NPB 平行程式在計算網格之效能評估

黃振維	楊大猷	吳勇均	賴冠州	楊朝棟
台中教育大學	台中教育大學	台中教育大學	台中教育大學	東海大學
數位內容科技學系	數位內容科技學系	數位內容科技學系	資訊科學學系	資訊工程與科學系
smallway11@gmail.com	hanamichi1111@gmail.com	firebird0103@hotmail.com	kclai@mail.ntcu.edu.tw	ctyang@thu.edu.tw

摘要

NAS Parallel Benchmark (NPB) 是 NASA 爲解流體力學而開發的平行程式，現在多用來做平行計算效能測試標準。本研究目的爲測試並觀察 NPB2.4 Benchmark 在網格環境中執行的結果與其負載上限；並說明依據工作性質的不同，計算資源的調配會影響執行時間以及執行效率。

關鍵詞：NPB, Grid, TIGER

1.前言

網格計算(Grid Computing)是新興的分散式計算系統，爲一個動態的、多個虛擬組織及不同管理者共同整合資源的環境，用來解決工業、科學以及分子運算等大型問題。網格環境可分享的資源從簡單的檔案傳輸到利用遠端的電腦、軟體、資料以及其他可利用網路存取的資源，進一步發現、分配並評估以用來分享、選擇並整合這些不同的資源，這些資源有可能是分散在不同地區由不同組織擁有的超級電腦、儲存系統和其他特定裝置。網格可應用在模擬環境、醫藥醫學、高等物理等領域[5]。

美國在 1984 年成立數值空氣動力模擬計畫(The Numerical Aerodynamic Simulation Program)，簡稱 NAS，其目的除了解決傳統的流體力學問題之外，也整合最新的硬軟體技術提供給 NASA 進行相關研究使用。

NAS 的主要目標是建置一個龐大的計算平台模擬航太飛行系統，以加速航太飛行設備的設計時程，此時必須開發電腦運算能力以加速模擬過程。而 NAS Parallel Benchmark(NPB) 即爲評估電腦運算效能之效能評估程式[7]。

NPB以Fortran和C撰寫，2.0版開始以MPI爲基礎開放原始碼提供下載，其中包含八個測試程式MG、CG、FT、IS、EP、BT、SP和LU。

本研究主要使用NPB2.4的MG、CG、IS以及EP來測試網格運算環境的負載上限。

2.相關研究

2.1 NAS Parallel Benchmark (NPB)

NAS 於 1991 年提出 NPB1，採用“Paper and Pencil”效能評估，其運算方式不受資料結構、數值方法、處理器的配置方式及記憶體的使用等的限制，單純考慮運算法則找到問題之最佳解[1]。

其後，於 1995 年發展出以 MPI 平行化程式爲基礎的 NPB2.0，並根據各程式之問題大小將其區分爲 Class S, A,與 B 三個層級；於 1996 年公佈 NPB2.2，其測試標準增加 Class W,與 C 兩個層級;之後2002年所發佈的NPB2.4版本再增加Class D 層級。

NPB2 爲一開放性原始碼，可供不同組織進行系統效能之測試。其原始碼主要以 Fortran 撰寫(除 IS 以 C 語言之外)。NPB2.0 版本改以 MPI 做爲平行計算環境之基礎，透過各種不同的 benchmark 來進行平行計算之效能評估。讓不同規格，不同平台的電腦藉由共同介面達到平行計算的目的。平行計算以區域分割爲主要的算則[4]。

NPB 有八種檢測標準，每種標準之 problem size 分有 S、W、A、B、C、D 六個 class，各個 benchmark 簡述如下[1,6]。

- **Integer Sort (IS)：**測試整數計算速度還有傳輸速率，不包含浮點運算。
- **Fast Fourier Transformation (FT)：**測試長途資料的通訊傳輸以及利用快速傅立葉解決 3D 偏微分方程（PDE）。
- **Multigrid (MG)：**爲求 3D Poisson 偏微分方程，主要檢測高度結構化的短距離及長距離之資料傳輸。
- **Conjugate Gradient (CG)：**與 MG 不同，CG 使用共軛梯度的方式解大型稀疏對稱有限矩陣的最小特徵值，使用不規則的長途資料傳輸和稀疏矩陣向量乘法非結構化的網格計算。
- **Lower-Upper (LU) diagonal：**測試 fine-grain 的區域傳輸，以 Symmetric Successive Over-relaxation (SSOR) 求解 Regular-sparse Block 5×5 的上三角與下三角矩陣系統，因 LU 以”block”方式執行 MPI，因此阻塞 bottleneck 產生在網路傳輸之處。
- **Scalar Pentadiagonal (SP) and Block Tridiagonal (BT)：**檢測計算量與傳輸通訊之間的平衡，執行 SP 與 BT 的 process 數量必須是 N 的平方個數，基本上 SP 與 BT

類似，NASA Ames 所用的程式如 ARC3D 就是 SP 和 BT 的典型代表，不過在資料傳輸和計算速度比方面，SP 比 BT 偏向傳輸密集度。

- **Embarrassingly Parallel (EP)**：屬典型的 Monte Carlo 應用，測試浮點數運算，EP 的特性是不計算處理器之間的傳輸，因此產生的結果可以等作於被測試系統的浮點運算上限。

目前平行計算已實際應用在各個研究領域，如大氣模擬、生物資訊、醫學、物理學等等甚至軍事研究方面；相較於超級電腦，Grid 提供更高的延展性及易於維護的優點，且網路技術的發展，讓計算資源、儲存空間以及使用者可以不受地理上的限制影響研究發展。因此在 Grid 環境上進行平行計算是未來需要大量運算資源的研究領域不可或缺的技術之一。

2.2 Taichung Integrating Grid Environment and Resource (TIGER)

Taichung Integrating Grid Environment and Resource(TIGER)基於Globus Toolkit version 4.0.1 建置分散式網格計算環境，目前TIGER參與單位包括東海大學（THU）、台中教育大學（NTCU）、修平技術學院（HIT）、大里高中(DL)，以及立人高中(LZ)， TIGER主網路骨幹爲1GB的TANET，TIGER的環境架構如圖1所示，環境軟體架構如圖2所示[3]。

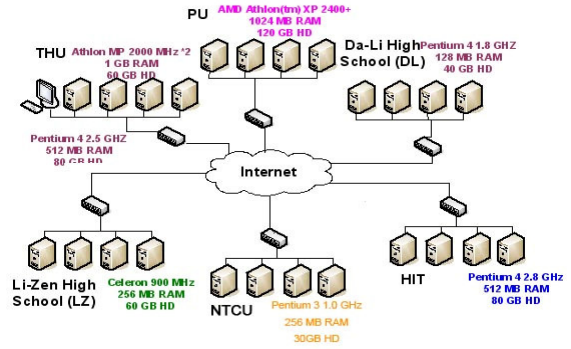

圖 1 TIGER 環境架構

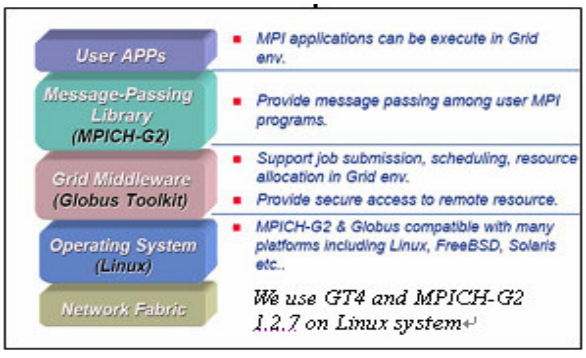

圖 2 TIGER 軟體架構

3.實驗設計

本校高效能計算實驗室於 2006 年六月參與 TIGER 網格環境，目前已加入四個 node 參與計算，未來將逐步增加計算資源。TIGER 環境以 Globus Toolkit 4.0.1 爲基礎，檢測程式爲 MPI-based NPB 2.4，FT 需 FORTRAN 90 編譯器，目前實驗環境尚未安裝；BT 及 SP 以 N^2 的 process 數爲測試對象；LU 以"block"方式執行 MPI 不爲本次研究範圍，故這次以其中的 MG、CG、IS、EP 四種 benchmark 作爲檢測標準。EP 不計算傳輸時間，單純計算各個 process 浮點計算的上限，可視爲 bag-of-tasks application 的測試；MG、CG、IS 則包含計算與傳輸時間，每個工作之間有相依性存在，problem size 爲 class S、A、B，處理器的數目分別設定爲 1、2、4、8、16，各個 problem size 的大小如表 1 所示。

爲了將 Grid 異質性及變動性的影響降到最小以便測得較精確的結果， MG、CG、IS、EP 以不同的 problem size 各執行十次取其平均值。

表 1 MG、CG、IS、EP problem size

Benchmark	Class S	Class A	Class B
Multigrid(MG)	32^3	256^3	256^3
Conjugate Gradient(CG)	1400	14000	75000
Integer Sort(IS)	65536	2^{23}	2^{25}
Embarrassingly Parallel(EP)	33554432	2^{28}	2^{30}

4.實驗結果

測試結果的平均數據如表 2、表 3 所示。當 problem size 一樣時，單一 process 的計算量會隨著 process 數的增加而減少，證明 grid 分散式計算環境確實能有效分散執行的工作量。當參與的 process 數量越多時，grid 便可將單一工作切割成更多小 size 的子工作分配給在不同地區，不同管理者的 process 執行，因此每個 process 不需要作太多的計算即可完成工作，本研究利用 EP 驗證此類型工作在 grid 環境上執行的狀況。

但是當子工作之間彼此有相依性存在時，總執行時間必須考慮 communication time 的影響。此研究採用 MG、CG、IS 三種 benchmark 做爲檢測的標準，實驗結果發現當工作有相依性時，工作切割越細，communication time 便隨之增高，造成雖然各個 process 計算量減少，但整體執行時間反而上升的情況發生；且因爲每次參與測試的 resource 會隨著時間不同而改變，因此造成整體執行時間有時會呈現不規則起伏，也反應了 grid 環境高度的異質性以及變動性。

表 2　average time in seconds

	1	2	4	8	16
EP-S	5.498	3.15867	1.58867	0.982	0.752
EP-A	86.99267	50.48933	25.37867	21.374	11.21333
EP-B	N/A	202.4833	101.682	86.122	43.31733
IS-S	0.03133	0.03067	0.03	2.65533	3.65067
IS-A	3.97267	3.49667	2.36067	21.58667	19.29133
IS-B	16.8367	16.54933	9.85133	78.58867	64.00467
MG-S	0.026	0.07267	0.08533	1.09067	6.618
MG-A	6.87533	7.08867	4.086	10.928	15.826
MG-B	33.37667	35.91333	21.4	69.70467	79.17
CG-S	0.260667	0.43467	0.40467	1.402	53.18467
CG-A	5.572667	3.89533	2.364	50.83867	97.12267
CG-B	220.5207	196.7233	114.2407	270.528	N/A

表 3　average Mop/s/process

	1	2	4	8	16
EP-S	6.11133	5.32	5.28467	4.50533	3.38067
EP-A	6.17067	5.318	5.28867	3.38867	3.2
EP-B	N/A	5.30333	5.28067	3.374	3.338
IS-S	21.288	11.882	4.80467	0.032	0.01667
IS-A	21.11267	12.006	8.904	0.49	0.28267
IS-B	19.98333	10.728	8.51867	0.53467	0.33
MG-S	329.98533	50.724	23.064	1.49067	0.706
MG-A	566.13067	274.638	238.20733	46.23733	18.054
MG-B	591.976	283.19267	236.16133	43.15067	18.3027
CG-S	266.03667	76.74333	41.24067	6.14333	0.94533
CG-A	269.64267	194.568	158.40733	7.3	2.24533
CG-B	248.238	139.07533	119.75067	25.32133	N/A

圖 2、圖 3 爲執行 EP 測試 bag-of-tasks application 的結果，整體計算時間會隨著 process 數增加而下降，每個 process 的計算量也會隨著 process 數增多而減少，顯示工作有效分配至各個 process 進行計算； problem size 屬於 class B 時單一個 process 的能力則無法負荷如此大的運算量。

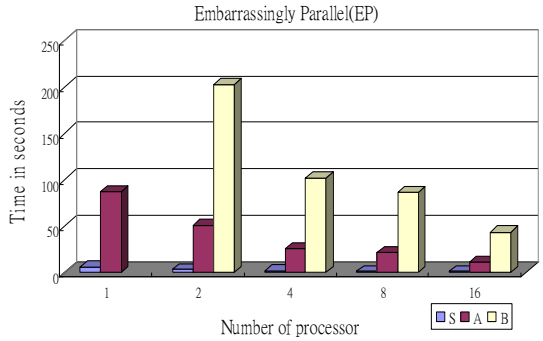

圖 2　EP 測試結果

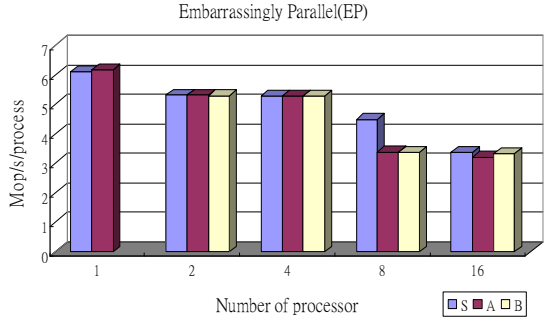

圖 3　EP 測試結果

圖 4、圖 5 是執行 IS 的結果，可以看到 process 數增加反而讓整體執行時間拉長，當 process 數爲 1、2、4 時的執行時間短且效率高，process 爲 8 和 16 個時，不僅完成時間大幅拉長，執行效率也很低，說明此種工作不適宜切割過細。

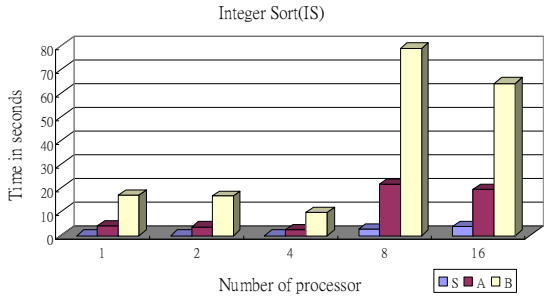

圖 4　IS 執行結果

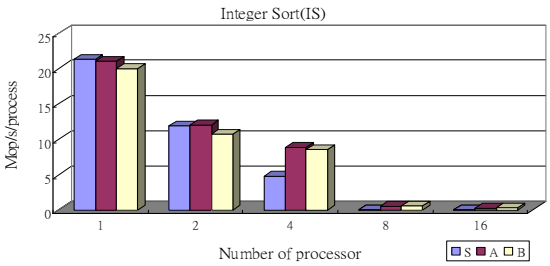

圖 5　IS 執行結果

圖 6、圖 7 是執行 MG 之後的結果，每個 process 平均分配到的計算量會隨著 process 增加而減少，但整體計算時間卻會隨著 process 數增加而增多，有可能是因爲 communication time 變多造成的結果。當 process 數等於 4 時的計算時間最短，process 數目越多的效率越來越低。

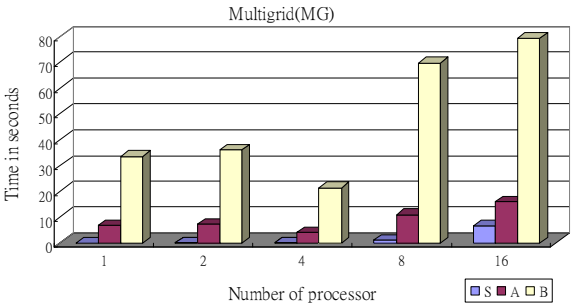

圖 6　MG 執行結果

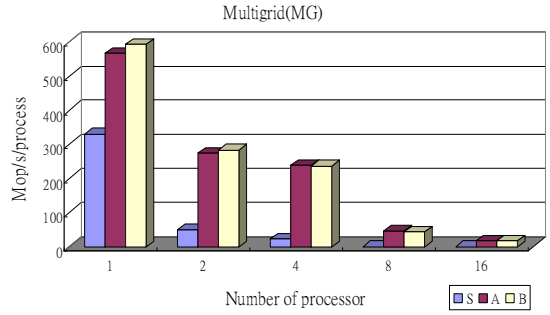

圖 7　MG 執行結果

圖 8、圖 9 爲執行 CG 的結果，可以看到和 IS 以及 MG 相同的情況，class B 雖然隨著 process 數的增加讓執行時間降低，但各個 process 的執行效率也會隨著數量增加而降低，因此造成 process 數大於等於 16 之後無法執行成功。

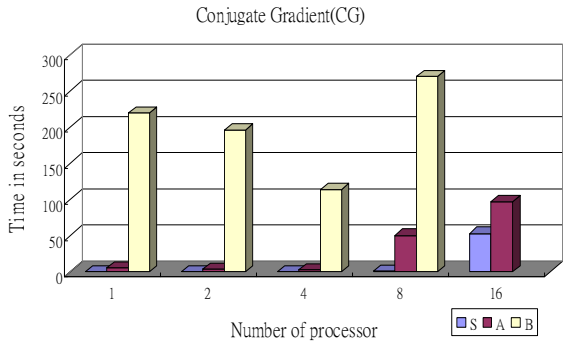

圖 8　CG 執行結果

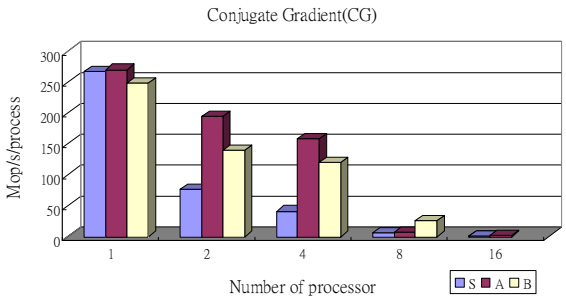

圖 9　CG 執行結果

5.結論與建議

本研究目的在於利用 NPB 平行處理程式來測試 TIGER 環境執行平行計算的負載上限。實驗發現，當執行平行計算工作時，除了工作計算所需的時間之外，切割成子工作的 communication time 也是影響執行時間以及 process 執行效率的重要因素；同時 TIGER 環境也隨著 benchmark 的 problem size 加大而出現負載上限的狀況。因此當執行平行計算工作的時候，必須依據工作性質的不同，根據計算環境的情況切割成最適當的大小，除了減少總體執行時間之外，各個 process 執行效率的經濟效益也要加以考慮；且在本次研究當中可以看到 grid 環境的異質性和高度的變動性對計算的影響。

本次實驗以 MPI-based 的 NPB 2.4 作爲標準，以 MG、CG、IS、EP 四種 benchmark 作爲檢測標準，problem size 爲 S、A、B，未來可增加檢測標準和擴大 problem size 以得到更詳細的結果。

未來利用 NPB 檢測 grid 計算環境時，必須考慮到 grid 高度的異質性和變動性，加入 WMS 以及 monitoring 的服務機制蒐集網路資訊以及各個 process 處理工作的詳細情況，相信對於分析平行計算工作會有很大的幫助，進一步研究 job schedule 用來發展 resource broker 讓使用者可以不需考慮工作切割以及選擇 process 的繁複工作，電腦即可自動完成且達到最高效率和最短時間的目標。

參考文獻

[1] 沈澄宇, 高效能計算簡介, 1998。

[2] 林志漢, 使用 GlobusToolkit 建立 Grid Computing 環境的初步結果, 2000。

[3] Chao-Tung Yang, Kuan-Ching Li, Wen-Chung Chiang, Po-Chi Shih, “Design and Implementation of TIGER Grid: an Integrated Metropolitan-Scale Grid Environment”, IEEE PDCAT, 2005。

[4] Henry Jin, Michael Frumkin, Parkson Wong, Huiyu Feng, “Update on the NAS Parallel Benchmarks”, Computer Sciences Corporation NASA Ames Research Center。

[5]Sai Rahul Reddy, Market Economy Based Resource Allocation in Grids, Master Thesis, Indian Institute of Technology, Kharagpur, India, May 2006.

[6] Rizwan Ali; Yung-Chin Fang; Victor Mashayekhi, Ph.D.; and Reza Rooholamini, Ph.D., “Evaluating High-Performance Computing Clusters Using Benchmarks”, 2001。

[7] “NAS Parallel Benchmarks Version 2.4 ”, NAS Technical Report NAS-02-007, October 2002。