Lab Insight

Run Apps up to 16X Faster: Storage Performance Comparison Lightbits vs. Ceph Storage

Authors:
Russ Fellows
Mohammad J Rabin

March 2022

Overview

Currently, two trends driving significant change in the IT landscape are cloud computing and container-based applications. Both are gaining traction due to their ability to run and operate applications with greater flexibility and often, lower costs than alternative technologies and methods. However, with all new opportunities there are significant challenges to realizing the potential benefits.

As new technologies continually emerge, they typically have an evolutionary effect on business operations. However, cloud computing as an operational model along with Kubernetes for managing cloud native applications are already transforming the way information technology is delivered, managed and consumed.

Moreover, the need for systems designed to support these environments requires the choice of compute, networking and particularly storage systems that can deliver highly scalable capacity and performance without the need for traditional administration. Kubernetes provides mechanisms for applications to transparently utilize multiple compute and storage resources. However, ensuring that applications retain access to persistent data demands that the underlying storage infrastructure is resilient while also meeting the capacity and performance needs for thousands of applications.

With this backdrop, Evaluator Group was asked to compare two storage systems suited for cloud native application environments.

- Ceph: Is open-source software defined storage often used in cloud native environments
- Lightbits Cloud Data Platform: Is a software defined storage designed for cloud environments

As shown below, Lightbits massively outperformed Ceph for all workloads tested, using identical hardware and storage media. The summary in Table 1 shows results using NVMe SSDs with QLC media. Testing was performed using workloads running as containers in an OpenShift Kubernetes environment.

	Ceph IOPs	Lightbits IOPs	Lightbits Advantage
4K-100% Read	1,032,428	4,068,462	3.94 X
4K-100% Write	30,728	515,697	16.78 X
8K-80% Read	90,363	1,129,335	12.50 X
16K-70% Read	35,740	372,009	10.41 X
32K-50% Read	19,797	111,852	5.65 X

Table 1: Lightbits vs. Ceph I/O rate Comparison with QLC media in Container Environment. (Source: Evaluator Group)

The remainder of this paper explores the technology and performance differences in more detail along with performance data when using NVMe SSDs with TLC media.

Emerging Technologies

As outlined previously, several trends are driving significant changes within the IT landscape. The move towards utilizing services including public clouds is changing where IT operations occur and how they are paid for, as well as having an impact on how applications are operated. The adoption of Cloud Native Applications (CNAs) is changing how applications are developed and managed and is driven in part by the migration to the use of services and clouds. Additionally, new technologies that enable access to NVMe storage via a fabric, along with new non-volatile memory technology provide significant performance benefits for systems that are able to leverage these capabilities.

Cloud Native Applications

Container based applications are designed to operate in cloud environments, including both on-premises and public clouds. Cloud native applications are designed as services that may be independently scaled, which are known as micro-services. By leveraging a micro-service architecture, CNAs are able to scale portions of applications as necessary, with little or no administrative input.

Additionally, CNAs may be deployed in one location, then moved to another based upon available resources or even based upon the price of the resources. Moreover, it is imperative that the persistent storage required by many of these applications also responds dynamically to changing requirements for capacity as well as performance.

NVMe over Fabrics

Non-Volatile Memory Express (NVMe), is an interface that enables solid-state drives (SSDs) to communicate using the high-speed PCIe bus, which provides low-latency access to solid-state devices. NVMe supports billions of commands with significantly less delay or latency compared to SCSI or SATA devices. In order to accommodate communications to external devices, NVMe over Fabrics extends the NVMe protocol to support accessing NVMe over a network fabric.

One of the newest options for NVMe over Fabric access is using the TCP transport protocol, known as NVMe/TCP which imposes far fewer technical burdens to implement, supporting commodity Ethernet without any specific hardware or switch settings. Overall, NVMe over TCP provides significantly lower latency and higher throughput than other storage protocols over TCP while still using existing networking infrastructure.

Persistent Memory

Persistent memory (PMem) combines storage features, specifically the ability to retain or “persist” data without requiring energy, together with memory bus access features such as very low latency and byte addressability. The two leading types of persistent memory are NVDIMM and Intel Optane Persistent

Memory. The primary difference between NVDIMM and Optane PMem is their storage density and cost, with Intel's Optane PMem having significant advantages vs. NVDIMM devices for both characteristics.

Intel PMem may be used either in "App Direct Mode" requiring application modifications, or "Memory Mode" which transparently expands system memory capacity at a lower cost than DRAM. Lightbits uses PMem in App Direct mode as a non-volatile write buffer, with writes and meta-data access directed to PMem, reducing wear on backing media and improving the system write performance.

Solution Overview

Storage running as software, or Software Defined Storage (SDS), using commodity hardware has been gaining traction over the past several decades. One appeal of this approach is the ability to operate resilient data center class storage in environments where dedicated, or proprietary hardware is not practical. Additionally, due to its flexibility, it is possible to use the same SDS stack across different environments and deployment sizes. As discussed, our evaluation used two different SDS options to provide persistent storage for container applications. An overview of their features is examined subsequently. The Lightbits solution leveraging Intel technologies is shown in Figure 1.

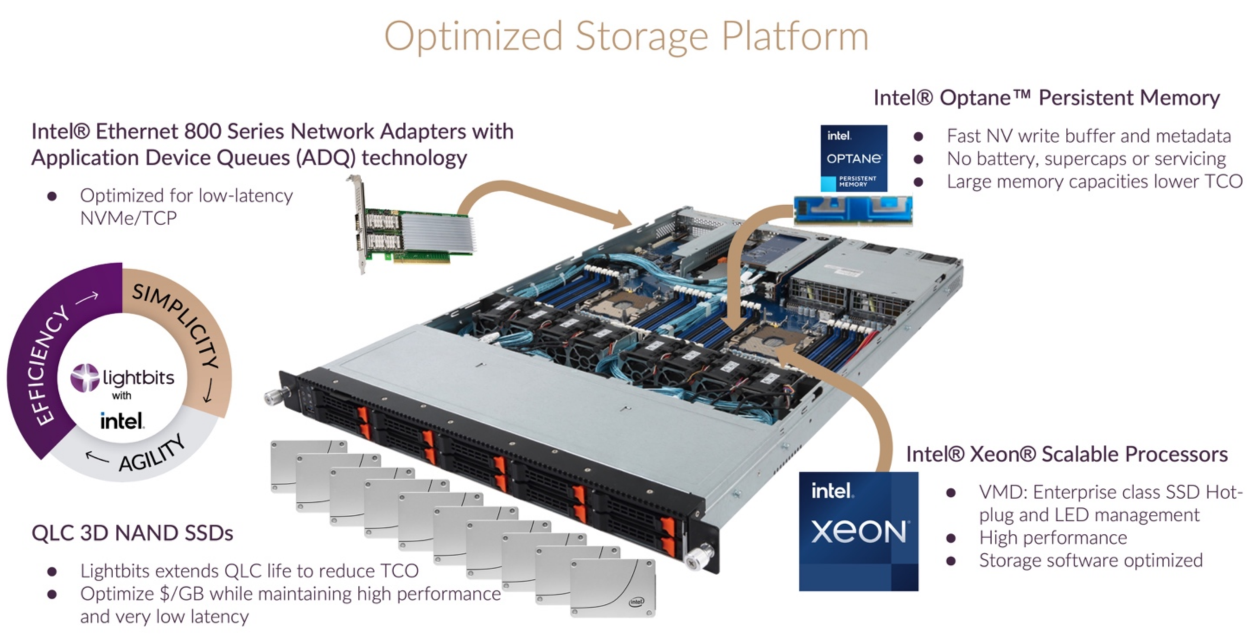

Figure 1: Intel Platform for Lightbits Software Defined Storage (Source: Intel)

Intel

Intel provides foundational platform elements for many of today's leading server and storage solutions. Components include Intel Xeon processors, Intel network adapters and Intel Optane™ technology delivered both as NVMe solid-state drives and as a memory form-factor Optane Persistent Memory (Pmem) devices. Additionally, Intel Ethernet 800 Series with Application Device Queues (ADQ) provide high performance, low-latency NVMe/TCP network interface cards.

Lightbits Cloud Data Platform

Lightbits Cloud Data Platform is software defined storage designed to utilize recent technologies, including NAND flash and Intel Optane persistent memory (PMem), and also supports NVMe over TCP using high speed 100GbE network interface cards. Lightbits uses a disaggregated architecture, enabling it to scale CPU, memory, PMem or NVMe devices independently as needed. This allows Lightbits to deliver scalable, enterprise class storage with performance that surpasses local NVMe devices.

Lightbits Intelligent Flash Management™(IFM) is a set of features which maximize the performance and extend the endurance of SSD's. Key features of IFM are write striping, IOP-less metadata access, Smart Garbage Collection, Append Write Strategy, and Parallel Read/Write Pipelines.

Coupling Intel's high-performance hardware platform with Lightbits delivers a scalable solution that reduces overall TCO. The Lightbits solution specifically leverages the following technologies:

- **Intel® Xeon® Processor Systems**
  - High performance CPU, memory, I/O and PCIe Gen4
  - Efficient storage-software using high-performance CPU architecture
  - Enterprise class SSD Hot-plug and LED support for NVMe media
- **Intel Optane Persistent Memory (PMem)**
  - Low-latency, non-volatile write buffer and metadata store
  - No batteries or capacitors required to maintain persistence
  - Large persistent memory capacities, lower TCO than alternatives
- **Intel Ethernet 800 Series Network Adapters**
  - Provide application device queues (ADQ)
  - High performance application device queues
  - Low latency NVMe/TCP without need for custom drivers

Access to Lightbits storage for Kubernetes managed containers as Persistent Volumes (PVs) is provided through the Kubernetes Container Storage Interface (CSI). The Lightbits CSI Plugin enables Kubernetes to store persistent volumes (PVs) in the Lightbits cluster using their "Elastic RAID" which provides erasure coded protection from data loss on a per node basis. Additional Lightbits services include compression, replication, volume snapshots, clones, and Role-Based Access Control (RBAC) for multi-tenant environments.

Ceph

Ceph is an open source, software defined storage offering designed with several distinct interfaces that enable access as block device, filesystem and object storage. When Ceph was designed, spinning hard-disk drives were the predominant storage media with NAND based SSDs playing no role in most main-stream storage systems.

During the ensuing 15-year period since Ceph's design, solid-state flash in various form factors has emerged as the predominant media for high-performance storage systems. Many elements of Ceph's original design remain, which were designed to improve the performance of spinning physical media. More recently, Ceph has been updated to utilize solid-state for meta-data operations with "BlueStore" and "RocksDB". To date Ceph has not been optimized for Persistent Memory and does not yet support NVMe over fabrics or NVMe over TCP.

Our testing used an open-source release of Ceph as detailed in the Appendix.

Storage Efficiency

The resiliency design of Lightbits provides additional benefits compared to some software defined storage systems. Lightbits utilizes erasure coding on each node to provide data availability with the loss of a storage device; together with replication between nodes, Lightbits is able to provide higher resiliency with less overhead and wasted storage capacity. During testing, Lightbits was configured to use erasure coding for data protection on each node along with making one additional copy of data on another node within the Lightbits cluster.

In contrast, Ceph was configured to create three data copies for data protection, which does not ensure data availability if a drive in one node along with an additional node experiences an outage. Although the storage efficiency of these two protection methods is similar, the Lightbits approach provides higher availability.

Performance of Container Workloads

Evaluator Group was asked to analyze two different storage offerings running as containers in a Kubernetes applications environment. We compared Lightbits software defined storage to Ceph, an open-source software defined storage offering for containers. The test environment was comprised of 12 Kubernetes nodes running container workloads against either a dedicated 3-node Lightbits cluster, or a 3-node Ceph cluster. A diagram along with additional details on the container application tests, the hardware, software and other aspects are documented in the Appendix.

Storage performance was measured using the well-known 'vdbench' tool to create workloads, using 8 container instances running vdbench per node, for a total of 96 instances of vdbench.

Testing consisted of 5 different access patterns and block sizes often found in performance sensitive applications:

- 4KB, 100% read, with 100% random access
- 4KB, 100% write, with 100% random access
- 8KB, 80% read / 20% write with 80% random access
- 16KB, 70% read / 30% write with 80% random access
- 32KB, 50% read / 50% write with 80% random access

These workloads were used to compare the performance of Lightbits with Ceph using QLC solid-state media as the primary storage media together with higher speed persistent media where appropriate for each storage system. In all cases, each of the 5 workloads ran multiple times to provide an average result used for comparison. For specific details regarding the configuration of both Lightbits and Ceph, please refer to the Appendix.

Evaluator Group Comments: Test results show massive performance advantages when using Lightbits, with up to 16 times the performance of Ceph for 4KB write workloads. This level of advantage is highly significant, particularly considering the number and type underlying media is identical in all comparisons.

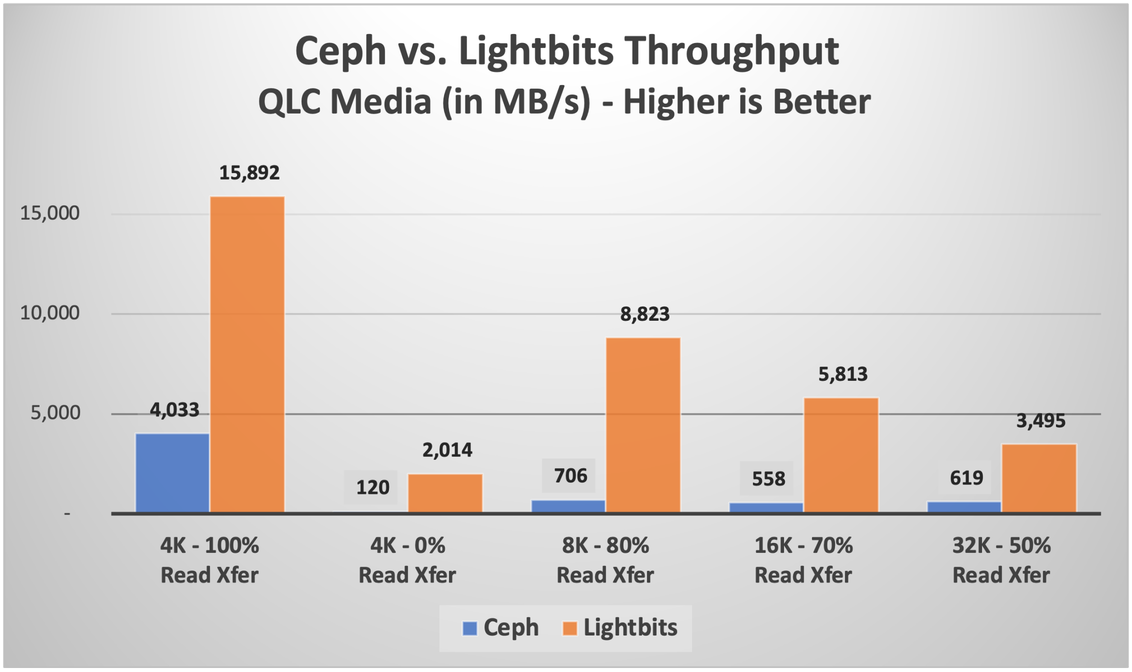

Figure 2: Ceph vs. Lightbits Throughput Comparison (source: Evaluator Group)

Figure 2 shows the throughput results for 5 different access patterns for both Lightbits and Ceph while using QLC media. The throughput results are directly comparable, as they account for the differences in block sizes and thus are shown all together.

Performance Details

The I/O rate (measured as I/O's per second, or IOPs) is a more common metric for small block workloads but makes comparisons between different block sizes difficult. Shown below in Figure 3 is a comparison of more typical application sizes of 8, 16 and 32 KB charting the I/O rate, measured in I/O's per second. At 32KB, Lightbits had a 5.7x advantage, at 16KB it was 10.4x higher and for 8K workloads Lightbits outperformed Ceph over 12 times.

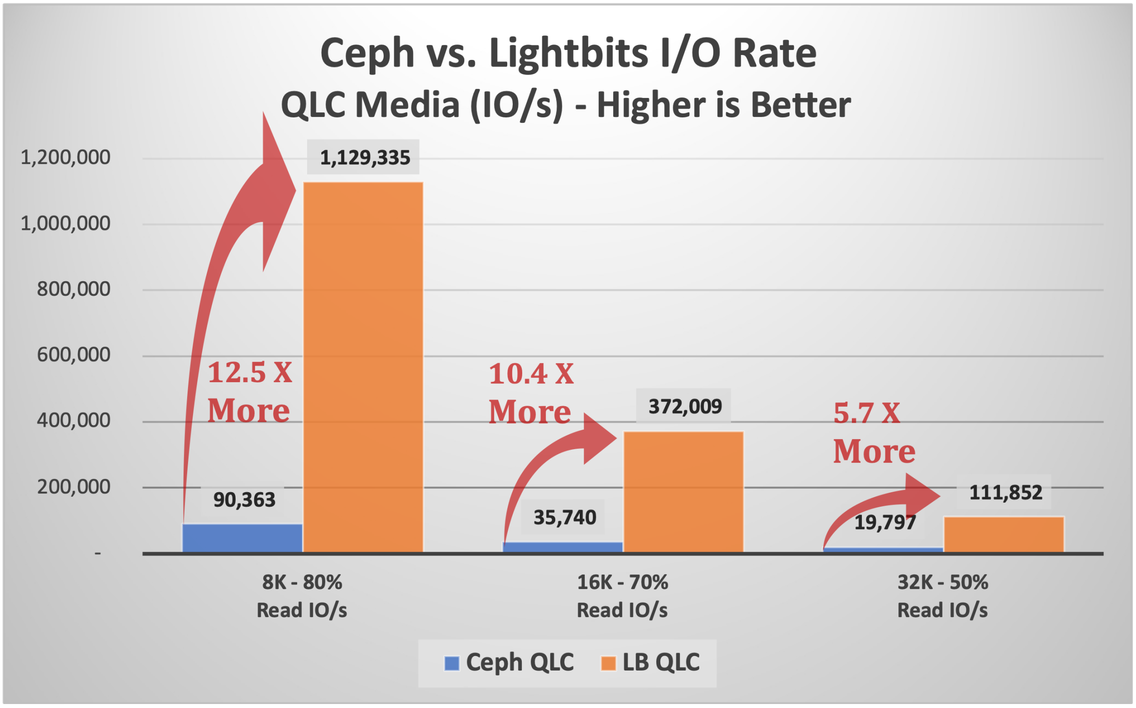

Figure 3: Lightbits vs. Ceph w/ QLC Performance – Throughput for Different Block Sizes (source: Evaluator Group)

The three workloads shown above are very common with databases or other transactional applications. Although many databases attempt to perform 16K or 32K I/O sizes, in many cases they will use smaller sizes if the changes are small or if there is a high transaction rate. Rather than waiting to coalesce several small I/O's, the database will perform 8K or even 4K transactions.

Shown below are I/O rate comparisons for 4KB read operations for both Ceph and Lightbits, using two different types of storage media, QLC and TLC media.

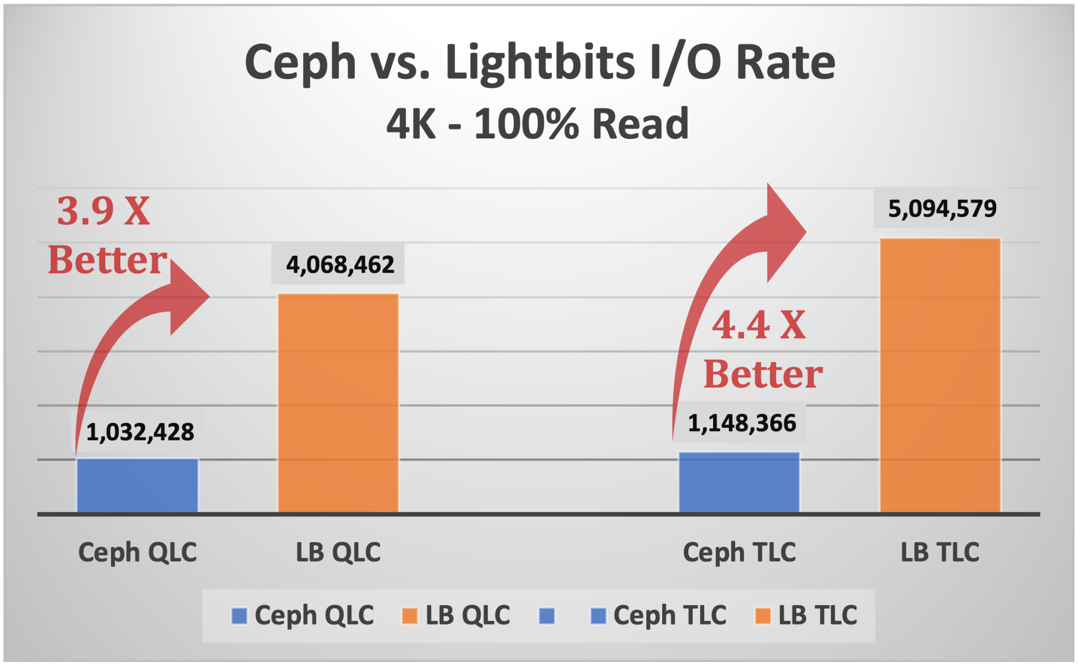

Figure 4: Ceph vs. Light bits 4KB Read I/O Rate Comparison for QLC and TLC media (source: Evaluator Group)

While 4KB read operations can occur, 4KB write operations are quite often used for database logging operations, with Lightbits providing nearly 17X better performance than Ceph while using QLC media, and nearly 7X better with TLC media as shown below in Figure 5.

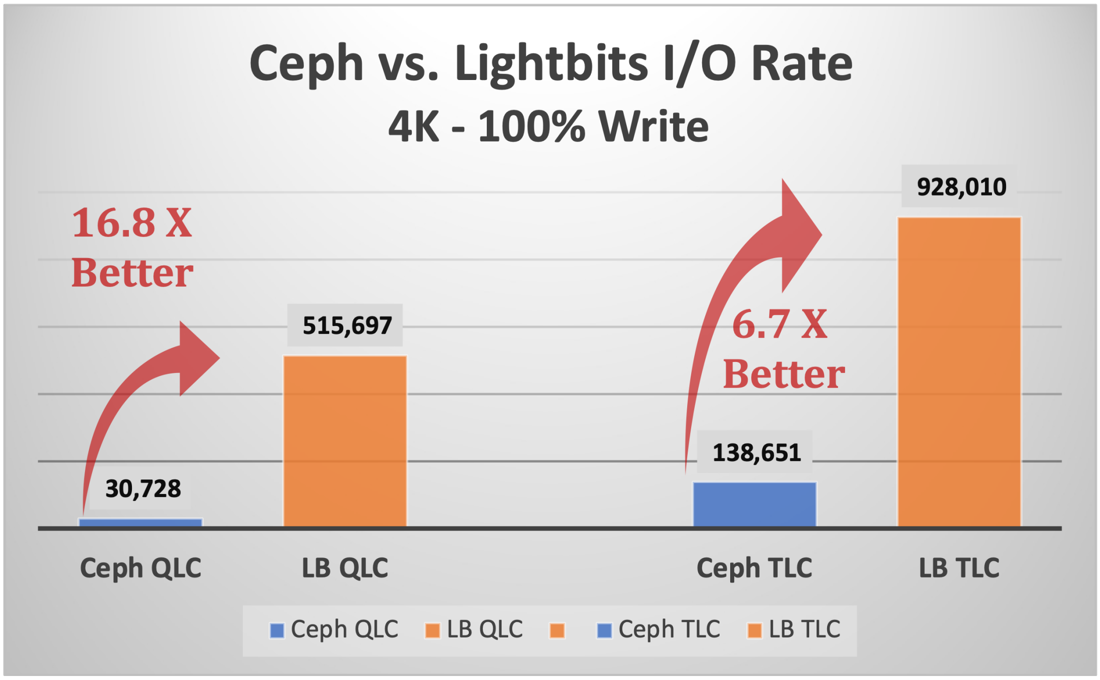

Figure 5: Ceph vs. Light bits 4KB Write I/O Rate Comparison for QLC and TLC media (source: Evaluator Group)

Evaluator Group Comments: Databases and other transactional applications rely upon low latency storage, particularly for transaction logs which must be written to persistent media before transactions complete. In many cases, logs are written with 4K I/O operations, a workload where Lightbits showed the most benefits over Ceph, a nearly 7X advantage with TLC media and almost 17X better with QLC.

Additional Performance Data

An overview of the performance results is shown below with three primary metrics indicated, including throughput, I/O rate along with the latency. The performance advantage for Lightbits over Ceph is so significant that an "x" factor is used to denote the advantage. This uses simple division (i.e. 515,697 / 30,728 = 16.78 x) to show the number of times better Lightbits performed. There are several takeaways:

- The results over all five workloads tested indicates between a 4x to 16x advantage for Lightbits
- Databases often write logs or journals with small blocks that are highly latency sensitive
- The performance advantage for Lightbits of 16.78x for 4KB writes is a massive difference

	QLC Media Comparison			TLC Media Comparison		
	Ceph QLC	Lightbits QLC	LB- X better	Ceph TLC	Lightbits TLC	LB- X better
4K-100% Read MB/s	4,033	15,892	3.94	4,486	19,901	4.44
4K-0% Read MB/s	120	2,014	16.78	542	3,625	6.69
8K-80% Read MB/s	706	8,823	12.50	3,542	16,219	4.58
16K-70% Read MB/s	558	5,813	10.41	4,959	15,449	3.12
32K-50% Read MB/s	619	3,495	5.65	4,802	9,837	2.05
4K-100% Read IO/s	1,032,428	4,068,462	3.94	1,148,366	5,094,579	4.44
4K-0% Read IO/s	30,728	515,697	16.78	138,651	928,010	6.69
8K-80% Read IO/s	90,363	1,129,335	12.50	453,419	2,076,046	4.58
16K-70% Read IO/s	35,740	372,009	10.41	317,367	988,724	3.12
32K-50% Read IO/s	19,797	111,852	5.65	153,677	314,774	2.05
4K-100% Read Lat. (ms)	1.5	0.4	3.95	1.3	0.3	4.46
4K-0% Read Lat. (ms)	19.8	1.1	17.27	4.1	0.6	6.72
8K-80% Read Lat. (ms)	22.6	1.4	16.34	3.4	0.7	4.59
16K-70% Read Lat. (ms)	22.3	2.2	10.27	2.4	0.8	3.12
32K-50% Read Lat. (ms)	11.5	1.9	6.19	1.2	0.6	2.06

Table 3: Lightbits vs. Ceph Performance Comparison (Source: Evaluator Group)

An additional note on testing is that the latency for Ceph was higher than acceptable for most workloads while using QLC media. This is due to using the same queue depth settings for all tests in order to provide a fair comparison. If the workload was optimized for Ceph QLC, a lower queue depth setting would be used, which would reduce the I/O rate, while also lowering the latency. If testing was done in this manner, Ceph's latency would be more acceptable, but its performance would be significantly lower, giving Lightbits an even greater advantage.

Final Thoughts

Performance is always an important consideration for any aspect of IT infrastructure, particularly when operating cloud environments or hosting cloud native applications. Although not every application or micro-service demands high I/O rates, providing resilient, scalable storage with high I/O rates and low latency is critical to running modern cloud infrastructure.

> ***Evaluator Group Comments: Using the same hardware configuration Lightbits outperformed Ceph by a significant margin while offering higher resiliency. Lightbits is optimized for Intel's high-performance technologies to deliver a powerful storage platform for cloud enabled Kubernetes environments.***

Several technological advances have occurred recently that enable IT consumers the ability to operate their infrastructure at scale, cost effectively. New storage media such as QLC provides an opportunity to support massive storage capacity for less demanding workloads with good performance when combined with Intel Optane Persistent Memory. Other technologies including new processors and end to end NVMe can also provide performance and cost benefits.

Intel specific technologies including Xeon CPUs, Intel network cards with NVMe over TCP together with Optane Persistent memory and high-density NVMe SSDs provide a strong hardware foundation for software defined offerings. Lightbits storage delivers local NVMe levels of performance due to the efficient use of Intel CPUs, network interfaces and persistent storage media, providing consistently low latency and data services.

The performance testing shown in this paper used tools and workloads that are typical for IT environments and applications. Further details are provided in the Appendix for those who desire a closer examination or wish to replicate these tests using their own hardware configurations.

> ***Evaluator Group Comments: Using test tools and workloads that are well understood by storage professionals, we found Lightbits delivered significantly better performance for modern applications. Lightbits can support latency sensitive applications using QLC media together with Optane PMem, while Ceph struggled to provide a fraction of the I/O's at significantly higher latencies.***

By leveraging new technologies such as Persistent Memory and end to end NVMe, Lightbits provides significantly better performance than Ceph while also having higher resiliency. Clearly, Lightbits is a better option for modern applications and cloud environments, providing up to 16 times the performance of Ceph, with lower latency.

Appendix

Test Environment Details

The test environment utilized the following hardware software and application workload items.

Testing was performed over a 4-month period, from September to December 2021.

Hardware and Infrastructure

- See Figure 6 below for an overview of the test configuration
    - An application cluster of 12- Kubernetes "worker nodes" running on bare-metal
    - A 3-node cluster of management / "master" Kubernetes nodes running as VMs
    - A storage cluster of 3 nodes running a software defined storage stack
- OpenShift version 4.6 was the application environment and orchestration

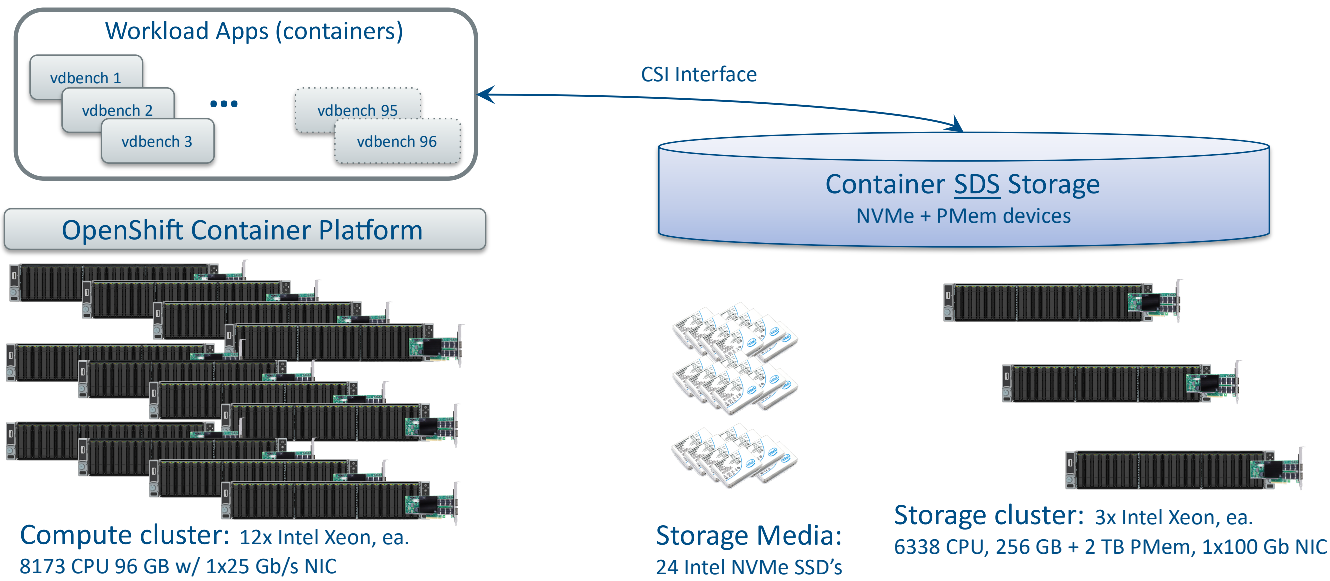

Figure 6: Test Setup for Workloads on OpenShift Accessing Storage Nodes (source: Evaluator Group)

Software Environment

- Red Hat OpenShift version 4.6 was used for the 3 management and 12 application nodes
    - The 12 application (OpenShift "worker") nodes all ran directly on hardware, so called "Bare Metal"
    - The 3 management (OpenShift "master") nodes all ran as VM's, but were installed as "Bare Metal"
- CentOS version 8.4 was the base OS used the 3 Ceph Nodes
- Open source Ceph version 16.2.6 aka "Pacific" was used for the Ceph SDS stack

- VMware 7.0; ESXi along with vCenter was used for the 3 VM's

SUT Configuration

- One, 12 node application cluster running Red Hat OpenShift
    - Application / worker node configuration:
        - 2 socket 1st Gen Intel Xeon Scalable systems (Intel Xeon Platinum 8173M)
        - 96 GB of Memory (12 x 8 GB DIMM)
        - Intel XXV710-DA2 NIC, w/ 2x 25 Gb/s connections per host
- Two storage systems were connected to the OpenShift application cluster
    - SDS #1, Ceph configuration:
        - 2 socket 3rd Gen Intel Xeon Scalable systems (Intel Xeon Gold 6338)
        - 256 GB of Memory (16 x 16 GB DIMM)
        - 2 TB of Intel Optane PMem (configured as WAL and Cache device)
        - Intel E810-CQDA2 NIC, w/ 1x 100 Gb/s connections per host
        - 8 x Intel NVMe storage media, with either
            - QLC SSD media: Intel SSD D5-P5316 @ 15.36 TB
            - TLC SSD media: Intel SSD D7-P5510 @ 3.84 TB
    - SDS #2, Lightbits configuration:
        - 2 socket 3rd Gen Intel Xeon Scalable systems (Intel Xeon Gold 6338)
        - 256b GB of Memory (16 x 16 GB DIMM)
        - 2 TB Intel Optane Persistent Memory 200 (configured as Lightbits write buffer)
        - Intel E810-CQDA2 NIC, w/ 1x 100 Gb/s connections per host
        - 8 x Intel NVMe storage devices, with either:
            - QLC SSD media: Intel SSD D5-P5316 @ 15.36 TB
            - TLC SSD media: Intel SSD D7-P5510 @ 3.84 TB

Client Configuration – Infrastructure

- A single Intel system running VMware ESXi was used for supporting infrastructure
    - The 3 VMs used as OpenShift "master" nodes
    - A "jump" host running Microsoft Server 2019 used for accessing infrastructure remotely

SDS Storage Configurations

Ceph Configuration

Note: *There is no "default" configuration for Ceph, every install may be different due to the availability of hardware, and other configuration parameters. See below for configuration:*

- Using "CentOS v 8.4" as the host OS for all three nodes
- Configure hardware as defined above, with 8 NVMe devices per node
- Download then configure Ceph "Pacific" on all 3 nodes
    - 2 TB of PMem was split into 2 partitions, one for WAL and the other for RocksDB

  - Each NVMe SSD was configured as 6 logical OSDs (setting of 6 OSDs per device)
- Ceph Device configuration
  - There were 8 SSD's per host, each partitioned into 6 regions
  - By using 6 partitions on 8 physical SSDs resulted in a total of 6 * 8 = 48 OSDs per system
  - Ceph was configured with the WAL residing on Optane PMem
- Default resource settings:
  - CPU was set to 1 CPU per OSD, Memory was set to 4 GB per device

Lightbits Configuration

***Note**: There is no "default" configuration for Lightbits. Lightbits and Intel engineering performed all setup operations for Lightbits. See below for configuration:*

- Using "Lightbits" as the host OS for all three nodes
- Configure hardware as defined above, with 8 NVMe devices per node
- Configure Lightbits on all 3 nodes
  - 2 TB of PMem assigned as cache device
  - Each NVMe SSD was used as a logical device

Application Workloads

- The "vdbench" tool was used to generate synthetic workloads with different block sizes and read to write ratios
  - Prior to other tests, a "write-fill" workload was run to completely write the entire storage capacity tested
  - Each test was run for a 15 minute "warmup" period followed by a 30 minute "measurement" period
  - The tests were run sequentially, with the completion of all 5 tests called a "set"
  - Each test set was then repeated 9 times, with the average of the 9 runs used for comparison
- The following 5 workloads were tested against storage configurations
  - 4KB, 100% read, with 100% random access
  - 4KB, 100% write, with 100% random access
  - 8KB, 80% read / 20% write with 100% random access
  - 16KB, 70% read / 30% write with 80% random access
  - 32KB, 50% read / 50% write with 50% random access

Overview of Test Process

- Setup servers and networking used for applications
- Setup systems and install VMware used for additional test infrastructure
- Install OpenShift on 15 systems (12 "worker" bare-metal hardware, and 3 "master" VM nodes)
- Install and configure software defined storage target clusters
  - Install Lightbits onto 3 nodes uses for Lightbits SDS storage

   - Install CentOS, then install Ceph Pacific on 3 nodes used for Ceph SDS storage
- Create container workload environment of 96 container instances of vdbench clients with 1 vdbench controller instance
- Run vdbench workload on controller container, which distributed workload to all 96 vdbench instances
- Gather the 9 result sets for each workload to find average I/O rate, throughput and latency

Comparison Using TLC Media

After completing testing using QLC media, both storage systems were reconfigured to used TLC media devices rather than QLC devices. TLC media is known to provide significantly better write I/O rates along with lower latencies. Once again, both storage systems were tested using the same workloads, running on the same server hardware, network and most importantly, an identical number and type of storage media. TLC based NVMe solid-state devices served as the primary storage media, with Intel Optane PMem used according to best practices for each storage stack.

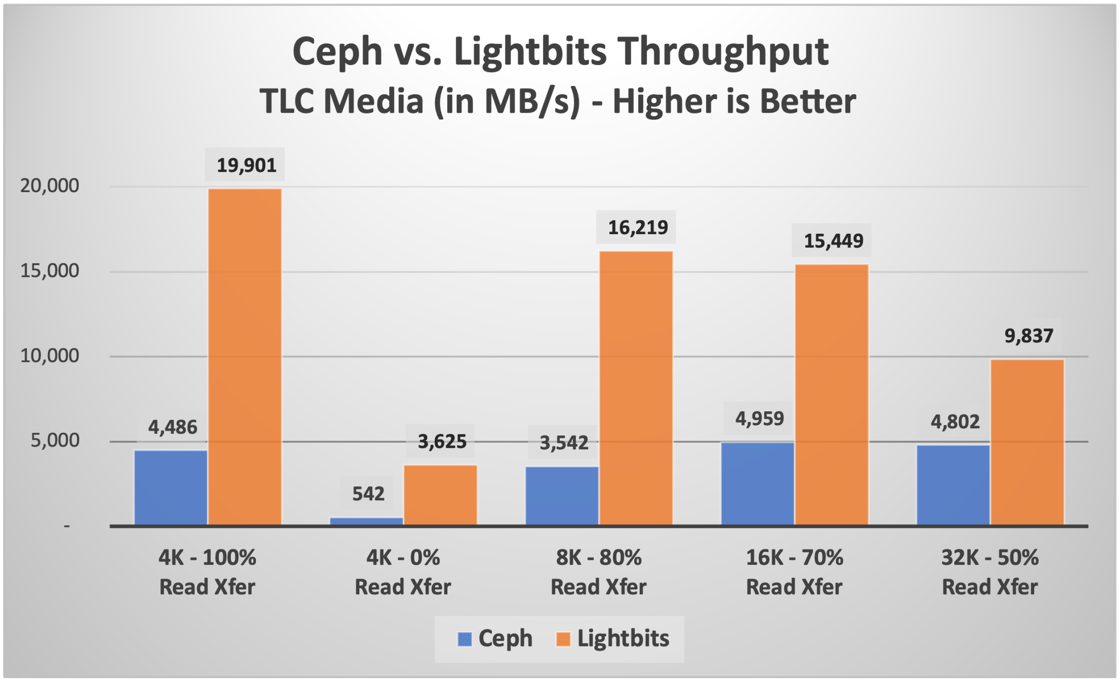

Figure 7: Lightbits vs. Ceph Throughput Comparison with TLC Multiple Workloads (source: Evaluator Group)

Figure 7 shows the throughput results for the same five tests while using TLC SSD media. Throughput results are shown in order to provide directly comparable results, even though I/O rate is a more common metric for showing individual small block workloads.

About Evaluator Group

*Evaluator Group Inc. is dedicated to helping **IT professionals** and vendors create and implement strategies that make the most value of their storage and digital information. Evaluator Group services deliver **in-depth, unbiased analysis** on storage architectures, infrastructures, and management for IT professionals. Since 1997 Evaluator Group has provided services for thousands of end-users and vendor professionals through product and market evaluations, competitive analysis, and **education**. **www.evaluatorgroup.com** Follow us on Twitter @evaluator_group*

This document was developed with Intel funding. Although the document may utilize publicly available material from various vendors, including Intel and others, it does not necessarily reflect such vendors' positions on the issues addressed in this document.