

PC Cluster 作為高效能科學及工程計算平台之效能評估

周朝宜

鄭守成

黃國展

張西亞

國家高速電腦中心

新竹市科學園區研發六路七號

E-mail: {b00cyc00, c00tch00, c00kch00, c00jhc00}@nchc.gov.t

摘要

本篇論文描述針對國家高速電腦中心 PC cluster 所進行的效能評估與分析。實驗的結果指出，PC cluster 具有優異的性能/價格比，並且對某些類型的應用程式而言，更有超越如 IBM SP 等傳統平行電腦執行效能的潛力。另外實驗結果也指出，PC cluster 的效能跟應用程式的特性息息相關。本篇論文的內容將可提供有興趣嘗試以 PC cluster 作為高效能計算平台的使用者有用的資訊，作為可行性分析及效能改進時之參考。

關鍵詞：PC cluster，高效能計算，效能分析，Benchmark，High Performance computing

一、前言

近年來，個人電腦 (PC) 的運算速度及記憶容量突飛猛進，直追傳統上的工作站甚至大型主機；相反地，價格卻不斷地調低，使得 PC 具有極佳的價格/效能比。同一時期內，網路相關技術亦有長足的進步。上述這些技術發展趨勢使得個人電腦叢集 (PC Cluster) 成為近年來學術界及產業界相當有興趣的一種新興高效能平行計算平台，國內外許多單位都正在嘗試此種新興平台的可行性 [16,18,19,22,23]。

國科會國家高速電腦中心 (以下簡稱本中心) 為因應國內各界對 PC cluster 與日俱增的興趣和需求，也積極地投入此領域之研究發展 [14]。本中心個人電腦叢集工作小組之任務目標包括：

- 與中心內各應用領域研究小組合作，評估 PC cluster 在各個應用領域內可發揮的潛在計算效能。
- 研究發展及推廣相關的平行程式發展技術。
- 進行 PC cluster 相關軟、硬體技術的研究發展與評估。
- 最重要的是希望能夠將 PC cluster 的建置與使用經驗推廣至國內有興趣的各個單位。

自今年初以來，本中心已成功地建置完成共 32 個 CPU 的 PC cluster，並針對 PC cluster 的計算效能進行多方面的分析。用來進行效能分析的效能評估程式集，範圍涵蓋國際上通用的標準程式集，如 Linpack [9]、NPB2 [6] 等，以及從各個領域收集來的各類科學與工程計算應用程式。本論文提供我們進行這一連串效能分析的經驗與結果。這些結果將可提供對 PC cluster 有興趣之高速計算相關人員一些有用的資訊，協助他們瞭解 PC cluster 在其應用領域內的潛力，以評估 PC cluster 是否足以成為傳統大型電腦之外，高速計算平台的另一種選擇。

二、NCHC PC CLUSTER

本中心目前有如圖一所示之兩組相連的 PC cluster。其中一個叢集由 16 部單 CPU 的 PC 所組成，另一叢集有 8 部雙 CPU 架構 (SMP) 的 PC。同一叢集內的 PC 經由一個 24 ports 的 Fast Ethernet switching hub 所連接，所以彼此間皆享有 100Mb/s 的通訊頻寬。兩個叢集間目前由一條 100Mb/s 的 link 所連接，因此跨叢集的通訊必須分享這 100Mb/s 的頻寬。將來，我們打算把兩個叢集的 switching hubs 堆疊起來，使得兩個叢集合而為一，成為一個具有一致性通訊頻寬的大叢集。

本中心 PC cluster 所使用的 CPU 是 Intel Pentium II 400 MHz 微處理器，每台 PC 配備 256 MB 的主記憶體以及 6GB 的硬碟。除了共 32 個 CPU 之 24 部專供計算的 PC 外，另外有兩台伺服器等級的 PCs，做為檔案伺服器、工作佇列管理、以及供使用者發展、編譯程式、及提出程式執行要求之用。每一台伺服器 PC 配備一顆 Intel Pentium II 400 MHz 微處理器、256 MB 主記憶體、以及 9GB 的硬碟。

本中心 PC cluster 所使用的作業系統是 Red Hat Linux 5.2 [17]，作業系統核心 (kernel) 已經升級至 2.2.0。程式語言編譯器包括：GNU C、C++、Fortran 77 編譯器、以及 PGIC、C++、Fortran 77、Fortran 90、及 HPF 編譯器 [10]。供撰寫平行程式之用的訊息傳遞程式庫 (message passing library) 有 MPI [3]、PVM [4]、及 BSP [7]。本叢集上目前使用美國佛羅里達州立大學超級電腦計算研究中心 (SCRI) 所發展的分散式佇列系統 DQS [11] 作為工作排程及管理之用。詳細的軟、硬體規格如表一所示。

三、基本效能參數分析

本節比較 NCHC PC cluster 與 IBM SP2 [8] 間基本效能參數的差異，IBM SP2 是目前國家高速電腦中心內最大的平行計算平台。首先，圖二顯示出 PC 的 Pentium II 400 MHz 微處理器有著比 IBM SP2 處理器較為高的時脈頻率；IBM SP2 最高的時脈頻率為 160 MHz。關於數值運算的效能，圖三顯示出 Pentium II 400 MHz 有較佳的整數運算效能，約為 IBM SP2 處理器的兩倍；想反地，其浮點數運算的效能就不及 IBM SP2，約只有其二分之一。

至於記憶體系統方面，如圖四中所顯示的，Pentium II 處理器跟 IBM SP2 的處理器比起來，有著較大的快取記憶體 (cache)。根據我們的實驗，有些應用程式可以受益於 Pentium II 處理器較大的快取記憶體，而達到較佳的執行效能。另一方面，根據表二及圖五，IBM SP2 有較高的主記憶體系統頻寬。因此，需要處理大量資料且主記憶體存取頻繁的應用程式，可以在 IBM SP2 上達到較好的執行效能。

另一個重要的效能參數是通訊頻寬，應用程式中所能使用到的通訊頻寬受到網路硬體、驅動程式、作業系統、通訊協定及相關軟體等因素的影響。圖六顯示使用MPI作為資料交換與訊息傳遞中介軟體之應用程式在這兩個平台上所能達到的通訊頻寬；MPI為目前使用最廣的資料交換與訊息傳遞中介軟體。圖六的結果指出IBM SP2配合它的high-performance switch [23]可以提供應用程式高於PC cluster之100Mb/s Fast Ethernet的通訊頻寬。這個結果隱含著一個訊息，必須進行頻繁的資料交換及工作同步化之平行應用程式將有可能在IBM SP2上獲得較佳的執行效能。

四、 NCHC 應用程式集

NCHC 應用程式集內共有八程式，涵蓋四個不同的應用領域：計算流體力學 (bem3d, ns3d)，計算固體力學 (finite, jcg3d)，物理 (pureg, hubksp)，以及化學 (modyn, mopac93) [24]。本節將描述用此應用程式集來進行效能分析的結果。表三顯示各個循序版程式在不同計算平台上的執行時間。表三的結果顯示出各個平台的相對執行效能依應用程式的不同而有所不同，深受諸如：機器指令比例 (instruction mix)、資料存取模式、資料大小等程式特性的影響。一般而言，能夠受益於較大之快取記憶體或是以整數運算為主的應用程式會在PC的平台上有着較佳的執行效能。

再看看平行處理的部分，圖七及圖八顯示 hubksp 及 modyn 這兩個應用程式的平行效能。針對這兩個程式而言，PC cluster比傳統的平行電腦呈現出更佳的執行效能，因為這兩個程式的特性能夠利用到PC架構的優點，並且計算過程中不需要太多的資料交換；在使用16個處理器時，通訊時間只佔整體執行時間的23%左右。分別言之，hubksp以整數運算為主，而modyn的資料量不大，可以放進PC的快取記憶體中。

相反地，如圖九所示，對那些以浮點數運算為主且需要進行大量資料交換的應用程式而言，傳統的平行電腦呈現出比PC cluster為佳的執行效能。圖九的 cw b 天氣預測程式是由中央氣象局所發展出來的東亞地區氣象預測系統 [2]。當這個程式在PC cluster上用到16個處理器時，有多達50%的執行時間是花在資料傳輸上。因為傳統的平行電腦如IBM SP2、HP SPP2200等，有較大的通訊頻寬，所以反映在圖九上的是傳統上的平行電腦有著比PC cluster為佳的效能延伸性 (scalability)。

五、 LINPACK

Linpack [9] 是超級電腦測試時最常用的評估軟體，主要是求解聯立方程式組之問題，分成兩種模式：一是求解100組聯立方程式，這不能修改程式，另一種是求解1000組聯立方程式，允許修改程式語法及演算法，使用者可依自己系統之特性調整問題的大小，其測試結果可以顯示該系統之可擴充性，目前的五百大 (Top 500) 超級電腦 [13] 即是以Linpack為標準，其詳細說明可參考 <http://www.top500.org/lists/linpack.html>。

圖十顯示在IBM SP及PC cluster上循序版本之Linpack程式之執行效能。這個結果顯示出針對這個以浮點數計算為主的程式，IBM SP2有著比PC cluster為佳的執行效能。圖中所顯示出的另一個結果為IBM SP2的倍精準度浮點數運算效能比單精準度浮點數運算為

佳；相反地，PC cluster的單精準度浮點數運算的效能比倍精準度浮點數運算為佳。

六、 NAS PARALLEL BENCHMARK

NAS Parallel Benchmark (NPB) [6] 是根據計算流體力學程式演化而來的效能評估程式集，主要是以Fortran 77及MPI寫成，包括八個部分：BT、SP、LU、MG、FT、EP、IS、CG，我們用來測試的版本是第2.3版。表四跟五描述NPB程式集的詳細型態與資料量大小。

圖十一顯示NPB程式集在NCHC PC cluster上的執行效能，結果顯示出，針對這個程式集，PC cluster至少具有16個處理器的延伸效能。圖十二以等級C的LU程式為例子，比較其在不同平台上的執行效能，結果顯示出PC cluster具有堪與其他平台匹敵的執行效能。而且，若以價格/效能比來看，圖十三指出，PC cluster在所有的測試平台當中，具有最優異的價格/效能比。

七、 SMP CLUSTER

包括Intel Pentium II處理器在內的大部分微處理器都支援多處理機PC架構，而且雙CPU的PC (two-way SMP PC)價格上只比單CPU PC稍貴一點。這個趨勢使得越來越多的PC cluster嘗試使用多CPU PC作為其基本組成架構 [1,5,15,20]。在國家高速電腦中心，我們也建立了一組由八部雙CPU PC所組成的叢集。本節將分析雙CPU PC叢集在執行效能上的特性，並與單CPU PC叢集做相對應的比較。

由於架構上的差異，雙CPU與單CPU PC叢集在記憶體頻寬、通訊頻寬、及整體執行效能等方面都有值得注意的差距；作業系統的支援在雙CPU架構的執行效率上也扮演著重要的角色。表十的第二及第三列比較雙CPU PC內兩個CPU間及不同PC間的通訊頻寬。雙CPU PC內兩個CPU間的通訊頻寬遠比兩部PC間的頻寬高出甚多。

雖然雙CPU架構有上述同一PC內高通訊頻寬的優點，但是卻也有著同一PC內的兩個CPU必須共享記憶體系統頻寬的缺點。所以雙CPU架構的整體利益依不同的應用程式而有所不同；尤其跟應用程式的資料存取、交換、傳輸等性質有密切的關係。

表六及七用Linpack 1000×1000來評估系統資源分享，尤其是記憶體系統頻寬分享，對雙CPU PC的執行效能所造成的影響。表中的第二列顯示出在一個單CPU PC上執行一個Linpack程式時的效能，第三列顯示在一部雙CPU PC上執行兩個Linpack程式時，每一顆CPU的平均效能。結果指出，因為系統資源分享的關係，雙CPU架構下每一顆CPU的平均效能跟單CPU PC比較起來低了不少。不過，當考慮效能/價格比的時候，如表七所示，雙CPU架構的PC就有可能表現得比單CPU PC來得好。

表八用NPB程式集進行與表六及表七相同的比較。表八的結果反映出與表七相同的趨勢。表九用兩個處理器的NPB平行版本來評估雙CPU架構PC的執行效能。針對每一個程式而言，第一列顯示使用兩部單CPU PC的執行效能，第二列顯示使用一部雙CPU PC的效能。這項測試綜合評估系統資源共享及高通訊頻寬優勢這兩項因素在執行應用程式時對雙CPU PC整體效能的

影響。結果指出，雙CPU架構雖然具有機器內部高通訊頻寬的優勢，但是由於系統資源共享所造成的影響，使得它的整體執行效能反而比不上兩部單CPU的PC。

對使用超過兩個處理器以上的應用程式而言，同一個雙CPU架構下的兩個處理器除了必須共享記憶體系統頻寬以外，同時也共享了網路系統的頻寬。表十的第三及第四列分析共享網路頻寬所造成的影響。第三列所顯示的情形為一部雙CPU PC內的兩個處理器同時與兩部單CPU的PC做資料交換，雙CPU PC內的處理器必須共享網路頻寬。第四列的情況為兩部雙CPU PC內的四個處理器同時互相在作資料交換，兩部PC內均有網路頻寬共享的現象。結果指出網路頻寬共享的影響不容忽視，雙CPU架構下的處理器所能用到的網路頻寬比單CPU PC內的處理器小得多，這個情形將對平行程式的執行效能產生一定的影響。這個結果同時反映出必須發展新的技術來提升雙CPU架構下的通訊頻寬，以改進雙CPU PC叢集的整體執行效能。

八、作業系統及編譯器

作業系統對PC cluster，尤其是SMP cluster，的效能有一定程度的影響。PC cluster系統上目前最通行的兩種作業系統為Linux及Microsoft Windows NT。為了比較這兩種作業系統可能造成的效能差異，我們將前一節所描述的測試在一部以Microsoft Windows NT為平台的雙CPU PC上也測了一遍。結果顯示，針對我們所測的效能評估程式及應用程式而言，這兩種系統並沒有造成明顯的效能差異。

另外一個影響執行效能的重要因素為程式語言編譯器。不同的編譯器採用不同的最佳化技術，因此所產生的機器碼也具有不同的執行效率。圖十四以一幅HINT [12]的效能圖形來描述不同的編譯器所可能造成的效能差異，數值越高表示執行效能越好。這個比較是在Linux的平台上所進行的。當程式需要的資料小於18060位元(Bytes)時，使用EGCS編譯之程式碼的執行效率最好，不過當程式所需要的資料量超過18060位元時時，使用PGI編譯之程式碼的執行效率最好。

九、結論

本篇論文呈現我們在NCHC PC cluster上關於效能評估與分析的經驗。這些結果將可提供有興趣以PC cluster作為高性能科學與工程計算平台的使用者不少有用的資訊，作為可行性評估或效能改進時的參考。實驗的結果指出，PC cluster可以較為優異的性能/價格比提供堪與傳統平行電腦相比擬的計算效能。對某些類型的應用程式而言，PC cluster甚至呈現出比傳統平行電腦更為優異的執行效能。這些結果顯示出，PC cluster的效能跟應用程式的特性息息相關。因此對應用程式特性詳細的瞭解與分析是使用PC cluster前的先決工作。

目前逐漸蔚為主流的SMP cluster有著比單CPU cluster更好的效能/價格比，不過在記憶體與網路系統方面仍有資源共享所造成的效能減低等問題必須克服，以進一步提升其整體效能。另外，MPI及PVM等平行軟體針對SMP架構所做的最佳化也可以進一步提升應用程式的執行效能，而編譯器的選擇也對程式執行效能有一定的影響。

奠基於目前效能分析的結果，接下來我們將調整網路架構並發展適合SMP cluster的平行程式撰寫技巧，以求提升使用者使用PC cluster時的整體效能；另外我們也計畫擴大PC cluster的規模，以探討建構大型叢集所可能面對的軟、硬體問題。

參考書目

- [1] F. Cappello, O. Richard, and D. Etiemble, "Performance Evaluation of Two Programming Models for a Cluster of PC Biprocessors," *1999 International Conference on Parallel and Distributed Processing Techniques and Applications*.
- [2] H. Y. Chang, S. C. Tcheng and K. C. Huang, "Performance of a Weather Prediction Model on Parallel Machines," *Proceedings of 1999 Advanced Simulation Technologies Conference, High Performance Computing Symposium*, The Society for Computer Simulation International, San Diego, USA, April 11-15, 1999.
- [3] A. Geist, A. Beguelin, J. Dongarra et. al., *PVM: Parallel Virtual Machine A Users' Guide and Tutorial for Networked Parallel Computing*, The MIT Press, 1994.
- [4] W. Gropp, E. Lusk, A. Skjellum, *Using MPI: Portable Parallel Programming with Message -Passing Interface*, The MIT Press, 1994.
- [5] HPVM Project, <http://www-csag.ucsd.edu/projects/clusters.html>
- [6] <http://science.nas.nasa.gov/Software/NPB>
- [7] <http://www.bsp-worldwide.org/>
- [8] <http://www.nchc.gov.tw/RESEARCH/spring/info/>
- [9] <http://www.netlib.org/linpack/index.html>
- [10] <http://www.pggroup.com>
- [11] <http://www.scri.fsu.edu/~pasko/dqs.html>
- [12] <http://www.scl.ameslab.gov/Projects/HINT>
- [13] <http://www.top500.org/>
- [14] NCHC PC Cluster Project Homepage, <http://www.nchc.gov.tw/RESEARCH/pccluster>
- [15] NT Supercluster in NCSA, <http://www.ncsa.uiuc.edu/SCD/Hardware/NTCluster/>
- [16] G. F. Pfister, *In Search of Clusters*, Prentice-Hall, Inc., 1998.
- [17] *Red Hat Linux 5.2: The Official Red Hat Linux Installation Guide*, Red Hat Software, Inc., 1998.
- [18] C. Reschke, T. Sterling, D. Ridge, "A Design Study of Alternative Network Topologies for the Beowulf Parallel Workstation," *Proceedings of the 5th IEEE Symposium on High Performance and Distributed Computing*, August, 1996.
- [19] D. Ridge, D. Becker, P. Merkey, "Beowulf: Harnessing the Power of Parallelism in a Pi le-of-PCs," *Proceedings of IEEE Aerospace*, 1997.
- [20] SMP Clusters in OSC, <http://www.osc.edu/whatsnew/>
- [21] T. Sterling, D. Becker, D. Savarese et. al., "BEOWULF: A Parallel Workstation for Scientific Computation," *Proc. Of the 1995 International Conf. On Parallel Processing (ICPP)*, August 1995.
- [22] T. Sterling, D. Savarese, D. Becker et. al., "Communication Overhead for Space Science Applications on the Beowulf Parallel Workstation," *Proc. of 4th IEEE Symposium on High Performance Distributed Computing (HPDC)*, August, 1995.

[23] C. B. Stunkel, D. G. Shea, B. Abali, M. G. Atkins, et al., "The SP2 High-Performance-Switch," IBM Systems Journal, vol. 34, no. 2, pp. 185-204, 1995.

[24] 鄭守成, "國家高速電腦中心效能評估彙總報告," 高速計算世界, Vol. 7, No. 2, Summer 1999.

處理器	Intel Pentium II 400 MHz
記憶體	NEC PC-100 256MB RAM
主機板	陞技AB-BX6主機板
計算專用 PC 硬碟	IDE, 6.5GB
伺服器 PC 硬碟	IBM DGHS-09U Ultra Wide 7200RPM 9.1GB
網路卡	3Com 3C905B-TX(UTP) PCI 10/100M
網路交換式集線器	3Com SuperStack II Switch 3300
作業系統	Red Hat Linux 5.2, kernel 版本 2.2.0
編譯器	GNU C, C++, Fortran 77, PGI C, C++, Fortran 77, Fortran 90, HPF
平行軟體	MPICH 1.1.2, PVM 3.4, BSP 1.4
叢集管理軟體	DQS 3.2.7
數學函式庫	PETSc 2.0

表一: NCHC PC Cluster 軟、硬體規格

NAME	kernel	Byte/iter	FLOPS/iter
<i>COPY</i>	$a(i) = b(i)$	16	0
<i>SCALE</i>	$a(i) = q*b(i)$	16	1
<i>SUM</i>	$a(i) = b(i) + c(i)$	24	1
<i>TRIAD</i>	$a(i) = b(i) + q*c(i)$	24	2

表二: STREAM 評估程式中的運算型態

	IBM SP2-160	HP SPP2-200	SGI Origin 2000	Fujitsu VPP300	Cray J916	Pentium II 400
Bem3d	387	534	229	103	374	320
Ns3d	1644	1971	2419	2250	4272	4718
Modyn	1148	1156	195	56	320	685
Mopac93	4737	9111	5583	5464	23124	12104
Pureg	266	246	250	346	1056	589
Hubksp	1413	1675	1146	6999	15106	890
Finite	214	772	689	194	343	692
Jcg3d	132	272	222	212	675	300

表三: NCHC 應用程式集執行時間(單位: 秒)

Name	Characteristics	Type
BT	Block Tridiagonal solver	Application
SP	Pentadiagonal solver	
LU	LU solver	
MG	Multigrid	Kernel
FT	3-D FFT PDE	
EP	Embarrassingly parallel	
IS	Integer Sort	
CG	Conjugate Gradient	

表四: NPB 程式集

Benchmark code	Class A	Class B	Class C
BT	64^3	102^3	162^3
SP	64^3	102^3	162^3
LU	64^3	102^3	162^3
MG	256^3	256^3	512^3
FT	$256^2 \times 128$	$256^2 \times 512$	512^3
EP	2^{28}	2^{30}	2^{32}
IS	2^{23}	2^{25}	2^{27}
CG	14000	75000	150000

表五: 不同等級 NPB 程式集之資料量大小

Machine	MFLOPS/Processor	MFLOPS/\$10000
One Job in Single CPU PC	43	8.6
Two Jobs in SMP PC	29	8.3

表六: 以倍精準度 Linpack 1000x1000 程式進行雙 CPU 架構的效能分析

Machine	MFLOPS/Processor	MFLOPS/\$10000
One Job in Single CPU PC	62.90	12.58
Two Jobs in SMP PC	47.83	13.67

表七: 以單精準度 Linpack 1000x1000 程式進行雙 CPU 架構的效能分析

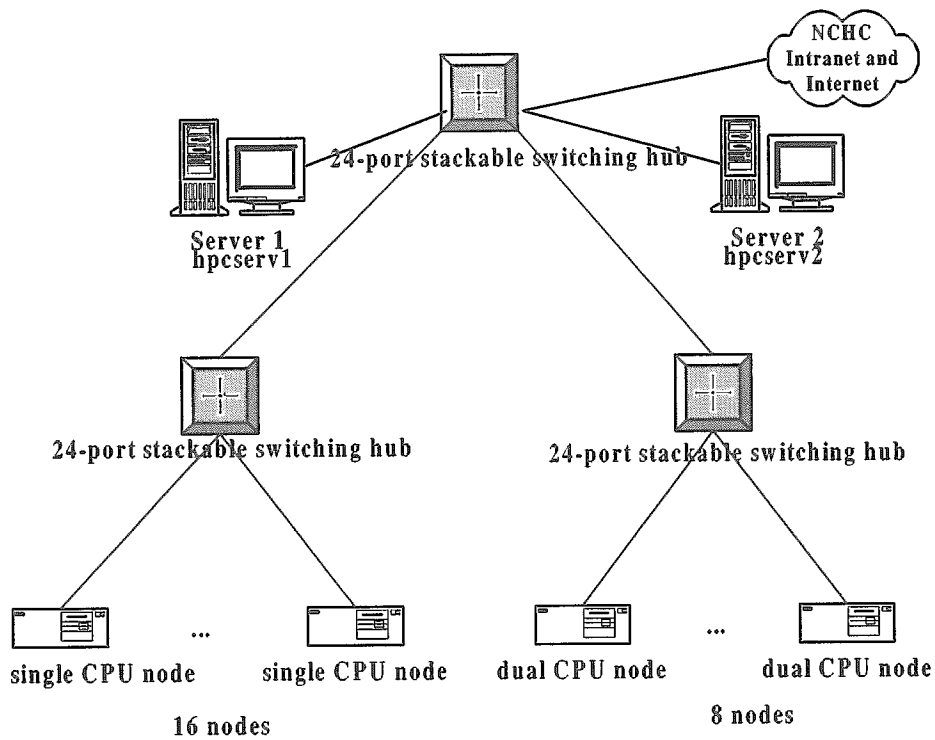
Name		M(FL)OPS /Processor	M(FL)OP/ \$10000
SP	One Job in Single CPU	47.50	9.50
	Two Jobs in SMP	35.79	10.22
LU	One Job in Single CPU	74.71	14.94
	Two Jobs in SMP	58.07	16.59
EP	One Job in Single CPU	1.11	0.22
	Two Jobs in SMP	1.10	0.31
IS	One Job in Single CPU	4.67	0.93
	Two Jobs in SMP	3.97	1.13
CG	One Job in Single CPU	34.23	6.85
	Two Jobs in SMP	28.92	8.26

Name		M(FL)OPS	M(FL)OP/ \$10000
LU	Two Single	143.91	14.39
	One SMP	106.77	15.25
EP	Two Single	2.23	0.22
	One SMP	2.17	0.31
IS	Two Single	3.61	0.36
	One SMP	5.55	0.79
CG	Two Single	55.88	5.59
	One SMP	52.01	7.43

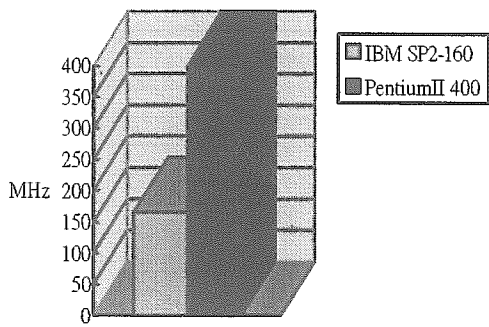
表八: 以 NPB 程式集進行雙 CPU 架構的效能分析 表九: 以平行版本 NPB 程式集進行雙 CPU 架構的效能分析

Message size	1KB	16KB	256KB	1MB	16MB
Between Single CPUs	2.5	8.6	10.5	10.7	10.8
Inside SMP	7.1	28.3	32.2	30.9	29.4
Between One SMP and Two Single CPUs	2.7	7.9	7.8	8.2	5.6
Between Two SMPs	1.7	5.4	6.7	7.6	5.5

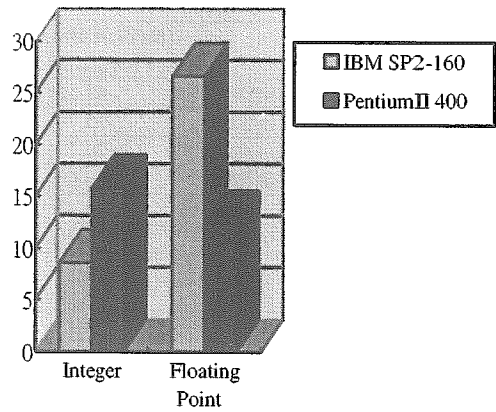
表十: 不同網路架構下的通訊頻寬



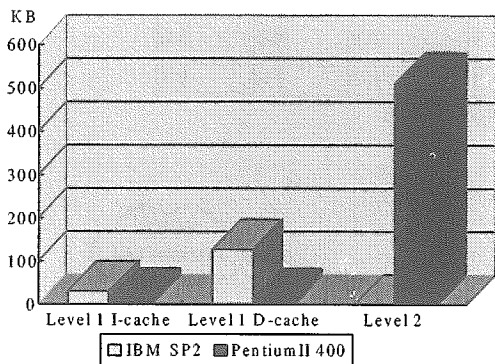
圖一: NCHC PC Cluster 網路架構



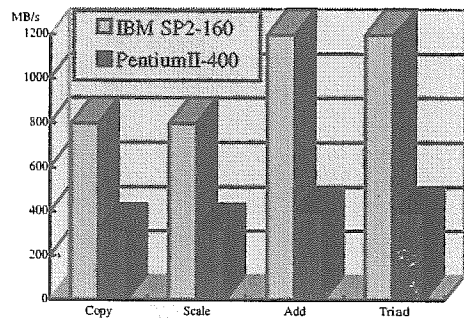
圖二: CPU 時脈頻率比較



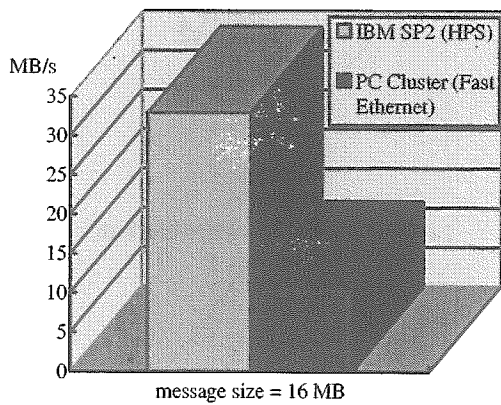
圖三: SPEC 95 評估程式集之效能比較



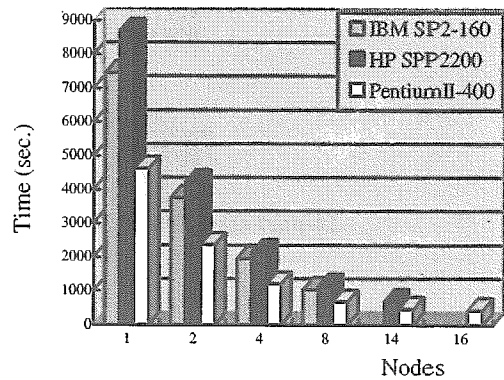
圖四: 快取記憶體大小比較



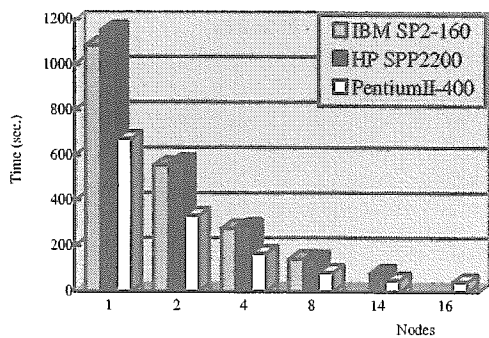
圖五: STREAM 評估程式效能比較



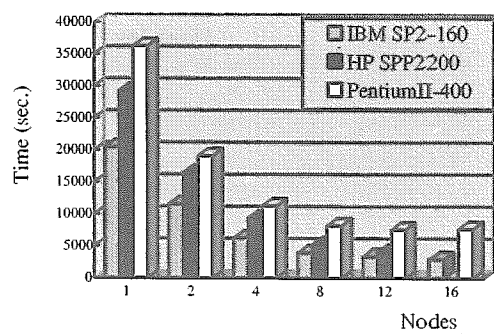
圖六: MPICH Ping-Pong 測試結果



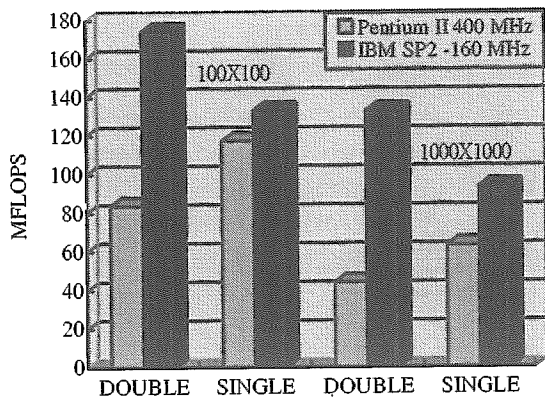
圖七: hubksp之平行效能



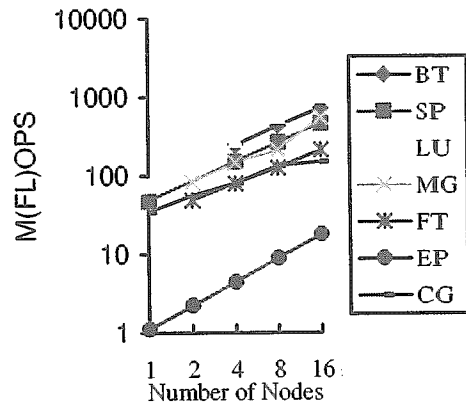
圖八: modyn之平行效能



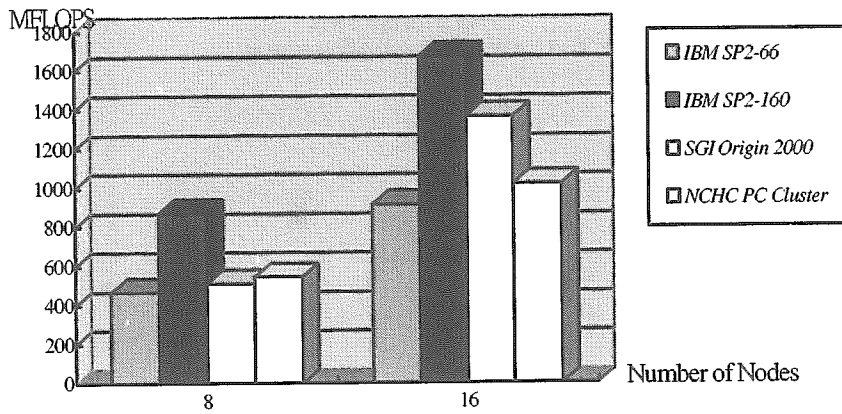
圖九: CWB氣象預報模式的平行效能



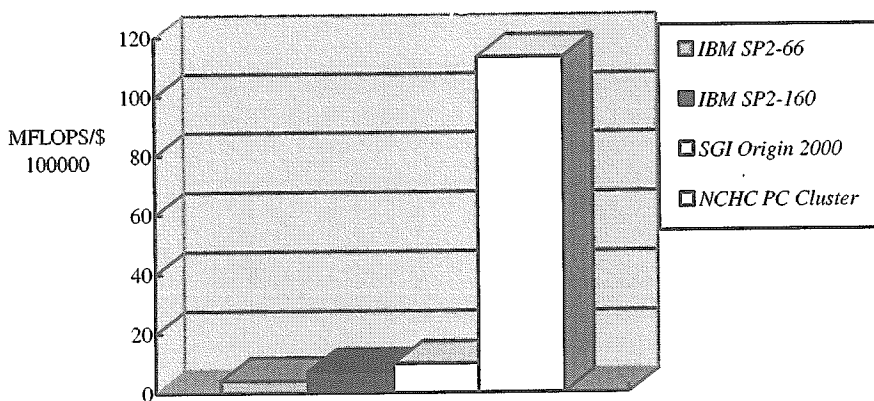
圖十: Linpack 執行效能



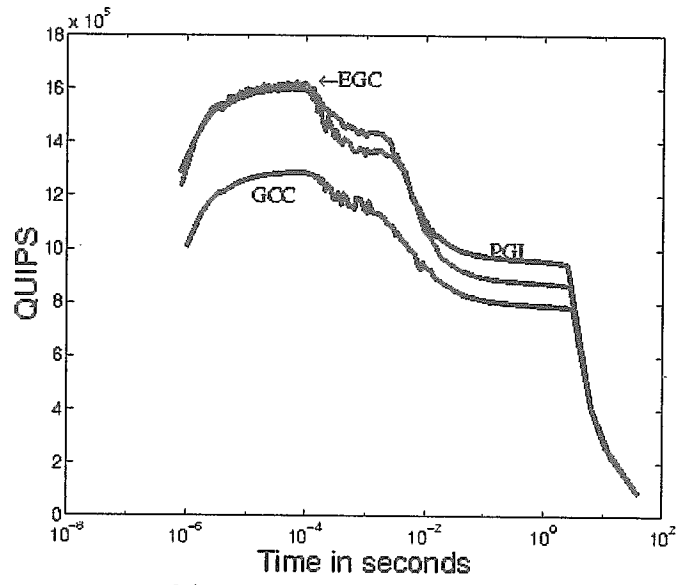
圖十一: NPB Class 在 PC cluster 上的執行效能



圖十二: Class CLU 在不同平台上之執行效能



圖十三: Class CLU 在不同平台上之效能/價格比



圖十四: 不同編譯器的HINT效能圖形