快閃記憶體磁碟陣列管理技術
Software Orchestrated Flash Array Technology (SOFA)

周宗廉
Tsung-Lian Chou

中文摘要

雲端儲存必須滿足多使用者同時快速存取資料的需求，這對於雲端運算系統是很重要的。快閃記憶體相對於傳統的硬碟有存取更快和消耗功率更低的優勢，但也有在隨機存取時效能很低以及有限的壽命[1]等缺點。這項快閃記憶體磁碟陣列管理技術，就是提供一個快閃記憶體磁碟陣列，在不影響使用壽命和資料安全性的同時，又能提供高效能的存取。

快閃記憶體磁碟陣列管理技術是在一台包含快閃記憶體磁碟陣列的一般商用儲存伺服器上，建構一個系統軟體。現今用來集合一群SSD最直覺的技術，便是使用硬體或軟體的磁碟陣列（Redundant Array of Independent Disks, RAID）。現有RAID5常用於儲存系統，但並非專門為快閃記憶體設計，在隨機寫入的狀況，對效能有很大的傷害。本技術使用虛擬實體位址轉換的技術，將隨機位址轉換成連續位址，突破現有RAID5技術的效能與壽命瓶頸，在讀取寫入效能上提升10倍，達到透過網路每秒一百萬次輸出入的能力。本技術並透過跨磁碟間的平均耗損機制，延長磁碟陣列的壽命達2倍。除了最佳化底層儲存和上層網路之外，本技術還提供了虛擬磁碟管理功能，可提供快照、複製、和自動精簡配置等功能，而不會對效能有任何影響。在虛擬磁碟管理功能中還有服務品質保證的功能，透過合理分配整體的吞吐量，可保證每個虛擬磁碟的最小頻寬。為了增加儲存空間的利用率，本技術也包含了針對虛擬磁碟空間離線壓縮的功能，有效提升空間利用率達兩倍。

Abstract

Cloud storage is vital to the cloud computing system, as it requires satisfying multiple users for accessing the data at high throughput. Flash memory features the advantages of being faster and lower power-consumption as compared to conventional hard disk, but also has the issues of low performance for random access and lifespan. Software Orchestrated Flash Array Technology (SOFA) has constructed a flash memory disk array, providing high random access while not affecting the lifespan and the safety of the data.

SOFA is to build a system software, running on a commodity storage server with a flash memory disk array. Nowadays the intuitive approach to aggregate a bunch of disks is hardware or software RAID. Currently RAID5 is widely used in storage systems, but it's not designed for flash memories, and the performance will be highly impacted at random writes. The technology adapts virtual to physical address re-mapping, transferring random addresses to sequential addresses, and breaks through the performance bottleneck of RAID5, getting 10 times the read / write performance to make 1 million IOPS through network. The technology also doubles the lifetime of disk array of RAID5 by Global Wear Leveling across disks. In addition to the optimization of storage and network

levels, the technology provides Volume Manager, capable of taking snapshot, volume clone, and thin provisioning without performance impact. There is Quality of Service (QoS) function embedded in Volume Manager, which guarantees the minimum bandwidth of each volume by well-arranging the whole throughput. To increase the utilization of storage capacity, the technology also includes off-line compression for data volumes, which doubles the available capacity of the storage system.

關鍵詞(Key Words)

固態硬碟(Solid State Drive；SSD)

快閃記憶體陣列 (All-flash Array；AFA)

陣列磁碟 (Redundant Array of Independent Disks；RAID)

每秒輸出入次數 (Input Output Per Second；IOPS)

服務品質保證 (Quality of Service；QoS)

虛擬磁碟空間管理程式 (Volume Manager)

1．前言

固態硬碟(Solid State Drive, SSD)因為高速的存取效能、體積小與省電的特性，在全球市場需求上快速的演進，而除了消費市場外，企業或雲端服務系統更是未來的趨勢，近年來採用與預估要採用固態硬碟的企業組織數量已經翻倍增加，固態儲存產品正在成為未來的標準設備，加上快閃記憶體的價格與硬碟(HDD)價差已經在逐年減少，根據2009年工研院產業經濟與趨勢研究中心(IEK)的資料顯示(如圖1)，SSD與HDD差距由6倍降為3倍，這使得固態儲存再也不是成本高昂、供應固定客群的特定產品，而是具備經濟吸引潛力商品。

此外，根據市調機構國際數據資訊(IDC)資料顯示，2014年企業級固態硬碟市場出貨量整體產值為40億美元，預計至2017年將會達到70億美元的產值，且因應雲端發展趨勢，國際大廠Google、臉書(Facebook)、亞馬遜網路書店(Amazon)及百度網路服務供應商正在增設資料中心，擴大將傳統硬碟更換固態硬碟需求。若國內廠商能結合像SOFA這樣的儲存軟體，建立自主軟體關鍵技術，將有機會協助整體產業升級，進軍正高度成長的雲端儲存應用市場。因此提供企業或雲端服務的高速整合儲存設備，是本研究致力的目標，並希望能升級國內硬體供應產業，提供系統軟體加值服務及轉型機會。

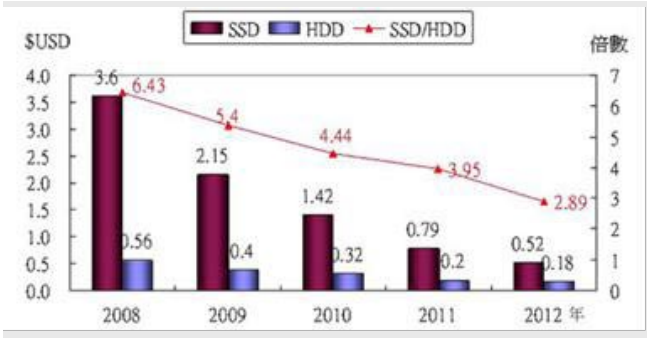

圖 1　2008至2012年每GB SSD/HDD成本價差

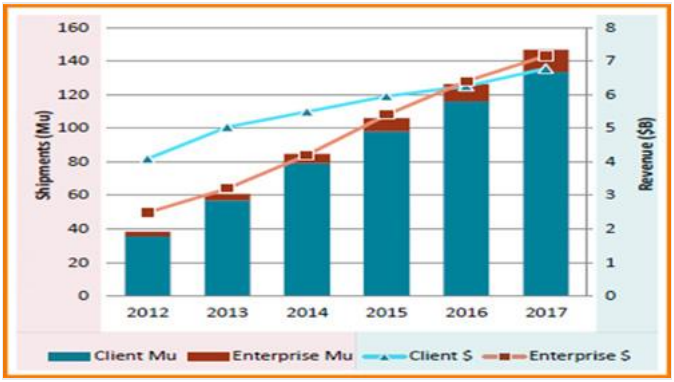

圖 2　2012至2017年全球SSD產值與出貨量 [10]

雲端需要之大容量且高隨機存取能力之固態儲存伺服器，這並不是單一個SATA-based 或是PCIe-based的SSD可以滿足的。因此，將多個SSD集合在一個儲存是無法避免的方向。而傳統儲存架構基礎建置的系統，都不是為了固態儲存或快閃儲存媒體而設計的，導致無法發揮

儲存技術的全部潛能，故本技術致力於發展專為固態儲存設計的軟體技術SOFA，並提供國際匹配的效能與技術。

本技術的操作方式是透過網路的通訊協定如iSCSI、SRP (SCSI RDMA Protocol) [2]或iSER (iSCSI Extensions for RDMA) [3]，來存取內部的儲存系統。如同圖3所示的測試環境，本技術運行在一台伺服器上，以三台PC當作client對伺服器進行隨機讀寫測試，在伺服器端可以量測到同時提供高達每秒一百萬輸出入次數(Input Output Per Second, IOPS)的儲存服務。

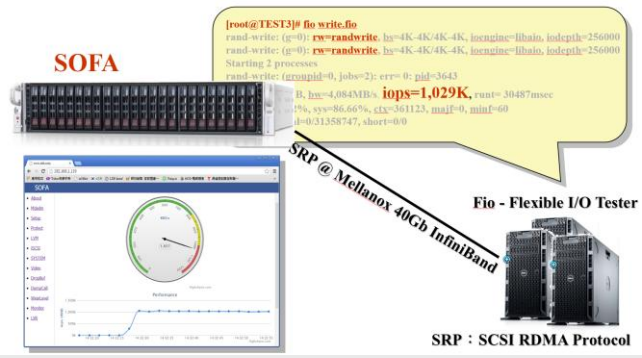

圖 3　SOFA的操作和測試方法

本技術的軟體架構如圖4所示，包含三大部分：網路通訊協定層(Network layer)、邏輯輸出入層(Logical IO layer)以及實體輸出入層(Physical IO layer)。由下一節開始，詳述各層分的功能和技術。

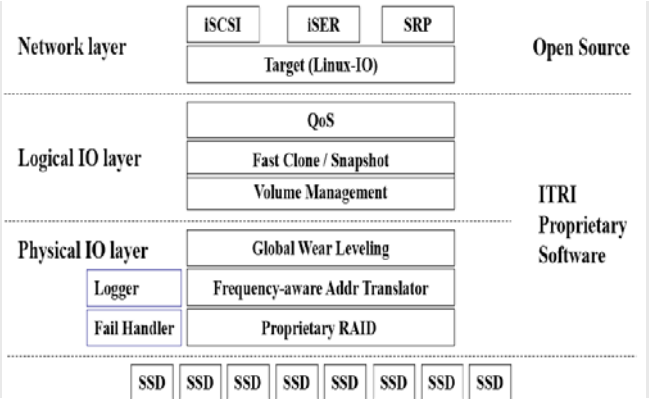

圖 4　SOFA的軟體架構

2．實體輸出入層

實體輸出入層主要是負責對底層NAND flash的處理。如前所述，用磁碟陣列(RAID)來集合一群SSD是普遍直覺的技術，但傳統RAID5的寫入方式如圖5

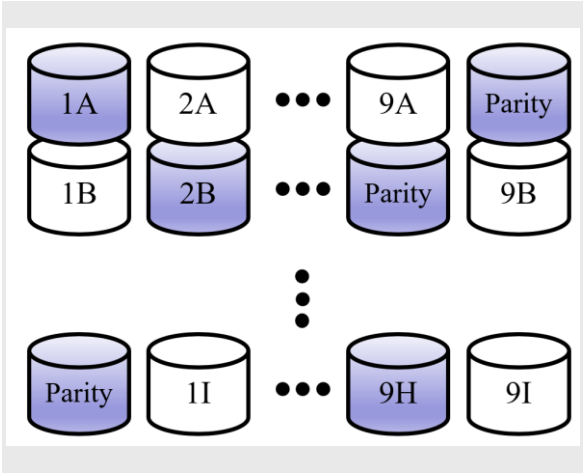

圖 5　標準的RAID5

所示，針對一個區塊隨機(random)寫入，需要讀取要寫入區塊以及同位檢查(parity)區塊，共兩筆讀取；之後要寫入待寫入區塊並重新計算新的同位檢查值再寫入，共兩個寫入。一個隨機寫入會製造四倍的讀寫次數，對於NAND flash的效能和壽命的傷害很大。

本技術的寫入方式則如圖6所示，為了解決NAND flash的寫入順序以及沒有抹除前不可重複寫入的困難，在架構上提出了動態位置映射，讓原本隨機的寫入位置被轉換成循序(sequential)位置。以十顆SSD做為一個讀寫單元來說，九筆隨機寫入會被轉換成九筆循序寫入，並在九筆寫入完成之後只需加一個多餘的同位檢查區塊。和傳統的RAID5相比，本技術整體所需要的讀寫次數少很多。

我們實際拿一個20顆SSD的伺服器來測試，其中每10顆當做一個保護列(stripe)，在80GB的區間隨機讀取40GB的資料。我們在測試中使用TPC-C的測試模式，目前，TPC-C作為一種由第三方非營利機構

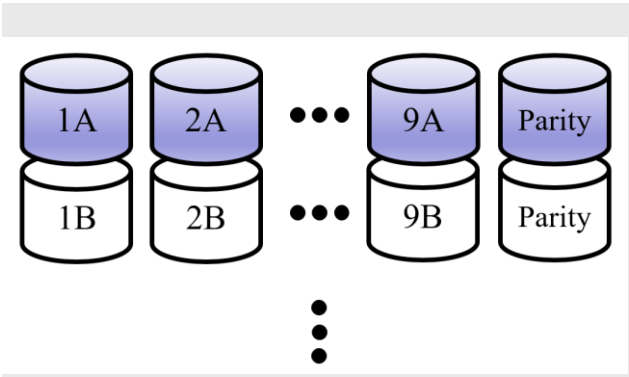

圖 6　SOFA的RAID保護機制

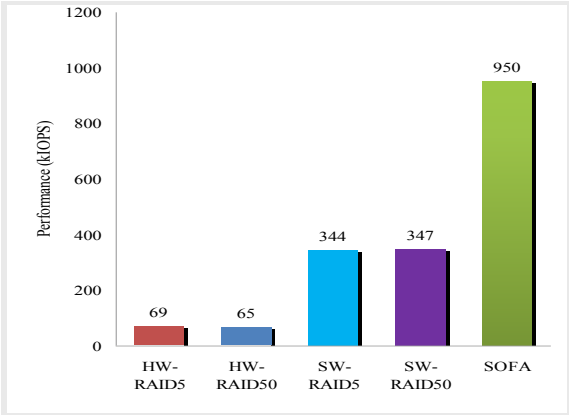

圖 7　SOFA和標準RAID5的讀取效能比較

(tpc.org)頒佈的基準測試指標，越來越多被各平臺廠商、應用系統提供商和最終用戶所引用，並被簡單地作為核心評估指標，應用於平台選型、系統規模設計等評估環節[4]。分別測試硬體RAID5、軟體RAID5和SOFA的效能，結果如圖7所示。

由測試結果可知SOFA和標準RAID5在讀取有數倍的效能差距。再用同樣的測試平台和測試模式測試隨機寫入如圖8所示，可以看到效能差距達到十倍以上。由於快閃記憶體的效能和壽命取決於寫入次數，SOFA有效降低多餘的寫入次數，因此可以提高效能和使用壽命。

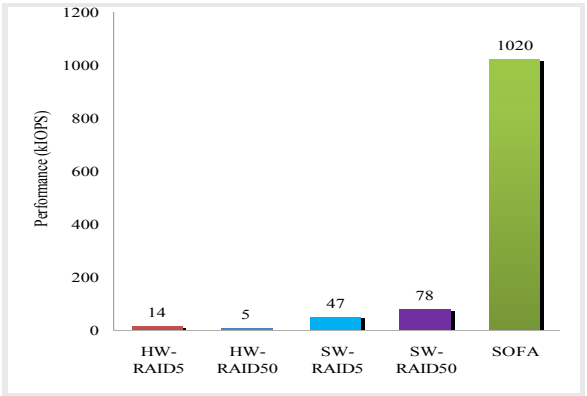

圖 8　SOFA和標準RAID5的寫入效能比較

再者，傳統RAID5不能保證進入此伺服器的寫入的行為被均勻打散到所有的磁碟，因此在系統運作一段時間後，部分磁碟會有較多的寫入次數，因為SSD是一種寫入次數有限制的磁碟，因此會導致部分磁碟較快損壞。考量到這種情形[5]，本技術集中管理寫入位置的分配，將隨機的寫入程序映射打散至各個磁碟，達到跨磁碟間的均勻寫入(global wear-leveling)，並且隨著映射程序，加上動態資料保護機制(data redundancy)，可保有資料保護機制又不會產生額外讀

圖 9　SOFA跨磁碟間的均勻寫入

寫導致效能受損。以圖9來看，傳統RAID5對磁碟寫入次數很不平均，最多和最少寫入次數會達到1.5倍的差距，而本技術對所有的磁碟寫入次數完全平均。

為了達到動態位置映射，將此映射表放在DRAM中並不符合成本，並且需考量斷電時所有mapping items喪失無法復原，為此建構日誌管理(log-based system)，其中的挑戰是要能儲存邏輯位置與實體位置對應關係的元資料(metadata)，但又不可以造成太多的額外元資料寫入影響負載，並且要在不正常斷電下快速的找回映射表重建資料。針對這個問題，我們提出有效的斷電還原機制，透過批次更新元資料(batch metadata)的概念，及時寫入元資料對應關係，減少寫入負擔；並透過遞層式的映射表，加速還原掃描的程序，使得系統對應關係可以完全復原。

動態位置映射後重複的位置不會覆寫，所以當每個抹除單位(Erase Unit, EU)寫滿時，要進行資料垃圾回收(Garbage Collection, GC)的程序，所謂的垃圾回收就是將該抹除單位上有效的資料複製到其他區域上，讓該抹除單位上的資料完全無效後，進行抹除的動作，因為每次的回收都會造成額外的讀寫運作(複製動作)，所以需要一套演算法決定如何挑選適當的回收抹除單位，使得額外的複製動作變最少，影響效能與壽命減到最低。其中挑選回收抹除單位時應考量如何減少回收量，本技術引入熱區群組(Hot-Page Grouping, HPG)的演算法，將資料根據邏輯位置寫入頻率做分類。若該資料的邏輯位置重複寫入時，則將該資料寫到溫度更高的區域上，反之若該區域上的資料被回收時，則複製到較冷的區塊。整個運作機制如圖10所示。

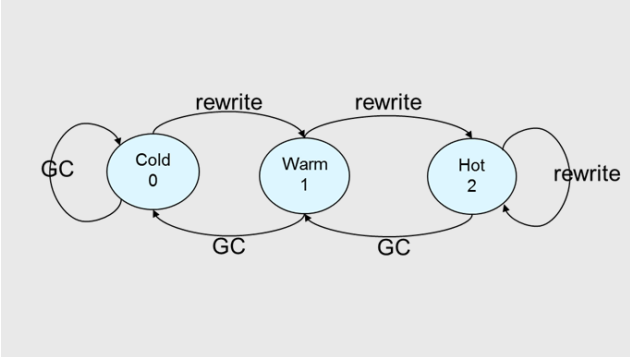

圖 10　Hot-Page Grouping機制

這樣的做法可以讓頻繁被寫入同時也是頻繁變成無效的資料，可以更集中於特定的回收區塊，而讓回收時需要搬移的有效資料變少，實際測試可以得到如圖11的結果。圖11的縱軸是在抹除單位裡面，有效資料頁面佔的比例，橫軸則是抹除單位的數量。沒有實作熱區群組的情況如綠線所表示，可以看到有很多的抹除單位當中，有效資料頁面所佔的比例分布比較平均，此時要回收無效資料的頁面，就要做比較多的抹除，並再把有效資料寫回，增加多餘的讀寫動作。另外，有實作熱區群組的情況如紅線所表示，大多數的抹除單位當中，有效資料頁面佔的比例很低，甚至幾乎都是無效資料，這些抹除單位可以直接回收而不需要再做有效資料寫回的動作；而有效資料比例很高的少數抹除單位，就暫時不做垃圾回收。由測試結果可以看出，用熱區群組的機制，可以讓無效資料大幅集中，減少多餘的抹除複製動作，間接降低效能與壽命影響。。

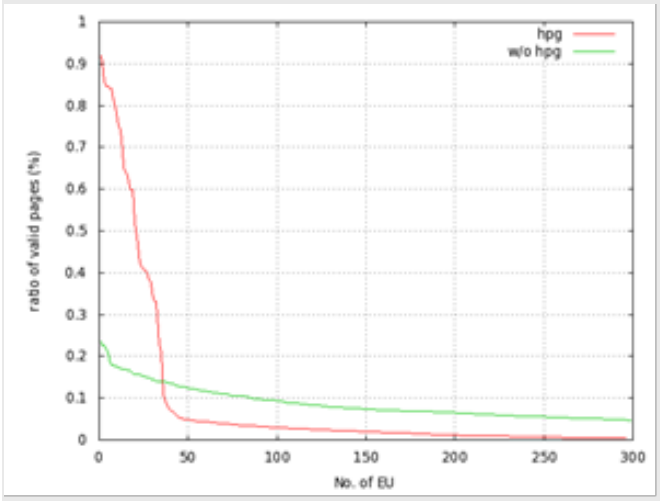

圖 11　Hot-Page Grouping效果

另外，本技術引入Stable Block Recycle (SBR)的方式，挑選有效資料量不再被更改的抹除單位，讓進行回收之後，減少搬移後的資料又變成無效的白工。圖12說明SBR的概念，不好的挑選如圖12左方所示，GC完的資料馬上成為無效；好的挑選如圖12右方所示，GC完之後不會馬上變成無效的資料。

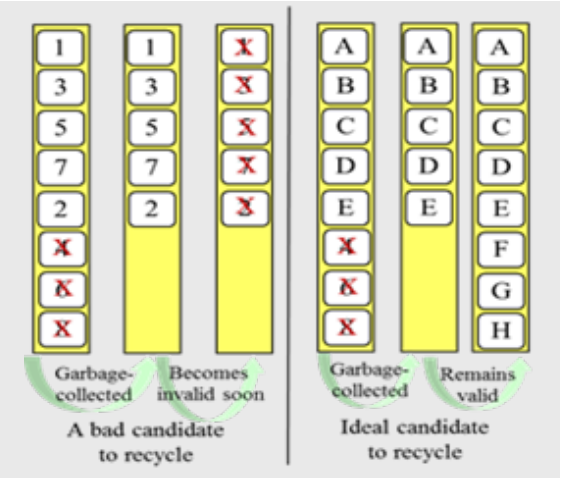

圖 12　Stable Block Recycle概念圖[6]

3．邏輯輸出入層

為了可以支援多用戶使用，在實體輸出入層之上，本技術針對實體輸出入層所提供整體的儲存空間，配備本團隊自行研發的虛擬磁碟空間管理程式，功能包含建立和刪除虛擬磁碟空間、快照、複製等功能，如圖13所示。

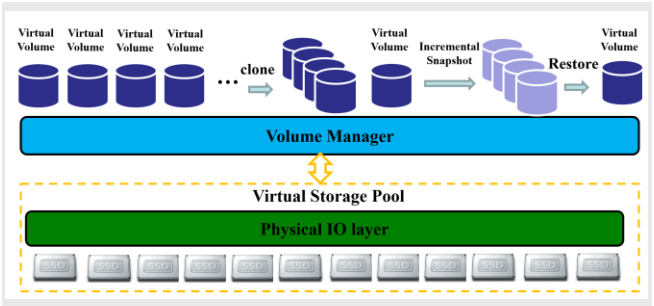

圖 13　虛擬磁碟空間管理程式

其中針對虛擬磁碟空間，使用的是自動精簡配置的機制，也就是將實際的實體儲存空間分配給有即時需求的虛擬磁碟空間；而不是一開始就配置全部的空間。之後，依照實際使用量自動提高儲存空間，避免浪費過多磁碟儲存設備。另外本技術在快照和複製虛擬磁碟空間的執行中，　並不需要複製資料或元資料，所以可已進行非常快速的快照和複製虛擬磁碟空間。而針對快照，除了完整備份之外，還有漸進式差

異性備份(incremental backup)，也就是說只針對上次提取快照到目前的差異來備份，更能節省備份的時間和空間。

目前業界常用開源碼的虛擬磁碟空間管理是Linux的Logical Volume Manager (LVM)，但是普遍認知LVM經過多次提取快照之後，使用效能就會受到嚴重的影響，甚至達到90%的降幅。因為LVM每筆快照資料都指向目前的區塊空間，對被提取快照的區塊寫入時，必須先把該區塊的資料讀出再寫入到新的區塊空間，然後才能真正寫入資料。於是一筆寫入就會產生一筆讀取和兩筆寫入；當被提取快照次數增加時，會產生多餘的讀取寫入增加的情形會更嚴重。本技術重新實作虛擬磁碟空間管理程式，對被提取快照的區塊寫入時，另外找一個區塊空間寫入資料，而不影響原先的區塊資料；另外每筆快照是以鏈結的形式彼此連結，因此不論提取多少快照，以上述機制寫入資料就不會影響快照的連結。這樣的設計，可以確保被提取快照區塊的寫入效能不會受到影響。本技術的做法以及和LVM的比較，如圖14所示。

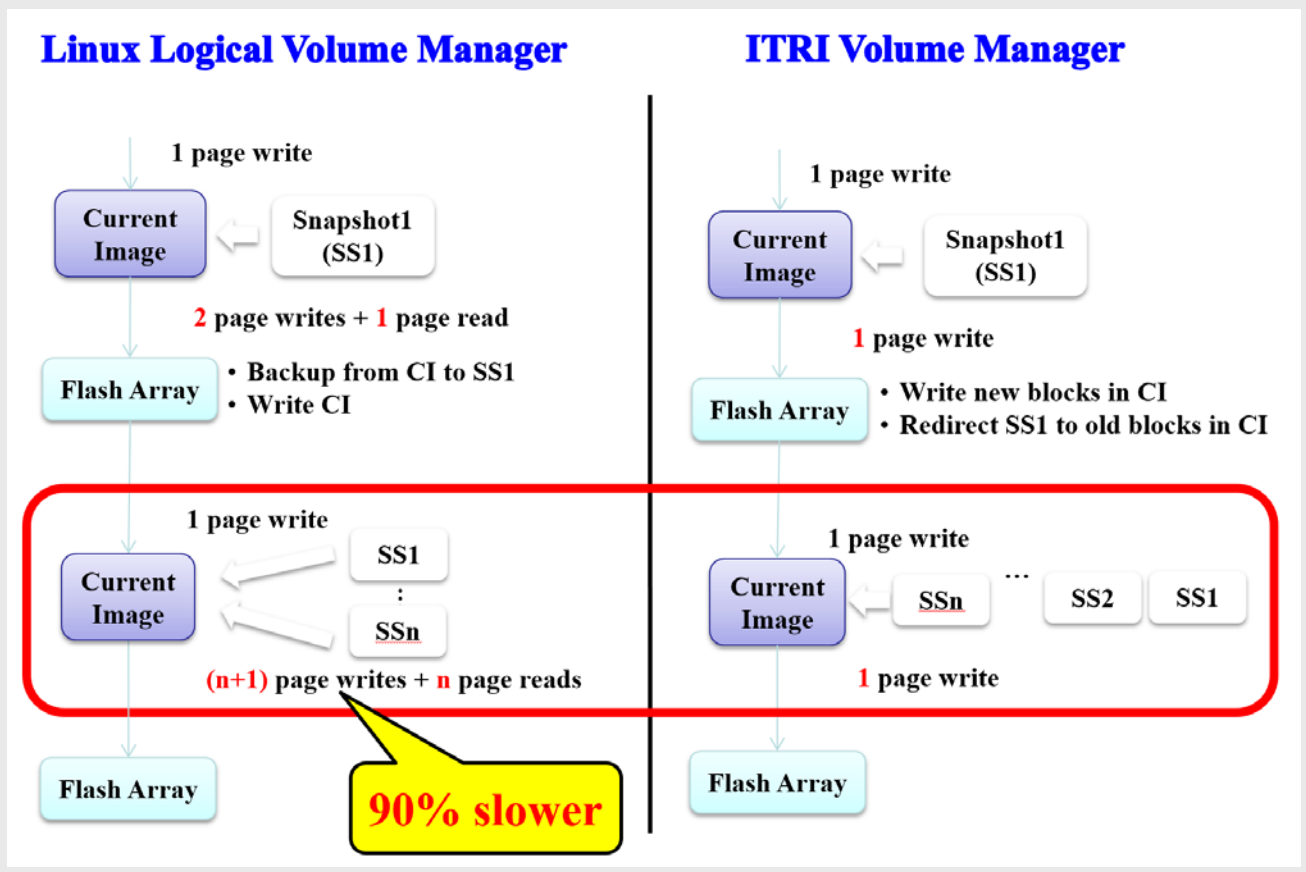

圖 14　SOFA和LVM的比較

另外，本技術還包含服務品質保證的功能。針對每個虛擬磁碟空間，可以設定最小保證頻寬；而非部分宣稱具備QoS的產品，僅僅是針對某些虛擬磁碟空間設定最大頻寬，防止單一虛擬磁碟空間佔去所有頻寬。服務品質保證的功能如圖15所示，圖中有三個虛擬磁碟空間，所設定的最小保證頻寬如紅色、藍色和橘色的標線，於是我們可以看到三個虛擬磁碟空間會依照設定值運行，頻寬會一直保持設定的最小頻寬之上。

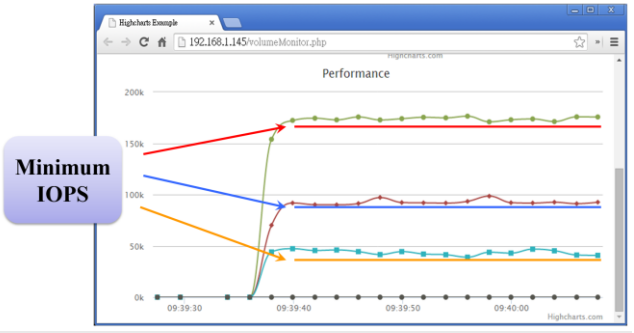

圖 15　QoS保證最小頻寬

而當某個虛擬磁碟空間暫時閒置的時候，該虛擬磁碟空間的頻寬可以分享出來，提供其他虛擬磁碟空間使用；但是在這種情形下，所有虛擬磁碟空間仍然被限制在所設定的最大頻寬之內。如圖16所示，紅色和綠色的虛擬磁碟空間分別的最小保證頻寬是藍色和黃色的標線，紅色的虛擬磁碟空間本來具有較高的保證頻寬，當兩者同時運行時，就依照各自保證的頻寬設定；但當紅色的虛擬磁碟空間閒置時，頻寬可以分享給綠色的虛擬磁碟空間使用。當然，閒置時間分享頻寬還是會有限度，本項功能可以協助營運商規畫更有彈性的分級收費機制，需要更高頻寬的使用者就必須繳交較高的費用。

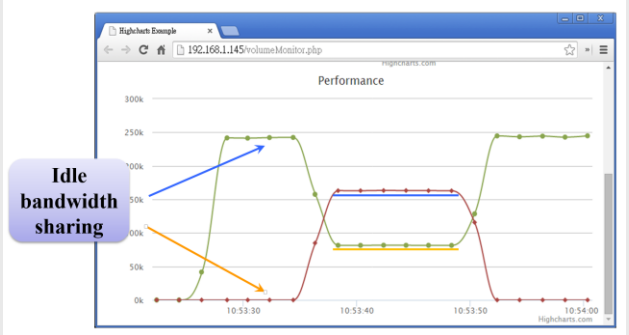

圖 16　QoS分享閒置頻寬

4．網路通訊協定層

有了針對快閃記憶體設計的實體輸出入層，以及適用於多用戶使用且對效能無損的邏輯輸出入層，在本機就可以建構高效能的儲存系統。因為雲端儲存系統還需要網路對外提供服務，本技術在軟體架構中還包括網路通訊協定層，透過網路傳送資料。這些通訊協定目前主流是來自於開源碼的iSCSI、SRP和iSER。本團隊針對開源碼的原始碼進行優化，切分執行續(thread)並對個別執行續分配不同的CPU資源。經過這些優化，單純以通訊協定層測試傳輸效能，在10Gb乙太的網路卡上，由原先的50K IOPS，提升到大約250K IOPS。以每個區塊4KB來計算，頻寬已經達到接近10Gb/s，符合理論值。另外在40Gb Infiniband網路卡，也由原先的500K IOPS，提升到大約一百萬IOPS，符合40Gb/s的頻寬理論值，如圖17所示。

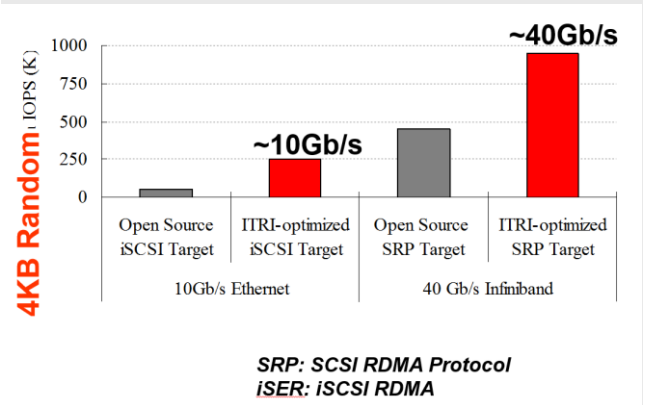

圖 17　網路通訊協定優化前後比較

5．動態資源配置

除了實體輸出入、邏輯輸出入和網路通訊協定三層軟體之外，為了要充分配置和利用硬體效能，本技術還有針對不同硬體平台，動態調整CPU資源的功能，以發揮最大的效能。

在分配硬體資源之前，我們先針對整套軟體區分成幾個群組，分別是實際讀寫磁碟的實體輸出入層、網路通訊協定層、更上層處理網路封包的網路層、以及其他包括伺服器上把磁碟連接到主機板的主機總線適配器(Host Bus Adapter, HBA)卡驅動程式。接著各別群組獨立測試，取得在無限大的CPU資源情況下，各模組所需要的最小CPU資源為何。以網路層來說，在使用三張10Gb乙太網路卡的情形下，每張網路卡使用三個輸出入貯列(Tx / Rx Queue)，各用一個core thread來處理，總共花費3個core threads，可以達到一百萬IOPS。網路封包頻寬足夠之後，再考慮網路通訊協定層的處理能力。以10Gb乙太網路使用的iSCSI來測試，兩個讀寫的執行緒需要12個core threads才能處理網路層傳輸的封包。接著考慮實際寫到磁碟的效能，實體輸出入層需要10個core threads，另外三張主機總線適配器卡也各需要一個core thread，才能達到和上層匹配的效能。如果沒有刻意分配CPU資源，得到的整體效能會和最佳配置有著極大的差距，如圖18所示。

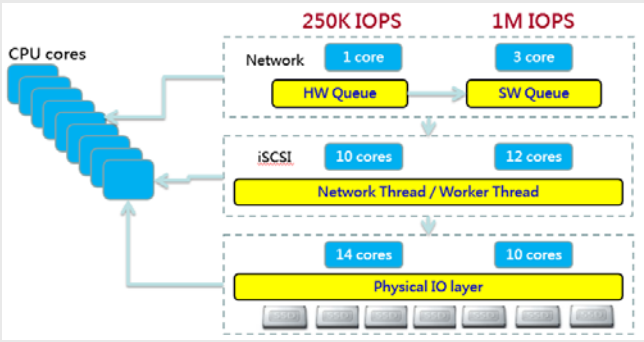

圖 18　CPU不同配置造成的效能差異

6．國際競爭者比較

本技術和目前世界領導品牌的全快閃記憶體磁碟陣列(All Flash Array, AFA)產品相比，無論是在功能或是效能，都毫不遜色。特別是在效能方面，在同等級產品中，本機能達到一百萬IOPS已經不容易；如果再加上由網路端得到的整體效能，能達到一百萬IOPS的產品更是寥寥可數，比較表如下。

	SOFA	Violin Memory	IBM FlashSystem V900	Pure Storage FlashArray 450	HP 3PAR StoreServ 7450
Flash unit	SSD	SSD	MLC flash module	SSD	SSD
Performance (4KB IOPS)	1M IOPS	1M IOPS	800K IOPS (Read/write 70%/30%)	200K IOPS	900K IOPS (Read)
Size	2U	3U	2U	4U	2U
Power (watt)	500	1600	625	1000	500
Proprietary RAID	Yes	Yes	No	Yes	No
Fast Clone / Snapshot	Yes*	No	No	No	No
QoS	Yes	No	No	No	No
Global Wear Leveling	Yes	Yes	No	Yes	No
Optimized Network	Yes	Yes	No	No	No

圖 19　本技術和業界產品比較[7][8][9][10]

*：正在進行中的功能，將於2016年底完成

7．結論

本技術所建構的快閃記憶體磁碟陣列，針對底層的快閃記憶體特性，做出獨特的設計，保留RAID5資料保護的精神，但避免了傳統RAID5對於快閃記憶體效能的傷害；不僅提升效能、維持使用壽命，也保護資料安全性。在多使用者的環境，更增加虛擬磁碟管理的功能，同時維持底層的高讀寫效能。針對雲端應用，優化了網路傳輸協定，達到整體效能的最佳化。最後，把硬體資源配置納入考量，以達到相容於各種硬體平台的目的。

現今SSD磁碟越來越普及，使用快閃記憶體磁碟陣列是不可抵擋的趨勢。綜合上述的技術突破以及達到的功能和效果，本技術將可以整合硬體對雲端運算提供高效能的儲存服務，同時也會大幅提升雲端使用者的體驗。

參考文獻

[1]Samsung NAND Flash, Revision 2.0, Samsung Electronics Japan, Tokyo, 1994/1995

[2] L. P. Chang. On e_cient wear-leveling for large-scale ash-memory storage systems. In the 22nd ACM Symposium on Applied Computing (ACM SAC).

[3]TPC-C. http://www.tpc.org/tpcc/default.asp

[4]SRP. http://www.cisco.com/c/en/us/td/docs/server_nw_virtual/open_fabrics_enterprise_distribution/ofed_host_driver/release1-1/user/guide/ofed_ug/srp.pdf

[5]iSER. http://conferences.sigcomm.org/sigcomm/2003/workshop/niceli/papers/NICELI-p10-iSER.pdf

[6] T. C. Chiueh, W. Tsao, H. C. Sun, T. F. Chien, A. N. Chang, and C. D. C! hen. Software Orchestrated Flash Array. In Proceedings of International Conference on Systems and Storage, ACM, pp. 1-11, 2014

[7] Violin Memory, Violin 7300 Flash Storage Platform with Concerto OS 7 Enterprise Data Services. https://www.violin-memory.com/wp-content/uploads/resources/Violin-Product-Overview-7300-FSP.pdf

[8]IBM FlashSystem 900. http://www-03.ibm.com/systems/storage/flash/900/overview.html

[9]Pure Storage FlashArray The All-Flash Enterprise Array. https://www.purestorage.com/content/dam/purestorage/pdf/datasheets/Pure_Storage_FlashArray_Datasheet.pdf

[10]HPE 3PAR StoreServ 7450 Storage. http://www8.hp.com/h20195/v2/GetPDF.aspx/c04111384.pdf

[11]Micro-Electronics http://www.mem.com.tw/article_content.asp?sn=1402270010

作者簡介

周宗廉

現任職於工研院資訊與通訊研究所資料中心系統軟體組。國立台灣大學電機工程碩士。專長於多媒體系統、高速電路設計、雲端儲存系統