Why NVMe/TCP is the better choice for your Data Center

Non-Volatile Memory express (NVMe)® has transformed the storage industry since its emergence as the state-of-the-art protocol for high-performance solid-state drives (SSDs).

Initially designed for high-performance direct-attached PCIe SSDs, NVMe was later expanded with NVMe over Fabrics (NVMe-oF) to support a rack-scale remote pool of SSDs. The industry has widely accepted that this new NVMe-oF model will replace the Internet Small Computer Systems Interface (iSCSI) protocol as the communication standard between compute servers and storage servers and become the default protocol for disaggregated storage.

Yet, the initial deployment options with NVMe-oF were limited to Fibre Channel and Remote Direct Memory Access (RDMA) fabrics.

What if we could offer a new, more powerful technology that provides the speed and performance of NVMe-oF without the prohibitive deployment costs and complexity?

NVMe over TCP (NVMe/TCP) extends NVMe across the entire data center using simple and efficient TCP/IP fabric.

This paper describes how NVMe/TCP is a better technology for existing data centers and the benefits it offers. These advantages include:

- Enabling disaggregation across a data center’s availability zones and regions
- Leveraging ubiquitous TCP transport with low latency, highly parallel NVMe stack
- No changes required on the application server side
- A high-performance NVMe-oF solution providing comparable performance and latency to Direct Attached SSDs (DAS)
- An efficient and streamlined block storage network software stack optimized for NVMe
- Providing parallel access to storage optimized for today’s multi-core application/client servers
- Standard NVMe-oF control path operations

1. Overview of NVMe/TCP

The NVMe specification has emerged as the state-of-the-art protocol for high-performance SSDs.

Unlike SCSI, iSCSI, SAS or SATA interfaces, NVMe implements a streamlined command model and multi-queue architecture that is optimized for many-core server CPUs. The NVMe-oF specification extended NVMe to share PCIe SSDs over the network, with the initial implementations using an RDMA fabric.

Today, Lightbits Labs is collaborating with Facebook, Intel, and other industry leaders to extend the NVMe-oF standard to support the TCP/IP transport that is complementary to RDMA fabrics.

Disaggregation with NVMe/TCP has the distinct advantage of being simple and highly efficient. TCP is ubiquitous, scalable, reliable, and ideal for short-lived connections and container-based applications.

Additionally, migrating to shared flash storage with NVMe/TCP does not require changes to the data center network infrastructure. No infrastructure changes means easy deployment across the data center because nearly all data center networks are designed to carry TCP/IP.

Broad industry collaboration on the NVMe/TCP protocol means the protocol is designed from the ground up with a wide ecosystem and support for any operating system and Network Interface Card (NIC) in mind. The NVMe/TCP Linux drivers are a natural match for the Linux kernel and use the standard Linux networking stack and NICs without any modifications.

The result is a promising new protocol, tailor made for hyperscale data centers, that is easy to deploy with no changes to the underlying network infrastructure.

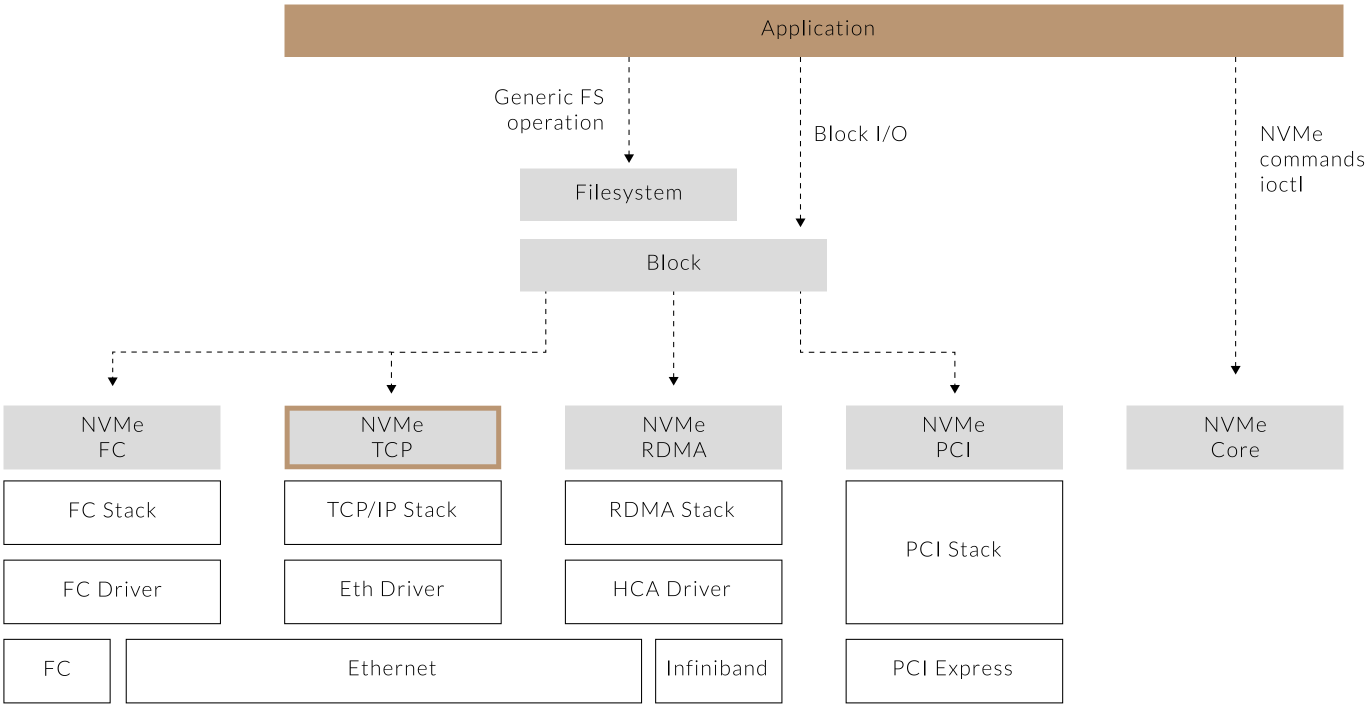

Figure 1: NVMe/TCP seamlessly integrated with existing NVMe protocols in the Linux kernel

2. How Data Centers Today Approach Storage

2.1. Direct-attached storage architecture and NVMe

The NVMe storage protocol was designed to extract the full performance out of solid state drives (SSDs).

The parallelism designed into the NVMe protocol helps achieve this performance. NVMe eliminates the single-queue iSCSI model. Instead, NVMe enables up to 64,000 queues between the CPU subsystem and storage.

SSDs are parallel devices that use many parallel communications channels to connect with many SSD memory locations, which means an SSD can efficiently receive data in large parallel streams. Before the NVMe/TCP protocol existed, the simplest way to take advantage of this parallelism was to install NVMe SSDs directly into the application server. In other words, you had to build your storage infrastructure using the DAS model.

With the DAS approach, an application benefits from:

- Multiple CPUs
- Multiple NVMe I/O queues
- Parallel SSD architecture

The challenge to the industry is to move the SSDs from the individual servers where you can have stranded capacity to a shared storage solution with improved utilization of your infrastructure — without losing the DAS performance gains. Therefore, the goal of all NVMe disaggregation technologies is to achieve DAS performance in a shared NVMe solution.

2.2. Previous generation IP-based storage architecture

Previously, the iSCSI standard was the only option for connecting to block storage over a TCP/IP network. It was developed at the turn of the century when most processors were single core devices.

In SCSI, there is a single connection between the application (initiator) and the storage (target). With iSCSI, there is also a single TCP socket connecting the client to the block storage server.

Today, data center processors are massively parallel multi-threaded devices. This complexity in today's processors demanded an overhaul for the available storage protocols. The result was the emergence of NVMe as a replacement for SATA and SAS (Serial Attached SCSI).

In all of these earlier protocols, their development was based around a serialized, spinning disk drive.

Non-volatile memory (NVM) is a parallel storage technology that has no need for a platter, or multiple platters, to spin underneath a magnetic head, or set of magnetic heads. With NVM storage devices, many memory locations are accessible in parallel and at lower latencies.

Undoubtedly, iSCSI is still relevant for use-cases with low to moderate storage performance demands. However, iSCSI does not meet the demand of I/O intensive applications that need low latency at large scale.

NOTE: To learn about head-to-head performance measurements between iSCSI and NVMe/TCP, send an email to info@lightbitslabs.com requesting a copy of Lightbits' whitepaper comparing iSCSI vs. NVMe/TCP.

2.3. Alternatives to NVMe/TCP disaggregation

RDMA and also Remote Direct Memory Access over Converged Ethernet (RoCE), as well as NVMe over Fibre Channel (NVMe over FC), are other networked storage protocols that attempt to solve the disaggregation problem. However, these alternatives require the installation of costly special hardware such as an RDMA capable NIC at both ends (Application Server and Storage Server). Also, with RDMA hardware installed, there is the complexity of configuring and managing flow-control within your RDMA capable switch fabric.

RDMA does offer performance that is suitable for some high-performance computing environments, but it demands increased cost and requires significant complexity for its deployment.

TCP/IP has proven to work reliably and efficiently in hyper-scale environments. NVMe/TCP inherits this reliability and efficiency, and it can co-exist as a complementary solution with RDMA or replace it completely.

3. Flash Disaggregation in the Data Center and the NVMe/TCP Solution

In DAS environments, drives are purchased before being deployed in servers or together with the servers, and their capacity utilization grows slowly over time. Additionally, to avoid the logistical headaches of running out of storage, DAS configurations are often intentionally over-provisioned.

In contrast, a data center that disaggregates storage from compute servers is more efficient. The storage capacity can be scaled independently and can be assigned to compute servers on an as-needed basis.

As the cost per GB of flash storage decreases, the disaggregated storage approach is more economical, and data center deployments' up-front costs are much lower. The over-provisioning overhead is avoided by dynamically assigning storage resources, dramatically reducing overall cost.

The NVMe/TCP solution unlocks the potential of a disaggregated high-performance solid-state drive (SSD) cloud infrastructure. It enables data centers to move from the inefficient direct-attached SSD model to a shared model in which compute and storage are scaled independently to maximize resource utilization and operational flexibility.

This new shared model utilizes the innovative NVMe/TCP standard. Lightbits invented the concept and is leading the development of this new standard.

NVMe/TCP does not compromise application performance. In fact, it often improves application tail latency thereby improving user experience and enabling the cloud provider to support more users on the same infrastructure. It does not require any changes to data center networking infrastructure or to application software. It reduces data center total cost of ownership (TCO) and makes it easier to maintain and scale hyperscale data centers. Lightbits Labs is working with other market leaders toward the industry wide adoption of this standard.

NVMe/TCP utilizes standard Ethernet network topologies and scales compute and storage independently to achieve maximum resource utilization and drive down TCO.

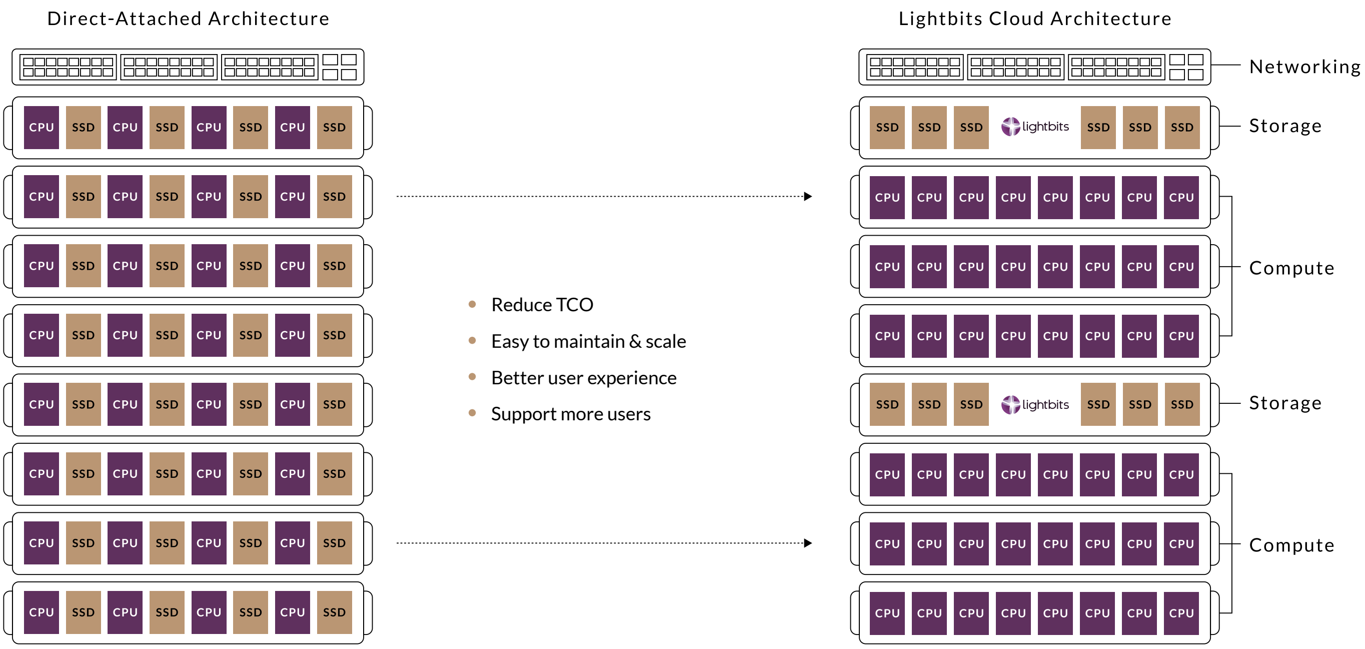

Figure 2: Moving from Direct Attached Storage (DAS) to disaggregated storage and compute

4. Lightbits Labs: Deploying NVMe/TCP in a Data Center

The Lightbits Labs™ solution offers estimated savings and performance advantages that include:

- Reduced tail latency by up to 50% versus Direct Attached Storage (DAS)
- Doubled SSD capacity utilization
- Data services with 2-4x better performance
- Scaling up to tens of thousands of nodes
- Support for millions of IOPS with average latency that is below 200μs

Lightbits solution achieves these improvements without compromising system stability or security.

- Physical separation of the application servers and their storage
  - Enables independent deployment, scalability, and upgrades
  - Enables storage infrastructure to scale faster than compute infrastructure
  - Increases efficiency of the application servers and storage
  - Simplifies management and drives down TCO through the Independent life-cycle management of application servers and storage hardware
- Delivers high performance and low latency that is comparable to internal NVMe SSDs
- Utilizes existing networking infrastructure, no changes required
- Enables disaggregation in multi-hop data center network architectures

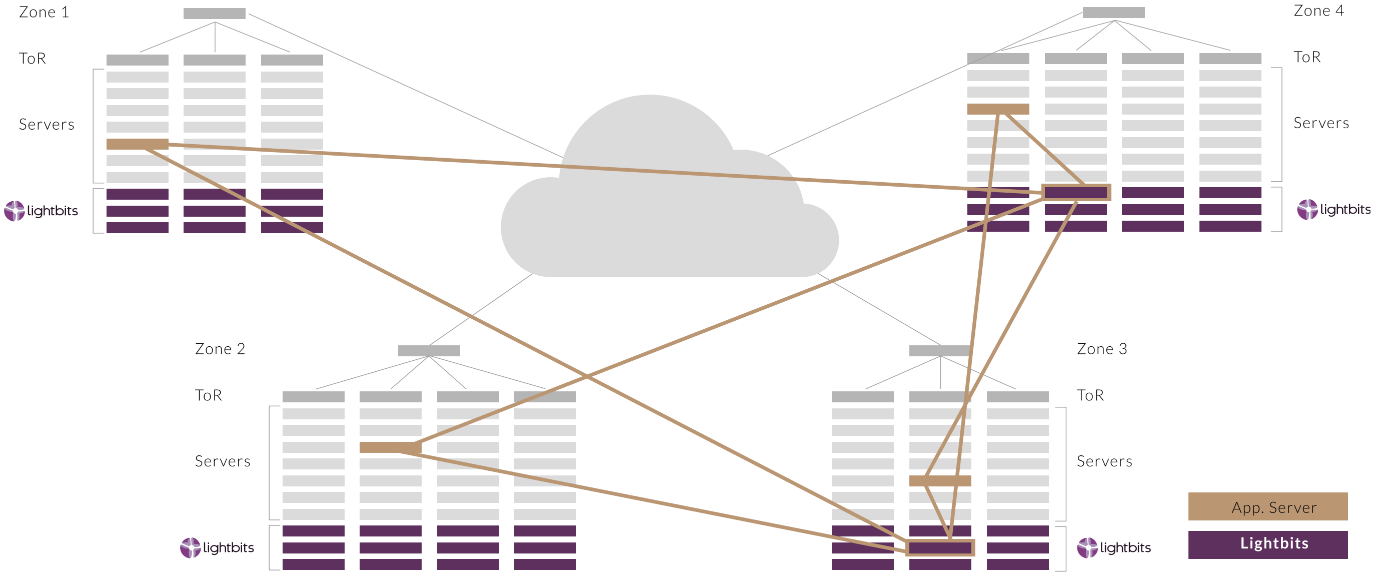

Figure 3: NVMe/TCP connecting storage nodes to application servers across the data center

5. How Lightbits' Storage Solution Works

Lightbits Labs offers a disaggregated flash platform for cloud and data center infrastructures.

Cloud-scale networking has exposed the extreme complexity that exists when tens or hundreds of thousands of compute nodes have isolated islands of direct-attached storage locked into each physical node.

The Lightbits' solution unlocks the potential of a disaggregated high-performance SSD solution. It enables data centers to move from the inefficient direct-attached SSD model to a shared model in which compute and storage are scaled independently to maximize resource utilization and flexibility.

When Lightbits Labs invented NVMe/TCP, we continued with the NVMe model that was used in DAS devices and then mapped it to the industry-standard TCP/IP protocol suite. NVMe/TCP maps the many parallel NVMe I/O queues to many parallel TCP/IP connections. The pairing between NVMe and TCP has resulted in a simple, standards-based, parallel architecture from end-to-end.

Lightbits Labs NVMe/TCP Utilizing The Parallelism of Modern Cloud Architecture

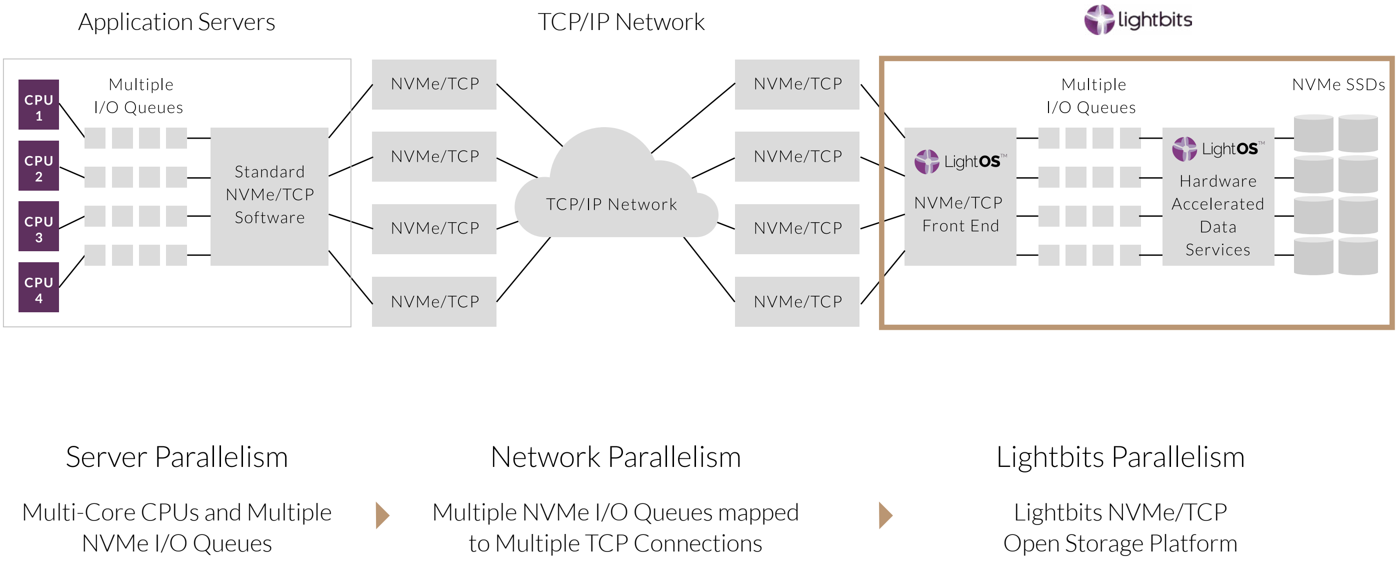

Figure 4: NVMe/TCP built for parallel cloud architectures

This new shared model utilizes the innovative NVMe/TCP standard that does not compromise latency, or require changes to networking infrastructure or application server software. Lightlabs is working with other market leaders toward the adoption of this new NVMe/TCP standard.

With the Lightbits Lab disaggregated storage solution, storage is thinly provisioned to application servers. Thin provisioning means that an administrator can assign a volume of any size to a client. And, only when the application server writes data is the underlying storage capacity consumed. Therefore, the storage gets used at the last possible moment—when it is needed. This further drives down costs by delaying the purchase of storage resources. Lightbits also provides a hardware-accelerated solution for data services at wire speed.

Therefore, storage costs can be cut to a fraction of comparable DAS solutions when using Lightbits' thin provisioning technology and hardware-acceleration for data services.

5.1. Flash-Friendly write algorithms

Flash media has very low latencies for both reads and writes. However, the flash controllers on SSDs must continuously perform "garbage collection" operations to provide free space for incoming writes. Unlike hard drives in which writes can overwrite existing data, flash drives only allow writing data to flash blocks that were previously unwritten or erased.

Garbage collection operations cause "write amplification". As the name implies, a single write issued by the application server can be amplified into many more writes onto the actual flash media by the SSD controller doing garbage collection. Write amplification increases wear on the flash drive, which reduces its long-term usage.

Additionally, background garbage collection causes increased latency for incoming I/Os, and garbage collection dramatically increases as more random writes are written to flash drives. Unfortunately, a high percentage of I/Os are random. Overall, this means users are not getting the best possible performance or flash endurance.

The Lightbits Lab solution addresses this issue through an intelligent management layer that manages pools of SSDs at different Quality of Service (QoS) levels. This solution reduces SSD background operations and makes I/Os faster and more efficient.

The LightOS^{TM} architecture tightly couples many algorithms together in order to optimize performance and flash utilization. This includes tightly coupling the data protection algorithms with a hardware-accelerated solution for data services and with our high performance read and write algorithms. Finally all IO is managed and balanced across pools of SSDs resulting in a massive improvement of flash utilization.

This design increases overall performance and reduces tail latencies, write amplification and reduces wear on the SSDs. This means LightOS delivers the maximum Return On Investment (ROI) for your flash storage.

5.2. High-performance data protection schemes

Disaggregation of the storage from application servers demands intelligent and efficient data protection that does not affect performance.

Lightbits incorporates high-performance data protection schemes that work together with the hardware-accelerated solution for data services and read and write algorithms.

Concerning how data is written to a pool of SSDs, the Lightbits data protection methods prevent excessive writes that expose the SSDs to greater wear, versus traditional RAID algorithms.

6. Summary

Lightbits Labs enables efficient flash disaggregation with the following benefits in implementation and operation:

- No need for any costly, specialized networking hardware. The Lightbits solution runs on standard TCP/IP networks.
- Uses TCP/IP to operate at rack scale over a LAN, or over multiple LANs, with no protocol restrictions.
- Providing performance and latency that are comparable to that of DAS, including tail latency that can be up to 50% better than DAS tail latency.
- Integrates high-performance data protection schemes with its hardware-accelerated solution for data services, along with read and write algorithms that ensure no performance compromise.
- Maximizes flash efficiency using a hardware-accelerated solution for data services that operates at full-wire speed with no performance impact.
- Implements thinly provisioned storage volumes that enables a “pay as you grow” consumption model.

Lightbits is the inventor of NVMe/TCP and the driving force behind its adoption.

With this new concept for efficient flash disaggregation, the same, or better than, DAS performance can now be realized with the Lightbits NVMe/TCP solution. Lightbits has created a modern IP storage architecture implementation that utilizes Application Server, NVMe, TCP, and SSD parallel architectures to their maximum potential.

With Lightbits Labs, cloud-native applications can achieve cloud-scale performance and cloud data centers can reduce their cloud-scale TCO.

Contact us for more information: **info@lightbitslabs.com**

About Lightbits Labs™

Today's storage approaches were designed for enterprises and don't meet developing cloud-scale infrastructure requirements. For instance, SAN is known for lacking performance and control. At scale, Direct-Attached SSDs (DAS) become too complicated for smooth operations, too costly, and suffer from inefficient SSD utilization.

Cloud-scale infrastructures require disaggregation of storage and compute, as evidenced by the top cloud giants' transition from inefficient Direct-Attached SSD architecture to low-latency shared NVMe flash architecture.

Unlike other NVMe-oF approaches, the Lightbits NVMe/TCP cost-saving solution separates storage and compute without touching network infrastructure or data center clients.

The Lightbits' team members were a key contributor to the NVMe standard and among the originators of NVMe over Fabrics (NVMe-oF). Now, Lightbits is crafting the new NVMe/TCP standard.

As the trailblazers in this field, the Lightbits' solution is already successfully tested in industry-leading cloud data centers.

The company's shared NVMe architecture provides efficient and robust disaggregation. With a transition that is so smooth, your applications teams won't even notice the change. They can now go wild with better tail latency than local SSDs!

Finally, you can separate storage from compute without drama.

www.lightbitslabs.com   info@lightbitslabs.com

US Office:

1830 The Alameda,
San Jose, CA 95126, USA

Israel (Kfar Saba) Office:

17 Atir Yeda Street,
Kfar Saba, Israel 4464313

Israel (Haifa) Office:

3 Habankim Street,
Haifa, Israel 3326115

The information in this document and any document referenced herein is provided for informational purposes only, is provided as is and with all faults and cannot be understood as substituting for customized service and information that might be developed by lightbits labs ltd for a particular user based upon that user's particular environment. Reliance upon this document and any document referenced herein is at the user's own risk.

The software is provided "As is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and non-infringement. In no event shall the contributors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings with the software.

Unauthorized copying or distributing of included software files, via any medium is strictly prohibited.