

在Grid系統上利用NGB驗證Grid Economy對工作排程的影響

賴冠州 黃國展 林士傑 黃振維 楊大猶 吳勇均

臺中教育大學資科系

kclai@mail.ntcu.edu.tw

摘要

格網系統目前已應用在各種研究領域以解決大量計算或龐大儲存資料的需求，未來格網技術可望成熟而普及至一般使用者。因此建立一個格網資源之交易機制讓格網資源提供者提供資源給使用者利用，而格網資源提供者則從使用者獲取格網虛擬貨幣為一個重要之研究方向。在這種交易機制下，Grid Economy model 可替格網系統建立一個良好的協定讓資源提供者及使用者達到各自目標。本研究主要目標在於驗證 Grid Economy model 可讓格網資源之管理達到較高的使用效率，且使用者能夠根據本身的 QoS 選擇適當的格網資源執行工作而達到經濟效益。本研究使用 NGB benchmark 測試 Taiwan Unigrid 效能並驗證 Grid Economy 原則確實對格網資源之管理及配置有所幫助。

關鍵詞：格網、經濟模式、資源分配、工作排程

1.前言

格網系統(Grid System)為一個透過網際網路連結動態、多個虛擬組織、不同管理者之分散式電腦系統，用以解決工業、科學及分子運算等具大量計算或龐大儲存資料之需求。格網系統提供的服務與分享的資源包含分散在不同地域由不同組織擁有的超級電腦、儲存系統和其他裝置所提供之計算能力、遠端電腦的軟體、資料及其他可透過網路分享的資源；格網系統需採取適當的資源分配策略分享、選擇並整合格網資源。格網系統之資源分配及管理之目標在於資源提供者與使用者之間必須建立一個協定以完成各自目標[8]。為了達成此一目標，資源提供者和使用者必須先了解格網系統的特性。

本研究利用 NASA 發展的 NGB benchmark 評估 Taiwan Unigrid 的系統效能並說明 Grid Economy model 對於資源分配及管理的重要性。我們發現針對不同性質工作以及使用者的 QoS，選擇適當的格網資源於執行工作時有助於提高 CPU 使用效率。當工作具有相依性，若以隨機方式分配資源執行工作時，次級效能的資源會因為執行效率低於平均執行時間，而使後續的資源使用延長等待時間，影響整體使用效率；當工作具獨立性質(例如 bag-of-task 程式)時，次級效能的資源雖不會影響其他資源的使用效率，卻同樣會拖累整體執行時間，同時該資源節點成為瓶頸所在。若使用者根據 QoS 選擇適當資源群組執行工作，則可讓次級效能的資源有更好的使用機會。

2.相關研究

2.1 Grid Economy

為了利用格網系統之服務滿足大量計算及大量儲存資料的需求，適當地調配格網資源及工作排程是最大的挑戰。傳統 System-centric 的資源分配策略強調資源使用效率的最大化，而不以使用者為思考中心的策略讓使用者無法根據本身的需求完成工作；而 User-centric 策略可讓使用者評估該如何分配資源滿足所需，因此 User-centric 的資源分配及管理方式是一種符合使用者經濟原則。在此原則下，工作排程及資源分配的決定都是動態的，可以隨時根據使用者的需求以及資源的加入/退出做彈性調整，以人類經濟法則，在 Grid 環境底下依據排程做資源分享以及分配調度。

在市場經濟原則下的格網系統必須有一個工具提供給使用者以及資源提供者表示他們的需求和

目標。使用Grid Economy的優點有[1]:幫助建立大型格網系統讓資源提供者遵循此交易模式提供資源;第二是讓使用者有個標準藉以評估本身需求強度,deadline越短表示需要越多資源在短時間內完成工作;此外Grid Economy幫助建立high-scalable的決定機制讓不同地理位置的使用者和資源提供者都可參與;最後,當此市場經濟行為模式建立之後,便產生一個標準,標準化使得商品能夠多樣化,不同的資源種類都可利用相同的經濟模式分享,因此除了計算資源,所有類型的資源都可藉由此模型進行交易。

Dr. Buyya 根據 Grid Economy 發展 Grid Architecture for Computational Economy (GRACE) [1]。資源提供者利用 GRACE 定義資源價格及管理策略;使用者則依賴 GRACE 的 Broker 與資源提供者互動。

本研究試圖驗證以 Grid Economy 的原則分配資源確實有效,不只希望使用者能在滿足 deadline 和 budget 的情況之下完成工作,同時藉由此原則讓資源發揮更大效益,未來格網技術發展更加成熟時,可藉由此交易原則,分享更多資源給使用者。

2.2 Grid benchmark

目前只有少數的 benchmark 專門設計用來檢測格網系統之效能,且因格網系統的變動性(dynamic nature)[2]使得要評估格網系統真正特徵仍有其困難度存在。

GridBench[6][12]為 CrossGrid 的子計畫,是基於 Globus toolkit 和 EU DataGrid 發展而來。GridBench 模擬各個格網元件包括 Computing Element (CE)管理 Worker(WN)執行計算,Storage Element (SE)負責儲存大量的資料,這些元件可能包含在一個或多個 Virtual Organizations(VOs)之內,各個 VO 提供 GIS、Resource Broker 等各式格網服務。GridBench 的目的在於希望根據計算能力、檔案傳輸等項目評估資源的效能;另一個目的在於提供使用者一個工具做為未來在格網環境上執行應用時的標準和依據。GridBench 以基於 XML 的

GridBench Definition Language(GBDL)描述執行[8][5][6][12],圖 2.1 為 GridBench 的架構圖。

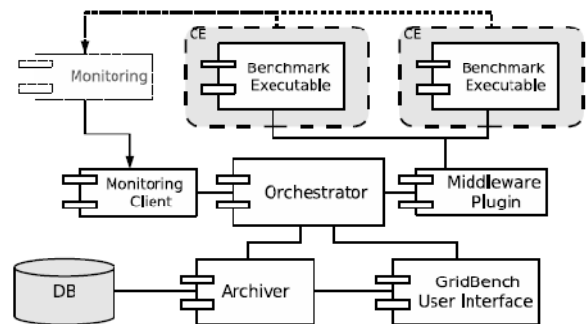


圖 2.1 Software Architecture of GridBench

NGB(NAS Grid Benchmarks)根據 NAS Parallel Benchmarks 發展做為 Grid benchmark,擁有四種 problem size 分別為 S、W、A 和 B。

Embarrassingly Distributed (ED) problem 多次執行相同的程式。在 ED 架構中,SP 根據 problem size 被執行多次,SP 之間沒有資料的傳輸各自具有獨立性,故 ED problem 為鬆散結構應用的代表檢測程式。

Helical Chain (HC) problem 是重複執行一系列包含 BT, SP,以及 LU 的程式,前一個執行的結果丟到後一個繼續執行。後一個 task 必須等待前一個 task 完成才可開始運作,具有相依性,因此 HC 使用來測試序列化應用的執行效能。

Visualization Pipeline (VP) problem 包含 BT、MG 以及 FT,分別扮演 flow solver、post processor、以及 visualization module respectively,以這三個模擬 pipeline 程序。Mixed Bag (MB) problem 類似 VP 但它是不對稱的,執行 MB 時,某些 tasks 的 workload 也許會比其他 tasks 來得大。

NGB 主要專注於格網系統執行分散式傳輸程序的能力。對於其他 Grid computing components 則未具體指出如何執行或選擇,例如 scheduling、grid resource allocation 和 security 等[8][15]。

在文獻[8]中嘗試利用 NGB 檢測 cluster grid 的效能試圖找出格網系統的特性,並發現傳統在 parallel computer system 上的應用不全然可移植到

格網系統執行。它以 system-centric 的角度探討格網效能，並未考慮使用者的需求，因此本研究希望以 user-centric 面向評估格網環境，利用 Grid Economy 原則調配資源以符合使用者的 QoS。本研究將實驗環境擴大，利用 Campus 與 Campus 之間連結而成的 Taiwan-Unigrid 做為實驗平台，以期更符合真實的格網環境，並試圖驗證 Grid Economy 理論應用對於格網環境的影響。

我們也曾經利用 NPB benchmark 檢測 TIGER Grid[16]的效能，發現次級效能的資源會拉長 turnaround time[17]，本研究根據 Grid Economy 原則選擇適當的資源執行工作以降低 turnaround time，而 deadline 時間較長時，選擇次級效能的資源執行可提高 CPU 使用效率。

3. 實驗設計

3.1 Taiwan Unigrid

Taiwan Unigrid 由清華大學(NTHU)等單位初始建置，目前有三十多所學校與研究單位參與計畫，遍及台灣各地，為目前台灣規模最大的格網計畫，圖 2.2 為 Taiwan Unigrid 環境架構圖。

Taiwan-Unigrid 主要提供龐大的格網計算資源供合作單位使用，並在其上進行格網相關技術研究、格網程式設計、執行大型應用程式等，藉以分享和推廣格網運算平台建置及應用經驗[14]。

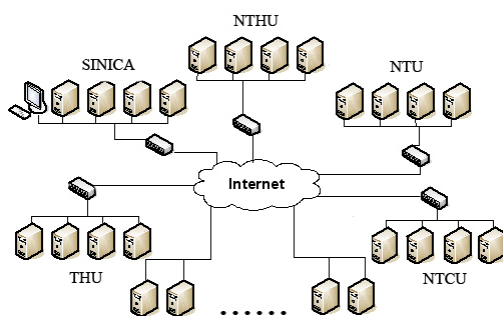


圖 2.2 Taiwan-Unigrid 環境架構圖

3.2 實驗環境設計

本研究使用 GridNPB3.0-JAVA-VERSION 做

為評估工具，problem size 設定為最小的 S 級。NGB benchmark 專注於探討效能的檢測，對於其他組成 Grid Middleware 如安全性等暫時不列入考慮。

本研究為初探性實驗，以 Taiwan-Unigrid 三分之一的機器為實驗對象。利用資源的 cpu_total_speed 為標準分成三個群組，將各 site 分為三個 class，10000 MHz 以上為 A class；6000~7000MHz 之間為 B class；1000~3500MHz 則為 C class。

實驗初始我們以隨機選取機器的方式在 Unigrid 挑選機器執行 ED.S、HC.S、MB.S、VP.S 各 20 次取平均值觀察 turnaround time，並紀錄每次執行時間的變動情況。接著從 A、B、C 三個群組選取機器執行 ED.S、HC.S、MB.S、VP.S 各 20 次同樣取平均 turnaround time。

4. 實驗結果

圖 4.1 為各個群組執行四種 problem 的平均執行時間。可以看到 A class 的執行時間優於其他群組，C class 平均執行時間最長，代表機器的計算能力確實影響每種 problem 的執行時間。但以隨機方式的執行時間只優於 C class，表示隨機選取機器執行沒有效率，尤其在執行 ED.S 的時候和 A、B class 差距最明顯，這是因為 ED.S problem 由九個 independent task 組成，彼此之間沒有進行 communication，各個 task 完成時間依據執行機器效能而定，最慢的機器執行時間便是 ED.S 整體執行時間。

執行 HC 可以發現隨機的品質在 A class 到 C class 之間起伏不定，若 task 由較好的機器執行，可縮短單一執行時間，但並不能保證在一定時間內完成。執行 MB 和 VP 時有某些 task 運算比例會較大，若此時以隨機方式選取到效能較差的機器負責執行，則該機器的負載過重便形成執行時間的瓶頸，因此隨機方式選擇機器同樣會有上下起伏的情況出現。因為無法預測 overhead 較大的 task 會由哪一台機器負責計算，對於使用者並無保障。

圖 4.2 到圖 4.5 為 ED.S、HC.S、MB.S、VP.S 四種 benchmark 各執行 20 次得到的結果，可以看到

以隨機方式選取節點時無法得到穩定的執行時間，如此不能保證滿足使用者的 QoS，且會因為某些機器效能過差形成瓶頸而拖累其他機器的執行效率。將機器以計算效能做簡單分類之後發現得到的執行時間相對穩定許多，這是因為機器效能相近，可以大幅提高服務的穩定度，同時也不會浪費機器的使用率。C class 機器表現不好是因為該群組機器已低於整體系統之平均效能，因此執行時間本就和其他群組的機器有所差距，但相對來說使用 C class 同樣能夠保證在一定時間之內執行工作完畢，若執行時間較長可採用此類機器執行工作。

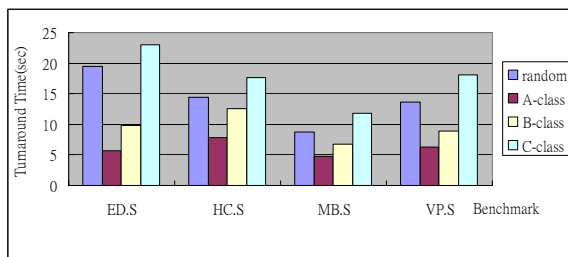


圖 4.1 四種 problem 平均執行時間

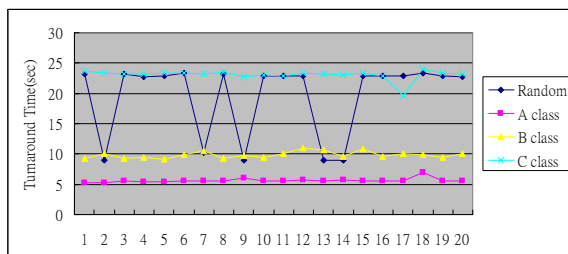


圖 4.2 ED.S 之 Turnaround time

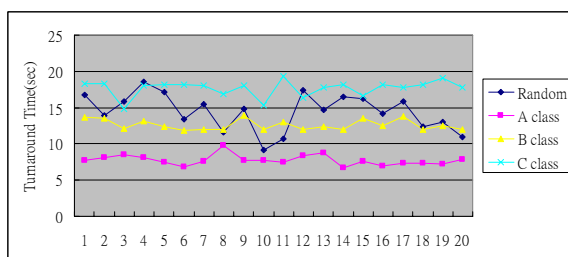


圖 4.3 HC.S 之 Turnaround time

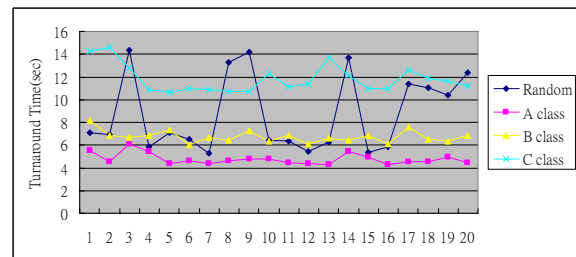


圖 4.4 MB.S 之 Turnaround time

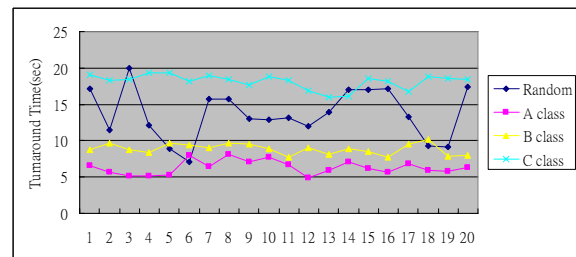


圖 4.5 VP.S 之 Turnaround time

一般來說，效能越好的機器價格必定越昂貴，而不同使用者的 deadline 和 budget 也不盡相同，因此使用者必須選擇最適合自己目標的資源執行工作才能達到最大經濟效益，但在此之前，資源必須能夠提供穩定的執行時間讓使用者藉此預估完成時間才能正確選擇適當資源。

利用計算效能因素將機器做簡單分類，若使用者 deadline 時間較短便選擇計算效能較高但 price 較高的資源做執行；若 deadline 較長或 budget 較低時就可選擇計算效能較低的資源執行。如此次級的機器不會拖累整體效能。雖然 B、C class 機器需要較多時間完成工作，但使用者可根據 benchmark 的結果和預估本身工作的完成時間，如此便能保證所有工作一定可在時間內完成，且使用者可根據資源的 cost 預估本身花費的多寡。

5. 結論與未來方向

我們發現以隨機方式選取資源執行工作有時會因效能較低的資源影響而降低整體效率，當利用效能相近的機器執行工作會提昇機器使用率且得到較穩定的結果。此外使用者可以根據 deadline, budget 以及 resource cost 選擇適當的群組執行工作，

不但可以滿足使用者的 QoS，對於工作結果也更有保障。

本研究以 NGB benchmark 做為檢測工具，發現不論使用者工作類型是否有相依性或獨立性，利用 Grid Economy 的原則選擇資源執行可擁有較高的經濟效益。未來研究方向希望能夠擴大至整個 Taiwan Unigrid 格網環境；並考慮加大 problem size 以獲得更精確的結果；另外考慮加入其他格網環境的因素，以便真實反映格網系統的特性。

本次研究採用 Grid Economy 中最常被考慮的三種因素：Resource cost、User deadline 以及 User budget，若使用者想在最短時間內完成工作就必須尋找 price 較高的資源執行花費較多的 budget；反之若使用者 budget 較少，便必須放寬 deadline 尋找 price 較低的資源執行，目前已有多種經濟模式演算法在發展並實際應用在格網系統或是利用模擬工具進行測試。為了更了解 Grid Economy 的作用，我們正在嘗試利用 Gridsim 工具模擬彈性競價模式在格網上工作排程的調配，希望找出更有效率的方法。

Acknowledgement

This work was sponsored in part by the National Science Council of the Republic of China under the contract number: NSC 95-2622-E-142 -001 -CC3.

參考文獻

- [1] Buyya R, Abramson D, Venugopal S. The Grid economy. *Proceedings of the IEEE* 2005; 93(3): 698-714.
- [2] Buyya R, Murshed M. GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing. *Concurrency and Computation: Practice and Experience* 2002; (May).
- [3] Buyya R, Stockinger H, Giddy J, Abramson D. Economic models for management of resources in peer-to-peer and Grid computing. *SPIE International Conference on Commercial Applications for High-Performance Computing*, Denver, CO, 20-24 August 2001.
- [4] Foster I, Kesselman C (eds.). *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan Kaufmann: San Francisco, CA, 1999.
- [5] G. Tsouloupas and M. D. Dikaiakos. Design and Implementation of GridBench. In *Proceedings of the European Grid 2005 Conference*. To appear as Volume in *Lecture Notes in Computer Science*, Springer, February 2005.
- [6] Gridbench. <http://www.crossgrid.org/products/otherproducts/gridbench.html>.
- [7] MPICH-G2. <http://www3.niu.edu/mpich/>.
- [8] Peng L, See S, Song J, Stoelwinder A, Neo HK. Benchmarks performance on Cluster Grid with NGB. *Proceedings of the High-Performance Grid Computing Workshop, 18th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Santa Fe, NM, April 2004. IEEE Computer Society Press: Los Alamitos, CA, 2004.
- [9] Pourebrahimi B, Bertels K, Kandru G.M., Vassiliadis S. Market-Based Resource Allocation in Grids. *e-science*, p. 80, Second IEEE International Conference on e-Science and Grid Computing (e-Science'06), 2006.
- [10] Pourebrahimi B, Bertels K, Kandru G.M., Vassiliadis S. Market-Based Resource Allocation in Grids. *e-science*, p. 80, Second IEEE International Conference on e-Science and Grid Computing (e-Science'06), 2006.
- [11] Sai Rahul Reddy P. Market Economy Based Resource Allocation in Grids. *Indian Institute of Technology, Kharagpur, India*, 2006.
- [12] Tsouloupas G, Dikaiakos MD. GridBench: A tool for benchmarking Grids. *Proceedings of*

- the 4th International Workshop on Grid Computing (Grid2003). IEEE Computer Society Press: Los Alamitos, CA, 2003; 60-67.
- [13] Unibox.
<http://gridv6.csie.chu.edu.tw/blog/unibox>
- [14] Unigrid. <http://www.unigrid.org.tw/>.
- [15] Van der Wijngaart RF, Frumkin MA. Evaluating the Information Power Grid using the NAS Grid Benchmarks. Proceedings of the High-Performance Grid Computing Workshop, 18th IEEE International Parallel and Distributed Processing Symposium (IPDPS), Santa Fe, NM, April 2004. IEEE Computer Society Press: Los Alamitos, CA, 2004.
- [16] Yang C.T., Li K.C., Chiang W.C., and Shih P.C., Design and Implementation of TIGER Grid: an Integrated Metropolitan-Scale Grid Environment. Proceedings of the 6th IEEE International Conference on PDCAT'05, pp. 518-520, Dec. 2005.
- [17] 黃振維, 楊大猷, 吳勇均, 賴冠州, 楊朝棟.
NPB 平行程式在計算格網之效能評估