1 Cluster 簡介

此篇文章最主要目的如何利用非官方支援的硬體來建置 IBM Cluster 1350；筆者認為就 Beowulf Cluster 的定義，IBM Cluster1350 應算是 Beowulf Cluster 的一種。那 Beowulf Cluster 又是什麼呢？首先筆者先介紹何謂 Cluster 及分類，還有 Beowulf Cluster 定義，最後再建置 IBM Cluster 1350。

1.1 Cluster 的分類

「Cluster」、「High Availability Cluster」、「High Performance Computing Cluster」這幾個名詞常讓人混淆。一般的定義，「所謂 Cluster 就是由一台以上的機器，為了某種特定需求所組成的架構」，根據不同的需求，可將 Cluster 分為以下三種：

- **High Availability Cluster**－增加伺服器和以網路為基礎的應用程式的高可用性及備援性，例如 IBM AIX HACMP 及 Red Hat Cluster Suite 中的 Cluster Manager。
- **Load Balancing Cluster**－將服務需求分派給多台伺服器，可視系統負載隨時彈性增加伺服器，例如 Red Hat Cluster Suite 中的 Linux Virtual Server（Piranha）。
- **High Performance Computing Cluster（高效能/平行運算叢集系統）**－所謂高效能/平行運算叢集系統就是讓你的應用程式可以使用到多台主機的運算能力（CPU、Memory..）讓程式很快地運算執行完畢，例如 IBM Cluster 1350、Beowulf Cluster。

1.2 平行運算的工作模式

由於這篇文章最主要目的為建置平行運算叢集系統，所以我們得先大致了解何謂平行運算，讀者可於參考東海大學「楊朝棟教授」叢集式處理電腦系統的網頁有很詳細的說明，筆者節錄其中的說明。

平行運算依照其工作的方式，大約可分為三種模式：第一種叫做「共用記憶體多處理器系統」（Shared Memory Multiprocessor System）模式，第二種稱為「分

散式記憶體多處理器系統」(Distributed Memory Multiprocessor System)模式，還有一種就是「叢集式處理系統」（Clustering System）模式。

- **「共用記憶體多處理器系統」模式**

共用記憶體多處理器系統又稱為對稱式多處理器電腦（Symmetric Multiprocessors，簡稱SMP），是傳統超級電腦的縮小版。SMP架構的特色是採用系統匯流排（System Bus）的方式將系統的CPU、記憶體及I/O相連接（圖1）。一般讀者所聽到的 2-Way,4-Way,8-Way..多處理器PC伺服器就是這種架構。對使用者而言，不用去管伺服器到底有幾顆CPU，你都是同一套作業系統打交道，作系統統會幫你把工作分配給負擔較輕的CPU。不過這種架構的缺點是因為CPU、記憶體及I/O等資源皆為共享，且系統匯流排（System Bus）的資料傳輸頻寬是固定的，不斷增加CPU時會造成瓶頸。雖然系統匯流排(System Bus)的通道越多，可以連結的CPU個數亦越多，但是成本亦越高。更何況要在主機版放入這麼多顆CPU而且又得使其正常運作對廠商來講可不是一件簡單的事。

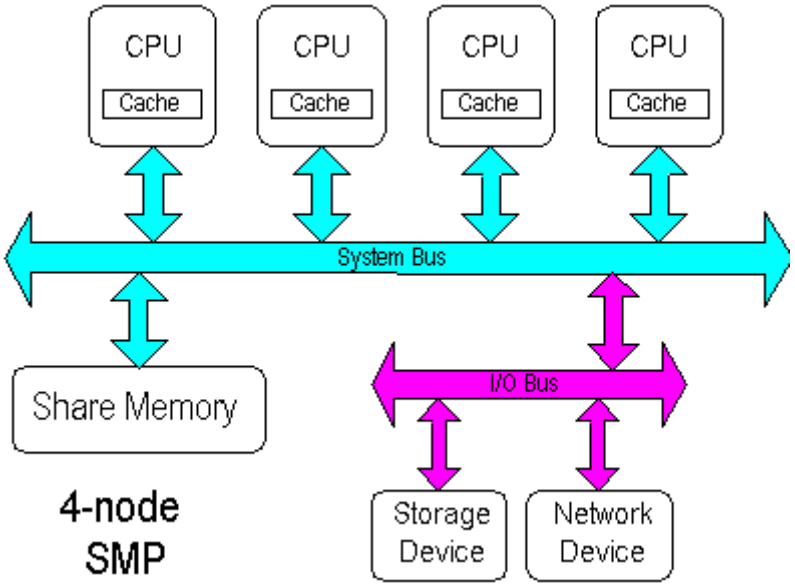

圖 1：SMP 電腦架構

- **「分散式記憶體多處理器系統」模式**

「分散式記憶體多處理器系統」是一種分工程度更高的模式，它是指同一台機器裡有多顆CPU，且各自擁有私屬的記憶體（Local Memory）。分散式記憶體多處理器系統又稱為巨量多處理器電腦（Massive Parallel Processors，簡稱MPP）。MPP架構是由數個CPU及記憶體之模組以某種拓樸學（Topology）的方式連接而成（圖2）。

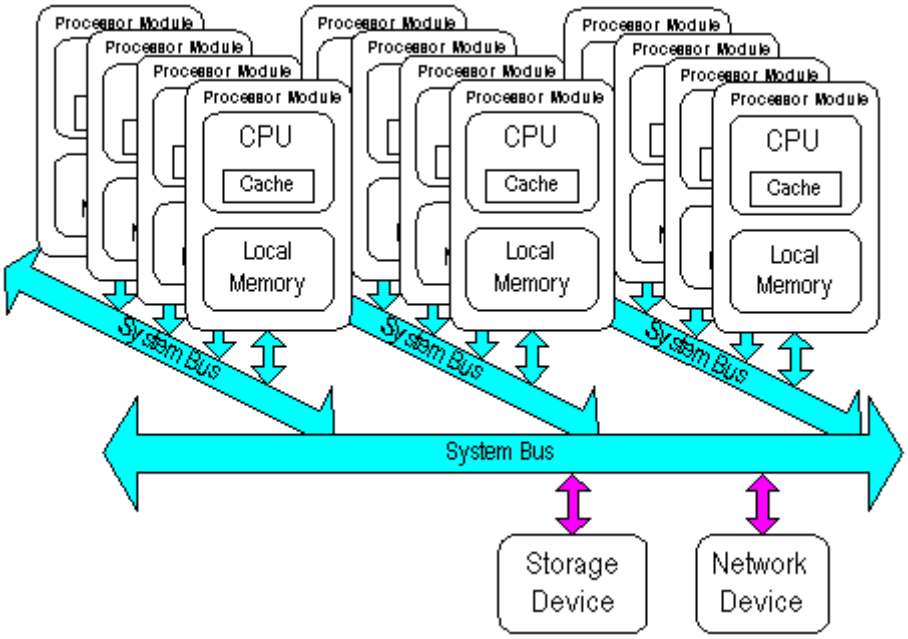

圖 2：MPP 電腦系統之架構

MPP 電腦系統架構裡，每一個含 CPU 及私屬記憶體(Local memory)的子系統皆為各自獨立，整個系統也是由單一作業系統管理，例如 IBM 的 SP2。下面是國家高速電腦中心（NCHC）IBM SP2 SMP(ivory)的長相及規格。
http://www.nchc.org.tw/htdocs/outside_service/hardware/ibm_sp_smp.php

圖 3：NCHC IBM SP2 SMP

機器名稱	IBM SP2 SMP(ivory)
機器架構	分散式平行電腦系統
主要使用領域	計算物理、計算電磁、流體力學
規格	---處理器個數：168(4CPUx42node) ---記憶體：184GB (4096x38+8192x4) ---磁碟機容量：1512GB+540GB
其他特別週邊裝置	SP Switch
安裝日期	---八十八年度： 332Mhz thin PCI SMP node(4cpu)x14 ---八十九年度： 375Mhz thin powre3 SMP thin node(4cpu)x42
用戶數量	500

每一個處理器模組（Processor Module = CPU + Local Memory）內的記憶體都是各自獨立作業，CPU 與 CPU 之間只能靠訊息傳遞（Message Passing）的方式來溝通，而重要的溝通問題則由 CPU 額外花時間來處理。由於每個 CPU 幾乎都具有完整的電腦架構（唯獨無法單獨處理 I/O），所以它們都各自執行其工作，只有透過彼此之間的通訊，才有辦法互相協調、互助合作，一般常聽說的 MIMD 就是指這種「多重指令、分散資料」的模式。

這種電腦系統可以連結的 CPU 個數相當多，可以到達 512 個、1024 個或更多。不像 SMP 系統，在 MPP 上做平行計算比較複雜，你必須把程式用到的資料和計算工作做適當的切割，把不同的資料區塊分別傳送到各個 CPU 的記憶體裡，然後各個 CPU 自行計算放在該 CPU 所屬記憶體裡的資料。如果要使用到其他

CPU 所屬記憶體裡所存放的資料時，就要經由 CPU 間的網路從該 CPU 取得該項資料後，才能夠繼續計算下去。

現階段 MPP 各個 CPU 之間高速網路的資料傳輸速度比 CPU 到自己記憶體內的資料傳輸速度要慢很多倍，因此 CPU 之間傳送的資料越少，平行計算的效率就越高。在 MPP 電腦系統上做平行計算必須使用訊息傳送（Message Passing）語言來撰寫，有 PVM（Parallel Virtual Machine）和 MPI（Message Passing Interface）等多種。如果你有一個循序程式（Sequential Code）要在 MPP 上平行化，你必須先學會一種訊息傳送語言，然後再把你的程式和該程式用到的資料做適當的切割、重組（Restructure），然後加入訊息傳送副程式改寫你的程式，經過編譯之後才能執行。

- **「叢集式平行電腦系統」模式**

「叢集式平行電腦系統」通常指的是把好幾台相同系統的電腦(或不同機型的電腦)用高速網路連結在一起，形成一個大電腦系統。叢集式處理系統是多個獨立電腦的集合體，每一個獨立的電腦有它自己的CPU、記憶體、和作業系統（圖3）。

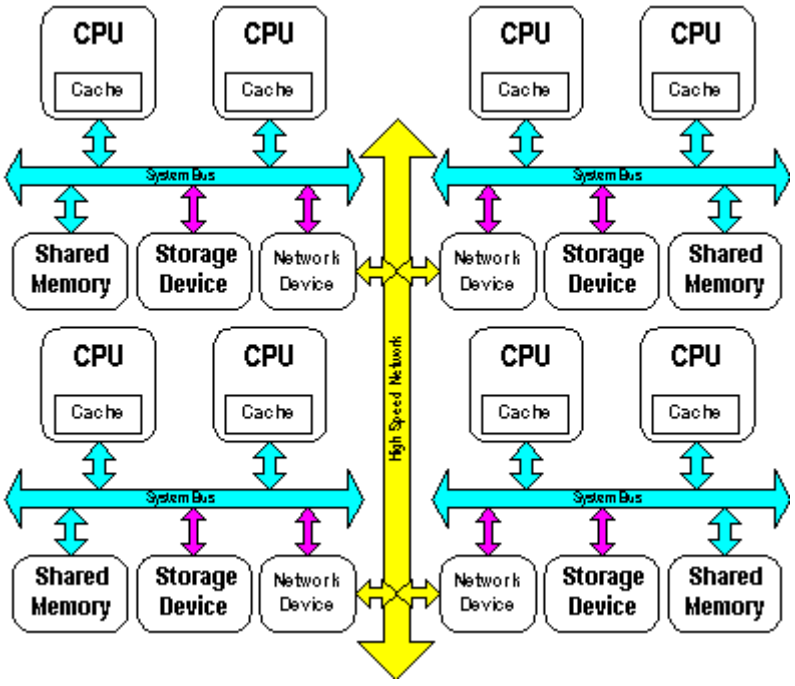

圖 4：叢集式處理電腦系統

這種工作方式常因各單機資料獨立，且無須交換大量資訊，必要時再透過網路交換訊息。目前常用的叢集式電腦系統是PC Cluster，最簡單的個人電腦群是透過乙太網路（Ethernet）或FDDI網路把若干個分散的工作站連結在一起，所以連

接各個工作站的網路便是瓶頸所在，不過由於高速乙太網路(1000Mbps)日漸普及，各個工作站（node）之間的訊息傳輸已獲得改良，所以PC Cluster有日漸走紅的趨勢。

現今的PC Cluster已經可以擴充到 256或512個CPU，在叢集式處理電腦系統上做平行計算是使用訊息傳送(Message Passing)語言來撰寫，常用的訊息傳送語言有PVM （Parallel Virtual Machine）和MPI（Message Passing Interface）。PVM及MPI這兩個Libraries都是使用TCP/IP來做為與其他電腦溝通的Protocol，其實就是用rsh指令來和其他電腦溝通。若要平行計算有較高的效率，除了有足夠的電腦資源可供利用之外，還要你的程式能夠被平行化才行。

2 Beowulf Cluster（北歐武夫 Cluster）

1994年夏季，美國太空總署（NASA）的CESDIS（the Center of Excellence in Space Data and Information Sciences）為了進行地球與太空科學研究計畫（the Earth and Space Science Project），而嘗試用低廉而易得的電腦相關零件，或稱為COTS （Commodity Off The Shelf），來組裝可支援平行計算的電腦系統，以應付該計畫所需處理的大量資訊處理與計算。

為此，將十六個Intel 100 MHz DX4以10Mbits/s Ethernet組裝成一個叢集（Cluster），並取名為Beowulf（戰狼或稱北歐武夫）。但是，Beowulf真正的內容定義則是到了1996年才確定下來。該年的九月，美國的Los Alamos National Laboratory（LANL）建置了一個名為Loki的叢集電腦系統，它是由16顆Intel Pentium Pro 200 CPUs所組成，網路則為100Mbps Fast Ethernet，作業系統採用Linux，平行運算與資料傳輸用MPI（Message Passing Interfacing）。

Beowulf網站（http://www.beowulf.org/）上對Beowulf Cluster的定義如下：

Cluster is a widely-used term meaning independent computers combined into a unified system through software and networking. At the most fundamental level, when two or more computers are used together to solve a problem, it is considered a cluster. Clusters are typically used for **High Availability (HA)** for greater reliability or **High Performance Computing (HPC)** to provide greater computational power than a single computer can provide.

Beowulf Clusters are scalable performance clusters based on commodity hardware, on a private system network, with open source software (Linux) infrastructure. The designer can improve performance proportionally with added machines. The commodity hardware can be any of a number of mass-market, stand-alone compute nodes as simple as two networked computers each running Linux and sharing a file system or as complex as 1024 nodes with a high-speed, low-latency network.

Class I clusters are built entirely using commodity hardware and software using standard technology such as SCSI, Ethernet, and IDE. They are typically less expensive than **Class II clusters** which may use specialized hardware to achieve higher performance.

Common uses are traditional technical applications such as simulations, biotechnology, and petro-clusters; financial market modeling, data mining and stream processing; and Internet servers for audio and games.

Beowulf programs are usually written using languages such as C and FORTRAN. They use message passing to achieve parallel computations. See Beowulf History for more information on the development of the Beowulf architecture.

通常一個典型的Beowulf Cluster是藉著一些訊息通訊程式庫來工作的，也就是說，在Beowulf上頭執行的平行程式，是架構在Messages Passing程式庫上頭，如此才能藉著網路來溝通每一台電腦，並把工作分配出去。而目前最常被使用的Message Passing Libraries有兩種，分別是PVM （Parallel Virtual Machine）及MPI（Message Passing Interface）。PVM及MPI這兩個Libraries都是使用TCP/IP來做為與其他電腦溝通的Protocol，其實就是用rsh指令來和其他電腦溝通。

3 IBM eServer Cluster 1350

資料來源：http://www-8.ibm.com/servers/eserver/tw/xseries/cluster/cluster1350.html

3.1 Cluster 1350 簡介

降低安裝Linux叢集所需的時間與資源

提供單點控管功能，簡化叢集管理工作並增強叢集的可用性

為高效能或商業運算工作及伺服器的整合需求提供極具規模彈性的解決方案

3.2 Cluster 1350 產品特色

3.2.1 傑出的價格/效能比

叢集系統一直以來都能在許多高效能工作上展現驚人的價格/效能比優勢。只要運用成本低廉的伺服器和開放原始碼(Open Source) 軟體，Linux 叢集就能進一步擴展這些優勢。

現今許多機構都使用硬體產品、標準的互連與網路技術、開放原始碼 (Open Source) 軟體及自行設計或協力廠商所提供的應用程式來建構其 Linux 叢集。在建構過程中，他們常會發現有相當多的資源必須投入於裝配、整合、測試、管理與支援該叢集中。因此，這些機構經常會碰到許多產品開發過程中常見的問題。

IBM 利用其在叢集式 UNIX 電腦方面的豐富經驗開發出 IBM Cluster 1350 來應付這些挑戰。Cluster 1350 使用先進的 IBM xSeriesTM Intel 處理器的伺服器節點及經過驗證的叢集管理軟體、以及高速互連的選購項目，能一次為您呈現 IBM 與協力廠商的最佳技術。因此，Linux 叢集的安裝速度便能大幅提昇，其支援也能大為簡化。

Cluster 1350 對於需要卓越的價格/效能比來處理高效能運算工作的工業、金融機構、生命科學、政府及教育機構而言，是一項相當理想的解決方案。它也是網頁伺服與合作或任何需要水平架構之應用方案的最佳選擇。

3.2.2 綜合的解決方案

客戶可以將 Cluster 1350 當成一個整合式的產品來訂購。它的組態設定工作非常容易，因此能幫助企業快速部署應用程式。選定了組態架構以後，IBM 會裝配並測試該叢集系統，以確保系統效能能達到特定的需求水準。IBM 會對整個叢集指定一個序號，作為所有相關服務的單一聯繫點。藉著研究、裝配、整合、測試與調整 Linux 叢集之時間與資源的降低，Cluster 1350 便能協助企業組織在部署 Linux 應用程式時，加速其開始投入生產的時間。除此之外，隨時都能在叢集中加入更多伺服器，以因應漸增的工作量、結合更多伺服器或增添新的應用程式。

IBM 提供 Cluster 1350 的安裝支援。另外，更高層級的支援包括選購的 Support Line for Linux Clusters，其參與人員皆為瞭解整個叢集環境的專家，包括 Linux、CSM for Linux 及 General Parallel File System (GPFS) for Linux，而非只瞭解單一元件的服務人員。

為了進一步簡化部署工作，IBM 提供了計畫管理支援來協調送貨及安裝等所有層面的細節，其中還包括硬體與軟體安裝服務。本方案亦適用令人心動的財務融資和租賃條款。

3.2.3 高效能叢集管理

IBM 提供 Cluster Systems Management (CSM) 這項先進的叢集管理軟體，讓以 Intel 處理器為主的 Linux 系統能從一個單一控制點來進行管理。如此一來，便可簡化叢集的管理工作，並提昇其規模擴增時的簡易程度，進而能改善系統管理員的效率。

CSM 包含一項基礎建設，能同時監視硬體和軟體事件，並可在適當時機啟動自動復原措施 CSM 這項具有高度可靠性的基礎建設和事件監視功能可以協助確保系統快速偵測並解決問題，因此能夠增進叢集的可用性。

CSM for Linux 係以 IBM Parallel System Support Programs for AIX 軟體產品為基礎所設計而成，並已部署於 IBM P 系列 (RS/6000) SPTM－世上最受歡迎的超級電腦中。CSM 包含數項以簡化 Linux 管理工作為目的的元件：

分散式管理伺服器：

能持續儲存叢集中每個節點的相關資訊，並維護每個節點的狀態。

事件反應資源管理器：

提供執行命令或描述語言 (scripts) 來回應使用者定義事件的能力。系統提供一組內容豐富的預設狀況與回應動作。許多資源都能加以監視，包括節點、轉接卡、檔案系統及程序。

遠端硬體控制

運用了 Cluster 1350 節點中的整合系統管理處理器。如此可讓系統管理員在遠端進行重置或開啟或關閉節點的電源。

組態設定檔管理：

提供節點之間常用檔案的儲存功能。CSM 會同步變更整個叢集的組態設定檔。

分散式介殼程式(shell)：

允許在遠端對所有叢集節點執行命令或描述語言 (scripts)，並且能夠自由選擇是否結合數部伺服器的輸出。分散式命令執行管理器是一項選購的圖形式使用者介面，它整合了分散式的介殼程式 (shell)，使節點和節點群組的管理工作更為簡單。

CSM 提供節點群組功能，非常便於對叢集中的伺服器子集合實施不同的管理規則。當叢集中支援多種應用程式的合併作業時，這是一項相當重要的考量。

有了節點群組功能，管理命令就能運用在各個節點、整個叢集，或由系統管理員定義的節點群組上。

藉著提供了單一的叢集控制點，CSM 便能急速簡化整個系統的管理工作，使伺服器合併方案更具成本效益。同時因為能執行描述語言 (scripts) 來因應常見的狀況，所以 CSM 也能協助叢集可用性的提昇。

3.2.4 先進的伺服器技術

Cluster 1350 係以來自 IBM 獨家的 X-ArchitectureTM 技術為基礎，該技術運用了一些 IBM R zSeriesTM 伺服器的可用性特性以及 IBM eServer pSeriesTM 系統的規模彈性。就其本身而論，這些符合產業標準且搭載 Intel 處理器的伺服器，其設計目的皆為以誘人的價格提供企業級的強大能力、規模彈性、控制與服務。

Cluster 1350 節點包含獨一無二的 Cable Chaining Technology (C2T) InterconnectTM 連接技術，能夠大幅降低每個系統所需的纜線數目，所以能加速升級速度，同時又能降低成本。除此之外，有一個整合系統管理處理器能讓 CSM 在遠端管理系統節點，增加伺服器的生產力。經由 CSM 命令的使用，系統管理員便能在記憶體、處理器、硬碟、風扇或電源所發生的各種事件中，指定要監視的事件以及要採取的行動。因此這些命令便能協助系統的效能與可用性達到顛峰狀態。

Cluster 1350 的標準架構包括 1 個管理節點、多達 512 個叢集節點及 32 個提供共享檔案儲存功能的儲存節點。對於需要更大規模或其他非標準架構的企業組織，IBM 亦提供另外的特別訂購程序。所有的節點均執行 Linux。

每套 Cluster 1350 亦包含一個可確保節點之間之通訊安全的管理 Ethernet VLAN、一個能使節點之間進行應用程式通訊的叢集 Ethernet VLAN，以及一個擁有遠端主控台功能的終端伺服器網路。叢集的標準配置具有一個 10/100 Mbps 乙太網交換機，也可選擇 10/100 Mbps 乙太網交換機、Gigabit 乙太網交換機或 MyrinetTM-2000 交換機。

叢集節點可以配置有單個或兩個 Intel XeonTM 處理器，記憶體可從 512MB 到 8GB。每個叢集節點都有一個或兩個硬碟，每個節點的儲存容量高達 440GB。管理節點還具有兩個 Intel Xeon 處理器，記憶體從 512MB 到 8GB，熱抽換磁碟儲存容量高達 440GB， 並帶有叢集管理介面卡。

選購的儲存節點可提供額外的儲存空間，以便設定額外的檔案系統儲存空間。儲存節點具有一個或兩個 Xeon 處理器，儲存容量從 512MB 到 8GB，熱抽換磁

碟儲存容量高達 440GB。在增加容量方面，這些節點可以支援外接式 Fibre Channel RAID 儲存子系統。

在高可用性方面，它們可以提供所有資料的備援路徑。標準架構所支援的儲存節點多達 32 個，透過特殊訂購則可提供更大的儲存架構。系統最少必須具備一個鍵盤/視訊/滑鼠 (KVM) 切換器。終端伺服器可提供遠端主控台支援。

3.2.5 系統擴充的可能性

Cluster 1350 提供許多選購元件來滿足不同機構組織的特定運算需求，包括互連技術的選擇。除了標準的 10/100 Mbps 乙太網路或 Gigabit 乙太網路外，機構組織也可以選擇 Myrinet 2000－來自 Myricom 公司的彈性互連技術。Myrinet 是一項具成本效益的高效能封包通訊與交換技術，目前已被廣泛運用於使用 Linux 作業系統的叢集中。它特別適用於具高效能或高可用性的叢集環境。

企業們也可以採用 GPFS for Linux 的優點。GPFS 是一個高效能的彈性檔案共享系統，能為 Linux 叢集環境內的所有節點提供高速資料存取功能。跨叢集中數個節點所執行的平行應用程式以及在單一節點中執行的序列應用程式都能使用標準的 UNIX 檔案系統介面來存取共享的檔案。再者，GPFS 也可以在磁碟機或伺服器故障時發揮故障復原功能。簡言之，GPFS for Linux 提供了世界級的檔案系統效能、規模彈性及可用性。它能配合 Linux 叢集的規模，並於叢集之外提供 NFS 匯出功能。

其他的叢集選購元件包括含 FAStT EXP500 擴充單元的 IBM FAStT200 儲存子系統，以及包含 FAStT EXP700 擴充單元的 IBM FAStT700 儲存子系統。Fiber Array Storage Technology 為需要高速傳輸與龐大資料。

3.3 Cluster 1350 規格

IBM @server® Cluster 1350 規格一覽表		
結構塊	**IBM @server® xSeries 345**	**IBM @server® xSeries 335**
節點類型	管理/叢集/儲存	叢集計算

外觀高度	機架 (2U)	機架 (1U)
處理器	2.6, 2.8 和 3.0 GHz Xeon 2-way 管理節點，1 或 2-way 叢集	2.6, 2.8 和 3.0 GHz Xeon 1 或 2-way 叢集計算節點或儲存節點
L2 快取記憶體	512KB	512KB
記憶體	512MB	512MB
磁碟擴充隔間	6 個 (熱抽換)	2 個 (熱抽換 SCSI)
I/O 插槽	5 個 PCI-X (1 個 32 bit, 4 個 64 bit)	2 個 PCI-X (64 bit)
磁碟控制介面及網路	內建 Ultra320 SCSI 和 2 個 Gigabit 乙太網路	內建 Ultra320 SCSI 和 2 個 Gigabit 乙太網路
系統連接管理 VLAN	一個標準 10/100 Mbps 乙太網路交換機	一個標準 10/100 Mbps 乙太網路交換機
系統擴充		
記憶體	8GB	8GB
SCSI 內部儲存容量	36.4 GB -- 440.4 GB	36.4 GB -- 146.8 GB
系統連接叢集 VLAN	—	可選擇 10/100 Mbps 乙太網路、Gigabit 乙太網路、 Myrinet-2000
搭配選購	Gigabit 乙太網路 SX Myrinet 133 MHz 系列 FAStT FC-2 主機適配器 ServerRAID 4Lx	Gigabit 乙太網路 SX Myrinet 133 MHz 系列 ServerRAID 4Lx
儲存	FAStT200 儲存控制器， EXP500 擴展儲存櫃 FAStT700 儲存控制器， EXP700 擴展儲存櫃	
作業系統	Red Hat Linux 7.3, 8.0 SuSE 8.0, 8.1, SLES 7, SLES 8 AS2.1	Red Hat Linux 7.3, 8.0 SuSE 8.0, 8.1, SLES 7, SLES 8 AS2.1
系統管理軟體	CSM for Linux 1.2 和 1.3	CSM for Linux 1.2 和 1.3
系統大小		
42U 基本或擴容機架	79.5" 高 x 25.5" 寬 x 43.3" 深 (2019.2 mm x 647.7 mm x 1099.8 mm), 575 磅 (260.9 kg)	
管理／儲存節點	3.36" 高 x 17.5" 寬 x 27.5" 深 (85.3 mm x 444.2 mm x 697.4 mm), 62 磅 (28.1 kg)	
叢集節點	1.72" 高 x 17.3" 寬 x 25.7" 深 (43.7 mm x 439.9 mm x 653.3 mm), 26 磅 (11.8 kg)	

可擴展性	要求一個管理節點，至少 4 個、最多 512 個叢集節點。另外，可以最多配置 32 個儲存節點。因此，最小配置包括 5 個節點。 （一個管理節點和四個叢集節點）。最大配置包括 513 個節點 （一個管理節點和 512 個叢集和儲存節點）。通過具體的投標過程可以提供大型配置。
服務	包括安裝服務。額外的 Linux 叢集支援選購服務。
保固	對大部分 IBM 元件的基本有限保固：3 年或 1 年，下一工作日回應、現場支援。

4 建置 Cluster 1350 on Non-Supported H/W

4.1 Cluster1350 架構

一個標準 Cluster1350 的架構應如圖 5，由圖及第 3.3 節可知建置 Cluster 1350 需用 IBM 特定幾款的伺服器。其實利用一般的 PC 也可建置 Cluster 1350，但是很多硬體監控的機制就無法使用，也無法自動收集各個 node 的硬體資訊，例如 MAC Address。

筆者接下來便介紹如何利用 Non-Sopported 的 PC 來建置 Cluster 1350；首先我們得建置 Management Server。在 Cluster1350 中的 Management Server 最主要用途有二：1.管理所有 node 為 Linux Node。2.扮演 Installation Server。只要把 Management Servere 建置好，其餘的 node（一般稱為 Computing Node）只要設定網卡開機，開機後便會根據 Management Server 上的自動安裝設定檔（kickstart 或 AutoYast 設定檔），當 1350 Cluster 安裝完成後，可以使用 dsh 或 rsh 跟所有的 Computing Node 溝通。這樣一來便建置好可以安裝 PVM（Parallel Virtual Machine）或 MPI（Message Passing Interface）的環境。

至於如何安裝 PVM （Parallel Virtual Machine）或 MPI（Message Passing Interface）及開發平行處理的程式，則不在本篇文章探討的範圍（這部份筆者也不懂，沒有接觸過 ^O^）。

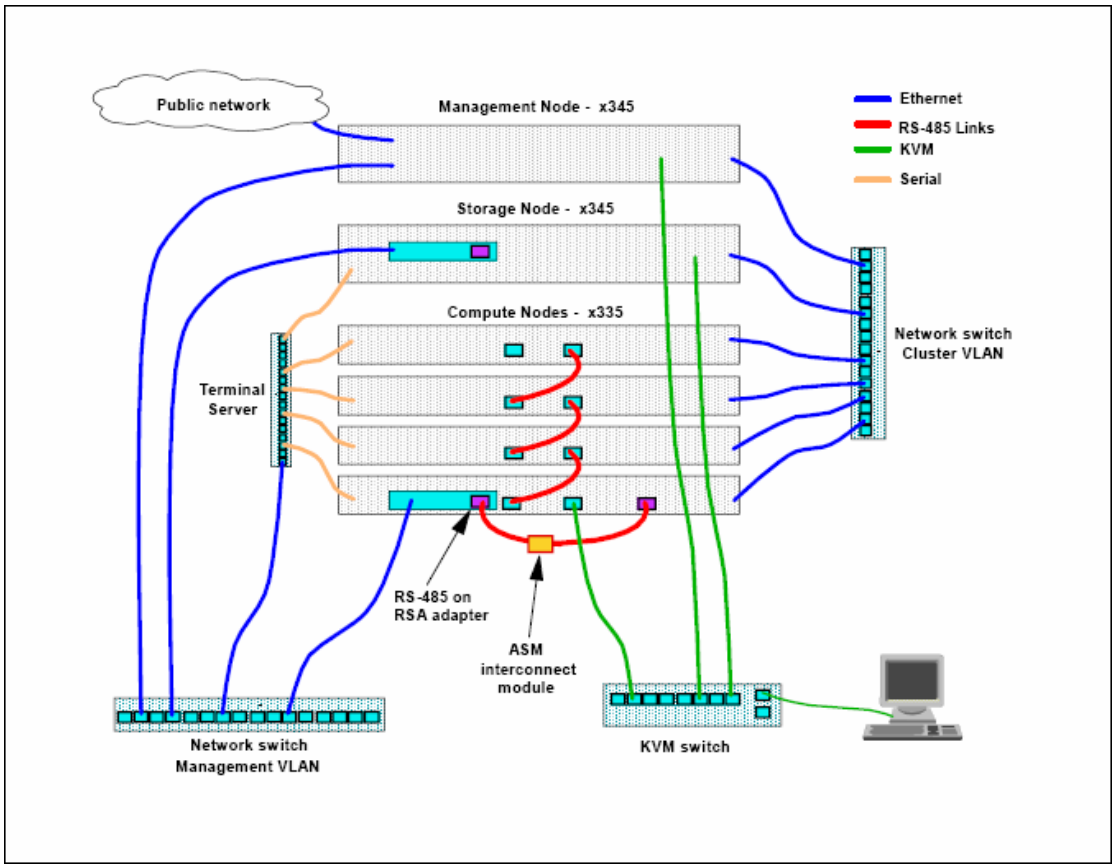

圖 5：Cluster 1350 架構圖

4.2 實作環境

Manager node : RHEL 4 U1

Computing node : RHEL 4 U1

4.3 實作步驟：

___ 1. Cluster 1350 規劃

建置 Cluster1350 之前得先規劃好所有主機的主機名稱、IP Address…，因為我們是要安裝在 Non-Supported 的 H/W 上，所以必須手動收集各個 node 上的 MAC Address。

hostname	Node Type	RAM	IP Address	MAC Address
manage.cluster.com	management	376MB	192.168.0.100	
node1.cluster.com	computing	256MB	192.168.0.101	**00:0c:29:DF:29:BA**
node2.cluster.com	computing	256MB	192.168.0.102	**00:0c:29:9C:64:AD**
node3.cluster.com	computing	256MB	192.168.0.103	**00:0c:29:9C:64:CA**

___ 2. 安裝 Linux 伺服器

讀者得先安裝一台 Linux 伺服器做為 Management Server 之用，可採用預設安裝，但記得先關閉 Firewall 及 SELinux，還有安裝完成後，硬碟至少得有 4~5GB 的空間。

___ 3. 下載 autoupdate rpm

ftp://ftp.mat.univie.ac.at/pub/teschl/autoupdate/，並將 autoupdate rpm 置於 /tmp/csmreqs 目錄下。

```
# mkdir /tmp/csmreqs
# cp autoupdate-cfg-redhat-5.4.1-1.noarch.rpm /tmp
```

___ 4. 修改/etc/hosts

```
127.0.0.1       localhost.localdomain   localhost
192.168.0.100   manage.cluster.com      manage
192.168.0.101   node1.culster.com       node1
192.168.0.102   node2.cluster.com       node2
192.168.0.103   node3.cluster.com       node3
```

___ 5. 建立/csminstall 目錄或檔案系統

/csminstall 目錄是用來存放 Computing Node 安裝時所需要的套件，約需 3GB 的空間。

```
# mkdir /csminstall
```

___ 6. 下載 Cluster 1350 管理軟體－CSM

http://techsupport.services.ibm.com/server/csm/download/home.html

下載 csm-linux-1.5.0.0.i386.tar.gz ，請將檔案置於 /tmp/csm 目錄，並將其解壓。

```
# mkdir /tmp/csm
# cd /tmp/csm
# tar -xzvf csm-linux-1.5.0.0.i386.tar.gz
```

___ 7. 安裝 csm.core 套件

```
# rpm -i /tmp/csm/csm.core-*
```

【註】安裝完畢後，請重新登入，讓新的環境變數生效。

___ 8. 安裝 CSM 軟體

```
# installms -p /tmp/csmreqs:/tmp/csm
```

請勿用 rpm 的指令直接安裝 csm 軟體，你必須執行 installms 指令來安裝 CSM 軟體。installms 的語法為 installms –p <autoupdate rpm 所在目錄>:<csm rpm 所在目錄>，執行過程中應會要求放入 RHEL 光碟片，安裝 Management Server 所需套件。

___ 9. 接受試用版 Licsense

```
# csmconfig -L
```

IBM 網站提供的 CSM 軟體有 60 天的試用期，不過讀者必須執行「csmconfig –L」接受 License。

___ 10. 定義 computing node 的資料

利用 definenode 指令，可將 computing node 的資料寫入 Management Server 的 1350 Cluster 資料庫

```
# definenode -n node1 InstallMethod=kickstart ConsoleSerialDevice=NONE
InstallAdapterName=eth0 InstallAdapterMacaddr=00:0c:29:DF:29:BA（同一行）

# definenode -n node2 InstallMethod=kickstart ConsoleSerialDevice=NONE
InstallAdapterName=eth0 InstallAdapterMacaddr=00:0c:29:9C:64:AD（同一行）

# definenode -n node3 InstallMethod=kickstart ConsoleSerialDevice=NONE
InstallAdapterName=eth0 InstallAdapterMacaddr=00:0c:29:9C:64:CA（同一行）
```

新增 node1~node3 定義後，可執行 lsnode 檢查是否新增 node 定義成功。

```
[root@manage ~]# lsnode
node1.cluster.com
node2.cluster.com
node3.cluster.com
```

___ 11. 修改 KickStart 範本檔

 # vi /opt/csm/install/kscfg.tmpl.RedHatEL-ES4

35 clearpart --all --initlabel ~~--drives=#CSMVAR:INSTALL_DRIVER#~~

因為 management 是利用 kickstart 的方式來自動安裝 computing node，而其參考的範本檔位於/opt/csm/install 目錄內；我們的作業系統為 RHEL 4 U1，請修改第 35 行將「--drives=#CSMVAR:INSTALL_DRIVER#」此段文字刪除，因為我們是在一般 PC 上安裝 CSM，所傳回的 INSTALL_DRIVER 的值會有錯誤，所以得將其刪除。若是讀者想自訂 Computing Node 預設安裝的軟體清單，請自行修改此檔案。

___ 12. 建置 Installation Server 並產生各個 node 的 kickstart 檔案及更新 dhcpd.conf

```
# csmsetupks -P
```

只要執行「csmsetupks –P」，CSM 便會要求你收入 RHEL 光碟，並根據之前所定義的 node 的資料，幫你更改 kickstart 檔案及 dhcpd f 伺服器的設定。

___ 13. 指定安裝那些 node。

```
# installnode –P –noreboot
```

installnode –P –noreboot 指令中，-P 的選項的意思為「PreManaged」即安裝所有狀態為 PreManaged 的 node。所謂「PreManaged」的代表這個 node 已定義但還未安裝。node1~node3 現在的狀態都是「PreManaged」，所以 installnode –P –noreboot 指令會幫這 3 個 node 修改相關的 PXE 設定檔。當 Computing node 安裝結束後，其狀態為變為「Managed」。

```
installnode –P –noreboot
```

___ 14. Power on node1~node3

這時候只要打開node1~node3的電源（記得設定網卡開機），node1~node3便會自動安裝。

___ 15. 測試 dsh 環境

待node1~node3安裝完畢後，可登入manage server，執行下列「dsh –a date」指令，應不用輸入密碼就可得到node1~node3上的時間。

```
[root@manage ~]# dsh -a date
node1.cluster.com: Wed Aug 23 16:05:15 CST 2006
node2.cluster.com: Wed Aug 23 16:05:15 CST 2006
node3.cluster.com: Wed Aug 23 16:05:15 CST 2006
```

5 參考資料：

1.「楊朝棟教授」叢集式平行電腦系統介紹：
http://www.hpc.csie.thu.edu.tw/ctyangweb/index.php?sub=claster

2.鳥哥的「簡易 Cluster 架設」:http://linux.vbird.org/linux_server/0600cluster.php

3.IBM Cluster Systems Management for AIX 5L and LinuxPlanning and Installation Guide
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12002.html

4.CSM for AIX 5L and Linux V1.5 Administration Guide
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7ad12002.html

5.CSM for AIX 5L and Linux V1.5 Command and Technical Reference
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7cm12002.html

6.Linux Clustering with CSM and GPFS 紅皮書
http://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/SG246601.html

作者簡介

林彥明（Alex Lin）：負責 Linux 相關技術支援工作。具有 RHCX（Red Hat 認證主考官）、RHCE、NCLP（Novell Linux 認證專家）、LPIC、IBM AIX Expert、IBM MQ、SCJP、SCWCD 等國際認證，曾參與建置臺灣第一套商業用 IBM Cluster1350 叢集系統及 RHEL 4 及 SLES 9 on zSeries、奇美、中華電信等 Linux 專案。