如何組裝一部平行電腦

陳昭安[*] 吳明佳[†] 胡進錕[‡]
中央研究院物理研究所
[*]e-mail: chenja@phys.sinica.edu.tw
[†]e-mail: mcwu@phys.sinica.edu.tw
[‡]e-mail: huck@phys.sinica.edu.tw

摘　要

本文簡單介紹為計算物理建構無磁碟平行計算叢集的程序。首先，我們將簡短地瀏覽 Linux 作業系統。依據這些知識，我們希望讀者對於整個架設的過程有一較佳之了解。

一、簡介

對於複雜系統的研究，計算模擬是一個強有力的工具。由於其強大的計算能力與逼真的動畫性能，傳統嘗試與錯誤（trial-and-error）的研究過程可以經由互動式的介面，藉由改變參數來加速，並且可以藉由電腦繪圖或動畫來了解模擬的結果。模擬也可以提供一個研究機會，對某些不容易觀察，或因為太昂貴而不能真實實現的過程做研究。

為了直接與有效率地解決複雜的問題，許多方面都是涉及計算物理的。其中之一是如何加速模擬的過程。一種加速的方法是使用或購買昂貴的平行超級電腦。不過，對於大多數研究群來說，這是很不實際的。而拜當今個人電腦（PC）的先進科技所賜，我們在硬體與軟體上，可以完全從大眾的角色，以負擔得起的預算，建構平行計算叢集。

最初的個人電腦叢集計畫，亦稱為比歐胡（Beowulf）專案，早於 1994 年初，便已經在國際航空和太空總署（NASA）太空資料與資訊科學卓越中心開始進行了。它通常是一個由一個主機或伺服器節點所組成的系統，而一個或更多個客戶端節點則經由乙太網路連接在一起。主節點控制整個叢集，並提供檔案給客戶端節點。主節點也是叢集對外界網際網路世界的控制台與閘道器。

類比歐胡叢集（Beowulf-like Cluster）的優點有：

- 硬體可以從多種來源獲得，這意味著低的價格與簡單的維護。
- 作業系統（LINUX）與平行程式套件（MPI、PVM 等等）兩種軟體皆可從網際網路中免費獲得。
- 這些軟體通常是依據電腦工業的標準。
- 在 GNU 之一般公眾執照下，原始碼對於每一個人都是免費的，這表示原始碼可以依據個人的需要作修改或改良。
- 可以從網際網路上找到大量建構比歐胡叢集的免費文件與指導。
- 考量性能價格比，它真的是便宜的。

去年（2001 年），我們在中央研究院物理所統計與計算物理實驗室（Laboratory of Statisitical and Computational Physics）建構了類比歐胡的平行計算叢集來測試這個構想 [4]。這個叢集包含一個主節

點、一個網路檔案系統（NFS）與上層網路檔案系統（NFS-root）伺服器節點，以及許多無硬碟的客戶端節點。所有這些節點都藉由乙太網路連接到伺服器集線器，以獲得平行計算的能力。此一叢集之硬體配備與軟體組態詳列如下：

硬體配備

・一個主機節點：

雙 Pentium III 1GHz 處理器，512MB 記憶體，三個 3Com 3c905c 乙太網路卡，一個 30G 硬碟，一個軟碟，一個 VGA 卡，一個螢幕

・一個網路檔案系統（NFS）與上層網路檔案系統（NFS-root）伺服器節點：

雙 Pentium III 1GHz 處理器，512MB 記憶體，兩個 3Com 3c905c 乙太網路卡，一個 30G 硬碟，一個軟碟，一個 VGA 卡，一個螢幕

・11 個客戶端節點：

一個 Pentium III 1GHz 處理器，512MB 記憶體，兩個 3Com 3c905c 乙太網路卡，無硬碟，一個軟碟，一個 VGA 卡（除錯用）

・兩個集線器

D-Link DES-1016R，D-Link DFE-916DX

軟體組態

・作業系統：RedHat 6.2

・網路開機：etherboot 4.0

・平行計算：Message Passing Interface mpich-1.2.1

・顯示器：X windows library, OpenGL library, Tc1/Tk

在本文中，我們將與讀者分享我們如何建構這個叢集的經驗。不過，實在很難提供讀者每一步精確的過程。原因很簡單，Linux OS 發展得很快，今天的安裝可能明天就沒用了。所以，我們將提供一般性的導引、基本的原理，以及如何從網際網路上，尋找額外的資訊，而不是提供詳細的敘述。為了成功地建構你們自己的叢集，一些關於 Linux OS 的知識是必要的。只有用這些知識，當發生錯誤時，才有足夠的信心解決它。總之，本文的大部分讀者是物理學家，而物理學家是不想盲目地做事的。

二、Linux 作業系統

建構無硬碟客戶端個人電腦叢集的主要挑戰，是如何啟動核心，以及如何從遠端伺服器安裝到上層檔案系統。既然客戶端節點沒有硬碟來對其核心與檔案系統提供主機服務，所以他們必須經由網路的連接，由其他的伺服器來提供。

此一伺服器節點需要 RedHat Linux OS 之完全安裝。

此一伺服器必須準備一個簡化的核心映像供客戶端節點下載。

此一伺服器必須準備最小化的上層檔案系統，供客戶端節點安裝。

客戶端節點必須具有網路開機磁碟，以從其軟碟機開機，然後從網路上取得核心映像。

客戶端節點從伺服器找到核心映像，將其下載、解壓縮，並且執行。

客戶端之核心映像將安裝在其上層檔案系統，以作為上層網路檔案系統。根據以上分析，我們需要知道：

・何謂 Linux 核心？

・如何為無硬碟客戶端建立核心映像？

・如何為無硬碟客戶端建立上層檔案系統？

・Linux 之啟動程序？

。啟動伺服器核心？

。啟動客戶端核心？

何謂 Liunx 核心？

Linux 核心的目的在於將硬體的複雜性與使用者隔離開來。它為使用者提供一組系統函數呼叫，

以避免處理硬體的細節。舉例來說，如果使用者想要從硬碟存取一個檔案，只要簡單地對核心發佈一個讀取系統的呼叫，核心便會處理細節，如移動磁碟讀取/寫入臂到硬碟的正確位置（磁軌、磁區），並將檔案的內容傳回給使用者。從這方面看來，你可以了解，如果你要處理電腦系統而沒有核心，那將會是個夢魘。

Linux 核心是一個多工，多使用者的系統。它包含許多元件，如程序管理、記憶體管理、檔案系統、裝置控制、網路連接等等。它藉由公平地分派中央處理器（CPU）、記憶體、輸入/輸出裝置、網路資源，來回應使用者的要求。總之，Linux OS 的核心是一大塊掌管處理所有這些要求的執行碼。如果系統想要有作用，第一件事情就是下載與執行此一核心。

依據應用，核心的大小可以很大或很小。舉例來說，如果你不需要 PCMICA，你可以不需要將其包含到核心中。通常，較大的核心提供更多的服務，但也會消耗更多的中央處理器（CPU）時間，而使系統慢下來。對於伺服器核心，它是大的，因為我們要求它做很多事。對於客戶端核心，它則是比較小的，因為它只是簡單地執行由伺服器指定給它的程式。

建立核心映像

建立核心映像的程序可以參見讀我（README）檔案，此一檔案係跟隨核心原始碼與核心（HOWTO）。

安裝核心：

・如果你安裝全部的原始碼，則下指令

cd/usr/src

gzip –cd linux-2.2.xx.tar.gz | tar xvf

使其全部放在適當的位置，其中 gzip 中的選項 -cd 意指解壓縮檔案，然後將輸出傳送到 stdout。

此處，請以最新核心之版本號碼取代「xx」。

・確認其中沒有舊的 .o 檔案與其附屬檔案：

cd /usr/src/linux

make mrproper

配置核心：

・執行 make config 以配置基本核心。Make config 需要 bash 來工作：它將會搜尋$ BASH，/bin/bash 與/bin/sh（以此順序）中的 bash，所以為了使其作用，其中之一必須是正確的。關於核心組態的進一步資訊，請參見 Documentation/Configure.help

即使你只是升級早期的版本，也不要跳過這個步驟。

・另一可選擇的組態指令是：

。make menuconfig

。make xconfig

。make oldconfig：依據你存在的./.config 檔案，預設所有的問題。

・檢查最上面的 Makefile，以供進一步的定址組態。

・最後，執行 make dep 以正確地安裝所有的附屬程式。

編譯核心：

執行 make zImage 或 make bzImage 以產生一壓縮之核心映像。

如果你將核心的任何部分配置成模組，你必須執行 make modules，緊接著執行 make modules_install。進一步資訊可閱讀 Documentation/modules.txt。

建立上層檔案系統

除了核心，你也需要上層檔案系統來為程式、組態，與資料提供主機服務。產生上層檔案系統包

含從系統中選擇需要的檔案來運作。

一個上層檔案系統必須包含支援完整 Linux 系統的每一樣東西。為了能夠做到這一點，磁碟必須包含 Linux 系統的最小需求：

- 基本檔案系統結構,
- 最小組目錄：/dev, /proc, /bin, /etc, /lib, /usr, /tmp,
- 基本公用程式組：sh, ls, cp, mv, 等等,
- 最小組組態檔：rc, inittab, fastab, 等等,
- 裝置：/dev/hd*, /dev/tty*, /dev/fd0,等等,
- 運作時的程式庫, 以提供公用程式使用的基本函數。

為了建立此一上層檔案系統，你需要一個大到足以容納壓縮前所有檔案的空白裝置。這有許多的選擇，在此我們選擇記憶體磁碟（ramdisk)。

使用記憶體磁碟（DEVICE=/dev/ram0)。在此一情況中，使用記憶體來模擬一個硬碟機。要學習如何使用記憶體磁碟，請參見「How to Use a Ramdisk for Linux」的連結： http://www.linuxfocus.org/English/November1999/article124.html

準備 DEVICE：

dd if=/dev/zero of=/dev/ram0 bs=1k count 4096

此一指令將裝置調零。將裝置調零是很重要的，因為後來檔案系統將會被壓縮，所以所有不使用的部分將會填零，以達到最大的壓縮。

接著，產生檔案系統。

Mke2fs –m 0 –N 2000 /dev/ram0

然後，產生安裝點並安裝裝置。

Mkdir –p /tmp/ramdisk

Mount –t ext2 /dev/ram0 /tmp/ramdisk

移植檔案系統

下列是對你的上層檔案系統合理的最小目錄組。

- /dev – 裝置檔案，執行輸入/輸出（I/O）所需
- /proc – 目錄結構，proc 檔案系統所需
- /etc – 系統組態檔案
- /sbin – 重要的系統二進位碼
- /bin – 系統之一部份基本的二進位碼
- /mnt – 維護其他磁碟的安裝點
- /usr – 額外的工具與應用程式

首先，產生以上所列的目錄。

cd /tmp/ramdisk

mkdir dev proc etc sbin bin mnu usr usr/lib

關於產生/dev

cp –dpR /dev/fd[01]* /tmp/ramdisk/dev

cp –dpR /dev/tty[0-6] /tmp/ramdisk/dev

或是

mknod console c 5 1

在 ramdisk.tar 中, 有關於上層檔案系統的詳細內容。最後，在你安裝所需要的所有程式庫之後，必須執行 ldconfig，以重新於上層檔案系統上，產生 /etc/ld.so.cache。此一快取會告訴載入程序哪裡尋找程式庫。你可以執行

ldconfig –r /tmp/ramdisk

當你已經完成建構上層檔案系統時，請不要安裝它，應將其拷貝成檔案並壓縮。

Umount /tmp/ramdisk

Dd if=/dev/ram0 bs=1k | gzip –v9 > rootfs.gz

轉移上層檔案系統

dd if=rootfs.gz of=/dev/fd0 bs=1k seek=KERNEL_BLOCK

Linux OS 的啟動程序

所有的個人電腦系統藉由執行唯讀記憶體（ROM）（具體地說，基本輸入輸出系統（BIOS））中的碼, 以從啟動磁碟的磁區 0, 磁柱 0 載入磁區，開始啟動過程。啟動磁碟通常是第一個軟碟

（/dev/fd0）或第一個硬碟（/dev/hda）。接著，BIOS 嘗試執行此一磁區。在大部分的可啟動磁碟上，磁區 0，磁柱 0 包含：

・來自啟動載入程序，並位於核心的碼，如 LILO，將其載入，並且執行以正常地啟動；或

・作業系統，如 Linux 核心的啟動。

當核心完全載入時，它將初始化裝置驅動程式與其內部的資料結構。一旦完全初始化，它會查詢在其映像中的位置，即所謂的記憶體磁碟代碼（ramdisk word）。此一代碼告訴其如何與何處找到其上層檔案系統。上層檔案系統通常只是被安裝成「/」的檔案系統。核心必須被告知何處尋找其上層檔案系統；如果它在那裡無法找到可定址的映像，它就停止不動了。

在某些啟動的情況中（通常是在從磁碟啟動時），上層檔案系統被載入記憶體磁碟，此一記憶體磁碟是以硬碟的方式，由系統存取的記憶體。而且，核心可以從軟碟機載入壓縮的檔案系統，將其解壓縮至記憶體磁碟，並允許更多的檔案存放到磁碟中。一旦載入並安裝上層檔案系統，你可以看到類似的訊息：

VFS：Mounted root （ext2 filesystem）readonly

一旦系統成功地載入上層檔案系統，它將嘗試執行 init 程式（位於/bin 或/sbin），init 讀取其組態檔案/etc/inittab，尋找指定 sysinit（/etc/rc.d/rc.sysinit）的線索，並執行副本（script）。此一副本是一組安裝基本系統服務的外殼（shell）指令，諸如硬碟中的 fsck，將所需之核心模組載入，初始化交換檔，初始化網路，並安裝/etc/fstab 中所提到的磁碟。

此一副本通常引用其他副本以完成模組初始化。舉例來說，在一般的 SysVinit 結構中，目錄/etc/rc.d 包含次目錄的複雜結構，其檔案詳細地說明如何使系統服務啟動與停用。可是，在啟動磁碟上，sysinit 副本通常是非常簡單的。

當 sysinit 副本完成控制並回到 init 時，接著進入預設執行階層，並由/etc/initab 與 initdefault 關鍵字所指定。

三、建立無磁碟個人電腦叢集之詳細程序

建立個人電腦叢集的程序可以分成兩個部分：

1.網路檔案系統（NFS）與上層網路檔案系統（NFS-root）伺服器安裝。

2.客戶端節點安裝。

網路檔案系統與上層網路檔案系統伺服器安裝

安裝伺服器是很直接的。有很多的方式安裝你的伺服器，無論是藉由安裝磁碟（由 RedHat 提供）或網路。為了簡單起見，讓我們假設在你的伺服器電腦上，有一個完整安裝的 RedHat 6.2。如果安裝程序是正確的，你應該具有網路連線的全功能 Linux OS。其次，你必須為客戶端電腦準備一片網路啟動磁碟片：

・從軟碟機完成網路啟動：

從 http://www.slug.org.au/etherboot 下載 etherboot-4.0 與 therboot-4.7. 24。從 etherboot-4.0/bin 取得 floopyload.bin，並從 etherboot-4.7.24/src/bin32 取得 3c905c-tpo. lzrom，然後鍵入下列指令以從網路製作啟動軟碟（你必須是超級使用者）

#cat floppyload.bin 3c905c-tpo.lzrom > /dev/fd0

注意：為了取得 3c905c-tpo.lzrom，你必須到 etherboot-4.7.24/src 執行 make。詳細資訊，請參見 INSTALL 指令。

・安裝 dhcp 伺服器程序

1.準備 /etc/dhcpd.conf 檔案，如：

#Sample configuration file for ISCD dhcpd

```
#
#Don't forget to set run_dhcpd=1 in /etc/init.d/dhcpd
#once you adjusted this file and copied it to /etc/dhcpd.config.
#
default-lease-time                21600;
max-lease-time                    21600;

option subnet-mast                255.255.255.0;
option broadcast-address          192.168.0.255;
shared-network WORKSTATIONS {
      subnet 192.168.0.0 netmask 255.255.255.0 {
      }
}
group      {
      use-host-decl-names    on;
      option    log-servers     192.168.0.254

      host pc1 {
      hardware ethernet       00:01:02:92:70:69;
      fixed-address           192.168.0.1;
      filename
"/tftpboot/pc1/vmlinuz.3c905nomodpc1";
      }
      host pc2 {
      hardware ethernet       00:01:02:91:43:OF;
      fixed-address           192.168.0.2;
      filename
"/tftpboot/pc2/vmlinuz.3c905nomodpc2";
      }
      host pc3 {
      hardware ethernet       00:01:02:92:70:18;
      fixed-address           192.168.0.3;
      filename
"/tftpboot/pc3/vmlinuz.3c905nomodpc3";
      }
      host pc4 {
      hardware ethernet       00:01:02:91:43:45;
      fixed-address           192.168.0.4;
      filename
"/tftpboot/pc4/vmlinuz.3c905nomodpc4";
      }
```

2.編輯/etc/rc.d/init.d/dhcpd 副本檔，尋找下行

daemon /usr/sbin/dhcpd

然後將其改成

daemon /usr/sbin/dhcpd eth1 （保持 eth0，不要修改它）

3.檢查檔案 /var/state/dhcp/dhcpd.leases 是否存在，如果沒有的話，則下指令

touch /var/state/dhcp/dhcpd.leases

以產生該檔案。

4.在/etc/rc.d/rc3.d 中，增加一個軟性連接

ln –s ../init.d/dhcpd S65dhcpd

現在，你可以藉著放進網路啟動軟碟片到客戶端個人電腦，並關閉電源來測試 dhcp 伺服器。

・安裝 tftp 伺服器之程序

1.檢查/etc/services，以確定下行是否存在：

tftp 69/udp

2.檢查/etc/inetd.conf，以確定下行是否出現：

tftp dgram udp wait root /usr/sbin/tcpd in.tftpd

3.再一次開始 inetd，以讀取新的組態檔案。

4.tftp daemon 係由 inetd 產生，可是你必須確定/etc/hosts.allow 包含下行：

ALL: 192.168.0.

或更具體地說，包含下列：

#bootpd: 0.0.0.0（對於 bootpd, 出現這一行）

in.tftpd: 192.168.0.

portmap: 192.168.0.

5. 在 /etc/hosts 中增加主機名稱，並與 /etc/dhcpd.conf 的內容一致。例如：

192.168.0.1 pc1

192.168.0.2 pc2

192.168.0.3 pc3

192.168.0.4 pc4

・以下列指令，在伺服器之/tftpboot 目錄上產生 pc1，pc2，pc3 與 pc4 的根目錄。

mkdir –p /tftpboot/pc1

mkdir –p /tftpboot/pc2

mkdir –p /tftpboot/pc3

mkdir –p /tftpboot/pc4

・為客戶端節點準備核心。當設定核心參數時，你必須確定你指定下列：

*沒有模組支援（為了簡單起見）。

*支援你的特殊網路卡, 舉例來說, 3com 3c905c。

*記憶體磁碟支援。

*BOOTP 支援。

*/proc 檔案系統支援。

*網路檔案系統（NFS）支援。

*網路檔案系統（NFS）上的上層檔案系統。

・當你從核心原始碼產生新的核心之後，舉例來說，將其命名為 bzImage。執行下列指令，以產生網路可啟動核心。（ /usr/local/bin/mkNetKernelA.bat）

./mknbi-linux –rootdir=/tftpboot/pc$1/pc$1root /usr/src/linux/arch/i386/boot/bzImage > /tftpboot/pc$1/vmlinuz.3c905nomodPc$1

・為每一個各戶端準備上層檔案系統。

將伺服器上層檔案系統拷貝到 /tftpboot/pc1, 並將任何不需要的檔案或套件刪除，以減少客戶端上層檔案系統的大小。修改網路設定、網路檔案系統設定，以及/tftpboot/pc1/etc 目錄中的其他設定。

四、結語：

無硬碟平行電腦的優點包括耗電能少、故障率低、造價低等等。有了組裝前述平行電腦的經驗之後，下一步我們將組裝一部更大的平行電腦，以便計算巨分子的熱力學性質與其結構[5]。

參考資料

1. Linux Documentation Project (http://www.linuxdoc.org)
   包含許多 Linux OS 之各方面的說明。本文之進一步細節可以在這裡找到。
2. The Beowulf Project（http://www.beowulf.org）
   比歐胡專案官方網站。
3. Etherboot（http://www.slug.org.au/etherboot）
   在乙太網路上啟動核心映像。
4. 中央研究院物理所統計與計算物理實驗室（Laboratory of Statisitical and Computational Physics）的網址為：http://www.sinica.edu.tw/~statphys/；其中之「Computer Facilities」有關於組裝平行電腦的進一步訊息。
5. 參考本期胡進錕與林財鈺的文章：*用平行電腦及解析方法研究巨分子*。