計畫名稱:高效能分散式共用記憶體平行系統

計畫編號 : NSC 90-2213-E-002-039

執行期間 :90年08月01日至91年07月31日

計畫主持人 : 賴 飛 羆 教授

中文摘要

由於 WWW 之廣受歡迎，造成 Internet 流量快速激增，因此如何節省網路頻寬,增加網路的速度也變成大家關心的議題。

由於網路代理伺服器可有效的改善網路壅塞,如何有效的增加網路代理伺服器的效能便是一個很重要的課題。故若能針對磁碟存取做大幅改進時,便能使網路代理伺服器更有效的運作，達到更大的工作效能。所以，我們實作網路代理伺服器以提升效能。

在 UNIX 系統中,使用 raw disk 來儲存資料,在磁碟存取效能上會比一般使用檔案系統較有優勢。因為當我們要對磁碟做寫入的動作時,使用檔案系統寫入時，要先寫到 UNIX buffer cache，再從 UNIX buffer cache 寫到磁碟，而使用 raw disk 可跳過 UNIX buffer cache,直接將資料寫入磁碟中。另外，使用 raw disk 也可省略掉對 i-node 的配置和維護所需花費的時間。故我們嘗試使用 raw disk 來存取物件,以期能改善磁碟存取的效能。

此外，我們使用 Vinum Volume Manager 以 Raid-0 的架構將 5 顆較小的 SCSI 硬碟串接成一顆大的 Vinum disk，除了可解決單顆硬碟容量不足的問題，並可利用 Raid-0 平行存取的優點，達到提升磁碟存取的效率。

更進一步地，我們使用非同步 I/O 來避免 proxy server 在進行磁碟存取時被 block 住的情形，增加平行化的能力，藉以改善其效能。

最後由實驗結果可知,經過我們在磁碟存取上的改善，可使得 peak request rate 相較於原來版本，增加 390 req/sec，即使與最新的 stable 版本相比，亦有 310 req/sec 的增進，在效能上有長足的進步。

由我們的整體研究可知,磁碟存取對於網路代理伺服器的效能而言,的確是一項很重要的影響因素。

英文摘要

Since Internet was developed, the traffic load of Internet has become heavier than before. But the bandwidth is always limited. For the exponential increasing of people surfing the Internet, how we reduce the usage of bandwidth is an important course for nowadays.

In Internet world, the proxy is a good choice for saving network bandwidth and reducing users' waiting time. How to improve the performance of proxy server is a good topic for research. Proxy server needs a large number of disks I/O accesses. If we can reduce the overhead of disk I/O, we may get large performance improvement of

proxy server. So, we establish a project to research this topic.

In most UNIX systems, it is a performance advantage to use raw device files for data storage. Raw devices bypass the UNIX file system cache, thus eliminating the overhead of copying the data from kernel file system I/O buffers to user buffers. Using raw devices also reduces the workload of the "syncer" daemon which flushes out file system dirty pages to disk. By using raw devices, less memory is used for the file system cache and file I/O buffers. This reduces system paging, and also makes more memory available for user and proxy processes, thereby reducing the chance of swapping. So, first, we try to use raw disk to be our cache disk for improving performance.

Second, we use the Vinum Volume Manager to stripe our 5 SCSI disks. That is RAID-0 structure. We can get multi-access disk advantage, and can solve disk space problem. Since SCSI hard disk still has no such large space in a single hard disk.

Furthermore, we use asynchronous I/O to avoid disk access blocking for improving parallelism.

From our test results, the peak request rate of our proxy server is 390 req/sec better than Squid 2.3 stable 4. There is still 310 req/sec improvement to Squid 2.4 stable 6. That is a great improvement.

So, from our research, we find disk I/O is a very important factor for performance of proxy server.

1.簡介

由於 WWW 之廣受歡迎，造成 Internet 流量快速激增，因此如何節省網路頻寬，增加網路的速度也變成大家關心的議題。雖然資料壓縮技術(data compression) 及 加 大 頻 寬 (network bandwidth)為解決此重要問題之有效方法，但以目前之壓縮技術，大約可有效地將資料壓縮成原來的 1/10，但目前此技術已達瓶頸，難有極大突破；而在加大頻寬方面，雖然有寬頻網路之發展，但由於網路多媒體應用(multi-media application)之漸趨流行，造成網路頻寬永遠不足，根據最近之研究，網路代理伺服器(web proxy server)的有效運用可降低網路頻寬，改善網路的傳輸延遲(latency)，為解決網路壅塞的有效方法。

由於網路代理伺服器可有效的改善網路壅塞，如何有效的增加網路代理伺服器的效能便是一個很重要的課題。在眾多因素中，輸出入裝置上的存取是一個很重要的瓶頸，其中尤其以磁碟及網路存取影響最大。磁碟讀寫動作為機械式的，其存取時間大約在 6ms~12ms 之間，較電子式的動態記憶體的 30ns~80ns 慢了許多，故磁碟讀寫動作常成為 web proxy server 系統效能上的瓶頸。若能針對磁碟存取做大幅改進時，便能使網路代理伺服器更有效的運作，達到更大的工作效能。

接續之前對 Squid 物件管理的改善，我們進一步對 Squid 做了以下的改善，並得到不錯的整體效能改進。

2. 以 raw disk 模式來儲存物件

在 UNIX 系統中，使用 raw disk 來儲存資料，在磁碟存取效能上會比一般使用檔案系統較有優勢。因為當我們要對磁碟做寫入的動作時，使用檔案系統寫入時，要先寫到 UNIX buffer cache，再從 UNIX buffer cache 寫到磁碟，而使用 raw disk 可跳過 UNIX buffer cache，直接將資料寫入磁碟中。另外，使用 raw disk 也可省略掉對 i-node 的配置和維護所需花費的時間。

在使用 raw disk 來儲存物件時，有下列一些事項需注意：

(1)最前面 10KB 需保留給系統儲存資料。

(2) 要注意避免發生 partitions overlapping。

(3)做備份時，要記得連同 raw disk 部分備份。

(4)使用 raw disk 前，必須做完整的備份動作。

在 raw disk 的使用上，最大的缺點在 partition 的大小沒有辦法動態調整，如果 partition 滿了，便需要再重新分割一個更大的 partition 供存取使用。另外，raw disk 的使用，對於使用者來說，是感覺不到任何差異的，使用者並不會知道系統使用的是一般的檔案系統或者是 raw disk。

3. 使用 Vinum Volume Manager

Vinum Volume Manager 是一套實作在 FreeBSD 中的開放軟體，透過 Vinum Volume Manager，我們可獲得以下優點：

(1)硬碟容量可彈性擴充

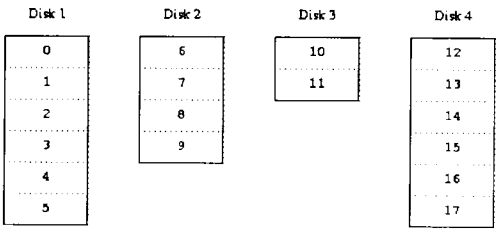
在我們的架構中，採用一個大檔案來儲存所有的物件，所以我們需要一個很大的硬碟來使用，尤其，若要使用速度較快的 SCSI 硬碟，單顆硬碟的容量更是有限，若透過 Vinum Volume Manager 則可彈性的擴充硬碟容量，可以將幾顆相同大小的小硬碟，併成一顆較大的硬碟。

舉例而言，以我們目前網路代理伺服器的設定，cache disk 大小為 35GB，我們只要使用 5 顆 9GB 的 SCSI 硬碟即夠使用，可達經濟實用之效果。

(2)透過多顆硬碟同時存取來增進效能

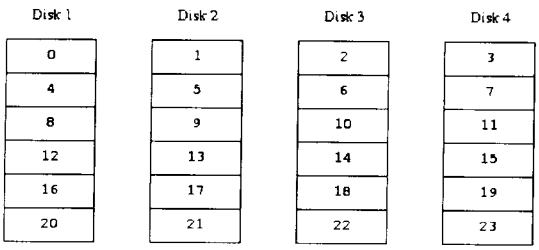
在磁碟的存取上，傳動裝置首先要作磁頭定位，接著要等到第一個要運作的 sector 移到讀寫頭下面。然後才能開始作資料傳輸。以上這些動作都是機械式動作，與傳輸時間比較起來，要花費的時間長很多，往往這些就是磁碟在存取上的瓶頸。

對於上述的問題，傳統的解決方式大部分偏向使用多個較小硬碟替代一個大的硬碟來作存取。每一個小硬碟都可以獨立的作磁頭定位及資料傳輸的動作，如此，便可以有效的增加 throughput。其存取順序如圖(一)：



圖一

上述方法雖可改善效能，但並無法保證每顆硬碟都可被平均存取，若能保證每顆硬碟均可被均勻存取，則效能上將會有更大的改善。此即 RAID-0 之架構，其存取順序如圖二：



圖二

利用 Vinum Volume Manager 的 stripe 功能，即可達到 RAID-0 架構，透過多顆硬碟的均勻存取，可降低因磁碟機械式動作的時間延遲而造成之瓶頸。

4. 使用 Asynchronous I/O

磁碟的讀寫動作可分為 Synchronous I/O 及 Asynchronous I/O，Synchronous I/O 在程式執行一個磁碟讀寫動作後，必須等待讀寫完成後才能繼續執行其他工作，而 Asynchronous I/O 則不須等待讀寫完畢便能繼續執行其他工作，讀寫完成後，作業系統會利用 signal 通知程式，程式便可處理讀寫完成後的後續動作。

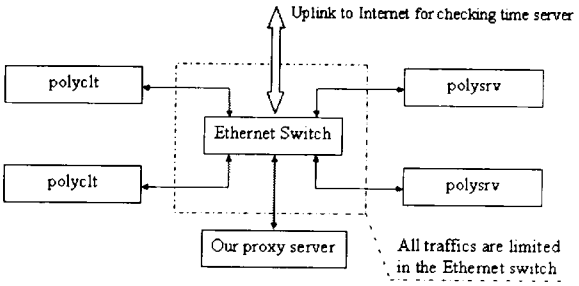
由此可知，使用 Asynchronous I/O 可節省等待的時間，增加 throughput，故我們亦透過 Asynchronous I/O 來增進 disk 存取的效能。

5. 實驗結果

我們使用 polygraph 2.5.4 作為測試軟體，其測試架構如圖三，其中 polyclt 用以模擬 client 端的行為，polysrv 用以模擬 server 端的行為。

測試平台如表一：

測試結果如表二，反應時間(Response time)與 peak request rate 的關係如圖四，由圖四觀察可知，當 peak request rate 愈高時，反應時間相對的會跟著拉長，其原因為反應時間會隨著 proxy server 的 loading 加重而變長。



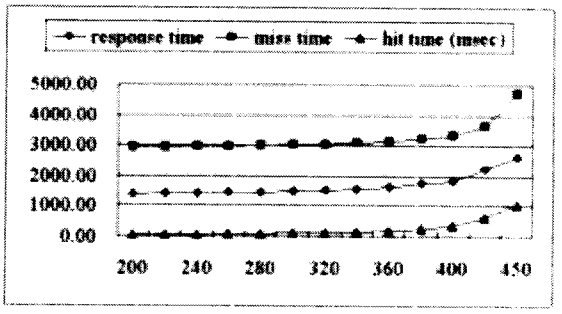
圖三

	Proxy Server	Server Machine	Client Machine
Hardware	CPU:PIII 650*2 Mem:1G MB HDD:SCSI 9G*6 SCSI card*2	CPU:PIII 800*1 Mem:256 MB HDD:IDE 30G*1	CPU:PIII 800*1 Mem:256 MB HDD:IDE 30G*1
Software	OS:FreeBSD4.1 Our proxy server	OS:FreeBSD3.4 Benchmark: polygraph2.5.4 Workload: Polymix-3.pg	OS:FreeBSD3.4 Benchmark: polygraph2.5.4 Workload: Polymix-3.pg

表一

peak request rate (req/sec)	response time (msec)	miss time (msec)	hit time (msec)	hit ratio (%)
200	1398.02	2941.2	58.08	56.36
220	1410.83	2950.64	70.12	56.28
240	1423.91	2960.64	79.14	56.18
260	1446.20	2981.7	90.22	55.90
280	1467.12	3001.92	103.91	55.77
300	1494.39	3025.85	121.88	55.52
320	1525.41	3052.63	141.93	55.20
340	1569.85	3092.8	172.65	54.89
360	1619.5	3140.61	206.12	54.53
380	1746.96	3232.94	277.7	53.75
400	1818.80	3338.05	365.06	52.91
420	2230.57	3659.79	645.54	52.52
450	2630.19	4720.21	1039.02	52.11

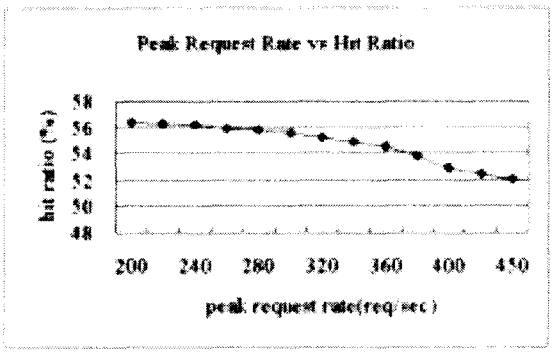
表二



圖四

Hit ratio 與 peak request rate 之關係如圖五，由圖中可觀察到，hit ratio 會

隨著 peak request rate 上升而下降，其原因為當 peak request rate 上升時，單位時間產生的 request 會上升，會產生 working set size 加大的效應，因而導致 hit ratio 的下降。



圖五

5.結論

經由與原始 Squid 測試結果的比較，可得表三。

	Squid 2.3 stable4	Squid 2.4 stable6	Ours
Peak request rate (req/sec)	60	140	450
Compare to Squid 2.3stable4	-	+80	+390
Compare to Squid 2.4stable6	-	-	+310

表三

Squid 2.3 stable4 是我們根據來修改的版本，而 Squid 2.4 stable 6 是目前最新的 stable 版本，經由表三的數據我們可發現，經過我們修改的版本，peak request rate 較原本的版本增加了 390 req/sec，即使與目前最新 Squid 的 stable 版本亦增加了 310 req/sec，故由此可驗證，經由對磁碟存取上的改善，可對 proxy server 的整體效能，可得到大幅的提升。

對於 proxy server 之效能上改進，我們整理出兩篇論文，已發表於 2002 International Computer Symposium [10] [11].

參考文獻

[1]. A. Rousskov and V. Soloviev. On Performance of Caching Proxies. In Proc. Of the 1998 ACM SIGMETRICS Conference, 1998.

[2]. Ahmed Alomari, "Oracle8i & UNIX Performance Tuning," Prentice-Hall.

[3]. Brian D. Davison, AWeb Caching Primer. IEEE Internet Computing, Volume 5, Number4, July/August 2001, pages 38-45.

[4]. Anirban Mahanti, Carey Williamson, and Derek Eager, Traffic Analysis of a Web Proxy Caching Hierarchy. IEEE Network, May/June 2000.

[5]. Charu Aggarwal, Joel L. Wolf, and Philip S. Yu, Caching on the World Wide Web, IEEE Transcations on Knowledge and Data Engineering, VOL 11, NO. 1, January/Feburuary 1999.

[6]. Device Polling support for FreeBSD http://info.iet.unipi.it/~luigi/polling

[7]. Edith Cohen, Haim Kaplan. Prefetching the Means for Document Transfer: A New Approach for ReducingWeb Latency. INFOCOM 2000.

[8]. Ing-Chao Lin, "I/O model and event handling in Web Cache", Department of Computer Science and Information Engineering, National Taiwan University.

[9]. Janeti. Egan, Thomas j.Teixeira, "Writing a UNIX Device Driver", John Wiley & Sons.

[10]. Mao-yu Jan, Yung-ching Weng, and Feipei Lai, "Performance Issues in Squid", ICS 2002.

[11]. Yen-Jen Chang and Feipei Lai, "Improve Web Proxy Performance by Alleviating Disk I/O Overhead", ICS 2002.