開發與建置交通大學異地備份儲存系統

楊詠仁　　王英鼎　　陳昌盛
國立交通大學 資訊技術服務中心
housten0219@mail.nctu.edu.tw、pcbug@mail.nctu.edu.tw、cschen@cc.nctu.edu.tw

摘要

本文旨在探討交通大學異地備份系統開發建置的經驗與測試評估分享。開發前需要確認三個重要關鍵問題：1.需要提供每月資料備份紀錄報表。2.實際網路環境及儲存設備是否符合備份傳輸需求。3.資料備份時如何同時寫入不同地區儲存設備。

開發過程中我們安排了兩個實驗，分別是網路測試以及儲存設備測試。網路測試部份，我們使用iperf[1]網路測試工具進行多次的測試，其中包含：跨校區單線(singal thread)以及多線(multi threads)單/雙向網路連線測試。儲存設備效率測試則是藉由測試兩地儲存設備同時寫入以及僅寫入台南校區儲存設備數據來進行觀察。網路測試時發現新竹與台南校區間的中華電信 VPN 線路有所異常，於是就提供相關數據給中華電信人員，並請中華電信協助調整網路設定，以達到符合當初購買的網路頻寬使用狀態。儲存設備測試發現兩地儲存設備同時寫入與僅寫入台南校區儲存設備數據效能是相差不遠的，平均寫入頻寬約 13Mbps，藉此證明本系統使用 ZFS[2]達到兩地同時寫入運作方式並不會因為兩地同時寫入，造成較差的效率。

關鍵詞：異地備份、國家高速網路中心、iperf、ZFS。

Abstract

This paper aimed to explore the issues on developing an automatic remote backup system for archiving the various important data archives of NCTU information systems. The three important concerns, before starting the development tasks, are as follows: 1. Ensure that the system (to be developed) should provide a monthly report of the backup activities conducted; 2. Ensure that the physical networking environments and the data storage equipments under considerations should meet the design requirements; 3. Ensure that the system should have the capability to write out the same data files onto two different data archives (i.e., one in the local site and the other in the remote site) simultaneously on conducting a remote backup.

Keywords: 異地備份、國家高速網路中心、iperf、ZFS。

1. 前言

當發生重大災難或是人為錯誤操作造成資料毀損或消失時，該如何將這些資料還原回來，這是非常重要的。為了達到資料復原，我們就必須要事先將重要資料多儲存幾份到不同地方，以備不時之需。備份的定義是將重要資料抄寫多份至不同的儲存媒體裝置上，異地備份則是多增加了一個明確要求，儲存媒體裝置需距離實體所在地 30 公里以上才符合異地備份的精神。

以往交通大學重要資料備份儲存於國家高速網路中心，考量到這些資料的保管性以及未來校內其他單位也需要異地備份服務的想法下，於是開始規劃進行異地備份系統建置以及備份程式軟體開發。圖 1 為交通大學異地備份系統整體架構圖，於台南校區以及新竹校區放置儲存設備(SAN)，作為異地備份儲存空間用，中間的網路透過中華電信 VPN 網路環境作為介接，網路總頻寬為 500Mb。異地備份系統為虛擬機器，並建置於新竹光復校區機房內，需備份的重要系統也是在新竹光復校區，將規劃備份機器使用 iSCSI[3]通訊協定將儲存空間掛載作為備份儲存空間用。

在異地備份系統開發過程中進行了網路以及儲存設備的效能評估測試，發現新竹與台南校區間網路是有所異常的，其測試方法與系統設計架構說明將於本論文後續章節詳細描述。後續安排內容如下：第二節為背景探討，將簡單介紹儲存設備類型，第三節為研究方法，將描述本系統開發前的評估事項，第四節為本系統架構說明，第五節為網路測試的數據分析，最後一節為本文之結論與未來方向。

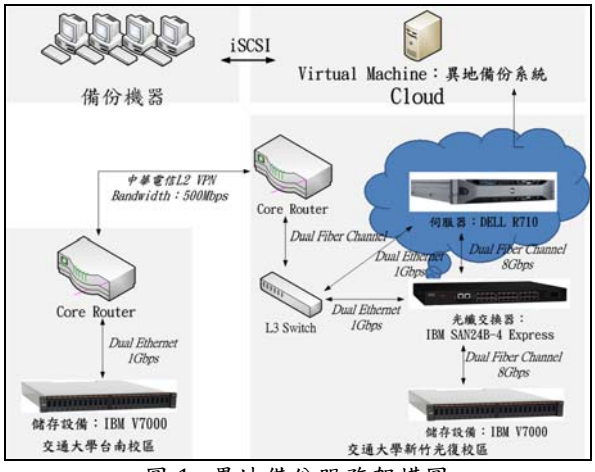

圖 1. 異地備份服務架構圖

2. 背景探討

2.1 設計考量

交通大學重要資料異地備份服務是與國家高速網路中心購買，思考到資料的保存以及隱密性問題，於是開始進行開發交通大學異地備份服務。現在市面上有許多提供免費備份服務，像是：GoogleDrive、Dropbox、SkyDrive…等，並無法符合備份需求，免費軟體因為必須要使用者登入執行狀況下才會進行備份動作，但重要系統並非 24 小時都在使用者登入狀況，所以免費備份軟體是無法運作的。商業等級的備份軟體(如：IBM Tivoli)設計為 Service 的模式運行，只要系統開機完成後，該服務就自動在背景執行備份程式。但因為大多數商業軟體僅能使用在一台機器上，並無法一套軟體多台機器使用，勢必需要額外的金錢進行購買。

為了確認開發異地備份服務是可行的，必須要先思考與確認以下項目：

1. 新竹與台南校區間的網路測試。
2. 資料儲存時該如何同時寫入兩地儲存空間。
3. 該選擇何種方式分享儲存空間。
4. 如何確保服務不中斷(備援機制)。
5. 該選擇何種儲存設備與磁碟。

上述 1~4 項目將於第 3 節詳細說明，以下小節將進行第五項的儲存設備介紹。

2.2 儲存設備介紹

儲存設備可以分成兩類：1. 區塊式儲存設備(Block Level Storage)。2. 檔案式儲存設備(File Level Storage)，以下將介紹這兩種類型種類儲存設備：

Direct Attached Storage(簡稱 DAS)，為區塊式儲存設備，主機需透過一張 Host Bus Adapter(HBA)主機介面卡與 Direct Attached Storage（簡稱 DAS）相連後(請參考圖 2)，即可取得一個或是多個磁碟裝置，接下來對這些磁碟裝置進行格式化動作即可以進行儲存空間操作。

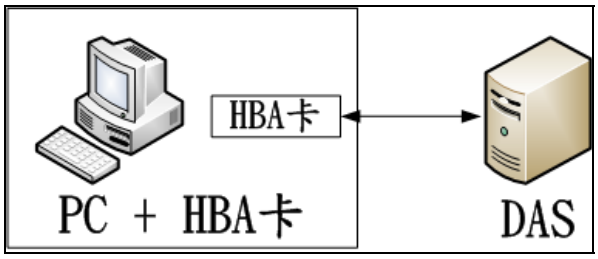

圖 2. DAS 架構圖

Storage Area Network[4]（簡稱 SAN），為區塊式儲存設備，其架構(請參考圖 3)需要搭配：光纖交換器(SAN Switch)、光纖介面卡(FC Host Bus Adapter)，SAN Storage。多數透過 Fibre Channel Protocols(FCP)或 Internet Small Computer System Interface(iSCSI)等通訊協定，提供儲存區塊空間。

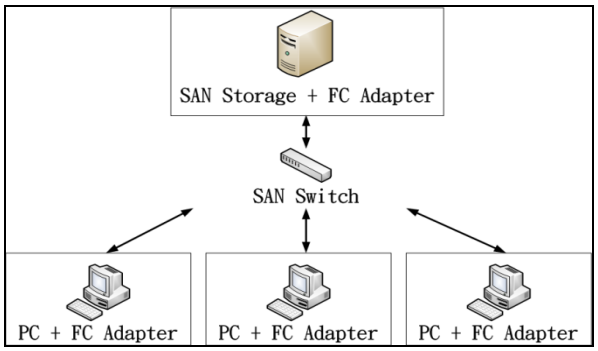

圖 3. SAN 架構圖

Network Attached Storage [5]（簡稱 NAS），為檔案式儲存設備，設備本身具備作業系統以及軟體，該軟體負責資料處理以及管理功能。UNIX 及 UNIX-like 系統使用 Network File System(NFS)或是 Samba 通訊協定，而 Windows 系統透過 Server Message Block(SMB)/Common Internet File System(CIFS)通訊協定。NAS 將儲存空間分享前，會透過本身的檔案系統進行處理，所以其他系統取得此儲存空間後，不需要對此儲存空間進行格式化動作，直接操作即可。系統上查看到的是掛載了一個資料夾而非一個磁碟裝置(其架構請參考圖 4)。

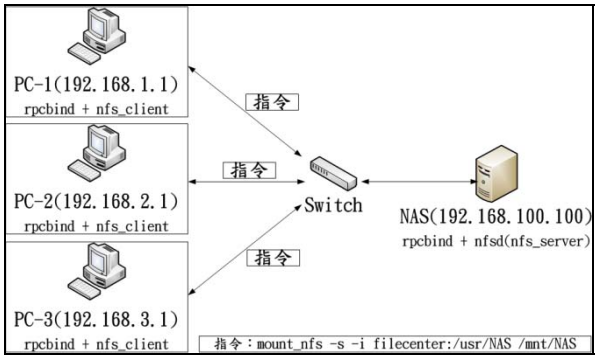

圖 4. NAS 架構圖

2.3 ZFS 檔案系統介紹

Zettabyte File System 簡稱 ZFS，由 Sun Microsystem 為 Solaris 作業系統開發的文件系統。其作業系統可以將多個磁碟裝置群組起來成為一個儲存池(zpool)，此儲存池的組成可以是 Raid 0 或是其他 Raid 模式組成，而這個儲存池內又可以進行切割小的空間分享給其他系統掛載使用，相當方便使用。

3. 研究方法

第二章節背景探討有提到五點的先期評估與測試，以下小節將進行說明。

3.1 新竹與台南校區網路測試

將於新竹及台南校區各別架設一台機器，並使用 iperf 網路測試工具進行網路測試，iperf 是一套免費的網路測試軟體，適用於 UNIX、Linux 與 Windows 作業系統環境。其軟體的運作模式為：主從式架構，由 Client 負責送流量給 Server，可以建立 TCP 和 UDP 的方式進行網路測試，也可以進行單向或是雙向的網路測試。

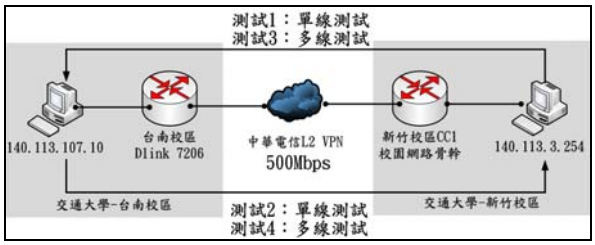

圖 5. 網路架構圖

其網路架構請參考圖 5，iperf 參數指令請參考圖 6，新竹與台南校區中間的線路使用中華電信 L2 VPN 網路環境進行對接，中間的頻寬為 500Mb，因為台南校區需要與其他學校介接，所以將 200Mb 的頻寬挪為使用，真正可用頻寬為 300Mb。其測試方式分成四項(請參考表 1)，每次測試 10 秒，相關指令參數請參考圖 6。

表 1. iperf 網路測試

No.	測試類型	client	server
1	單線	140.113.3.254	140.113.107.10
2	單線	140.113.107.10	140.113.3.254
3	多線	140.113.3.254	140.113.107.10
4	多線	140.113.107.10	140.113.3.254

Server	Command
	iperf.exe -s -w 1M -M 1540
Client(單線)	iperf.exe -c 140.113.107.10 -i 1 -P 1 -t 10 -l 1518
Client(多線)	iperf.exe -c 140.113.107.10 -i 1 -P 10 -t 10 -l 1518

-w TCP window size
-M 設定 TCP maximum segment size
-i 每隔 1 秒將數據顯示出來
-P 設定多少 parallel client threads
-t 監視測量數據時間為 120 秒
-l 設定 buffer(read/write)長度

圖 6. iperf 參數指令

3.2 儲存資料同時寫入兩端儲存空間

我們基於異地備份原則下，於交通大學新竹及台南校區各建置一台儲存設備。首先異地備份系統透過 iSCSI initiator 軟體將新竹及台南校區儲存設備分享出來的 Volume 掛載起來，這時系統端可以取得二個儲存裝置，接下來透過 ZFS 檔案系統提供之工具 zpool 將這兩個儲存裝置做成 Raid 1 模式的儲存池，之後提供備份機器使用的 Volume 再透過 ZFS 檔案系統提供之工具 zfs 建立成功後透過 istgt 軟體即可以分享儲存空間使用，之後寫入的儲存資料即可以達成新竹及台南端儲存空間保存(請參考圖 7)。

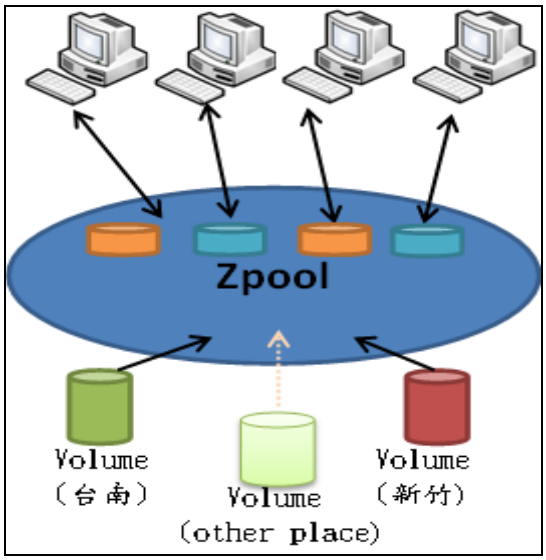

圖 7. 儲存空間組成示意圖

3.3 使用 iSCSI 方式分享備份儲存空間

分享儲存空間的方式有很多種，像是第二章節所提到的 NFS、Samba 以及 iSCSI 方式。NFS 以及 Samba 的概念是使用者透過網路取得儲存空間後就可以進行掛載，就像是在檔案系統上建立了一個額外的資料夾，不需要帳號密碼就可以使用，而且還可以提供檔案共享的使用，此共享空間可以同時多人掛載使用。

使用者透過 iSCSI 方式取得儲存空間時，作業系統會視為一個實體硬碟，這時就需要對此硬碟進行格式化等動作。iSCSI 分享的儲存空間只允許線上 1 個人掛載使用。

此次交通大學異地備份案例，備份的資料大多數屬於個人資料，所以必須要考慮資料隱密性的問題。如選擇 NFS 的方式的話， Server 端可以直接接觸到使用者上傳的資料。此次服務也不需要檔案共享的使用，所以選擇使用 iSCSI 的方式。

3.4 儲存設備與硬碟的選擇

此次服務是以備份為主，將提供大容量儲存空間為主要需求，所以採用 3TB 7,200 轉，頻寬 6Gbps 雙通道之 3.5 吋近線 SAS(NearLine-SAS)介面之磁碟機共 24 顆(新竹及台南校區各使用 12 顆)。

異地備份系統組成為：2 台儲存設備(SAN Storage)，1 台伺服器，1 台光纖交換器。為了增加儲存可靠性，建立 Raid 6 磁碟陣列。伺服器與光纖交換器則建置於新竹校區。

3.5 備援機制

為了確保異地備份服務可靠性，儲存設備、伺服器、光纖交換器與路由器連接上各建立兩條連線，避免因介面卡問題，造成服務中斷的情形。

其中一端儲存設備如有異常(網路中斷、系統故障…等因素)，因為 zpool 組成為 Raid 1 架構模式，可以透過 zfs 提供之功能將台南端掛載的儲存空間先進行 offline，備份的儲存服務仍可以持續進行，而台南端的儲存設備也可以同時進行維護或是

修復等動作，待設備恢復正常後，即可以重新將儲存空間進行 online，接下來 zfs 會將變動的部分更新至重新掛載的儲存空間，同步完畢後，資料的儲存將恢復成同時兩地的寫入及讀取，在同步運行階段，當有資料讀取寫入時，服務並不會中斷，須同步的儲存空間則繼續進行同步，原本 online 的儲存空間則繼續提供讀取寫入的服務。

4. 系統實作

4.1 異地備份儲存系統組成

異地備份儲存系統建置於虛擬機器，備份服務的組成分為三項：作業系統、網頁以及使用者端備份程式。作業系統使用 FreeBSD，統透過 iscontrol 軟體將新竹與台南端的 SAN Storage 磁碟陣列掛載起來，並透過 ZFS 檔案系統以及 istgt 軟體分享儲存空間給備份機器使用。使用 Apache 作為網頁伺服器，並使用 PHP+MySQL 開發應用網頁。使用者端程式安裝以及備份程式使用 Python 語言開發 (請參考表 2)。

表 2. 異地備份系統資訊

Virtualization Software	VMware ESXi 4.10
CPU	2 core
Memory	32GB
Guest OS	FreeBSD 9.1(amd64)
Program Language	PHP5 Python 2.7.5
Software	Apache PHP5 phpMyAdmin MYSQL istgt ZFS gmultipath iscontrol

網頁角色的權限設定分為四個類型，分別是：SuperAdmin、Admin、User 以及 BackupNode，以下將介紹這四個角色的功能說明。

- SuperAdmin：系統最高權限管理者，主要負責管理群組、設定群組可使用的空間以及管理群組管理者帳號(Admin)。
- Admin：群組管理者，系統第二高權限的管理者，主要負責管理該群組(管理者/一般使用者)帳號、設定使用者可使用的空間。
- User：一般使用者，主要負責設定備份機器使用空間、管理 BakupNode 排程設定。
- BackupNode：備份機器，透過 BackupNode 系統資訊來取得儲存空間。

表 3. 異地備份系統角色資訊

角色	功能
SuperAdmin	建立/刪除/管理 • Group 儲存空間 • GroupAdmin • Account 儲存空間 • Account 建立/刪除 • BackupNode 排程設定
Admin	建立/刪除 • Account 儲存空間 • Account 建立/刪除 • BackupNode 排程設定
User	建立/刪除 • BackupNode 儲存空間 建立/刪除 • BackupNode 排程設定
BackupNode	取得 Volume 抓取排程 執行操作指令(備份、還原、刪除備份資料)。

儲存空間使用 ZFS 提供工具 zpool 將台南與新竹端儲存磁碟建立一個 Raid 1 儲存池(簡稱 zpool)，共分為四部份，分別是：RootPool、GroupPool、UserPool 以及 BackupNodeVolume。BackupNodeVolume 類行為 volume，其他三類都是檔案系統。真正分享給備份機器使用的是 BackupNodeVolume，透過 zfs 工具可以設定每個 Pool 可以使用多少容量，這樣就可以很明確地給定不同使用者的容量需求設定(請參考圖 8)。

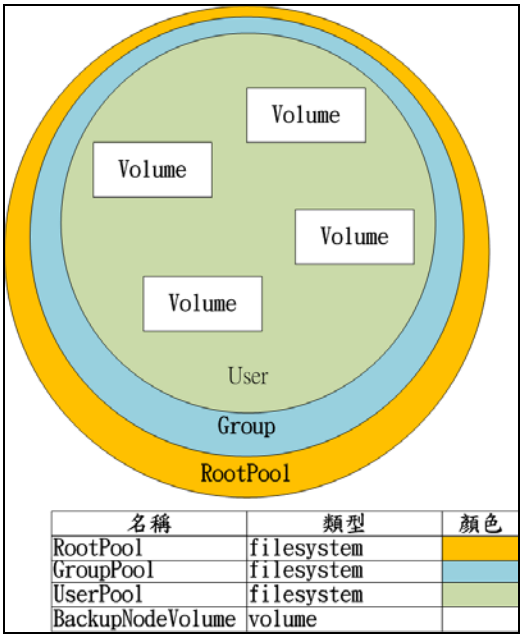

圖 8. 儲存空間結構圖

4.2 系統運作流程：

系統運作可以分為三個部分：基礎環境建置(請參考圖 9)、軟體安裝執行(請參考圖 10)以及排程設定(請參考圖 11)。基礎環境設定分為以下三個階段：

第 1 階段：SuperAdmin 建立初始化 Group，包含：GroupPool 以及 GroupAdmin 建立。SuperAdmin 新增群組資訊，群組資訊包含：群組名稱以及設定此群組可以使用的空間設定，後端程式判斷正確後，會透過 zfs 指令將 GroupPool 建立出來。SuperAdmin 新增帳號並設定此帳號權限為此群組管理者並設定此帳號可以使用的空間設定，後端程

式判斷正確後，會透過 zfs 指令將 UserPool 建立出來。通知此群組管理者，服務已經開通。

第 2 階段:Admin 建立群組使用者。群組 Admin 可以新增新的使用者，其權限可以設定為一般使用者(User)或是群組管理者(Admin)並設定此帳號可以使用的空間設定，後端程式判斷正確後，會透過 zfs 指令將 UserPool 建立出來。

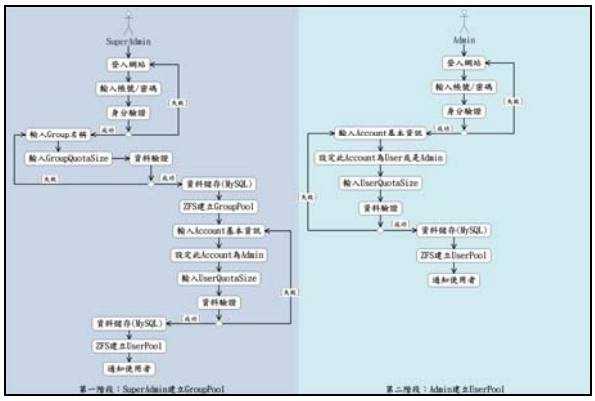

圖 9. 基礎環境設定活動圖

軟體安裝分為以下 4 個階段：

第 1 階段：身分確認並建立備份機器資訊。執行安裝程式會需要輸入使用者帳號及密碼，接下來會將此資料回傳至 Server 進行身分確認，確認完成後，會將此備份機器系統資訊回傳 Server 記錄。

第 2 階段：安裝 iscsi initiator 程式。因本備份服務是透過 iscsi 方式分享儲存空間，使用者端需要 iscsi initiator 程式，若檢察系統上無此軟體，將會協助下載並進行安裝。

第 3 階段：設定 volume 空間大小。安裝程式會詢問使用者此備份機器可以使用多少的儲存空間，使用者輸入完畢後，程式會將此資訊回傳 Server，確認無誤後，後端程式會透過 zfs 將 volume 建立，並將此備份機器允許連線資訊新增到 istgt 設定檔內，並完成連線設定。

第 4 階段：格式化儲存空間。安裝程式會透過 iscsi initiator 將儲存空間掛載起來，透過 Diskpart.exe 程式將此儲存空間進行格式化動作，格式化完成後將儲存空間卸載，並啟動備份服務。

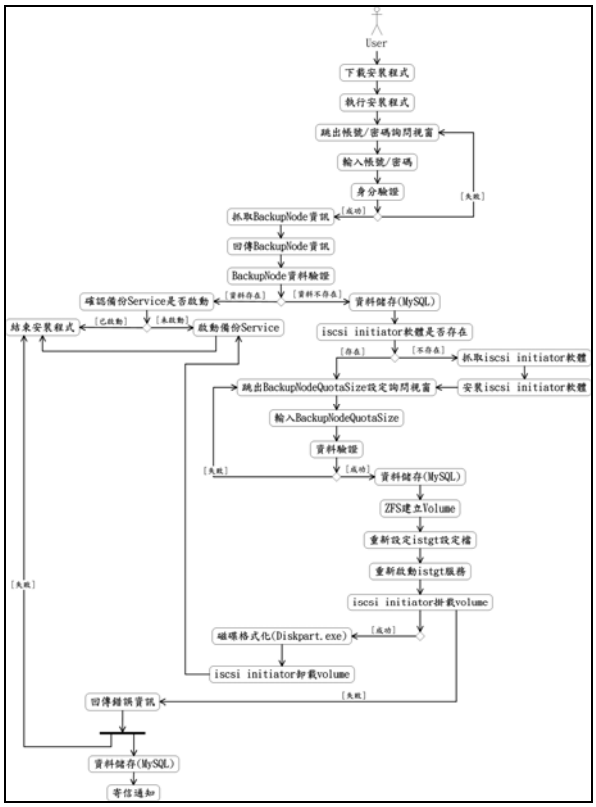

圖 10. 軟體安裝設定活動圖

排程設定部分，較為簡單，使用者透過網頁點選須建立排程的機器，就可以進行排程設定，網頁設定完成排程後，備份程式就會在預定的時間點進行備份，備份時將會使用 7-zip 程式將所需要備份的檔案進行壓縮加密，以確保資料安全性。

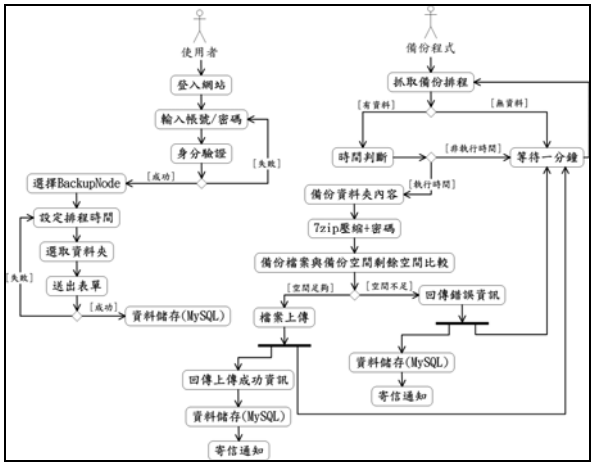

圖 11. 備份排程設定活動圖

5. 實驗結果與數據分析

5.1 網路測試數據分析

2012 年 8 月份開始陸續透過 iperf 進行多次的網路測試(測試數據請參考表 4)。在 2012-08-13 測試時發現，從新竹校區送流量至台南校區的數據都比較低，所以就請中華電信協助確認中間線路設定，於 2012-08-21 再次進行網路測試，不論是新竹往台南送流量，或是反向回來，數據都比較接近

了，但新竹往台南多線(multi thread)數據仍是不佳的，所以再次跟中華電信反映，在 2012-08-27 再次進行一次測試，這次數據平均為每秒 50Mbps，接下來幾天的測試數據跟 08-27 測試的數據差異不大。根據 2012-08-27 多線的數據可以證實，新竹與台南校區間的總頻寬是可以使用到 500Mbps，但中華電信有額外設定每一個連線使用的頻寬上限為 50Mbps 左右。

表 4. iperf 網路測試數據資料

Date	Type	Bandwidth (Mbits/sec)	
		單線	多線(平均)
08-13	新竹->台南	8.99	13.9
	台南->新竹	65.6	18.1
08-21	新竹->台南	32.3	16.5
	台南->新竹	49.7	50.2
08-27	新竹->台南	52.1	50.8
	台南->新竹	52.3	49.8

5.2 預估與實際上傳數據比較分析

既然知道台南校區與新竹校區中間 VPN 網路傳輸頻寬，接下來就可以透過數據推算檔案上傳完成的時間，需要預估時間有三個階段(流程示意圖請參考圖 12)，將三個階段時間加總就可以預估上傳結束時間。

1. BackupNode 將資料上傳至異地備份系統(網路傳輸時間)：資料大小(單位需轉換為 MB)*8/校內網路傳輸頻寬(100Mbps)。
2. 異地備份系統資料傳送至儲存設備(網路傳輸時間)：資料大小(單位需轉換為 MB)*8/網路頻寬，如僅寫入新竹校區儲存設備，網路頻寬為 100Mbps，反之，僅寫入台南校區儲存設備，或是兩地儲存設備網路頻寬為 50Mbps(中華電信 VPN 線路頻寬)。
3. 儲存設備將資料寫入硬碟空間(Disk I/O)：
   其演算法為：資料大小(單位需轉換為 MB)/磁碟寫入頻寬，如僅寫入新竹校區儲存設備，磁碟寫入頻寬為 11MBps，反之，僅寫入台南校區儲存設備，或是兩地儲存設備磁碟寫入頻寬為 3MBps(請參考圖 13)。

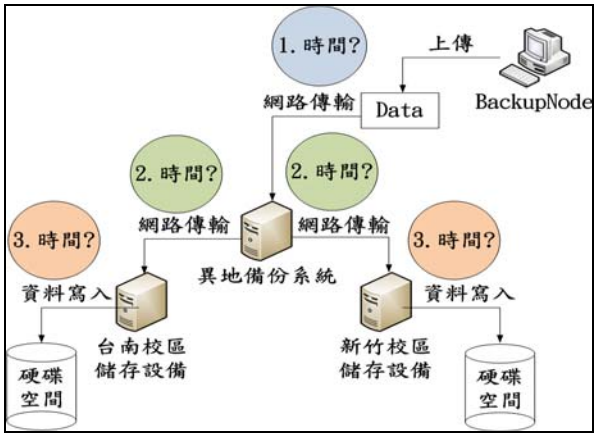

圖 12. 檔案上傳流程示意圖

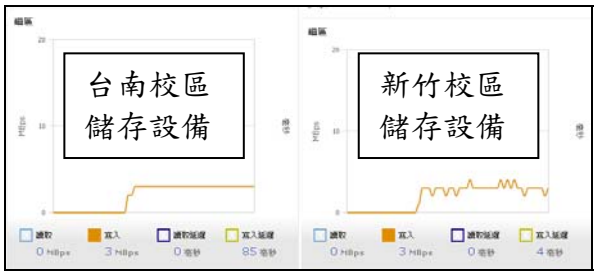

圖 13. 儲存設備寫入效能監控圖

以檔案大小為 3.4GB 為例，推估兩地同時寫入完成時間約： 3.4*1024*8/100 + 3.4*1024*8/50 + 3.4*1024/3 = 1997 秒，與實際備份上傳紀錄時間相差約 3%(請參考表 5)。

接下來將繼續進行額外的測試，測試僅寫入台南端儲存設備，與兩地同時寫入花費時間來進行比較，從表 4 可以發現，兩地同時寫入所花費的時間比僅寫入台南端儲存設備多快 40 秒左右的時間，所以，可以證實透過 ZFS 方式同時寫入兩端儲存設備的架構是可行的。

表 5. 檔案備份上傳數據資料

Type	FileSize	Time(sec)		Bandwidth (Mbits/sec)	
		實際	預估	實際	預估
兩端寫入	3.4GB	2133	1997	13.1	13.9
	682MB	398	391	13.7	13.9
新竹寫入	3.4GB	885	874	31.4	31.9
	682MB	183	172	29.8	31.7
台南寫入	3.4GB	2093	1997	13.3	13.9
	682MB	398	391	13.7	13.9

6. 結論

現階段系統已完成百分之 80，基本的設定儲存空間、儲存空間掛載、檔案上傳備份、資料還原以及刪除檔案基本功能都已經可以正常運作，接下來將會進行大檔案傳輸，以及增加備份機器數量進行同時上傳等測試。

本系統將會持續性修正與改進，朝簡單方便可擴充性的系統架構，提供更快速檔案上傳功能以及更安全的儲存環境方向進行。

參考文獻

[1] Iperf,
http://iperf.fr/
Retrieved Aug 1 2013

[2] ZFS,
http://en.wikipedia.org/wiki/ZFS
Retrieved Aug 1 2013

[3] iSCSI RFC,
http://tools.ietf.org/html/rfc3721
Retrieved Aug 1 2013

[4] Storage Area Network,
http://en.wikipedia.org/wiki/Storage_area_network
Retrieved Aug 1 2013

[5] Nas,
http://en.wikipedia.org/wiki/Nas
Retrieved Aug 1 2013