爲什麼 NVMe/TCP 是資料中心的更優選擇

■文：Lightbits Labs 供文

自從 NVMe 作為高性能固態硬碟 (SSD) 的最新協議出現以來，已經改變了儲存行業。

NVMe 最初是為高性能直連式 PCIe SSD 設計的，後來以 NVMe over Fabrics (NVMe-oF) 的形式進行了擴展，以支持機架級 (rack-scale) 的遠端 SSD 儲存池。業界普遍認為，這種新的 NVMe-oF 模式將取代 iSCSI 協定，作為計算伺服器和存儲伺服器之間的通信標準，並成為解耦合儲存 (disaggregated storage) 方案的預設協議。然而，NVMe-oF 最初的部署選項僅限於光纖通道 (Fibre Channel) 和遠端直接記憶體存取 (Remote Direct Memory Access，RDMA) 結構。

如果我們能夠提供一種新的、更強大的技術，既能提供 NVMe 的速度和性能，又不需要高昂的部署成本和複雜性，將會如何？

NVMe over TCP (NVMe/TCP) 就可以使用簡單高效的 TCP/IP 結構將 NVMe 擴展到整個資料中心。

本文將描述 NVMe/TCP 如何成為面向現有資料中心的一種更優技術及其可提供的優勢。這些優勢包括：

■支援跨資料中心可用區域的解耦合
■利用無處不在的 TCP 傳輸和低延遲、高並行的 NVMe 協議棧 (Protocol Stack)
■無需在應用伺服器端進行更改
■可提供類似直連式 SSD (DAS) 性能和延遲的高性能 NVMe-oF 解決方案
■針對 NVMe 優化的高效、精簡的塊儲存網路軟體棧
■可並行訪問針對當今多核應用 / 客戶伺服器優化的儲存
■標準的 NVMe-oF 控制路徑操作

NVMe/TCP 概述

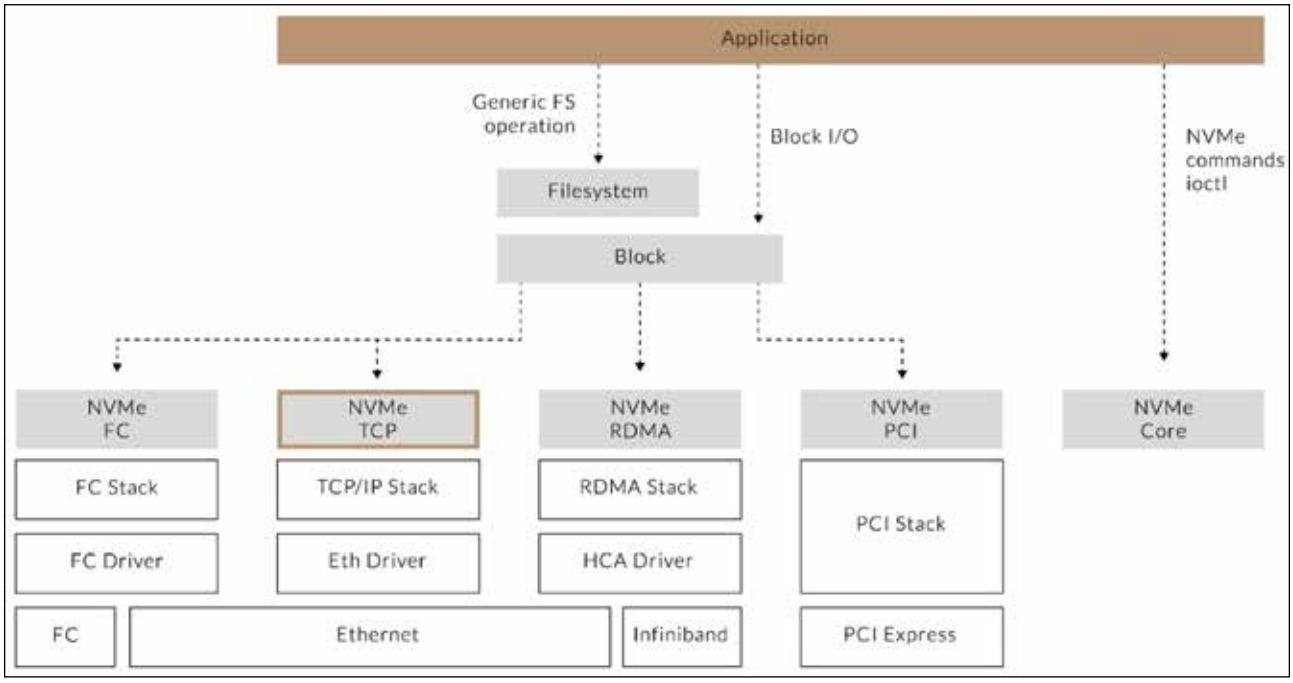
NVMe 規範已經成為高性能 SSD 的最新協議。

與 SCSI、ISCSI、SAS 或 SATA 介面不同，NVMe 實現了針對多核伺服器 CPU 優化的簡化命令模式和多佇列 (queue) 體，系結構。NVMe-oF 規範擴展了 NVMe，實現了通過網路共用 PCIe SSD，其最初是使用 RDMA 結構來實現的。如今，Lightbits Labs 與 Facebook、英特爾和其他行業領先企業合作來擴展 NVMe-oF 標準，以支援與 RDMA 結構互補的 TCP/IP 傳輸。

基於 NVMe/TCP 的解耦合存儲方案具有簡單、高效等明顯優勢。TCP 具有普及性、可擴展性和可靠性，對於短暫連接和基於容器的應用而言是一種理想的選擇。此外，通過 NVMe/TCP 遷移到共用快閃記憶體 (Flash) 也不需要更改資料中心的網路基礎設施。無需更改基礎設施意味著可以輕鬆地跨資料中心進行部署，因為幾乎所有資料中心網路都被設計為可支援 TCP/IP。

基於 NVMe/TCP 協定的廣泛行業合作意味著該協議從設計之初就具有廣闊的生態系統，並且支援任何作業系統和網路介面卡 (NIC)。NVMe/TCP Linux 驅動程式原生匹配 Linux 內核，可以使用標準的 Linux 網路通訊協定棧和 NIC，無需任何修改。

圖 1:NVMe/TCP 可與 Linux 內核中的現有 NVMe 協議無縫整合 (seamless integration)



這種很有前景的新協議為超大規模資料中心量身定制，可以在不改變底層網路基礎設施的情況下輕鬆實現部署。

現在的資料中心如何處理存儲

直連式存儲架構與 NVMe

NVMe 存儲協定旨在從固態驅動器 (SSD) 中提取全部性能。

NVMe 協議中所設計的並行能力有助於實現這種性能。NVMe 並未使用單一佇列的 iSCSI 模式。取而代之的是，NVMe 在 CPU 子系統和存儲之間可支援多達 64000 個佇列。

SSD 是使用多個並行通信通道與多個 SSD 存儲位置相連接的並行設備，這意味著 SSD 可以在大規模的並行流中高效地接收資料。在 NVMe/TCP 協定出現之前，利用這種並行模式的最簡單方法就是將 NVMe SSD 直接安裝到應用伺服器上。換句話說，你必須使用 DAS 模式來構建自己的存儲基礎設施。

使用 DAS 方法，應用可以受益於以下方面：

■多個 CPU

■多個 NVMe I/O 佇列

■並行 SSD 架構

對業界而言，挑戰在於將 SSD 從可能具有多餘容量的獨立伺服器轉移到具有更高基礎設施利用率且不會損失 DAS 性能收益的共用存儲解決方案。因此，所有 NVMe 解耦合技術的目標都是在共用 NVMe 解決方案中實現 DAS 性能。

前一代基於 IP 的存儲架構

以前，iSCSI 標準是通過 TCP/IP 網路連接至塊儲存的唯一選擇。它是在世紀之交開發的，當時大多數處理器都是單核元件。

在 SCSI 中，應用 (啓動器 initiator) 和儲存 (目標器 target) 之間只有一個連接。對於 iSCSI，也是只有一個 TCP 通訊端 (socket) 將用戶端連接至塊儲存伺服器。

現在，資料中心的處理器都是大規模並行多執行緒 (multithreading) 元件。當今處理器的這種複雜性要求對可用的儲存協定進行徹底改革。其結果就是 NVMe 作為 SATA 和 SAS (串列連接 SCSI) 的替代者出現了。所有那些早期協定的開發都是基於一

個串列的旋轉型磁碟機。

非易失性記憶體 (NVM) 是一種並行儲存技術，它不需要一個或多個碟片在一個或一組磁頭下面旋轉。使用 NVM 存放裝置，可以並行訪問許多存儲單元，且具有較低的延遲。

毫無疑問，iSCSI 仍然適用於具有低到中等存儲性能要求的應用場景。然而，iSCSI 卻不能滿足 I/O 密集型應用的要求，這類應用需要在大規模下實現低延遲。

其他替代方案和 NVMe/TCP 解耦合方案

RDMA、基於聚合乙太網的 RDMA (RoCE)，以及基於光纖通道的 NVMe (NVMe over FC)，也是試圖解決解耦合問題的其他網路儲存協定。然而，這些替代方案要求在兩端 (應用伺服器和儲存伺服器) 都安裝昂貴的特殊硬體，例如具備 RDMA 功能的 NIC。此外，安裝了 RDMA 硬體之後，在你具備 RDMA 功能的交換結構中配置和管理流控制也是很複雜的。

RDMA 確實提供了適用於某些高性能計算環境的性能，但它要求更高的成本，並且需要進行非常複雜的部署。

TCP/IP 已被證明可以在超大規模環境中可靠、高效地工作。NVMe/TCP 繼承了這種可靠性和效率，它可以作為 RDMA 的互補解決方案與之共存，也可以完全取代 RDMA。

資料中心中的快閃記憶體解耦合和 NVMe/TCP 解決方案

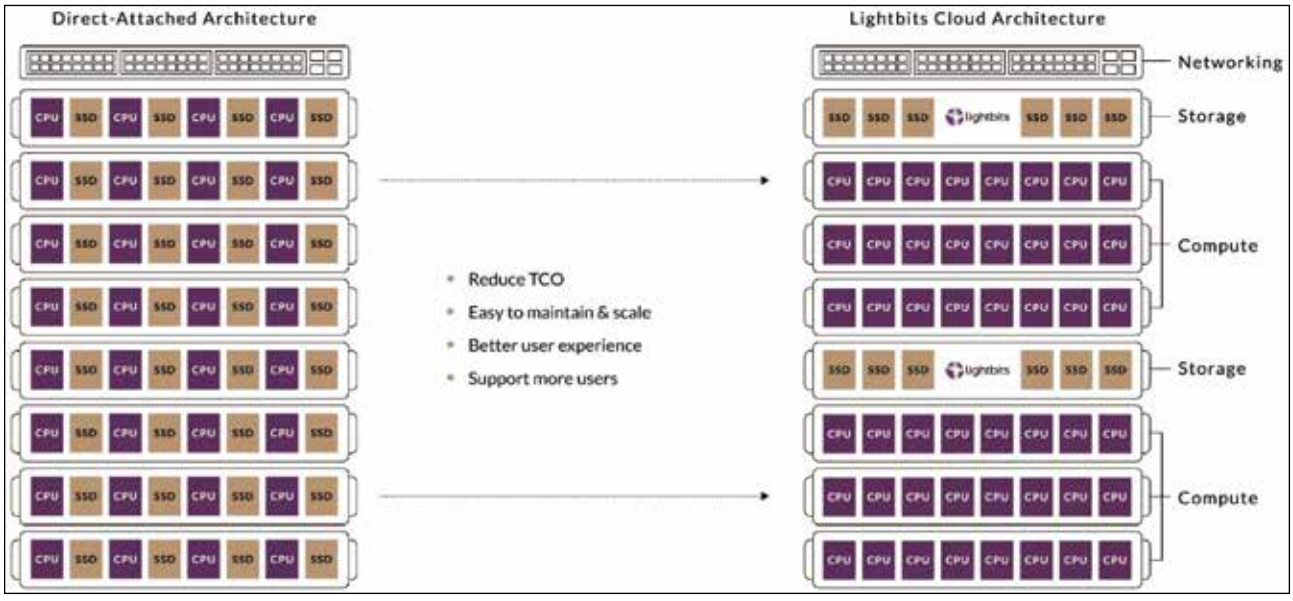
在 DAS 環境中，驅動器是在部署到伺服器中之前購買的或與伺服器一起購買的，隨著時間的推移，它們的容量利用率增長得很緩慢。另外，為了避免儲存用盡這種尷尬的局面出現，DAS 常常會有意將容量配置為過剩的狀態。

相比之下，將儲存從計算伺服器中分離出來的資料中心會更加高效。這樣，儲存容量就可以獨立地進行擴展，並且可以根據需要分配給計算伺服器。

隨著每 GB 快閃記憶體成本的降低，解耦合儲存方法更加經濟高效，而且資料中心部署的前期成本也要低得多。通過動態分配存儲資源，可以避免過度配置 (over-provisioning) 開銷，從而大大降低總體成本。

NVMe/TCP 解決方案釋放了基於解耦合高性能固態硬碟 (SSD) 的雲基礎設施的潛力。它使資料中心能夠從低效的直連式 SSD 模式轉為一種共用模式，在該模式中，計算和儲存可以獨立擴展，以最

圖 2: 從直連式儲存 (DAS) 轉為解耦合儲存和計算



大限度地提高資源利用率和運行靈活性。

這種新的共用模式採用了創新的 NVMe/TCP 標準。Lightbits Labs 發明了這一概念，並且正在引領這一新標準的發展。

NVMe/TCP 不會影響應用的性能。實際上，它通常會改善應用的尾部延遲，從而提升用戶體驗，並使雲服務提供者能夠在相同的基礎設施上支持更多用戶。它也不需要對資料中心網路基礎設施或應用軟體進行任何更改。它還可以降低資料中心的總體擁有成本 (TCO)，並使維護和擴展超大規模資料中心變得更容易。Lightbits Labs 正與其他市場領導者合作，以實現該標準在行業中的廣泛採用。

NVMe/TCP 利用標準的乙太網網路拓撲，獨立地擴展計算和儲存，以實現最高的資源利用率，並降低 TCO。

Lightbits Labs：在資料中心部署 NVMe/TCP

Lightbits Labs 的解決方案提供了如下性能優勢：

■與直連式儲存 (DAS) 相比，尾部延遲減少多達 50%

■SSD 容量利用率翻倍

■資料服務的性能提升 2-4 倍

■可擴展至數萬個節點

■可支援實現數百萬 IOPS 的性能，平均延遲低於 200 µs

Lightbits 解決方案在不影響系統穩定性或安全性的情況下可實現如下改進：

■應用伺服器及其存儲的物理分離

-- 支持獨立部署、擴展和升級

-- 支援儲存基礎設施比計算基礎設施更快地擴展

-- 提高應用伺服器和儲存的效率

-- 通過對應用伺服器和存儲硬體進行獨立的生命週期管理，可簡化管理並降低 TCO

■提供與內部 NVMe SSD 相當的高性能和低延遲

■可利用現有的網路基礎設施，無需進行更改

■可在多跳 (multi-hop) 資料中心網路架構中實現解耦合

Lightbits 存儲解決方案的工作原理

Lightbits Labs 為雲和資料中心基礎設施提供瞭解耦合快閃記憶體平臺。

當數萬或數十萬計算節點將直連式儲存的多個孤島鎖定在每個物理節點中時，雲級網路就會暴露出其所存在的極端複雜性。

Lightbits 的解決方案釋放瞭解耦合高性能 SSD 解決方案的潛力。它使資料中心能夠從低效的直連式 SSD 模式轉為一種共用模式，在該模式中，計算

圖 3:NVMe/TCP 可以跨資料中心將儲存節點連接至應用伺服器

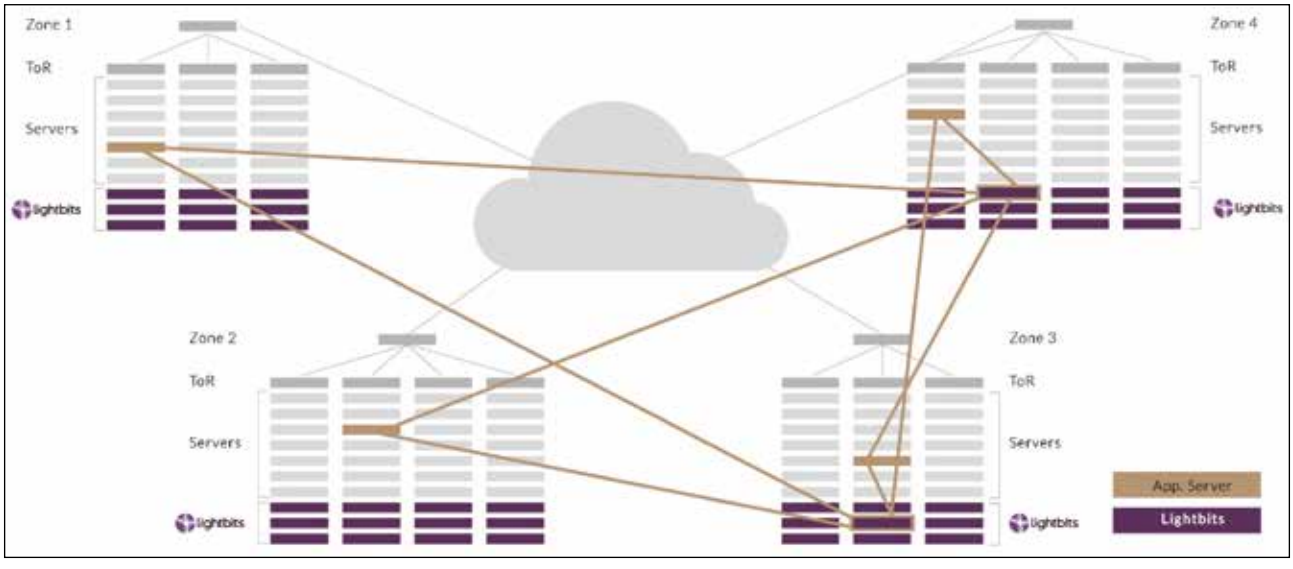
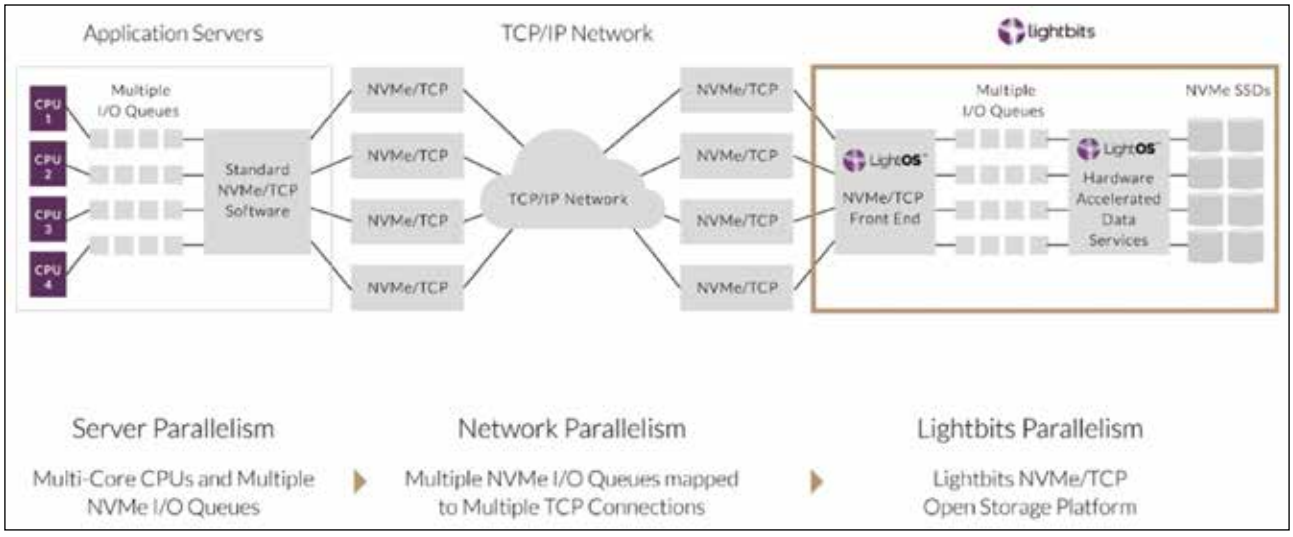


圖 4: 針對並行雲架構打造的 NVMe/TCP



和儲存可以獨立擴展，以最大限度地提高資源利用率和靈活性。

在 Lightbits Labs 發明 NVMe/TCP 時，我們繼續使用 DAS 設備所用的 NVMe 模式，然後將其映射到行業標準的 TCP/IP 協定組中。NVMe/TCP 是將多個並行的 NVMe I/O 佇列映射到多個並行的 TCP/IP 連接上。這種 NVMe 和 TCP 之間的配對可以實現一種簡單的、基於標準的、端到端的並行架構。

這種新的共用模式使用了創新的 NVMe/TCP 標準，該標準不會影響延遲，也不需要更改網路基礎設施或應用伺服器軟體。Lightbits Labs 正在與其他市場領導者合作，以推動這一新的 NVMe/TCP 標準的採用。

利用 Lightbits Labs 的解耦合儲存解決方案，儲存可以精簡的方式配置給應用伺服器。精簡配置意味著管理員可以將任意大小的卷分配給用戶端。而且，只有當應用伺服器寫資料時，才會消耗底層儲存容量。因此，儲存在最後一刻（即需要它的時候）才會被使用。這將延遲對更多儲存資源的購買，從而進一步降低成本。Lightbits 還為以線速運行的資料服務提供了一種硬體加速解決方案。

因此，當使用 Lightbits 精簡配置技術和面向資料服務的硬體加速方案時，存儲成本可以降低到只有性能相當的 DAS 解決方案成本的一小部分。

適合快閃記憶體的寫演算法

對於讀和寫操作而言，快閃記憶體介質的延遲都很低。但是，SSD 上的快閃記憶體控制器必須持續執行“垃圾收集”(GC) 操作，以便為即將到來的寫操作提供可用空間。與硬碟驅動器的寫操作可以覆蓋現有資料不同，快閃記憶體驅動器只允許將資料寫入以前未寫入或已擦除的快閃記憶體塊中。

垃圾收集操作會導致“寫入放大”。顧名思義，SSD 控制器執行垃圾收集時，應用伺服器發出的單個寫操作會被進行垃圾收集的 SSD 控制器在實際的快閃記憶體介質上放大為更多的寫操作。寫入放大會增加快閃記憶體驅動器的耗損，這將影響它的長期使用。

此外，後臺的垃圾收集會導致即將到來的 I/O 的延遲增加，並且隨著更多隨機寫操作被寫入快閃記憶體驅動器，垃圾收集會顯著增加。不幸的是，很大比例的 I/O 都是隨機的。總的來說，這意味著使用者無法獲得最好的性能或快閃記憶體耐久性。

Lightbits Labs 的解決方案通過一個智慧的管理層來解決這一問題，該層以不同的服務品質 (QoS) 等級來管理 SSD 池。這種解決方案減少了 SSD 後臺操作，並使 I/O 更快速、更高效。

LightOS 架構將多種演算法緊密結合在一起，以便優化性能和快閃記憶體利用率。這包括將資料

保護演算法與用於資料服務的硬體加速解決方案以及我們的高性能讀寫演算法緊密結合在一起。最終，跨 SSD 儲存池管理和平衡所有 I/O，從而極大地提高快閃記憶體利用率。

這種設計提高了總體性能，減少了尾部延遲、寫入放大和 SSD 上的耗損。這意味著 LightOS 可以為你的快閃記憶體儲存提供最高的投資回報率 (ROI)。

高性能資料保護方案

要想實現儲存與應用伺服器的分離，還需要智慧、高效且不影響性能的資料保護功能。

Lightbits 結合了高性能資料保護方案，其可與用於資料服務的硬體加速解決方案和讀寫演算法一起工作。

就如何將資料寫入 SSD 儲存池而言，相比傳統的 RAID 演算法，Lightbits 的資料保護方法可以防止過多的寫入， 以避免 SSD 遭受更多耗損。

總結

Lightbits Labs 實現了高效的快閃記憶體解耦合方案，在實施和運行方面具有以下優點：

- 不需要任何昂貴的私人網路絡硬體，Lightbits 解決方案運行在標準的 TCP/IP 網路上。
- 使用 TCP/IP 以機架規模在一個或多個局域網上運行，沒有協定方面的限制。
- 提供與 DAS 相當的性能和延遲，包括尾部延遲比 DAS 尾部延遲低 50%。
- 將高性能資料保護方案與其用於資料服務的硬體加速解決方案，以及可確保性能不受影響的讀寫演算法結合在一起。
- 通過用於資料服務的硬體加速解決方案最大限度地提高快閃記憶體效率，該解決方案以全線速運行，且不影響性能。
- 實現了精簡配置的儲存卷 (storage pool)，支援“按需付費”的消費模式。

Lightbits 是 NVMe/TCP 的發明者，也是其廣泛採用的推動者。

作為一種新理念的應用，Lightbits 的 NVMe/TCP 解決方案可以實現高效的快閃記憶體解耦合，從而獲得與 DAS 相當甚至更好的性能。Lightbits 創造了一種現代的 IP 儲存架構實現方式，可以最大限度地發揮應用伺服器、NVMe、TCP 和 SSD 並行架構的潛力。通過 Lightbits Labs 的解決方案，雲原生應用可以實現雲級性能，雲資料中心可以降低其雲級 TCO。

COMPUTEX 2021 BC Award 得獎名單公布 展現新世代數位轉型解決方案

科技大展 COMPUTEX TAIPEI 共同主辦單位台北市電腦公會 (TCA) 表示，為協助廠商展現創新研發能量，增加科技新品曝光，作為海內外買主採購指標的 COMPUTEX 官方獎項 Best Choice Award (又稱 BC Award)，邁入第 19 屆，近日以線上首播方式正式對外公布 19 件得獎產品，得獎廠商包括 AMD、華擎科技、宏正、圓剛科技、群邁通訊、元太科技、宜鼎國際、凱捷國際、佐臻、神達數位、微星、NVIDIA、普萊德、瑞昱、Silicon Labs、Supermicro、三緯國際等海內外科技大廠。

得獎產品涵蓋電競、AI、IoT、HPC、影音串流、車用科技、3D 列印、XR 等數位轉型科技應用解決方案，而年度大獎 (Best Choice of the Year) 則由微星 MEG Aegis Ti5 獲得！

TCA 指出，綜合評審團相關專家意見，會發現不論是在消費市場或是垂直市場，科技產品必須要能夠明確點出如何提升企業運作或個人工作效率，並解決使用者面對不同應用所產生之痛點，才能獲得評審團的專家認可，進而獲得 BC Award 獎項肯定。