東海大學

資訊工程研究所

碩士論文

指導教授: 楊朝棟博士

基於 VMware 虛擬儲存區域網路的分散式共用儲存效能評估

Performance Evaluation of a Distributed Shared Storage Base on VMware Virtual SAN

研究生: 陳勇綸

中華民國一〇五年六月

東海大學碩士學位論文考試審定書

東海大學資訊工程學系 研究所

研究生 陳 勇 綸 所提之論文

基於 VMware 虛擬儲存區域網路的分散式共用儲存效能評估

經本委員會審查，符合碩士學位論文標準。

學位考試委員會

召 集 人 _____ 簽章

委 員 _____

_____

_____

_____

指 導 教 授 _____ 簽章

中華民國 105 年 6 月 27 日

摘　要

近年來隨著伺服器虛擬化的蓬勃發展，延伸出許多應用議題，如：PaaS、IaaS、SaaS 等雲端服務概念；但無論是那種虛擬化技術，在後端的儲存 I/O 效能及高可用性都是服務品質的重要關鍵因素。軟體定義資料中心的概念也隨著虛擬化技術的演進而提出了新的看法；愈來愈多的軟體定義原則被提出，例如由儲存虛擬化進而延伸出的軟體定義儲存（Software Defined Storage）。在發生軟硬體錯誤時讓系統能繼續維持服務可用，並將有問題的執行個體重新導向正常運作的執行個體，使系統服務不停頓並能夠穩定持續地提供資訊服務。在愈來愈多的核心資訊服務集中化管理後，如何能夠有效率的提升資料儲存在後端的處理讀取及寫入效能，這將會是每一種虛擬化技術或是雲端架構最為重要的議題。本論文將探討 Virtual SAN 分散式儲存系統，在不同原則下的佈署差異，並了解在不同的擴充方法：「垂直擴充（Scale Up）」或「水平擴充（Scale Out）」下，資料節點的成長對於儲存資料讀寫效能的差異性，以供真實的生產環境規劃參考。

關鍵字：Virtual SAN, 軟體定義儲存, 高可用性, 儲存虛擬化

Abstract

Server virtualization has been very prosperous in recent years. It extends many application topics, such as PssS, IaaS and SaaS. But whether what kind of virtualization technology are, the important key factor in the quality of services are storage I / O performance at the rear end and high availability. There are some new concepts of software-defined data center along with the evolution of application virtualization technology has been propose: more and more software-defined principles were suggested. For example, Software Defined Storage from storage virtualization. While an error occurred in hardware and software , it can continue to maintain service available, and redirect problematic instances to the normal one. Without stopping the system service and continue to provide stable information services. It would be the most important issue for each virtualization technology or cloud infrastructure: How can we enhance the efficient of data stored in back-end processing read and write performance after more and more centralized management of core information services.This paper discuss the difference of Virtual SAN distributed storage system, deployed at different principle. Learn about the effect of performance in data reading and writing while data storage node growth on different methods of expansion: “vertical expansion (Scale Up)” or “horizontal expansion (Scale Out)”. It will be valuable for real production environment planning reference.

Keyword：Virtual SAN, Software-Defined Storage, High availability, Storage Virtualization

致謝詞

完成碩士學業，一直是過去以往的夢想。過去就讀技職體系出生的我，能夠再次攻讀碩士學位除了要感謝家人的支持外，我要非常的感謝我的指導老師楊朝棟教授！他除了在學習的過程中不斷的引導我學習的方向，更對我的本篇論文研究不遺餘力的提供一個良好的研究環境，讓我可以在做學術研究的同時也能直接應用研究的成果到職場的工作領域上，充實專業知識的同時也能減輕了工作上的學習壓力，這些研究經驗對未來的我來說可謂彌足珍貴。

在求學的過程中，一開始總是朦朦朧朧的對論文的方向毫無準備，加上現在也不是當年專門的學生，求學的同時還有工作崗位上的事務需要處理，回到家中也有一個小家庭要照顧，因此能夠花在學習及做研究上的時間可說是相當有限；雖然一開始就定好目標，希望自己能夠在二年內畢業，但其實內心是相當的沒有底氣。還好在求學的過程中，除了指導教授楊老師的鼓勵及指導外，還有 HPC 實驗室的學長們可以請教！在這裡真的要特別感謝實驗室的廣欽學長，他在我實驗遇到瓶頸時給了我許多寶貴的建議，也因此我才能順利的完成這些複雜的研究實驗。

這些學習回憶，還有跟同學們的一起學習過程，都讓我深深的感受到當初選擇來東海念資訊工程碩士在職專班並選擇楊教授做為指導老師，真的是在正確不過的選擇了！最後要感謝的，當然就是在背後默默支持我的老婆亞儒，還有我最可愛貼心的大女兒麗安及小兒子羿通，有他們的支持才能讓我放心的順利完成這個人生的夢想和目標。感謝所有人！

Table of Contents

List of Figures

List of Tables

Chapter 1

簡介

1.1 研究動機

隨著現代 IT 技術的發展，往往是一個持續以新的技術、新的概念，去滿足既有需求的不斷更動取代過程，利用新的技術與架構，實現過去舊有的服務無法提供的低成本效益及簡易性、方便性的提升。而儲存設備繼 10 年前磁碟替換磁帶的基礎架構更新潮之後，現在愈來愈多以針對關鍵應用、備份、歸檔為主的儲存設備，更是因應效能的需求而加入了固態硬碟 SSD 做為資料 I/O 交換的提升主要設備。

以傳統 IT 基礎架構而言，伺服器系統和儲存系統的角色涇渭分明，分別負責應用程式執行提供使用者服務和儲存後端資料的功能，然而，隨著資料中心整體應用的運算架構全面走向通用化 x86 平臺，這兩種系統可搭配的硬體規格，差異已經越來越小，任何一臺伺服器，只要安裝了儲存系統軟體，幾乎就跟儲存設備沒有什麼兩樣。再加上近年來吹起的超融合（Hyper Converged）與軟體定義（Software Defined）的應用風潮，這些陸續運用了伺服器虛擬化技術、多節點的叢集運算與容錯服務架構，以及軟體化自定義化的儲存系統，動態調度資料中心運算資源與儲存資源的自動化工作，而這樣的發展再度模糊了伺服器與儲存之間的界線。

因此，儲存系統不只是走向軟體化、虛擬化、叢集化，以及支援由軟體來定義的動態調配資源作法，同時還要能支援超大規模儲存環境（Hyper Scale Storage）的建置能力，使用規模與儲存容量不只是能做到橫向擴充（Scale-out），甚至需達到無限擴充的程度。

並且儲存系統也不能只是單純的儲存設備，也要因應企業雲端虛擬化架構的結合，配合整體的資訊虛擬化服務以提供高可用性、軟體定義的自動化管理服務。減少企業的資訊管理複雜度、降低企業整體持有成本（Total Cost of Ownership，TCO）。

1.2 論文目標與貢獻

本論文主要在探討 VMWARE Virtual SAN 虛擬化儲存服務的架構，如何應用 Virtual SAN 結合關鍵底層虛擬化，並有效提升關鍵核心服務資料的存取效能。並研究在不同的網路架構及不同的軟體定義原則下，虛擬機器資料吞吐的差異性。並應用 Virtual SAN 在實務生產環境中，如何提升服務的高可用性，在不同原則下的佈署差異，並了解在不同的擴充方法：「垂直擴充（Scale Up）」或「水平擴充（Scale Out）」下，資料節點的成長對於儲存資料讀寫效能的差異性，以供真實的生產環境規劃參考。

1.3 論文架構

第二章將說明本論文的研究背景，介紹目前軟體定義資料中心及 Virtual SAN 相關技術，並探討現有的高可用性技術及可能面對的問題。第三章則描述系統的架構、建置實作。第四章說明實驗的環境、方法以及實驗的結果，並依照實驗的結果討論在不同的網路架構及不同的軟體定義原則下的效能異差。在最後，第五章將說明本論文依照實驗結果做出的結論及未來可能的研究方向。

Chapter 2

研究背景與相關研究

2.1 研究背景

2.1.1 Software-Defined Data Center，SDDC

所謂軟體定義資料中心的概念，是藉由虛擬化及相關可程式化條件控制技術，將所有的硬體資源集結成為一個資源池（Resource Pool），放棄或是不經由人工的直接手動控制，改而透過可程式化的軟體原則來掌控。當 IT 人員在運用與控制硬體及系統資源時，不再細部介入伺服器要如何操控、安全該如何防護、網路又要如何串連、空間該如何切割，讓資源都美好且和諧地運作。

而在傳統的資料中心架構中，所具有的只是一系列伺服器、儲存、網路、安全等個體的集合，資訊人員在對作業環境做資源分配時，需要直接且清楚的描述應用作業系統對於基礎設施技術的一系列要求；事實上每一個作業環境都是由中央處理器（CPU）、作業系統（Operation System）、儲存資源池（Storage Pool）、網路（Networking）、安全（Security）及管理系統（Management System）在縱向上的連結集合體。對管理人員而言建置資料中心的過程，資訊人員必須先選擇硬體平台、資料庫、中介軟體及管理軟體，甚至必須熟悉相關的網路設備技術、系統安全的分析等，再從中選擇合適的建置方案後才能開始執行建置，這通常需要花費數週甚至是數個月才能部署一個新的作業環境。

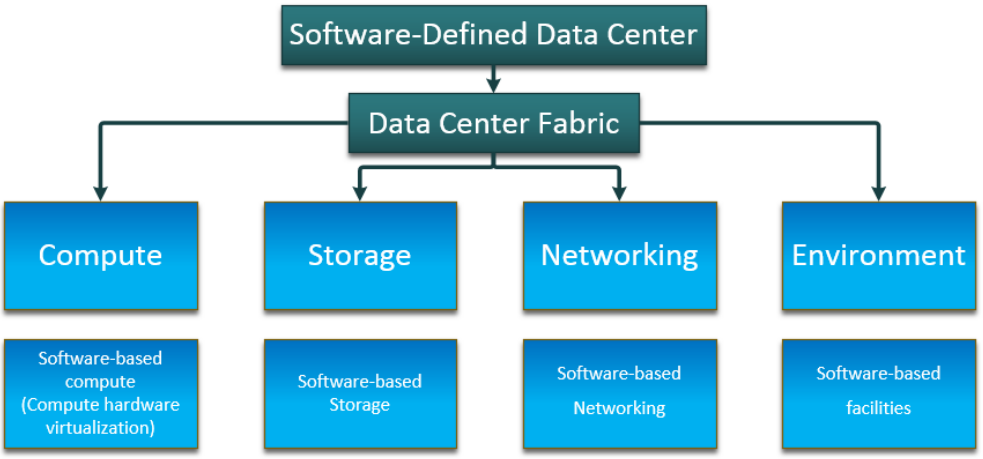

FIGURE 2.1: Software-Defined Data Center，SDDC 架構圖

VMware 在 VMworld 2012 年度大會時提出軟體定義資料中心（Software-Defined Data Center，SDDC）的願景，在 VMware SDDC 的願景裡面，資料中心所建構的私有雲及公有雲基礎便包含了運算能力 Compute)、儲存 (Storage)、網路架構 (Networking)…等都將被虛擬化，如 Figure 2.1。而所有運作元件皆完全虛擬化的資料中心，將使 IT 管理團隊更加的靈活，不但降低操作的複雜度及營運成本，同時還可以提高資源的可用性及靈活度，可以大幅的縮短及降低企業或組織的服務上線流程及時間。

而在虛擬的數據中心擁有很多的虛擬化 IT 設備，例如 IT 網路技術員可於主機 Hypervisor 之中安裝虛擬網路設備，所以虛擬機器可先通往虛擬網路設備，然後才連接真實網路世界。虛擬網路設備與實體一樣都會有交換機、防火牆、路由器和負載平衡器等等，如果虛擬機器不需要連接外部網絡，只是與內部虛擬機器溝通，但又要有防火牆設備阻隔一些流量，所以此時選擇虛擬防火牆是最好不過的，因為是於此時網路流量僅在主機虛擬網路之中流通，亦能夠解決外部或實體網路延遲的問題。

軟體定義資料中心 SDDC 帶給我們很多方便，但是有沒有想過什麼方法去管理如此龐大的虛擬數據中心架構？所以如 Figure 2.2 中，在軟體定義資料中心 SDDC 之中亦包含了自動化管理程序，將所有的虛擬設備由自動化中央平台去管轄。

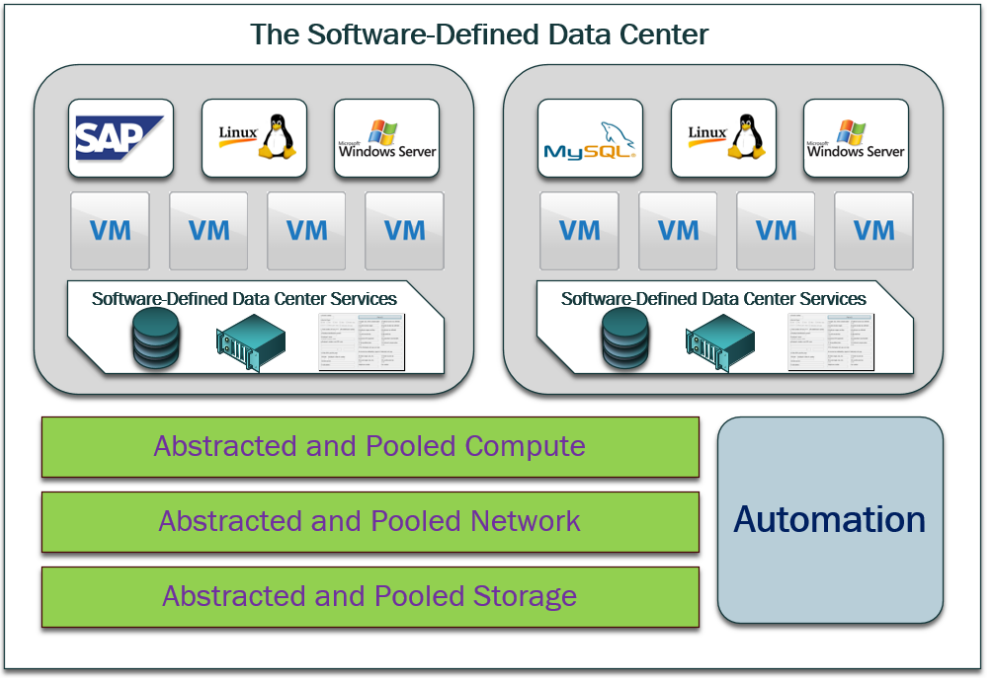

FIGURE 2.2: Software-Defined Data Center 必備的基本框架

2.1.2 Software-Defined Storage，SDS

在當前的資訊化架構中，儲存資源一直是每個資訊服務最重要的一環，除了要考慮容量、效能及高可用性外，儲存資源的有效分配運用，在當前的資料中心環境當中，更受到管理員極大的重視，因為不論是為了開拓創新的資訊業務型態所引發的新型系統建置需求，或導入發展已久、已經很成熟的各式 IT 服務應用，都需要儲存空間來保管資料或作為分析之用，而且受到近年來巨量資料的資訊需求因素，所需配置的容量更是越來越大；再加上，許多儲存設備受限於所使用的技術，往往只能一臺臺去建置，長期下來，許多公司會面臨儲存資源無法整合的難關，明明每一臺儲存設備的性能和容量都還有很大的可用空間，卻如同一座座孤島般只能各自為政，無法得到妥善的有效利用。

而且，為了確保系統的服務品質，很可能每一臺儲存設備只能讓特定的應用程式來存取，且如果企業資訊環境當中，建置了不同品牌的儲存設備，它們之間也不一定能夠整合在一起管理，所以想要做到儲存資源的集中管控，達成的困難度很高。

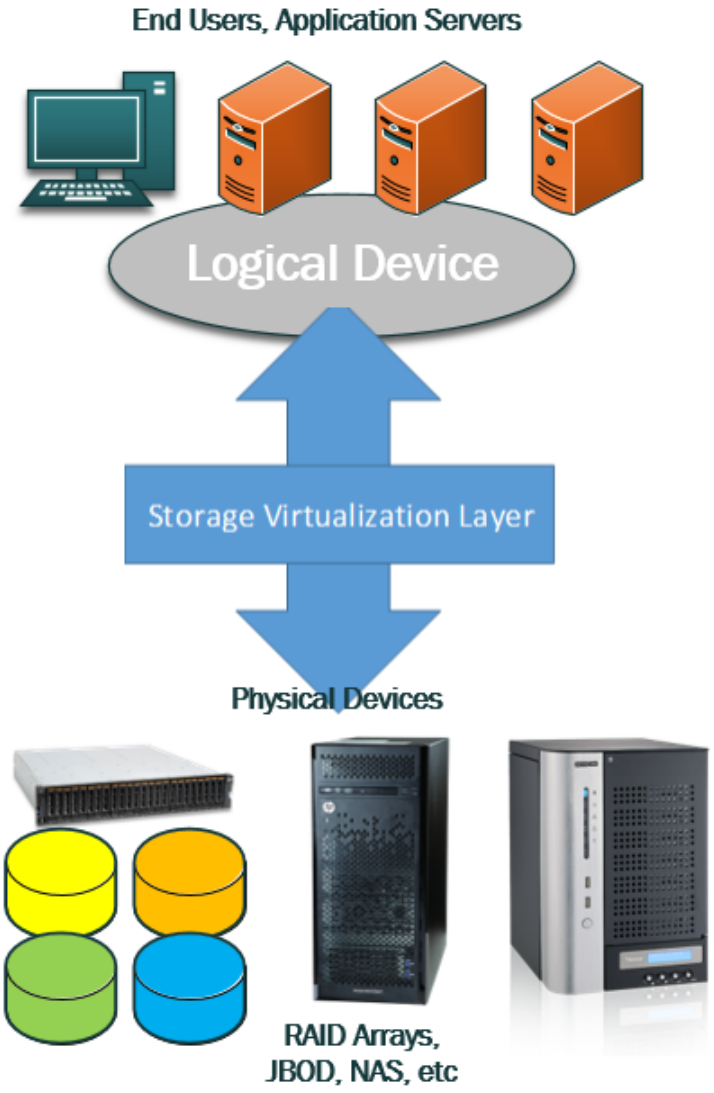

FIGURE 2.3: 儲存虛擬化架構（Storage Virtualization）

那除了不斷的加買無法完全利用的儲存設備外，還是花費大筆金額買具有高擴充性的單台設備，難道就沒有一個好的解決方法了嗎？於是就有廠商提出了儲存虛擬化（Storage Virtualization），利用虛擬化的特點來集中管理所有的儲存設備，如 Figure 2.3。但這仍然不是完全的軟體定義儲存架構。

而除了資源集中化管理之外，軟體定義儲存的核心精神更是透過軟體控制的方式，達成儲存資源的自動化及資源化；它允許你將儲存硬體抽象化並將所有的資源集中，並能達成在虛擬化環境下的「垂直擴充（Scale Up）」及「水平擴

充（Scale Out）」的彈性架構。而根據儲存網路工業協會 SNIA 的分析，他們更提出軟體定義儲存 SDS 架構應會具有以下十種特性：

1. 用戶可以自行建立 SDS。廠商會提供他們所推出的軟體與通用硬體設備，然後將環境建置起來。
2. 可以搭配任何硬體設備，或者是搭配強化的硬體功能特製設備。
3. 能夠支援橫向擴充的儲存設備，而不是典型的垂直擴充型儲存設備。
4. 盡可能地提供儲存資源與其他資源的共用功能。
5. 能以漸進方式建立儲存與資料服務的解決方案。
6. 可提供自動化管理機制。
7. 提供使用者自助式服務的操作介面。
8. 能涵蓋服務等級管理的形式，讓中繼資料能加上標記，進而使儲存與資料服務可根據這些資訊來套用，執行的粒度初期可能會很大，但未來可發展成具有更細部的服務等級調節能力。
9. 管理者可以訂定管理儲存與資料服務的政策。
10. 對於原本已經整合的儲存與資料服務，可以有自行分離的彈性。

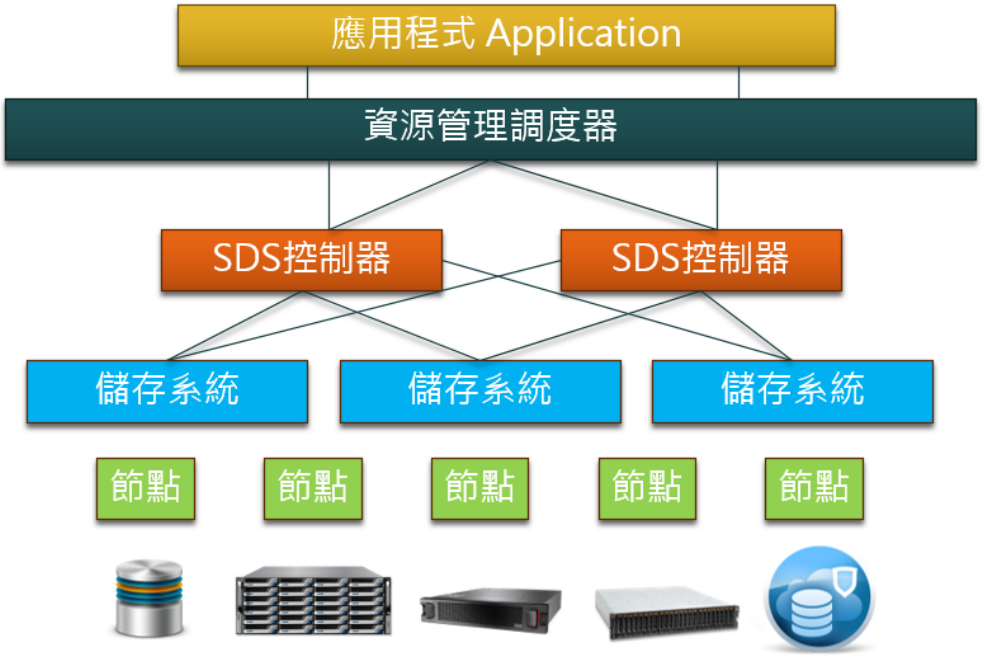

FIGURE 2.4: 軟體定義儲存架構（SDS）

2.1.3 分散式檔案系統

相對於儲存在本機端的檔案系統而言，分散式檔案系統（Distributed File System）或是網路檔案系統（Network File System），則是一種允許檔案透過網路分別跨多台主機上儲存的檔案系統。典型的分散式檔案系統會將檔案儲存在各個不同的伺服器，並透過統一的操作介面，讓使用者感覺不出檔案儲存在不同的機器上面，有如一般的檔案系統操作。如 Figure 2.5，使用者透過系統操作介面（User Interface），並且利用客戶端程式（Client Application）將操作命令利用網路發送出去，而伺服器接受到存取命令時，則將使用者要求資源提供給客戶端。

而在一個分享的磁碟檔案系統中，所有節點對資料儲存區塊都有相同的存取權，在這樣的系統中，存取權限就必須由客戶端程式來控制。分散式檔案系統包含的必要功能會包含透通的資料複製複寫功能與資料的容錯設計。也就是說，即使在檔案系統中有一小部份的資料節點離線或是服務伺服器硬體故障，對整個檔案系統來說系統仍然可以持續運作提供檔案服務而不會有資料損失無法存取的問題存在。

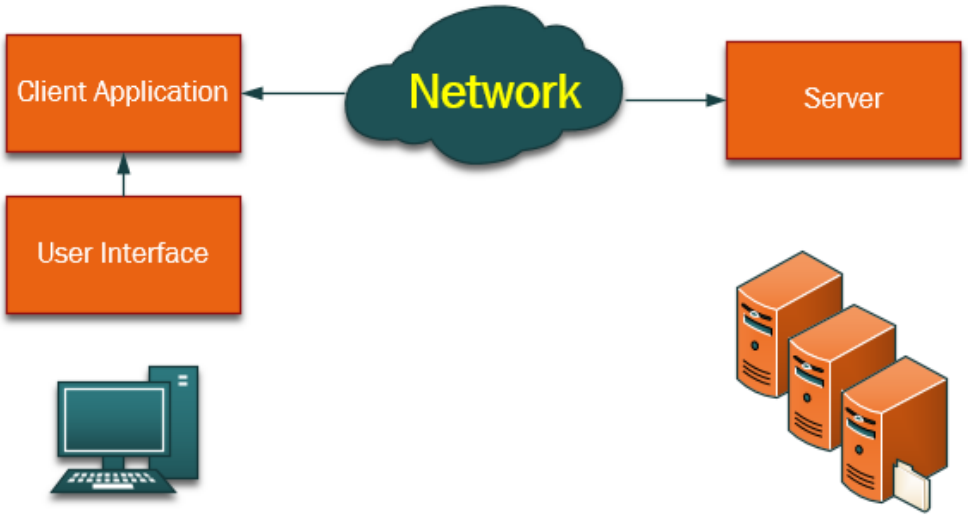

FIGURE 2.5: 分散式檔案系統基本架構

分散式檔案系統也具備高效能、高容錯、高可靠、高可用與高擴充的特性。它除了將資料分散的儲存在不同的資料節點之外，每個資料節點同時也具備分

散式處理計算的能力，所以原本只能在單一主機處理的程式邏輯與資料處理，可以分散到數以萬計的大量計算節點上運行處理，等每個節點計算出結果後，再將處理完成的個別結果結合起來，產生最終的結果，這種運作方式有效率的節省了資料存取與計算的排隊時間，提升了處理效能，例如 Google 搜尋在關聯式資料庫中無法達到這樣的效能與精確度。

在設計分散式檔案系統時通常會遇到的問題其中之一就是命名問題（nameing），命名（nameing）是邏輯物件（logical objects）與實體物件（physical objects）之間一種對應（mapping）；對於檔案本身來說，使用者在使用檔案時所知道的是檔案的名稱，而檔案本身儲存在什麼路徑則是由系統來決定，這中間需要一種對應（mapping），將檔案名稱對應到儲存的地點。例如就像是實體磁碟上的磁軌所儲存的資料方塊（Block），或是網路上的某台資料節點伺服器。

命名通常需要達到二個基本要求：

1. 地點透明化（location transparency）：檔案的名稱不需要包含顯示檔案儲存所在的路徑或是地點。
2. 地點獨立性（location independence）：當檔案的儲存路徑改變時，檔案的名稱不需要隨之改變。

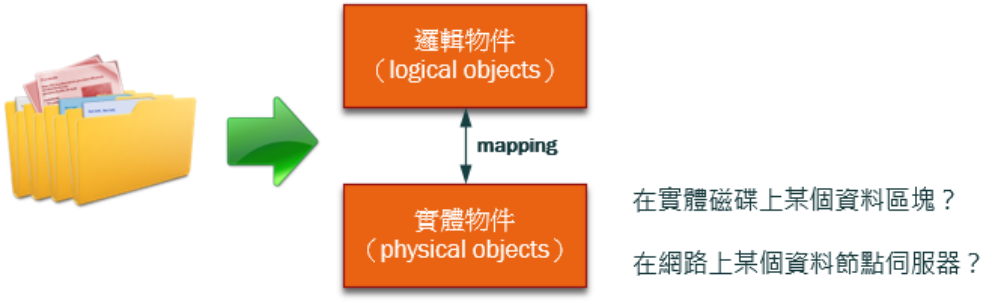

FIGURE 2.6: 分散式檔案系統命名的問題？

目前有三種命名方式可以參考：

1、將每個伺服器固定代號，並在寫入資料時與檔案名稱做對應。

2、將遠端伺服器透過掛載（mount）的方式，讓控制端（Controller）的伺服器可以直接使用遠端伺服器的儲存空間資源。

3、使用全域命名（Global Name Structure）的方式來包含所有的檔案，不過在實做的困難度比較大。

另外就是遠端檔案的存取方式上面，一般傳統的檔案存取是透過對本機實體磁碟的存取來取的檔案的寫入讀取權限，對分散式檔案系統而言，遠端檔案的存取由於不是在本機端的單一磁碟，所以需要透過遠端資料存取協定的機制，例如 RPC、HDFS、SMB 等。使用者使用遠端的檔案讀取及寫入時必須先送出請求，透過命名服務（nameing service）找到提供存取服務的伺服器，在控制端的伺服器收到存取要求後，再將使用者需要的檔案傳送過去。Figure 2.7 為常見的分散式檔案系統架構，通常在檔案存取時也可以運用快取（cache）來提升存取的讀寫效能，增加 disk I/O 的 IOPS。而在遠端檔案存取時，也可以運用 cache，除了減少 disk I/O 的使用外也可以降低網路的使用傳輸。

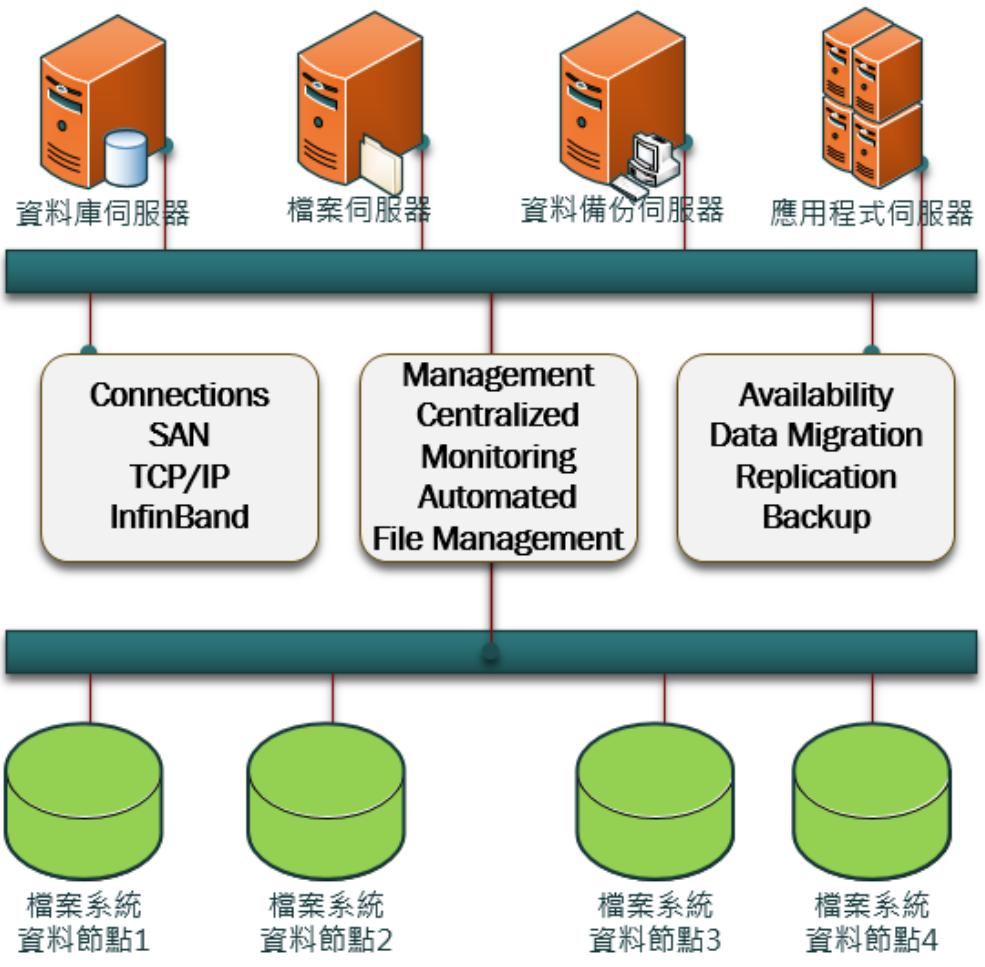

FIGURE 2.7: 分散式檔案基本架構

2.1.4 VMWare Virtual SAN

傳統儲存網路中的儲存設備（Networked Storage）依採用之協定與應用分為 SAN（Storage Area Network）與 NAS（Network Attached Storage）兩種架構，但二者皆為透過單一儲存硬體的方式在管理實體的儲存元件，並透過網路去連接掛載使用。而 vSAN 分散式儲存系統則有別於過去單一儲存硬體管理的方式，它將分散在不同主機上的實體儲存元件虛擬化後，在加入一個共同的儲存資源上面提供虛擬化環境直接使用。

vSAN 也是一種混合式硬碟系統，它利用本機的快閃記憶體裝置（PCIe Flash / SSD）做為資料讀寫的 Cache，同時結合本機的傳統機械式硬碟做為存放資料的所在。因此在生產環境中能夠有效的提升虛擬機器的資料讀寫服務 I/O 的效能。vSAN Datastore 為虛擬主機群的本地儲存資源所匯集而成，而效能由每一台虛擬主機成員當中的快閃記憶體裝置決定整體的執行效能，儲存資源的大小則由一般機械式硬碟空間來決定。因此當有新的虛擬主機成員加入時，整體的資料讀寫效能及儲存空間也會隨之增加，也就是「水平擴充（Scale Out）」，或者直接在原有的主機上增加硬體資源達成「垂直擴充（Scale Up）」。

在 vSAN 的運作架構當中，至少需要三台的 ESXi 成員主機提供 vSAN 儲存資源，且這三台主機都需滿足 vSAN 儲存架構的最小要求「一顆或一片快閃記憶體裝置（PCIe Flash / SSD），以及「一顆傳統機械式硬碟（SAS / SATA）。當 vSAN 儲存空間建立後，其他的虛擬主機即使不提供儲存資源也能直接使用，如 Figure 2.8 所示。

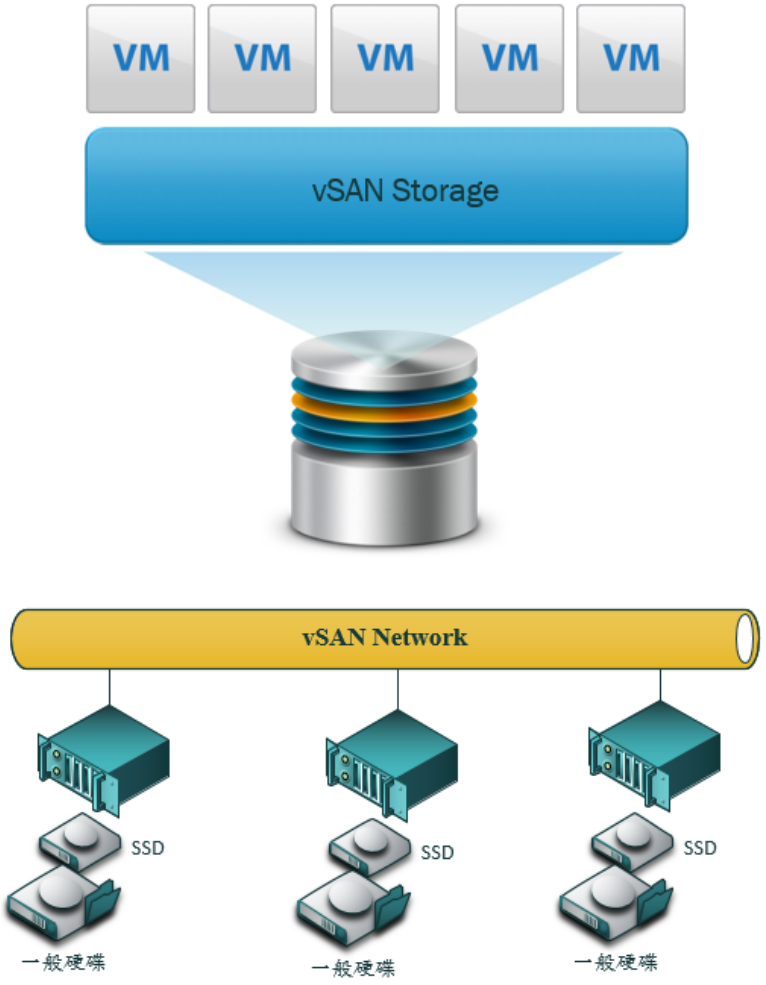

FIGURE 2.8: vSAN 分散式檔案基本架構

vSAN 分散式儲存系統透過「儲存原則管理（Storage Policy-Based Management, SPBM）」來控制佈署 VM 虛擬主機到 vSAN Storage 的機制，任何儲存在 vSAN 當中的虛擬機都需要套用儲存原則才能運作。並且有別於傳統儲存設備的實體硬碟 RAID 陣列保護，在 vSAN 當中則是透過儲存原則來做虛擬硬碟層級的陣列保護，如：RAID 0 機制的 Stripe、RAID 1 的 Mirror 功能。

目前來說在 vSAN 儲存原則當中，可供控制設定的共有五項如下：

(1) 容許的故障次數

簡單來說就是虛擬主機儲存物件的可容忍故障數量，其預設值為 1，可供設定的範圍為 0 到 3。當其值為 1 時，系統會自動將虛擬主機儲存物件額外產生一份副本，邏輯架構如 Figure2.9 所示，並且系統會自動達成如 RAID1 的副本機制，如 Figure 2.10 所示。

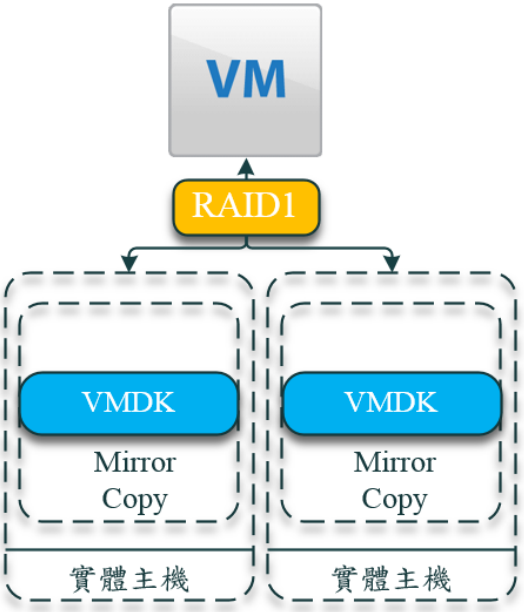

FIGURE 2.9: 容許的故障次數邏輯架構

FIGURE 2.10: 容許的故障次數為 1 時

(2) 每個物件的磁碟等量區數目

有別於容許的故障次數是做為儲存物件的副本，每個物件的磁碟等量區是將儲存物件做分割，以到類似於 RAID0 的 Stripe 機制。其預設值為 1，可供設定的範圍為 1 到 12，如 Figure 2.11 所示。

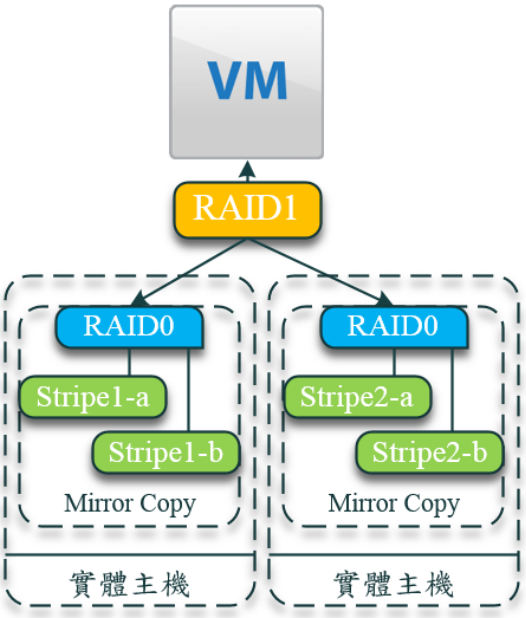

FIGURE 2.11: 容許故障 1 和磁碟等量區 2 邏輯架構

(3) Flash 讀取快取保留區

vSAN 將資料的讀寫置放於 Flash 快閃記憶體裝置中，以加快資料讀取及寫入的速度，因此也可以在這裡設定要將儲存物件的百分之多少直接保留在 Flash 元件中，預設值為 0，可供設定的範圍為 0

(4) 物件空間保留區

因為在 vSAN Storage 中，預設所有的儲存物件皆為精簡佈建（Thin），因此你可以在這裡設定在 vSAN Storage 中要保留多少的空間給儲存物件本身，預設值為 0，可供設定的範圍為 0

(5) 強制佈建

當 vSAN 儲存架構中不滿足你前面所設定的儲存原則時，是否要先強制佈建虛擬機，待之後硬體擴充時再自動依設定的儲存則調整。

2.2 相關研究

在本研究中，我們使用了 Iperf 來模擬真實世界的伺服器網路工作負載。Iperf 被廣泛選擇為用於測量 TCP 和 UDP 頻寬性能的 I/O 測試的一個標準。我們測試中使用 Iperf 運行用戶端可能的密集型多工處理程序。並且在最終測試結果出來前，在每個小單元測試完畢後，下一個測試必須等待五分鐘，直到確認網路封包處理程序已經結束。此種測試方式是參考了 Danhua Guo 所使用的測試方法 [4]，並且 Danhua Guo 也在該篇文章中提到，關於高速網路卡應用在多核心處理器的環境中，在其底層的作業系統預設安裝的情況下，並無法完全啟用及分配網路 I/O 的處理序到適合的處理器核心上，無論是在一般常見的本機安裝作業系統上或是虛擬化的環境中。如 Figure 2.12 所示，所以我們最後在調整優化高速網路卡時，便參考了相關的原廠文件做出適當的調整。

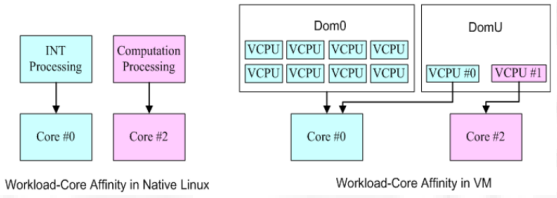

FIGURE 2.12: Workload-Core Affinity

另外我們也參考了 Weijun Xiao 文中所使用的測試方法 [29]，使用 IOMeter 並調整 Outstanding I/Os 為 16，並測試區塊從 4KB 到 256KB 之間的效能變化，主要測試的是針對資料讀取的效能。

Iometer 主要可以用來模擬量測硬碟 I/O 資料讀寫的速度及效能，透過不同的 pattern 可以模擬主機儲存媒體真正在被存取時 I/O 的表現及 IOPS。Iometer 是在 1998 年 2 月 17 日的 IDF(Intel Develpoer Forum)Intel 所提出來的，但是目前已經移轉到 Open Source 來開發。2001 年 11 月，Iometer 在 SourceForge.net 註冊，開發項目從 2003 年月 2 月起又重新啟動，當然項目維護者變成了獨立的一個內部工作組。IOmeter 包括 2 個程式，Iometer.exe 和 Dynamo.exe。其中 Iometer 是控制端程式並且是圖形介面，讓你可以簡單調節參數和顯示測試結

果，而 Dynamo 就是讓測試儲存媒體產生壓力測試的主程式，Iometer 來主動控制 Dynamo 程式產生工作負載。

另外我們使用 VMware 原廠所提供用於測試及檢測 Virtual SAN 的測試工具，參考文件中提到此種測試會產生每台主機 10 到 20 個的 VMDK 文件（將由 Virtual SAN 分佈到每個第二層實體硬碟上面）[34]。文件產生完成後，測試將會同步並行處理所有主機實體磁碟上的 VMDK 虛擬磁碟檔案，並產生資料 I/O 讀寫的工作負載。測試完畢之後，VMDK 文件就會被刪除。由於 Virtual SAN 前端所使用的 Cache 快取磁碟，會隨著測試時間的拉長而留存較為準確的測試資料，因此測試就必須得分成幾種測試模式，以真實的模擬生產環境中可能遇到的狀況，而非單純的拉長測試資料讀寫的時間，追求測試結果的高讀取快取命中數值。

Chapter 3

系統設計與實作

3.1 建置環境

本實驗環境使用 VMware vSphere 6.0 配合 Horizon View 6 實做 Virtual SAN 虛擬儲存架構及 Virtual Desktop 佈署情境，並分為三個不同的環境以測試在不同的條件下虛擬分散式儲存的效能差異，以提供實務生產環境下效能條調教設定的參考。

儲存讀寫 I/O 測試則使用 IOMeter 軟體及 VMware 原廠 Virtual SAN Health Check Plugin。

3.1.1 多節點實驗環境說明

本實驗環境使用英業達 Zion 伺服器做為實驗機器，英業達 Zion Server 為機櫃集成的類高密度伺服器設備，同一個 Rack 內可同時容納二台實體伺服器，同一個機箱內可同時裝載 35 個 Rack 及 70 台伺服器，本研究實驗同時最高使用十台 Zion 伺服器搭建本環境測試。

其中由於 Zion Server 預設規格僅支援單一儲存硬碟裝置，所以我們需要另外手動銲接非原廠的 SATA 資料 Y 型接頭，以讓單台伺服器可以同時安裝二

顆實體磁碟，以滿足 Virtual SAN 的基本建置需求。單台伺服器規格如 Figure 3.1。

TABLE 3.1: 多節點實驗環境設備規格

硬體設備型號：**Inventec Zion Server**		
CPU	Intel(R) Xeon(R) CPU E5540 Quad-Core 2.533GHz	2 顆
RAM	4GB DDR3-1066 ECC Register	12 支
SSD	Intel 730 Series 240GB SSD	1 顆
HDD	WD 2003FYYS SATA 2 2TB 7.2K	1 顆
LAN	1Gbe	4Port

3.1.2 異質網路環境說明

本實驗環境使用 HP ProLiant DL380 Gen9 伺服器做為實驗機器，HP ProLiant DL380 Gen9 為 2U 類型伺服器，單台伺服器最高可以安裝 24 顆實體硬碟，本研究實驗使用共三台實體 HP 2U DL380 Gen9 主機組成 Virtual SAN 環境，設備規格如 Figure 3.2。

TABLE 3.2: 異質網路實驗環境設備規格

硬體設備型號：**HP ProLiant DL380 Gen9**		
CPU	Intel(R) Xeon(R) CPU E5-2640 v3 8-Core 2.60GHz	2 顆
RAM	16GB DDR4-2133 ECC Register	8 支
SSD	Intel 730 Series 480GB SSD	2 顆
HDD	600GB 10K 6Gbps SAS 2.5	12 顆
LAN1	HP Ethernet 1Gb 331i Adapter	4Port
LAN2	HP Ethernet 10Gb 560SFP+ Adapter	2Port
LAN3	Mellanox MT27520 ConnectX-3 Pro	1Port

3.1.3 全快閃式儲存環境說明

本實驗環境使用自組式 PC 做為實驗機器，並使用常見的一般家用型 SATA 7200 轉硬碟和個人用快閃式裝置 SSD 來組合 Virtual SAN 環境，本架構使用四台主機，並會在此實驗中測試「垂直擴充（Scale Up）」及「水平擴充（Scale Out）」的效能成長幅度。設備規格如 Figure 3.3。

TABLE 3.3: 全快閃式儲存實驗環境設備規格

硬體設備型號：**DIY 自組 PC**		
CPU	Intel® Core™ i7-5960X 8-Core 3GHz	1 顆
RAM	16GB DDR4-2133	4 支
SSD	Intel 535 Series 480GB SSD	2 顆
HDD	WD 2003FYYS SATA 2 2TB 7.2K	2 顆
LAN1	Intel 82574L 1Gbe	1Port
LAN2	Mellanox MT27520 ConnectX-3 Pro	1Port

3.2 系統架構與實作

3.2.1 多節點實驗環境架構

在多節點的實驗環境中，我們除了測試資料讀寫的 I/O 效能外，也搭建了 VMware Horizon View 的 Virtual Desktop 環境，以測試虛擬桌面的佈署效能。網路部分因為英業達 Zion 伺服器單台主機有四個 1GbE 千兆網路介面，所以我們會使用 3 個 1GbE 網路介面，實做配合 802.3.ad with LACP（Link Aggregation）的技術來提升 Virtual SAN 分散式儲存環境的資料傳輸效能。

並且由於 Virtual SAN 是符合 Software-Defined Storage（SDS）的產品，其中一個重要的因素就是它可以透過軟體定義的儲存原則來設定虛擬磁碟本身的容錯、切割，甚至是保留前端 Cache Tier 的 SSD 快取比率，這些多重的因素考量下，該如何設定才能得到管理人員想要得到的效能改善，這也是我們在這個實驗章節要討論的。

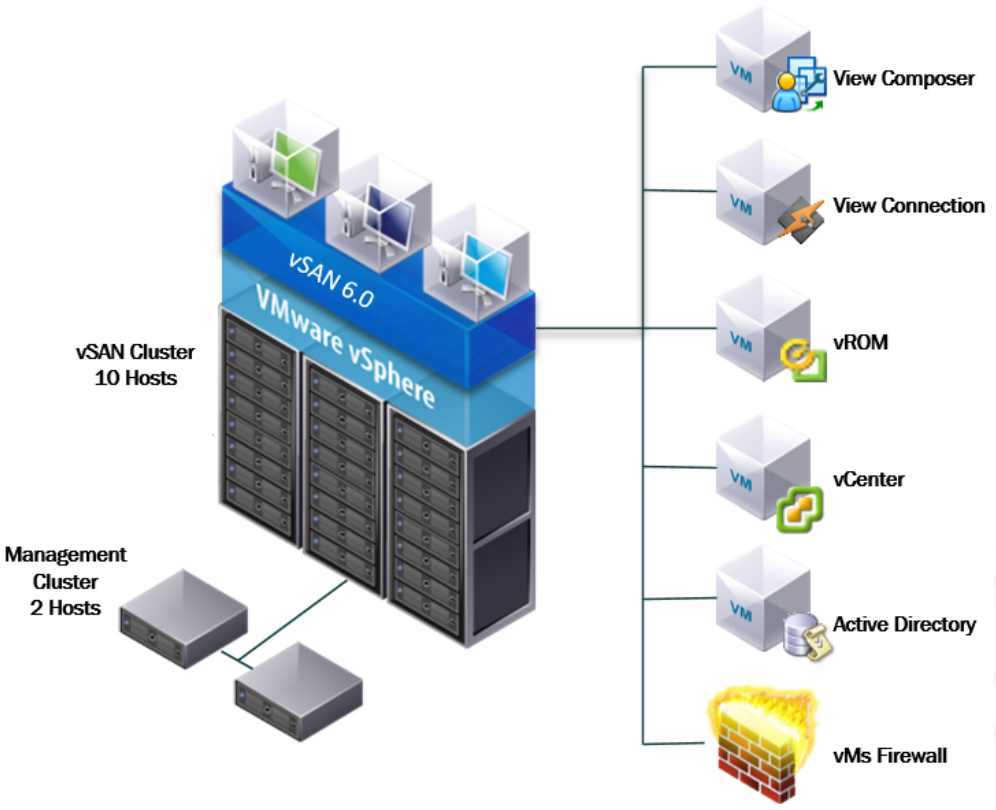

FIGURE 3.1: 多節點實驗環境架構圖

3.2.2 異質網路環境架構

在異質網路環境測試中，我們使用了二種不同的網路介面以測試 Virtual SAN 在高速網路下的效能成長幅度，並測試 Mellanox 的 40GbE QSFP+ 光纖網路卡對於虛擬化 Hypervisor 底層的硬體相容性及網路效能，因為組成 Virual SAN 分散式儲存環境效能的一個最重要先決條件便是高速的網路傳輸，在原廠 VMware 的文件中更是指定在生產環境應使用 10Gbase 的萬兆高速乙太網路卡來搭建 Virtual SAN 儲存環境。

因此在此環境中，我們為了滿足測試異質高速網路的需求，我們得先排除儲存裝置磁碟本身的效能瓶頸，已避免造成不是網路頻寬不足而是儲存裝置的資料讀寫效能不足的錯誤測試因素。

另外此環境中我們使用較高階的商用型快閃式裝置以及企業等級的 SAS 萬轉硬碟，並在 Virtual SAN 分散式儲存的環境中每台主機設定分為二個磁碟群

組，每個磁碟群組同時使用 6 個 SAS 萬轉硬碟配合 Intel 730 快閃式裝置 SSD。網路環境架構如 Figure 3.5 中所示。

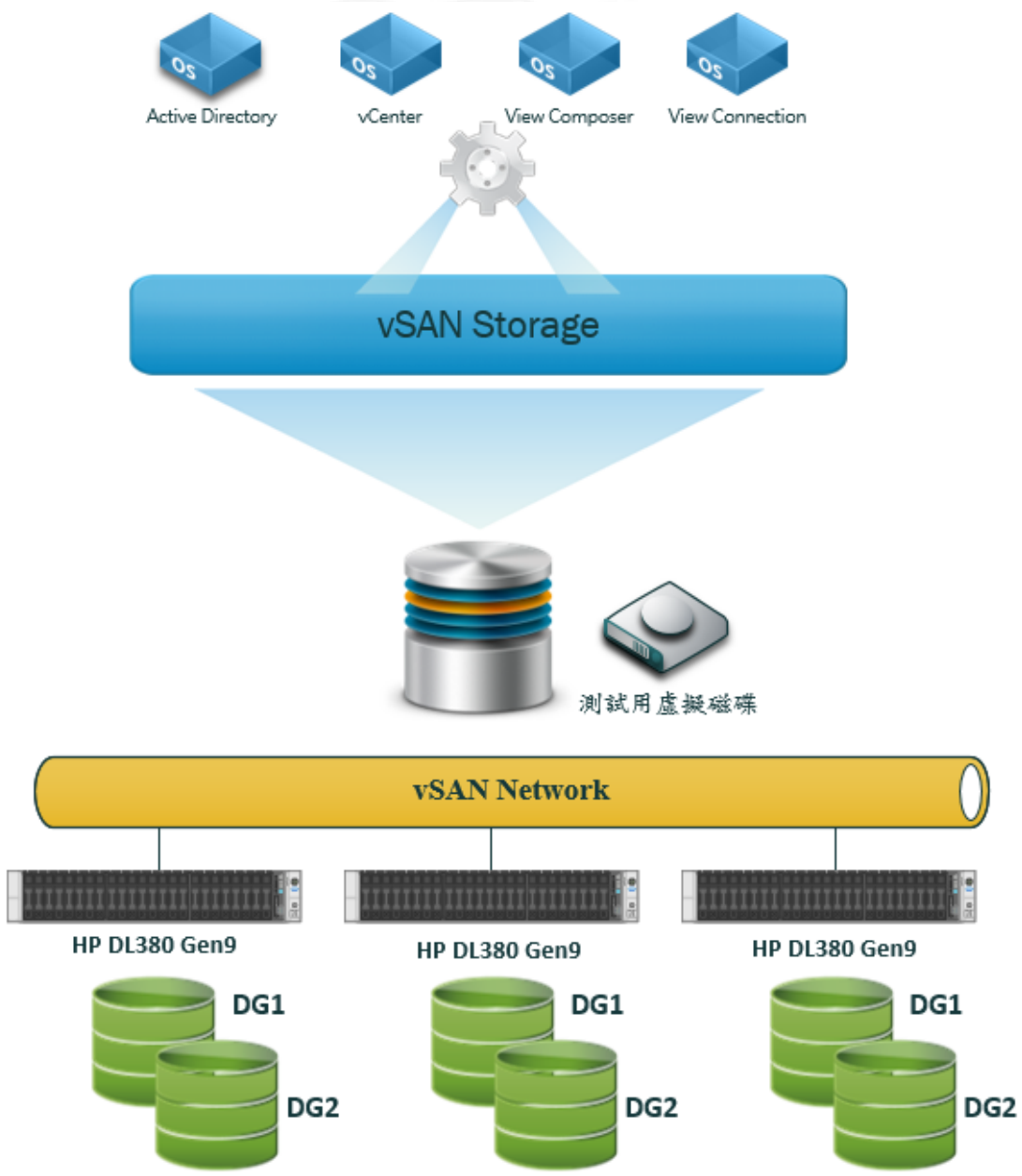

FIGURE 3.2: 異質網路實驗環境架構圖

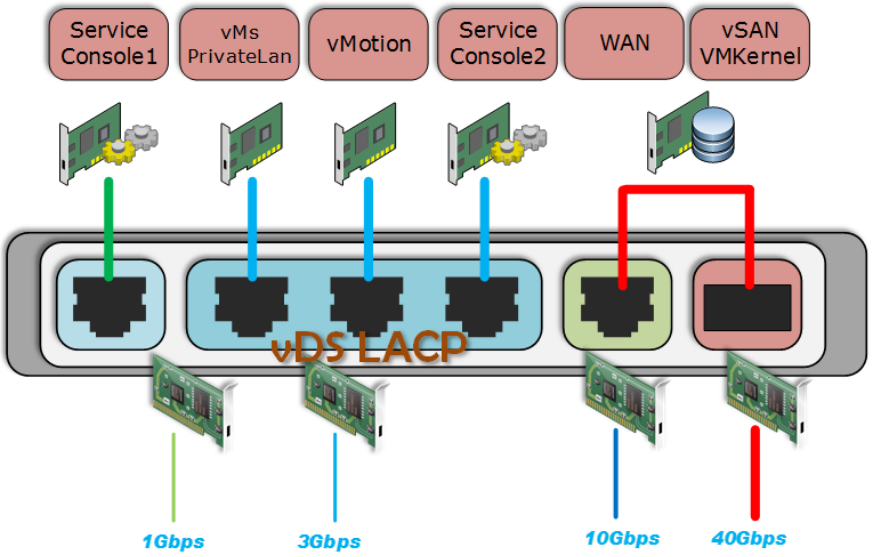

FIGURE 3.3: 異質網路實驗環境網路卡設定

3.2.3 全快閃式儲存環境架構

在全快閃式儲存環境的測試中，我們使用了四台資料節點來組成測試環境。並同樣配合 Mellanox 40GbE QSFP+ 光纖網路介面來組成 Virtual SAN 分散式儲存環境，已確保不會遇到網路瓶頸而影響研究資料的正確性。

因為在這個實驗架構中，我們所要了解的是如果將所有的磁碟置換成快閃式裝置 SSD，那麼 Virtual SAN 儲存環境的效能將會是爆發式的增長？或是有限幅度的提升？另外在 Virtual SAN 儲存環境中擴充硬體設備，又該如何選擇使用「垂直擴充（Scale Up）」或「水平擴充（Scale Out）」的應用時機。

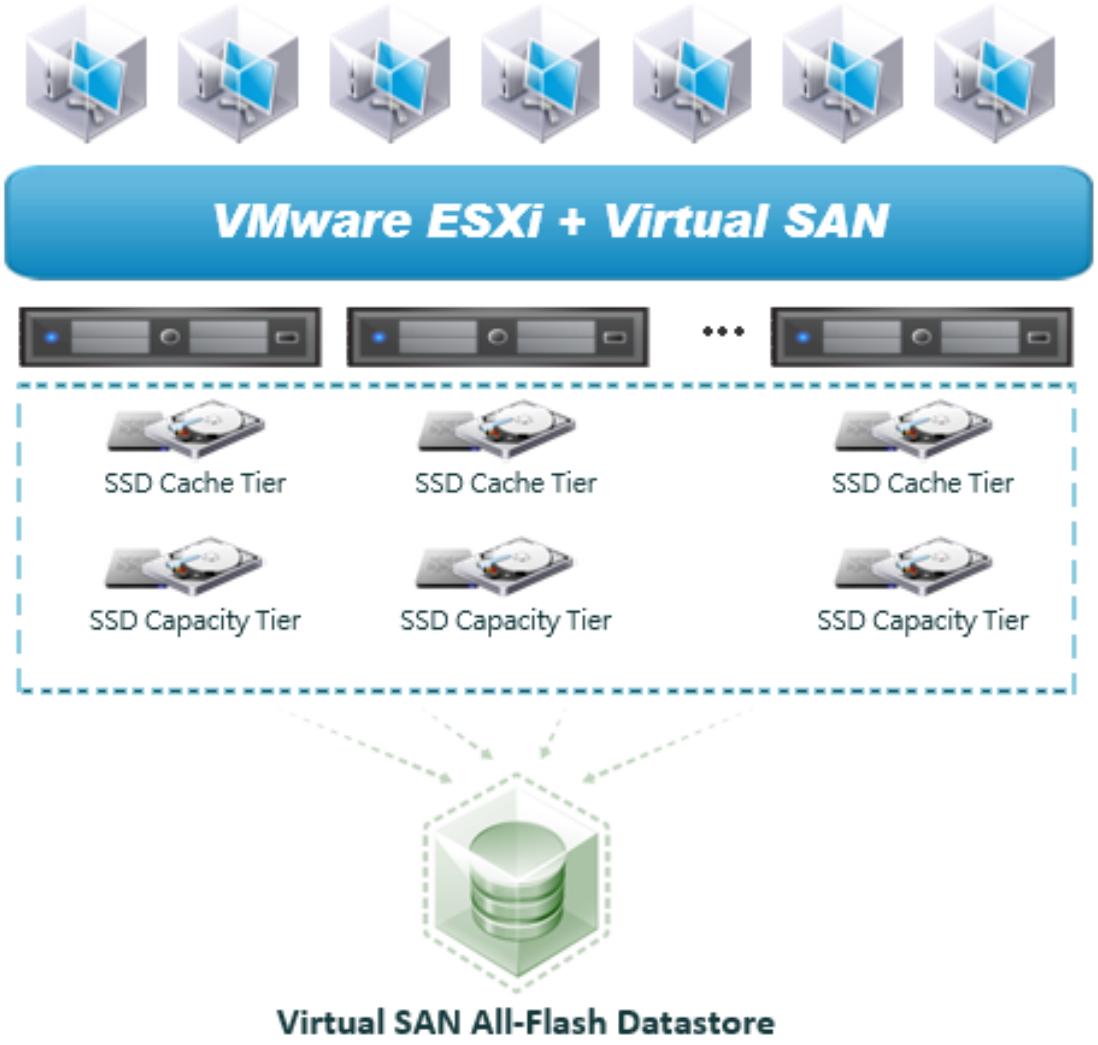

FIGURE 3.4: 全快閃式儲存環境架構

Chapter 4

實驗環境與結果

4.1 實驗環境

VMware Virtual SAN 的架構主要是將多台伺服器內建的儲存硬碟空間視為一個統一的資源池，讓伺服器內部的硬碟空間成為彼此共享的儲存資源。而伺服器叢集的虛擬硬碟資料會互相複製，當其中一個硬碟因故不能運作時，VM 還能夠從其他的硬碟中讀到資料。由於 VSAN 的架構沒有本地的觀念，因此資料存放可能不會是在原本應用服務運行的伺服器上。由於 Virtual SAN 支援「垂直擴充（Scale Up）」及「水平擴充（Scale Out）」的彈性架構，所以理論上在愈多的資料伺服器節點上面，可以得到更為顯著的效能提升。

因此我們的實驗就是希望能得到在不同的擴充型式上，無論是「垂直擴充（Scale Up）」或「水平擴充（Scale Out）」，效能的成長是屬於線性成長亦或是有其他需要考量的因素存在。在原廠 VMWare 的官方文件中，Virtual SAN 1.0 版最高可支援 32 節點叢集，根據 VMware 內部標準測試，可達 200 萬每秒輸入輸出運行（IOPS）的讀取工作負載，讀寫工作負載為 640,000 IOPS；如果是 NVMe 固態硬碟的部分，透過此新式通訊介面可提供 VSAN 更佳的運作效能，據 VMware 官方的測試結果顯示，可達每台 VSAN Node「10 萬 IOPS」，也就是在 32 Nodes 的 VSAN Cluster 運作架構中，提供高達「320 萬 IOPS」的儲存效能。

因此我們會同時測試全快閃式儲存的環境中，效能的成長是否能如預期般的明顯，而現在多數的中小企業生產環境中，主流網路多數仍然都是 1GbE 等級的乙太網路，那僅有 1GbE 的網路是否又符合搭建 Virtual SAN 的需求呢？我們實驗中也會包含利用 1GbE 網路做網路聚合模式，配合 10GbE 等級乙太網路和 Mellanox 的 40GbE 特殊乙太網路光纖卡規格做測試，以求最出最適合 Virtual SAN 的企業生產環境規劃建議。

4.2 實驗方法

4.2.1 多節點環境實驗

我們將建置好的 vSAN 資料節點叢集，額外加入了一台同規格的伺服器，但並未貢獻本機的儲存空間，只透過 vSAN 資料交換網路掛載了 vSAN Storage。然後新增測試用的虛擬機並存放於本機的硬碟儲存空間，然後先透過磁碟資料讀寫測試軟體 IOMeter 測試本機的硬碟讀寫速度，以此做為接下來測試的對照組，我們並且設定 Outstanding I/Os 為 16，以多工方式提高設定的精準度，測試結果如 Table 4.1。

Table 4.1: 一般內接式硬碟 IOPS 測試值

Type	Latency(ms)	AVG IOPS
4 KB 100% Read	1.09	14693
64 KB 100% Read	4.19	3818
256 KB 100% Read	31.87	501

然後在新增一個 500GB 的虛擬磁碟存放於 vSAN Storage 上面，參考 Figure 4.1 架構。

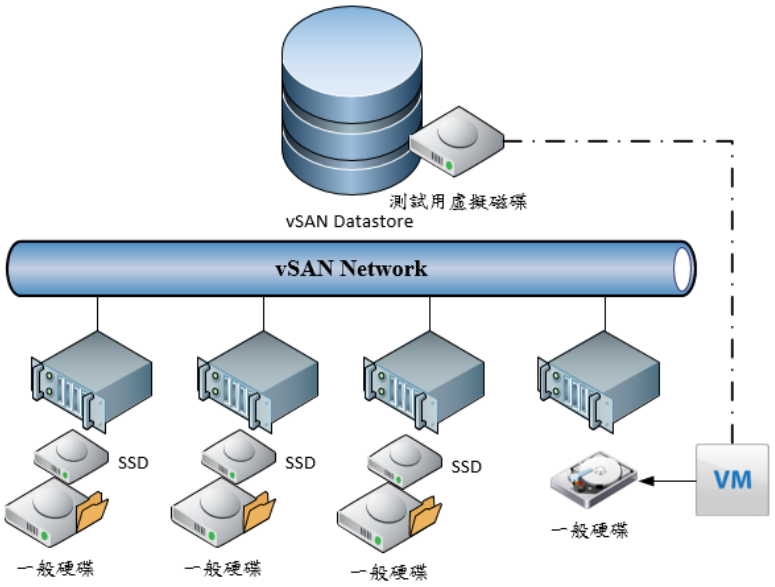

FIGURE 4.1: 多節點實驗環境架構圖

這樣在測試的過程中可以避免虛擬機本身的資料 I/O 影響了存放於 vSAN Storage 的虛擬磁碟資料讀寫效能，來增加測試數值的穩定性；然後我們將可以進一步透過調整 vSAN 資料節點網路的 802.1ad LACP 的網路聚合條件由一張 1GbE 網路卡逐步調整至三張 1GbE 網路卡，以此瞭解網路速度對 vSAN 分散式儲存的影響。然後我們也調整切割磁碟等量條件於測試用的虛擬磁碟上，以瞭解此分散式儲存方式是否對效能有所幫助。

最後我們再加入傳統 NAS 透過 iSCSI 方式掛載 10 顆 2TB 硬碟組合的 RAID 10 陣列磁碟，然後透過 Horizon View 6.1 同時佈署生 60 台測試虛擬機器，然後跟透過 vSAN 分散式儲存的方式佈署做比較，以此檢視兩者的效能差異。

4.2.2 異質網路環境實驗

由於在分散式儲存環境中，資料的交換速度仰賴於網路傳輸的快慢，因此不同頻寬的網路介面，相對會影響分散式儲存的資料讀寫效能。前面我們透過線路聚合（802.13.ad with LACP）的技術實測了 3 張千兆位元速率的網路卡聚合，並以此應用在 Virtual SAN 分散式儲存的環境中，而在這個異質網路環境的實

驗中，我們將實測更高速率的網路介面，測試的基準將拉高至 VMWare 原廠建議的 10GBps over Ethernet 以及較為特別的 Mellanox 40G 高速光纖網路介面。

Mellanox MT27520 ConnectX-3 是 QSFP+ 介面，QSFP（Quad Small Form-factor Pluggable）：四通道 SFP 介面 (QSFP)，QSFP 是為了滿足市場對更高密度的高速可插拔網路解決方案的需求而誕生的一種光纖介面，這種 4 通道的介面每通道 10Gbps 的速度支援四個通道的同時資料傳輸，傳輸速率可達到了高達 40Gbps 的網路傳輸速度。目前為了滿足 QSFP 的高速網路需求，他必需至少得安裝在支援 PCI Express 3.0 x8 的介面上，而且由於是透過多通道的並發鏈接的“轉換線纜”技術，所以也必需得透過軟體平台的支援才能達到 40GBps 的高速網路。

而在 VMWare 的環境中，我們先使用 IPerf 實測了點對點的網路頻寬，發現了幾個會影響後續實驗結果的問題。如 Table4.2 是 Intel 10GBps 的測試結果，Table 4.3 是依預設值安裝 Mellanox 40GBps 的測試結果。

TABLE 4.2: Intel 10Gbps 網路卡測試結果

Intel 10G Iperf Test Result(Gbits/sec)						
Number of parallel	AVG Bandwidth	First	Second	Third	Fourth	Fifth
1	5.688	5.83	5.46	5.9	5.37	5.88
6	9.13	9.18	9.04	9.23	9.14	9.06
12	9.122	9.18	8.91	9.06	9.25	9.21
24	9.044	9.01	9.06	8.96	9.16	9.03

TABLE 4.3: Mellanox 40Gbps 網路卡預設值測試結果

Mellanox 40G Default Installation Iperf Test Result(Gbits/sec)						
Number of parallel	AVG Bandwidth	First	Second	Third	Fourth	Fifth
1	7.382	7.94	6.39	8.73	6.45	7.4
6	14.18	13.8	14.7	13.7	14.7	14
12	14.44	14	14.7	14.2	14.5	14.8
24	14.04	13.7	14.3	13.9	14.3	14

由於 Mellanox 40G 的網路效能測試結果不佳，這勢必會影響我們後面的實驗，因此在查閱相關文件後，我們調整了幾個 ESXi 的參數，得到了如 Table 4.4 的測試數值。

TABLE 4.4: Mellanox 40Gbps 網路卡最佳化測試結果

Mellanox 40G Optimization Iperf Test Result(Gbits/sec)	
Number of parallel	AVG Bandwidth
1	18.6
6	33.5
12	32.2
24	32.5

到目前為止我們在配合多個處理程序的情況下已經得到了超出 10Gbps 三倍的網路效能，但這仍然沒達到 40Gbps 的網路頻寬，在繼續查詢原廠的說明文件後，我們得知了如果要達到 40Gbps 的結果，我們得調整虛擬機器底層的參數設定，並配合設定虛擬機裡面作業系統的多線程處理器並行程序，才可以達到。測試結果如 Table 4.5。

TABLE 4.5: Mellanox 40Gbps 網路卡 Ubuntu 作業系統測試結果

Mellanox 40G Virtual Machine Iperf Test Result(Gbits/sec)	
Number of parallel	AVG Bandwidth
6	35.7
12	37.6
24	36.3

雖然我們得到了接近 40Gbps 的測試數據，但這是基於透過虛擬機器對虛擬機器的測試，而 Virtual SAN 是直接原生在 Hypervisor 底層的分散式儲存應用，且還並不了解 Virtual SAN 的網路使用是否可以配合多個 CPU 核心來並行處理多個處理序，所以對 Virtual SAN 所使用的網路來說，仍只有可能最底是 18 19Gbps 或最高 32Gbps 的頻寬可用。

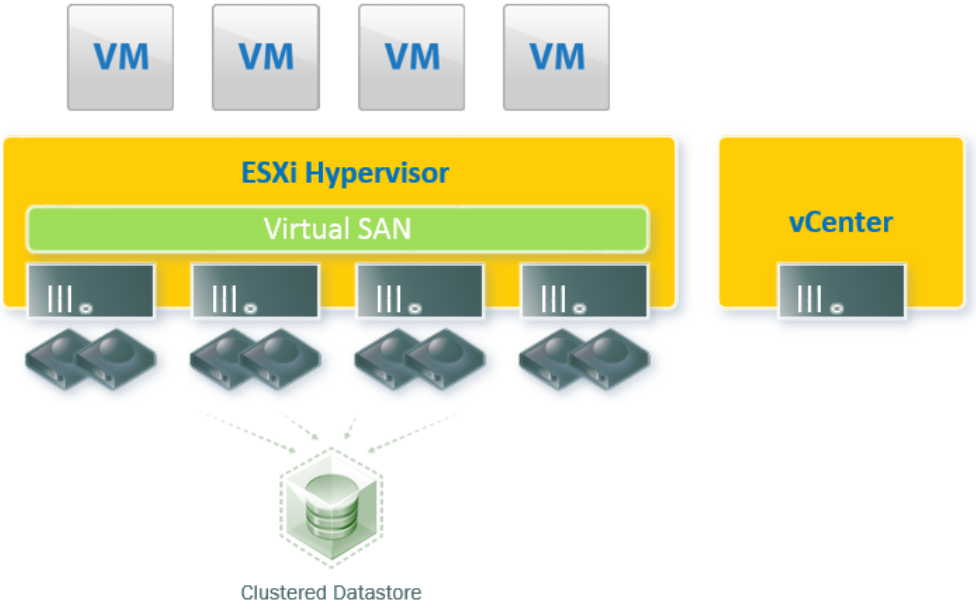

FIGURE 4.2: Virtual SAN 整合 Hypervisor 融合架構

而這一次的測試將採用 VMWare 原廠所提供的 Virtual SAN Health Plugin 測試工具，此工具除了可以檢測 Virtual SAN 環境是否服務正常外，也可以提供許多資料讀寫的效能測試。實驗將以下列五種測試方法實做：

1. Basic sanity test, focus on Flash cache layer

   此種測試方法模擬常見的工作狀態，並在每台測試主機產生 1GB 的測試檔，分別同時做 70% 的讀取以及 30% 的寫入動作。

2. Stress test

   此種壓力測試方法將會做多重的同時讀寫動作，並在每台測試主機產生 1TB 的測試檔。

3. Performance Characterization - 100% Read, optimal RC usage

   此種測試方法將大量的偏重使用前端的快取緩存裝置，並在每台測試主機產生 10GB 的測試檔。

4. Performance Characterization - 100% Write, optimal WB usage

   此種測試方法將大量的偏重使用寫入緩存來做測試，並在每台測試主機產生 5GB 的測試檔。

5. Performance Characterization - 70/30 Read/Write, high I/O size, optimal flash cache usage

此種測試方法將使用 64k 的 Block 大小來做測試，並同時做 70% 的讀取及 30% 的寫入動作，並在每台測試主機產生 30GB 的測試檔。

4.2.3 全快閃式儲存環境實驗

在 Virtual SAN 的架構中，我們知道是在前端使用快閃式裝置 SSD 來加速分散式儲存的讀寫效能及回應速度，那如果後端的機械式傳統硬碟也使用快閃式裝置 SSD 呢？是否會得到更好的讀寫效能還是會受限於 Virtual SAN 演算法的限制，反而造成效能下降？在 Virtual SAN 1.0 版中我們並無法實驗這個研究，因為版本的限制所以無法實做。但在 Virtual SAN 6.0 的版本中，VMWare 已經可以實做這個設定，所以我們將在這個章節透過 Mellanox 40G 的光纖網路卡來做測試。

另外，在 Virtual SAN 分散式儲存的環境中，因為支援「垂直擴充（Scale Up）」及「水平擴充（Scale Out）」，在前面的章節中我們測試了「水平擴充（Scale Out）」帶來的效能成長，而在全快閃式儲存環境實驗中，我們也會同時比較傳統硬碟做「垂直擴充（Scale Up）」的效能成長幅度。

這一次的測試將同樣採用 VMWare 原廠所提供的 Virtual SAN Health Plugin 測試工具，並多增加以下四種測試方法實做：

1. Performance Characterization – 70/30 Read/Write, realistic, optimal flash cache usage

此種測試方法將使用 4k 的 Block 大小來做測試，並同時做 70% 的讀取及 30% 的寫入動作，並在每台測試主機產生 30GB 的測試檔。

2. Performance Characterization - 100% Read, Low RC hit rate / All-Flash demo

此種測試方法將 100% 使用快閃式裝置做測試，並在每台測試主機產生 1TB 的測試檔。

3. Performance Characterization - 100% Streaming Reads

此種測試方法將會模擬一個持續性讀取的串流服務，像是媒體伺服器；並在每台測試主機產生 1TB 的測試檔。

4. Performance Characterization - 100% Streaming Writes

此種測試方法將會模擬一個持續性寫入的服務，像是備份伺服器；並在每台測試主機產生 1TB 的測試檔。

4.3 實驗結果

4.3.1 多節點環境實驗結果

4.3.1.1 網路速度影響 vSAN 分散式儲存的效能比重

我們先依 vSAN Storage 的預設儲存原則對測試用虛擬磁碟做 IOMeter 效能測試，結果得出來的結果如 Figure 4.3 所示。結果在預設的儲存原則下，發現調整網路卡的聚合數量，幾乎對效能沒有任何影響；我們猜測可能的原因是因為測試用虛擬磁碟資料的讀寫都來自於同一台資料節點的主機，然後在單一主機的資料吞吐量無法滿足單 1GbE 的速度，所以造成 IOPS 幾乎沒什麼改變的情況。

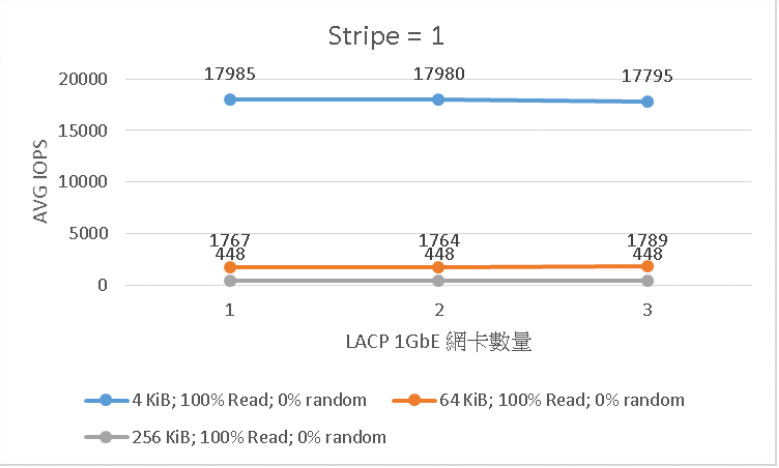

FIGURE 4.3: Stripe = 1 測試網卡聚合數量

所以我們嘗試改變調整切割磁碟等量的數量，希望以此測試當資料同時存放在多個資料節點時，調整網路卡聚合的數量，是否會對結果造成改變。測試

出來的結果發現，在 4KB 測試時幾乎還是不會有什麼效能上的差異，但在測試 64KB 時，同一時間聚合 2 張 1GbE 的網路卡，效能有了提升 51% 的顯著提升，但在增加聚合 3 張 1GbE 網路時，效能卻又跟 2 張的結果一致，測試 256KB 時結果同樣。

因此我們可以得知，在資料同時讀寫較大區塊的檔案時，網路的速度將會影響讀寫的效能。測試結果如 Figure 4.4 所示。

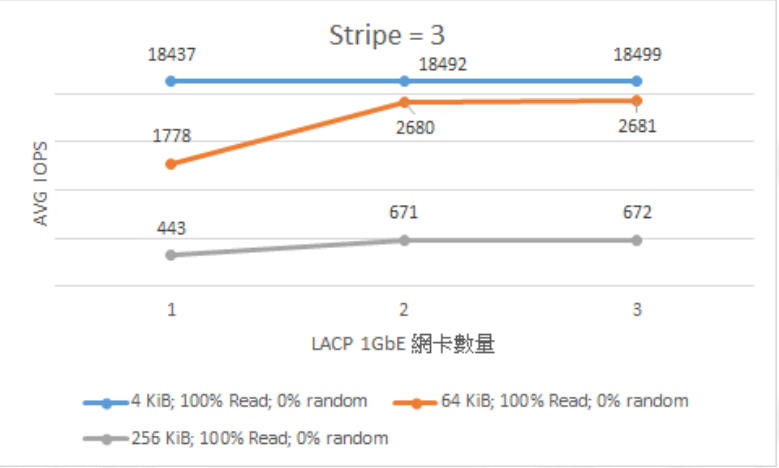

FIGURE 4.4: Stripe = 3 測試網卡聚合數量

4.3.1.2 切割磁碟等量區對效能的影響差異

由前面的測試結果，我們得知了透過調整磁碟等量數目的原則來增加虛擬磁碟存放的資料節點主機時，在網路速度可以滿足資料吞吐量的同時可以提升資料讀寫的 I/O 效能。但這是否代表在主機數量足夠的同時，將虛擬磁碟切割的愈多份存放在多台主機上就可以不斷的提升資料讀寫效能呢？

我們實際測試了調整等量區的數量從 1 至 10，從 Figure 4.5 中可以得知在測試 4KB 區塊讀取效能的時候，除了在資料節點數目 1 時效能高於一般內接式磁碟 21%，但資料節點提升到 2 到 10 的時候，效能只提升了約 3% 就停住成長的趨勢，因此我們可以得知如果在小區塊的檔案讀取上面，增加虛擬磁碟存放的資料節點數量並沒太大意義。

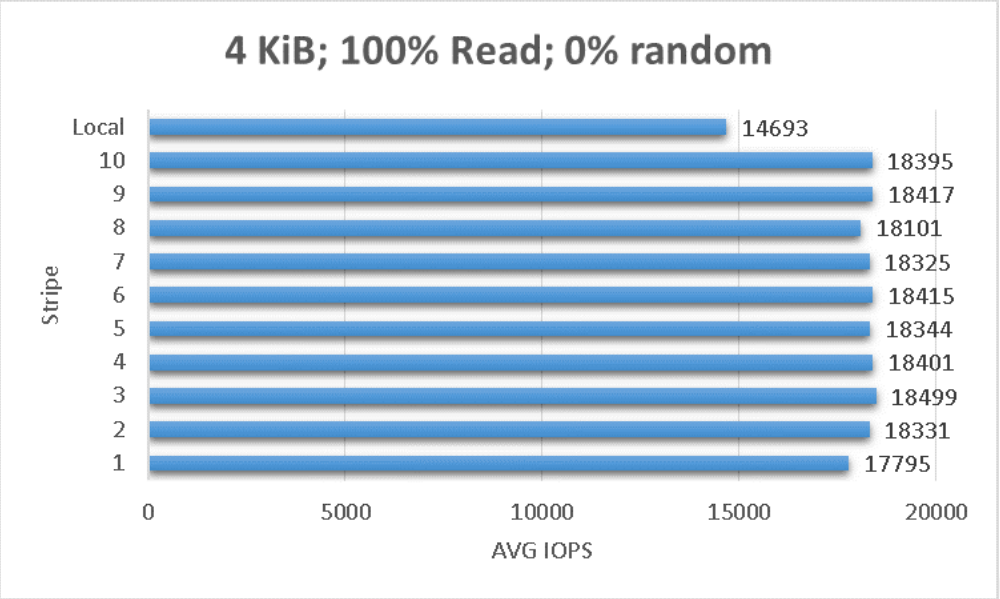

FIGURE 4.5: IOMeter 4KB 測試磁碟等量數量

接下來我們進一步測試資料區塊提升至 256KB 的讀取效能，從 Figure 4.6 中可以發現效能的提升在某些情況下可能並不一定顯著，可能的原因有資料的讀寫是來自於該資料節點中的快閃記憶體或是機械式磁碟，兩者的讀寫效能自然並不一致；但可以得知的是在資料區塊較大的時候，增加虛擬磁碟存放的資料節點數量對於資料 I/O 的效能是有所幫助的，如 Stripe 為 10 時就比 Stripe 為 1 時，快了約 2.4 倍的 IOPS。

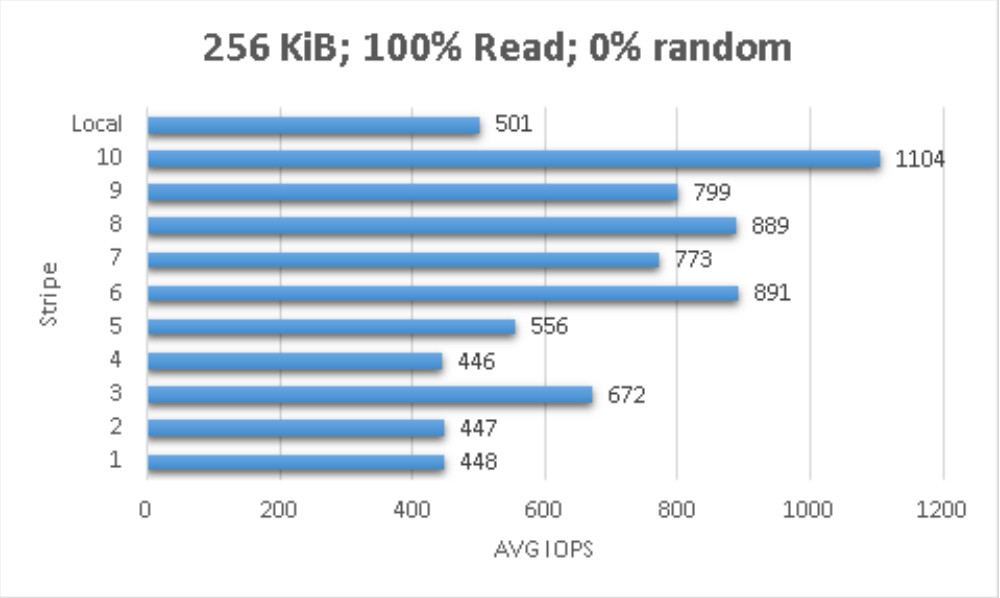

FIGURE 4.6: IOMeter 256KB 測試磁碟等量數量

4.3.1.3 實際佈署大量虛擬機測試

接下來我們實際用生產環境中見的大量虛擬機器佈署來測試傳統 NAS 與 vSAN 分散式儲存環境的效能差異，我們使用同樣支援 vSAN 分散式儲存的 Horizon View 6.1 版本來一次產生 60 部虛擬桌面環境，並從前面 Figure 4.6 測試切割磁碟等量數量中測試效能較佳的資料節點 3、資料節點 6、資料節點 10 的方式來分別比較其完成時間。測試出來的結果請參考 Figure 4.7，而由測試結果中可以得知，在資料節點數量 10 時比傳統 NAS 透過 iSCSI 方式連接快了約 1.9 倍，資料節點 6 時也比傳統 NAS 快了 1.6 倍，但資料節點僅有 3 時反而由傳統 NAS 快了 1.3 倍。

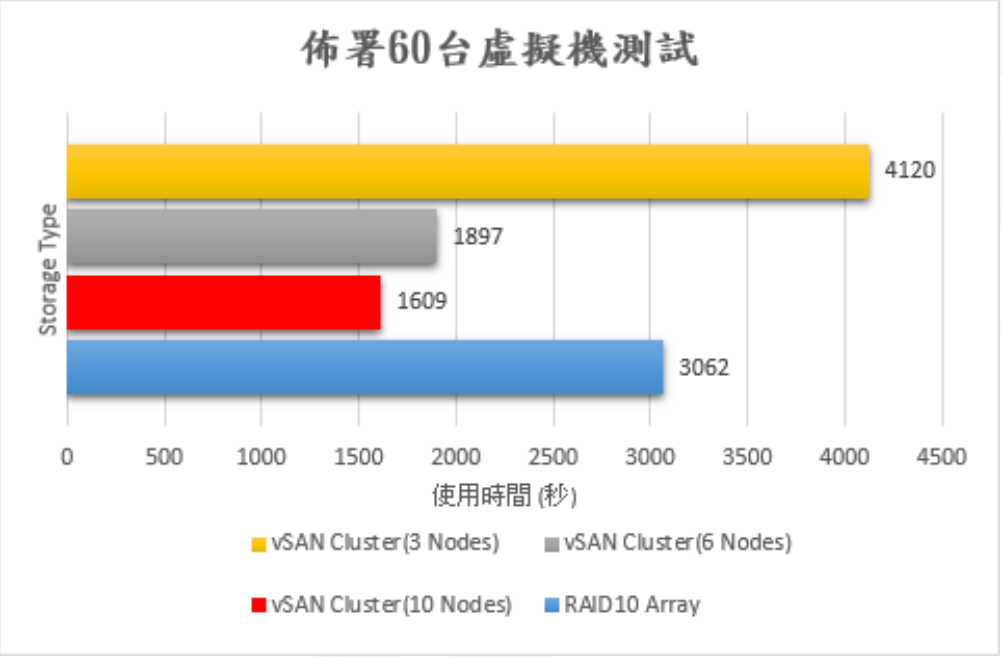

FIGURE 4.7: 傳統 NAS 與 vSAN 分散式儲存效能比較

因此為了瞭解在資料節點 3 時效能不佳的主要原因，我們將虛擬桌面環境的佈署流程中，幾個實際影響完成時間的關鍵因素表列出來，請參考 Figure 4.8。

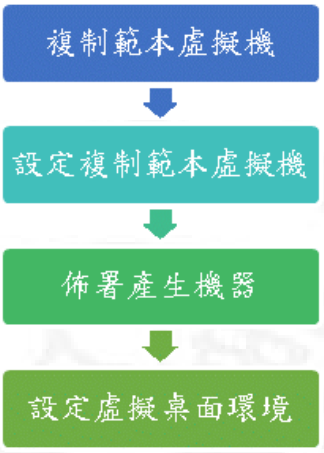

FIGURE 4.8: 虛擬桌面環境（VDI）佈署主要流程

由 Table 4.6 中可以更進一步得知，在單獨產生虛擬機器的時間上，資料節點 10 快了傳統 NAS 約 2.57 倍，資料節點 6 時快了傳統 NAS 約 2.27 倍，資料節點 3 時也快了傳統 NAS 約 1.82 倍，但在安裝及設定虛擬機中的桌面環境時，此時過多的資料 I/O 讀寫要求，已超過快閃記憶體的 Cache 容量，造成頻繁的實體機械式硬碟做讀寫動作，在僅有三個資料節點的情況下，反而慢於傳統 NAS 由 10 顆硬碟組成的 RAID 10 磁碟陣列。

TABLE 4.6: 佈署虛擬桌面完成時間表 (秒)

Storage Type	Copy	Setting	Deploy	Setup
NAS RAID10	368	188	827	1679
vSAN(10Node)	479	205	322	603
vSAN(6Node)	394	222	364	917
vSAN(3Node)	532	216	453	2919

4.3.2 異質網路環境實驗結果

接下來我們更換測試環境，重新進行異質網路環境的測試。在實驗方法中已經有提到因為 Mellanox 40G QSFP+ 的網路卡，在 ESXi Hypervisor 的底層作業協定中，在我們調整最佳化過後也僅能跑到 22GBps 左右，因此接下來我們使用 VMware Virtual SAN Health Check Plugin 做測試時，也可以檢視 Virtual SAN 分散式儲存的資料讀寫效能是否也可以達到二倍於 10G 的成長幅度。

我們先做資料讀寫最為頻繁的壓力測試（Stress Test），此種測試方法將會做多重的資料同時讀寫動作，並在每台測試主機產生 1TB 的測試檔。從 Figure 4.9 中可以看出如果以 10G 的網路介面做為比較基準的話，在更換成 Mellanox 40G QSFP+ 的網路卡後，的確在磁碟資料的 IOPS（Input/Output Operations Per Second）讀寫及資料的吞吐量 Throughput 上面，有明顯的效能提升效果；而在平均等待時間（Average Latency）上面，也有明顯的降低。

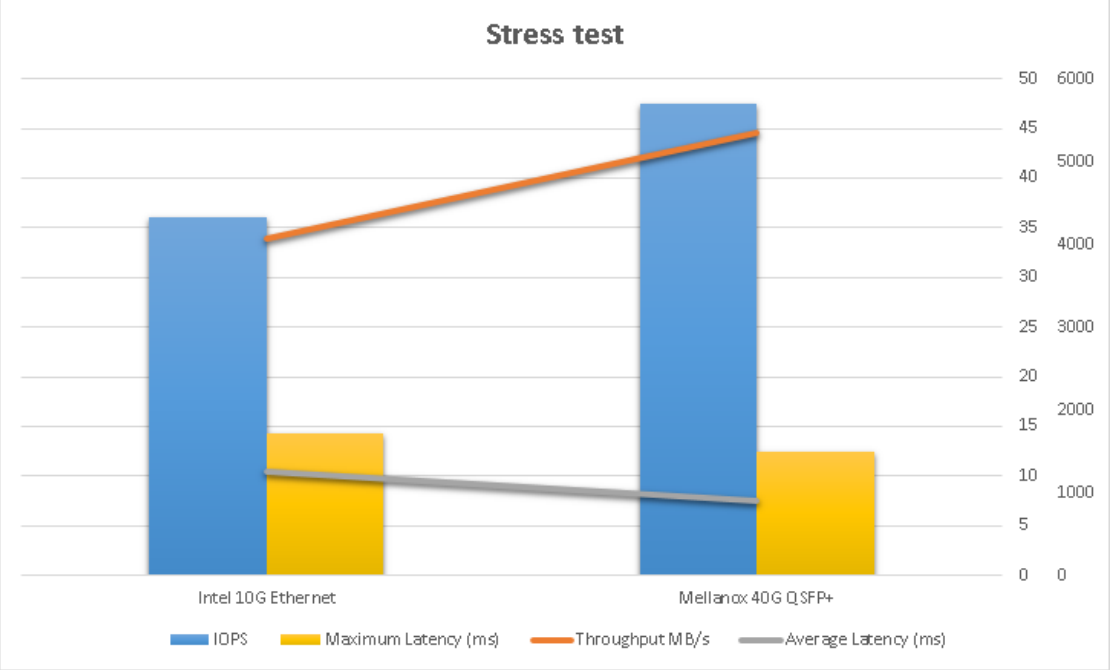

FIGURE 4.9: 10G 和 40G 網路介面的資料讀寫壓力測試

接下來我們進一步跑完所有的測試，如 Figure 4.10 中包含 Basic sanity test, Stress Test, Performance Characterization - 100% Read, Performance Characterization - 100% Write, Performance Characterization - 70/30 Read/Write, high I/O size；我們可以進一步得到一個結論，如果是在基本測試 Basic sanity test 中，此種測試方法模擬常見的工作狀態，並在每台測試主機產生 1GB 的測試檔，分別同時做 70% 的讀取以及 30% 的寫入動作。而在測試結果裡面，相對於 10G 的 IOPS 數值，在更換成 Mellanox 40G QSFP+ 後，Virtual SAN 磁碟的效能成長了約 17%。

而在 Performance Characterization - 100% Read & Write 中，因為多是使用快取緩存以及寫入緩存的關係，資料的交換並不頻繁的結果反而效能成長幅

度僅約 10%。而在大區塊資料讀寫測試中 Performance Characterization - 70/30 Read/Write high I/O size，此種測試方法將使用 64k 的 Block 大小來做測試，並同時做 70% 的讀取及 30% 的寫入動作，並在每台測試主機產生 30GB 的測試檔。測試結果 Virtual SAN 效能成長則是 13%。在壓力測試 Stress Test 裡面，整體效能成長幅度是最大的，大量資料的同時讀寫和多工並行的執行序，讓整體的 IOPS 效能有明顯的 32% 成長。

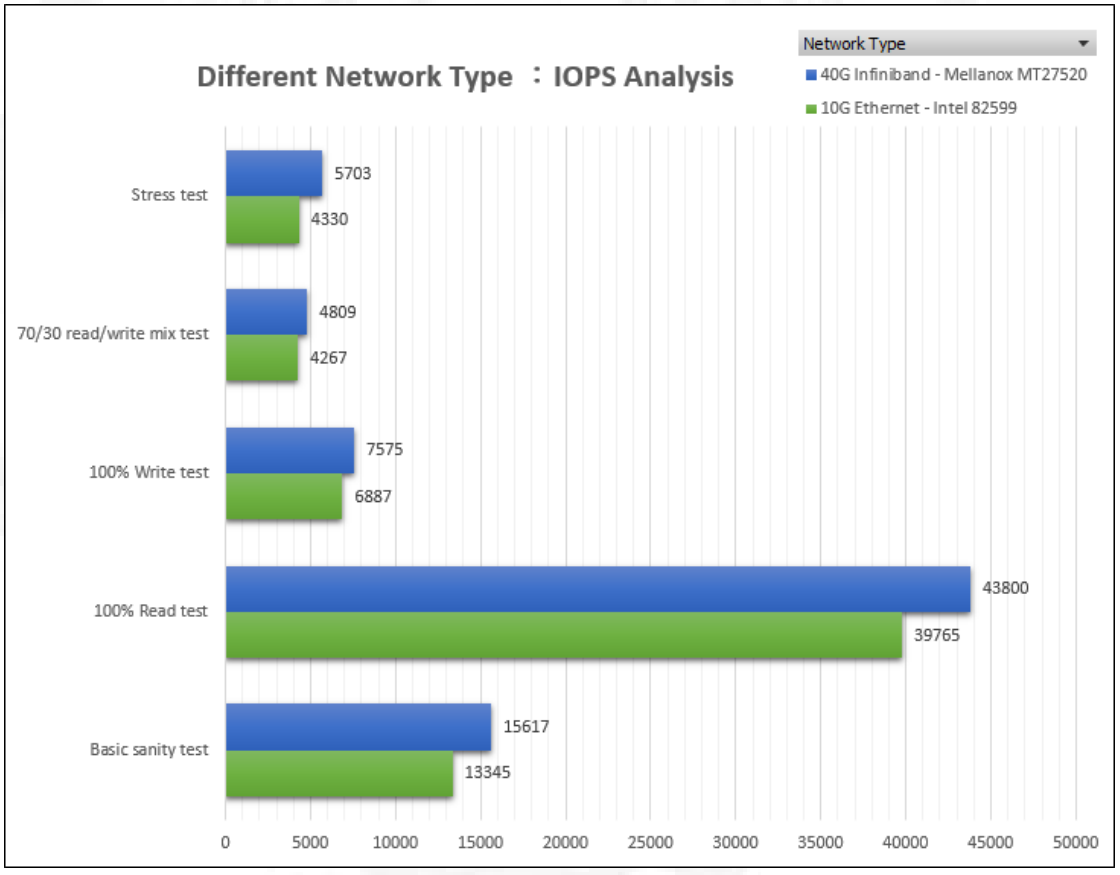

FIGURE 4.10: 10G 和 40G 網路介面的 IOPS 測試結果

在資料吞吐量（Throughput）上面，如果是在較大區塊（64KB）的傳輸上面，透過 10G 的網路介面最高可以來到 267MB，如果是 Mellanox 40G 的光纖網路卡則可以來到 301MB；其他在讀取上面也有不錯的效能表現。不過如果是在寫入及壓力測試上面，資料吞吐量（Throughput）的數值並不高，這有可能是因為在 Virtual SAN 預設的設定中，70% 的快閃式裝置 SSD 空間皆是用來做讀取快取，而僅有 30% 的空間是拿來做寫入緩存。

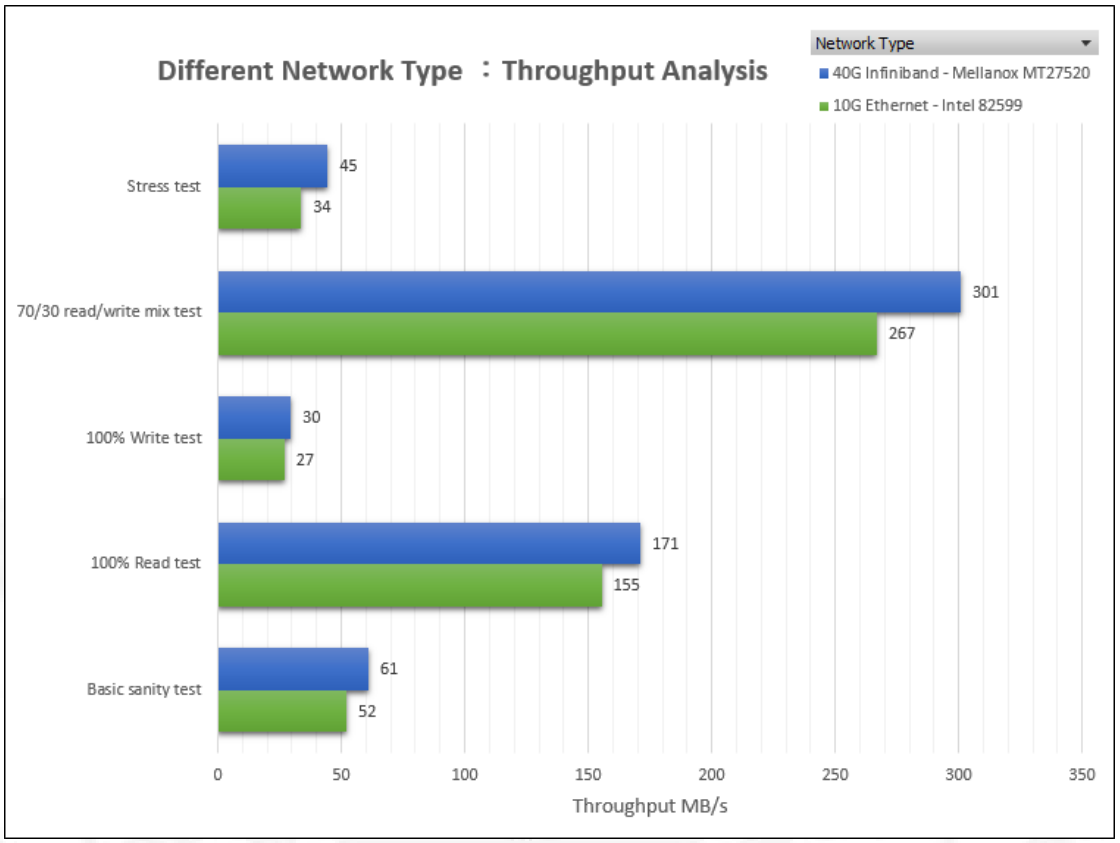

FIGURE 4.11: 10G 和 40G 網路介面的 Throughput 測試結果

而在延遲時間 Latency 上面，也是反映出明顯的降低整體的回應時間。如 Figure 4.12 中，延遲時間的差異由其是在讀寫都承受極大壓力的壓力測試中相差最大。

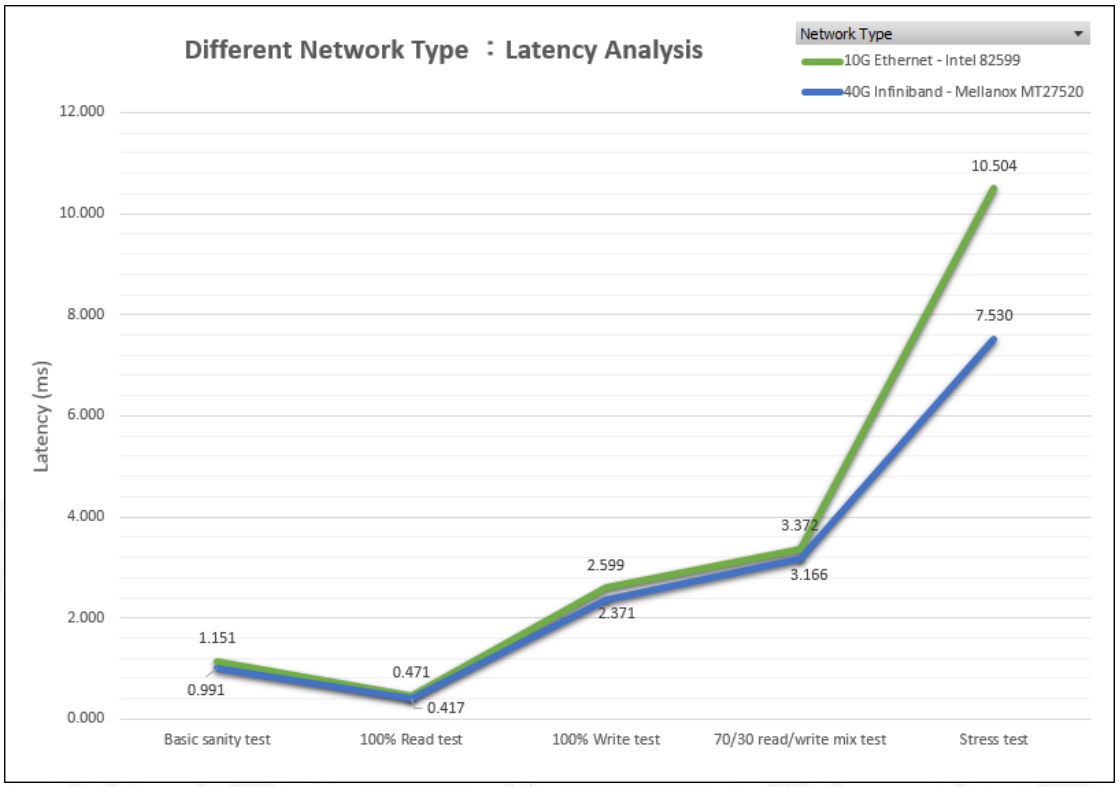

FIGURE 4.12: 10G 和 40G 網路介面的 Latency 測試結果

4.3.3 全快閃式儲存環境實驗結果

在 Virtual SAN 1.0 版分散式儲存環境中，僅有一種 Hybrid 架構就是由第一層使用快閃式裝置 SSD 做為 Cache Tier，後端的資料節點第二層 Capacity Tier 使用的是傳統的機械式硬碟。但在 Virtual SAN 從 1.0 更新到 6.0 版後，則多了一種 All-Flash 的架構，他將第一層、第二層前後端的儲存裝置皆改為快閃式裝置 SSD，在理論上此種架構能夠解決傳統硬碟資料讀寫時所碰到的 IO 瓶頸。

所以我們首先重新搭建了一個四個節點的測試環境，並在每個資料節點使用二顆快閃式裝置 SSD，一個做為 Cache Tier，一個做為 Capacity Tier；而在另一個做為比對的 Hybrid 環境，我們使用同樣的硬體規格，但後端的 Capacity Tier 改為傳統的 7200RPM SATA 硬碟。我們先測試 Basic sanity test, 100% Read optimal RC usage, 100% Write optimal WB usage, 100% Read Low RC hit rate / All-Flash，這五個部分，測試結果如 Figure 4.13。

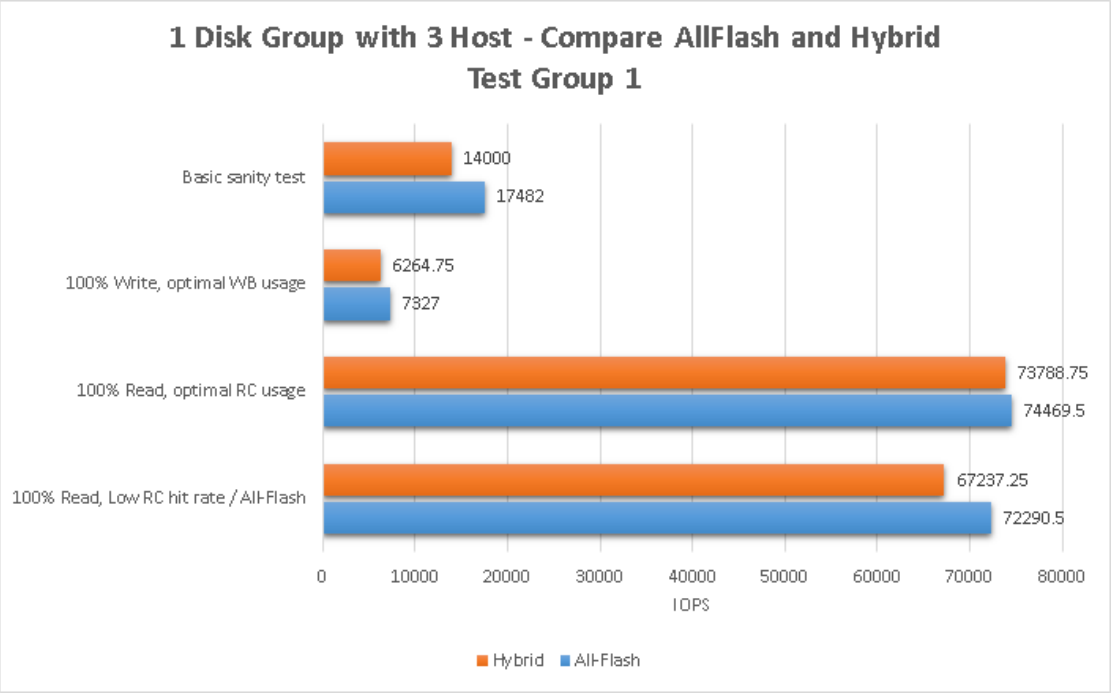

FIGURE 4.13: 1DiskGroup with 3 Host 測試結果 1

結果僅在 Basic Sanity Test 項目中 All-Flash 的架構比 Hybrid 的架構成長了 25% 的效能，而在 100% Read optimal RC usage 中僅成長了 1% 的效能，而 100% Read Low RC hit rate / All-Flash 的測試，此種測試方法將 100% 使用快閃式裝置做測試，並在每台測試主機產生 1TB 的測試檔，此種優先使用快閃式裝置的測試方式，Virtual SAN 效能則成長了 8% 的效能；這表示如果測試項目集中在資料的讀取上，且測試資料優先來於第一層的 Cache Tier 時，因為同樣都是使用快閃式裝置 SSD，所以效能的成長其實並不明顯。但如果是在 100% Write optimal WB usage 的項目中，因為寫入第二層 Capacity Tier 裝置的不同，所以寫入速度可以提高 17% 的效能。但這樣的效能成長並如我們所預期一般，因此我們又在做了 Stress test, 70/30 Read/Write realistic, 70/30 Read/Write high I/O size, 100% Streaming Reads, 100% Streaming Write 這五項的測試，測試結果如 Figure 4.14。

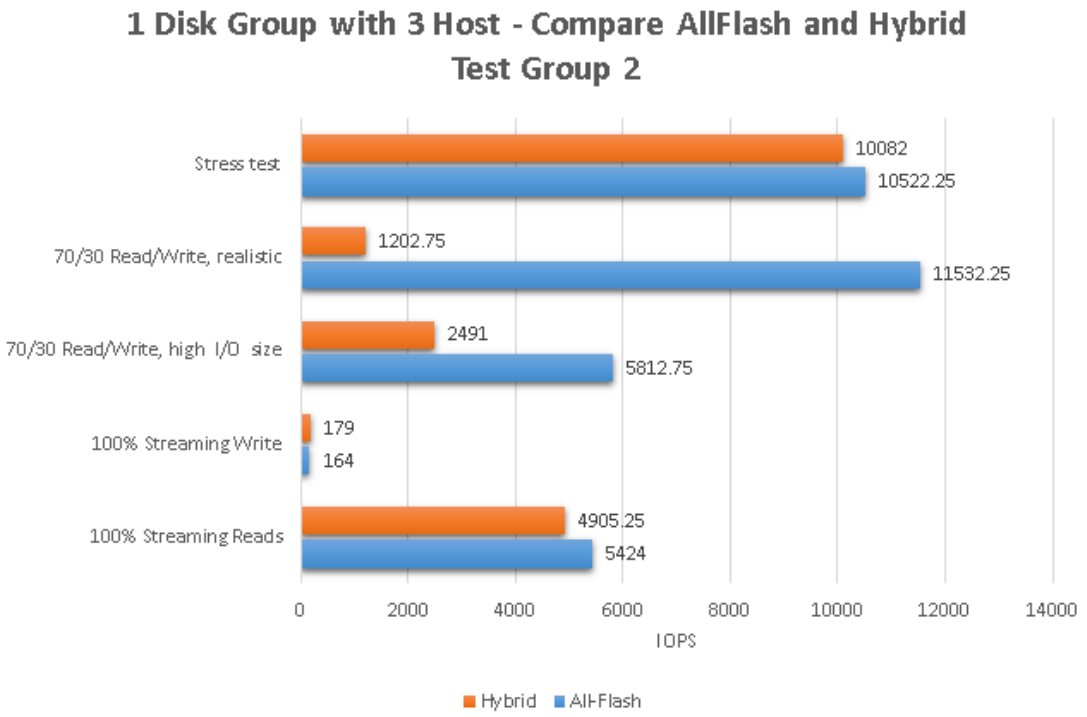

FIGURE 4.14: 1DiskGroup with 3 Host 測試結果 2

其中我們發現了 70/30 Read/Write realistic 以及 70/30 Read/Write high I/O size，這二個項目中，資料讀寫的效能成長分別高達 859% 以及 133%，這二種測試方法的差異僅在於測試資料切割的大小，前者使用 4K Block Size，而後者使用 64KB Block Size，且同樣都在每台主機產生 30GB 的測試檔案；這表示在 4K Block 區塊的資料交換中，快閃式裝置的加速效能遠遠高於傳統機械式硬碟！但在 100% Streaming Write 持續性資料寫入的測試中，此種測試方法將會模擬一個持續性寫入的服務，像是備份伺服器；並在每台測試主機產生 1TB 的測試檔。All-Flash 架構卻發生了負成長，小輸了 8% 於 Hybrid 架構。測試結果效能反而降低，推斷可能是因為我們所使用的快閃式裝置單顆大小僅 480GB，還小於此測試項目所產生的 1TB 測試檔案，所以因為資料節點的分割存放影響了測試的效能結果。

以上我們已經知道了在同樣資料節點下，全快閃式架構相對於 Hybrid 架構的效能成長幅度。接下來要更進一步的來測試對於 Virtual SAN 分散式儲存的「垂直擴充（Scale Up）」或「水平擴充（Scale Out）」，何者的擴充方式是對於資料讀寫的效能成長是較有幫助的。

我們先以基本的三個資料節點配合一個磁碟群組（DG）做基本對照，然後逐步增加「水平擴充（Scale Out）」到四個資料節點配合一個磁碟群組，最後在增加「垂直擴充（Scale Up）」的部分到四個資料節點配合二個磁碟群組。我們一樣執行了八種測試模式，並分二個部分做分析。

如 Figure 4.15 中所示，可以看到在所有的測試當中，如果是做「水平擴充（Scale Out）」則效能則都是線性成長，由其是在強調寫入測試的模型，效能提升幅度更是較其他讀取的測試更來的高達 69% 之多，但在「垂直擴充（Scale Up）」的部分，在部分測試模型項目中效能提升的情況並不明顯，由其是在 Stress 壓力測試中，更是出現了些微的負成長。

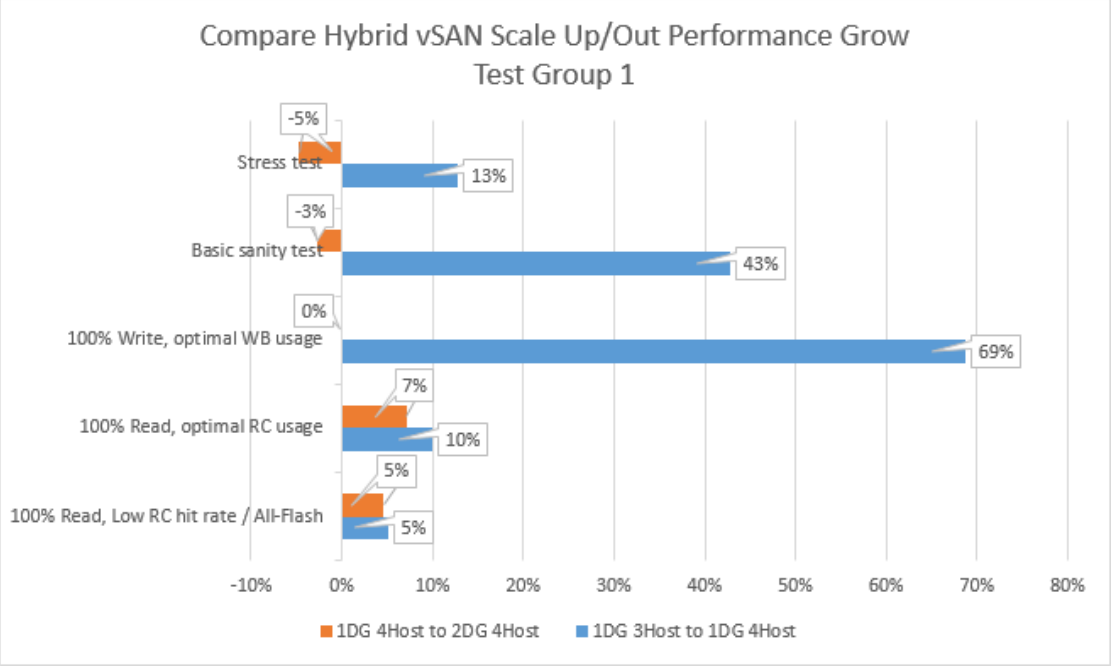

FIGURE 4.15: Compare Hybird vSAN Scale Up/Out 測試結果 1

那是否「垂直擴充（Scale Up）」的部分對效能就沒有幫助了嗎？我們再做了第二組測試，如 Figure 4.16 所示。第二組測試跟第一組測試的差別是，第一組測試中主要集中在專門的讀取及寫入，並集中在前端 Cache Tier 的使用上；而第二組比較偏向常見的服務應用，如持續性的讀取像是媒體伺服器、持續性的寫入像是備份伺服器，而也分析了 4KB Block Size 和 64KB Block Size 的讀寫效能。

我們得到的結果迥異於第一組的結果，除了在持續性的讀取上面效能仍差於「水平擴充（Scale Out）」外，其他的部分資料讀寫的效能，「垂直擴充（Scale Up）」皆優於「水平擴充（Scale Out）」。在 4KB Block Size 的測試中更是一次上升了 480% 之多，這也表示了在磁碟群組增加的情況下，對於資料的讀寫效能是有所幫助的。因為前端 Cache Tier 的使用率也會提升。

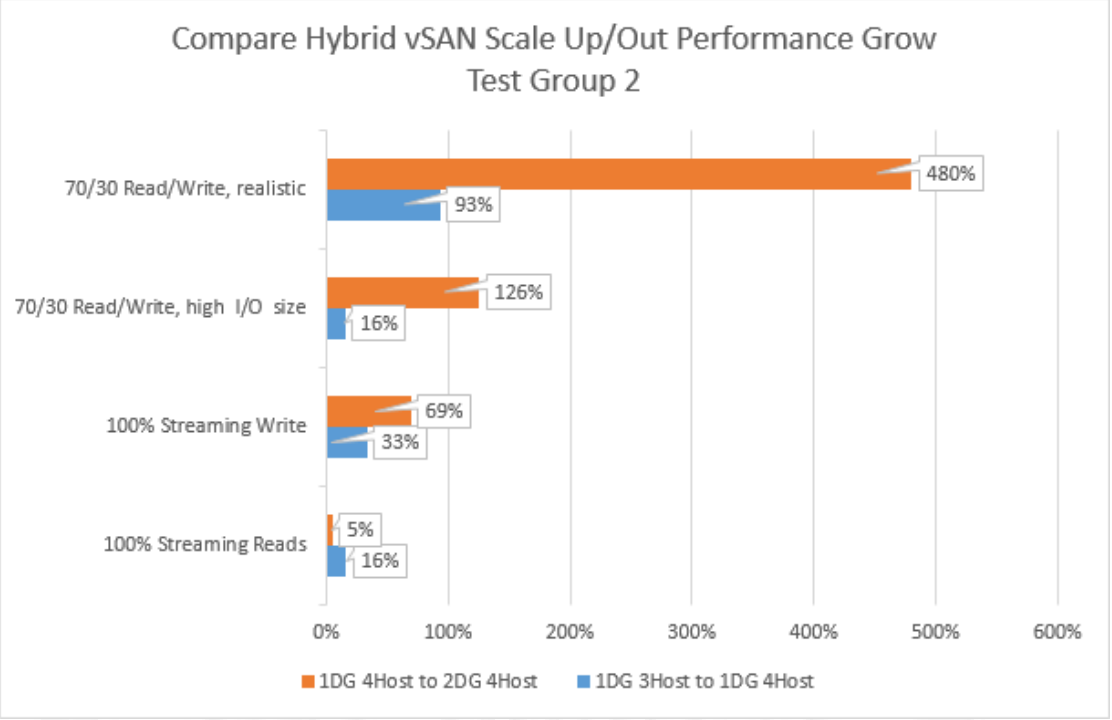

FIGURE 4.16: Compare Hybird vSAN Scale Up/Out 測試結果 2

最後我們比對使用一般的千兆位元網路與 Mellanox 40G QSFP+ 做比對測試，如 Figure 4.17，可以發現同樣都是全閃存的架構下，二者的 IOPS 資料讀寫效能僅差距了 22 倍，這也進一步驗證了我們前面所做的網路效能測試。在 ESXi Hypervisor 底下，Virtual SAN 並無法完全配合多核心並行處理多個處理序的方式，所以 Mellanox 40G QSFP+ 的光纖網路卡，在此時僅能使用到 22GBps 左右的網路頻寬，雖然網路卡本身還有足夠的空間可以使用，但 Virtual SAN 融合嵌入 Hypervisor 底層後，也只能使用到 22Gbps 的網路頻寬，不然我們的測試數值應該還有往上提升的空間。

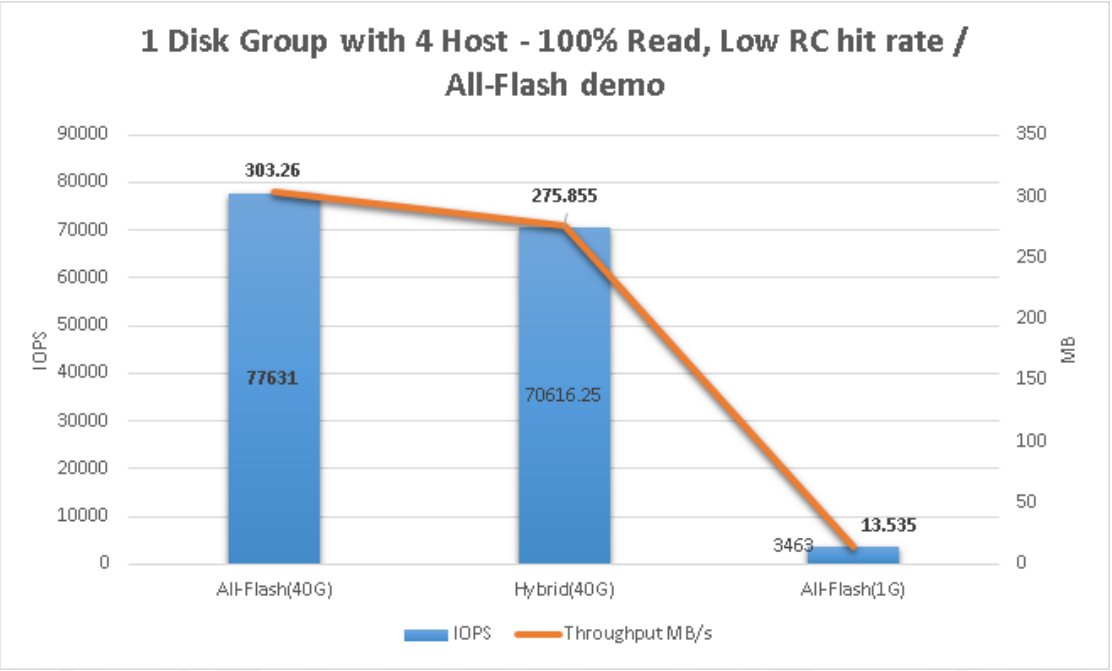

FIGURE 4.17: Compare 1GbE Network Interface Performance

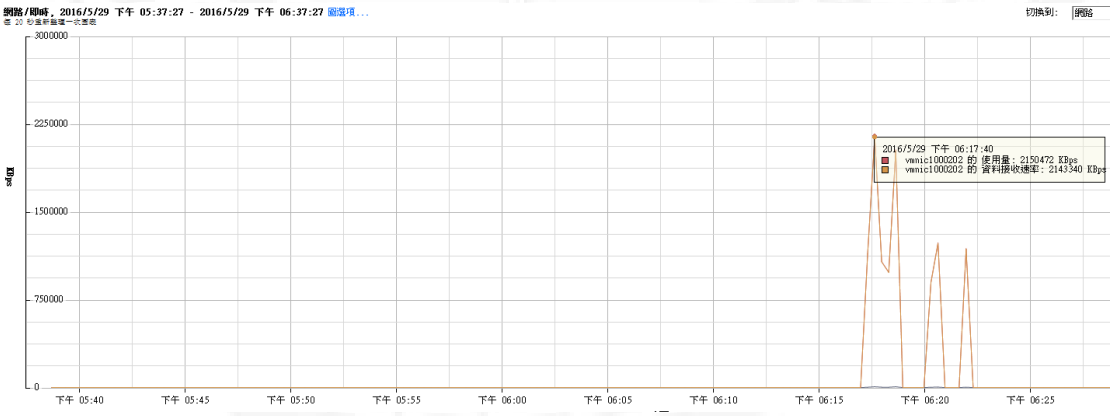

FIGURE 4.18: 網路頻寬利用率

Chapter 5

結論與未來方向

5.1 結論

雲端服務普及的現在，愈來愈多的公有雲、私有雲以及混合雲的架構產生，而這些雲端服務背後的底層多是使用虛擬化的技術做為平台，在關鍵核心服務集中化的同時，這些重要的服務更是需要使用愈來愈多的硬體資源，其中對於影響服務效能最為重要的指標就是資料的儲存架構。由其是在資訊平台的高可用性上面，除了需要做到自動備援、自動修復外，在儲存設備的擴充、維護更換上面更需要朝向「垂直擴充（Scale Up）」或「水平擴充（Scale Out）」都能支援在線執行。

而應用在虛擬桌面佈署的環境裡面，Virtual SAN 的虛擬機的佈署速度在資料節點數量 10 時，比傳統 NAS 做 RAID10 的磁碟陣列並透過 iSCSI 連接方式更快了約 1.9 倍。而在佈署 Virtual SAN 的成本和選擇高單價的 NAS 儲存櫃上，Virtual SAN 分散式儲存架構更可以有效的降低企業整體持有成本及有效提升虛擬桌面相關應用的可行性。

另外我們研究結果中也發現，可以提供在生產環境實務使用以及管理 Virtual SAN 分散式儲存系統管理人員的一份參考資料；如果是應用 Virtual SAN 在複雜且大量虛擬機資料讀寫的環境裡面，那麼在做系統環境擴充的時候

選用「水平擴充（Scale Out）」的方式會得到較佳的效能提升。而如果是在應用多媒體伺服器資料存放或是應用在檔案伺服器存放的 Virtual SAN 應用環境中，那麼選用「垂直擴充（Scale Up）」的方式則會得到更好的效能提升。

本論文提出了以 Virtual SAN 分散式儲存架構做為後端的資料儲存平台，除了能夠結合底層的虛擬化架構外，更能透過 Software-Defined Policy 的方式支援多重的虛擬機資料保護，提高虛擬機可用性的同時也能分散磁碟資料讀寫的效能瓶頸，進而做到軟體定義資料中心的實際應用。而我們研究中也進一步提出以 Mellanox 40G QSFP+ 的高速光纖網路介面，應用在 Virtual SAN 分散式儲存架構，並透過調整及優化 Mellanox 40G QSFP+ 光纖網路介面後，可以得到超出使用原先原廠所建議的 10G Ethernet 網路規格更好的資料傳輸讀寫效能，所以應用在生產環境中的 Virtual SAN 分散式儲存架構，其實使用更高速度的 40G QSFP+ 網路卡會是更好的選擇。

5.2 未來方向

透過 Virtual SAN 分散式儲存環境，可以有效的解決過去本機的儲存資源無法妥善利用的問題，也可以降低建置虛擬化環境平台的成本與預算，透過我們的研究可以知道在網路頻寬設備速度提升的同時，也能提升分散式儲存環境的資料讀寫效能。但目前 Mellanox 40G QSFP+ 的卡在虛擬化平台 Native 環境下的支援度仍不高，Virtual SAN 僅能利用約一半的有效網路頻寬，未來如能透過更新 Virtual SAN 版本或網路卡韌體優化的方式提升有效使用頻寬，將可以使 Virtual SAN 分散式儲存環境的整體效能及可靠度再度提升。

參考文獻

[1] J. Li, D. Zhou, and H. Dong. Research of the improved identity authentication system based on pki in virtual san. In *Internet Technology and Applications, 2010 International Conference on*, pages 1–4, 2010.

[2] Y. Deng, F. Wang, K. Zhou, and S. Wu. Virtual storage image implementation in a san system to improve storage capacity, fault tolerance and bandwidth. In *INTERMAG 2006 - IEEE International Magnetics Conference*, pages 611–611, 2006.

[3] X. Su, M. Wu, and J. Xu. A novel virtual storage area network solution for virtual desktop infrastructure. In *2014 International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pages 204–208, 2014.

[4] D. Guo, G. Liao, and L. N. Bhuyan. Performance characterization and cache-aware core scheduling in a virtualized multi-core server under 10gbe. In *Workload Characterization, 2009. IISWC 2009. IEEE International Symposium on*, pages 168–177, 2009.

[5] Y. L. Chen, C. T. Yang, S. T. Chen, K. C. Chang, and W. C. C. Chu. Environment virtualized distributed storage system deployment and effectiveness analysis. In *Trustworthy Systems and Their Applications (TSA), 2015 Second International Conference on*, pages 94–99, 2015.

[6] P. Wang, R. E. Gilligan, H. Green, and J. Raubitschek. Ip san - from iscsi to ip-addressable ethernet disks. In *Mass Storage Systems and Technologies,*

2003. (MSST 2003). Proceedings. 20th IEEE/11th NASA Goddard Conference on, pages 189–193, 2003.

[7] Junfei Wang, Jiwu Shu, and Bigang Li. San-mds: A high performance disk based on memory device for san. In *International Conference on Autonomic and Autonomous Systems (ICAS'06)*, pages 27–27, 2006.

[8] Yue Jiang, Hai Jin, and Xiaofei Liao. Devstore: Distributed storage system for desktop virtualization. In *Computer Science Education (ICCSE), 2013 8th International Conference on*, pages 341–346, 2013.

[9] V. Lasky. Implementing resilient remote laboratories with server virtualization and live migration. In *Remote Engineering and Virtual Instrumentation (REV), 2013 10th International Conference on*, pages 1–6, 2013.

[10] M. Roohitavaf, R. Entezari-Maleki, and A. Movaghar. Availability modeling and evaluation of cloud virtual data centers. In *Parallel and Distributed Systems (ICPADS), 2013 International Conference on*, pages 675–680, 2013.

[11] T. J. Liu, Chun-Yan Chung, and Chia-Lin Lee. A high performance and low cost distributed file system. In *2011 IEEE 2nd International Conference on Software Engineering and Service Science*, pages 47–50, 2011.

[12] S. Sasaki, R. Matsumiya, K. Takahashi, Y. Oyama, and O. Tatebe. Rdma-based cooperative caching for a distributed file system. In *Parallel and Distributed Systems (ICPADS), 2015 IEEE 21st International Conference on*, pages 344–353, 2015.

[13] C. Akinlar and S. Mukherjee. A scalable bandwidth guaranteed distributed continuous media file system using network attached autonomous disks. *IEEE Transactions on Multimedia*, 5(1):71–96, 2003.

[14] S. C. Deshmukh and S. S. Deshmukh. Improved load balancing for distributed file system using self acting and adaptive loading data migration process. In

Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on, pages 1–6, 2015.

[15] J. Yu, W. Wu, and H. Li. Dmoosefs: Design and implementation of distributed files system with distributed metadata server. In *Cloud Computing Congress (APCloudCC), 2012 IEEE Asia Pacific*, pages 42–47, 2012.

[16] X. Zhou and L. He. A virtualized hybrid distributed file system. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference on*, pages 202–205, 2013.

[17] A. Glagoleva and A. Sathaye. A load balancing tool based on mining access patterns for distributed file system servers. In *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*, pages 1248–1255, 2002.

[18] T. Zhao, Z. Zhang, and X. Ao. Application performance analysis of distributed file systems under cloud computing environment. In *Information Science and Control Engineering (ICISCE), 2015 2nd International Conference on*, pages 152–155, 2015.

[19] T. L. S. R. Krishna, T. Ragunathan, and S. K. Battula. Improving the performance of read operations in distributed file system. In *Computational Intelligence and Communication Networks (CICN), 2014 International Conference on*, pages 1126–1130, 2014.

[20] C. M. Wang, C. C. Huang, and H. M. Liang. Asdf: An autonomous and scalable distributed file system. In *Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on*, pages 485–493, 2011.

[21] Y. Gong, Y. Xu, Y. Lei, and W. Wang. Varfs: A variable-sized objects based distributed file system. In *High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and*

Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on, pages 148–153, 2015.

[22] P. Chrobak. Implementation of virtual desktop infrastructure in academic laboratories. In *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*, pages 1139–1146, 2014.

[23] Mohiuddin Solaimani, Mohammed Iftekhar, Latifur Khan, and Bhavani M. Thuraisingham. Statistical technique for online anomaly detection using spark over heterogeneous data from multi-source vmware performance data. In *2014 IEEE International Conference on Big Data, Big Data 2014, Washington, DC, USA, October 27-30, 2014*, pages 1086–1094, 2014.

[24] Deng Liu, Ningfang Mi, Jianzhe Tai, Xiaoyun Zhu, and Jack Lo. VFRM: flash resource manager in vmware ESX server. In *2014 IEEE Network Operations and Management Symposium, NOMS 2014, Krakow, Poland, May 5-9, 2014*, pages 1–7, 2014.

[25] F. Wang, Y. Liu, B. Lei, and J. Li. Benchmark driven virtual desktop planning: A case study from telecom operator. In *Cloud and Service Computing (CSC), 2012 International Conference on*, pages 204–211, 2012.

[26] D. H. Tran, T. D. Nguyen, E. N. Huh, and C. S. Hong. A performance comparison of in-memory virtual desktop environment. In *Network Operations and Management Symposium (APNOMS), 2014 16th Asia-Pacific*, pages 1–4, 2014.

[27] Trung Nguyen, Prasad Calyam, and Ronny Bazan Antequera. Benchmarking in virtual desktops for end-to-end performance traceability. In *IFIP/IEEE International Symposium on Integrated Network Management, IM 2015, Ottawa, ON, Canada, 11-15 May, 2015*, pages 1268–1273, 2015.

[28] R. S. R. Pasnoori, H. Swapnil, and B. Radhakrishnan. Survey on application level tools for ssd benchmark validation. In *Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on*, pages 341–346, 2013.

[29] W. Xiao, X. Lei, R. Li, N. Park, and D. J. Lilja. Pass: A hybrid storage system for performance-synchronization tradeoffs using ssds. In *2012 IEEE 10th International Symposium on Parallel and Distributed Processing with Applications*, pages 403–410, 2012.

[30] L. Rui and R. Guocan. Design and application of teaching resources network storage system. In *Electrical and Control Engineering (ICECE), 2011 International Conference on*, pages 6478–6481, 2011.

[31] Xuejiao Fang, Jianxi Chen, Feng Ye, Dan Feng, and Jieqiong Li. Introduction of metadata-request queue with immediate response for i/o path optimizations on iscsi-based storage subsystem. In *Networking, Architecture and Storage (NAS), 2015 IEEE International Conference on*, pages 100–105, 2015.

[32] S. Wu, J. Chen, D. Feng, and B. Mao. Implementation and evaluation of the block i/o interface between the iscsi target and the storage device. In *Convergence Information Technology, 2007. International Conference on*, pages 1451–1456, 2007.

[33] Bo Mao, Dan Feng, Suzhen Wu, Jianxi Chen, Lingfang Zeng, and Lei Tian. Raid10l: A high performance raid10 storage architecture based on logging technique. In *Computer Systems Architecture Conference, 2008. ACSAC 2008. 13th Asia-Pacific*, pages 1–8, 2008.

[34] VMware. Vmware virtual san health check plugin guide. `http://www.vmware.com/files/pdf/products/vsan/VMW-GDL-VSAN-Health-Check.pdf`, 2015. [Online; v6.0].

附錄 A

硬體環境準備與安裝

一、英業達 Zion 伺服器改裝

由於我們第一個實驗環境使用的英業達伺服器預設的情況下僅支援一個 SATA 介面硬碟，但建置 Virtual SAN 環境中的資料節點存放伺服器必要條件之一，至少同時需要二個儲存 I/O 介面，以安裝至少一個快閃式磁碟 SSD、一個傳統機械式磁碟 HDD，且伺服器並無提供一般的 SATA 電源接頭，僅提供專用的 4Pin 電源接頭；故我們得先採買 SATA 的分接頭以及 4Pin 的專用接頭，以手工方式自行製作專門的 SATA 一轉二接頭提供給英業達的伺服器使用。

FIGURE A.1: Zion 伺服器專門的 4Pin 電源接頭

FIGURE A.2: 一般 SATA 電源接頭

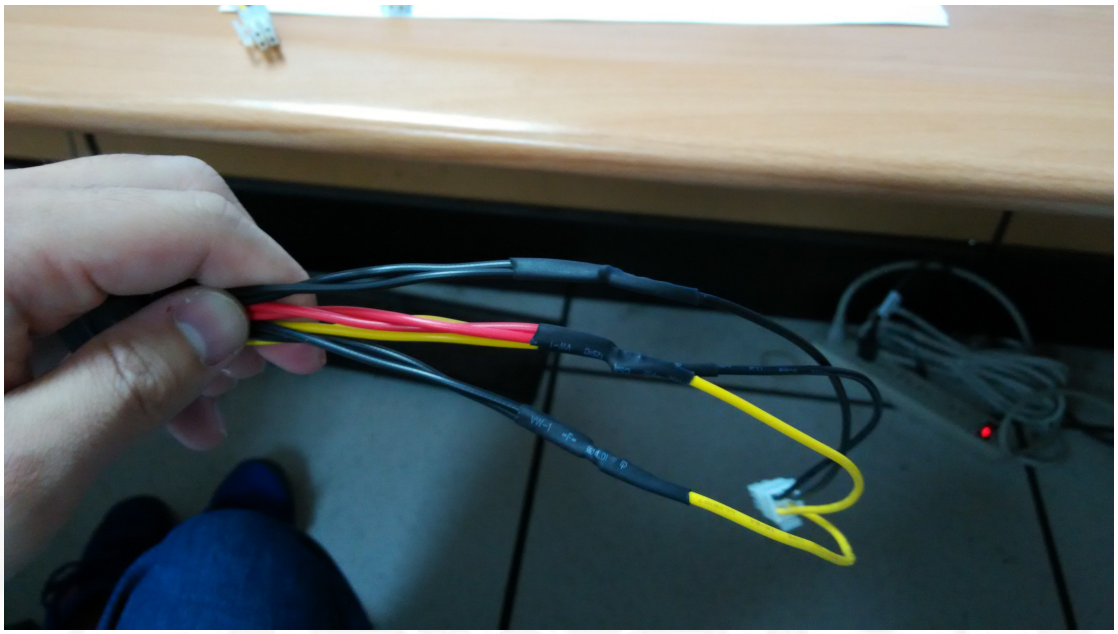

FIGURE A.3: 以熱縮套管合併二個 SATA 電源接頭

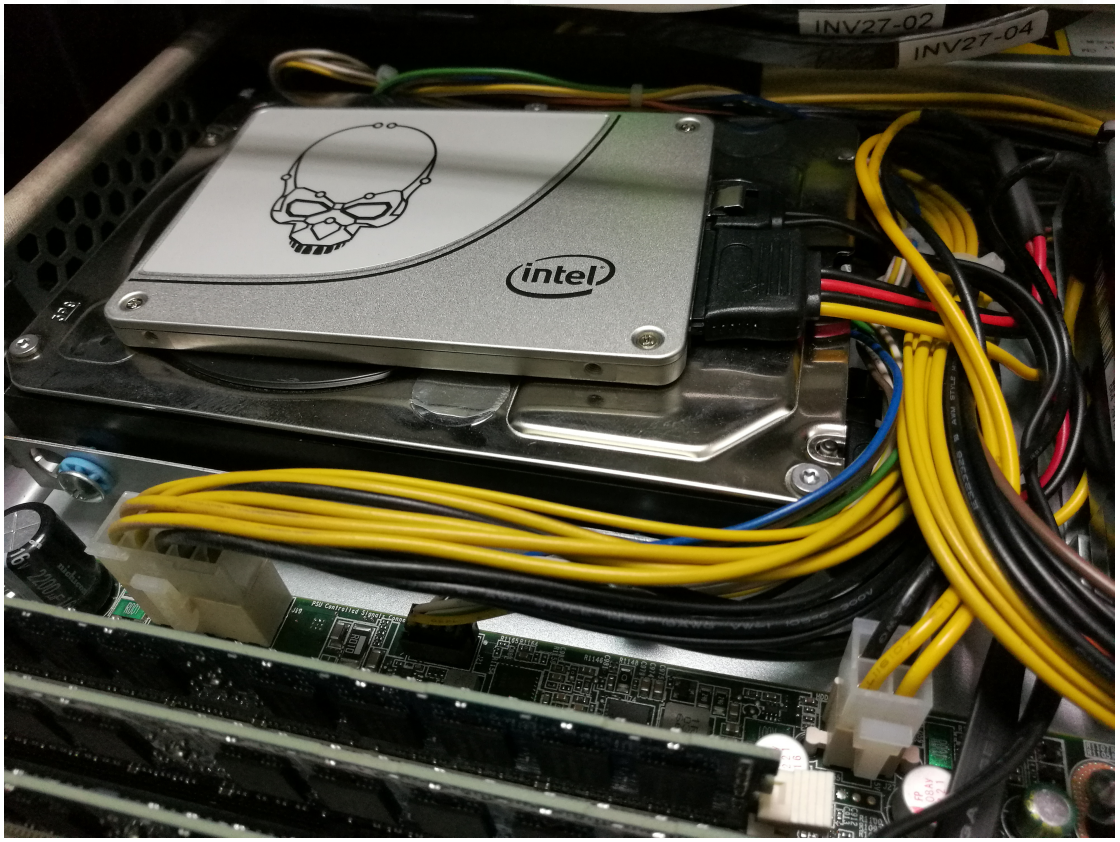

FIGURE A.4: 改裝完成圖

附錄 B

VMware vSAN 環境建置與設定

在安裝虛擬化 vSAN 架構環境之前，我們要先確定要安裝 vSphere 環境的實體伺服器，是否已經滿足下列硬體及運作環境需求： 至少三台 ESXi 主機（需為 5.5 Update 1 版本以上） 需要建置 vCenter 管理主機（可採用 Linux Base vCenter Server Appliance） 每一台 ESXi 主機至少應具備 6GB 以上的記憶體空間 每一台實體主機至少應具備一顆傳統機械式硬碟機（SATA 或 SAS 皆可） 每一台實體主機至少應具備一顆 SSD 快閃記憶體裝置硬碟 Hypervisor ESXi 需安裝在其他的開機媒體上（如另外的硬碟或 USB 碟） Hypervisor ESXi 主機的硬碟控制器，需支援 Pass-Through 或 JBOD 模式 至少需要 1GbE 專用 Ethernet 網路

環境硬體規格說明： 伺服器主機：英業達小型 1U Zion 伺服器 處理器：Intel(R) Xeon(R) CPU E5540 @ 2.53GHz * 2 記憶體：48GB DDR2 儲存裝置：WD 2GB 7200RPM & Intel 730 Serials 480GB 網路介面：Gigabyte Interface * 4

一、佈署 vCenter Server 管理環境

因為我們實驗環境中由於希望降低一般生產環境中 vCenter Server 佈署的複雜度，我們將直接採用 OVF 佈屬的方式直接安裝 Linux Base vCenter Server Appliance 管理主機。先連線到預備部屬 VMware vCenter Server 的 ESXi 主機

上，接著直接從選單中，選取『Deploy OVF Template』並直接依預設值完成導入 OVF Template 步驟；接著將該 vCenter Server 虛擬機 Power ON 開機。

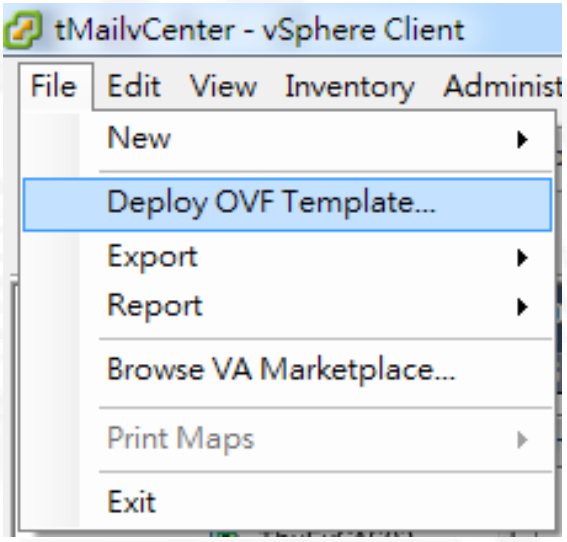

FIGURE B.1: 佈署 OVF 範本 1

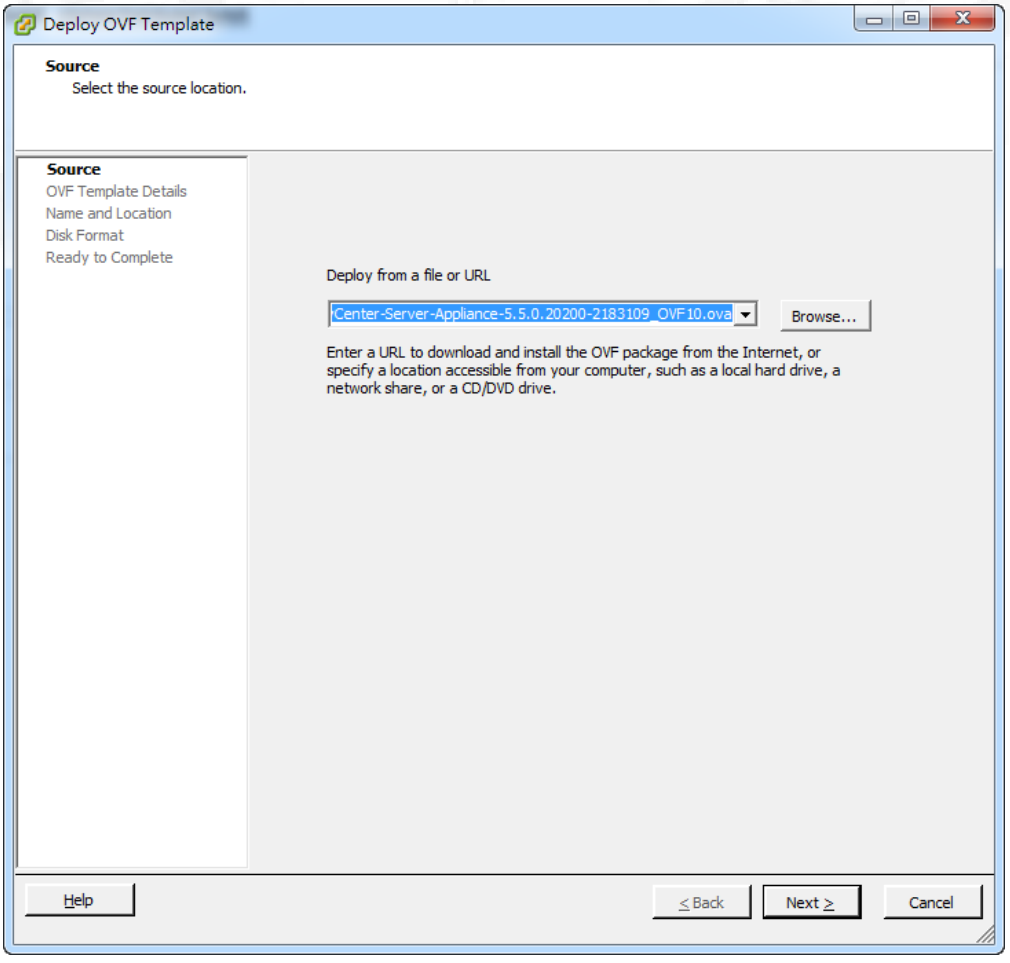

FIGURE B.2: 佈署 OVF 範本 2

由於開機後的 vCenter Server 虛擬機網路預設值是採 DHCP 設定，故如安裝之網路環境無 DHCP 伺服器提供自動分配 IP，則請先打開虛擬機主控台（Open Console），並登入 vCenter Server 作業系統。預設登入帳號為 root，密碼為 vmware。

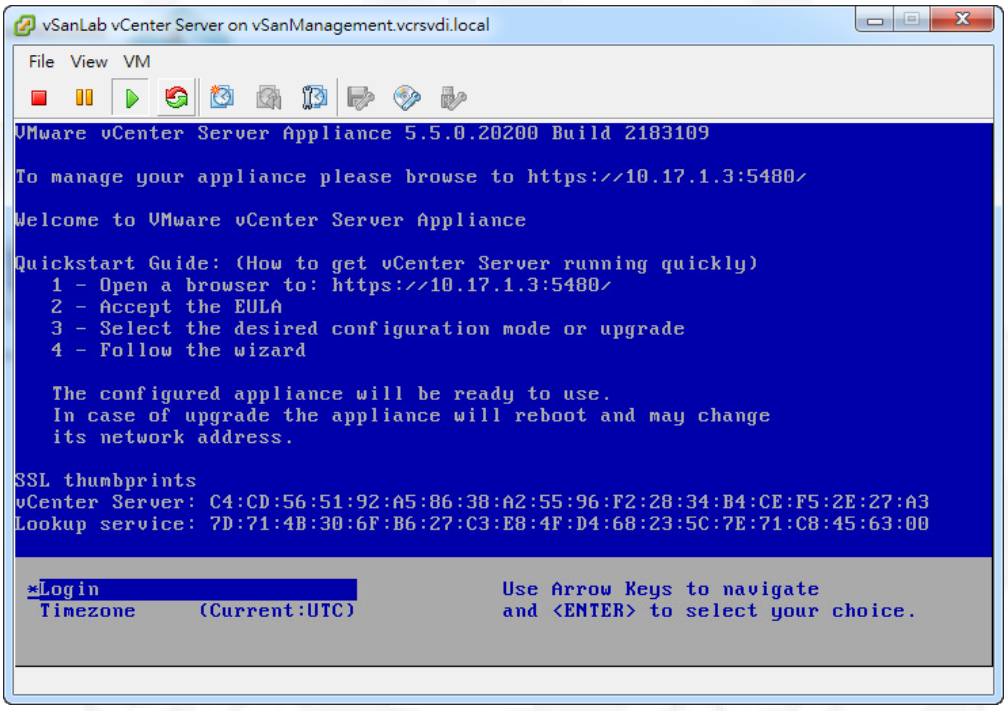

FIGURE B.3: 佈署完成 Console 畫面

接著執行 vami config net 指令，以設定主機管理 IP 位址。

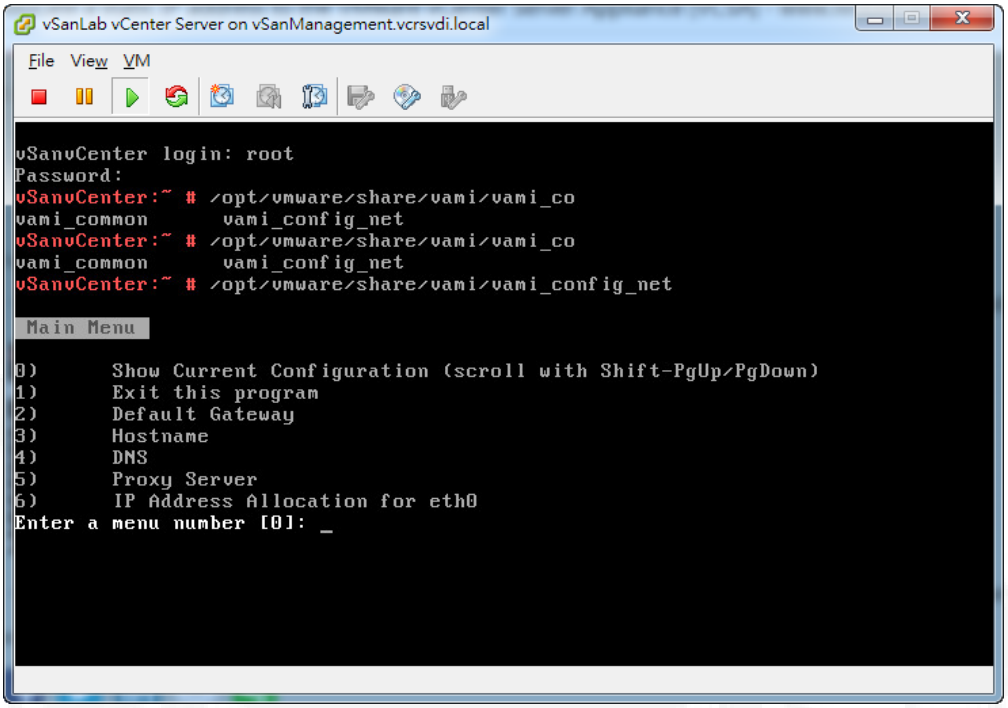

FIGURE B.4: VAMI Config Net 指令畫面

接著請打開瀏覽器輸入『https://static-ip-address:5480』，並登入 vCenter Server 服務管理網站。預設值登入帳號為 root，登入密碼為 vmware。

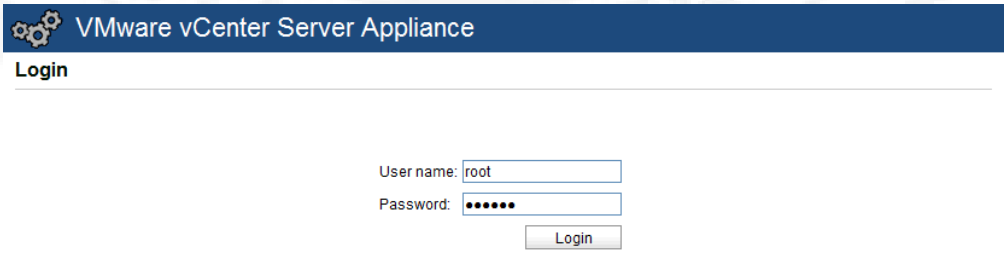

FIGURE B.5: Web Console 畫面

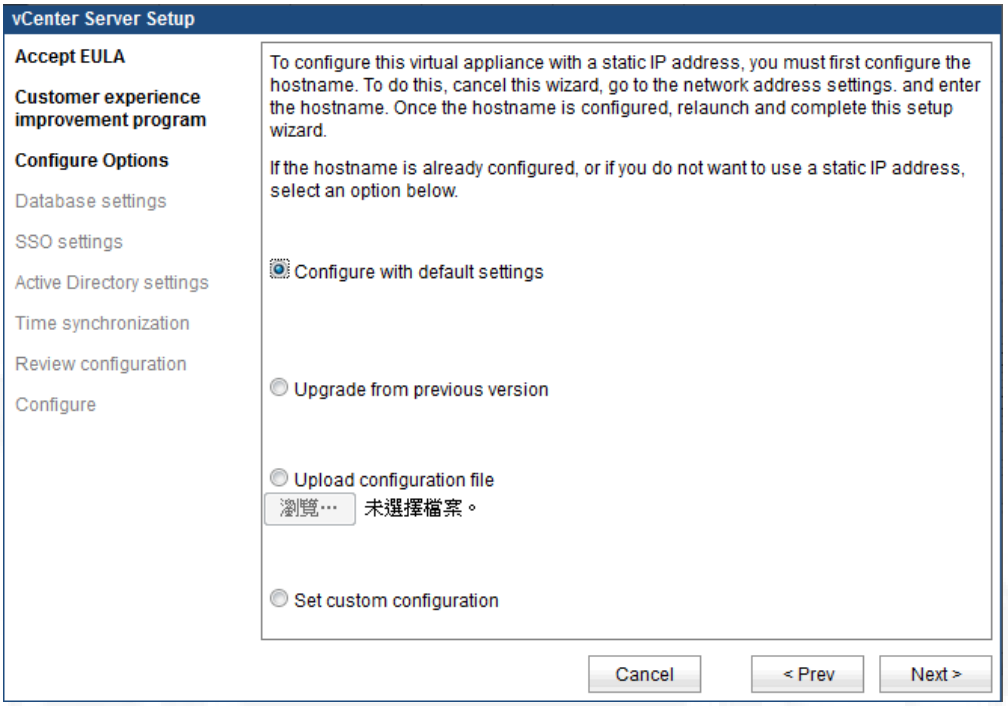

FIGURE B.6: 初始化設定畫面

選擇預設設定進行下一步設定

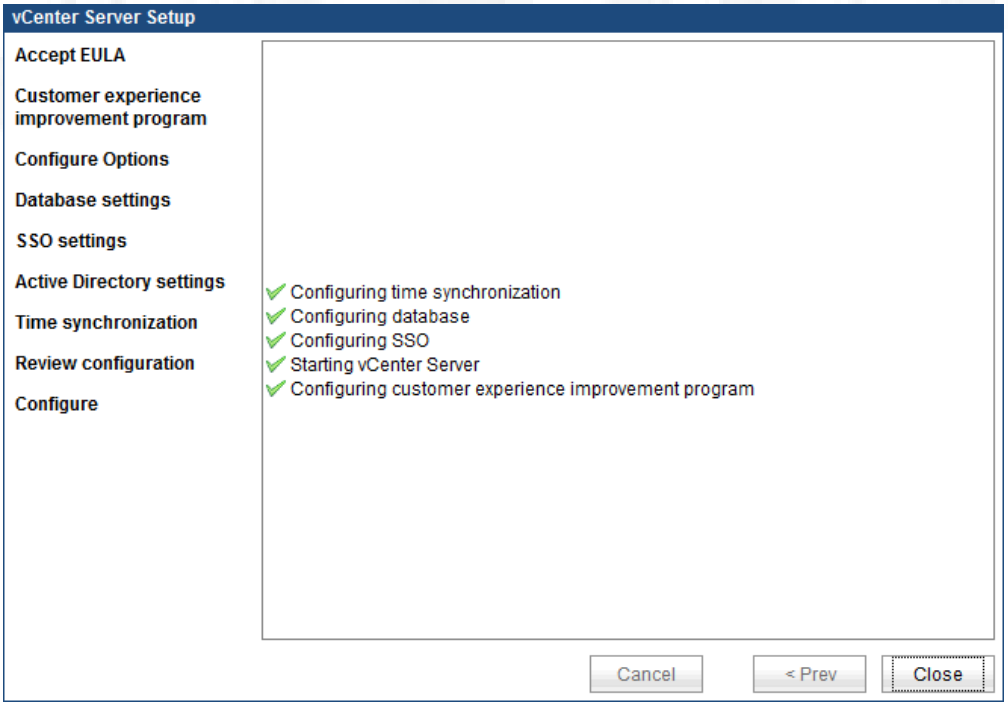

FIGURE B.7: 完成初始化設定

按 Close 完成佈署

vCenter Server 安裝完成後，後續如有需要變更登入認證為 Active Directory 登入或是變更登入密碼，以及更新 vCenter Server Appliance 版本，皆可透過登入此管理網站做後續管理及維護動作。

接著我們將接續設定叢集節點的相關設定，由於我們環境共有四部 ESXi 主機，故將叢集環境分為管理節點 1 台主機、資料節點 3 台主機，以滿足 vSAN 環境最小的安裝需求；登入 vCenter Server 的方式有二，分別為安裝本機端的 VMware vSphere Client 工具來登入，或是直接透過瀏覽器的 VMware vSphere Web Client 來登入。但我們後續要進行的進階相關網路設定及 vSAN 環境設定，目前僅能透過 Web Client 的方式操作，故我們將以 Web Client 做為主要操作工具。

登入 vCenter Server 網址為『https://static-ip-address:9443』

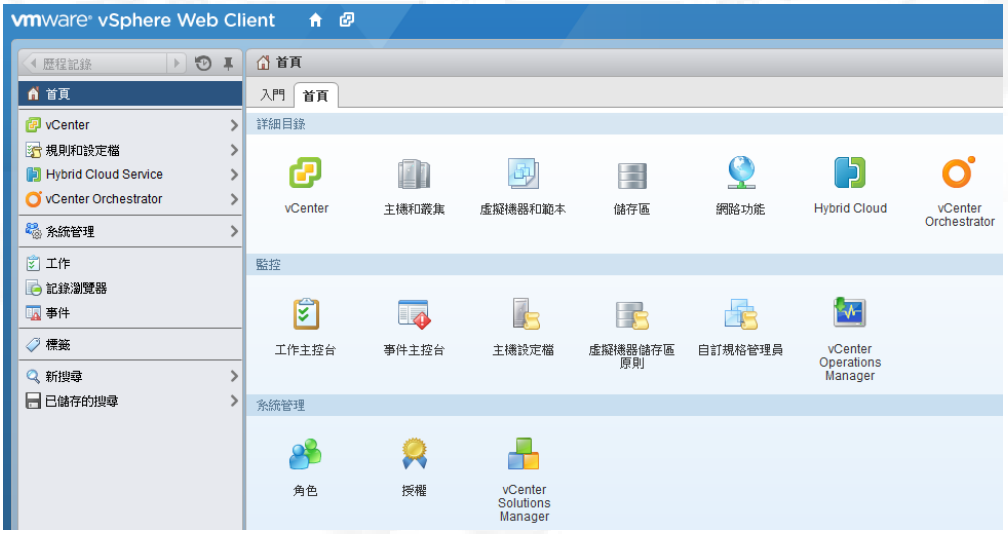

FIGURE B.8: Web Client 管理主頁

登入 vCenter Server 後，我們需要先建立一個資料中心，才能進行建立叢集的動作。故從首頁上直接點選『vCenter』→『主機和叢集』，再點選『建立資料中心』並給予一個識別名稱。

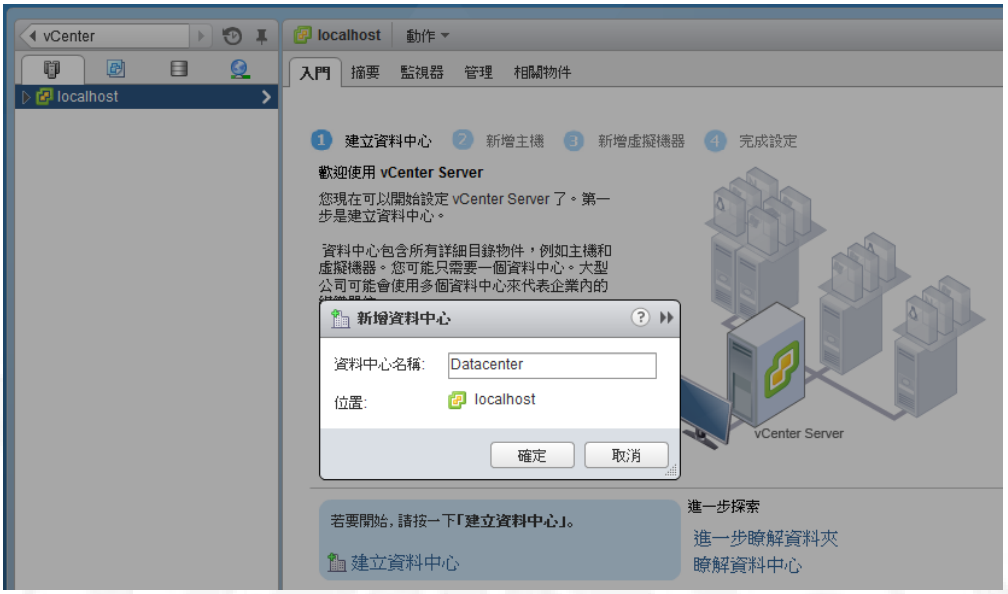

FIGURE B.9: 建立資料中心

接著再點選剛剛建立好的資料中心，再點選『建立叢集』，命名為 Management，此處可看到在建立叢集的位置已經有一個『虛擬 SAN』的功能可供勾選，但此處我們先不勾選，因目前先建立的是給管理主機使用的叢集。如企業要提供給管理主機可好的高可用性服務，則後續可在加入二台主機至該叢集後，再勾選 vSphere HA 服務或 DRS 功能，以提供管理服務主機更好的高可用性。此處由於我們僅加入單一主機，故可先不勾選啟用相關高可用性功能。

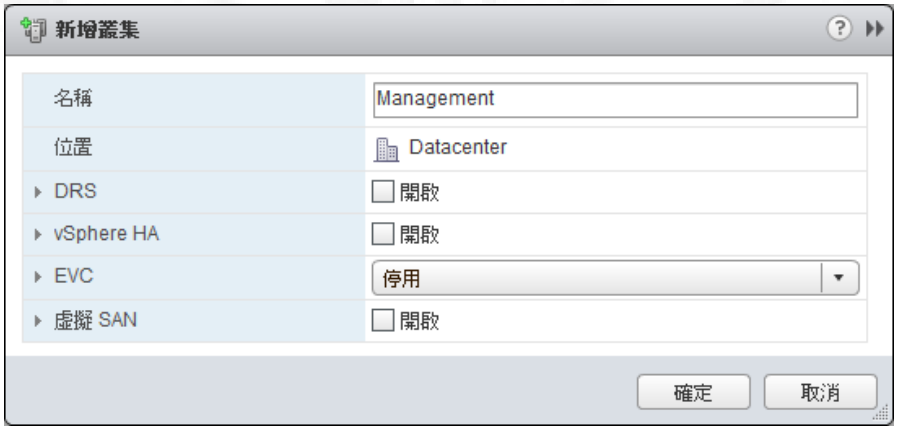

FIGURE B.10: 建立管理叢集

接著我們建立預定做為資料節點的叢集，用相同的方式建立命名為 DataNode 的資料節點叢集；此時可先勾選『虛擬 SAN』功能，預設 vSAN 是

採自動將 Host 本機磁碟加入至儲存區，此時是單一 Host 單一磁碟群組模式，如要建立單一 Host 多磁碟群組模式，則可在此處下拉選單改為『手動』；此處我們採自動模式。

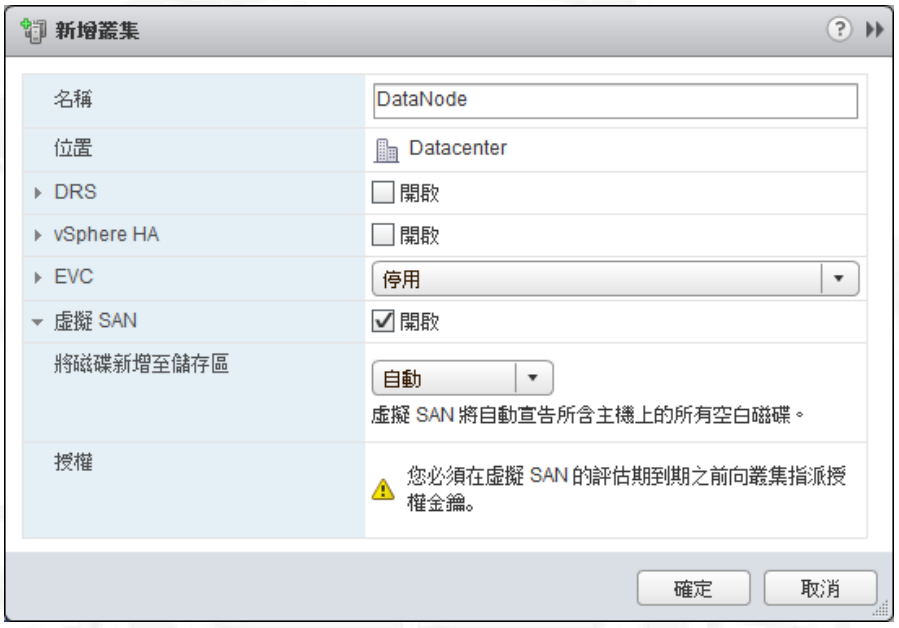

FIGURE B.11: 建立 vSAN 資料叢集

叢集新增完畢後，便可以開始將 ESXi Host 主機加入到叢集中，只要 vCenter Server 可以連接通訊的到 ESXi Host 主機，便沒有任何的問題。

二、設定 vDS（vNetwork Distributed Switch）分散式交換器

網路環境在 vSAN Cluster 中相當重要，因為每一台 ESXi 主機上的 vSAN 資料 I/O 流量都需要透過 VMkernel Port 來傳輸，因此一致性且正確的網路組態設定，是部署 vSAN 環境的關鍵之一，同時由於 VM 虛擬主機的儲存資源都是散落在不同的 ESXi 主機當中（分散式儲存架構），所以網路環境的設計也將是影響網路傳輸能否順暢的重要關鍵。所以 VMware 官方才會建議在生產環境中，採用 10GbE 等級的網路環境。

在 vSAN 環境中，可以支援 vSS 及 vDS 二種網路虛擬交換器；但由於我們希望透過我們的研究來自訂 NIOC（Network I/O Control）網路流量網制，及使用進階的 LACP 網路聚合功能來組合一般的 1GbE 網路環境，故我們將採用

vDS 分散式交換的設定方式來完成；此種方式也有助於前面所提到的一致性的網路設定，以盡可能減少降低網管人員在設定 ESXi 主機的網路錯誤可能性。

從首頁上點選『網路功能』後，再點選資料中心即可建立『Distributed Switch』，如下圖方式建立：

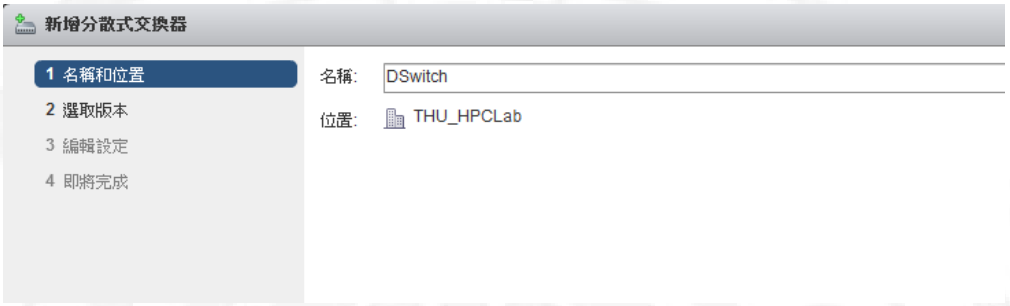

FIGURE B.12: 建立 vDS 交換器名稱

FIGURE B.13: 選擇 vDS 交換器版本

接著維持預設值上行數目 4，啟用 Network I/O Control，但將建立預設連接埠群組的選項拿掉。

FIGURE B.14: 建立 vDS 交換器選項設定

完成後，我們將建立三個分散式連接埠群組，分別提供給 VMotion、vSAN Data Flow、VMs Network 使用，並在後續設定相關的 NIOC 網路流量控制給 vDS 分散式網路交換器使用。

FIGURE B.15: 建立 vDS 交換器 VMKernal 連接埠

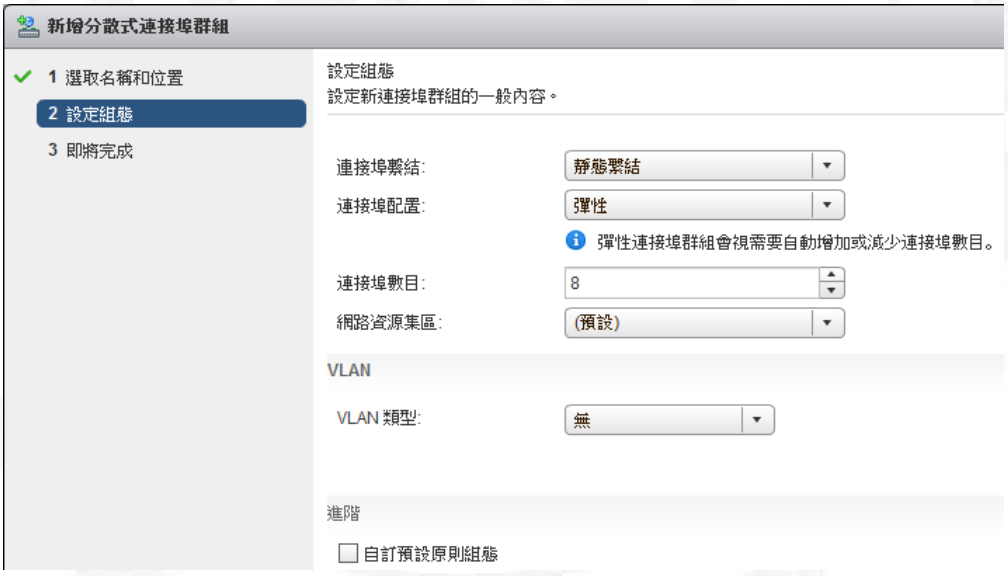

FIGURE B.16: 建立 vDS 交換器 VMKernal 設定畫面

完成後網路拓撲如下圖。

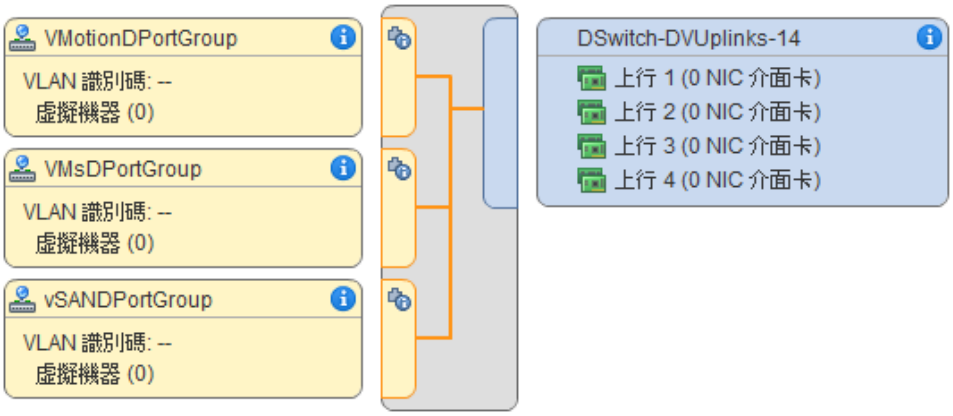

FIGURE B.17: vDS 交換器網路拓撲

對於前面我們所在新增 vDS 分散式交換器時所選擇的分散式交換器 5.5.0 版本，最好的差異在於過去傳統的 LACP（Link Aggregation Control Protocol）設定方式僅能作用於 vSwitch 上面，而新版本的進階加強版 LACP 則可直接設定作用於連接埠群組，因此可以有更多的彈性應用方式。但唯一需要注意的是，LACP 功能須同時設定於實體網路交換器上面，如實體網路交換器不支援，則設定將不會生效。

請先從 vDS 左邊選單上選擇『LACP』，並點選加號新增。

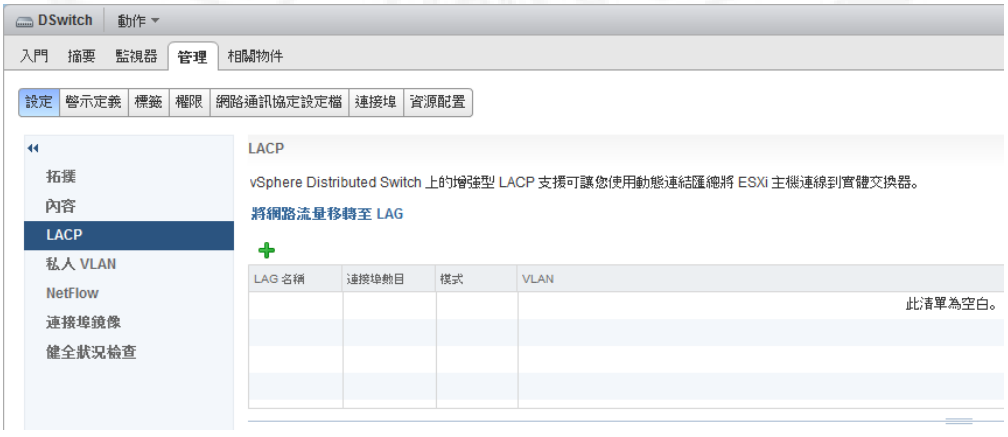

FIGURE B.18: 新增 vDS 交換器 LACP 連接埠

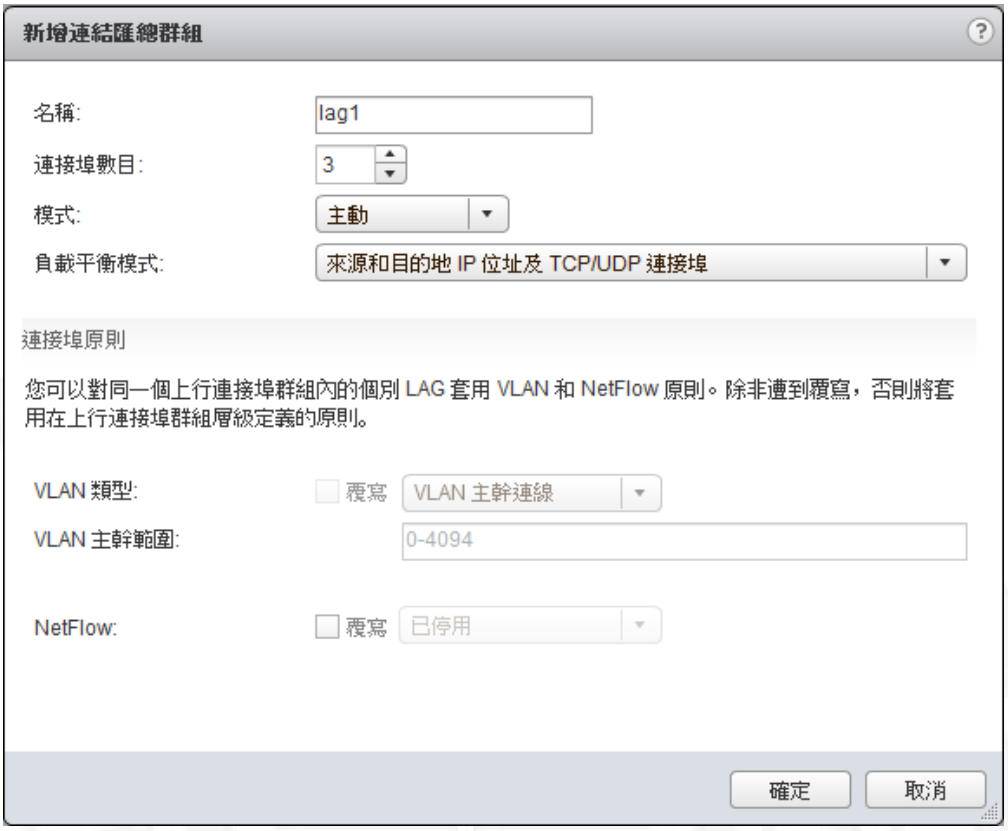

FIGURE B.19: vDS 交換器 LACP 連接埠設定

並且我們也需要同步調整實體連接第二層交換器 L2 Switch 的 Link Aggregation Groups(LAG) 設定。

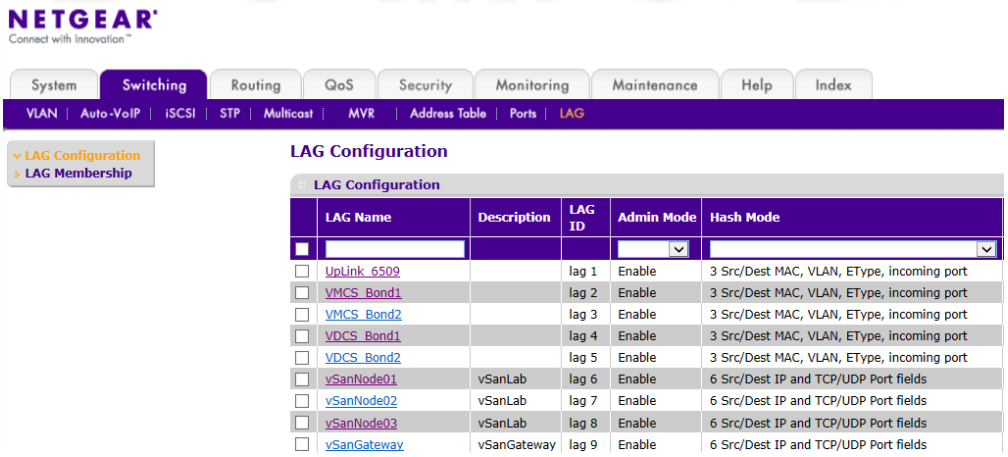

FIGURE B.20: 實體交換器 LAG 設定

為了確保在網路頻寬發生壅塞的時候，不會因為任何一種服務類型佔用大部分的網路頻寬，進而影響到其他的服務類型，因此我們的研究在測試過各種

實務上會遇到的狀況後，提出了建議的 NIOC 網路流量控制設定範本給企業參考套用使用。在此應用情境中，我們將會在 NIOC 機制中，為每項服務類型以『共用（Share）』比例的方式進行設定，以便網路資源充足時能盡量使用網路頻寬，而網路頻寬雍塞時也能保證每項服務類型，不致於中斷服務。

下圖為我們的測試過得出的建議 NIOC 設定值：

服務類型	共用比例
管理網路（Management Network）	20
即時移轉網路（vMotion Network）	50
VM 虛擬主機網路（VMs Network）	30
vSAN 資料網路（Virtual SAN Network）	100

FIGURE B.21: 建議 NIOC 設定值

我們可以從 vCenter Server 中，點選首頁 → 網路功能 →vDS：資源配置的方式，進入網路流量控制的組態設定頁面，如要調整 vSAN 的「共用率值」則點選該項目後，再按下編輯圖示即可調整。

預設啟用網路流量控制後的設定，設定內容為無限制網路流量、實體網路介面卡共用率為 50、沒有 QOS Tag，其中說的無限制網路流量表示當網路頻寬充足的情況下，可以最大限制的使用所有網路頻寬，但是當網路頻寬發生雍塞時，便會依實體網路介面卡的預設值「50」來分配 50

所以我們才會建議在共用網路介面的情況下，我們應合理的保留 vSAN 資料移轉網路所需的網路頻寬，以盡量增進系統傳輸的整體穩定性及效能。

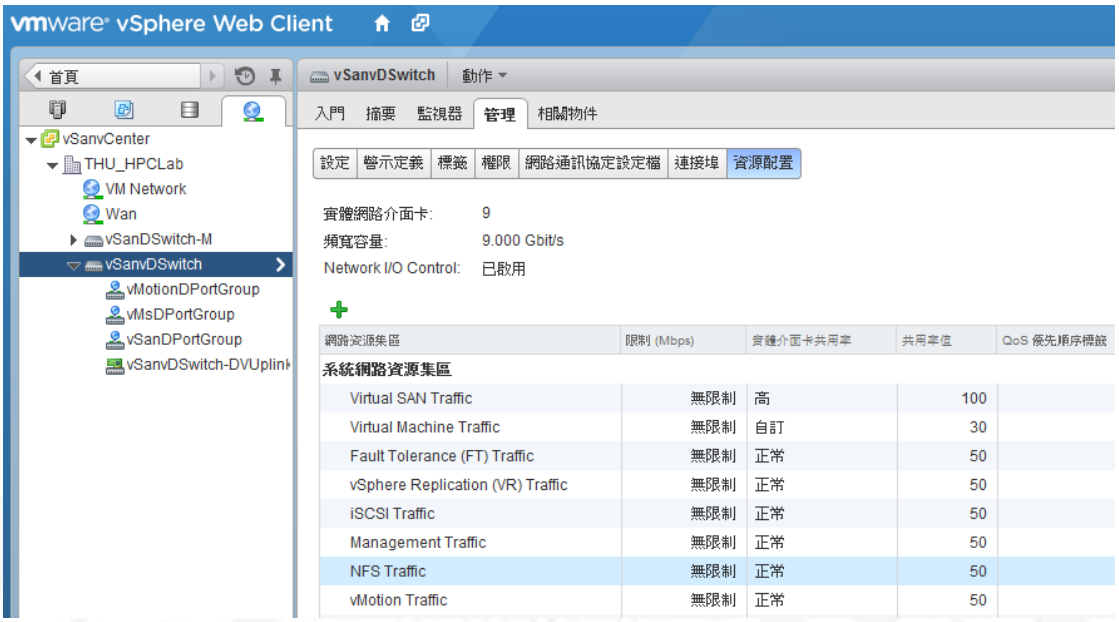

FIGURE B.22: NIOC 流量管理畫面

三、建立 vSAN Cluster DataStore 儲存裝置

在目前的 vSAN 版本中，每個資料節點叢集僅能支援建立一個 vSAN DataStore，而建立 vSAN Cluster 的方式，其實在前面已經提到與傳統建立 vSphere Cluster 的方式是相同的，也就是說可以與 DRS（Distributed Resource Scheduler）及 HA（High Availability）機制整合在一起。因此在前面當我們在建立資料節點叢集時勾選『虛擬 SAN』項目，後續在加入資料節點的 ESXi Host 主機時，系統即會自動幫我們建立預設的 vSAN DataStore；唯一要注意的是前面所提到的主機及網路硬體規格相容性的問題即可。

而我們的研究範例架構，剛好以滿足 vSAN 主機架構為例，資料節點三台主機、每台主機分別安插一個固態 SSD 硬碟及一般機械式 SATA 硬碟，所以完成的架構如下所示：

FIGURE B.23: 叢集狀態磁碟群組檢視

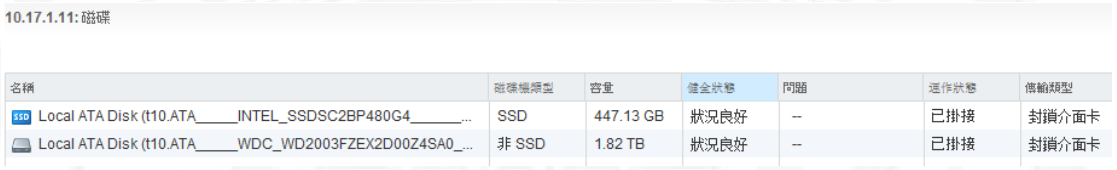

FIGURE B.24: 單一機器磁碟群組檢視

FIGURE B.25: vSAN 總容量檢視

目前最多支援 32 台 ESXi 成員主機，每一台 ESXi 成員主機最多可以有「5 組磁碟群組」，而每一組磁碟群組中需至少包含一個固態型 SSD 硬碟以及 1 至 7 顆的一般機械式硬碟。由於我們前面是採自動建立 vSAN 磁碟架構，所以若後續為了要增進效能或是擴充儲存規格，可先將『虛擬 SAN』改為手動方式，再手動建立磁碟群組即可。如下圖：

FIGURE B.26: 手動建立 vSAN 磁碟

在完成前面建置過程中各項調整設定後，我們的 vSAN Cluster 網路拓撲如下：

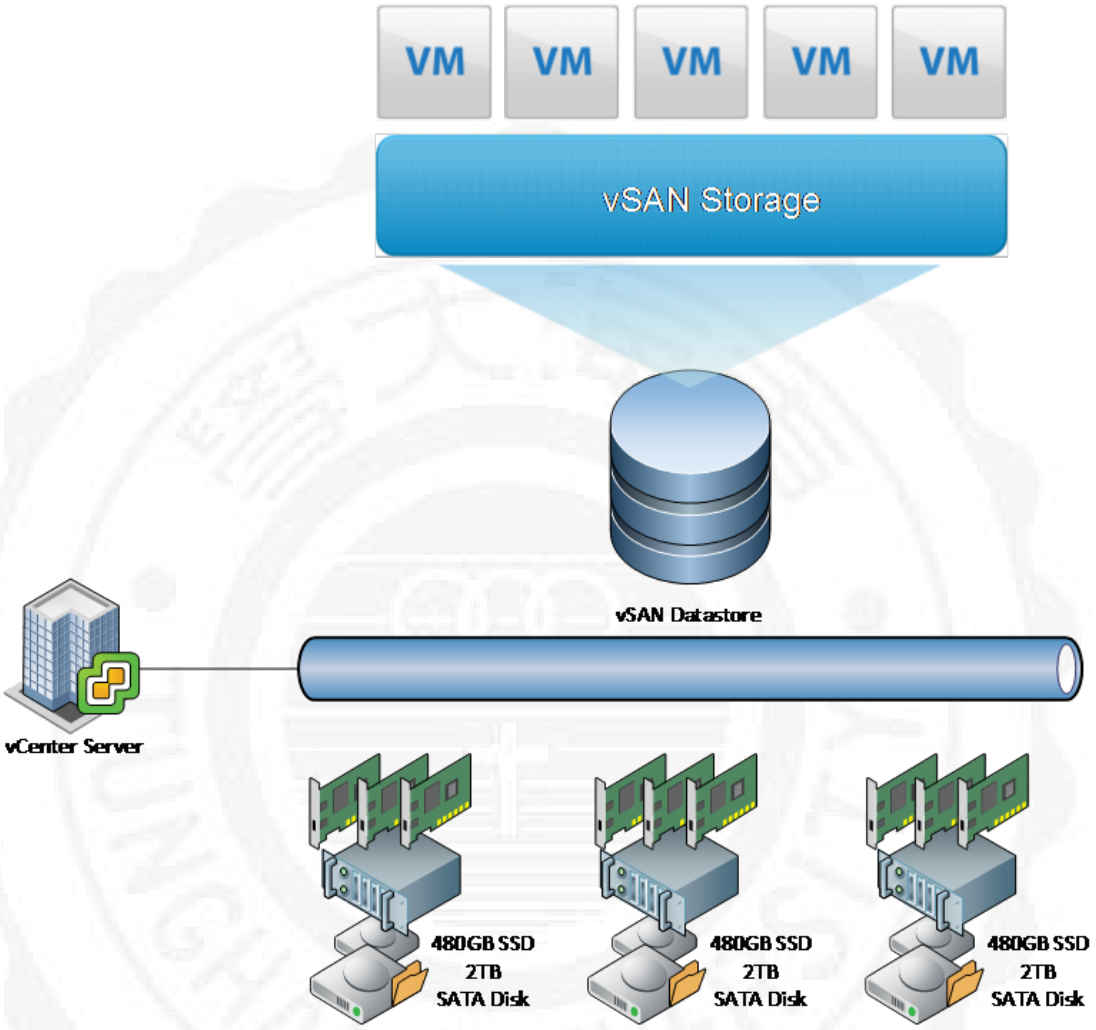

FIGURE B.27: vSAN 測試環境架構

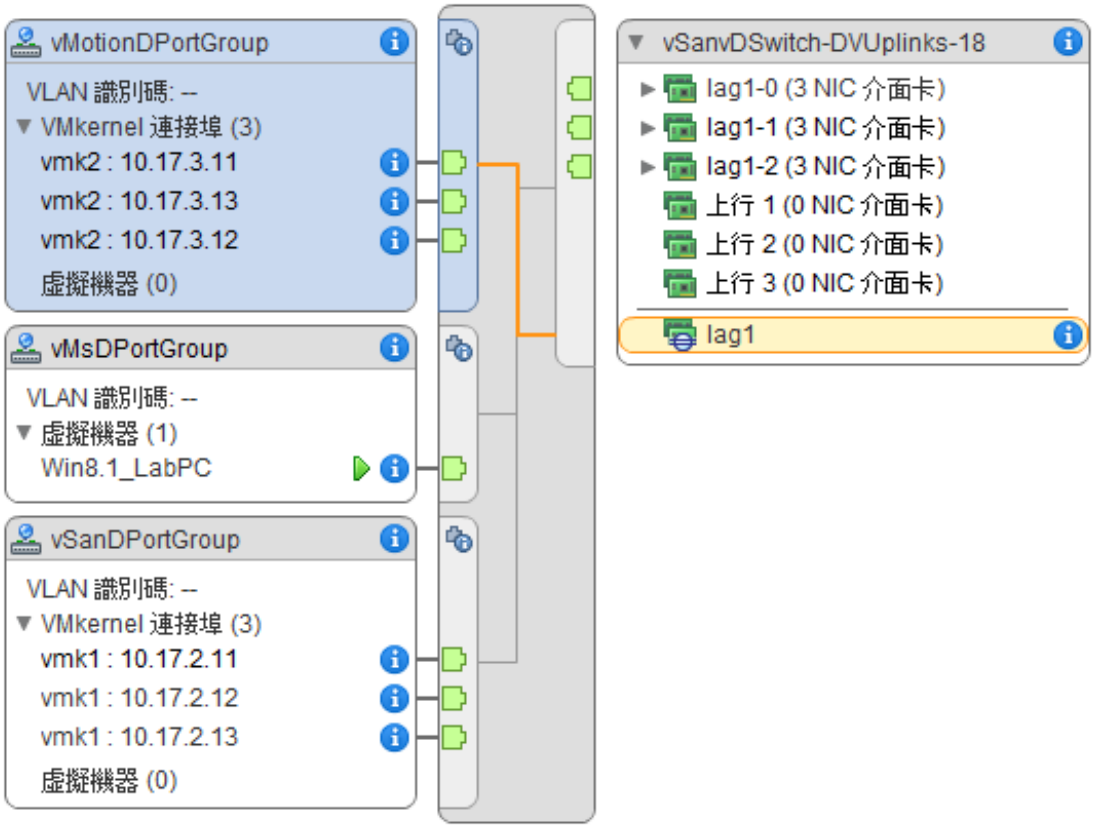

FIGURE B.28: vDS 網路拓樸架構

附錄 C

Mellanox 40G QSFP+ 優化設定

在優化調整 Mellanox 40G QSFP+ 的網路卡在 vShere ESXi 底層之前，我們得先確認主機的 BIOS 設定是否已經符合以下三項設定：• Set CPU max performance • Set C-state disable • Set PCIe fix to Gen3

因為大多數的主機 BIOS 預設值都是為了符合一般生產環境中的綠色節能要求，所以在 CPU 的時脈上都會以加以限定，而 PCIe 匯流排的設定在某些情況下都是自動偵測，但偵測出來的結果有時會不一定正確；所以我們會先做上述的調整，以優化調整主機最大效能在 BIOS 的設定中。

而在 vSphere 6.0 以上的版本中，系統底層已經內建了 Mellanox 原廠的驅動程式 Driver，並且此驅動程式版本是由 VMWare 官方所認證過的。但我們仍需要確認系統是否正確載入驅動模組，以免後續的操作無法執行。

```
[root@HPC239:~] esxcli software vib list |grep nmlx4
nmlx4-core                     3.0.0.0-1vmw.600.0.0.2494585         VMware  VMwareCertified   2016-04-06
nmlx4-en                       3.0.0.0-1vmw.600.0.0.2494585         VMware  VMwareCertified   2016-04-06
nmlx4-rdma                     3.0.0.0-1vmw.600.0.0.2494585         VMware  VMwareCertified   2016-04-06
[root@HPC239:~]
```

FIGURE C.1: 確認 ESXi 載入驅動模組

其中在 Mellanox 原廠的文件中有提供，RDMA 模組的載入正確與否，雖不影響網路卡使用，但卻會在多工的環境下影響系統效能，所以我們也要一併確認 RDMA 模組。

```
[root@HPC239:~] vmkload_mod -l |egrep 'mlx|ib_|rdma'
nmlx4_core                1     288
nmlx4_en                  1     200
rdma                      5     196
nmlx4_rdma                1     92
```

FIGURE C.2: 確認 RDMA 模組載入

確認 ESXi 目前可使用的網路卡清單。

```
[root@HPC239:~] esxcfg-nics -l
Name         PCI          Driver      Link Speed     Duplex MAC Address       MTU    Description
vmnic0       0000:00:19.0 e1000e      Up   1000Mbps  Full   38:2c:4a:71:43:98 1500   Intel Corporation Ethernet Connection (2)
 I218-V
vmnic1000202 0000:02:00.0 nmlx4_en    Up   40000Mbps Full   e4:1d:2d:e3:9e:e1 1500   Mellanox Technologies MT27520 Family
vmnic2       0000:02:00.0 nmlx4_en    Down 0Mbps     Half   e4:1d:2d:e3:9e:e0 1500   Mellanox Technologies MT27520 Family
```

FIGURE C.3: 確認 ESXi 網路卡清單

Enable NetQueue support ESXi 在預設的情況下，VMWare 的 NetQueue 是 Disable 的，這是因為在傳統過去的虛擬化項目中，常見的建議值是安裝多個 1Gbps 的網路卡，並分別分配給多個虛擬機個別使用，這最大化了虛擬機的網路效能；但如果是一張 1Gbps 的網路卡同時共享給多個虛擬機使用，此時網路卡的效能並不會被考慮進去，因為 1Gbps 的頻寬比對系統的 I/O 需求來說，要擔心的是網路卡的頻寬不足而非無法塞滿網路卡的頻寬。

但在高速的網路介面如 10Gbps 或 40Gbps 的網路卡上面，此時因為網路卡的頻寬可能已高於系統的 I/O 最大值，所以此時要考慮的就是要打開網路卡的排隊駐列需求，以求最大化的使用。

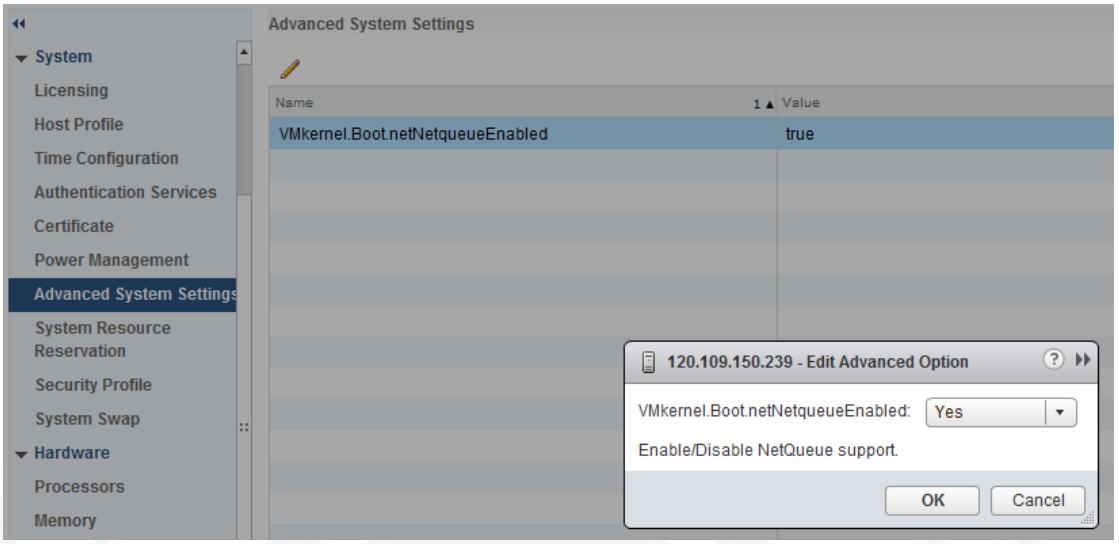

FIGURE C.4: 啟用 NetQueue

接下來我們要調整二個參數，以讓網路對於多核心的 CPU 可以最佳化。

```
[root@HPC239:~] esxcli system module parameters set -m nmlx4_en -p "num_rings_per_rss_queue=4"
[root@HPC239:~] esxcli system module parameters list -m nmlx4_en |grep num_rings_per_rss_queue
num_rings_per_rss_queue  int   4      Enable RSS
   When this value is != 0, RSS is enabled with 1 RSS Queue that manages num_rings_per_rss_queue Rx Rings
```

FIGURE C.5: 啟用 num rings per rss queue

```
[root@HPC239:~] esxcli system module parameters set -m mlx4_en -p "netq_num_rings_per_rss=4"
[root@HPC239:~] esxcli system module parameters list -m mlx4_en |grep netq_num_rings_per_rss
netq_num_rings_per_rss  uint  4      Number of rings per RSS netq
```

FIGURE C.6: 啟用 netq num rings per rss

接下來調整 ESXi 主機的 CPU 效能設定，以符合前面 BIOS 的設定。

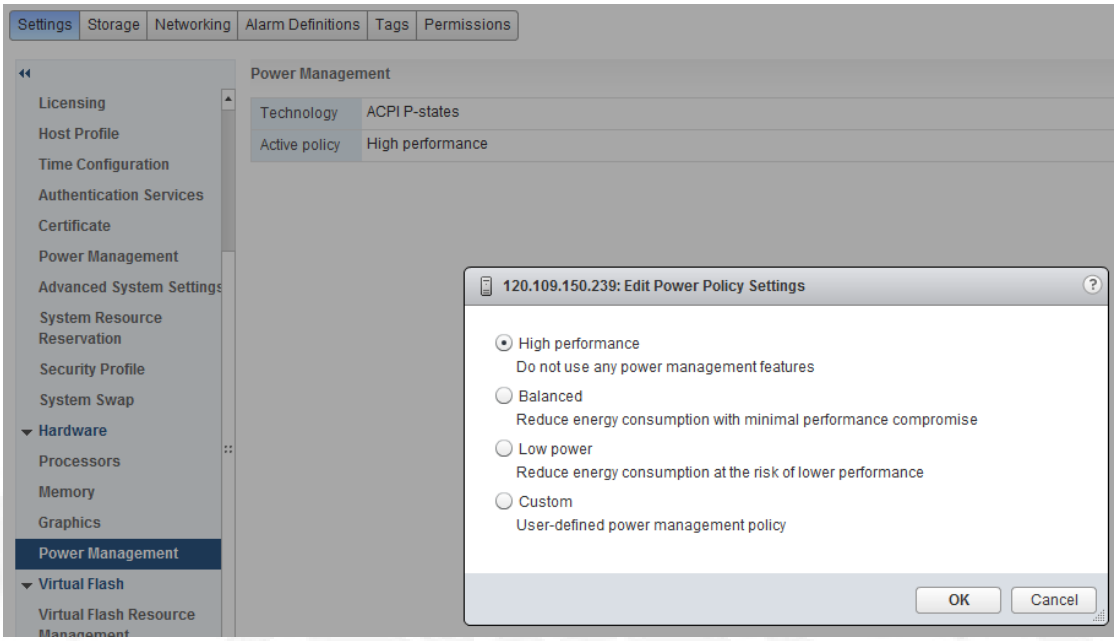

FIGURE C.7: 設定 ESXi 主機的效能設定

最後，如果要從虛擬機器這邊也啟用多核心支援高速網路，須要在從虛擬機這手動增加一個參數到 VMX 設定檔中。

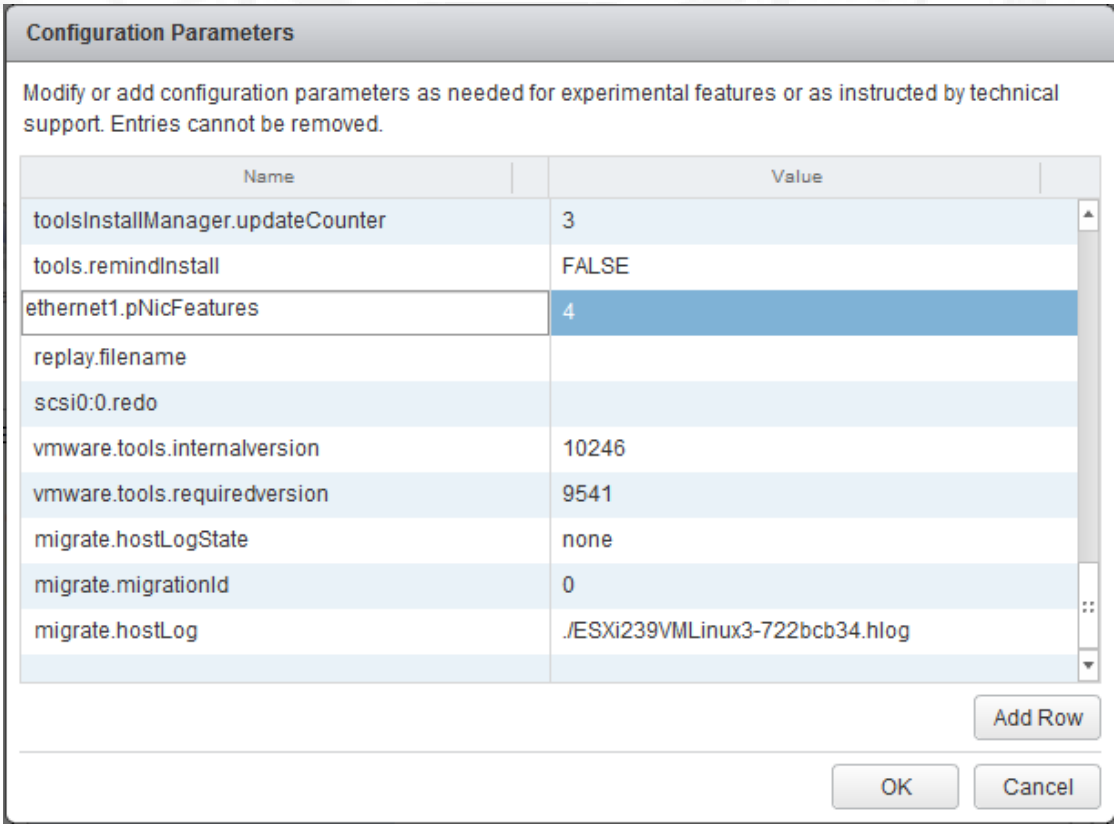

FIGURE C.8: 啟用虛擬機的多核心網路卡支援