【漫步在雲端】系列報導-5

雲端計算～榨乾你的電腦效能

高雄區農業改良場 鄭文吉

※本文已於2012年12月發表於農業世界雜誌352期75-80頁※

前言

上期以平行運算的概念介紹雲端處理的運作方式，希望能讓大家了解，並不是什麼東西都適合採用雲端處理，而是要看你想處理的事情的特性而定。對一般人日常生活或工作上所遇到的需求來說，不管是打報告、整理資料、上網站、聽音樂、看影片、玩遊戲...其實隨便找台最便宜的個人電腦就足夠應付了。

問題是，有些東西就是要花很長的時間才能處理完畢，以致我們會希望電腦的運算速度可以再快一點，不要讓人等那麼久。但問題是那種新電腦不知幾時才會出現，因此不如想辦法讓現在的電腦效能提昇，這樣還比較實際一點。

有很多方法可以更有效率的運用我們的電腦效能，以下舉幾個例子說明：

一、 科學研究：

雖然一般人不會有很多資料需要計算，但對於科學研究人員來說，經常需要處理十分龐雜的數學計算。例如進行蒙地卡羅模擬試驗時，就需要同時探討多種變因下不同變化等級的可能結果，然後從中找出最佳處理組合。假設現在有個試驗包含 6 種變因、每種變因各有 10 種變化等級，這樣就會有 10^6 = 100 萬種變化組合，如果每一種組合的計算要花 1 秒鐘，這樣就得花 100 萬秒～大約 11 天半的時間才能算完所有的組合。這時若能運用平行處理的觀念，先借到 10 台電腦，然後將計算工作分成 10 部分放在不同電腦上分別計算，之後再把結果整合起來，就可以把計算時間減少到只剩十分之一；如果能借到的電腦越多，所花的時間就可以越精簡。

上面只是舉例，比那個更複雜的科學研究多得很。例如天文學家想找出能夠證實外星智慧生物存在的證據，於是利用世界最大的無線電望遠鏡(如圖 1)掃描整個天空採集無線電信號，然後從中比對不正常的部份，這些就有可能是外星智慧生物所發出的。由於所收到的資料量實在太龐大，用天文台的電腦根本分析不完，又買不起超級電腦來用(想想

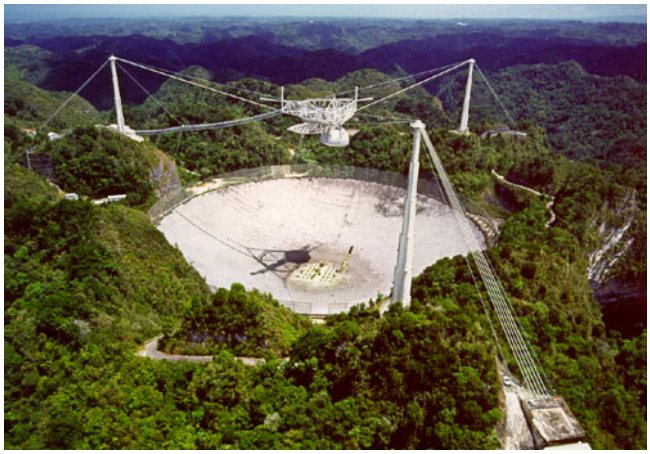

圖 1. 位於波多黎各的阿雷西沃山谷中的阿雷西博天文臺(Arecibo Observatory)無線電望遠鏡，直徑 350 公尺。

上期介紹的那台由幾千台伺服器組成的紅杉超級電腦，就知道有多貴)。因此就有人想到，不如把資料切成很多小片段，然後號召全世界使用者幫忙分析。使用者只要註冊並下載他們設計的程式執行之後，程式就會自動到網站下載資料片段來做分析，算完會自動回傳結果，再下載另一個片段繼續分析(如圖 2)。

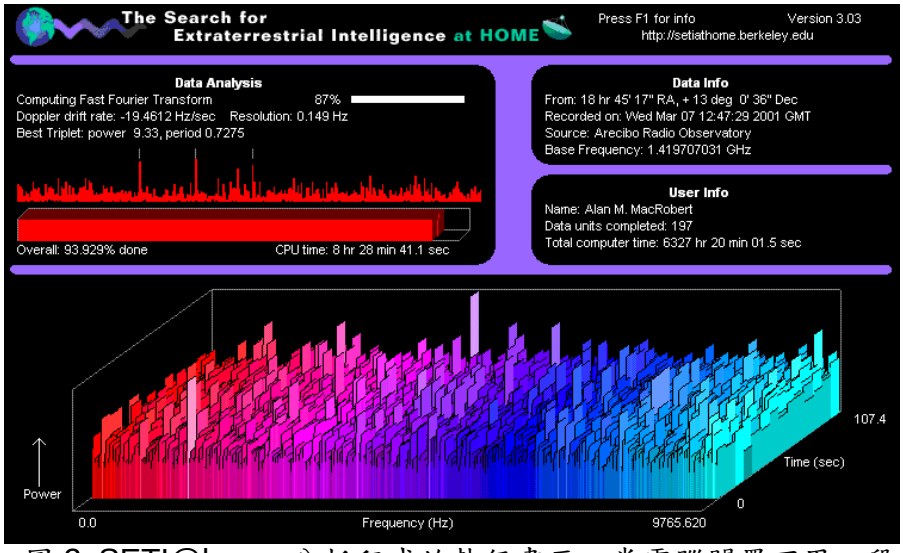

圖 2. SETI@home 分析程式的執行畫面，當電腦閒置不用一段時間後，就會出現這個畫面當作螢幕保護程式。

這就是有名的「在家搜尋外星智慧文明計畫」(SETI@home, Search for ExtraTerrestrial Intelligence at Home)，目前有 325 萬台世界各地的電腦加入運算，整體速度達 563TeraFLOPS (每秒 563 兆次浮點運算)，比 2007 年世界最快的超級電腦 IBM Blue Gene/L (速度 478TeraFLOPS)還快。不過可惜的是，雖然用這種方式分析了這麼多年，到現在還是沒找到能夠證實外星智慧生物存在的證據，所以現在這個工作還在繼續進行中，有興趣的朋友也可以加入幫忙分析。

這樣集合一堆使用者電腦幫忙進行科學計算的作法，稱為「分散式計算」。上期提過，我們的電腦計算功能並沒有完全發揮，平常幾乎都在閒置狀態，而這些分析程式就是利用這些剩餘的計算能力做計算，因此對使用者根本沒有影響；但研究人員卻可以藉此得到足以比擬超級電腦的計算能力，而能進行以前因為經費不足而不敢嘗試的科學分析。因此除了前面提到的天文研究外，像是數學、物理、氣象、蛋白質結構及新藥物研究等需要大量計算工作的研究領域，也有很多採用分散式計算進行分析。而負責維運 SETI@home 計算平台的加州大學柏克萊分校，也將這個平台加以擴充，提供各領域研究者使用，稱為「柏克萊開放式網路計算平台」(Berkeley Open Infrastructure for Network Computing，簡稱 BOINC)。

如果各位有興趣，願意把你的個人電腦閒置的運算能力貢獻出來做更有意義的事情，可以到 BOINC 平台網站(http://boinc.berkeley.edu)逛逛，網站雖是英文的，但可以更改語言變成簡體中文，所以還是加減看得懂。網站裡面列出目前參與分散式計算的各種研究計畫，只要選取你感興趣的項目，然後把程式下載回來安裝執行就行了。程式會利用電腦執行的空檔進行分析，不會影響電腦正常運作，卻能真正把電腦運算能力完全發揮出來～反正平常電腦開著也是要用電，不是嗎？

二、 資料搜尋：

上面所提的科學計算似乎有點深奧，那就來點平易近人的好了。我們知道，資料搜尋是常見的電腦處理動作，它需要從一大堆資料數據裡逐一抽出資料片段，然後和我們設定好的條件作比對，如果有符合的部分就另外收集起來整理成結果。所以，整個搜尋過程其實就是大量的資料傳輸和比對動作，因此如何加快搜尋速度，也成為電腦科學家研究改進的重點。常用的方法是先將資料作排序，或者抽取其中的關鍵字做成索引檔(Index)，這樣就可以使速度大幅提昇。

除了各種加快搜尋速度的資料庫技術外，如果能應用平行處理的觀念，將資料分配給很多台電腦一起搜尋，這樣更能加快搜尋速度。舉例來說，大家第一次用 Google 搜尋資料時，大概會被那種驚人的搜尋速度震撼到，居然不到一秒鐘就找到成千上萬符合條件的網頁。這當然不是 Google 根據你要找的關鍵字到全世界網站一個一個比對出來的結果，而是事先定期掃描各網站，把網頁內容收集回來整理建檔來提供搜尋，所以就不用每次都要遍尋全世界網站。

就算可以省下重新瀏覽網路的時間，但由於所有網站內容的資料庫實在太

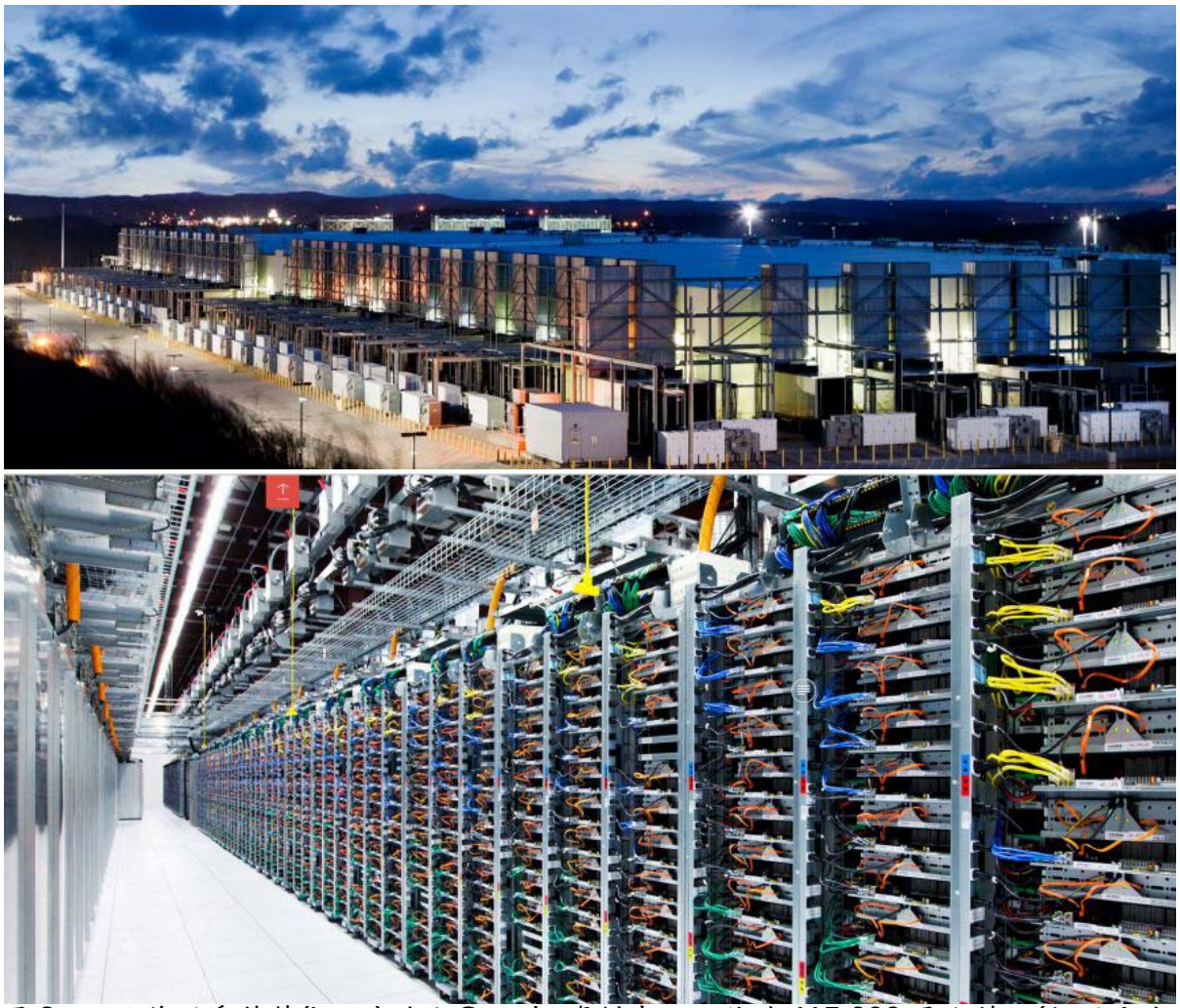

圖 3. 位於美國愛荷華(Iowa)州的 Google 資料中心，佔地 115,000 平方英呎(超過 1 公頃)，外觀像倉庫(上圖)，裡面堆滿成千上萬台伺服器(下圖)。Google 已宣布，未來還要在智利、香港、新加坡和台灣繼續增建這樣的資料中心。

龐大了，只靠一台電腦來搜尋也得花很多時間；更何況同時有那麼多人上來搜尋，而且每個人找的東西還都不一樣？因此 Google 採取了很多措施來加快搜尋速度。首先 Google 在世界各地建了 8 座資料儲存中心，每個資料中心都有成千上萬台的主機(如圖 3)。不論使用者是要搜尋、寄信(gmail)、看影片(youtube)，查地圖(google map)…，都可以先找最近的資料中心解決，以減少跨地區甚至跨國的網路傳輸流量。至於資料庫本身則拆散分布在那一大堆主機中，動用所有的主機一起幫忙找，每一台只搜尋其中一部分就行了，因此可以達到平行處理效果。此外，Google 還事先建立好一大堆(應該有上百萬種)常見的可能搜尋指標，如果使用者剛好用到這些關鍵字，就可以直接拿出來用。因此我們常常只打一個字，後面就出來一堆相關關鍵字給你選，而且越常見的熱門關鍵字會放越前面，選了就直接給答案，根本不用找。正因為有這麼多的電腦一起工作，並配合各種提昇搜尋效率的措施，才能將效率提升到如此驚人的地步。

舉 Google 作例子，並不是說大家都應該先蓋一棟那種資料中心才能提升效率，畢竟那是動輒上億的投資，可不是一般企業或政府機關花得起的，一般個人就更別提了。不過，我們還是可以參考 Google 所用的策略來改善資料搜尋效率，例如把資料依照業務性質分散放在不同課室的電腦就近查詢，就可以改善內部區域網路傳輸效率，避免全部集中在電腦機房主機造成雍塞現象。又如農委會有許多農業改良場和農林漁牧業試驗所，平時任務就是協助輔導轄區的農作物研究改良，當地農友如果有問題，也是就近到改良場或試驗所網站查詢資料，或找研究人員協助解決。因此若能適度把網站分散放在各機關，不但可以紓解網路流量，不會因為資源過度集中造成運作障礙等問題。萬一農委會主機故障或停電，更可避免因此造成全國農友都無法查詢資料的狀況。

三、 檔案分享

或許有人會說，Google 的規模太大，不是一般人玩得起的。沒關係，這邊再介紹一個更生活化的應用。假設現在你想把一個大型檔案分享給很多人，由於檔案太大無法透過 Email 寄送，傳統的方法是先架設 WWW 或 FTP 網站，把檔案放上去，再把網址和帳號密碼告訴對方等他來下載，稱之為主從式架構(Server-Client)。然而這樣你除了要會架設網站，更要面臨頻寬問題。因為一般我們家裡使用的網路都屬於上傳下載速度不一樣的 ADSL，雖然平常下載速度很快，但若要把資料分享給別人，就需要用到上傳的頻寬；如果要分享的對象很多，就會分散流量，使傳輸速度更慢(如圖 4 上)。不但大家都要花很多時間，你的電腦也要一直開著等所有的人都下載完畢才行。如果你的檔案很熱門許多人都想要，就會一直有人上來想要下載，那你的電腦就沒法關機休息了。

為解決這種需要，便有人發展出 P2P(peer-to-peer)資料傳輸技術，例如 eMule、BT 和 Foxy 等，都是利用這種原理進行分享檔案。它的原理是先把要分享的檔案切成一大堆片段，然後分別傳輸給所有想下載的人，片段下載完的人就再繼續下載另一個片段，但同時也把自己收到片段提供給其他人下載。因此，

每個人下載的同時也在幫忙分享給別人，這樣不但讓資料來源變多，速度也比只靠原始分享者提供頻寬給大家下載快很多，而且越多人下載，速度反而越快(如圖 4 下)。

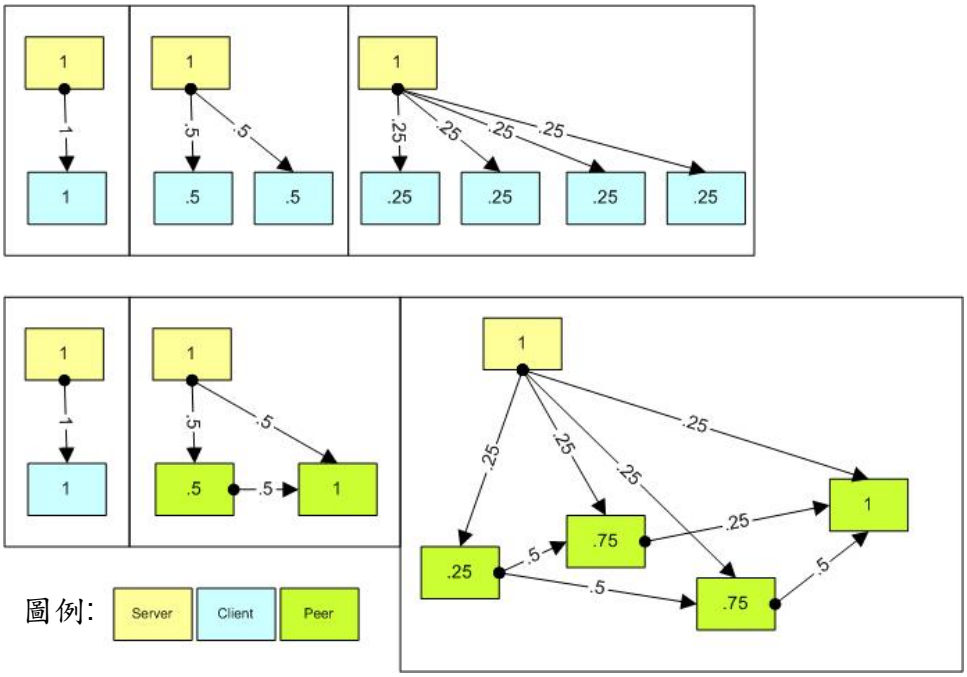

圖 4. 主從式(上)和 P2P(下)傳輸模式比較
主從式：只有 Server 提供資料，隨下載 Client 數量增加，流量也分散變慢。
P2P：除了 Server 外，所有下載者(Peer)也都協助分享資料，因此越多人下載，速度反而越快。

舉例來說，假設每個人的網路頻寬都是下載 10M/sec 上傳 1M/sec，現在我有一個 100MB 的檔案要分享給 10 個人，如果採用傳統主從式架構，因為我的上傳頻寬只有 1M/sec，平均分配給 10 人，每人下載速率只剩 100K/sec；如果改用 P2P 技術，先把檔案分割為一大堆小片段，一開始因為只有我這邊才有資料，因此 10 個人都用 100K/sec 的速度分別下載一個片段。等片段下載完畢後，每個人再從我這邊下載另一個片段，但同時也將自己取得的片段分享給其他人，因此每個人都可以從其他人那邊再下載另外 9 個片段，下載速度立刻提升 10 倍變成 1M/sec，而且下載的人越多，等於分享的人也越多，速度反而越快。如果又有人中途加入想要下載，由於前面 10 個人已經下載很多片段，這人就可以同時下載數十個片段，下載速度甚至可以達到 10M/sec 的全速。

P2P 技術另一個優點是，由於有其他人幫忙分享，因此最初的檔案提供者不用等到全部的人都下載完畢才能關機(那要等到什麼時候？)，只要等所有片段確定都已經被人下載，他就可以關機睡覺去。剩下的部分因為已經分散到所有使用者那邊，就算我這個原始檔案提供者不在，交流的動作還是可以繼續下去。

對檔案分享來說，P2P 技術確實是非常有效率的技術，可以善用所有人的

網路頻寬來提高下載速度，但缺點也在此。由於你下載的同時也在上傳片段給其他人，因此若沒有設定流量(預設是全部開放)，就可能把你的上傳速度完全佔掉，影響你作其他正常工作(例如寄信)的速度。此外，由於 P2P 技術是靠所有使用者提供資料片段來源和網路頻寬，因此除了目前正在下載中的檔案，還可以設定哪些資料夾的檔案是要分享的，但若不小心設定，就可能把你所有檔案、私密照片或影片也都流傳出去，新聞就曾報導過很多這類的案件。所以 P2P 軟體雖然好用，但使用時一定要小心，如果沒有把握，其實建議還是不要用比較安全。

四、 虛擬電腦

前期提過，現在的個人電腦雖然都擁有多核心處理器及大容量的記憶體和硬碟空間，但由於軟體不一定支援多核心運算，以致平常並沒有將效能完全發揮出來。這時，也可以將現有的處理器、記憶體和硬碟空間撥出一些，安裝成一台虛擬電腦(Virtual PC)。它就像剛買來的新電腦一樣，你可以在裡面安裝作業系統和應用軟體，然後進行分析運算，而不用再買一台電腦。例如 SAS 統計分析軟體只能使用單核心進行運算，不管電腦配備多少核心的處理器都沒用。然而，如果處理器核心夠多、記憶體夠大，我們就可以在電腦裡面設置很多台虛擬電腦，再分別各安裝一套 SAS，這樣就可以同時進行很多個統計分析工作，讓電腦的計算效能充分發揮。

除了讓軟體可以重複執行外，我們還可以在虛擬電腦裡面安裝完全不同的作業系統，而不用另外買一台電腦。這樣就有很多優點，例如有些舊版軟體沒法在新的 Windows 7 作業系統執行，只要另外安裝一台 Windows XP 的虛擬電腦，就可以把這些舊版軟體放在裡面繼續運作(如圖 5)。或者你對 Linux、FreeBSD 或 Ubuntu 等非 Windows 作業系統有興趣，可以用這種方式安裝在虛擬電腦裡來玩。或者你想嘗試某些可能會造成電腦損壞的動作，例如想教人如何格式化硬碟、安裝作業系統，或者下載到可能帶有病毒的檔案不敢打開，甚至想學電腦病毒設計...都可以在虛擬電腦裡面作測試，就算真的出問題導致虛擬電腦不能用了，頂多再把它重新安裝就行了，對真正的電腦完全沒有影響。

圖 5. 在 Windows 7 作業系統的電腦裡啟動 Windows XP 作業系統的虛擬電腦，連開機畫面都跟真正的電腦一模一樣。

結語

本期介紹幾項對一般人來說可以提昇電腦運算處理效率的事情，包括加入

分散式科學運算計畫，把你的電腦閒置不用的計算能力貢獻出來幫忙分析，讓你的多核心處理器的運算效能完全發揮；或者把需要處理的資料加以分割，放到不同電腦去計算或搜尋，以提升效率；以及利用 P2P 技術集合很多人的電腦網路頻寬來分享檔案，讓每個人的網路頻寬充分被利用，提高下載效率；另外也可以面安裝虛擬電腦，讓同一套軟體可以重複被執行，以充分發揮電腦效能等。這些應用其實都不是什麼新發現，早在「雲端」兩個字問世前就早已經被大家使用了。但這也是我一直強調的，只要能善用電腦和網路資源來提高處理效率，或者彈性運用整體設備的效能相互彌補，讓資源更有效地發揮出來，就可以算是雲端應用；至於裡面應用什麼技術，或者資料是在哪一台電腦處理，那就不用管了。以這種定義來看就可以發現，前面所提的這些行之有年的應用方式，共同的特性就是讓你的電腦處理效能真正發揮出來。因此，不管是加入尋找外星生命計畫，或者把資料分散在不同電腦做搜尋，或者用 P2P 軟體分享檔案，甚至安裝虛擬電腦來把一台電腦當成很多台來用，這些動作當然也可以算是雲端應用，不是嗎？

然而，對那些講雲端技術的專家學者來說，大概不會認為這些也算是雲端應用吧？他們會提出一堆專有名詞，然後說一定要用這樣那樣的技術才算是雲端處理。因此，下一期開始，我們就為大家介紹這些雲端處理模式背後的運作原理，以及如何將很多電腦設備的效能整合成雲端主機，敬請期待。