

虛擬之自動化半導體廠-子計畫二： 虛擬半導體廠之異質資料庫資訊伺服器(3/3)

計劃編號: NSC 89-2218-E-002-086

執行期限: 89年8月1日至90年7月31日

主持人: 許永真副教授
共同主持人: 許鈞南博士

國立台灣大學資訊工程學研究所
中央研究院資訊科學研究所

一、摘要

本篇為虛擬半導體廠之異質資料庫資訊伺服器子計畫之結案報告。本子計畫著重在各種異質資料庫之整合，並應用於虛擬半導體廠的各項資訊源整合。虛擬半導體廠中的資料源包括晶圓廠內的物料、機台、客戶、製程配方、批貨、下單等各種異質性資訊源。在面對多而複雜的各種資訊時，異質性資料庫伺服器必須能彙整各個系統間的資料流，同時也要能提供使用者簡便的使用方法及與各資訊源的溝通管道。

目前異質資訊系統主要研究的課題包括：在分散式架構下維護資料的一致性、定期自動地更新資料、由混亂的資料堆裏找出有用的資料等。本子計畫所設計之異質性資料庫伺服器系統，能彈性地建構資料領域並能整合各個不同性質的資料源。每一個資訊源均包含與其他資訊源不同的內容，且資訊源的資料完整性及查詢能力也各不相同。因此，必需將每個資料源做整理，定義出基本查詢運算、各種查詢及資料更新方法，以決定資料或運算元對映處理模式。同時，也必需研究如何規劃出適合的查詢步驟，並且成功地整合來自各資訊源之資料。

為解決資料之異質問題，每一資訊源均透過特別設計之轉換工具，即包覆程式(wrappers)，使所有資訊源都可以使用共同的格式來交換資訊，並且利用包覆程式使異質資料庫資訊伺服器系統能輕易整合並取得資料。

Abstract

The information server has to integrate all data streams between professional systems and support an easy-of-use interface for communication. Processes of producing wafers are very complex and elaborate. Resources of a semiconductor factory may include materials, machines, employees, recipes, paths of robots moving, ordering forms, customers, etc.

Moreover, data consistency of distributed data, regularly automatically updating data and retrieving useful data from chaos are more concerned than before by a good information system. Therefore, SIMS system with a domain model and IS models is supported to connect data resources within the same domain. Every data resource has different contents and query ability, and mappings of data and operators should be defined to accurately describe every data resource. Then query plans can be easily decided which data resources are related to, what proper query steps are and how to integrate query results.

To solve the problem of heterogeneous data, every data resource has its own information wrapper that makes SIMS integrate data and all information resources exchange data based on the same format.

二、計劃緣由與目的

人們在面對龐大而複雜的資料時，往往束手無策；尤其在半導體晶圓廠，其製

程十分複雜，通常需要使用及參考到許多不同類型的資料庫及其他相關資訊。因此，異質資料庫的整合除了提供有效率的查詢介面外，也有助於半導體業者對於各種不同資料的分析，使業者在激烈的競爭中保持優勢。

虛擬半導體廠的異質資料庫資訊伺服器必須彙整廠中各專屬系統間的資料流，以提供使用者一個簡便的溝通管道。除此之外，異質資料庫資訊伺服器也必需考慮到下列問題：

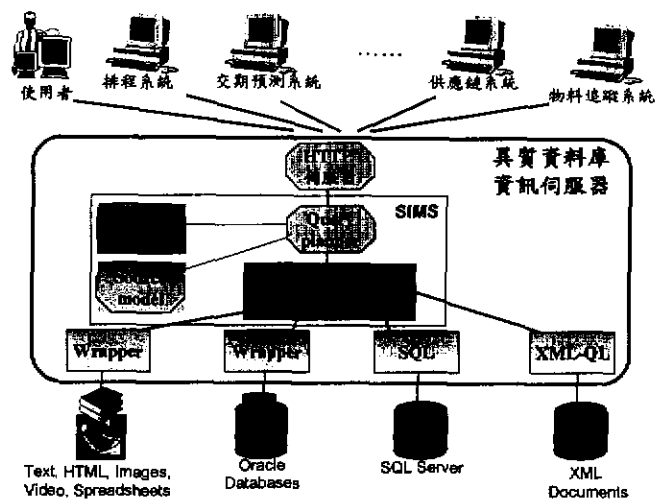
- ◇ 同一資料在不同的資訊源中，可能會有不同的名稱，甚至有不同的資料屬性或是不同的表達型式。
- ◇ 每個資訊源中可能有性質相似的資料重覆出現，但由於資訊源的來源均不相同，因此，在建立資料關聯性時必須考慮到這個問題。
- ◇ 資訊源的儲存型態及資料表現模式均不相同；也就是儲存資料的型式可能是關連式資料庫、物件式資料庫、文字檔或是其他特殊型式。
- ◇ 在晶圓廠中，各資訊源的資料可能隨時會有異動產生，造成資料的不易掌控性。
- ◇ 某些資訊源的資料來源可能因外力或人為干擾，造成資料不穩定或不準確。

因此，我們必需正確地描述每個資訊源，並有效地選擇與查詢相關的資訊源，規劃出適合的查詢步驟，同時成功地整合來自各資訊源之資料。資訊源之描述則應包含其內容、查詢能力及資料完整性等。另外需將每個資料源做整理，定義出基本查詢運算，以便解決資料或運算元對映處理模式。

為解決資料之異質問題，每一資訊源均透過特別設計之轉換工具，即包覆程式(wrappers)，使所有資料源都能使用共同的格式來交換資訊。

三、研究方法及成果

圖一所示為異質資料庫資訊伺服器的系統組織架構圖。除了使用者介面外，我



圖一、系統組織架構圖

們的系統也支援跨平台的工作介面，使各相關子系統的資訊，可以彼此交流。

異質資料庫資訊伺服器與 client 端的溝通管道，均以 HTTP 通訊協定為主；不論 client 端是否為瀏覽器，或是其他的代理程式及主機。藉著 HTTP 通訊協定，將 client 端的需求傳送至異質資料庫資訊伺服器的 query planner 中，同時也必需參考應用領域模 (domain model) 及資料源模型 (source model) 的定義，以便產生出一個查詢計劃，這個查詢計劃將交由 query plan optimizer/executor 來執行。由於資訊源來自各個伺服器，且型式各異。因此，必需利用包覆程式(wrapper)與各個資訊源溝通並將資訊源的資料轉換成所需的資訊。最後，查詢的結果將傳回對應的 client 端。

關於 client 端的介面以及異質資料庫資訊伺服器處理的相關資料，將於隨後的小節中有更清楚的詳細說明。

(一) Client 端介面

我們提供了以瀏覽器為主的圖形使用者界面。藉由瀏覽器上的圖形及文字說明，讓使用者對系統內的資料能夠一目了然，並且可以直接在圖上點選並且自動產生相對應的 SIMS 查詢運算式，以避免直接輸入文字所造成的混淆與錯誤。

對於其他子系統的代理程式，我們也設計了以 HTTP 通訊協定為準的溝通管道，使其他子系統的代理程式與異質資料庫資訊伺服器之間可以有效的交換資訊；另外，為了配合各種作業系統平台，我們

也提供跨平台的設計方案。

不論是那一種 client 端介面，他們的查詢都會透過代理程式傳送到異質資料庫資訊伺服器。代理程式會所欲查詢的資料以 XML(eXtensible Markup Language)為基礎的描述語言來說明。傳給異質資料庫資訊伺服器的內容主要有資料型態及查詢的資料內容。XML 為標記式語言的一種，它主要用來描述資訊的型式，並且也可以作為另一個新的標記式語言的基本架構。

Client 端的代理程式根據已建立好的資訊模型，執行對應的動作，並且透過 HTTP 通訊協定，將查詢的指令或相關資訊傳給異質資料庫伺服器。

(二) 異質資料庫伺服器

我們的異質資料庫伺服器主要包含 HTTP 伺服器、SIMS 及包覆程式。分別說明如下：

i) HTTP 伺服器

在經由 HTTP 通訊協定收到 client 端的查詢後，必須將收到的資訊轉換成 SIMS 的語法，才能把查詢的訊息傳送給 SIMS，在收到使用者的查詢後，SIMS 會自行決定完成該查詢所需之資料源，同時規化出一個查詢計劃(query plan)，並且將執行查詢計畫的產生結果回傳給對應的 client 端。

HTTP 伺服器與 SIMS 之間的溝通是利用 TCP 協定。在 SIMS 端，必須先啟動一個 daemon 程式，程式啟動後，client 端的查詢才能透過 socket 執行 SIMS 指令。當 client 端發出查詢資料的需求時，資料將透過 HTTP 的共用傳輸介面程式，將收到的資料轉成 SIMS 的查詢語法。在產生 SIMS 查詢語法後，以 TCP client 向 daemon 程式之 socket server 發出初始化 domain model 和 source model 的指令。

ii) SIMS

異質資料庫伺服器的主體方面，我們採用南加大發展中同性質的計劃的 SIMS (Services and Information Management for decision Systems, or Single Interface Multiple Sources)系統當作實驗平台。

SIMS 非常適合作為 client 端程式與分散式異質資料庫的溝通媒介。當 SIMS 收到查詢的需求時，query planner 會根據 domain model 及 source model 的描述產生查詢計劃(query plan)。Query plan optimizer/executor 根據查詢計劃將送出對應的子查詢，這些子查詢將會由包覆程式執行，在收到查詢結果後，SIMS 也會將結果傳回 client 端。

iii) 包覆程式

SIMS 與包覆程式之間的溝通透過一個簡稱為 KQML (Knowledge Query and Manipulation Language)的訊息交換格式標準。透過 KQML Agent Name Service 將子查詢送給對應的包覆程式，包覆程式除了會產生子查詢所需要的查詢語法外，還會取得某些必要資訊，例如資料庫系統的位置、資料庫名稱及資料庫使用者等資訊。在取得與資訊源相關的訊息後，包覆程式便透過 JDBC 到對應的資料庫系統取得所需的資料，再包裝成 SIMS 所需之格式傳回給 SIMS。

針對同一種資訊源型態，只需設計一個對應的包覆程式。我們已經成功的設計出可以應用在 Oracle、MS Access 及 MiniSQL 等資料庫系統的包覆程式。

除了常見的資料庫類型以外，我們也設計了全球資訊網網頁的包覆程式，我們的包覆程式使用一個以 XML 為基礎的描述性語言來定義網頁，藉由這個語言撰寫而成的描述文件，可以使包覆程式在全球資訊網上收集所需的資訊。

(三) 異質性資料源

由於半導體製程過程極為複雜，各子系統中均需要儲存大量的資訊，這些資訊包括晶圓廠內的物料、機台、人員管理資料、製程配方、無人自動搬運的路徑、批貨、下單等資料。因此，我們需要整合存放有數百個表單且資料量十分龐大的資料庫。

我們已能處理 Oracle、MS Access、MiniSQL 及全球資訊網的資訊源，這些資訊源分別儲存在不同的主機及作業系統

中。我們同時也建立了一個小型的測試環境，分別將顧客、產品、員工、批貨、製程、作業單元及訂單等資料存放於數台主電腦中；每台電腦的資料庫系統及資訊內容均不相同。

四、結論與討論

本子計畫之目的為發展一資訊伺服器以便整合分散於多個異質資料庫中的資訊，讓使用者透過符合標準且簡易的操作界面來得到較完整之資訊。

這三年來本子計畫成功地開發完成了異質資料庫之轉換模組，同時也定義了具高階查詢功能的描述式語言，利用描述式語言的說明來達成資訊伺服器技術及異質性資料庫之間的資料查詢。在對全球資訊網與各個資料庫間的介面技術，也有深入的研究。本子系統的每個模組均已建置完成且通過系統測試。

五、參考文獻

- [1] Greg Barish, Craig A. Knoblock, Yi-Shin Chen, Steven Minton, Andrew Philpot, and Cyrus Shahabi. TheaterLoc: A Case Study in Building An Information Integration Application. In Proceedings of IJCAI-99 Workshop on Intelligent Information Integration, 1999.
- [2] Chun-Nan Hsu and Chien-Chi Chang. Finite-State Transducers for Semi-Structured Text Mining. In Proceedings of IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications, 1999.
- [3] Ion Muslea, Steve Minton, and Craig Knoblock. STALKER: Learning Extraction Rules for Semistructured, Web-based Information Sources. In Proceedings of AAI-98 Workshop on AI and Information Integration, 1998.
- [4] Nicholas Kushmerick, Daniel S. Weld, and Robert Doorenbos. Wrapper Induction for Information Extraction. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97). 1997.
- [5] Jeffrey D. Ullman. Principles of Database and Knowledge-Base Systems. Computer Science Press, 1988.
- [6] WebMethods, Inc. Phillip Merrick and Charles Allen. Web Interface Definition Language (WIDL), September 22, 1997.
- [7] William W. Cohen. Integration of Heterogeneous Databases Without Common Domains Using Queries Based on Textual Similarity. SIGMOD 98. 1998.
- [8] Thomas Kistler, and Hannes Marais. WebL AV a programming language for the Web. In Computer Networks and ISDN Systems (WWW7), 30 April 1998.
- [9] Luca Cardelli and Rowan Davies. Service Combinators for Web Computing. DEC SRC Research Report 148. June 1997.
- [10] Laurent Miclet. Syntactic and structural pattern recognition, chapter 9, World Scientific, 1990.
- [11] S. Nestorov, S. Abiteboul, and R. Motwani. Extracting schema from semistructured data. In Proceedings of the SIGMOD International Conference, pages 295--306, 1998.
- [12] Yannis Papakonstantinou and Pavel Velikhov. Enhancing semistructured data mediators with document type definitions. In International Conference on Data Engineering, 1999.
- [13] Tim Bray and Jean Paoli and C. Sperberg-McQueen, Extensible Markup Language (XML) 1.0, W3C Recommendation, 1998.
- [14] William Cohen, Integration of Heterogeneous Databases Without Common Domains using Queries Based on Textual Similarity, In Proceedings of SIGMOD'98, 1998
- [15] Richard Connor and Keith Sibson, HCL - a language for Internet Data Acquisition, Workshop on Internet Programming, ICCL '98, 1998