Phantom Cluster 於科學計算之效能分析

謝志偉、李金泓、吳長興、王順泰

國家高速網路與計算中心

E-mail: david.hsieh@nchc.org.tw

摘要

近年在各種計算加速器與 SoC 的應用下，各種混和協同式計算架構吸引了許多科學家的目光，而隨著綠色節能的觀念風行全球，這樣的概念也開始逐漸影響到高性能計算系統的設計，未來電力以及熱能的消耗在高性能計算系統的建置將是一個重要的考量因素。在本文中所介紹的 Phantom Cluster 提供了節能及經濟的高性能計算系統平台，另外在本文中也利用了 HPL 等的效能測試軟體來評估此新型平台的效能表現以及可靠度，另外就節能方面來分析傳統叢集系統與 Phantom Cluster 的價格-效能與耗電-效能比。

關鍵詞：高性能計算、效能分析、Phantom Cluster、節能

一、前言

自從 1994 年使用開放源碼作業系統的 Beowulf Cluster[6] 由 NASA 的 Donald Becker 等人發展出來後，個人叢集電腦(PC Cluster)，因為硬體建置便宜、安裝設定方便以及具有發展成熟的訊息傳遞程式庫(MPI)[7]等特性，所以 PC Cluster 在高性能計算系統中具有不可取代的角色直到今天。另外在全球五百大超級電腦排名(TOP500)[11]中，PC Cluster 在 1997 年開始出現於 TOP500 排名內，而後一直到今年六月份的第三十三屆 TOP500 的排名中依舊可以看到 PC Cluster 的身影，且它的佔有率是逐年提昇直到最新的資料中可以發現 PC Cluster 在所有的高性能計算系統中佔了百分之八十二(圖 1)，而其他架構的高性能計算系統卻逐年下降，由此可見 PC Cluster 從出現後到現今一直在高性能計算中佔有舉足輕重的地位。

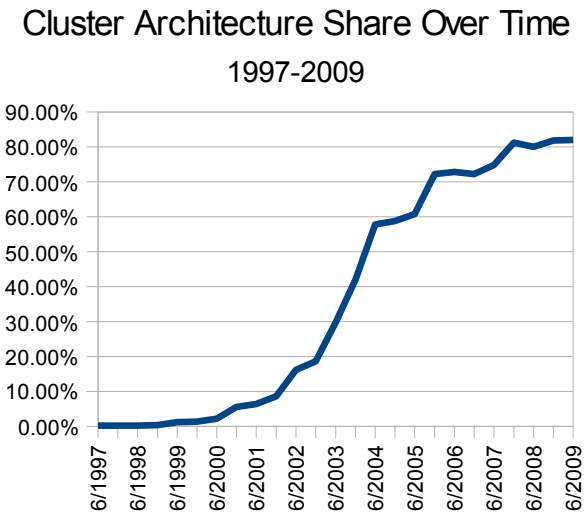

圖 1. 1997-2009 Cluster 在 TOP500 的佔有率

隨著計算需求的增加所消耗的電力成本以及散熱問題也隨之而來，在過去大家只關注在高性能計算系統所得到的效能，很少注意到系統本身所耗費的電力成本以及熱能，在現今綠色節能及全球經濟風暴的影響之下，節能指標在日後將是建置高性能計算系統時會被列入的參考數據之一。TOP500 網站除了在統計全球超級電腦的排名外，因應全球綠色節能的潮流，TOP500 網站在 2008 年開始將耗能指標列入參考項目之一，另外在 2008 年成立的 Green500 網站[10]，就是依高性能計算系統的節能指標所做排名的網站。

目前另一個值得注意的是資源的使用，由國家高速網路與計算中心(NCHC) 目前對外服務 FormosaII HPC Cluster 的資源使用率高達八成的情形下，對從事科學計算研究及程式開發初期或許只需要數十個節點來驗證其程式的正確性，但往往需要在提供高速計算的服務系統佇列中排隊等候許久才會執行，如此一來在程式的除錯及驗證往往就失去了先機。故本文提出使用電腦教室閒置的資源來做為高速計算的解決方案，以提供一些計算時間不長，卻需要數十個 PC Cluster 節點來驗證或測試的計算，另外本文提供在此平台上的效能測試以驗證其可行性。

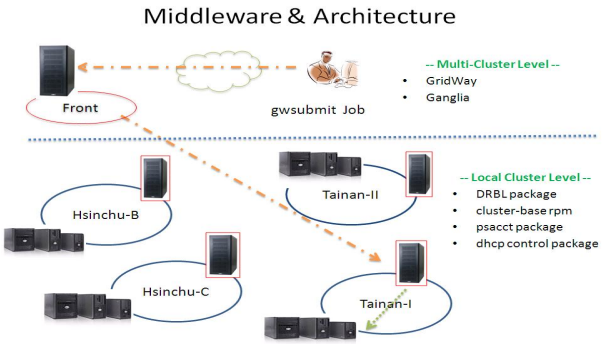

圖 2. Phantom Cluster 系統架構圖

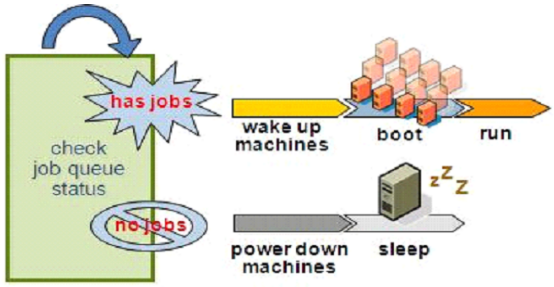

圖 3. Phantom Cluster 工作流程圖

二、NCHC Phantom Cluster

Phantom Cluster 計劃[9]是由 NCHC 的 HPC Cluster 團隊共同規劃與建置，最初計劃目的是提供不熟悉 PC Cluster 安裝的使用者可以利用此專案來快速佈建與管理 PC Cluster，使用者只要專注平行程式的開發、測試即可。在經過長期的開發之後，Phantom Cluster 逐漸形成如今的架構，而現在的 NCHC 的 Phantom Cluster 主要可以提供使用者在測試性質、小規模、短時間的計算量所建置的特殊高性能計算平台。

Phantom Cluster 平台是基於 NCHC DRBL 專案[8]的一種無碟多叢集系統(Diskless Multi-Cluster System)上，最初的規劃就是為了不影響現有電腦教室的環境設定，以及重復利用電腦教室未上課時的資源來做計算，例如：假日或晚上，如此一來即可節省購買新機器的經費又可以不浪費電腦教室閒置的資源。目前於 NCHC Phantom Cluster 的組成主要是由多間電腦教室叢集系統所組成(圖 2)，基於此架構下，本系統建置都已經包裝成 RPM 套件，並且也已經客製化常見的叢集系統中介軟體：MPICH 平行函式庫、 Torque 排程系統和 Ganglia 系統監控軟體。

Phantom Cluster 的系統工作流程如下：當使用者工作派送至工作佇列後，系統就會將工作派送到分散各地的電腦教室並利用電腦教室閒置的時間(圖 3)，依照其計算所需的節點數來喚醒足夠數量的計算節點來進行運算。當工作完成後，系統會將喚醒的工作計算節點做關機動作以節省電力之消耗，因此本系統擁有動態節能，充分利用異地電腦教室電腦閒置時間。另外在計算的過程中不會變更原本計算節點安裝的系統，所以當白天上課時學員打開個人電腦就可以上課。此系統只要是熟悉 Linux 系統的管理者，透過簡單的 RPM 安裝指令，就能迅速安裝好計算環境，省下自行摸索編譯 MPI 函式庫與設定排程系統的時間，所以並不會增加系統管理員的負擔。Phantom Cluster 特色有以下幾點：

1. 節能: 僅在計算時喚醒所需之計算節點。

2. 經濟: 利用空閒電腦教室資源。

3. 省時: 系統安裝、管理容易。

表 1. SIRAYA 硬體規格

處理器	AMD Opteron 275 DualCore 2.2GHz x 2
記憶體	8GB DDR400 Registered/ECC SDRAM
網路介面	10Gb/s Infiniband 、Gigabit Ethernet
節點數	80

表 2. IBM 1350 硬體規格

處理器	Intel Woodcrest 3.0 Ghz Dual-Core processor × 2
記憶體	16 GB PC2-5300 667MHz FBD 240-pin ECC DDR2
網路介面	20Gb/s Infiniband
節點數	512

表 3. Phantom Cluster 軟、硬體規格

處理器	Intel Core2 Duo 2.83GHz
記憶體	8G DDR2
網路卡	Intel(R) PRO/1000
作業系統	CentOS 5.2 X86_64
平行編譯軟體	OpenMPI 1.3.2
數學函式庫	INTEL MKL 9.1.023

三、系統效能測試與分析

本文中就 LINPACK[12] 及 NCHC 效能測試應用程式集[1、2、3]，等效能測試軟體來評估 Phantom Cluster 在高性能計算上的性能表現並且我們以 NCHC 自行建置的 FormosaII HPC PC Cluster(表 1) 以及 IBM 1350 Cluster(表 2) 來做為比較對象。SIRAYA PC Cluster 是由 80 個計算節點所組成，建置其共為二個階段：第一階段是由 Gigabit Ethernet 所組建 Linpack 效能達 848*GFlops*，第二階段擴建了 Infiniband 網路後其 Linpack 效能達 1.2*TFlops*。我們分別從系統效能、耗能及價格比來做分析。而 Phantom Cluster 測試平台是使用 NCHC 電腦教室中的 32 台 PC 所組成，詳細資料如表 3。

Linpack 是傳統用來測量超級電腦效能的效能評估軟體，而目前 TOP500 就是使用 Linpack 的效能指標來做為超級電腦系統排名的測試依據來源。在 Linpack 中主要是在求解線性系統問題的軟體，而求解線性系統問題常出現於工程計算上，最初是由美國 Argonne 國家實驗室提出，之後由美國田納西大學教授 Jack Dongarra [4、5] 設計成為用來評估超級電腦的效能評估軟體，其測試結果可以用來評估系統的擴充性能。而另一項測試則是使用了 NCHC 效能測試應用程式集來評估 Phantom Cluster 在不同領域的高速計算問題的表現，我們選用了在物理(hubksp)以及大氣科學(nonh3d)等應用領域的程式來做為在不同計算領域的測試評估。

在 Linpack(圖 4)的表現下 Phantom Cluster 的 Rmax 可達 271.9*Gflops*，此效能已達 SIRAYA-GE 效能的三成以上。另外在物理問題 hubksp(圖 5) 中的效能中顯示在 Phantom Cluster 的系統架構之下其延展性效能是符合高性能計算上的需求，在大氣科學 nonh3d(圖 6) 的表現上來看，其效能表現在 32 個節點上有 10 倍以上的加速，顯示 Phantom Cluster 的系統架構在處理類似的問題上也有不錯的平行性能。

從價格-效能比上來看，建置一座效能達 1.2*Tflops* 的 PC Cluster 計算系統所能得到的 C-P 值僅有 0.059 *Gflops/cost*，而 Phantom Cluster 只是利用電腦教室閒置的時間來做計算，所以可以忽略其建置成本，二者相較之下 Phantom Cluster 所得到 C-P 值遠大於建置大型叢集系統。再則由每*KW*所得到的效能來看(圖 7)，Phantom Cluster 在節能表現相對較優於傳統電腦叢集系統。

四、結論

在本文中提出的 Phantom Cluster 具有節能、經濟、安裝及管理容易的高性能計算系統，另外在 Linpack 效能的效能測試上 Phantom Cluster 有 271.9*Gflops* 的表現，另外在 NCHC 效能測試應用程式集效能上顯示此平台適合於科學平行計算問題，而在價格-效能比以及效能-耗電比都比傳統叢集系統來的較具優勢。總體上來看 Phantom Cluster 平台可以適用於小規模的平行計算平台或測試使用，利用電腦教室閒置的電腦，就可以不用花費大筆經費就可以擁有高速計算系統。

五、參考文獻

[1] 周朝宜、鄭守成、黃國展、張西亞，”PC Cluster 作為高效能科學及工程計算平台之效能評估”，NCS'99，pp. A503-A510，1999。

[2] 張西亞、王順泰、周朝宜、陳德民、吳長興、李金泓、謝志偉”，高效能電腦叢集的發展與趨勢”，物理月刊，第廿九卷，第五期，pp. 936-947，2007。

[3] Kuo-Chan Huang, Hsi-Ya Chang, Cherng-Yeu Shen, Chaur-Yi Chou, Shou-Cheng Tcheng, ”Benchmarking and Performance Evaluation of NCHC PC Cluster”, HPCAsia, Vol. 2, pp.923-928, Beijing, China, May 14-17, 2000.

[4] J. Dongarra,”Performance of Various Computers Using Standard Linear Equations Software”, Technical Report CS-89-85, University of Tennessee, 1989.

[5] J. Dongarra, J. Bunch, C. Moler and G. W. Stewart,”LINPACK Users Guide”, SIAM, Philadelphia, PA, 1979.

[6] Thomas Sterling , Donald J. Becker , Daniel Savarese , John E. Dorband , Udaya A. Ranawake , Charles V. Packer , ”BEOWULF: A Parallel Workstation For Scientific Computation”, Proceedings of the 24th International Conference on Parallel Processing (ICPP), pp. 11-14, 1995.

[7] William Gropp, Ewing Lusk, and Anthony Skjellum, ”Using MPI: Portable Parallel Programming with the Message-Passing Interface”, The MIT Press, 1999.

[8] http://drbl.nchc.org.tw/

[9] http://sourceforge.net/projects/phantomcluster/

[10] http://www.green500.org/

[11] http://www.top500.org/

[12] http://www.netlib.org/linpack/

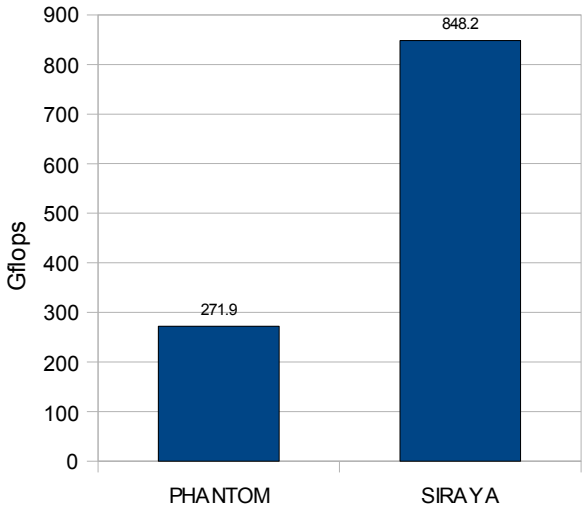

圖 4. Linpack Performance

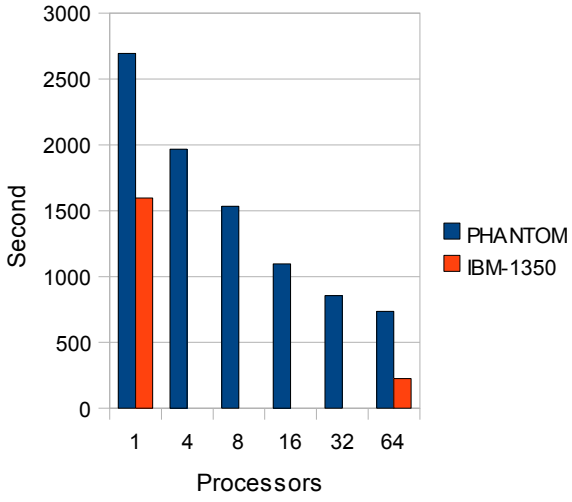

圖 5. hubksp 的平行效能

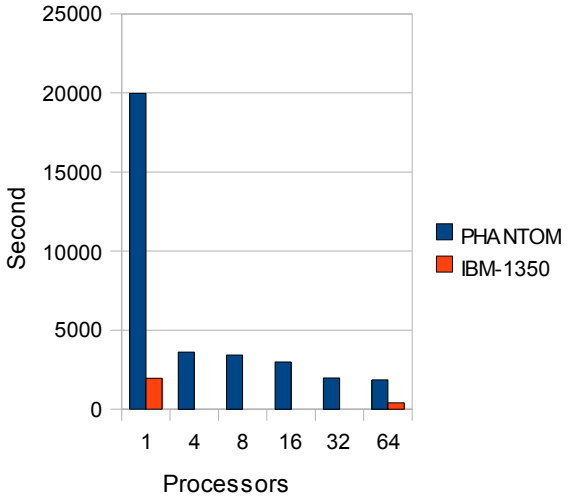

圖 6. nonh3d 平行效能

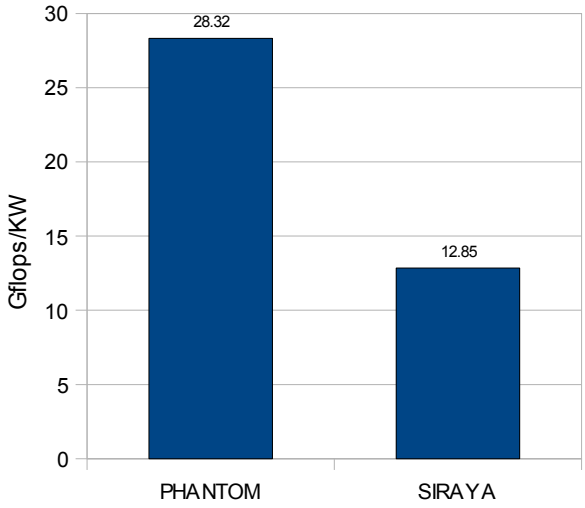

圖 7. 耗電-效能比較圖