行政院國家科學委員會專題研究計畫　成果報告

即時傳呼系統的設計與效能分析研究

計畫類別：個別型計畫
計畫編號：NSC93-2622-E-009-004-CC3
執行期間：93 年 05 月 01 日至 94 年 04 月 30 日
執行單位：國立交通大學資訊工程研究所

計畫主持人：吳毅成

計畫參與人員：徐健智、翁仕全、黃德彥、許俊彬

報告類型：精簡報告

處理方式：本計畫為提升產業技術及人才培育研究計畫，不提供公開查詢

中　華　民　國 94 年 8 月 30 日

國科會補助提升產業技術及人才培育研究計畫成果精簡報告

學門領域：資訊工程二

計畫名稱：即時傳呼系統的設計與效能分析研究

計畫編號：NSC 93-2622-E-009-004-CC3

執行期間：93 年 5 月 1 日到 94 年 4 月 30 日

執行單位：國立交通大學資訊工程研究所

主 持 人：吳毅成

參與學生：

姓　名	年　級	已發表論文或已申請之專利	工作內容
徐健智	博士班		分析並設計高效率、可延展的即時線上傳呼系統架構。
翁仕全	碩士班	已完成相關碩士論文	模擬分析即時線上傳呼系統架構之執行效能及可延展性。
黃德彥	碩士班		收集現有即時線上傳呼系統使用者行為模式之相關資料。
許俊彬	碩士班		收集並研究即時線上傳呼系統之相關資料。

合作企業簡介

合作企業名稱：群想網路科技股份有限公司

計畫聯絡人：趙先台

資本額：20,000,000 萬元

網址： http://www.ethink4u.com/　　電話：(03) 535-4357

合作企業產品簡介：

目前，群想網路科技自行研發出之技術，主要可分為兩個部份針對多人互動網路應用的一套系統平台，及企業入口網站 (Enterprise Information Portals，EIP)。

針對多人互動網路應用，本公司開發了一套網路互動平台。以此系統為主，我們已經開發了下列產品：

- 聊天室模組：含一般聊天室及活動聊天室。
- 線上遊戲模組：含麻將等三十多款遊戲，目前正在開發３Ｄ的多人線上遊戲。
- 手機遊戲模組：含麻將等數款遊戲。

- 無線 PDA 遊戲模組：含麻將等數款遊戲。

在管理上，本公司利用技術的優勢，來開發許多管理工具來協助管理，這也是可以維護如此多的會員人數的原因之一。

在 EIP 方面，群想網路科技開發了以下產品：

- 社區模組：社區模組可協助網站經營者與網友互動或促進網友們之間的互動。模組含會員系統、留言板、電子報、討論區、佈告欄。
- 電子商務模組：電子商務模組是協助公司導入電子商務，簡化運作流程，降低營運成本。模組含廠商資訊、新消息發佈系統、WAP 相關技術與產品研發。

群想網路科技由於擁有優良的技術團隊，在國內算是少數專以技術開發為重心的公司。因此，在技術的發展上可說是不遺餘力。在遊戲平台上，本公司已經取得與許多知名的網站合作計畫，如台灣新浪，香港新浪，美國新浪，Hinet，尋夢園網站等。目前尖峰時刻的同時上線人數接近 10,000 人，會員人數更高達 1,000,000 人，可說是國內相類似遊戲的廠商中極為成功的一家。網址在：http://www.cycgame.com。其他產品也都有相當不錯的成績。在企業入口網站方面（Enterprise Information Portals，EIP），也獲得桓基科技等公司的 EIP 計畫專案。

研究摘要(500 字以內)：

本計畫的主要研究目的是，研究設計高效率(Highly Efficient)及可延展(Scalable)的即時線上傳呼系統(Presence and Instant Messaging System；PIM) 架構。我們的目標是能支援愈多使用者愈好，如數十萬到數百萬同時線上使用者。為了達到這些目標，本計畫的研究將採用 BSD UNIX 的 KQUEUE 這些新的技術，代替傳統的 Threads 及 select。

本計畫的主要完成之工作項目，主要包含如下部份。

1. 收集並研究即時線上傳呼系統之相關資料。
2. 收集現有即時線上傳呼系統使用者行為模式之相關資料。
3. 研究並設計高效率、可延展的即時線上傳呼系統架構。
4. 模擬分析即時線上傳呼系統架構之執行效能及可延展性。

藉由此計畫與合作廠商的實際應用經驗，讓我們了解 PIM 系統的實際應用於業界的效果與能力。

人才培育成果說明：

本計劃對參與人才的培育有以下的成果：

(1) 在理論方面：

(a) 理論研究 UNIX 作業系統底層的設計，如 BSD UNIX 的 KQUEUE 及 LINUX 的 dev/poll 這些新的技術，並研究分析系統的連線上極限。

(b) 研究如何模擬極高連線數（如上數十萬到數百萬）的使用者行為。

(c) 研究高效率、可延展的即時線上傳呼系統架構。

(2) 實際方面：

(a) 達到單一伺服器就能支援到數十萬同時連線使用者這個目標，及可延展的高效率 PIM 系統。這可大為減少系統成本及維護成本。

(b) 目前有愈來愈多的企業，需要 PIM 系統來作企業內部的資訊交流，如何分析出使用者的行為模式也是非常的重要資訊。

對參與工作人員之訓練：

(1) 獲得研究現有即時線上傳呼系統內部元件與運作之方式。

(2) 獲得蒐集使用者行為模式資料的方法及經驗。

(3) 獲得即時線上傳呼系統架構之研究設計的方法及經驗。

(4) 獲得模擬分析系統的方法及經驗。

(5) 獲得分析系統效能及可延展性的方法及經驗。

(6) 參與者也有發表碩士論文及相關論文的經驗。

技術研發成果說明：

一、前言

近幾年來，網際網路(Internet)的蓬勃發展深深影響著我們的生活方式。網際網路的發達，使得資訊的傳播無遠弗屆，也使得資訊的傳播更為快速有效率。網際網路的應用大致可分為兩大種類：一、尋找資料：使用者在家裡就可使用瀏覽器(browser)等工具，便可以連上網際網路尋找所需要的資料。二、提供通訊管道：如網路電子郵件、電子看板(BBS)、網路聊天室等為許多人與朋友聯繫最常用的方法。

現代生活在快速的步調下，電子郵件的回應速度，有時無法滿足現代人的需求。因此一種更新的通訊方式－－線上即時訊息(Presence and Instance Messaging，以下簡稱 PIM)，因應而生。

一般的 PIM 系統，如 MSN Messenger、ICQ、Skype、Yahoo Messenger，都允許使用者維持一份線上好友名單(buddy list)。為了讓使用者能夠知道好友是否在線上，以及傳訊息給在線上的好友，一般的 PIM 系統會實作下列三個基本功能：

1. 訂閱他人的狀態：當我們連上線上即時訊息的伺服器後，可以訂閱我們有興趣的對象，隨後就可取得這些人目前的狀態
2. 當狀態改變時通知所有訂閱者：當我們狀態改變時，必須通知所有訂閱我們的人。
3. 傳送訊息給其他使用者：我們可以透過伺服器傳送即時訊息給正在線上的其他使用

者，他人也可利用相同的方法傳送即時訊息給我們。
除此之外，一般 PIM 系統還會實作線上聊天室、檔案傳輸、視訊會議(teleconferencing)等功能。

二、研究目的

本計畫的主要研究目的是，研究設計高效率(Highly Efficient)及可延展(Scalable)的 PIM 系統架構。由於這類的應用，使用者的數量通常很高，但每個使用者的 Loading 並不會很高，因此一台伺服器應該要能支援愈多使用者愈好，如數十萬到數百萬。為了達到這些目標，本計畫的研究將採用 BSD UNIX 的 KQUEUE 這些新的技術，代替傳統的 Threads 及 select。並研究及分析高效率及可延展的 PIM 架構。

三、文獻探討

當今較著名的線上即時傳呼系統有 MSN Messenger、ICQ、Yahoo Messenger、AOL Messager、YamQQ、Skype 等軟體。這些系統功能已經相當完善，傳輸檔案，語音視訊交談等功能都可透過線上即時傳呼系統做到。但是由於考量到市場的競爭，較少做公開的線上即時傳呼系統的系統效能分析。我們就所收集的文件，分別來討論以下三種熱門的 IM 系統：ICQ、AOL Messager、MSN Messager。

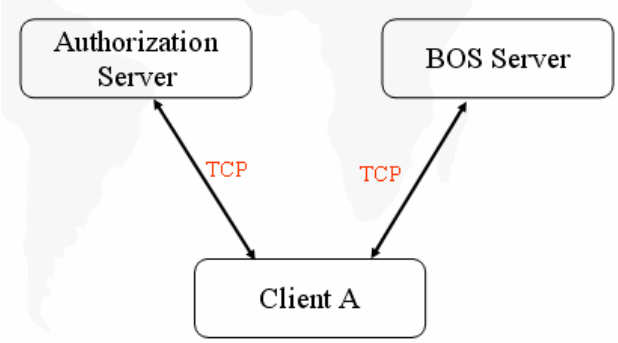
3.1. ICQ

根據 ICQ 所公布的資料，其主要的通訊方法有兩類：Client 與 Server 間通訊及 Client 與 Client 間通訊。Client 與 Server 間的通訊協定為 UDP，每個 UDP 封包加密後從 Client 端送到 Server，但是 Server 回傳的訊息卻不作加密動作。由於是用 UDP 在傳輸封包，因此需要實作具有 reliability 的連線控管機制：一端接收到封包之後，需回傳訊息(ACK)回去。如此一來，ICQ 封包也需加入自己的 Header，如 SEQ_NUM、COMMAND、SESSION_ID 等欄位，以便 Server 辨別封包是由哪個使用者所送出來的。很明顯，這樣的傳輸，程式開發的成本相當高。

Client 與 Client 間的通訊協定主要為 TCP，系統在使用者上線的時候，會告知其好友的狀態、IP、Port 等資料，當使用者需要傳送訊息時，便會先以 UDP 傳輸自己的 IP 及所開啟的 port 給對方，對方再根據資料連線過來建立 TCP 連線，之後便用 TCP 傳輸，進行聊天的動作，當使用者無法直接連線的時候，所要傳送的訊息便轉由 Server 轉送。

ICQ 架構的優點在於：高效率及伺服器不需額外佔用連線數目。但其缺點有：需實作偵測斷線機制、程式開發的成本相當高、訊息易被攔截攻擊。

3.2. AOL

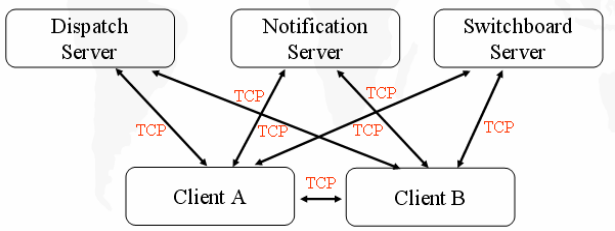


AOL 的系統構造圖如上圖，採用由 AOL 自己所發展出來的 IM 系統通訊協定 Oscar (Open System for Communication in Realtime)。AOL 是多 Server 架構，主要分為兩種 Server：Authorization Server 及 BOS (Basic Oscar Service) Server。Authorization Server 負責認證的動作，以及回傳給使用者 Cookie，以便使用者連線至 BOS Server 時確認身份。BOS Server 主要是提供線上狀態及即時訊息服務。Server 與 Client 間的通訊是以 TCP 來傳輸。系統運作的流程主要分為兩個大步驟：(一) 連線至 Authorization Server，(二) 連線至 Authorization Server 所建議的 BOS Server。使用者先連線至 Authorization Server，此 Server 會回傳 ACK 的訊息。使用者再發出 Authorization Request，Server 所回傳的 Authorization Response 中會帶有 Cookie 以及 Client 所要連線的 BOS Server 位置，此時 BOS Server 由 Cookie 便可辨認身份，不需要再作認證的動作。

AOL 架構的優點在於：無須實作偵測斷線機制、減少被攻擊的可能性。但其缺點是 BOS Server 間的通訊量極大，人數多時會影響其可延展性。

3.3. MSN Messenger

MSN Messenger 是微軟公司在 Windows 平台上所提供的，在 linux 上也有實作一套功能較少的版本，2000 至 2001 的使用者人數由九百萬人增加到一千八百萬人。



由於 MSN Messenger 有網站公布其通訊協定，因此我們可以知道更確切的資訊，它彼此間都是用 TCP 連線來傳送訊息，主要的通訊都是在 Client 與 Server 間，只有部分功能會使用 Client 與 Client 間互連，如檔案傳輸。其 Servers 主要分成三種：Dispatch Server、Notification Server、Switchboard Server。Dispatch Server 負責認證及分配使用者到適當 Notification Server。Notification Server 負責線上狀態服務，記錄使用者的線上狀態及所在位置等資訊。Switchboard Server 負責即時訊息服務，轉送使用者的聊天訊息，與其他 Server 間的通訊量極少。原則上，MSN 架構保有 AOL 架構的優點，此外 Switchboard Server 影響增加可延展性。

四、研究方法

本計畫的主要研究方法及項目如下，除了第一項以敘述於第二節外，我們分述於各子節：

1. 收集並研究 PIM 系統之相關資料。
2. 收集現有即時線上傳呼系統使用者行為模式之相關資料。
3. 研究並設計高效率、可延展的即時線上傳呼系統架構。
4. 模擬分析即時線上傳呼系統架構之執行效能及可延展性。

1.1. 收集即時線上傳呼系統之相關資料

為了了解使用者使用 Messenger 的行為，我們記錄合作廠商員工使用 Messenger（MSN Messenger）的行為。我們的記錄方式，是將每位使用者 Messenger 的所有以下動作都記錄（包括發生時間）到一個 log 檔案。

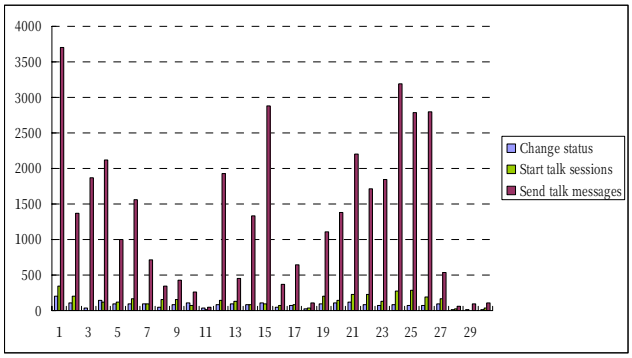
1. Change status:使用者登出入系統或改變狀態。
2. Start talk sessions:使用者開始啟動一個 talk session（如開 talk 視窗）。
3. Send talk messages：使用者送任何訊息給朋友。

	# packets	Percentage
Change status	2408	5.31%
Start talk sessions	3998	8.81%
Send talk messages	38964	85.88%
Total	45370	100.00%

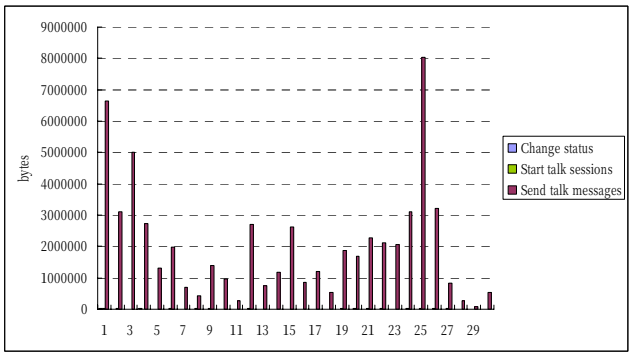
表一：該整個月之封包流量

	# bytes	Percentage
Change status	481603	0.72%
Connect talk server	799600	1.19%
Send talk message	65808112	98.09%
Total	67089315	100.00%

表二：該整個月之位元(byte)流量



表三：該整個月每日之封包流量



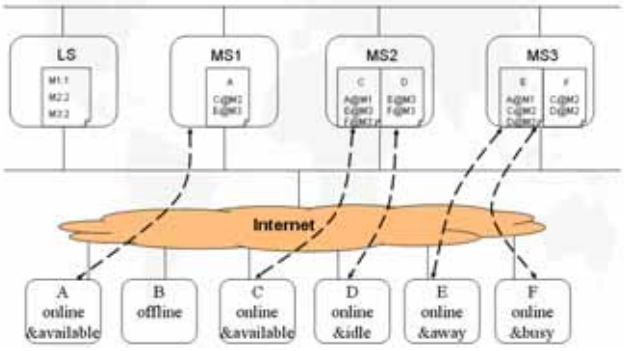
表四：該整個月每日之位元(byte)流量

我們將收集一個月的 log，整理成如上前兩圖表，分別顯示封包流量及位元流量。另外，我們亦將每日的流量細分於上後兩圖表。

這些圖表顯示一個有趣的結果，那就是大部分的使用者行為是：使用者改變狀態及開新的 session 次數及數量，都遠低於使用者談話的次數與數量，這結果對我們的系統架構設計有相當的影響。

1.2. 即時線上傳呼系統架構

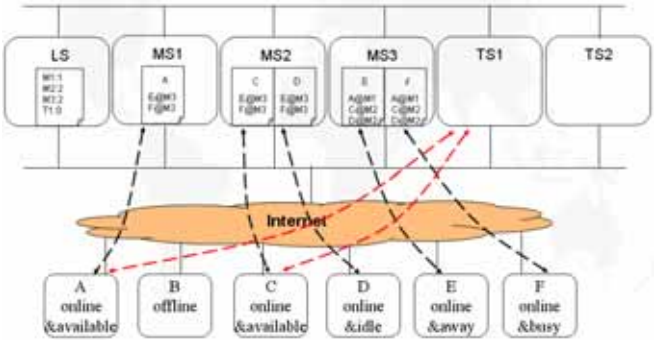
首先，傳統的 Threads 及 select 對大量使用者的 services，都會有嚴重的 overhead。為了解決這類問題，我們採用 BSD UNIX 的 KQUEUE 這些新的技術，代替傳統的 Threads 及 select，來減少 overhead。



圖一：無 TS 的系統架構圖

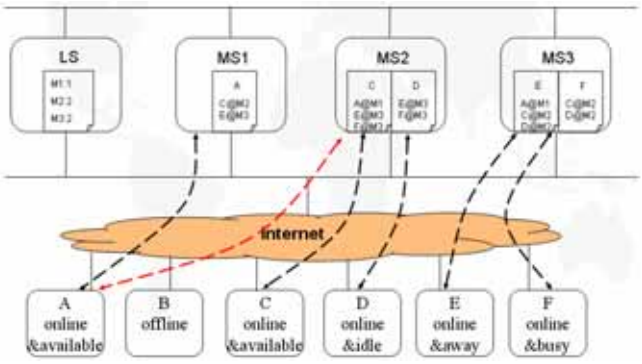
對多主機系統，我們先將採用類似 MSN 的架構，使用三種 servers。我們稱為 Login servers (LS)、Messaging severs (MS)、talk servers (TS)。若無 TS 的情形，系統架構如上圖。LS 的任務，除了要做 login 確認外，同時可以幫忙平均分配 loading。

然而，從前子節中，我們知道 talk 訊息的流量極大，若無法讓大部分使用者的好友都在同一台 MS 的話，MS 間的流量會非常的大。由於這個原因，通常在 server 端，加上一個 local LAN (如圖上端)，是有幫助的。但是，使用者數量更大時，還是會無法 scalable。因此，加入 TS 專門負責 talk messages 應該有助於多台伺服器的 Scalability。這是 MSN 的做法，其架構如下圖。



圖二：含 TS 的系統架構圖

對上圖而言，問題是系統必須要知道 TS 與 MS 之間的比例才能將 Load 分配的好。因此我們建議一個方法，是平均分布使用者於各個 MS，然後當一個使用者 A 想要與 C talk 時直接到 C 的 MS。由於平均分布的關係，每台 MS 分配的 talk 流量也會平均。另外，若 MS 的 Load 真的很高時，在分到 TS 去。因此我們架構圖如下。



圖三：系統架構圖及 talk 訊息

1.3. 模擬分析即時線上傳呼系統架構之執行效能

我們模擬的方法是，將上述的合作廠商的員工使用 Messenger 的 Log 檔案，取出較為繁重的 30 分鐘出來模擬其訊息分布。由於這是一段繁重的使用，以保守原則，這應具有代表性，代表最繁重的狀況。

在這段資料中，除了團體內的 20 個使用者之外，外界的好友還有 11 個人，而我們要用此資料來模擬的話，需要讓 31 個人都上線。以保守原則，我們讓這 31 個人都上線，但只算 20 個使用者。然後，我們複製這一組使用者，及其 log 資料。

首先我們模擬 Messenger 在單一一台主機的情形。主機環境如下：

- CPU: 使用 AMD Athlon™ XP 2000+。
- Memory: 達 1.5G bytes。
- OS: 使用 FreeBSD 4.9-RELEASE
- Network adapter: 使用 3Com 3c905B-TX Fast Etherlink XL。

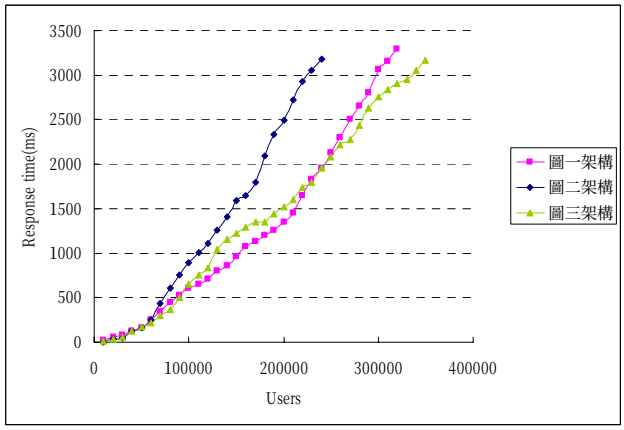
然後，由 0 人開始，每次增加 10000 人，取得平均回應時間及伺服器負載量(CPU

loading)，另外我們還計算前 10%高的回應時間平均，來當作我們判定系統是否負擔過重的判斷因素。因為一般使用者所可以忍受的延遲大約在三秒以內，所以當前 10%高的回應時間平均超過三秒時，我們便說此系統超過負荷。

根據我們的實驗，單一伺服器在服務 120000 個使用者時，前 10%高的回應時間平均為 3013.45ms，剛好超過三秒。

現在考慮多台主機情形。我們用三台來實驗。由於 Locality 不是我們系統的重點，每組的使用者將被平均分散到不同機器，來增加 MS 間的流量。我們的實驗實際上有考慮許多因素如下：

- 使用 Local LAN 與否。
- 使用 TCP 或 UDP。
- 對 UDP，使用廣播(broadcast)與否。
- 維護使用者資料在 LS 或在各個機器上（利用廣播來處理）。
- 用圖一、圖二、或圖三架構來處理 talk 訊息。



圖四：三種架構之數據結果

我們對以上各種組合，除了明顯效率不好的狀況外，我們均有做實驗分析。在此，我們將只顯示最後一項的分析比較結果。若用圖一或圖三架構，我們用三台 MS；若採用圖二，我們用一台 MS 及兩台 TS。數據結果如上圖。我們的分析如下：

- 對圖三架構，在服務 340000 個使用者時，前 10%高的回應時間平均為約為三秒。依據之前的定義，我們說此系統可支援 340000 使用者，比單一機器的 120000 使用者相比，接近 3 倍。另外圖一架構，在服務 300000 個使用者時，前 10%高的回應時間平均為約為三秒。另外圖二架構，在服務 230000 個使用者時，前 10%高的回應時間平均為約為三秒。
- 對圖一與圖三架構，由於圖三架構可大量減少 Local LAN 的 traffic，當人數變得愈大時，效率愈佳。
- 對圖二架構，由於無法對 MS 及 TS 的 Load 做最佳的分析評估，以致於 Load 的分配不平均，導致執行效率最低。

以上結果顯示我們的架構有較佳的延展性。我們將會用更多的主機來驗證。

五、成果與討論

本計劃對萃取容錯能力做了研究與分析。主要成果如下：

1. 收集即時線上傳呼系統之相關資料。
2. 收集現有即時線上傳呼系統使用者行為模式之相關資料。我們分析出談話佔絕大多數的 Load。
3. 研究並設計高效率、可延展的即時線上傳呼系統架構。
4. 模擬分析即時線上傳呼系統架構之執行效能及可延展性。
5. 依據我們的定義，我們模擬實驗結果顯示，一台一般的主機可支援到 120000 個使用者。三台可支援到 340000 個使用者，有相當不錯的延展性。

技術特點說明：

本計畫的技術在於研究 PIM 系統的極限，並分析及驗證圖三架構的效能優越性。研究結果的特點如下：

- 對單一一台基本配備的主機（如上所述），以保守方式模擬，可支援達 120000 使用者同時上線。
- 對多主機系統（三台配備如上），圖三架構可支援達 340000 個使用者同時上線。這比圖一架構的 300000 個使用者，及圖二架構的 230000 個使用者都好很多。以上結果顯示我們的架構有較佳的延展性。我們相信這樣的結果對想開發類似 PIM 系統的業界有相當的助益。

可利用之產業及可開發之產品：

本計畫除了有模擬 PIM 的雛形系統外，我們的成果對想開發類似 PIM 系統的業界提供有用的發展方向。此模擬系統經修改後，應可開發成業界的 PIM 產品。

推廣及運用的價值：如增加產值、增加附加價值或營利、增加投資/設廠、增加就業人數……等。

由於過去的研究少有對 PIM 做研究分析，本計畫所作的一些結果，對學界及業界均有一定的幫助。因此，自評此計畫對學術研究及台灣相關業界有相當的助益。