High-Performance Distributed Platforms for Computation and Data-Intensive Applications
高效能分散式計算與資料應用平台

張耀中*、林育珊、鄭淳瀚、廖祥瑞
*國立台東大學資訊工程學系
國立台東大學資訊管理學系
{ycc,ysl}@nttu.edu.tw

摘要

Taiwan UniGrid 集合國內各大專院校之學術研究團體，共同建置一計算能量互援之格網運算平台供合作單位使用，並於此平台進行格網相關技術研究、格網程式設計、格網應用程式等，藉以分享與推廣格網運算平台建置及應用經驗。本研究採用美國 San Diego Supercomputer Center 所發展之 Storage Resource Broker(SRB)系統，SRB 為整合異質與分散儲存資源之中介軟體(Middleware)，提供虛擬資料管理環境，使用者能存取分散儲存資源。本研究負責 Taiwan UniGrid 台東節點之建置與系統實作，研發高效能之資料管理系統(Data Management System)，各節點透過高速網路緊密連結，達到高可用性(High Availability)、高可靠度、資源分享及強大計算能力之目標。

關鍵詞：Taiwan UniGrid、Storage Resource Broker、效能分析、格網運算。

Abstract

Taiwan UniGrid project concentrates several universities and academic research centers to construct grid computing environments for sharing computation power, including grid technology research, grid programming, grid applications and etc. The goal of grid computing is to coordinate resources under network environment. In this project, Network Technology and Information Security Laboratory (NTIS) of National Taitung University connects several grid systems between universities and NCHC to build "An Experiment Platform of National Computing Grid", which is used to popularize the concept of grid computing to academia and industry. The data management system and grid environment also provide high availability, reliability, research sharing and powerful computation.

Keywords: Taiwan UniGrid, Storage Resource Broker, Benchmarks, Grid Computing.

1. 前言

本研究藉由資料格網技術與 Storage Resource Broker(SRB)[1,2]溝通平台實作格網架構系統及資料儲存分配與管理應用系統，透過 inQ client[3]中介軟體連接 SRB 平台，進行資料儲存與測試，並藉由 Taiwan UniGrid 系統架設驗證格網運作效能。

為了讓東部地區透過格網技術來整合資源，國立台東大學網路技術與資訊安全實驗室 (NTIS)負責 Taiwan UniGrid 在台東地區的節點建置與實作高效能的資料管理系統(Data Management System)。本論文之第 2 節簡介格網計算相關議題，第 3 節提出 Taiwan UniGrid 整合架構，第 4 節介紹 Taiwan UniGrid 效能分析方法與測試結果，第 5 節總結本文。

2. 格網計算介紹

Grid 技術之發展，自 1996 年開始，配合需求殷切的先端基礎研究，成為大規模分散資源整合與共享之完整且通用解決之方案。然而許多個別需求的應用，例如：Data Grid：利用其運算技術，以管理並共享分散各地的資源及設備；SRB：跨越不同作業系統平台，以連結各個不同節點之資料庫來獲得所需的資料；共通規範的訂定，如：Globus[4]、GSI、OGSA(Open Grid Service Architecture)，開放網格服務體系結構，主要協調 Globus Toolkit 的開發工作，透過 Internet 與 Web Service 存取應用程式，串連不同架構的電腦系統。目前一般較常被應用的格網技術有以下三種[5]：

- 運算格網(Compute Grid)

運算資源分享是目前常應用的一種格網技術，高效能的分散式計算加速計算複雜度高的工作與資料集之運算效率，並減少工作時間，且能夠整合各分散的計算資源，執行單一計算設備無法處理之獨立的子工作。

- 資料格網(Data Grid)

資料格網主要為建立一共用之虛擬檔案資料庫。在目前的研究中，本團隊以 inQ 和 MySRB 此兩種中介軟體作為應用，建立龐大的資料庫系統，作為分享 Grid 中所有資訊的橋樑，功能遠勝於 FTP，且在存取資料的時候，如同存取本身電腦資料一般，讓任一節點之使用者均可在此機制中共享資料。

- 應用程式格網(Application Grid)

應用程式格網是格網技術的長期願景，未來所有的電腦都連線網際網路，在合理的授權範圍下，Grid 系統中所有電腦之應用程式均可相互溝通，此概念類似 Web Service 的方式，最後都會朝向在網路中不同節點應用程式進行互動。Web Service 的技術將會逐漸地發展成 Grid Service 方式，目前 Microsoft.Net、Sun 之 Sun ONE 與 IBM WebSphere 等均朝向應用程式資源格網的應用方向發展。

然而，在 Grid 系統和應用相關的問題與挑戰如[6]所述：

- 安全防護方面的挑戰(使用者登入到 Earth System Grid)

  Challenge：新增一個單一使用者登入介面至 Grid 的應用中，而那些使用者只需要一組簡單的帳號與密碼的結合即可登入。同時，只要更動安全設定程序，便能變更 Grid 內部的安全性。

  Solution：可在帳號與密碼部分分別設定與管理，如增設管理與使用等級，分別控管儲存、讀取、複製、管理功能，低於某等級者，則無法進行與使用功能。

- 監視/管理方面的挑戰(Earth System Grid 的監視系統)

  Challenge：當服務失去作用與服務所紀錄的效益在後續的分析中過時，公佈系統管理人，藉著監視關鍵性的系統元件，增進分散式系統全面效益。

  Solution：藉此監視系統的控管，由系統管理人緊急關閉失去作用的服務，以避免 Grid 系統中資料的損失。

- 資料方面的挑戰(快速地移動在 TeraGrid 上的資料)

  Challenge：當轉換一個大檔案或者一組小檔案之時，給予使用者一個坐落在廣大網域間之網點的方法來轉換檔案，此方法適用於 10-30Gb/s 網路連結的最大容量。

  Solution：資料檔案的轉換為 Grid 中諸多功能之一，而在偌大的空間中進行檔案的轉換與複製，難免無法保證不會遺失。因此，套用 DRS 的功能，以保證其穩定性，確切地將資料送至目的地。

- 虛擬組織管理方面的挑戰(管理計算結果及其應用在 EGEE 與 OSG 的需求)

  Challenge：一個虛擬組織從一個資源提供者協商一組配置，然後根據虛擬組織的策略，將此配置散佈開來遍及其成員。

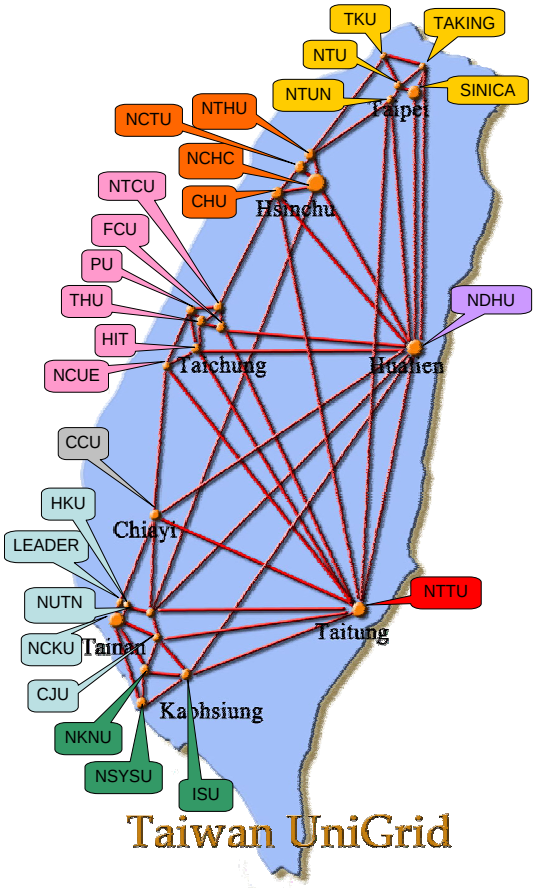
  Solution：虛擬組織管理的挑戰等同於 Grid 中安全與資料監控。

3. Taiwan UniGrid[7]

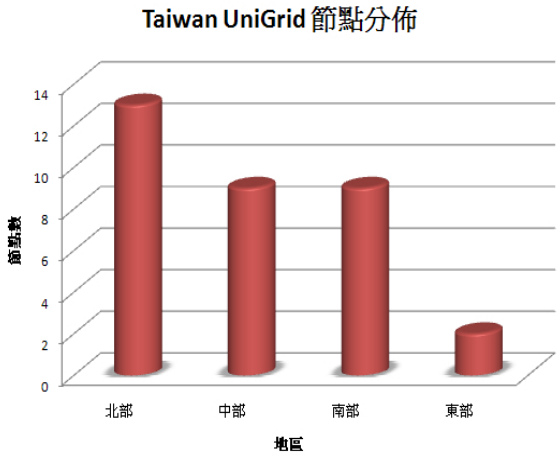
格網技術其主要核心概念為結合分散於各地的電腦運算資源、儲存裝置以及其他各種類型的資訊系統，彙整網路相關的計算資源，構成一虛擬整合應用環境。使用者可突破空間的障礙，建立彼此可共享資訊、資源、應用工具與知識的虛擬組織。透過此一虛擬組織，建構一安全穩定的資料儲存與共享機制，並且可以有效地管理分散於各機構之異質資源，以促進資源分享與利用，並解決更大尺度的問題；再者，使用者也可突破時間上的障礙，在格網中存取與分享其他節點資料庫，以期更有效率地進行資料的搜尋、整合、分析及儲存與管理[8-13]。Taiwan UniGrid 建置節點如圖一所示，包含北區：台大、清大等 13 所大專院校；中區：彰師、逢甲等 9 所大專院校；南區：中正、成大等 9 所大專院校；東區：東華、台東 2 所大專院校，總計共 33 所學術研究單位參與此計畫，計畫目標分述如下[14]：

1. 分散式超級計算(Distributed Supercomputing)：利用 Grid 整合各地的計算資源，透過許多電腦與網路，橫跨各領域、單位，整合成單一資源來使用，合力解決一個原本在單一計算設備與能力上，無法處理和滿足複雜計算或應用需求的大型計算問題。
2. 高生產力計算(High-throughput Computing)：利用整個 Computational Grid，處理大量獨立或非緊密相關的計算工作，在處理過程中，不依賴於單一計算資源處理，而是平衡整體 Grid 中所有獨立資源之負載量，目的在使整個 Grid 達到原本各計算資源分散開時無法達到的最大生產力，並使 Grid 中的計算資源能發揮最大效用。
3. 隨選計算(On-demand Computing)：利用 Grid 中分散各地的豐富計算資源，達成即時需求的工作，在任何配置下實作與交互存取各地的計算資源，以期滿足其即時性的需求。
4. 資料密集計算(Data-intensive Computing)：整合存在 Grid 內各處的資料，以計算及通訊密集的方式，達成大規模計算量之資料共享，建制一計算系統，整合使用者所需資訊，如產生大量觀察或計算量應用之高能物理與天文物理等領域。

5. 協同計算(Collaborative Computing)：著重 Grid 系統之人跟人之間的互動，著重於跨領域、跨邊界的資源共享，而非集中於中央控管的資源，協助各地研究或工程人員共同觀測及討論。



圖一、Taiwan UniGrid 建置情形



圖二、Taiwan UniGrid 節點分布長條圖

3.1 GridFtp[15]

GridFtp 是一種安全可靠度、高頻寬、廣域之資料傳輸協定。GridFtp 協定是基於 FTP 這一廣受歡迎的資料傳輸協定，源自於意識到格網環境需要一種快捷、安全、有效而且可靠的傳輸機制。計算格網之應用是如此的龐大和分散，以致於需要一種健壯的傳輸機制。GridFTP 提供下列功能而滿足了使用者需求：

- 並行資料傳輸：使用多個 TCP 流量比使用單個 TCP 的流量更能提升使用頻寬，並行資料傳輸由 FTP 命令擴展和資料通道擴展提供支援。
- 格網安全性基礎設施(Grid Security Infrastructure, GSI)與 Kerberos 認證支援：由用戶控制各種資料完整性和機密性級別的設置。此種功能為傳送文件提供了認證功能，具備資料完整性與機密性。
- 資料傳輸的第三方控制：支援為大型分佈式社區管理大型資料集，使第三方能控制與存取伺服器。
- 分塊資料傳輸：能夠將數據分割置放在多個伺服器上，進而提升傳輸頻寬。GridFTP 是透過定義在格網論壇（Grid Forum）草案中的擴展協議以支援分塊資料傳輸。
- 部分文件傳送：與標準 FTP 要求應用程式傳送整個文件不同，新型 FTP 命令支援傳送文件的某些特定區域。
- 可靠的資料傳輸：故障恢復方法可以處理瞬間網路斷訊和伺服器故障，同時可以重新啟動失敗的傳送。
- 手工控制 TCP 緩衝區大小：支援獲取最大 TCP/IP 頻寬。
- 集成檢測（Instrumentation）：支援返回重新啟動之功能。

3.2 資料管理[16]

網頁服務之格網資源分配與管理元件是由網頁服務資源架構(Web Services Resource Framework)所允許的網頁服務，透過 WS GRAM 來對格網工作資源進行搜尋、呈現、監測等功能，然而它並不是一個工作排程而且一種共同的協定，用以提供客戶端與 Grid 系統資源之間的溝通，WS GRAM 是指對大量任務定位出何處為可靠的運算、監督狀態、管理和檔案等重要功能：

- 提供協調單一任務程序、協調多元任務子工作及多元任務處理功能。
- 協調任務移動路徑。
- 提供選擇客戶端的詳細狀態。
- 監控與管理執行任務。
- 對批次排班系統提供彈性與一致介面。
- 執行任務前後的檔案管理。

3.3 Data Replication Service (DRS)

DRS 是一由 GT 4.0 所提供的技術服務，最早出現在 GT 3.9.5 試用版發表中。此元件的主要功能是允許使用者確認一組存在 Grid 環境中所要求的檔案，由一個或多個資源位置藉著轉換檔案進行資

料檔案的複製，並且登記新的複製檔案於 Replica Location Service。DRS 遵照 WS-RF 的規格，呈現出需求複製活動的當前狀況與允許使用者為了監督資源狀況與資源屬性。DRS 是建立在 GT 4.0 Java WS Core 與使用 Globus RLS 以尋找與登記資料複製，並於 Globus RFT 轉換檔案。

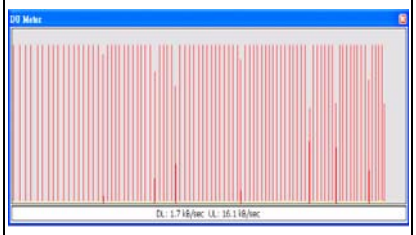
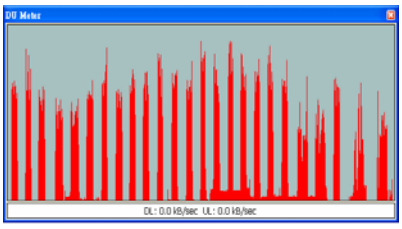
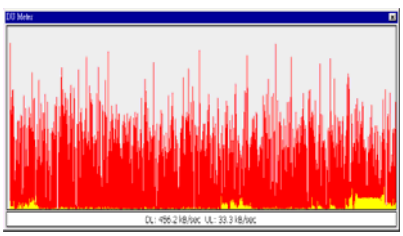
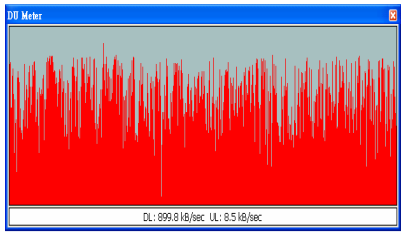
4. 效能分析

本研究使用 Microsoft inQ 作為主要的 SRB 介面，但由於 MySRB 提供較便利之網頁視窗作業環境，因此實做系統於 MySRB 伺服器。而目前 MySRB 伺服器架設在 SDSC(San Diego Supercomputer Center)，因此在速度上及與節點相連接效率較 inQ 差；至於另一種介面 Scommand 因牽涉到語法及指令使用複雜性，在實際應用上對使用者操作之介面仍需調整；而另一種 hfs 介面目前只提供 SRB 瀏覽功能，並無法支援下載上傳等功能。hfs 可提供 SRB 的瀏覽功能，運用 jargon 的 java 程式，需再利用 jargon 的 API 撰寫其他相關功能之 API，造成使用者不便。綜合目前介面軟體，本研究以 inQ 為主要效能分析方法，採用 inQ 與 Scommand 測試與比較格網傳輸環境之效能及穩定性。

4.1 InQ 實驗：

使用 inQ 以及 FTP 軟體分別在學術網路與一般網路下載時的傳輸速率，測試的檔案大小為 926MB，實驗結果如下表：

表一、InQ 流量圖

	流量圖
(SRB介面) 學術網路	
(SRB介面) 一般網路	
(FTP介面) 學術網路	
(FTP介面) 一般網路	

表二、InQ 傳輸速率

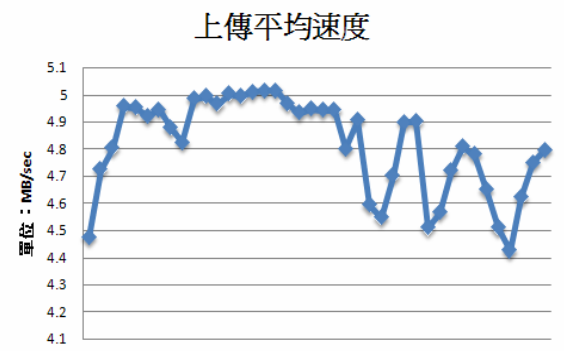
	下載完成時間	線路傳輸速度	下載速度排名
(SRB 介面) 學術網路	**33分30秒**	**1G/s**	**2**
(SRB介面) 一般網路	**1時10分30秒**	**8M/s**	**4**
(FTP介面) 學術網路	**54分49秒**	**1G/s**	**3**
(FTP介面) 一般網路	**26分20秒**	**8M/s**	**1**

由上表顯示經由SRB傳輸檔案呈現較平均的分配，雖然每個區段會間隔一段時間，但每個時段下載的量卻很平均，尤其是學術網路的使用，更能使資料的下載更快速及平穩。由於台東大學NTIS實驗室MCAT主機架設在東華大學主機之下，因此使用SRB傳輸需由台東大學連結台灣高品質學術研究網路(TWAREN)，再連線至東華大學。而相較於一般使用FTP軟體下載，直接連結到東華大學的FTP站台，相較之下速度較快，但從下載時的穩定度與傳輸量，使用SRB確實能提供較平均以及穩定之下載效率。

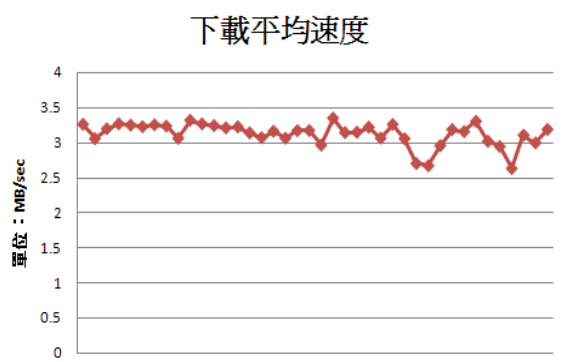
4.2 Scommand 實驗：

Scommand 則是一套 command line 之應用軟體。實驗過程為：連線到實驗主機，利用終端機指令，開啟 Scommand，使用 Scommand 內建指令，開啟後先輸入指令“$ Sinit”兩次，Sinit 指令是開啟與 Server 之間的連線。接著輸入“$Sls”查看 Server 端檔案列表，使用“Sput”指令將檔案傳送到 Server 端，檔案名稱前面加上“-v”代表顯示資訊。使用“Sget”指令將檔案由 Server 端下載回主機，檔案名稱前面加上“-v”代表顯示資訊。

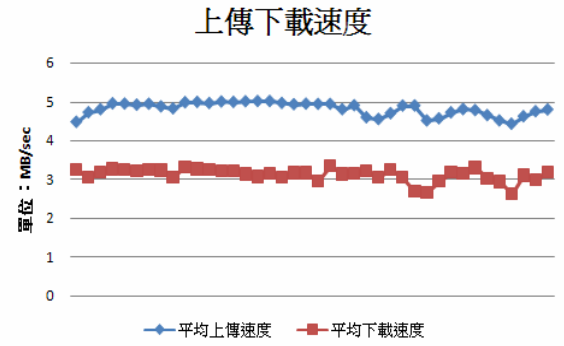
使用 Scommand 的上傳下載指令，各自測試 40 次，檔案大小 90.619MB，實驗結果如下圖：



圖三、Scommand 上傳速度圖



圖四、Scommad 下載速度圖



圖五、上傳下載速度圖

由圖三、四、五之比較顯示當使用Scommand作為上傳功能時，其速度的波動較下載時大，在40次上下載傳輸測試中，下載的穩定度大於上傳的穩定度，與前述inQ測試有相近的結果。當使用SRB下載檔案時，均能提供使用者較為穩定且平均之下載傳輸環境，但也因此產生使用SRB上傳時波動幅度較大之問題，可能與網路、伺服器或其他相關因素及節點個數有關。

5. 結論

本研究成果包含Taiwan UniGrid台東節點之建置與系統實作，研發高效能之資料管理系統(Data Management System)，各節點透過高速網路緊密連結，達到高可用性(High Availability)、高可靠度、資源分享及強大計算能力之目標。由研究成果可得知，在格網環境下使用SRB作為資料傳輸達成較穩定傳輸效果，且在資料完整性與安全性上較強效益。未來Taiwan UniGrid計畫規劃連結全台之MCAT節點，此分散式計算平台所提供資訊交流與資料傳輸效益將更為可觀。

致謝

本研究成果由國科會計畫 NSC 95-2218-E-143-001、NSC 95-2221-E-143-003-MY2 及 NSC 96-2221-E-143 -001 提供部份研究經費補助，並感謝國立東華大學資訊工程學系張瑞雄教授之網路創新技術實驗室提供 Grid 相關技術支援，國立台東大學資訊工程學系林正祐、陳群元、洪家祥與石振豪同學共同開發與維護此應用平台。

參考文獻

[1] SRB, http://www.sdsc.edu/srb/index.php.
[2] San Diego Supercomputer Center, “Overview of the SDSC Storage Resource Broker”, University of California San Diego, May 2004.
[3] inQ, “Beginner’s User Guide to inQ and SRB”, http://www.sdsc.edu/srb/index.php/InQ
[4] Globus, http://www.globus.org/.
[5] Getting the MPICH implementation, http://www-unix.mcs.anl.gov/mpi/mpich1/download.html
[6] Grid Computing, http://home.kimo.com.tw/mis2000_graduate/network_data/grid_computing.htm
[7] The Taiwan UniGrid Project, http://www.unigrid.org.tw/
[8] 蕭志棋，「自由軟體應用於格網與叢集計算環境的經驗 」，第四屆建立開放應用環境論壇，94年8月
[9] 蕭志棋，「叢集與格網計算環境DRBL實地展示」，95年3月
[10] 張瑞雄、王智敏、林春福、黃甯園、林皓嶔，「應用格網技術於數位典藏系統」，94年2月
[11] 黃國展、張西亞，興國管理學院電子商務學系，「計算格網平台上不同工作負載分擔模式之效能評估」，第四屆離島資訊技術與應用研討會，pp.1，94年5月
[12] 東海大學物理系，「Taiwan UniGrid之效能分析」，95年10月
[13] 陳德民，中原大學資訊管理學系碩士論文，「以格網服務為基礎之問題解答環境」，93年7月
[14] International Symposium on Grid Computing 2003 & TW Grid Workshop, http://www2.twgrid.org/event/isgc2003/tw/index-c.html
[15] IBM，「用GridFTP傳送文件」，http://www-128.ibm.com/developerworks/cn/grid/gr-ftp/
[16] Data Management of Globus http://www.globus.org/toolkit/docs/4.0/data/key/