

iSCSI 網路儲存：技術、協定與產品

古佳育 林義能 林盈達

摘要

檔案系統可以透過兩種方式要求儲存空間的服務：file I/O 和 block I/O。file I/O 透過 NFS、CIFS 等網路協定，而 block I/O 則由檔案系統直接對儲存裝置做處理，儲存裝置的類型除了磁碟、磁帶外，還有一個重要的網路儲存技術-SAN(Storage Area Network)。現在的 SAN 指的是 FC SAN(Fibre Channel SAN)，有著價格高、不易架設、距離限制以及需要光纖通道技術以管理和維護等問題，也因此造成另一種 SAN 的技術-IP SAN 的興起。在眾多的相關技術中，真正能實現 IP SAN 的首推 iSCSI，亦即將 SCSI 協定映射到 TCP/IP 上面來。SCSI 為了能夠同時處理多個 I/O 裝置之間的操作，命令集較為複雜，但卻也較適用於需要處理大量資料的伺服器。然而網路不比匯流排，除了本身使用的 SCSI 命令集，還需要一些額外的命令來做控制。

iSCSI 的產品分為硬體和軟體兩大類，我們從功能性、效能、符合性、互通性，以及儲存設備等角度來探討其品質的差異。未來 iSCSI 面臨的最大挑戰在於效能上是否能夠滿足需求，可能的解決方向是用 iSCSI 的 ASIC 來處理 iSCSI 命令和未來的高頻寬網路技術來提升頻寬解決需求的問題。

1. 簡介

大量的資料與其備份常常造成儲存空間不敷使用和管理不易的問題。NAS(Network Attached Storage)與 SAN(Storage Area Network)就是在這樣的環境和趨勢之下應運而生的網路儲存技術。相對於傳統在伺服器上直接連接儲存設備的方式(DAS, Direct Attached Storage)，網路儲存技術指的是伺服器和儲存裝置之間透過網路作連接。

圖 1 說明 DAS、NAS、SAN 和檔案系統之間的關係。NAS 的概念像一台檔案伺服器，把所有儲存裝置連接在起來，而需要儲存空間的伺服器再透過 NFS、CIFS 等等協定向 NAS 請求儲存空間的服務，值得注意的是以檔案為單位(file I/O)處理請求。SAN 是獨立於 LAN 的儲存區域網路，伺服器和 SAN 的溝通是用 client-server model；提出請求的一端稱為 initiator，提供儲存服務的一端則稱為 target。有著可以連接數以萬計(理論上可達 1600 萬)的儲存裝置，且會自動做資料備份和對儲存空間做分配和管理的能力，SAN 儼然成為一個容量極大且功能強大的智慧型磁碟。

表 1 比較它們彼此之間的不同。從表中可以看出 SAN 是 block I/O，而 NAS 則是 file I/O。block I/O 比起 file I/O 更為靈活，因為關鍵的資料儲存通常是以資料庫為基礎進行儲存和管理，而其主要資料型式便是 block；從作業系統的角度來看，block I/O 會被辨識為本地磁碟，可以直接進行操作，反之 file I/O 則必須在 file 和儲存裝置的命令之間做轉換，所以在大量的資料處理上來看，block I/O 會比 file I/O 快得多。再來就是儲存裝置的集中化管理，這不但能有效的分配和管理儲存空間和提

但是距離上卻受到限制，而且 FC SAN 需要專門的光纖通道技術人員管理和維護，IP 網路則是相對較為普及，比較不會有管理和維護上的問題。

IP SAN 技術

目前一般認為，SAN 的發展有三個階段，分別是

1. FC SAN 的擴展，亦即將目前既有的 FC SAN 透過 IP 網路互連。
2. 有限區域的 IP SAN。架設有限區域的 IP SAN，多屬小型產品，而且 IP SAN 不會取代 FC SAN。
3. 全球化的 IP SAN。

IP SAN 能使用的技術有很多，我們挑出最常見的 FCIP 與 iFCP 來和 iSCSI 在表 3 作個比較。在此之前，先讓我們對這三種技術做概念性的描述。

FCIP(見圖 2-1)把光纖通道訊框封裝在 TCP/IP 裡面，允許在 FCIP 橋接器之間建立 FCIP 隧道。FCIP 的目的是把 FC SAN 透過 IP 網路整合在一起，以解決傳輸距離上的問題。伺服器向本地的光纖通道交換器送出請求，然後透過名稱伺服器 and FCIP 橋接器通過 IP 網路傳送到目的地的 FCIP 橋接器，再傳送資料給 SAN 做正確的處理，故此技術適用於階段一或二；但是 FCIP 缺乏隔離故障的能力，如果發生錯誤很可能會擴散到其他的 SAN，再加上隧道中斷的話也不會自動重建，造成一般企業沒有採用的意願。

iFCP(見圖 2-2)也是把光纖通道訊框封裝在 TCP/IP 裡面，不同的是 FCIP 將 SAN 整合成一個很大的 SAN，對伺服器而言使用的仍是一個 SAN，而 iFCP 提供的是 iFCP 閘道器之間的連接，閘道器有 IP 位址，可以讓 iFCP 可以透過路由器傳遞正確的目的，每個 iFCP 閘道器下都是一個獨立的自治區，這種做法使得 iFCP 具有隔離故障的能力，並且包含有錯誤偵測與恢復機制，比起 FCIP 是更成熟的技術。更重要的是，iFCP 閘道器可以取代光纖通道交換器，將 IP SAN 的使用逐步導入，是一種相當具有戰略性的技術，適用於階段一和二。

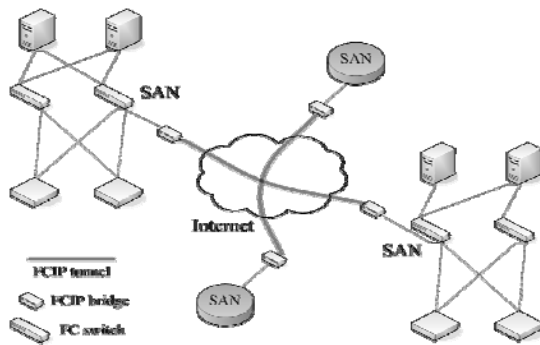


圖 2-1：FCIP 架構圖例。

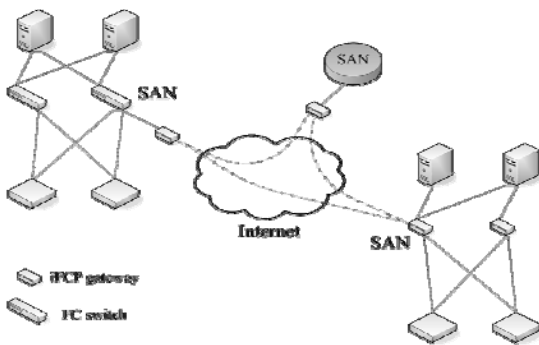


圖 2-2：iFCP 架構圖例。

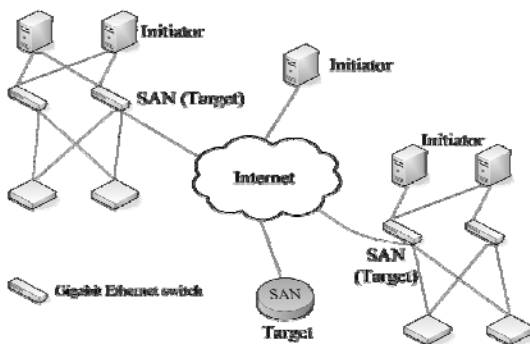


圖 2-3：iSCSI 架構圖例。

	FCIP	iFCP	iSCSI
Mechanism	FC command over IP	FC command over IP	SCSI command over IP
Model	integrates with SANs	FCIP gateway to iFCP gateway	client-server (initiator-target)
Error handling	N/A	error detection fault isolation recovery	error detection fault isolation recovery
Security	N/A	IPsec	IPsec
Future development phase	phase1&2	phase1&2	phase2&3

表 3：iSCSI 與其他 IP SAN 技術比較表。

iSCSI(見圖 2-3)是將 SCSI 命令封裝在 iSCSI 裡面，再透過 TCP/IP 傳送到世界上任何一個角落。然後由 target 依序解開 IP 標頭、TCP 標頭和 iSCSI 標頭，以取得 SCSI 命令，最後再回應 initiator 的請求。iSCSI 完全不需要光纖通道的技術和設備，是完全運作在 IP 網路上面的技術，適用於階段二和三。最後讓我們利用表三將以上三種技術做個比較。我們發現 iFCP 和 iSCSI 都有提供完善的錯誤偵測、恢復機制以及安全性的處理，但是 iSCSI 可以不需要光纖通道的設備，完全運作在 IP 網路上，相較之下，iFCP 比較像是過渡時期的技術。

3. 制訂 iSCSI 的背景與想法

iSCSI 的概念是 SCSI over IP，簡單的說，就是把 SCSI 命令用 TCP/IP 透過網路來傳遞，使我們可以利用網路使用遠端的儲存設備，但是這個構想是從何而來？我們透過觀察 iSCSI 協定和 SCSI 技術與命令集，來探討它制定的背景與想法。

iSCSI 與 SCSI 之間的關聯

SCSI 的全名是 Small Computer System Interface，是一種連接主機和外圍裝置的接口，在主機和儲存裝置之間處理 I/O 命令。SCSI 在處理多個裝置的環境之下效能特別突出，這是因為 SCSI 本身的設計考量，SCSI 的設計理念是能夠處理多個裝置之間的溝通，為了同時處理多個裝置的要求，應用了多工作業，命令排序等等方法，讓 SCSI 控制器可以在裝置等候 I/O 的時間去處理別的裝置的命令。SCSI 初期就制定了相當詳細的規範，包括匯流排寬度、資料傳輸方式和所需要的控制命令等等，後來為了支援越來越多的裝置，再加上必須有完全向後相容性，所以規格(控制單元和命令集)也就越來越複雜。

SCSI 技術有詳細而完整的標準，是一個具有高成熟度的儲存技術，而且是為了處理多個裝置的設計，很符合網路儲存技術的需求，但是卻有著 SCSI 匯流排長度上的限制。IP 網路同樣是具有高成熟度的技術，有著高頻寬以及距離不受限制的特點，不受限制的距離解決了 SCSI 最大的問題。最重要的是，SCSI 的溝通方式是 client-server model，和 IP 網路不謀而合，因而會有把 SCSI 搬到 IP 網路上來的想法。

從圖 3 裡面我們可以看到 SCSI 標準間的結構，SCSI 標準都被 architecture model 所規範，SCSI transport protocol 則和命令集區分開來，primary command set 是所有的 SCSI 裝置都必須實作的指令集，Device-Type specific command set 則是針對不同裝置的延伸命令集。iSCSI 屬於 SCSI transport protocol，被 SAM-2(SCSI-3 Architecture Model-2)所規範，使用的命令集是 SPC-3 和 SBC。SBC 是針對處理磁碟 block 所發展出來的命令集，因為在 iSCSI 的架構下，initiator 會將 target 端看成是一顆 SCSI 磁碟，所以只需要對 block 做處理的命令集。

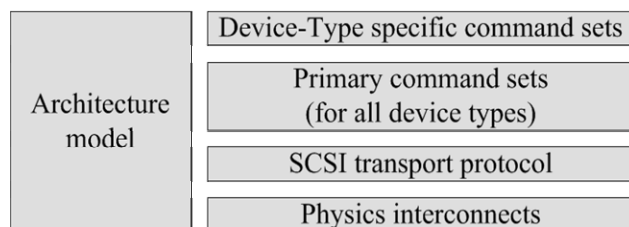


圖 3： SCSI 標準的結構圖。

iSCSI 協定概述

iSCSI 直接把 SCSI 映射的 TCP/IP 上面，也是採用 client-server model。client 端稱為 initiator，server 端稱為 target，傳輸的方向是以 initiator 的角度來看的，而裝置

之間溝通的方式是 request-response model。從圖 4 我們了解了 iSCSI 的基本架構。

每個 iSCSI node 可以是 initiator 或是 target，initiator 和 target 之間要先透過 network portal(TCP/IP connection)建立 connection，之後必須通過 CHAP 認證，然後才在兩者之間建立 session。一個 session 中可以有幾個 connections，在 initiator 端 network portal 是以 IP 識別的，在 target 端則是以 IP 和 port 做識別。在 session 建立之前是屬於 Login Phase，之後才是進行實際傳輸的 Full Feature Phase。

緊接著讓我們來看看 iSCSI 的命令集架構，iSCSI 的 request 和 response 列於表 4。有些 request 和 response 是為了透過網路傳輸所做的處理，舉例來說，iSCSI 的 recovery 主要有兩種，分別是 command 和 connection 的 recovery，拿 command recovery 當例子，雖然網路傳輸方面的 error handling 有 TCP 處理，但是 iSCSI 還是提供了 iSCSI PDU 的 lose 和 digest error 的處理。SNACK request 就是 initiator 發生 lost data、timeout 或是 digest error 時，要求 target 重新傳送傳輸資料。最重要的部份在於使用的 SCSI 命令集，也就是 SPC-3 和 SBC，SCSI 的命令可分為 control plane 和 data plane，control plane 尤其複雜，這是因為 SCSI 必須處理很多個裝置之間的溝通，並不只是單純的做讀寫的動作。我們在表 5 為這些命令做個分類。

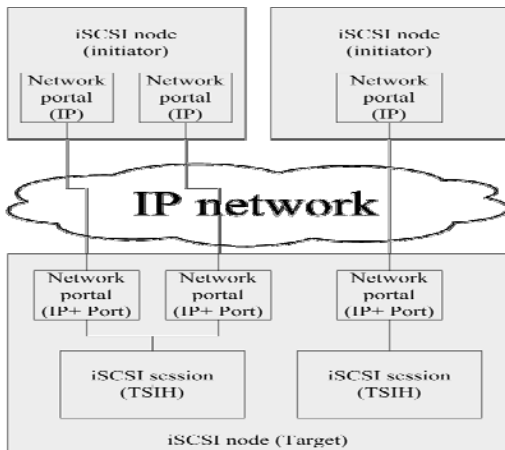


圖 4：iSCSI 架構圖例。

	request/response	functionality
iSCSI payload	Test	parameter negotiation vehicle
	Login	set up the session and connection parameter
	Logout	close the connection for recovery or others
	SNACK	request retransmission of data from the target
	Reject	the target to report an iSCSI error condition
	NOP-IN/NOP-OUT	as a ping mechanism
Both	Asynchronous message	carry AENs and iSCSI asynchronous messages
SCSI payload	SCSI-command	carry the SCSI CDB and all the others
	SCSI-response	carry return value and others
	Task management function request	provide an initiator to control iSCSI functions
	Task management function response	Carry an indication of function completion
	SCSI data-initiate-out	the main vehicles carry SCSI payload
	R2T	the target request the initiator for output data

表 4：iSCSI 協定的 request 和 response。

SPC-3(all device types)		
Control plane (obtain device information)	ACCESS CONTROL IN	REPORT ALIASES
	ACCESS CONTROL OUT	REPORT DEVICE IDENTIFIER
	CHANGE ALIASES	REPORT LUNS
	INQUIRY	REPORT PRIORITY
	LOG SELECT	REPORT SUPPORTED OPERATION CODES
	LOG SENSE	REPORT SUPPORTED TASK MANAGEMENT FUNCTIONS
	MODE SELECT(10)	REPORT TARGET PORT GROUPS
	MODE SELECT(6)	REPORT TIMESTAMP
	MODE SENSE(10)	REQUEST SENSE
	MODE SENSE(6)	SET DEVICE IDENTIFIER
	PERSISTENT RESERVE IN	SET PRIORITY
	PERSISTENT RESERVE OUT	SET TARGET PORT GROUPS
	PREVENT ALLOW MEDIUM REMOVAL	SET TIMESTAMP
	READ ATTRIBUTE	TEST UNIT READY
	READ MEDIA SERIAL NUMBER	WRITE ATTRIBUTE
	RECEIVE COPY RESULTS	
	Data plane	EEXTENDED COPY
READ BUFFER		WRITE BUFFER
RECEIVE DIAGNOSTIC RESULTS		
SBC(direct-access block devices)		
Control plane	CHANGE DEFINITION	REASSIGN BLOCKS
	COMPARE	REBUILD

(obtain disk information、request service actions of disk)	FORMAT UNIT	RECEIVE DIAGNOSTIC RESULTS	
	INQUIRY	REGENERATE	
	LOCK-UNLOCK CACHE	RELEASE (10)	
	LOG SELECT	RELEASE (6)	
	LOG SENSE	REPORT LUNS	
	MODE SELECT(10)	REQUEST SENSE	
	MODE SELECT(6)	RESERVE (10)	
	MODE SENSE(10)	RESERVE (6)	
	MODE SENSE(6)	SEEK(10)	
	MOVE MEDIUM	SET LIMITS(10)	
	PERSISTENT RESERVE IN	SET LIMITS(12)	
	PERSISTENT RESERVE OUT	START STOP UNIT	
	PRE-FETCH	SYNCHRONIZE CACHE	
	PREVENT-ALLOW MEDIUM REMOVAL	TEST UNIT READY	
	READ CAPACITY	VERIFY	
	READ ELEMENT STATUS		
	Data plane	COPY	WRITE BUFFER
		COPY AND VERIFY	WRITE LONG
READ BUFFER		WRITE SAME	
READ DEFECT DATA (10)		WRITE(10)	
READ DEFECT DATA (12)		WRITE(12)	
READ LONG		WRITE(6)	
READ(10)		XDREAD	
READ(12)		XDWRITE	
READ(6)		XDWRITE EXTENDED	
SEND DIAGNOSTIC		XPWRITE	
WRITE AND VERIFY			

表 5： iSCSI 使用到的 SCSI 命令集列表。

SPC-3 是對所有 SCSI 裝置都適用的命令集，SBC 則是針對磁碟 block 所做的動作，裡面的命令主要可分成維護裝置及獲得資訊、要求裝置進行某些動作以及資料傳輸，前兩者皆為 control plane，故歸在一類，後者為 data plane，現在我們各舉一個例子做簡單的說明。LOG SENSE 是對 target 要求 target 底下裝置使用的統計資訊。FORMAT UNIT 是要求磁碟進行格式化的動作。EXTENDED COPY 則是能把一組裝置中的資料複製到另一組裝置。

4.iSCSI 相關產品

iSCSI 並不是近幾年才發佈的新技術。早在 2001 年 1 月就由 IBM 和 Cisco 提出草案標準和推出了硬體產品。而在正式通過 iSCSI 標準之後，越來越多的廠商願意投入 iSCSI 產品的研發，微軟更是在標準推出的同年，不斷地推出產品。標準的制定通過和微軟的大力支持可以說是 iSCSI 成長的催化劑。

軟體產品

微軟可以說是 iSCSI 的幕後推手，不但推出了 Windows Storage Server 2003、Windows iSCSI Target，更為了推動 iSCSI，推出免費的 Windows iSCSI initiator。其他廠商有 FalconStor 的 iSCSI Storage Server、DataCore 的 SAN melody、RocketDivision 的 StarWind、StarPort 等，另外還有一些 open source 的 Linux 套件，分別列於表 6。

Product type	Product name	Operating system	License
Initiator	Windows iSCSI initiator	Windows XP/2000/Server 2003	Microsoft
	StarPort	Windows XP/2000/Server 2003	RocketDivision
	UNH iSCSI	Linux	GNU GPL
	Linux-iSCSI	Linux	GNU GPL
	Open-iSCSI	Linux	GNU GPL
Target	Windows iSCSI Target	Windows XP/2000/Server 2003	Microsoft
	StarWind	Windows XP/2000/Server 2003	RocketDivision

	UNH-iSCSI	Linux	GNU GPL
	iSCSI Target	Linux	GNU GPL
	iSCSI Storage Server	Windows XP/2000/Server 2003	FalconStor
	SANmelody	Windows XP/2000/Server 2003	DataCore

表 6：iSCSI 軟體產品列表。

硬體產品

iSCSI 的硬體產品主要有 iSCSI 配接卡(HBA)、交換器、負載平衡器(Director)、儲存系統等類別。這些設備和一般以太網路設備的不同處在於，設備本身必須支援 iSCSI 協定，有些還必須具有做為 iSCSI SAN 的入口的功能。以交換器來說，就必須能夠連結儲存裝置與設備和支援 iSCSI 轉換成 SCSI 命令。從表 7 我們可以看出，iSCSI 儲存系統的產品最為豐富，這類產品多半是磁碟陣列或是磁帶櫃，具有作為 iSCSI SAN 入口的功能和管理系統。

iSCSI SAN 目前仍然不夠普及，所以大部份的消費都集中在企業等級。雖然可以靠軟體模擬，但是為了效能上的考量，特別是 target，還是需要添購硬體設備。對小規模的 iSCSI SAN 而言，iSCSI 儲存系統會是一個蠻不錯的選擇。現在 iSCSI 配接卡的價格仍然不低，普及度也還不夠，大部份族群的使用者都還沒有需求。再者 initiator 可以靠軟體模擬，所以 iSCSI 配接卡銷售情況可能還不如預期，但是仍有潛力，因為軟體模擬的效果目前仍未獲得肯定。

Product Type	Manufacturer	Model
ASIC	Intel	IOP331
	QLogic	ISP4010
iSCSI HBA	Adaptec	7211F
	Intel	Pro/1000 IP Storage Adapter
iSCSI Bridge	ATTO	IP Bridge 2500
	POTOMAC	iSCSI Bridge
Switch	SANRAD	V-Switch 2000
	Cisco	MDS 9216i Multilayer Fabric Switch
Director	McDATA	Interpid 10000 Director
	Cisco	MDS 9509 Multilayer Director
Router	HP	dubbed StorageWorks SR2122 iSCSI storage router
	IBM	TotalStorage SAN16M-R multiprotocol SAN router
Stroage System	EMC	CLARiiON CX500i
	IBM	IP Storage 200i ; TotalStorage DS400
	Dell	Dell/EMC AX100
	HP	StorageWorks MSA1000
	NetAPP	FAS900;NearStore R200

表 7：支援 iSCSI 的硬體產品列表。

影響 iSCSI 產品品質的因素

影響 iSCSI 產品的因素包含功能性、效能、符合性、互通性、以及儲存裝置本身。功能性是指產品提供的功能正確無誤以及管理介面是否簡單使用，例如是否能正常的建立連線、傳送檔案等等；效能著重在生產量與反應時間，例如輸入和輸出的速率以及傳送很多小檔案(例如相片)的反應時間。符合性是指產品對於標準規範符合的程度，例如是否在 Login Phase 做 CHAP 認證，才進入 Full Feature Phase 等；UHN-IOL 及 Ixia IxANVL 有提供對於 IETF 標準規範的測試套件。另外，對於網路儲存軟體管理，SNIA 制定了 SMI-S 標準，使得網路儲存軟體能管理實作此標準的裝置；互通性指的是不同廠商在同一個網路環境下產品共同工作的能力，這也是廠商最關切的重點之一，SNIA 的技術中心有提供收費的互通性測試，並且會舉

辦插拔大會，讓許多不同的廠商齊聚一堂測試互通性及符合性；最後儲存裝置是指磁碟陣列或磁帶機等等，是影響 iSCSI 產品品質最重要的因素。

軟體實際操作

連線之後會多出一顆磁碟機，可以像對本地磁碟機一樣的操作。我們利用 IOmeter(UNH-IOL 釋出的免費測試軟體)來比較 RAM disk、在 localhost 的 iSCSI RAM disk 以及在 LAN 的 iSCSI RAM disk 三者效能和反應時間的簡單數據。IOmeter 被設定為對磁碟做 sequential access，只做 read 的動作，每筆 I/O 要求大小為 2KB。由表 8 我們可以發現僅僅是對 localhost 的 iSCSI RAM disk 做存取已經和本機的 RAM disk 有了相當的差距，透過 LAN 的結果更是和 RAM disk 相差了十倍以上，可見處理 iSCSI 命令和網路頻寬對效能的影響很大，純粹靠軟體模擬出來的 iSCSI 效能可能還是差強人意，未來希望能以專門處理 iSCSI 的 ASIC 和 HBA 來降低 CPU 的負擔和未來高頻寬 Ethernet 的問世來解決頻寬的問題。

	Average I/O response time(ms)	Mbps	I/O/s	Environment
RAM disk	0.0109	173.75	89309.84	Initiator : Windows iSCSI Initiator 2.0 Target : RocketDivision StarWind
localhost iSCSI	0.1329	14.56	7457.44	Operating system : WindowsXP CPU : AMD XP 1.3G
LAN iSCSI	0.6436	3.02	1568.61	RAM DDR400 256MB × 3

表 8：實際操作結果。

5. 結論

網路儲存技術的世代已經來臨，從集中管理和不受連接距離限制的等等好處，我們可以看見未來網路儲存技術必定會深深地和我們的生活結合在一起，其中又以 SAN 最有潛力。block I/O 的方式適合大型資料庫的管理和維護，server-free backup 大大的降低了伺服器本身的負擔，SAN 的專業儲存環境已經深受肯定，iSCSI 的出現，更是帶入 IP SAN 的關鍵，不但降低了 SAN 的建置價格，而且解決了距離的問題，透過 IP 網路無遠弗界的連接，也讓 SAN 更容易普及。

iSCSI 把 SCSI 映射到 TCP/IP 上來，利用 SCSI 本來為了處理多個裝置之間的 I/O 所設計的架構和命令集。我們藉此了解了 iSCSI 制定的背景與想法，對 iSCSI 有了更深一層的了解。最後我們看到 iSCSI 的產品，除了硬體之外，在廠商的大力推動之下，各式各樣的模擬軟體也紛紛出籠，甚至在廣受歡迎的 Linux 作業系統上也有 open source 套件支援。隨著新產品的一一推出，產品測試變成不可或缺的一環。我們可以從功能性、效能等五個角度來探討對產品品質造成的影響，相信在不久的將來，我們可以看到 iSCSI 的蓬勃發展，為儲存技術寫下新的一章。

參考文獻

- [1] rfc3720 (IETF Standard Document), <http://www.faqs.org/rfcs/rfc3720.html/>.
- [2] SCSI standards architecture, <http://www.touchbriefings.com/pdf/22/lohmeyer.pdf/>.
- [3] IOmeter Documents, <http://www.iometer.org/doc/documents.html>.
- [4] iSCSI technologies white paper, http://www.storusint.com/storage_protocols/iscsi/iSCSI%20White%20Paper.pdf.
- [5] SAN & NAS comparisons, <http://www.nas-san.com/differ.html>.
- [6] LAN-free & server-free Backup, <http://www.brocade.china.com>.
- [7] SCSI standards architecture, <http://www.t10.org/scsi-3.htm>
- [8] An experience of iSCSI, <http://www.tn.edu.tw/chukk/iSCSI.html>.