff matrices $P = [p_{ij}]_{m\times n}$ and $Q = [q_{ij}]_{m\times n}$. The rows of P and Q matrices represent the set of strategies for p while the columns represent the set of strategies for q. In other words, p_{ij} (q_{ij}) represents the payoff for p (q) as p chooses strategy i and q chooses strategy j, where $1 \le i \le m$ and $1 \le j \le n$.

A game can be also classified into a zero-sum and a nonzero-sum game. Specifically, if $p_{ij} + q_{ij} = 0$ for all $1 \le i \le m$ and $1 \le j \le n$, it is a zero-sum game; otherwise, it is a nonzero-sum game. The number of payoff matrices for a zero-sum two-player game is one since the payoff matrix for one player can be easily obtained from the other player. Note that if $p_{ij} + q_{ij}$ are constant, it is also zero-sum.

In a two-player noncooperative nonzero-sum game, a configuration $\{u_i^*, v_j^*\}$ is said to be a Nash equilibrium solution if the values of payoff for p and q under $\{u_i^*, v_j^*\}$ are maximal. In other words, no player can change its strategy to get a better payoff.

It has been shown that a two-player non-cooperative nonzero-sum game may or may not have a Nash equilibrium solution [4]. However, in the proposed admission control game, we will show that it always has at least one Nash equilibrium solution, which will be presented in Section III-E.

C. Game Formulation

The proposed admission control game in the system is non-cooperative in nature. In addition, the game is also nonzero-sum. It is noted that a zero-sum game has the property that the increment for one particular payoff value will cause the decrement of the corresponding payoff value so that the summation can be zero. If a new user sends a flow request to an AP in under-loading conditions, admitting the new flow will not degrade the ongoing flows and will gain revenue for the AP. Also, the admitted user will also feel satisfied and gain its revenue. That is, both the values of payoffs for the AP and the mobile user are increasing. Therefore, the admission control game is nonzero-sum.

The two players for the admission control game are as follows:

- p: the current serving AP (service provider) of the class-k user;
- q: a class-k user that intends to initiate a new flow.

The player p has the set of strategies $\{u_1, u_2\}$, where u_1 is to admit the flow request and u_2 is to reject the flow request. Similarly, the player q has the set of strategies $\{v_1, v_2\}$, where v_1 is to leave the current serving AP and subscribe to another one and v_2 is to stay at the current AP.

We now define the payoff matrices $P = [p_{ij}]_{2\times 2}$ for p,which is given as

$$P = \begin{bmatrix} R + (R_k - L) - L_k & R + (R_k - L) \\ R - L_k & R \end{bmatrix}.$$

The notations used in P are defined as follows.

- R: the sum of revenues for the current ongoing flows;
- R_k: the revenue obtained by admitting a class-k user;
- L_k: the revenue loss obtained when a class-k user leaves;
- B_k: the *bandwidth violation ratio* for a class-k user. B_k is defined as

$$B_k = \begin{cases} 0 & \text{if } \rho_k/n_k - r_k > 0 \\ \frac{r_k - \rho_k/n_k}{r_k} & \text{if } \rho_k/n_k - r_k \le 0 \end{cases},$$

  where ρ_k/n_k is the average available bandwidth for a class-k user. The bandwidth violation ratio is zero whenever the average available bandwidth is greater than the data rate, i.e., $\rho_k/n_k - r_k > 0$. Otherwise, the bandwidth violation ratio is defined as $\frac{r_k - \rho_k/n_k}{r_k}$. The estimation for ρ_k can refer to [6] and is not state here;
- L: potential revenue loss obtained by admitting a class-k user. When an AP admits a new flow to transmit in its coverage range, it may cause potential revenue loss for the AP. It is because that each time a new flow is allowed to transmit in the system, the system may be overloading and then the QoS satisfactions of ongoing flows will be degraded;
- $S(B_k)$: the QoS satisfaction ratio that represents the degree a class-k user satisfies the current quality. The satisfaction ratio is a function of bandwidth violation ratio. It is noted that $0 \le S(B_k) \le 1$. The detailed derivation for $S(B_k)$ will be shown in Section III-D.
- $O(B_k)$: the QoS dissatisfaction ratio that represents the degree a class-k user dissatisfies the current quality. $O(B_k)$ also represents the inclination that a class-k user intends to leave the current AP. It is noted that $O(B_k) = 1 - S(B_k)$.

The element p_{21} represents that the class-k user will leave the current AP and the current AP will reject the flow request. Hence the value of p_{21} is simply the sum of revenues of the current ongoing flows minus the revenue loss obtained by the leaving of the class-k user, i.e., $R - L_k$. The value of p_{22} is simply R since the class-k user does not leave the current AP and may keep trying to request later.

The values of p_{12} should consider the potential loss and hence the value of p_{12} is $R + R_k - L$. Similarly, the value of p_{11} is $R + R_k - L - L_k$. We estimate the potential loss in terms of potential departures of ongoing users. The departures stem from the QoS dissatisfaction for the current quality. For class-k users, the potential revenue loss is $n_k O(B_k) L_k$ and the total potential revenue loss for all users is hence $L = \sum_{i=0}^{K-1} n_i O(B_i) L_i$.

Now let us discuss the payoff matrix $Q = [q_{ij}]_{2\times 2}$. Let U_k, (V_k) be the revenue obtained when the flow request is admitted (rejected). Apparently, $U_k > V_k$. The payoff matrix Q is given as

$$Q = \begin{bmatrix} U_k - Y_k & U_k - wO(B_k) \\ V_k - Y_k & V_k - wO(B_k) \end{bmatrix},$$

where Y_k and w are the revenue loss when the class-k leaves the current AP and the multiplier factor, respectively. Y_k may be the money need to pay for the leaving plus the money need to subscribe to wireless network access with another service provider. Other explanation for Y_k is also possible by considering more realistic market mechanisms.

The element q_{11} (q_{21}) corresponds to the configuration that the AP admits (rejects) the flow request and the class-k user leaves the current AP. Hence the value of payoff for q_{11} (q_{21}) is $U_k - Y_k$ ($V_k - Y_k$). The elements q_{12} and q_{22} should minus an additional term $wO(B_k)$ since the user which chooses to stay at the current AP may need to tolerate the possible dissatisfaction. The term w is a multiplier factor and larger than 1.

Once the payoff matrices for p and q are defined, the next step is to show that if this game owns a Nash equilibrium solution or not, which will show in Section III-E. Before we show the existence of Nash equilibrium solution, we derive two important ratios – QoS satisfaction ratio and QoS dissatisfaction ratio – in the following subsection.

D. QoS Satisfaction Ratio

We characterize the degree of QoS satisfaction for a class-k user as a function of bandwidth violation ratio. That is, the degree of QoS satisfaction for a class-k user depends on the quality of available bandwidth. We approximate the QoS satisfaction ratio, $S(B_k)$, by using the Sigmoid function [9]. The Sigmoid function has been used to characterize users' satisfaction in recent literature [7], [11]. We modify the Sigmoid function to involve user's satisfaction with respect to the bandwidth violation ratio, i.e.,

$$S(B_k) = \frac{1}{1 + e^{-\alpha_k(\beta_k - B_k)}},$$

where α_k and β_k are the steepness and the center of curve. α_k and β_k are used to indicate the *sensitivity of bandwidth violation* and the *toleration of bandwidth violation* for traffic class k. The greater the value of α_k is, the lower sensitivity of QoS degradation for traffic class k is. Fig. 1 shows the plot of QoS satisfaction ratio versus bandwidth violation ratio, where α_k=20, 30, 40 and 50, respectively. In Fig. 1, we know that the steepness of the curve is increasing as α_k is increasing and hence the sensitivity of bandwidth violation is decreasing. Similarly, the higher value of β_k is, the higher toleration for the bandwidth violation is, as depicted in Fig. 2. Once the bandwidth violation ratio is greater than β_k, the degree of users' satisfaction (i.e., QoS satisfaction ratio) will be degraded dramatically.

Once $S(B_k)$ has been defined, the QoS dissatisfaction ratio $O(B_k)$ is simply given as $1 - S(B_k)$.

E. Equilibrium Solutions

Before discussing the existence of Nash equilibrium solution, we first explain the concept of domination.

Definition 1: In a two-player non-cooperative game, a strategy u_i^* dominates another strategy u_j^* of p if it obtains better playoff regardless of what the strategy q chooses, where $1 < i, j < m$.

Theorem 1: The proposed admission control game possesses at least one Nash equilibrium solution.

Proof: We consider the proof in three cases:

- $R_k - L > 0$
  In this case, u_1 dominates u_2 and hence p will always choose strategy u_1. Therefore, q will choose v_1 if $U_k - Y_k > U_k - wO(B_k)$; otherwise q will choose v_2. The resulting configuration may be $\{u_1, v_1\}$ or $\{u_1, v_2\}$.
- $R_k - L = 0$
  It is trivial that four configurations are all possible.
- $R_k - L < 0$
  This reasoning process is similar with $R_k - L > 0$ and the resulting configuration may be $\{u_2, v_1\}$ or $\{u_2, v_2\}$.

■

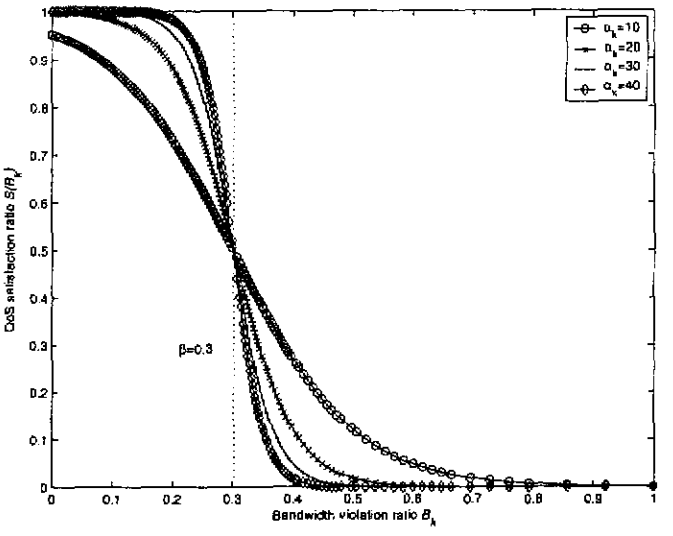


Fig. 1. QoS satisfaction ratio $S(B_k)$ vs. bandwidth violation ratio B_k (α_k=10, 20, 30 ,40; β_k=0.3).

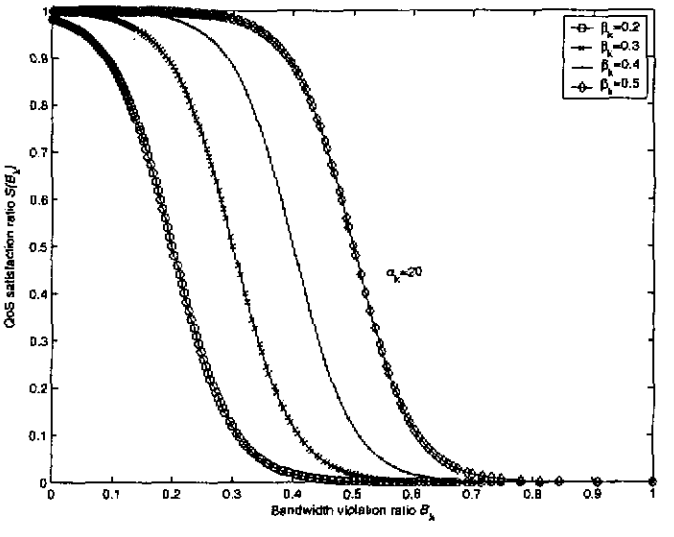


Fig. 2. QoS satisfaction ratio $S(B_k)$ vs. bandwidth violation ratio B_k (β_k=0.2, 0.3, 0.4 ,0.5; α_k=20).

IV. Performance Evaluation

We simulated the MAC layer protocol based on EDCF. The underlying physical layer we adopt is the IEEE 802.11b standard [2] and the channel bit rate is assumed to be 11 Mbps where multi-rate capability is not supported [1]. The values of

TABLE II

PARAMETERS USED IN THE SIMULATION

PHY header	48 μs
Preamble duration	144 μs
Average channel bit rate	11 Mbps
Propagation delay	1 μs
T_{SIFS}	20 μs
Length of time slot	10 μs
MAC header	272 bits
RTS	160 bits
CTS	112 bits
ACK	112 bits
RTSThreshold	1000 byres
R_0, R_1, R_2	20, 10, 6
L_0, L_1, L_2	10, 5, 3
Y_0, Y_1, Y_2	30, 25, 15
U_0, U_1, U_2	5, 5, 5
V_0, V_1, V_2	0, 0, 0
w	5

TABLE III

TRAFFIC CHARACTERISTICS.

	Voice	Video	FTP
Codec algorithm	G.729	H.263	N/A
Average packet inter-arrival time (ms)	20	40	best effort
Payload length (byte)	20	500	1500
Bit rate (Kbps)	8	100	best effort
CW_{min}	8	16	32
CW_{max}	16	32	1024
AIFSN	1	1	2
Minimum required bandwidth (Kbps)	8	100	0
Maximum delay toleration (ms)	5	5	∞

parameters used in the simulation are summarized in Table II. There are three traffic classes in the simulation: voice, video and FTP. We assume that voice, video and FTP are mapping to traffic class 0, traffic class 1 and traffic class 2 respectively. The traffic characteristics for these three traffic classes are shown as Table III. It is noted that a FTP flow has no QoS guarantee, i.e., the minimum required bandwidth is zero and the maximum delay toleration is infinite.

The proposed admission control game was simulated to show the performance under an overloading condition, i.e., the system cannot admit any class-k flow to transmit for $0 \leq k \leq 2$. In order to create an overloading environment, the traffic load is continuously offered in the order – one voice flow, one video flow and three FTP flows – until the system cannot admit any class-k flow for $0 \leq k \leq 2$.

For simplicity and convenience, the sensitivities of bandwidth violations and tolerations of bandwidth violations for voice and video flows are assumed to be equal, i.e., $\alpha_0 = \alpha_1$ and $\beta_0 = \beta_1$, and are denote by α^* and β^* respectively. In order to reflect the high toleration of FTP flows, we let $\alpha_2 = \infty$ and $\beta_2 = 1$.

TABLE IV

NUMBER OF ADMITTED FLOWS ($\alpha^* = 10$).

β^*	n_0	n_1	n_2
0.1	6	2	2
0.2	11	5	6
0.3	10	8	19
0.4	10	9	22
0.5	11	10	22
0.6	12	11	25
0.7	13	12	28
0.8	15	13	31
0.9	16	14	37
1.0	17	15	40

In Table IV, we show the number of admitted flows in the proposed admission control game, in which the sensitivities of bandwidth violation for voice and video flows are set to 10, i.e., $\alpha^* = 10$. In Table IV, we tune the tolerations of bandwidth violations for voice and video flows. It can be seen that as β^* is greater, the number of admitted flows for each traffic class is greater. System administrators can tune the value of β^* by considering the tradeoff between QoS guarantees of ongoing flows and the revenue.

In Fig. 3, we show the average throughput for each traffic class in the proposed admission control scheme under different values of sensitivity of bandwidth violations and toleration of bandwidth violations. In Fig. 3(a), the average throughput for voice flows is higher as α^* is smaller. Similarly, Fig. 5(b) and Fig. 5(c) which show the average throughput for video flows and FTP flows also have the same situation with Fig. 5(a). As seen from Fig. 5(a), Fig. 5(b), and Fig. 5(c), the average throughput for each traffic is decreasing as β^* is increasing. The reason is that the number of admitted flows is increasing as β^* is increasing.

In Fig. 4, we show the average packet delays for different traffic classes under different values of sensitivity of bandwidth violations and toleration of bandwidth violations. The average packet delay for each traffic class is greater as β^* is greater. Also, the average packet delay for each traffic class is greater as α^* is smaller.

Fig. 5 shows the bandwidth violation ratios for voice and video flows. As β^* is greater, the number of admitted flows is greater and hence the bandwidth violation ratio is greater.

V. CONCLUSION

In this paper, we analyzed admission control problem by using game theoretical framework. The system environment we considered was within a local region in which multiple service providers compete for providing wireless Internet services. EDCF protocol was implemented in all APs and mobile stations to provide service differentiation. In the competitive environment, we addressed on the resource management problem that how to increase the revenue for service providers while fulfill most part of QoS satisfactions of ongoing flows.

In the further research, we will consider both real-time and best effort traffic and investigate joint admission and rate

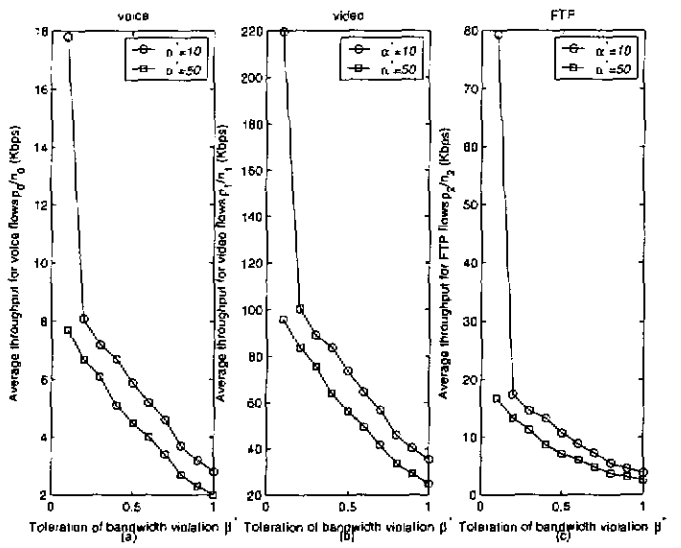


Fig. 3. Average throughput vs. toleration of bandwidth violation. (a) average throughput for voice flows. (b) average throughput for video flows. (c) average throughput for FTP flows.

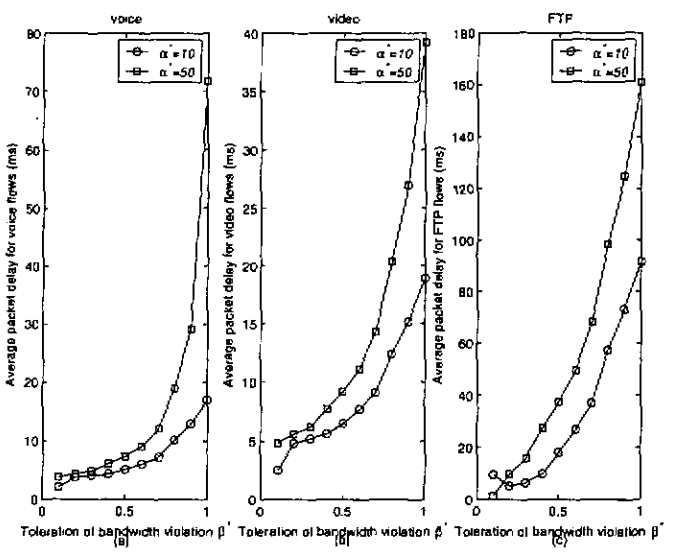


Fig. 4. Average packet delay vs. toleration of bandwidth violation. (a) average packet delay for voice flows. (b) average packet delay for video flows. (c) average packet delay for FTP flows.

controls to not only provide QoS for real-time traffic but also ensure throughput for best effort traffic.

REFERENCES

[1] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard 802.11, 1999.

[2] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-speed Physical Layer Extension in the 2.4GHz Band*, IEEE Standard 802.11b, 1999.

[3] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications:Medium Access Control (MAC) Enhancements for Quality of Service (QoS)*, IEEE Standard 802.11e/D4.1, February 2003.

[4] T. Basar and G. T. Olsder, *Dynamic Noncooperative Game Theory*, 2nd Ed., Society of Industrial and Applied Mathematics, 1999.

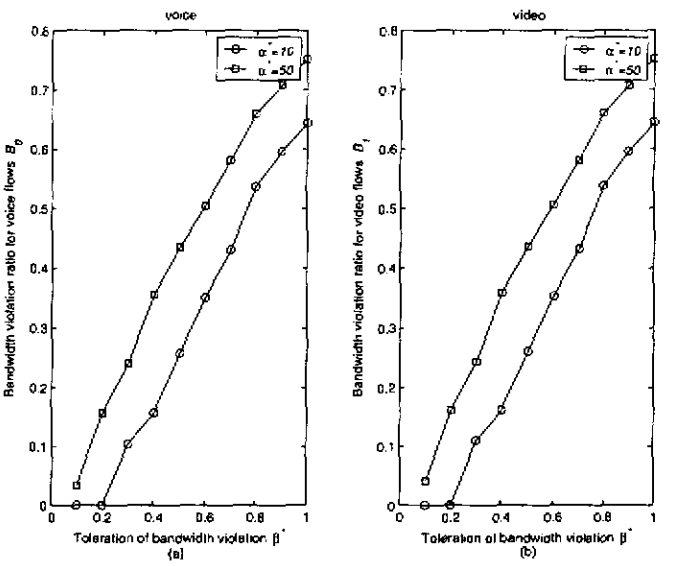


Fig. 5. Bandwidth violation ratio vs. toleration of bandwidth violation. (a) bandwidth violation ratio for voice flows. (b) bandwidth violation ratio for video flows.

[5] D. Gu and J. Zhang, "QoS enhancement in IEEE 802.11 wireless local area networks," *IEEE Communication Magazine*, vol. 41, pp. 120-124, June 2003.

[6] Y. L. Kuo, C. H. Lu, H. K. Wu, and G. H. Chen, "An admission control strategy for differentiated services in IEEE 802.11," *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, vol. 2, San Francisco, December 2003, pp. 707-712.

[7] H. Lin, M. Chatterjee, S. K. Das, and K. Basu, "ARC: An integrated admission and rate control for CDMA Networks based on non-cooperative games." *Proceedings of ACM International Conference on Mobile Computing and Networking (Mobicom)*, San Diego, California, Septmber 2003, pp. 326-338.

[8] R. B. Myerson, *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press, 1991.

[9] D. Seggern, *CRC Standard Curves and Surfaces*, Boca Raton, FL, CRC Press, pp. 124, 1993.

[10] U. Varshney, "The status and future of 802.11-based WLANs," *Computer*, vol. 3, pp. 102-105, June 2003.

[11] M. Xiao, N. B. Shroff, E. K. P. Chong, "Utility-based power control in cellular wireless systems", *Proceedings of the IEEE International Conference on Computer Communication (INFOCOM)*, vol. 1, 2001, pp. 412-421.

附錄二

室內用寬頻無線資訊接取系統子計畫(一)
系統架構、射頻、中頻及類比前端電路設計
System Architecture and RF/IF Analog Front-End Circuit Design

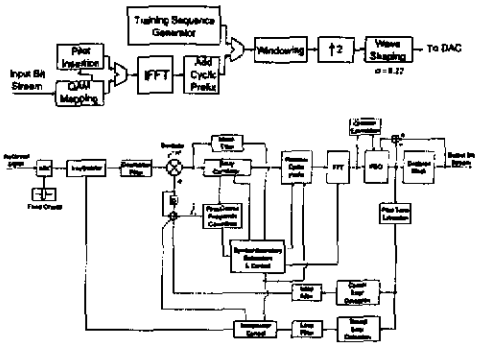
計畫編號：NSC 93-2213-E-002-001
執行期間：93/01/01～93/12/31
主持人：王志宏 Email: ckwang@cc.ee.ntu.edu.tw
執行機構及單位名稱：國立台灣大學電機工程學研究所

研究重點與研究成果

- 研究重點
  - IEEE802.11a 無線區域網路傳收機系統架構設計與模擬
  - IEEE802.11a/b/g 5.8/2.4GHz 雙頻帶射頻接收機架構與電路設計
  - IEEE802.11a 無線區域網路射頻接收機與基頻系統電路整合研究
- 研究成果
  - IEEE802.11a 無線區域網路傳收機系統架構設計與電路實做
  - IEEE802.11a/b/g 5.8/2.4GHz雙頻帶射頻接收機架構與電路之設計與實做
  - IEEE802.11a 無線區域網路射頻接收機前端整合測試電路板模擬
- 論文發表
  - Chun-Chih Hou, Ching-Chi Chang, Chorng-Kuang Wang, "A Dual-band IEEE 802.11a/b/g Receiver Front-end Using Half-IF and Dual-Conversion," *AP-ASIC 2004*, Fukuoka, Japan, Aug., 2004.
  - Chun-Chih Hou, Ching-Chi Chang, Chorng-Kuang Wang, "A Variable-length DHT-based FFT/IFFT Processor For Vdsl/Adsl Systems," submitted to *APCCAS 2004*.

第二年研究成果（I）

- IEEE802.11a OFDM傳收機系統架構設計與模擬



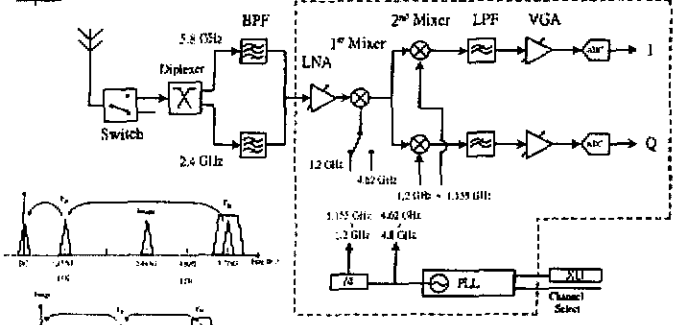
•Performance summary

SNR Requirement for PER<10%

	0 ns	50 ns	100ns
BPSK	9.7 dB	11.6 dB	12.0 dB
QPSK	12.8 dB	15.0 dB	15.9 dB
16QAM	20.7 dB	22.6 dB	23.0 dB
64QAM	26.2 dB	28.0 dB	29.3 dB

第二年研究成果（II）

- 雙頻帶接收機前端射頻電路架構

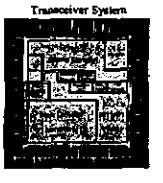


- ☐ I/5 IF for IEEE 802.11a
- ☐ Half-IF for IEEE 802.11b/g

第二年研究成果（III）

- 基頻系統晶片與射頻前端電路晶片設計與實做

- IEEE802.11a Baseband Transceiver System



- IEEE802.11a/b/g RF Front-end LNA and 1st Mixer

- IEEE802.11a Baseband Transceiver System

Data Rate (Mbps)	6, 9, 12, 18, 24, 36, 48, 54
Modulation (OFDM)	(BPSK, QPSK, 16QAM, 64QAM)
CFO Tolerance	± 20ppm (± 232.2 KHz)
TFO Tolerance	± 20ppm (± 2.5 KHz)
Technology	TSMC 0.25 μm 1P5M CMOS
Chip Area	3.5 × 3.5 mm² (2.8 × 2.8)
Gate Counts	302199
DAC/ADC	10 Bits
Clock Rate	40 MHz
Power Dissipation	198 mW
Power Supply	1.8 V
Package	128-pin CQFP

- IEEE802.11a/b/g RF Front-end LNA and 1st Mixer

		Simulation	Measure
2.4 GHz	LNA+Buffer Gain	14 dB	11 dB
	LNA+1st Mixer Gain	20 dB	12 dB
	LNA+Buffer NF	4 dB	9 dB
	LNA+1st Mixer NF	3.1 dB	8 dB
	LNA+1st Mixer IIP3	-19 dBm	-3 dBm
	S11	-20 dB	-4.3 dB
	ISR	30 dB	15 dB
	Up-converted LO noise	No	No
	Power (core only)	24 mW	24 mW
5.7 GHz	LNA+1st Mixer Gain	18.8 dB	3.5 dB
	LNA+1st Mixer NF	3.8 dB	25 dB
	LNA+1st Mixer IIP3	-13.4 dBm	-4 dBm
	S11	-15 dB	-7.5 dB
	ISR	30 dB	35 dB
	Power (core only)	24 mW	24 mW

第三年計畫之工作規劃

- Wireless MAN IEEE802.16系統架構設計
- 多頻帶射頻LNA電路以CMOS設計
- 多頻帶射頻mixzer電路以CMOS設計
- IEEE802.16無線寬頻網路射頻接收機與基頻系統電路整合研究

預期研究成果

- Wireless MAN 通訊系統架構設計
- 高頻射頻訊號處理晶片的設計與實做
- 雙頻射頻訊號處理晶片的設計與實做
- 射頻訊號處理晶片電路板的整合與驗證

室內用寬頻無線資訊接取系統子計畫二－基頻處理電路設計
Baseband Processing IC Design

計畫編號：NSC93-2213-E-002-001
執行期間：93/01/01～93/12/31

主持人：闕志達教授 Email: chiueh@cc.ee.ntu.edu.tw
執行機構：國立台灣大學電機工程學系暨研究所

研究重點與研究成果

- 研究重點
  - 高速無線區域網路之實體層演算法（MIMO-OFDM）比較與測試
  - 高速無線區域網路之接收機基頻處理架構設計
  - 無線都會區域網路(802.16a)之實體層演算法研究
  - 無線都會區域網路(802.16a)之接收機基頻處理架構設計
- 研究成果
  - 高速無線區域網路之實體層演算法（MIMO-OFDM）模擬結果
  - 高速無線區域網路封包格式、接收機同步與通道估計演算法設計與模擬結果
  - 無線都會區域網路(802.16a)之實體層演算法模擬結果
- 論文發表
  - Chi-Yeh Yu, Zih-Yin Ding, and Tzi-Dar Chiueh, "Design and Simulation of a MIMO-OFDM Baseband Transceiver for High Throughput Wireless LAN," submitted to *APCCAS 2004*.
  - Sang-Jung Yang, Yi-Ching Lei and Tzi-Dar Chiueh, "Design and Simulation of IEEE 802.16A OFDM Mode Subscriber Station Physical Layer Inner Transceiver," submitted to *APCCAS 2004*.

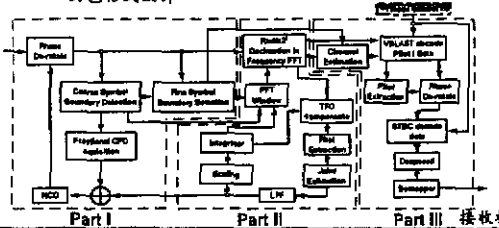
第二年研究成果（I）

- MIMO-OFDM系統封包格式、實體層參數與接收機架構

封包格式設計

實體層參數

Modulation type	BPSK, QPSK, 16QAM, 64QAM
FFT size	64
Subcarrier in use	52 (48 data, 4 pilot subcarriers)
Bandwidth	20MHz (Occupied bandwidth 16.6MHz)
Subcarrier spacing	312.5KHz (=20MHz/64)
FFT period	3.2us (=1/312.5KHz)
Guard interval duration	0.8us
OFDM symbol duration	4.0us (=3.2+0.8us)
Error correcting code	Convolutional code, r=1/2 [g0=133(oct), g1=171(oct)]
Coding rate	1/2, 2/3, 3/4
Operating frequency	5.2GHz band



Part I　Part II　Part III　接收機架構

第二年研究成果（II）

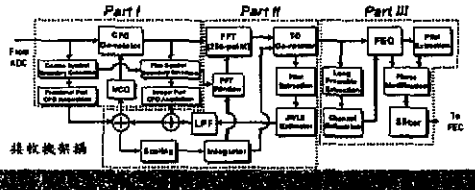
- 802.16a 實體層參數、使用頻寬與接收機架構

使用頻寬列表

BW(MHz)	NFFT	OFDMA PHY (256)				
		Tb(us)	Tg(us)			
			Tb/32	Tb/16	Tb/8	Tb/4
1.75	128	128	4	8	16	32
3.5	256	64	2	4	8	16
7.0	512	32	1	2	4	8
14.0	1024	16	0.5	1	2	4
28.0	2048	8	0.25	0.5	1	2

實體層參數

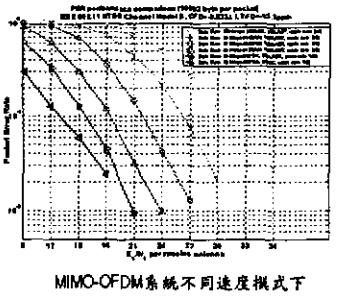
RF frequency	2-11GHz
FFT Size	256
Effective subcarriers	192
Bandwidth (MHz)	BW
Guard time (us)	Tg
Data time (us)	Tb
Symbol time (us)	Tg + Tb
Subcarrier spacing (kHz)	x1
Sampling rate (MHz)	Fs = BW x 8/7
Maximum data rate (Mbps) (BW=28MHz, 64QAM, code rate 3/4)	104.73



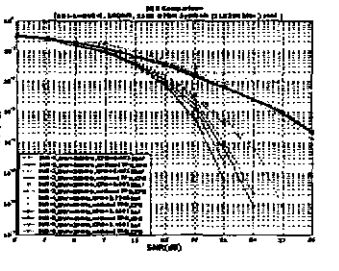
接收機架構

第二年研究成果（III）

- 系統模擬結果



MIMO-OFDM系統不同速度模式下封包錯誤率（PER）表現



802.16a OFDM系統在有/無載波頻率飄移及時脈飄移效應下位元錯誤率（BER）表現

第三年計畫之工作規劃

- 高速無線區域網路之2對2基頻收發機電路設計與驗證
- 高速無線區域網路之4對4系統之進階設計
- 無線都會區域網路之基頻收發機電路設計及驗證

預期研究成果

- 高速無線區域網路之2對2基頻收發機電路FPGA驗證
- 高速無線區域網路之4對4系統收發機架構
- 無線都會區域網路基頻收發機電路FPGA驗證
- 無線都會區域網路基頻收發機IP實作

室內用寬頻無線資訊接取系統-子計畫三-
適用於無線區域網路之加解密引擎IP設計及實現
Design and Implementation of Digital IP for Cryptographic Engine in Wireless LAN Applications

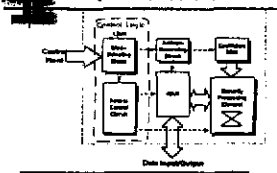
計畫編號：NSC 93-2213-E-002-001
執行期限：93/01/01~93/12/31

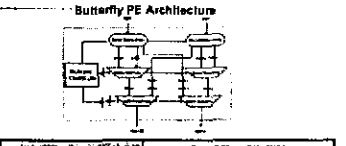
主持人：吳安宇教授 Email: andywu@cc.ee.ntu.edu.tw
執行機構：國立台灣大學電機工程學系暨研究所

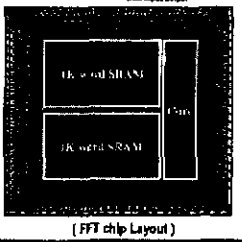
子計劃三：適用於無線區域網路之加解密引擎IP設計及實現
主持人：吳安宇教授(台灣大學)

- 研究重點
  - 適用於無線區域網路之數位IP模組演算法推導及電路架構設計及實現
  - 符合系統晶片滙排流之可重覆使用之數位IP模組設計及驗證模型
- 研究成果
  - 低複雜度反傅利葉/傅立葉架構設計
  - 李德-所羅門編碼器電路設計實作
  - 進階加密標準架構設計
  - 建立符合系統晶片滙排流之可重覆使用之數位IP模組
  - 以ARM為基礎之進階加密引擎電路設計

子計劃三：適用於無線區域網路之加解密引擎IP設計及實現
主持人：吳安宇教授(台灣大學)



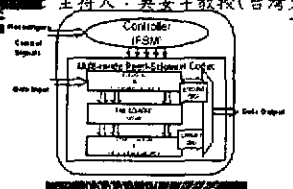


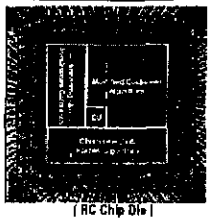


(FFT chip Layout)

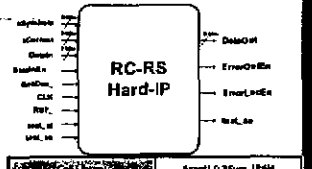
(Chip Summary)

子計劃三：適用於無線區域網路之加解密引擎IP設計及實現
主持人：吳安宇教授(台灣大學)



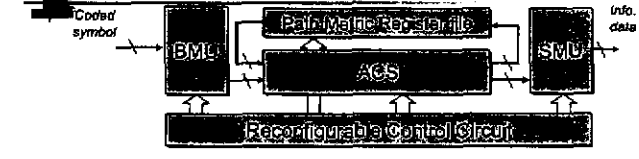


(RC Chip Die)



(Chip Summary)

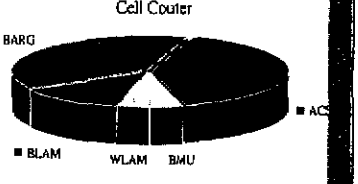
子計劃三：適用於無線區域網路之加解密引擎IP設計及實現
主持人：吳安宇教授(台灣大學)



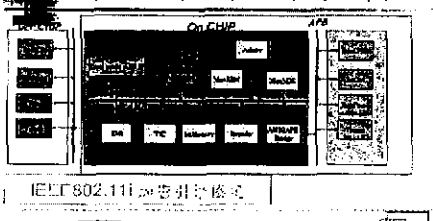
Cell counter

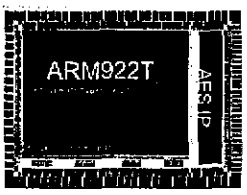
	Cell	Area	FlipFlops
BARG	BARG	409(37%)	264
ACSU	ACSU	448(45%)	1280
BMU	BMU	42(3%)	3
SMU	WLAM	41(3%)	26
	Ram	0(0%)	0
	BLAM	138(12%)	84
Total		1082(100%)	1789

Altera FLEX10K200E(EPF10K200SRC240)

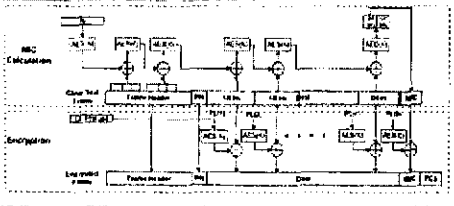


子計劃三：適用於無線區域網路之加解密引擎IP設計及實現
主持人：吳安宇教授(台灣大學)





IEEE 802.11i 加密引擎核心



Cell Library	Artisan 0.18um tpl4m
Voltage 電壓	1.8V
Core size 核心面積	4.1x2.6mm²
Die size 晶片面積	4.8X3.3mm²
I/o pad 腳位	128pin package
power 功率	780mW@100MHz
Ram size 記憶體	4X256Bytes

子計劃三：適用於無線區域網路之加解密引擎IP設計及實現
主持人：吳安宇教授(台灣大學)

- 晶片製作
  - 李德-所羅門編、解碼器(Reed-Solomon Codec)
- 論文發表

- Tsun-Shan Chan, Jen-Chih Kuo, and A. Y. Wu, "A Reduced-complexity Fast Algorithm for Software Implementation of the IFFT/FFT in DMT Systems," in EURASIP Journal on Applied Signal Processing, vol. 2002, no. 9, pp. 961-974, Sept. 2002 (SCI Expanded, Full paper).
- A. Y. Wu and Cheng-Shing Wu, "A Unified View for Vector Rotational CORDIC Algorithms and Architectures based on Angle Quantization Approach," in *IEEE Trans. Circuits and Systems Part-I: Fundamental Theory and Applications*, vol. 49, no. 10, pp. 1442-1456, Oct. 2002 (Full paper, SCI, EI).
- H. Y. Hsu and A. Y. Wu, "VLSI Design of a Reconfigurable Multi-mode Reed-Solomon Codec for High-Speed Communication Systems," In Proceedings. IEEE Asis-Pacific Conference on ASICs, pp.359-362, 2002.

室內用寬頻無線資訊接取系統子計畫四：
適用於數位用戶迴路及無線區域網路的DSP設計
Stream-Based DSP core for xDSL and Wireless LAN

計畫編號：NSC 93-2213-E-002-001
執行期間：93/01/01 ~ 93/12/31

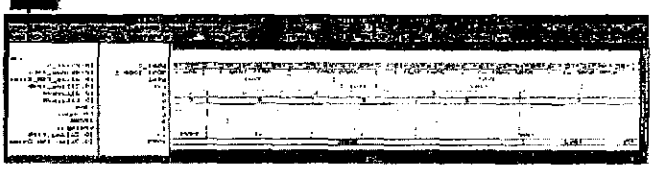
主持人：周世傑教授 Email: jerry@ee.ncu.edu.tw
執行單位：國立中央大學電機工程學系(所)

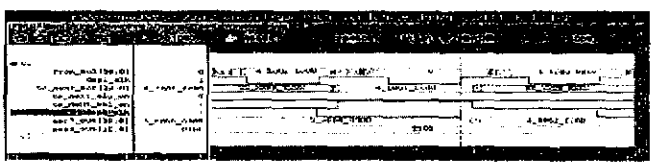
研究重點與研究成果

- 研究重點
  - 系統可參數化架構設計及核心設計
  - 串流式架構(Stream-Based) DSP設計
  - 無線網路及多重(Multi-DSP)DSP之評估
  - DSP 核心之IP化實現
- 研究成果
  - 系統可參數化架構設計及精進設計完成
  - 串流式架構設計完成
  - 無線網路及多重(Multi-DSP)DSP之評估完成
  - DSP 核心之IP化實現完成
- 論文發表中
  - Ya-Lan Tsao, Shyh-Jye Jou and Wei-Hao Chen "Buffering Hardware Nested Loop of Parameterized and Embedded DSP," Circuits and Systems, 2003. ISCAS The 12th Workshop on Synthesis And System Integration of Mixed Information technologies
  - Ya-Lan Tsao, Shyh-Jye Jou and Wei-Hao Chen "Buffering Hardware Nested Loop of Parameterized and Embedded DSP," IEE Electronic Letters
  - Ya-Lan Tsao, Yu-Chun Lin, Wei-Hao Chen, Bo-Shiang Huang and Shyh-Jye Jou "A module generator for parameterized DSP core," 2004 IEEE Asia-Pacific Conference on Circuits and Systems
  - Ya-Lan Tsao, Yu-Chun Lin, Wei-Hao Chen, Bo-Shiang Huang and Shyh-Jye Jou "A module generator for parameterized DSP core," IEE Proceeding
  - Ya-Lan Tsao, Jun-Xian Teng, Hao-Ching Lin and Shyh-Jye Jou," Low Power Correlator of DSP Core for Communication System," 2004 IEEE Asia-Pacific Conference on Circuits and Systems

第三年研究成果（I）

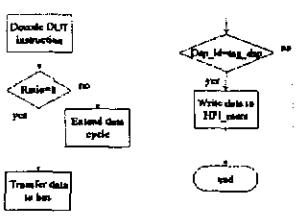
- Stream Based I/O 在執行速度上驗證





第三年研究成果（II）

- Stream Based I/O的整合測試驗證(I)

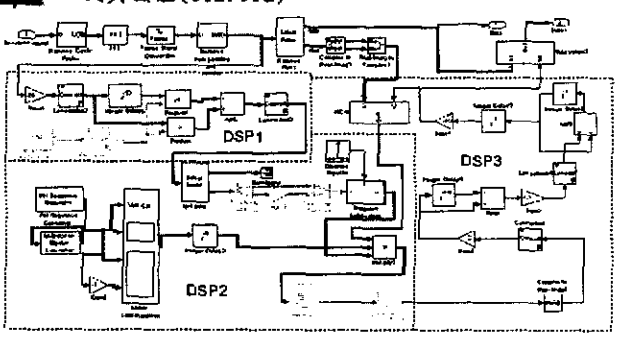


	No stream IO	Stream IO
Area	103240 (gate count)	105714 (gate count)

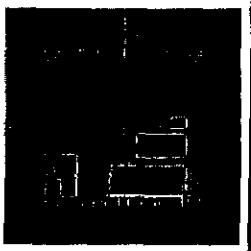
Area overhead is only 2.4%

第三年研究成果（III）

- Stream Based I/O在Wireless LAN的整合測試與驗證(802.11a)



第三年研究成果（IV）

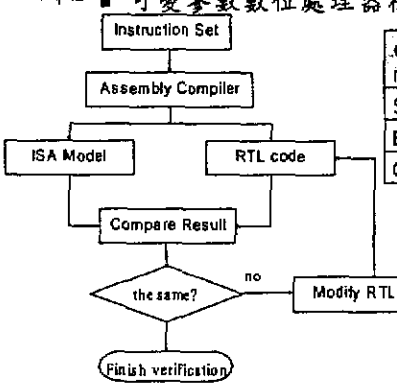


DSP chip in 0.25um standard cell

Technology	TSMC 0.35um SPQM	TSMC 0.25um 1PSM
Cell Library	Avant!0.35um Cell Library	Artisan 0.25 Cell Library
Supply	3.3V	2.5V
Pipeline	6-stage	6-stage
Max. Clock	100MHz	110MHz
Gate Counts (memory is not included)	49,895	75611
On-Chip Memory	512*24bit, 512*16bit, 64*16 bit	512*24bit, 512*16bit, 64*16 bit
Power Consumption @100MHz	740mW	275.96mW

第三年研究成果（V）

- 可變參數數位處理器核心之IP化實現



coverage metric	total number	coverage percentage(%)
Statement	8541	98.3
Branch	1251	94.2
Condition	14	100

Test_vn_merge

室內用寬頻無線資訊接取系統 子計劃五 –
射頻和混合訊號系統晶片之測試技術
Test Methodology for Analog and RF Front End Circuit

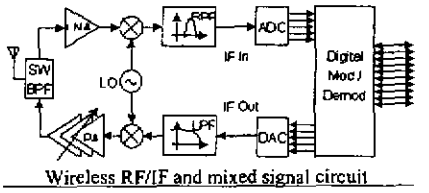
計畫編號：NSC 93-2213-E-002-001
執行期限：93/01/01 ~ 93/12/31
主持人：朱紹祖教授 Email: ccsu@cn.nctu.edu.tw
執行機構：國立交通大學電機與控制工程學系

研究重點與研究成果

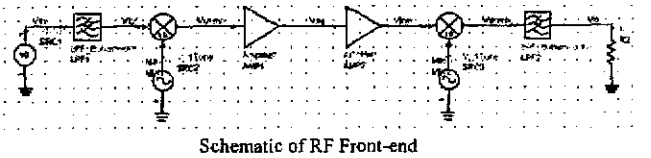
- 研究重點
  - 測試、模擬無線網路傳送接收器射頻與中頻架構
  - 提出無線網路傳送接收器測試模擬方法
  - 推導參數變動與效能變化的關係與數學、統計分析
- 研究成果
  - 無線網路傳送接收器射頻與中頻電路模擬環境建立
  - 推導參數變動與效能變化的關係
- 論文發表
  - Hung-kai Chen, Chauchin Su,"Convolution Based RF Transceiver Testing Methodology", IMSTW 2004

第二年研究成果（I）

- 射頻、中頻模擬環境建立



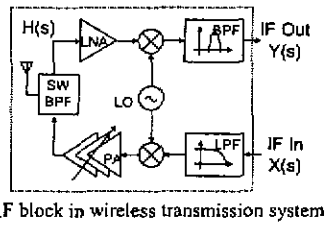
Wireless RF/IF and mixed signal circuit


Schematic of RF Front-end

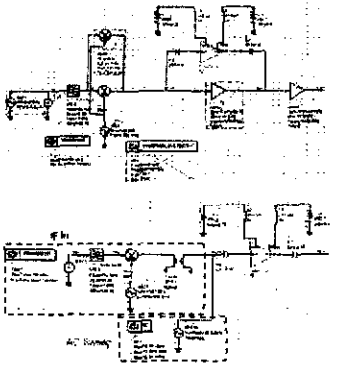
第二年研究成果（II）

- 理論與測試訊號產生方法

Convolution : $y(t) = x(t)*h(t)$
$Y(s) = X(s)\,H(s)$
$H(s) = Y(s) / X(s)$
$H(s)db = Y(s)_{db} - X(s)_{db}$ (log scale)

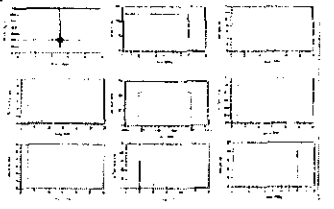


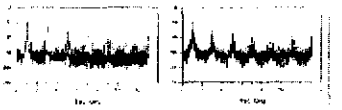
RF block in wireless transmission system

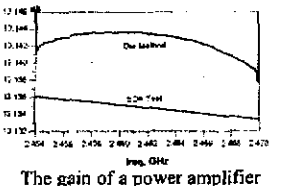


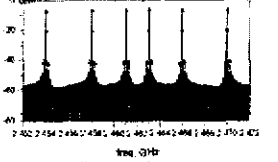
第二年研究成果（III）

- 參數變動與效能變化的關係


Ideal RF Front-End simulation results


Ideal Mixer　　IP3=1 dBm


The gain of a power amplifier


The multitone test

第三年計畫之工作規劃

- 整合及測試混合信號模組(ADC/DAC) 來產生中頻所需之訊號
- 整合802.11a 之混合信號、中頻、射頻模組
- 測試整合之802.11a 模組
- 整合與測試無線區域網路傳送接收器

預期研究成果

- 建構無線網路測試方法
- 數學推論參數變化與效能關係
- 整合類比數位轉換器與數位類比轉換器來產生中頻所需之訊號
- 整合與測試無線區域網路傳送接收器之混合信號、中頻、射頻模組

室內用寬頻無線資訊接取系統子計畫六：
無線網路品質服務分析及無線隨意網路公平排程機制

計畫編號：NSC 93-2213-E-002-001
執行期限：93/01/01 ~ 93/12/31

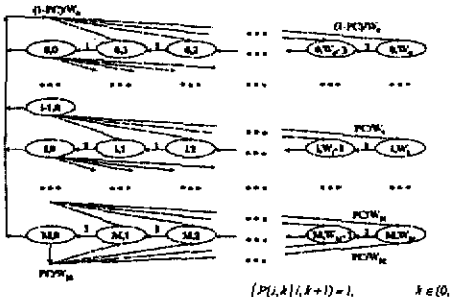
主持人：廖婉君 教授 Email: wjliao@cc.ee.ntu.edu.tw
執行機構：國立台灣大學電機工程學系暨研究所

研究重點與研究成果

- 研究重點
  - IEEE 802.11e WLAN 之 QoS
  - 無線隨意網路之QoS機制
- 研究成果
  - IEEE 802.11e 之 priority-based 及 fairness-based QoS 機制
  - 無線隨意網路中之公平排程機制
- 論文發表
  - Hsi-Lu Chao and Wanjiun Liao, "Fair Scheduling with QoS Support in Ad Hoc Wireless Networks," accepted by IEEE Transactions on Wireless Communications, 2004.
  - Hsi-Lu Chao and Wanjiun Liao, "Channel Compensation for Multimedia Wireless Ad Hoc Networks with Channel Errors," accepted by IEEE Transactions on Wireless Communications 2004.

第二年研究成果（I）

- QoS in IEEE 802.11e WLAN

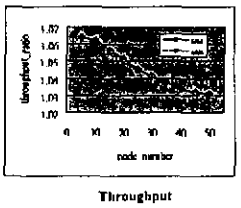


• Markov Chain Model for AC Backoff Window Size

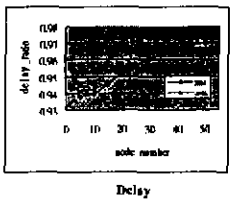
$$\begin{cases} P(i,k\mid i,k+1)=1, & k\in(0,W_i-1), i\in(0,M) \\ P(0,k\mid i,0)=(1-PC)/W_0, & k\in(1,W_0), i\in(0,M) \\ P(i,k\mid i-1,0)=PC/W_i, & k\in(1,W_i), i\in(1,M) \\ P(M,k\mid M,0)=PC/W_M, & k\in(1,W_M) \end{cases}$$

第二年研究成果（II）

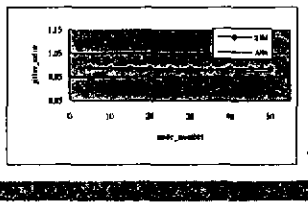
- Throughput, Delay and Jitter Analysis in IEEE802.11e



Throughput



Delay

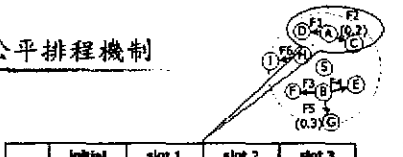


Jitter

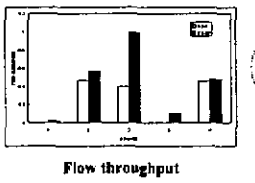
第二年研究成果（III）

- 無線隨意網路 公平排程機制

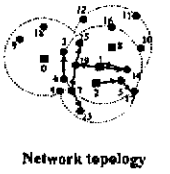
$$\rho=\frac{f}{n}\sum_{j=1}^{n}x_j,\quad x_j=\begin{cases}1 & \text{if } \varphi(i)\ge 0\\ 1-\alpha\times\dfrac{|\varphi(i)|}{Resv(i)} & \text{if } \varphi(i)<0\end{cases},\quad x_i=\frac{\sum_{j=1}^{n}x_j^2}{\sum_{j=1}^{n}x_j}$$



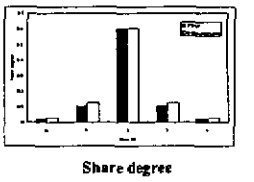
		Initial			slot 1			slot 2			slot 3		
	FID	C	U	E	C	U	E	C	U	E	C	U	E
Best effort	F1(b)	0	0	0	0	0	0	0	0	0	0	0	0
Guaranteed	F2(r)	0.2	0	-0.2	0.2	0	-0.2	0.4	1	0.6	0.6	1	0.4



Flow throughput



Network topology



Share degree

第三年計畫之工作規劃

- IEEE802.16上之研究
  - 效能分析模型
  - MAC 機制與TCP的交互作用
  - QoS 機制
  - 對IEEE802.11與IEEE802.16並存之網路架構分析其交互影響與效能上的變化
  - 可靠性群播機制

預期研究成果

- 提出IEEE802.16 MAC之效能分析模型
- 針對IEEE802.16上不同的QoS實作機制建立分析模型並比較其優缺
- 分析IEEE802.16 MAC 對上層協定的影響
- 針對WiFi及WiMAX並存的網路架構，提出解決方案，以達效能的最佳化。
- 提出BWA上的Reliable Multicast機制

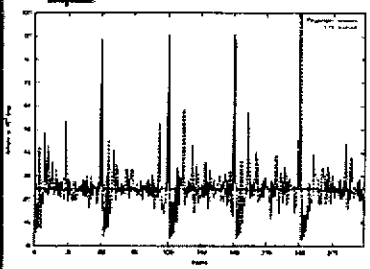
室內用寬頻無線資訊接取系統 子計畫七 –
異質網路間之服務品質保證及視訊應用
Dynamic Resource Management and Adaptive Video Streaming Technology over wireless broadband

計畫編號：NSC 93-2213-E-002-001
執行期限：93/01/01 ~ 93/12/31
主持人：吳曉光 教授 Email: hsiao@csie.ncu.edu.tw
執行機構：國立中央大學資訊工程學系

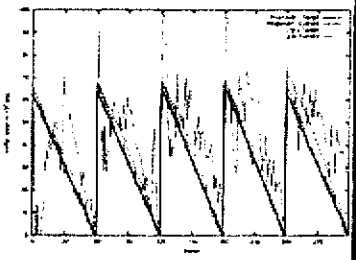
研究重點與研究成果

- 研究重點
  - 視訊串流的可調適性速率控制設計與佈局
  - 可靠多點傳輸在無線網路上的改進
- 研究成果
  - 視訊串流的可調適性速率控制設計與實做
  - 可靠多點傳輸在無線網路上改進的模擬與分析
- 論文發表
  - Eric Hsiao-kuang Wu, Meizhen Chen, "JTCP, Jitter-based for Wireless Networks", accepted for IEEE Journal of Selected Area Communications (JSAC)
  - Shi-Yang Chen, Eric Hsiao-kuang Wu, Yi-Shuang Chen, " A New Approach Using Time-Based Model for TCP-Friendly Rate Estimation", appear in Proceedings of IEEE ICC2003
  - Eric Hsiao-kuang Wu, Hsu-Te Lai, Meng-feng Tsai, "Low Latency and Efficient Packet Scheduling for Streaming Applications", will appear in Proceedings of IEEE ICC 2004
  - Eric Hsiao-kuang Wu, Chien-Shao Tseng, Chung-Yuan Chang, Meng-feng Tsai, "TMRC: A Load-Adaptive TCP-Friendly Rate Control Protocol for Real-time Multimedia Applications" will appear in Proceedings of IEEE ICC2004

第二年研究成果（I）
■ 視訊串流的可調適性速率控制

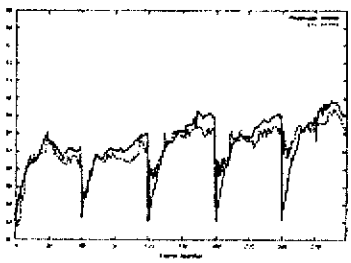


畫面位元率的控制結果比較

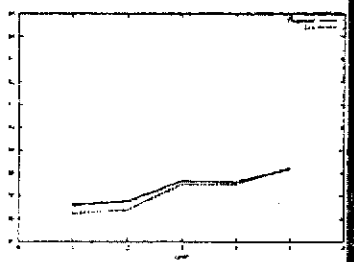


緩衝區飽滿度的控制結果比較

第二年研究成果（II）
■ 視訊串流的可調適性速率控制

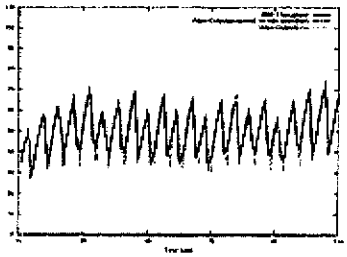


畫面間畫質的控制結果比較

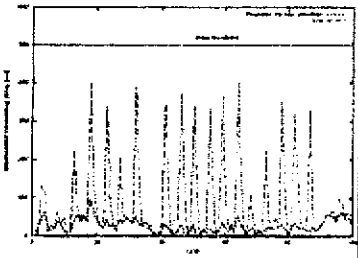


群組畫面間畫質的控制結果比較

第二年研究成果（III）
■ 視訊串流的可調適性速率控制

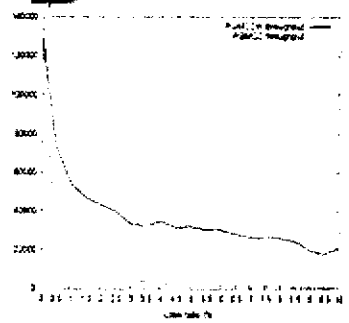


視訊輸出量與網路傳輸量控制結果

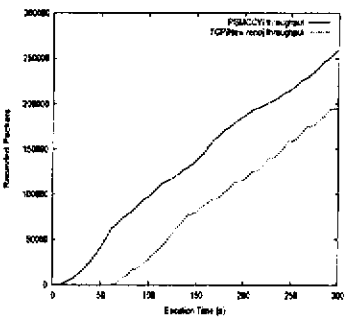


每秒畫面傳送延遲時間的控制結果

第二年研究成果（IV）
■ 可靠的多點傳輸在無線環境下的改良



無線網路環境下的傳輸效能



保持原先TCP Friendly的特性

預期研究成果

- Wireless LAN上即時視訊串流系統架構設計
- 即時視訊串流系統視訊延遲控制的設計與實作
- 即時視訊串流系統錯誤控制機制的設計與實作
- 無線網路上可靠的多點傳輸增進傳送效率
- 增進多點傳輸效能同時，保持與主流網路協定共享頻寬的特性

附錄三

Outline

- 應用MIMO-OFDM技術之高速無線區域網路基頻訊號處理架構(C程式)

應用MIMO-OFDM技術之高速無線區域網路基頻訊號處理架構(C程式)

- 技術移轉項目：
  - 應用MIMO-OFMD技術之高速無線區域網路基頻收發機電路架構，內含
    - MIMO-OFDM基頻傳送機模擬程式
    - MIMO通道模擬程式（採用IEEE 802.11 TGn之MIMO Channel model D）
    - MIMO-OFDM基頻接收機模擬程式
- 預期利用範圍及預期產品
  - 高傳輸率無線區域網路裝置及其數位接收處理積體電路

應用MIMO-OFDM技術之高速無線區域網路基頻訊號處理架構(C程式)

- 功能描述：
  - 本技轉之MIMO-OFDM基頻收發機為pre 802.11n，支援5GHz的UNII頻帶與20MHz的頻寬，採用兩根傳輸天線與兩根接收天線進行空間多工傳輸。
  - 最高速度模式為IEEE 802.11a最高實體層（PHY）傳輸率的兩倍（54 x 2 = 108 M bps）。而系統在未來進一步的速度延伸上，可藉由調整錯誤更正碼率至7/8，並配合有效的MAC-PHY轉換率（80%）來達成IEEE 802.11n的傳輸率要求－媒體存取控制層（MAC）傳輸率100 Mbps = 2 x 48 x 6 x (7/8) x (80%) / 4。

應用MIMO-OFDM技術之高速無線區域網路基頻訊號處理架構(C程式)

- 功能描述：
  - 目前系統支援六種速度模式：12、24、48、72、96、108 Mbps。各速度模式使用不同的MIMO傳輸方式（VBLAST與STBC）、不同的星座圖大小（QPSK、16QAM、64QAM）以及不同之錯誤更正碼率（Punctured Convolutional Code Rate = 1/2、2/3、3/4）。
  - 系統使用的MIMO通道模型為典型辦公室環境，MIMO channel model D。此通道模型為非直接路徑（NLOS），RMS delay spread為50 ns，具有三個cluster。且各通道之Rayleigh fading係數具有相關性。另外，系統一併模擬傳送端與接收端不完美效應如power amplifier (PA) non-linearity、carrier frequency offset (CFO)、phase noise、I/Q imbalance、DC offset與timing offset (TO)。

應用MIMO-OFDM技術之高速無線區域網路基頻訊號處理架構(C程式)

- 功能描述：
  - 系統於初始同步過程中利用短前置碼及長前置碼估得coarse and fine symbol boundary、fractional CFO，並經由快速傅利葉轉換（FFT）將訊號轉換至頻域，以進行殘餘CFO及TO之追蹤。而具有正交性的長前置碼，在轉至頻域後，可藉由線性組合來完成通道估計。且在低訊噪比時，通道估計結果可再藉由頻域平均來改善。最後，資料回復是利用VBLAST等化器、進行ordered successive cancellation，再通過決策器來決定對應的位元。外接收機接者負責錯誤更正碼解碼，解得原始資料位元。
  - 系統可支援IEEE 802.11a與pre IEEE 802.11n封包格式偵測。更進一步，系統除了2x2的傳送接收(VBLAST/STC)外，尚可向下相容1x1、1x2(MRC)額外兩種模式，以支援 802.11a。

附錄四

國科會補助研究計畫之主持人出席國際會議報告

93 年 8 月 17 日

報告人姓名	汪重光	學校系所	國立臺灣大學電機工程學系暨研究所
會議期間及地點	Aug. 4~5, 2004 日本 福岡	計畫編號	NSC-93-2213-E-002-001
會議名稱	(中文)2004 亞太先進整合系統電路會議 (英文)2004 Asia-Pacific Conference on Advanced System Integrated Circuits (AP-ASIC 2004)		
發表論文題目	(中文) (1). 應用於寬頻通訊有效應體實作之 64-QAM 低中頻傳收機 (2). 一雙頻射頻電路採半中頻及雙降頻架構於 IEEE802.11a/b/g (英文) (1). A Hardware Efficient 64-QAM Low-IF Transceiver Baseband for Broadband Communications (2). A Dual-band IEEE802.11a/b/g Receiver Using Half-IF and Dual Conversion		

一、參加會議經過

會議分兩天來舉行，總共有21個sessions，我所要報告的paper分別在第11與18的session，各在這兩天的下午。第一天一大早就到會議報到處領取收據及相關資料，隨後碰到了幾位台灣來的朋友，接著就進入會議室聆聽各論文的報告。由於這21個sessions分三個會議室來進行，因此就無法聽取所有的論文報告，所以我們個別選取了相關的領域來聽取。第一天早上只有一個session，我選擇了Session 3:Analog signal procession circuits來聽取，用過午餐後，我選擇了Session 8: Analog techniques for RF來聆聽，而這個會議室也是下個我將報告的session之會議室，目的也是先來適應一下環境，雖然報告對我來說是還算駕輕就熟的，但必須以英文來報告，對我來說多少都有點擔心。接下來的Session 11:High speed and wide-bandwidth links，我的論文排在第七篇，"A Hardware Efficient 64-QAM Low-IF Transceiver Baseband for Broadband Communications"，報告過程很順利，而且有位來自台灣元智大學的黃教授還在報告後舉手提問題與我討論。

第二天的早上我選取了Session 15:High speed links來聽取，主要的原因是一位熟悉的朋友在這個session報告。接著用過午餐後，下午的第一個session時段是海報展覽，我大約看了幾個有興趣的議題後即到室外去練習一下我自己的報告內容。下午的第二個session時段我則到Session 18:RF front-end circuits去，這個session只有四篇papers而台灣來的就佔了三篇，而且正巧的我們都互相認識，而我的排在第二篇，"A Dual-band IEEE 802.11a/b/g Receiver Front-end Using Half-IF and Dual Conversion"。順利報告完後，接著的是Invited Talk及Panel Discussion，之後就結束了這次的會議。

二、與會心得

這次的AP-ASIC是在日本福岡舉行，雖不像旅遊雜誌上所描述的東京般繁華，但其多姿多采的民族與文化氣息仍是令人難以忘懷。除了重要都會的繁榮外，還有周邊城鎮的人文勝地，讓我重新地去深入認識日本。我們利用會議前的幾天參觀了豪世登堡、小蒼城與太宰府，讓我重新地對日本人豎然起敬，日本人的文化傳承之用心與敬業的態度，是其成爲工業與經濟強國的主因。

在會議的過程來說，以英語溝通爲主，很明顯地台灣人大部份的英文能力比起日韓人士來得好些，但也有許多台灣的報告者以唸稿的方式報告，真覺得有失禮儀的地方。會議的各個session中，我選擇了些有關系統整合與射頻系統的部份來聽取，很明顯地可以發現，亞太區以超大型積體電路的發展在全世界的水準來說並不會落後於其他國家，尤其是在一些先進的系統議題

上，甚至與歐美並駕齊驅。而台灣在這次AP-ASIC的投稿率又超出全數的大半，換句話說台灣的研發能力已達世界之流。較爲遺憾的是，此次會議中看到許多日本大學或研究室來參加的學者或研究員，他們不一定有發表研究結果，但卻是抱著學習新知的理念而來，這是值得國人學習的地方。

三、建議

1. 參與國際學術會議對於國內研究人員有相當大的幫助，除了可以了解目前國際上技術的發展現況，也可以提昇自身的視野。國內研究人員應積極參與國際大型會議，這是一件非常珍貴且深具意義的事。
2. 由於國際會議報名時間通常都很早，希望公布獲得補助的時間可以提早些，以便及早準備出國的相關事務。
3. AP-ASIC 慣例上是在台灣、日本、韓國及大陸舉行。若國內能充分利用舉辦大型國際會議的機會，除了達到促進國際交流的目的以開闊國內研究生的眼界，同時也能推展國內旅遊，使世人增加了解寶島台灣的特色。

四、攜回資料名稱及內容

1. AP-ASIC 2004 Program及Proceedings各一本
2. 會議論文CD-ROM 一份，內含所有發表論文內容

五、其他

台灣的觀光業缺乏像外國如此精致地包裝，所以在觀光業上無法發達起來。其實台灣是有許多吸引人之處，只需要相關單位加以適當的管理與宣傳，當達到一定知名度時，舉辦Conference 的話，就能吸引國際更多人士來參與。

A Hardware Efficient 64-QAM Low-IF Transceiver Baseband for Broadband Communications

Ching-Chi Chang, Muh-Tian Shiue*, and Chorng-Kuang Wang
Graduate Institute of Electronics Engineering, and Department of Electrical Engineering,
National Taiwan University, 106 Taipei, Taiwan R.O.C.
*Department of Electrical Engineering, National Central University, 320 Jung-Li, Taiwan R.O.C.
E-mail: ckwang@cc.ee.ntu.edu.tw

Abstract—**This paper presents a hardware efficient VLSI design of digital baseband for 64-QAM communication systems over the last-mile cable network. This VLSI system design involves a cost-efficient architecture of the adaptive equalizer and a two-phase linear architecture of the pulse shaping filters, which reduce the hardware requirement by a factor of four comparing with traditional quadrature direct form FIR filters. In this design, the two-fold carrier recovery loop possesses a pull-in range of $\pm 100kHz$ (i.e. $\pm 18,500ppm$ of the symbol rate) and $-82dBc$ jitter suppression. Based on the proposed multi-staged LMS-based fractionally-spaced equalizer, the receiver realizes the symbol spaced timing recovery in a $\pm 200ppm$ tolerance of the symbol rate. The acquisition time of the proposed 64-QAM blind adaptive system is $7ms$, and the transceiver reaches the operation speed of the case for $32.28Mb/s$ 64-QAM low-IF digital CATV system over NTSC 6MHz bandwidth channels. Using $0.35\mu m$ CMOS technology, the transceiver design occupies a chip area $5.5mm \times 5.5mm$ and power consumption 1.35W (1.0W for RX) when the power supply is 3.3V.**

I. Introduction

Fig.1 shows a typical functional block diagram of QAM transceiver for last-mile broadband digital communication systems, such as digital CATV [1], SCM-VDSL [2] and Wireless MAN [3]. The transmitter mainly consists of an FEC encoder, a symbol mapper, a Nyquist pulse shaping filter (PSF), a quadrature mixer, and the analog front end (AFE). In the receiver, a digital quadrature mixer and two identical low-pass filters (LPFs) are used to demodulate the received passband signal. The coherent carrier recovery (CR) loop is employed to combat the performance degradation from non-coherent receiving. The timing recovery (TR) loop for symbol timing synchronization provides the ADC a proper sampling instant. In order to combat the ISI caused by the band-limited channel, an adaptive equalizer with blind adaptation is utilized.

In a conventional architecture of a baseband adaptive QAM decision feedback equalizer (DFE), both feed forward (FF) and feedback (FB) filters contain tens to hundreds adaptive complex taps. This requires a significant hardware on an SoC chip. In order to reduce the chip area, the equalizer can employ a multiplication-and-accumulation (MAC) unit with higher clocking rate by paying the power consumption and circuit complexity [4]. Practically, this advantage will be eliminated when the equalizer is designed to combat various channel conditions with variable equalization tap numbers [1][5]. In a coherent receiving, it is then imperative to extract frequency and phase of the carrier from the incoming signal [6]. Most of the data modems require fast carrier frequency/phase acquisition in the startup and high jitter suppression in the steady state. In practice, it is hard to be satisfied in a simple CR loop, especially when the RF local oscillator (LO) frequency offset has to be taken into account in RF transmission systems [1][3]. In order to reduce the transceiver chip area, the PSF in the transmitter and the LPF in the receiver can be replaced by image rejection lowpass filters (IRLFs) [7]. The IRLF is a linear-phase half-band filter with an I/Q-path interleaving scheme. The hardware cost of the IRLF can be further reduced by eliminating the long delay line which is proportional to the number of coefficient taps.

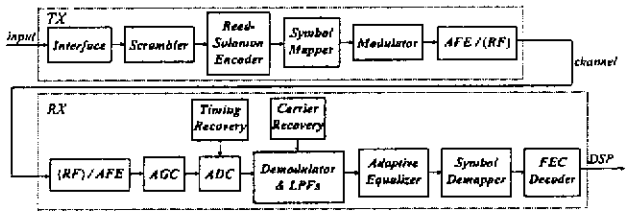


Fig. 1. Functional block diagram of QAM transceiver for broadband communications

Based on the QAM transceiver architecture shown in Fig.1, this paper presents a hardware efficient architecture for the multi-stage LMS-based blind equalizer. The proposed pipeline architecture of the equalizer shown in Fig.3 processes some parallel parts sequentially to reduce the hardware requirement by a factor of four comparing with the architecture implemented directly from the transfer function. Taking the advantage of the $T/2$ fractionally spaced equalizer, the TR loop shown in Fig.4 utilizes a baud-rate decision-directed early-late timing extraction scheme to achieve a $\pm 200ppm$ timing frequency tolerance within $7ms$ acquisition. Since it is mandatory to provide wide locking range and low jitter variation of carrier recovery in a high constellation QAM RF transmission system, the proposed two-fold CR loop can achieve a locking range $\pm 100kHz$ (i.e. $\pm 18,500ppm$ of the symbol rate) and jitter suppression $-82dBc$ for carrier retrieval in a single architecture as shown in Fig.7. In order to further reduce the hardware complexity, the proposed architecture shown in Fig.2 for the two-phase linear FIR filter can actualize higher hardware efficiency than an IRLF can do.

The organization of this paper is as follows. The transceiver design is presented in Section II. Section III presents VLSI implementation considerations of the transceiver. Section IV shows the simulation results, and the conclusion will be drawn in section V.

II. Transceiver Architecture

In this transceiver VLSI architecture design, the 64-QAM $4.035MHz$ low-IF digital CATV system is demonstrated. Based on the various digital CATV channel models [8], all the functional blocks are designed to overcome the non-ideal transmission circumstances.

A. Transmitter

In the proposed QAM transmission system, in order to avoid the drawback of analog demodulation, such as DC offset and mismatch between quadrature paths, the low-IF $4.035MHz$ modulator is implemented in digital domain. In addition, the anti-sinc FIR filter

with linear phase is used to compensate the inherent DAC distortion. The square-root raised-cosine filters with a roll-off factor 0.1152 are employed for both transmitter PSFs and receiver LPFs in the CATV system. The proposed two-phase linear FIR filter architecture shown in Fig.2 translates the linear phase FIR filter into a linear phase halfband filter. Using the two-phase clocking technique, the in-phase data are latched by the negative clock edge in upper registers while the quadrature data are latched by the positive clock edge in the lower registers. The mutiplexers (MUXs) are controlled by the clock phase to pass the in-phase data to the adders when the clock is high. In contrast, the quadrature data is passed to the adders when the clock is low. Comparing with the traditional quadrature direct form FIR filter, which is composed of $2N$ multipliers and $(2N-2)$ adders, the hardware complexity is reduced to $(N+1)/2$ multipliers and $(N-1)$ adders, where N is the tap number of the quadrature PSF. The power consumption is approximately half of the traditional quadrature direct form FIR filter.

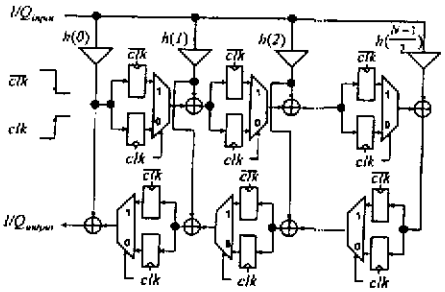


Fig. 2. The proposed two-phase architecture for the pulse shaping filter

B. Adaptive Equalizer

Based on the multi-stage and signed delay-LMS algorithm [8], the modified LMS algorithm is shown in the following.

$$C(n+1) = C(n) + \mu \cdot r(n-1) \cdot sgn(e^*(n-1)) \qquad (1)$$

where $C(n)$ is the adaptive complex coefficient. μ is the updating step size, which is designed to be a number of power of two for scaling the signal $r(n-1) \cdot sgn(e^*(n-1))$ in each adaptive stage with a factor of two to eliminate multiplications. $e(n)$ and $r(n)$ are the error message and the received signal, respectively, and the complex multiplier is replaced by a two's complementer. A stop-and-go flag mechanism [9] associated with μ is introduced for blind updating. The reduced constellation algorithm (RCA) is adopted to blindly start the equalizer in the acquisition state.

If the direct form and fully parallel process are employed to construct complex filter banks, it will consume $(4N+4M)$ real multipliers and adders, where N and M are FF and FB tap numbers respectively. In order to reduce the hardware cost and circuit complexity at the same time, the 8-stage pipeline architecture shown in Fig.3 is proposed. The clk in Fig.3 is 8-times of symbol clock rate, i.e. 43.04MHz. The complex-value convolution is sequentially performed by four real value convolutions during the preliminary three hardware stages. After decreasing delays, the real value convolution results are ready at $T3$ to $T6$, and complex-value sums, slicing, equalized signal SNR calculations and $1/8$ decimation are performed in the final stages, $T6$ and $T7$. The proposed pipeline architecture of the blind adaptive equalizer consumes $(N+M)$ real multipliers and adders which is only a quarter of the architecture implemented directly from the transfer function. The programmable numbers N and M can be easily set up according to the channel condition by the hand-shaking protocol during the startup of the transmission.

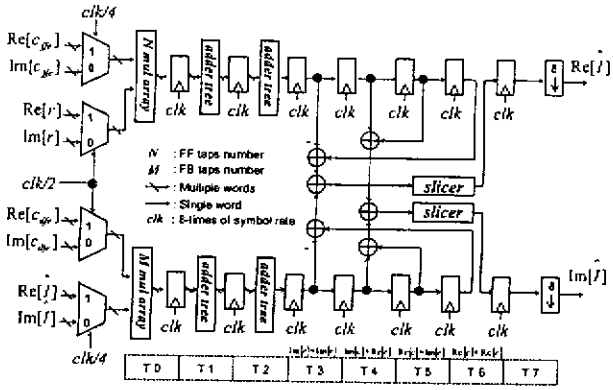


Fig. 3. The proposed 8-stage pipeline architecture for the adaptive blind equalizer

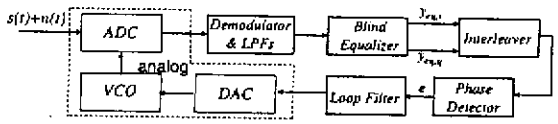


Fig. 4. The receiver architecture using mixed-mode baud-rate timing recovery loop

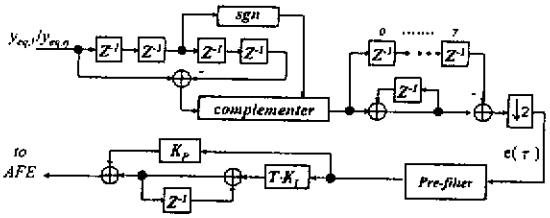


Fig. 5. The phase detector and the loop filter of the timing recovery loop

C. Timing Recovery

Instead of suffering signal SNR degradation in an all-digital TR loop with an interpolator, the mixed-mode TR loop architecture shown in Fig.4 is devised. Practically, the stochastic estimation technique cooperated with a digital phase-locked-loop (PLL) is employed to maintain system stability. Based on the baud-rate early-late timing estimation algorithm shown in Eq.(2) [10], the decision-directed early-late timing estimation derived in Eq.(3) is proposed.

$$e(\tau) = Re\{A_n \cdot [y_{eq}(T,\tau) - y_{eq}(-T,\tau)]\} \qquad (2)$$

$$e(\tau) = E\{sgn(\hat{I}_n(\tau)) \cdot [y_{eq,n+1}(\tau) - y_{eq,n-1}(\tau)]\} \qquad (3)$$

where A_n and y_{eq} are transmitted and equalized signals, respectively, and $\hat{I}_n$ is the decision result. The digital multiplier can be avoided by using the sign bit of the slicer output. Fig.5 shows the circuit of the phase detector and the loop filter of the baud-rate TR loop. Preceded the proportional and integral (PI) loop filter, an IIR pre-filter is employed to suppress the noise power. This pre-filter is of 5 to 10 times wider bandwidth than the loop filter, and the closed loop architecture can be treated as a second order PLL.

D. Carrier Recovery

The proposed CR loop has three operation stages [6]. In the acquisition state, the two-fold CR loop is configured to be the prior CR loop, a modified Costas loop shown in Eq.(4), and starts at time A as shown in Fig.6(a). At time B, the two-fold CR loop is switched to the posterior CR loop when the adaptive equalizer converges roughly. Meanwhile, the DC component of the prior CR loop output, ω_{dc}, is extracted and added to the initial frequency of the numerical-control-oscillator (NCO). The posterior CR loop begins with the NCO frequency at $\omega_0 + \omega_{dc}$ as shown in Fig.6(b). In this tracking state, the equalizer operates in an ISI-affected converged mode and the CR performs decision-directed maximal-likelihood (DD-ML) phase estimation derived in Eq.(5). At time C, the equalizer fully converges and the CR loop is further switched to a decision-directed minimum-mean-square-error (DD-MMSE) phase estimation shown in Eq.(6). Using the equalized signal, the CR loop provides high jitter suppression in the steady state. The proposed architecture shown in Fig.7 is hardware efficient since the three-staged operation and the two-fold architecture share most of the circuit functions.

$$\hat{\phi}_{ML} = -Im\{sgn(\hat{I}^*) \cdot y_l(t)\} \tag{4}$$

$$\hat{\phi}_{ML} = -Im\{\hat{I}^* \cdot y_{eq}(t)\} \tag{5}$$

$$\hat{\phi}_{MMSE} = E\{\left[y_{eq}(t) - \hat{I}\right]^* \cdot sgn(y_{eq}(t))\} \tag{6}$$

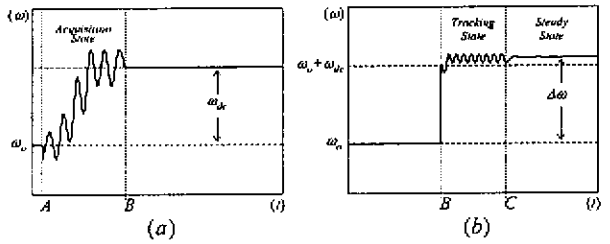


Fig. 6. The behavior of two-fold carrier recovery loop (a) Prior carrier recovery loop (b) Posterior carrier recovery loop [6]

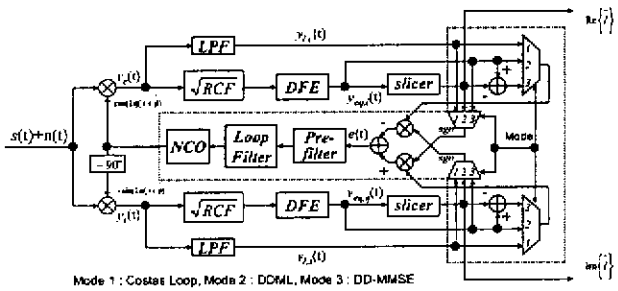


Fig. 7. The proposed architecture to realize the three-staged two-fold carrier recovery loop [6]

III. VLSI Implementation Considerations

Three approaches of the VLSI design are employed. The non-regular functional blocks are written in a high level description language (HDL) and synthesized using standard cell methodologies. Then, the clock trees and control signals are manually synthesized to ensure timing constraints. Finally, the dedicated multipliers and adders are synthesized by CAD tools with specifications of delay, area, styles and power. In addition, a self-test circuit is also built in this chip for automatic verification. The head-end transmitter function is also included for system self testing. Using TSMC $0.35\mu m$ 1P4M CMOS technology, the chip occupies $5.5mm \times 5.5mm$ chip area as shown in Fig.8.

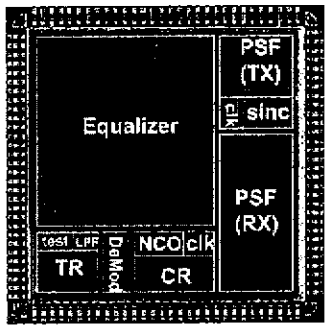


Fig. 8. The layout for $4.035MHz$ Low-IF digital baseband of 64-QAM CATV Modem

IV. Simulation Results

Using the post-layout extracted circuit under 4 process corners, the digital baseband transceiver is simulated in all transmission channel models. The tolerance of the carrier offset in a digital CATV system is $\pm 100kHz$, i.e. $\pm 18,500ppm$ of the symbol rate, as shown in Fig.9. Fig.10 shows the frequency convergence of the TR loop when the symbol timing offset is as large as $\pm 200ppm$. As soon as the receiver enters the steady state, the carrier jitter can be reduced down to $-82dBc@100kHz$ away from the carrier frequency, as shown in Fig.11. Fig.12 shows the mean-square-error (MSE) of the equalized signal. The constellation for decision in the steady state is shown in Fig.13. The acquisition time is $7ms$, which is much less than the requirement of $100ms$ in a CATV system. Table I summarizes the simulation performance of this 64-QAM Low-IF transceiver digital baseband chip.

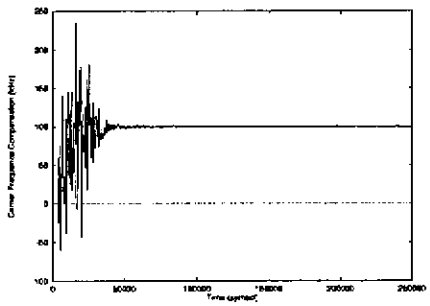


Fig. 9. The $+100kHz$ compensation frequency of the carrier recovery loop

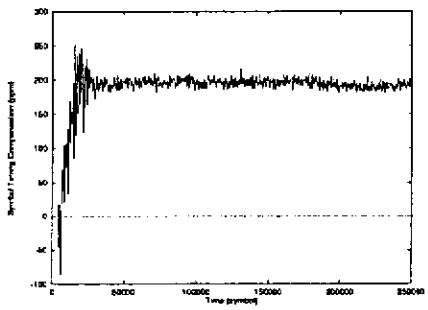


Fig. 10. The $+200ppm$ compensation frequency of the timing recovery loop

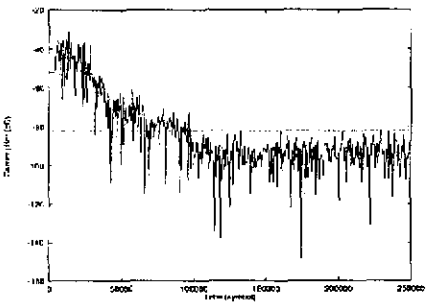


Fig. 11. The carrier jitter performance

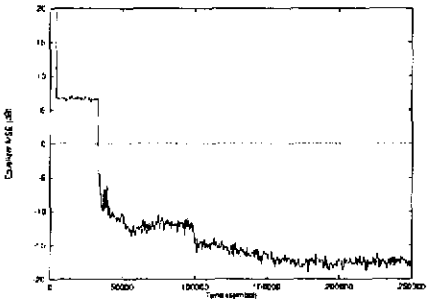


Fig. 12. The MSE of the blind adaptive equalizer

V. Conclusion

This paper presents a hardware efficient VLSI architecture of digital baseband for 64-QAM low-IF digital CATV system with data rate $32.28Mb/s$. Taking the advantage of the proposed multi-staged LMS based blind equalizer, timing offset tolerance $\pm 200ppm$ of the baud rate TR loop and pull-in range better than $\pm 18,500ppm$ with $-82dBc$ jitter suppression of the two-fold CR loop are demonstrated. Both architectures of the adaptive equalizer and the two-phase linear pulse shaping filter contribute the most hardware reduction in this transceiver chip. Since single carrier modulation (SCM) technology is considered one of the main candidates for the Wireless MAN system [3], the proposed architecture, which contains the transmitter, synchronous loops and the programmable blind equalization, can be applied to fixed broadband wireless access networks.

References

[1] Digital Audio-Visual Council, "Lower layer protocols and physical interfaces," *DAVIC 1.2 Specification Part 8, Rev. 4.2*, 1997.

[2] T1E1.4, "Vdsl technical specification, part 2: Technical specification for a single-carrier modulation (scm) transceiver," *Working Group T1E1 (DSL Access), Greensboro, NC*, February 2001.

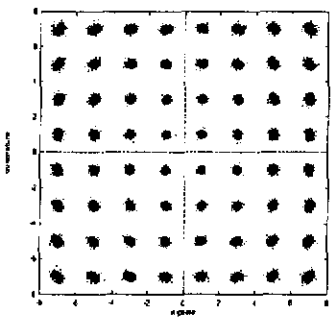


Fig. 13. The signal constellation for decision in the steady state

TABLE I

The 64-QAM transceiver digital baseband simulation performance

Application System	Low-IF digital CATV Modem
System Clock	$43.04MHz$
Symbol Rate	$5.38MHz$
IF Frequency	$4.035MHz$
PSF/LPF Tap Number	39
Anti-sinc Filter Tap Number	9
Equalizer Tap Number	FF:21, FB:19
Timing Recovery Tolerance	$\pm 200ppm$
Carrier Recovery Tolerance	$\pm 100kHz$
Carrier Jitter Performance	$-82dBc@100kHz$
Equalizer SNR Performance	$30dB$
Receiver Acquisition Time	$7ms$
Process Technology	TSMC $0.35\mu m$ 1P4M
Gate Count	280,195
Supply Voltage	3.3V
Power Consumption	1.35W (RX:1.0W)
Chip Size	$5.5mm \times 5.5mm$

[3] IEEE Standard for Local and Metropolitan Area Networks, "Part 16: Air interface for fixed broadband wireless access systems-amendment 2: Medium access control modifications and additional physical layer specifications for 2-11 ghz," April 2003.

[4] C. F. Wu, M. T. Shiue, C. C. Huang, and S. J. Jou, "Qam/vsb dual mode equalizer design and implementation," *AP-ASIC*, pp. 323 -326, Aug 1999.

[5] T1E1.4, "Very-high bit-rate digital subscriber lines (vdsl) metallic interface, part 1: Functional requirements and common specification," *Working Group T1E1 (DSL Access), Vancouver, Canada*, February 2002.

[6] C. C. Chang, C. C. Lin, M. T. Shiue, and C. K. Wang, "A wide pull-in range fast acquisition hardware-sharing two-fold carrier recovery loop," *ISCAS*, vol. 4, no. 12, pp. 358–361, May 2001.

[7] L. K. Tan, J. S. Putnam, and F. Lu, "A 70-mb/s variable-rate 1024-qam cable receiver ic with integrated 10-b adc and fec decoder," *IEEE J. Solid State Circuit*, vol. 33, no. 12, pp. 2205–2218, Dec 1998.

[8] M. T. Shiue, *Transceiver VLSI Design for High Speed Local Access Modems*, Ph.D. thesis, National Central Univ., Taiwan R.O.C., Sept 1998.

[9] G. Picchi and G. Prati, "Blind equalization and carrier recovery using a stop-and-go decesion-directed algorithm," *IEEE Trans. Commun.*, vol. COM-35, no. 9, pp. 877–887, Sept 1987.

[10] K. H. Mueller and M. Muller, "Timing recovery in digital synchronous data receivers," *IEEE Trans. Commun.*, vol. COM-24, pp. 516–531, May 1976.

A Dual-band IEEE 802.11a/b/g Receiver Front-end Using Half-IF and Dual-Conversion

Chun-Chih Hou, Ching-Chi Chang, and Chorng-Kuang Wang

Department of Electrical Engineering and Graduate Institute of Electronic Engineering,
National Taiwan University, Taipei, Taiwan 10617, R.O.C.

Abstract—This paper presents a dual-band receiver front-end architecture which combines the dual-conversion [4] and half-IF techniques for IEEE 802.11a/b/g. The proposed architecture receives dual-band signal with single receiver chain to reduce components count such as the 2nd down-conversion mixer and VCO. The required LO frequencies of the dual-band application can be synthesized by only one VCO in combination with a divide-by-four circuit. An LC tank aided mechanism of mixer is also proposed to deal with the flicker noise. The core portion of the front-end receiver has been fabricated in 1P6M 0.18μm CMOS technology. The delivered gain, noise figure, and IIP3 are 20 dB, 3.5 dB, and −13 dBm, respectively simulated. The chip occupies an area of 1.21×1.46mm^2 and the power consumption is 24mW under the supply voltage of 1.8V.

Index Terms—WLAN, LNA, mixer, half-IF, flicker noise, dual-conversion, dual-band.

I. Introduction

In the market of wireless local area network (WLAN), multi-standard transceivers have been the trend. The IEEE 802.11b standard at the 2.4 GHz ISM band provides data rates up to 11 Mbits/s [2]. The 802.11a standard at the 5GHz U-NII bands provides data rates up to 54 Mbits/s using OFDM modulation [1]. Released in 2003, the 802.11g standard, operated at the same band of 802.11b, uses the modulation method of 802.11a and provides data rates up to 54 Mbits/s [3]. Many products in the market can support each of the standards simultaneously, and the employed transceiver architectures are usually the direct-conversion ones.

As for the published dual-band architectures, most of them receive the two-band signal by two signal paths and the direct-conversion architecture is mostly adopted, too [5-7]. In this paper, a new architecture is proposed that can receive signal at two-band, 5.8GHz and 2.4GHz, with the same receiver chain. Due to the elaborate frequency conversion planning, the troublesome image problem is relaxed in this design. The organization of this paper is as follows. The receiver architecture is described in Section II. In Section III, the circuit topologies meeting the requirements of this architecture are presented. Finally, the simulation results and conclusions are given in Section IV and V, respectively.

II. Receiver Architecture

The proposed receiver front-end architecture is shown in Fig. 1. The components inside the dashed frame can be implemented into a single chip, and those outside the frame are the lumped components, which are switch, diplexer, and band-select band-pass filters. Two band-pass filters instead of one are used in order to provide sufficient rejection of the out-of-band interferences for each band. When the desired signal is at 5.8 GHz, the frequency conversion method, shown in Fig. 2., is used [4]. The 3.465GHz image is 2 GHz far away from the desired signal, and it can be filtered out by the band-pass nature of the on-chip amplifiers. The first local oscillator (LO) frequency, which is 4.62 GHz, is four times of the second LO, and thus they can be synthesized by a single voltage controlled oscillator (VCO) and a divide-by-four circuit. Because the VCO frequency is different from the signal frequency, the LO pulling problem will not occur.

This dual-band receiver architecture can also be switched to half-IF conversion to receive 2.4GHz RF signal, as shown in Fig. 3. The utilized 1.2 GHz LO frequency can be obtained by the same circuits composed of one VCO with 4.8-5 GHz tuning frequency and a divide-by-four circuit. As shown in Table I, the VCO frequency range of the two bands is close, so a VCO with wide tuning range can deliver the required LO frequencies for dual-band application. At the same time, the 90° phase shifter can be eliminated due to the utilization of dividers. Another benefit of this dual-band receiver is that the IF frequencies are close for this dual-band signal, so there is an ease to implement the circuits after the first down-conversion. As for the image signal around the zero frequency, the attenuation contributed by amplifiers and lumped components such as antenna and band-select filter is very large, and thus the image rejection ratio can be as high as 60dB [8]. Two problems that should be taken into account are the upconversion of flicker noise to IF frequency and the DC-offset that comes from LO-to-IF feedthrough. The circuits proposed in the next section will deal with the first problem. For the DC-offset problem, a high-pass filtering offset-cancellation circuit can solve it [8]. The SNR degradation due to the filtering is not significant because for IEEE 802.11a/g, the OFDM modulation leaves the center sub-carrier empty, and for IEEE 802.11b, spread-spectrum modulation is used.

This program is supported by the National Science Council (NSC), Taiwan, R.O.C.

III. Circuit Topologies

The only RF circuits that need to be carefully designed for our receiver architecture are the LNA and the first mixer. The design of other components is the same as the single band ones. For example, the VCO required is just a wide-tuning-range one. Besides, the second mixer and the amplifiers at baseband have single-band nature because the IF bandwidth is not very wide, which has been shown in the last section.

The LNA schematic is shown in Fig. 4. Because the LNA is designed to have the ability to receive and amplify the two-band signal, it has a load that contains two inductors and two capacitors [9]. The function of this LC tank can be comprehended as follows. When the frequency is low, the L and C in series are capacitive in effect. When the frequency is high, they are inductive in effect. Thus, the LC tank has two peaks in the frequency domain. The input matching of the LNA is achieved by off-chip microstrip lines, which are combinations of some open-stubs on the PCB board.

The mixer schematic is shown in Fig. 5. It is a gilbert-cell type mixer with source degeneration resistors to enhance the linearity. The mixer is a wide-band device in nature, so signal can drive the input ports and LO ports directly. Also, because signal frequency at the output of the mixer is about 1 GHz for both the two-band operation, the output port is single-band in nature and thus only one inductor is used, instead of the two-inductor case in LNA.

The mixer in [8] prevents the up-conversion of flicker noise of the transconductance stage to IF, but ignores the flicker noise of the switching pairs. The mixer proposed here uses an LC tank to achieve the filtering function. The resonant frequency of the LC tank is at the signal frequency such that it blocks the signal from passing through it. As for the flicker noise around zero frequency, it sees a short because the LC tank has an inductor in parallel. The circuit can thus be reduced to the half circuit as shown in Fig. 6. The topology of mixer is destroyed, and the flicker noise in neither the transconductance stage nor the switching stage will be up-converted to IF.

The linearity of amplifiers is an important consideration for wireless communication. When the signal received by the antenna is not very large, the third-order inter-modulation signal located at the signal frequency band because of the adjacent-channel interferences and the non-linearity of amplifiers must be low enough in order not to degrade the SNR. For WLAN, the IIP3 of the receiver chain should be at least −26dBm to meet this requirement. Meanwhile, when the input signal is large, which is at most −30dBm, −10dBm, and −20dBm for IEEE 802.11-a/b/g respectively, the amplifiers may be saturated. Moreover, for IEEE 802.11a/g that have 52 subcarriers, another add-on power back off of 10dB is needed to accommodate the large peak-to-average ratio (PAR). Thus, a gain-adjusting mechanism is necessary for the front-end amplifiers. A variable resistor can be placed in parallel with the load inductor, and in this case it is replaced by a digitally switched PMOS, as shown in Fig. 7. Discrete gain steps can also be achieved by many of such PMOS placed together.

IV. Simulation Results

The input return loss of the LNA is shown in Fig. 8., in which two bands matching is achieved, with bandwidth larger than 200 MHz in ISM band and larger than 300 MHz in the upper U-NII band. The S21 of LNA is shown in Fig. 9. The gain at each band is about 10 dB. The image rejection ratio (IRR) provided by this amplifier is larger than 60 dB for upper U-NII band, and 30 dB for ISM band. The noise figure of the LNA and mixer operated at 2.4 GHz when no LC tank is inserted in the LO switching pair is shown in Fig. 10. The up-converted flicker noise has a significant influence on the noise figure. After the LC tank is inserted, the influence from the flicker noise vanishes, as shown in Fig. 11.

Table II is the summary of the simulation results of the LNA and mixer. From the comparison between the two circuits with and without LC tanks to block the up converting of flicker noise, it can be observed that the noise figures for both bands improve. However, the conversion gain and IIP3 is not the case. This is because that the gain depends on the quality factor of the LC tanks and the linearity depends on the harmonic termination [10], which is not the same for the two bands.

The power consumption of the core circuit is 24mW. It has been fabricated by TSMC 1P6M 0.18μm CMOS. The chip area is 1.21×1.46 mm^2 and its microphotograph is shown in Fig. 12.

IV. Conclusion

A dual-band receiver front-end architecture is proposed in this paper. It combines the advantage of dual-conversion and half-IF receivers to provide IRR of 60dB and 30dB for U-NII and ISM band signal respectively. The IF frequency range of 1.14 to 1.24GHz relax the component requirement after the first downconversion. The dual-band LNA and mixer have been designed to deliver a conversion gain of 20dB and IIP3 of −14dBm with a power consumption of 24mW. The troublesome flicker noise problem in half-IF conversion has been solved by a circuit topology proposed in this paper.

Acknowledgement

The authors would like to thank to the technical support by Chip Implementation Center (CIC), Taiwan, R.O.C.

References

[1] IEEE Std 802.11, Wireless LAN Medium Access

Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band, Sept. 1999.

[2] IEEE Std 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer Extension in the 2.4 GHz Band, Sept. 1999.

[3] IEEE Std 802.11, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 4: Further Higher Data Rate Extension in the 2.4 GHz Band, June 2003.

[4] D. Su, M. Zargari, P. Yue, S. Rabii, D. Weber, B. Kaczynski, S. Mebta, K. Singh, S. Mendis and B. Wooley, "A 5GHz CMOS Transceiver for IEEE 802.11a Wireless LAN," in *IEEE Int. Solid-State Circuits Conf. Tech. Dig.*, Feb. 2002, pp. 70-405.

[5] M. Hotti, J. Kaukovuori, J. Ryynanen, K. Kivekas, J. Jussila, and K. Halonen, K, "A Direct Conversion RF Font-end for 2-GHz WCDMA and 5.8-GHz WLAN Applications," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 45-48, June 2003.

[6] B. U. Klepser, M. Punzenberger, T. Ruhlicke, and M. Zannoth, "5-GHz and 2.4-GHz Dual-band RF Transceiver for WLAN 802.11a/b/g Applications," *IEEE Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 37-40, June 2003.

[7] M.Y. Wang, R.R.-B. Sheen, D.T.-C. Chen, and R.Y.J. Tsen, "A Dual-band RF Front-end for WCDMA and GPS Applications," in *Proc. 2002 Int. Symp. Circuits and Systems*, vol.4, 2002, pp. 113-116.

[8] B. Razavi, "A 5.2-GHz CMOS Receiver with 62-dB Image Rejection," *IEEE J. Solid-State Circuits*, vol. 36, pp. 810-815, May 2001.

[9] H. Hashemi, H and A. Hajimiri, "Concurrent Multiband Low-noise Amplifiers - Theory, Design, and Applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, pp. 288-301, Jan. 2002.

[10] J. S. Fairbanks and L. E. Larson, "Analysis of optimized input and output harmonic termination on the linearity of 5 GHz CMOS radio frequency amplifiers," *Proceedings of Radio and Wireless Conference, 2003, RAWCON '03*, pp. 293-296, Aug. 2003.

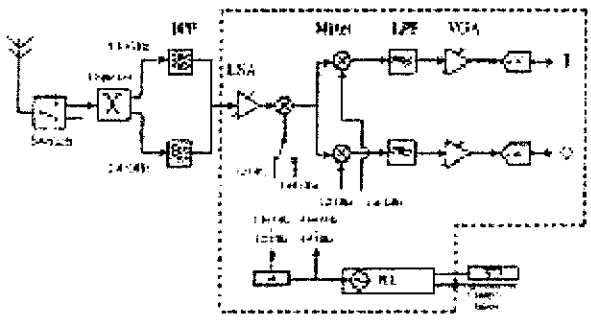


Fig. 1. Block diagram of the receiver.

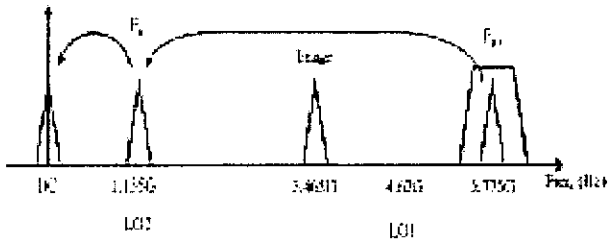


Fig. 2. Frequency planning of 5 GHz signal.

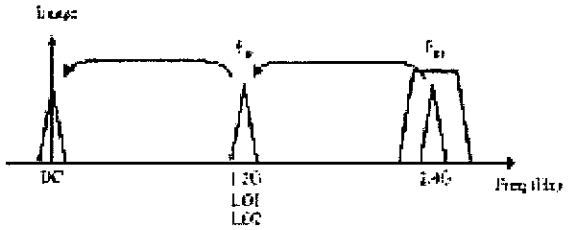


Fig. 3. Frequency planning of 2.4 GHz signal.

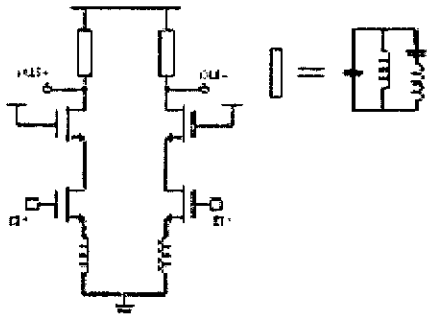


Fig. 4. Schematic of the LNA.

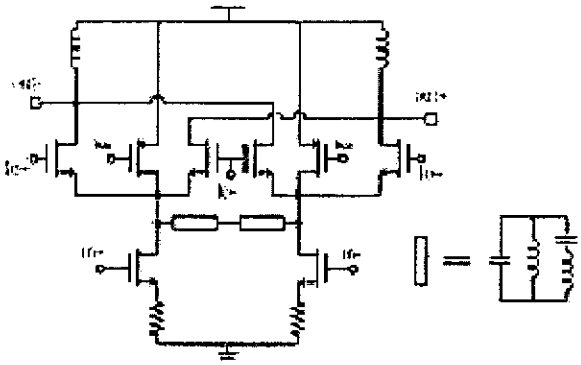


Fig. 5. Schematic of the first mixer.

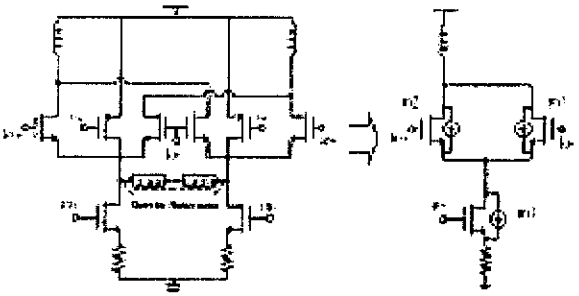


Fig. 6. Filtering mechanism for flicker noise.

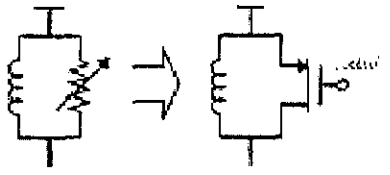


Fig. 7. Gain-adjusting circuit.

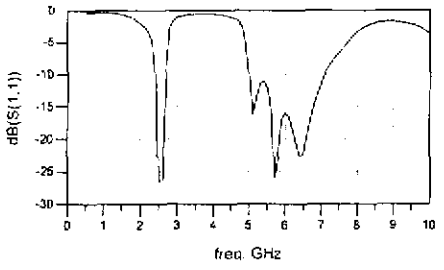


Fig. 8. Input return loss of the LNA.

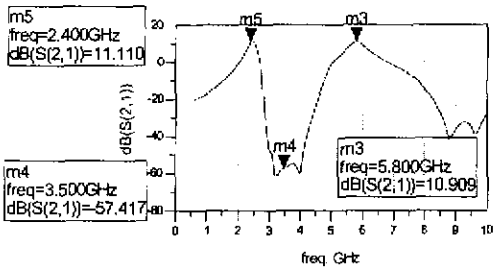


Fig. 9. S21 of the LNA.

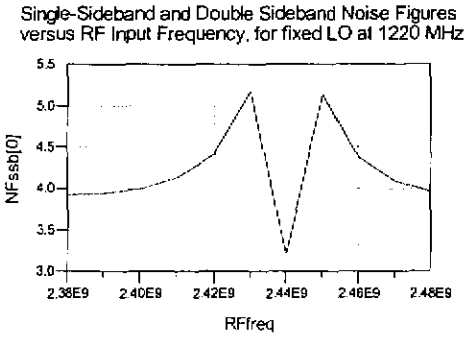


Fig. 10. Noise figure at the 2.4 GHz band, without LC tank inserted.

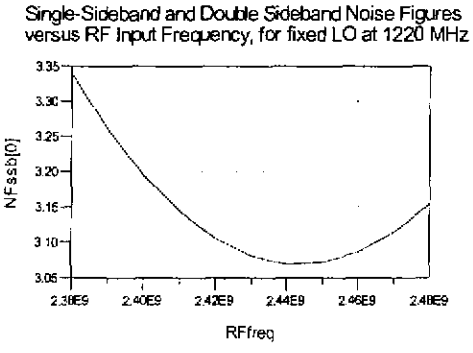


Fig. 11. Noise figure at the 2.4 GHz band, with LC tank inserted.

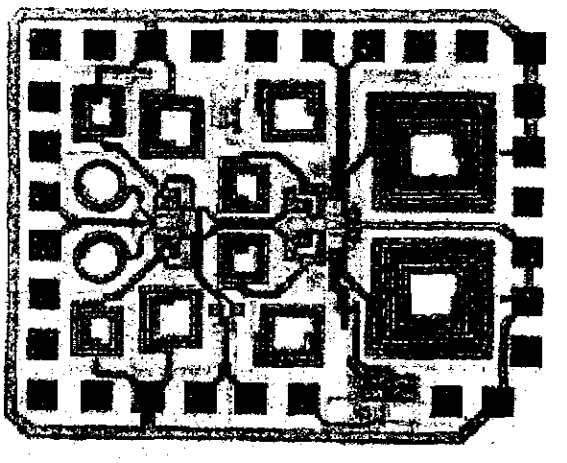


Fig. 12. Chip photograph.

TABLE I
FREQUENCY RANGE OF EACH STAGE

Frequency Band	Frequency Range	VCO Range	IF Range
Upper U-NII Band (GHz)	5.725-5.825	4.58-4.66	1.145-1.165
ISM (GHz)	2.40-2.4835	4.8-4.967	1.2-1.242

TABLE II
SIMULATED LNA AND MIXER PERFORMANCE.

Parameters		LNA+Mixer without flicker noise filtering circuitry	LNA+Mixer with flicker noise filtering circuitry
Supply Voltage		1.8 V	1.8 V
Power Consumption (without biasing and output driver circuits)		24 mW	24 mW
2.4 GHz	Gain	21.3 dB	20 dB
	NF	>4.5 dB	3.1 dB
	IIP3	-16.5 dBm	-13.4 dBm
	LO Power	-2 dBm	-2 dBm
5.8 GHz	Gain	16.4 dB	18.8 dB
	NF	4.25 dB	3.55 dB
	IIP3	-8.8 dBm	-11.4 dBm
	LO Power	-2 dBm	-2 dBm
Technology		TSMC 1P6M 0.18μm CMOS	
Chip Area			1.21×1.46 mm^2

附錄五

NSC IT Program Meeting Minutes

Date	2004 / 2 / 13 .
Time	10:00 -- 12:00
Location	台大電機新館 R 142
Attendee	教授: 吳安宇 吳曉光
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : the front-end architectures of dual band multi-standard signal direct conversion and multi-conversion Comment: 802.11a/b/g 業界已有產品，爲何還要做? Comment: 研究價值在哪? Comment: 既然direct conversion較好，那爲何要選用multi-conversion來做? 子計畫二(Prof. Chiueh) : 802.11n preamble design the same short preamble and double long preamble Comment: antenna數目會在standard上規定嗎?這個數目會和long preamble 長度有關嗎? Comment: 二個天線的位置，距離等等的影響 802.16a project的進度及schedule報告 Comment: 台灣環境與歐美測出來的通道效應不同，如何改? 子計畫四(Prof. Chou) : 估出power，要送四月0.25um的晶片 比較用dc_shell及nanosim測出的power大小及所需時間差 Comment: 分成三大區塊模擬時，如何互相溝通? Comment: 用了三個DSP，cost會不會太大了? 子計畫六(Prof. Liao) : reliable multicast demo with and without reliability的效果 Comment: 可用在電話上嗎?或是不要互動的video stream? 子計畫七(Prof. Wu - NCU) : embedded application-MP3 player on dual processor platform OMAP1510 ARM+DSP+IDE tool Comment: 有沒有一個OS是可以同時控制ARM及DSP的? Others: 開學後爲隔週五報告 國科會期末計畫報告 子計畫三及五的同學，下次務必出席報告進度 下次meeting在台大(2/27)(五)

下次meeting(2/27)在台大

NSC IT Program Meeting Minutes

Date	2004 / 2 / 27 .
Time	10:00 -- 12:00
Location	台大電機新館 R 142
Attendee	教授: 汪重光 周世傑 蘇朝琴 廖婉君
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : half-IF receiver and architecture image rejection ratio, LO-IF feed-through and flicker noise Propose mixer architecture for solving flicker noise problem Comment: 上次送的chip量的如何?下次報告一下 子計畫二(Prof. Chiueh) : coarse symbol, fine symbol boundary detection frequency and timing synchronization design preamble detection by SNR threshold Comment: frequency and timing synchronization的順序，估測範圍和精確度 Comment: moving average如何決定長度及smooth到何種程度? Comment: frequency filter的效應及如何設計? Comment: 爲何delay spread不是對稱的? 子計畫三(Prof. Wu - NTU) : AES IP architecture and Synthesis EASY platform Dual mode AES design→ counter mode and CBC mode 60-70% of layout is CPU (AES) Comment: chip summary的performance列太少了 Comment: 這麼大的chip，經費來源和可過的機會有多少? 子計畫四(Prof. Jou) : 利用3個DSP做OFDM的receiver 因爲有多個DSP，所以寫assembly code時要注意長度相同 Comment: 三個DSP分別做什麼? 子計畫五(Prof. Su) : 做 802.11a/b/g 的 testing 找業界現有的產品及量測方式，特別是 RF 的部份 做 table 來比較各種參數及系統 Comment: RF 有很多種不同架構，每種都能測嗎? Comment: 這種 testing 與傳統的方式有何差別?優點爲何? Comment: 整個模擬環境建好了嗎?可以參數化嗎? Others: 新一年度的經費已經下來了，感謝大家過去一年的幫忙 若是paper與此計畫相關，發表時記得要寫上"國科會" 新進儀器的training時間及地點，改天會發信通知大家 下次meeting在台大(3/12)(五)

下次meeting(3/12)在台大

NSC IT Program Meeting Minutes

Date	2004 / 3 / 26 .
Time	10:00 -- 12:00
Location	台大電機新館 R 142
Attendee	教授: 汪重光 闕至達 吳安宇 吳曉光 周世傑
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : 加上high-pass filter消除LO-IF feedthrough Comment: high-pass filter一個order夠不夠? Ans : 夠 Comment: high-pass filter的corner frequency需要很準嗎? Ans: 不用 子計畫二(Prof. Chiueh) : 用兩倍長度的long preamble來解決channel estimation 最後可能不做低速的mode來節省硬體 可能用4x4來達到100Mbps的function requirement Comment: 圖若是uncoding的話要顯示出來 Comment: 4x4面積是2x2的幾倍? Ans:約4倍,約是linear的關係 子計畫三(Prof. Wu - NTU) : 上次slot line的問題非skin effect，而是由製程角度來看的問題 加入metal slot來解決error。 在 CVD(chemical vapor deposition) 時線太寬，造成電流不均 Comment: slot是否可有equivalent electric field circuit? 子計畫五(Prof. Su) : Build RF environment：single stage & double stage,以transmitter爲主(ideal) Simulation：gain compression. Comment: 是否要來現有的example來design? Comment: testing有哪些突破？ Comment: simulation show的圖對不對?應該要做一些比較的圖。 Comment: simulation error是否爲取的點數太少或 tool 的問題? 子計畫六(Prof. Liao) : 目標在QoS demand service達到其BW的需求。 用SFQ或WFS來決定傳送哪個封包。 Comment: draft是第幾版? 子計畫七(Prof. Wu - NCU) : MPEG4-Rate Control. Non-zero ratio和video bit ratio呈線性。

下次meeting(4/9)在台大

NSC IT Program Meeting Minutes

Date	2004 / 4 / 9 .
Time	10:00 -- 12:00
Location	工綜大樓B04教室
Attendee	教授: 闕志達
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : Half-IF + Dual Conversion (LNA + Mixer) Comment: 爲什麼選擇half IF? Ans : 可以用相同的LO 多了 flicker noise 的影響,加新的電路解決 Comment: 要有和別人比較的結果 Comment: 有沒有要做後面IF的部分 子計畫二(Prof. Chiueh) : Symbol boundary detection and CFO acquisition Comment: 檢查maximum delay spread → maximum excess delay Symbol boundary : worst case有兩段64點short preamble(Coarse)和完整(兩段128)的long preamble(Fine) 由模擬結果求出需要30套 correlator bank 子計畫三(Prof. Wu - NTU) : 把ASE放進platform core size: 4438 x 3105 μm^2 die size: 5 x 3.5 mm^2 子計畫四(Prof. Jou) : Statement coverage, branchcoverage Frequency offset lock loop OFDM receiver系統模擬尚未完成 Comment: 三個DSP的運算量 子計畫六(Prof. Liao) : Throughput analysis 新的架構內contention window在collision後必需乘兩倍 只能求數值解 Simulation results 子計畫七(Prof. Wu - NCU) : Throughput simulation results (考慮 random loss) Congestion window v.s. random loss Maximum through v.s. random loss

下次meeting(4/23)在台大

NSC IT Program Meeting Minutes

Date	2004 / 4 / 23 .
Time	10:00 -- 12:00
Location	台大電機新館 R 142
Attendee	教授: 蘇朝琴 汪重光 闕至達 吳曉光 周世傑
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : Dynamic Range of Rx chain. Trade-off between LPF and ADC Comment: 把LPF spec改的實際一點 Comment: show一下simulation result 子計畫二(Prof. Chiueh) : 加進outer的BER simulation. 在subcarrier加入spreading. V-blast algorithm Comment: delay spread 也許不用那麼大 Comment: 是否要多加一個error correction的機制 子計畫三(Prof. Wu - NTU) : AES IP core design Clock tree issue in post-simulation Comment: 哪裡low voltage, 哪裡low power? 要有數據 子計畫四(Prof. Jou) : Final place and route – clock tree balance 子計畫五(Prof. Su) : 目前還在找合適的電路 Comment: 先決定用什麼architecture. 子計畫六(Prof. Liao) : 802.11e Delay and Jitter Analysis comment: SIFS的時間? Ans: 16 μ s Simulation 和 analysis match High class會影響到 low class 子計畫七(Prof. Wu - NCU) : MPEG4的分析 Adaptive rate control Comment: improvement的目標是多少? Comment: 可以多show一些和別人比較的結果

下次meeting(5/21)在台大

NSC IT Program Meeting Minutes

Date	2004 / 5 / 21 .
Time	10:00 -- 12:00
Location	台大電機新館 R 142
Attendee	教授: 蘇朝琴 汪重光 吳曉光 周世傑
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : Review of IEEE 802.11a Rx Chain Budget. Testing of 5.7GHz dual band LNA & Mixer Trade-off between LPF and ADC Comment: 把chip不能work的理由找出來，用simulation證明。 子計畫三(Prof. Wu - NTU) : Hierarchical Verify Flow Gate clock 可以 low power. 但要小心clock skew, 且fault coverage較低 子計畫五(Prof. Su) : Class C PA Comment: 有沒有spec可以知道simulation對不對?. 子計畫七(Prof. Wu - NCU) : simulation of WPGMCC 10M fairness Comment: 傳送的速率有variation,可以容忍多少? Comment: 指出和別人比較好在哪裡? 下次報告：Group 2, 4, 6

下次meeting(6/5)在台大

NSC IT Program Meeting Minutes

Date	2004 / 6 / 4 .
Time	10:00 -- 12:00
Location	台大電機新館 R 142
Attendee	教授: 汪重光 吳安宇
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫二(Prof. Chiueh) : Transfer C++ code to sample-base format. Forward Error Correcting Blocks Comment: show出tracking loop多久可以track到。 子計畫四(Prof. Jou) : Stream Based I/O transfer between DSP Comment: 在最大的Block內不要再加logic control Comment: power overhead 是多少? 子計畫六(Prof. Liao) : 改進目的: robust and scalable, one mechanism scalable in both greoup memver and group size Distributed greedy algorithm: remove and move.. 下次報告：Group 1, 3, 5, 7

下次meeting(6/18)在台大

NSC IT Program Meeting Minutes

Date	2004 / 6 / 18 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一(Prof. Wang) : Dual band RF receiver summary 子計畫三(Prof. Wu - NTU) : AES hardware architecture System block diagram of CCMP. 把額外的component拿掉，面積縮小爲4.8mm x 3.3mm FPGA，無法把整個platform放上去 子計畫五(Prof. Su) : 把LNA和Mixer拆開來模擬 Comment: 去看看工業界如何test. 子計畫七(Prof. Wu - NCU) : 同時有很多node有loss,看其對performance的影響 看loss和其他bottleneck對performance的影響 下次報告：Group 2, 4, 6

下次meeting(6/28)在台大

NSC IT Program Meeting Minutes

Date	2004 / 6 / 29 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光 吳曉光 吳安宇
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫二 z(Prof. Chiueh) : High throughput 2x2 MIMO OFDM transceiver Dual Mode: SIMO and MIMO 在802.11n 定義的data rate是指在MAC layer要達到100Mb/s, 因此在physical layer 至少要120Mb/s的data rate, 可能方法: code rate由3/4 → 7/8 子計畫四 曹亞嵐(Prof. Jou) : Instruction set verification 在test時把clock分開, 不能有gated clock 改進一些IO來改善迴轉率 Comment: 最後要做啥?還要多久完成? Comment: carrier phase tracking loop是如何tracking? Comment: 可以容忍多少晃動? 下次報告：Group 1, 3, 5, 7

下次meeting(7/13)在台大

NSC IT Program Meeting Minutes

Date	2004 / 7 / 13 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一 李尉彰(Prof. Wang) : 802.16a屬於broad band 還是 narrow band? Comment: 將來要用哪一種architecture for multiband? Comment: 2.3/3.-- /5.—GHz有什麼挑戰?有什麼問題? Comment: 焦點在 2 — 11GHz or 60GHz? 子計畫三 汪威(Prof. Wu - NTU) : Comment: standard要多少encryption? Comment: 802.16的security,不同standard是否有不同? Comment: 512btit的encryption的hardware engine價值? Line up 802.16a 子計畫七 楊惟仁(Prof. Wu - NCU) : Dynamic resource management adaptive video stream. Comment: smoothing details? Comment: 有沒有考慮P2MP? 下次報告：Group 2, 4, 6

下次meeting(7/27)在台大

NSC IT Program Meeting Minutes

Date	2004 / 7 / 27 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光 吳曉光 吳安宇 廖婉君
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫二 楊尚融(Prof. Chiueh) : 加入Guard interval length detection. 加入 channel gain update. Comment: 如何知道channel gain是否會收斂? Comment: coarse和fine的symbol boundary的error是多少? 子計畫四 魏庭楨(Prof. Jou) : Status report 子計畫六 劉得煌(Prof. Liao) : DOCSIS MAC Control Mechanism 802.16是從1.1的版本改過來，因爲有QoS 下次報告：Group 1, 3, 5, 7

下次meeting(9/7)在台大

NSC IT Program Meeting Minutes

Date	2004 / 9 / 7 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光 闕志達 吳曉光
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一 李尉彰(Prof. Wang) : 802.16a link budget Comment: IQ imbalance和DC offset 有沒有spec? 子計畫三 汪威(Prof. Wu - NTU) : Symmetric key (fast), Public key(slow) Public key(three categories) 1. encryption/decryption 2. digital signature(optional) 3. key exchange Comment: 如何決定要用software還是hardware? Ans:以速度和硬體來考量 Comment: 做好後要和誰比較?如何比較? 子計畫七 Ben Tsai(Prof. Wu - NCU) : Comment: 目的是要達到如何的performance? 希望改進TCP的protocol使RTT的performance影響降低 下次報告：Group 2, 4, 6

下次meeting(9/21)在台大

NSC IT Program Meeting Minutes

Date	2004 / 9 / 21 .
Time	12:30 -- 14:30
Location	台大電機新館 R 124
Attendee	教授: 闕志達 吳曉光 吳安宇
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫二 丁姿吟(Prof. Chiueh) : Comment: Channel model的delay spread是多少? Comment: SNR達到的PER是否合理，夠不夠好? 802.11a 可能改用LDPC code或turbo code 子計畫六 陳一強(Prof. Liao) : IEEE 802.14 v.s. DOCSIS(win) Comment: DOSIS如何和802.16 link? 下次報告：Group 1, 3, 5, 7

下次meeting(10/21)在台大

NSC IT Program Meeting Minutes

Date	2004 / 10 / 21 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 吳安宇 周世傑
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一 李尉彰(Prof. Wang) : Multi-band transceiver for 802.11a/b/g, DCS and GSM 重點在用一顆LNA去選所要的頻率 子計畫三 游自強(Prof. Wu - NTU) : Introduction to LDPC codes Parity-check matrix Very sparse parity-check matrix Irregular LDPC 在code length 10^7可非常接近shannon limit 子計畫五 Hung Kai(Prof. Su) : Single tone test Comment: 用較寬頻的訊號去測 子計畫七 Wei-Li Chang(Prof. Wu - NCU) : QoS : Unsolicited grant service Real-time polling service Non-real time polling service Best effort service Comment:是否要開始做? 下次報告：Group 2, 4, 6

下次meeting(11/4)在台大

NSC IT Program Meeting Minutes

Date	2004 / 11 / 4 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光 闕志達 吳曉光 廖婉君
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫二 雷貽晴(Prof. Chiueh) : Comment: IFFT的bit數可否少一點? Comment: 找一些和別人比較的資料? Comment: AGC的range沒那麼大 子計畫四 (Prof. Jou) Comment: Power consumption 如何? Multi-rate, multi-stage IFIR design 子計畫六 劉得煌 (Prof. Liao) : 802.16a : (1) support ARQ (2) support both PMP and Mesh mode Comment: 如何看performance? Ans: latency, throughput, overhead Comment: 每個station要提供多少node? Ans: 約200~600 Comment: 802.16e和802.20差在哪? Ans: 訴求不同, 對象相同 下次報告：Group 1, 3, 5, 7

下次meeting(11/18)在台大

NSC IT Program Meeting Minutes

Date	2004 / 12 / 02 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 吳安宇 闕志達 汪重光 吳曉光
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫一 李尉彰(Prof. Wang) : Contour design methodology 子計畫三 游自強(Prof. Wu - NTU) : LDPC decoding algorithms Sum-Product algorithm performance最好，complexity最大。 LDPC has two constrains: 1. 任兩columns不能有超過一個的1的overlap。 2. 不能有short loop。 Log domain和probability domain的BER Performance差別? No difference 子計畫五 Hung-Kai Chen(Prof. Su) : LNA test Comment: 目的是要做到如何的testing? Comment: 進度到哪? 子計畫七 張偉立(Prof. Wu - NCU) : 802.16, 802.16a 下次報告：Group 2, 4, 6

下次meeting(12/16)在台大

NSC IT Program Meeting Minutes

Date	2004 / 12 / 16 .
Time	10:00 -- 12:00
Location	台大電機新館 R 124
Attendee	教授: 汪重光 吳安宇
Agenda	A. Status Review B. Others
Minutes	Status Review: 子計畫二 丁姿吟 **(Prof. Chiueh) :** Comment: TGn Sync爲什麼會贏? Ans:因爲RF和802.11a相同, 只要改baseband Comment: 天線的限制在哪?performance是否可以無限制增加? 子計畫四 **Tz - Jie Hung (Prof. Jou)** module generator multi-stage比single stage好 comment: 有沒有其他公司有同樣的東西?要收集資料 子計畫六 劉得煌 **(Prof. Liao) :** PMP mode, Mesh mode, hybrid mode Comment: mesh是如何組成? Comment: channel model? Ans: 只要幾個parameter即可 下次報告：Group 1, 3, 5, 7

下次meeting(1/27)在台大