如何用好你的 SSD

■作者：陳定寶
Lightbits Labs 解決方案架構師

在過去十幾年中，CPU 的性能提升了 100 倍以上，而傳統的 HDD 硬碟 (Hard Disk Drive) 才提升了 1.5 倍不到，這種不均衡的計算存儲技術發展，極大地影響了 IT 系統整體性能的提升。直到固態硬碟 SSD (Solid State Drive) 被發明出來，其性能有了顛覆性的提升，才解決了存儲的瓶頸問題。然而，SSD 作為一項新技術，仍然存在一些固有的缺陷，如何充分發揮 SSD 的優勢，是一個值得研究的方向。下面從性能、持久性、使用成本等方面對此話題做一些探討。

一、如何充分發揮出 SSD 的性能

首先，我們來看看傳統 HDD 的使用方式：

1. 協定一般都採用 SAS、SATA 介面；
2. Linux 的 IO 調度需要用電梯演算法來對 IO 進行重排以優化磁頭的路徑；
3. 企業級存儲通常使用 Raid 卡做資料保護。

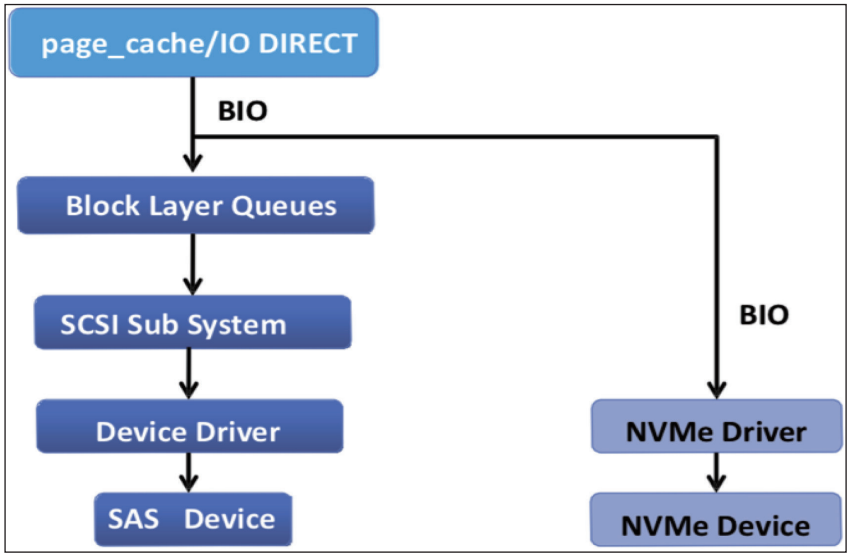
在介面協定方面，隨著 SSD 的發明，NVMe 協議應運而生。相較於 SAS、SATA 的單佇列機制，NVMe 最多可以有 65535 個佇列，並且直接採用 PCIe 介面，消除了鏈路和協定瓶頸。

在控制卡生態方面，各大廠商也紛紛推出自己的 NVMe 控制卡晶片，有 PMC (現屬於 Microchip)、LSI、Marvel、Intel、慧榮以及國內的得瑞等，技術也已經非常成熟。

在 Linux 驅動和 IO 協議棧方面，也做了相應的優化，如圖 1 所示，NVMe 驅動可以直接繞過那些傳統的、專為 HDD 設計的調度層，大大縮短了處理路徑。

到目前為止，為了充分發揮 SSD 的性能，上面提到的三個傳統 HDD 的問題中前兩個已經得到了解決，然而在企業級市場上，基於 NVMe 的 Raid 始終沒有太好的方案。傳統企業最廣泛使用的 Raid5/Raid6 資料保護機制 (N+1, N+2)，通常是把資料條帶化分片，然後計算出冗餘的 Parity Code (同位碼)，將資料存放到多塊硬碟，寫入新資料通常是一種“讀改寫”的機制。這種機制本身就成為了性能瓶頸，並且“讀改寫”對 SSD 的使用壽命有很大的損耗。另外，因為 NVMe 協議把控制卡放到了 NVMe 盤的內部，IO 都由 NVMe 盤內部的 DMA 模組來完成，這就給基於 NVMe 的 Raid 卡設計帶來了更大的困難。目前市場上這類 Raid 控制卡可用方案也很少，並

圖 1：NVMe 驅動可以直接繞過那些傳統的、專為 HDD 設計的調度層，大大縮短了處理路徑。



且性能上也無法發揮出 NVMe 的優勢，因此沒能被廣泛使用。

基於目前這種狀況，很多企業級存儲方案仍然在使用 SAS/SATA 的 SSD 加傳統的 Raid 卡，這種方式又會出現前面已經解決的兩個問題，SSD 的性能得不到充分發揮。

然而，這樣的情況也在發生改變，由 Lightbits Labs 發明的 NVMe over TCP (NVMe/TCP) 存儲集群解決方案就對這個問題做了很好的處理。該解決方案通過自主研發的一塊資料加速卡，採用 Erasure Code (糾刪碼) 機制可以做到超過 1M IOPS 的隨機寫性能，並且可以避免 " 讀改寫 " 帶來的使用壽命損耗。另外，Lightbits 提出了 Elastic Raid 機制，該機制提供彈性的 N+1 保護 (類似於 Raid5)，相較于傳統的 Raid5 需要熱備盤或者需要及時替換損壞盤，該機制在一塊硬碟發生損壞之後能自動平衡形成新的保護。比如一個節點內原先有 10 塊盤，採用 9+1 的保護，當某塊盤損壞後，系統會自動切換成 8+1 的保護狀態，並且把原先的資料再平衡到新的保護狀態，從而在可維護和資料安全性方面實現了大幅提升。此外，該資料加速卡還能做到 100Gb 的線速壓縮，顯著提高了可用容量，進而能大幅降低系統使用成本。

二、如何提升 NVMe 盤的持久性

目前使用最廣泛的 SSD 是基於 NAND 顆粒的，而 NAND 一個與生俱來的問題就是持久性 (endurance)。並且隨著技術的發展，NAND 的密度也越來越高，最新一代已經到了 QLC (4bits per Cell)，同時每個 Cell 可被擦寫的次數也在減少 (1K P/E Cycles)。發展趨勢如圖 2 所示。

另外，對 NAND 的使用有一個特點，就是可擦除的最小單位比較大，如圖 3 所示，寫的時候可以 4KB 為單位往裡面寫，但是擦除的時候 (比如修改原有資料) 卻只能以 256KB 為顆粒來操作 (不同的 SSD 大小不一樣，但原理都一樣)。這就容易形成空洞而觸發 SSD 的 GC (Garbage collection) 資料搬移，進而導致所謂的寫放大現象，對盤的持久性會產生進一步影響。

在企業級存儲中，通常使用 Raid5/6 這種 " 讀改寫 " 的機制，會對盤的寫運算元量進一步放大，一般使用場景下大約是直接寫入方式的 2 倍損耗。此外，很多 Raid5 還會啓動 Journal 機制，對盤的使用壽命會進一步損耗。

最後，對於最新的 QLC 來說，使用中還需要考慮另一個因素——Indirection Unit (IU)。比如有些

圖 2：NAND 發展趨勢

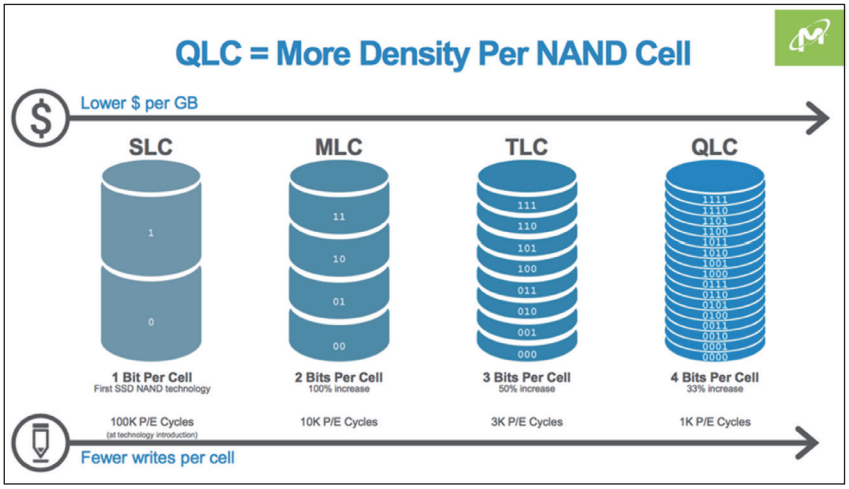
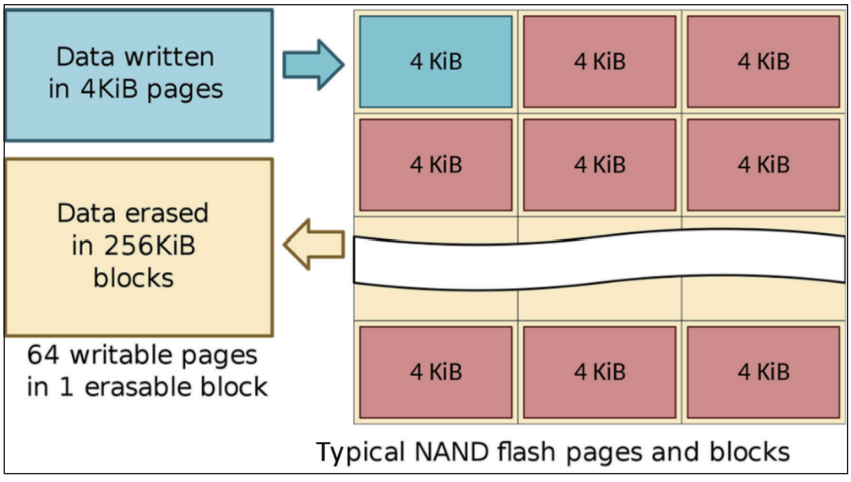


圖 3：NAND 使用有一個特點，就是可擦除的最小單位比較大



QLC 盤使用 16KB 的 IU，如果要寫入較小的 IO，也會觸發內部“讀改寫”，對使用壽命又多一重損傷。

由此可以看出，基於 NAND 的 SSD 還是比較嬌弱的。不過，只要能正確地使用，還是可以避免這些問題。比如以某常用的 QLC 盤為例，通過如下兩組關於性能和持久性相關的參數可以看出，在持久性上順序寫是隨機寫的 5 倍，而性能更是 26 倍：

- 順序寫 0.9 DWPD, 隨機 4K 寫 0.18 DWPD；
- 順序寫 1600 MB/s, 隨機 4K 寫 15K IOPS(60MB/s)。

通過上面的分析發現，能把盤使用在一個最佳的工作狀態至關重要。好消息是目前一些先進的解決方案，比如 Lightbits 的全 NVMe 集群存儲解決方案就可以解決這個問題。該方案通過把隨機 IO 變成順序 IO 的方式，以及獨有的 Elastic Raid 技術避免了 Raid“讀改寫”的弊端，從而能大幅提高盤的持久性及隨機性能。

三、如何降低使用成本

由於 SSD 相對於 HDD 而言是一項新技術，再加上產業的生產規模和需求量的矛盾，目前價格相比 HDD 仍然偏高。那麼如何降低 SSD 使用成本就變得非常重要。

降低使用成本最重要的一環就是要把 SSD 充分使用起來，無論是容量還是性能。不過就目前而言，大多數 NVMe 盤都是直接插在應用伺服器上使用，而這種方式非常容易造成大量的容量和性能浪費，因為只有這台伺服器上的應用才能使用它。根據調研發現，使用這種 DAS (Direct Attached Storage，直連式存儲) 方式，SSD 的利用率大概在 15%-25%。

針對這個問題比較好的解決方法是近幾年來市場上被廣泛接受的“解耦合”架構。解耦合之後，把所有的 NVMe 盤變成一個大的存儲資源池，應用伺服器用多少就拿多少，只要控制總數量夠用就行，可以非常容易地將利用率推到 80%。另外，因為資源集中起來，可以有更多的手段和方法用於降低成本，比如壓縮。例如，平均應用資料壓縮比在 2:1，就相當於多了一倍的可用容量，也相當於每 GB 價格降了一半。當然壓縮本身也會帶來一些問題，比如壓縮本身比較費 CPU，另外很多存儲解決方案在開啓壓縮之後性能就會大大降低。

針對壓縮方面的問題，Lightbits 的 NVMe/TCP 集群存儲解決方案可以通過存儲加速卡來予以解決。該卡可以做到 100Gb 的線速壓縮能力，並且不消耗 CPU，不增加延遲。利用這樣的解決方案，壓縮功能幾乎沒有額外的成本。此外，正如前面在介紹提高持久性時所提到的，Lightbits 解決方案能提高使用壽命並支持使用 QLC 盤，從整個使用週期來看，在使用成本方面也會有非常大的降低。總的來說，通過解耦合提高使用效率，壓縮提高可用容量，優化提高使用壽命或啓用 QLC，經過這樣的重重提升，SSD 的使用成本可以得到極大的控制。

以上從性能、持久性、使用成本三個方面分析了如何用好 SSD 盤，可以看到要用好 NVMe SSD 盤還是不容易的。因此，對一般使用者而言，選擇一個好的存儲解決方案就至關重要。為此，以色列創新公司 Lightbits 以充分發揮 NVMe 盤的最大價值為使命，發明了 NVMe/TCP 協議，並推出了新一代的全 NVMe 集群存儲解決方案，可以說明使用者輕鬆地將 SSD 盤用好。 CTA