redhat | SAMSUNG

WHITEPAPER

HIGH-PERFORMANCE WORKLOADS WITH SOFTWARE-DEFINED STORAGE AND NVMe SSDs

INTRODUCTION

Over the past several years, cloud computing has proven to be a transformational technology, driving many improvements in how IT organizations deliver applications and services to their respective businesses. One major result of this shift to the cloud has been broader adoption of distributed computing approaches across an increasingly large portion of the enterprise portfolio.

Enterprise storage has followed a similar trend. Not long ago, storage was almost exclusively deployed via expensive, proprietary, and monolithic storage devices offered by a handful of large vendors. Fast forward several years and we see that a new choice has emerged. Today it is common for enterprises to deploy a variety of applications using software-defined storage, taking advantage of its reduced cost, greater agility, and improved scalability compared to traditional storage options.

To date, software-defined storage has distinguished itself in the domain of capacity-constrained applications, such as large-scale object and media serving, where the ability to easily scale a storage system has been paramount. As software-defined storage and underlying infrastructure technologies have advanced, opportunities to target a much broader set of use cases are emerging.

One particularly exciting development has come in pairing software-based storage with solid-state drives (SSDs) based on the NVM Express (NVMe) interface specification. Because NVMe SSDs significantly push the performance, latency, throughput, and value frontiers of flash-based storage, they help users extend software-defined storage deployments to support traditional performance-constrained workloads such as relational and NoSQL databases, virtualized environments, telco and financial services applications, and more. Organizations that take advantage of software-defined storage combined with NVMe SSDs for these applications have a new lever for driving growth, reducing cost, and increasing agility without sacrificing performance and reliability.

Red Hat, the provider of Red Hat® Ceph Storage, an industry-leading software-defined storage solution, and Samsung, the world leader in enterprise SSDs and semiconductor memory technology, have partnered to ensure that this combination is validated, supported, and readily accessible to customers.

SOFTWARE-DEFINED STORAGE AND PERFORMANCE-BOUND WORKLOADS

Over the past decade, software-defined storage has grown from a niche technology used by large internet companies to a broadly accepted storage approach for a growing number of capacity-focused applications.

Examples of the most commonly deployed software-defined storage use cases today include:

- **Rich media repositories.** Software-defined storage is increasingly used to support "active archive" applications because of its ability to support vast collections of video files, images, documents, or other objects for serving or streaming to web- and mobile-based clients.
- **OpenStack®.** Because it enables enterprises to deliver thinly provisioned block storage for virtual machine images in an OpenStack private cloud, as well as object and file storage services, Red Hat Ceph Storage has long been the de facto standard storage solution for OpenStack cloud environments.
- **Operational (log) data.** The proliferation of networked servers, devices, and sensors has made log file storage a growing challenge for enterprises. Software-defined storage is ideal for log storage because stores can start small and grow as logs do.

While these capacity-constrained workloads were a natural starting point for enterprises adopting software-defined storage—indeed, in many cases there were no economically viable alternatives available—the advantages of software-defined storage extend beyond them.

"Software-defined storage promises reduced costs, improved agility, and easier management relative to more hardware-defined legacy architectures that make it an excellent fit for '3rd Platform' computing environments," says Laura DuBois, vice president at IDC. "It is used extensively for workload consolidation, and we are seeing an increasing number of organizations using it with what they consider to be mission-critical workloads when they deem the deployment model to be a good fit with their applications."

High-performance workloads present a substantial opportunity for reducing storage cost and complexity in the enterprise. Perhaps the most common examples of a performance-constrained workloads are databases, such as the open source MySQL, MariaDB, PostgreSQL, and MongoDB. The typical enterprise has many instances of these databases, or their commercial cousins, and spends significant sums on their underlying storage and overall operations. Other examples of performance-bound workloads include virtual desktop infrastructure (VDI) and high-performance computing (HPC) applications.

In addition, in the OpenStack domain, a large pool of fast storage is a key enabler of a wide variety of emerging use cases such as big data analytics and data lakes, Database-as-a-Service (DBaaS), network functions virtualization (NFV), and a variety of financial services and telco applications such as ticker plants and other market data systems in the former and over-the-top (OTT) video delivery services like Network DVR and other edge services in the latter.

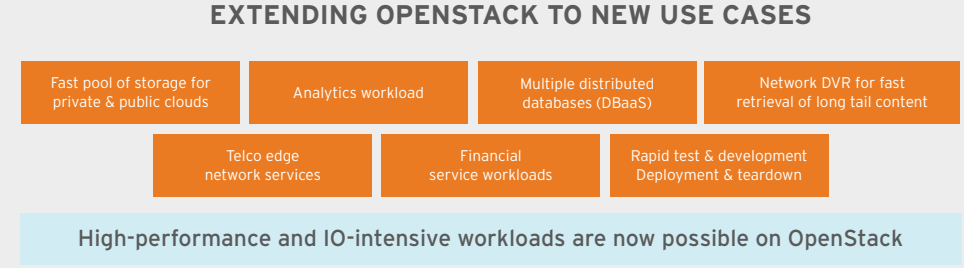

Figure 1. High-performance OpenStack storage enables new use cases

TECHNOLOGY

Requirements of performance-constrained workloads

Whereas scale is the primary consideration for many capacity-constrained applications currently deployed on software-defined storage, performance-bound workloads like databases have a different set of requirements. Typically, these systems are concerned primarily with optimizing performance metrics such as the number of IO operations per second (IOPS), transactions per second (TPS), or query response times.

When we also consider the fact that high IOPS and high-throughput workloads frequently support mission-critical business functions, the list of requirements for supporting them is significant. Any storage system, software-defined or otherwise, must deliver the following to support mission-critical IOPS- or throughput-optimized workloads:

- **Low latency**. The amount of time it takes to return the first bit from the storage device must be low.
- **High bandwidth**. The number of IOPS or the amount of data transferred in each unit of time (e.g., per second) must be high.
- **High availability.** The system must ensure the continuous availability of data.
- **Elasticity and scalability.** The system must allow workloads to start small and grow or shrink as needed.
- **Low cost.** To maximize the return on the enterprise's investment in these systems, they must minimize the cost per IO operation or per megabyte of data transferred.

COMMON WORKLOAD CHARACTERISTICS

IOPS optimized	•Lowest cost per IOP •Highest IOPS •Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster)	•MySQL/MariaDB-based apps on Openstack •Block storage
Throughput optimized	•Lowest per MBps* •Highest MBps •Highest MBps per BTU •Highest MBps per watt •Meets minimum fault domain recommendation (single server is less than or equal to 10% of the cluster) *MBps=Mbytes per sec	•Digital media server workloads •Block or object storage

RED HAT CEPH STORAGE AND SAMSUNG NVMe SSD INNOVATIONS ENABLE HIGH-PERFORMANCE WORKLOADS

While software-defined storage systems like Red Hat Ceph Storage have many inherent advantages over monolithic storage in creating high-performance storage systems—for example, the fact that Ceph "parallelizes" IO operations across multiple storage devices—there is still an underlying dependency on infrastructure and device performance. Low network and storage latency, as well as high network and storage bandwidth, are critical to supporting performance-optimized workloads in a software-defined storage environment.

In fact, support for high IOPS and high-throughput workloads has required innovation in both the storage device and software layers of the stack. Networking advances such as the growing ubiquity of 40-and 100-gigabit ethernet, while critical, are beyond the scope of this paper.

Together, the innovation offered by Samsung NVMe SSDs and Red Hat Ceph Storage allow enterprises to field software-defined storage environments supporting primary storage scenarios for performance-bound workloads such as relational databases.

SAMSUNG NVMe SSDs

Early solid-state disk drives preserved the serial AT attachment (SATA) interface that was already in use with magnetic hard disk drives (HDDs). Doing so helped spur adoption by ensuring compatibility with existing servers and was not a severe limitation because the immature flash drives available at the time were both smaller and slower. As drive capacities increased however, SSDs hit a wall with SATA, which limits performance to 6 Gbps (SATA 3.0).

To overcome the limitations inherent in SATA, an industry working group was formed to develop the modern interface standard that has come to be called NVM Express. NVMe uses the PCI Express (PCIe) bus to deliver much greater bandwidth and lower latency than SATA interfaces. This allows NVMe to support the IOPS densities (IOPS/GB) required for high-capacity SSDs, and in the context of software-defined storage is a key enabler of high-performance workloads.

Samsung is the world leader in enterprise SSDs, having introduced the world's first NVMe SSD in July 2013. Samsung now ships more Enterprise SSDs than anyone else, including highly efficient NVMe SSDs designed specifically for scale-out storage architectures, which offer 4x the performance of SATA SSDs at the same cost.

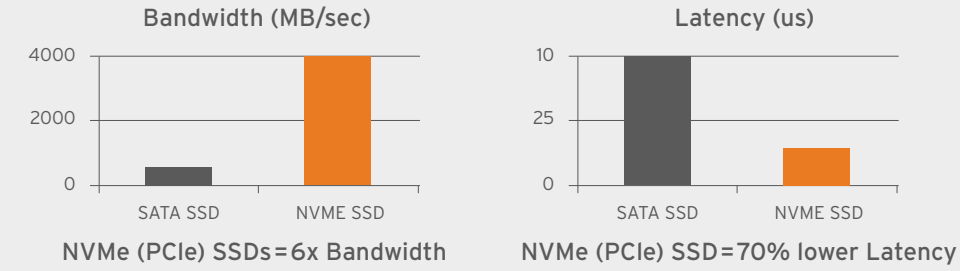

Figure 2. Samsung NVME SSD performance and low latency benefits

Samsung Enterprise NVMe SSDs offer a wide variety of technical and economic advantages over SATA SSDs and HDDs, including:

- **Low latency.** The NVMe protocol has a three times the latency advantage (3us vs. 10us) over SATA, which was designed with legacy HDDs in mind.
- **Greater throughput and IOPS.** NVMe SSDs can provide up to six times the bandwidth, more than four times the read performance, and up to eight times the IOPS as SATA-connected SSDs, and over 4,000 times the IOPS as magnetic drives.
- **Cost parity.** Today's NVMe SSDs provide all their performance advantages at the same cost per gigabyte as SATA SSDs.
- **Greater power efficiency.** NVMe enterprise storage drives are 77% more efficient than SATA SSDs on an IOPS per watt basis.
- **Dual-port connectivity.** Samsung Enterprise NVMe SSDs can offer high-availability dual-port connections, enabling traditional storage area network/network-attached storage (SAN/NAS) scale-up architectures to reduce latency far below what is possible with legacy SAS connectivity.
- **Greater reliability.** NVMe SSDs provide outstanding reliability thanks to the absence of moving parts in their solid-state storage design.

RED HAT CEPH STORAGE

Red Hat Ceph Storage is uniquely mature among software-defined storage offerings. In the decade since its introduction as an open source technology, Ceph has benefited from both continued feature development and significant business acceptance. These two advantages form a cycle that continues to accelerate Ceph's development and contribute to its ability to support mission-critical, performance-bound workloads.

Red Hat's backing, combined with the open source nature of the product, has led to the development of a robust ecosystem of hardware and software partnerships with technology leaders such as Samsung. These technical collaborations have led to the elimination of a wide variety of bottlenecks across the software and hardware stack. Ceph components that have evolved to benefit from fast access to flash-based storage via the NVMe protocol include:

- **RADOS Block Device (RBD).** The RBD provides block device interfaces to a Ceph storage cluster. Block storage is usually consumed by databases and other workloads that benefit from raw IOPS performance and low latency per transaction. In a recent version of Ceph, RBD memory allocation has been specifically tuned to improve random IO performance on RBD devices while reducing memory usage.
- **RADOS Object Gateway (RGW).** The RGW provides object interfaces to a Ceph storage cluster via Amazon S3 and Swift-compatible application programming interfaces (APIs). It also manages object metadata such as Access Control Lists (ACLs) and indexes, the maintenance of which requires frequent communication between the RGW and the object store. Fast NVMe SSDs increase the efficiency of this communication and improves RGW scalability.
- **RADOS.** The RADOS cluster, also called the Ceph Object Store, is the underlying object store in a Ceph deployment. Over the past several Ceph releases, new features and architectural improvements have led to a general increase in RADOS performance, specifically when used in conjunction with NVMe SSD media.

In addition to these enhancements to the traditional Ceph architecture, Ceph's BlueStore, a next-generation block-based storage subsystem currently in Technical Preview, promises to even further increase Ceph performance for block applications. By storing end-user data directly onto a BlueStore-managed block device and bypassing the filesystem, BlueStore achieves a 4-5x performance boost for typical performance-constrained workloads.

Due to its powerful controlled replication under scalable hashing (CRUSH) algorithm, Ceph is unique among software-defined storage technologies in its ability to direct individual workloads to the portion of a Ceph cluster with appropriately configured underlying hardware. This allows cost- and capacity-constrained workloads to utilize HDD-based servers, throughput-based workloads to be directed to a mix of NVMe SSD and HDD servers, and high-performance workloads to be pushed to NVMe-configured servers. The ability to direct workloads in this way will remain important for some applications, since standard magnetic drives retain a raw cost-per-GB advantage over flash, while flash drives have gained a total cost of ownership (TCO) advantage over a three-year term.

ADVANTAGES

RED HAT CEPH STORAGE AND SAMSUNG NVMe SSD SOLUTION

To validate the operating characteristics of performance-optimized, all-flash software-defined storage clusters and their suitability for high-performance workloads, Samsung and Red Hat have developed a reference architecture for Ceph running on Samsung NVMe SSDs.

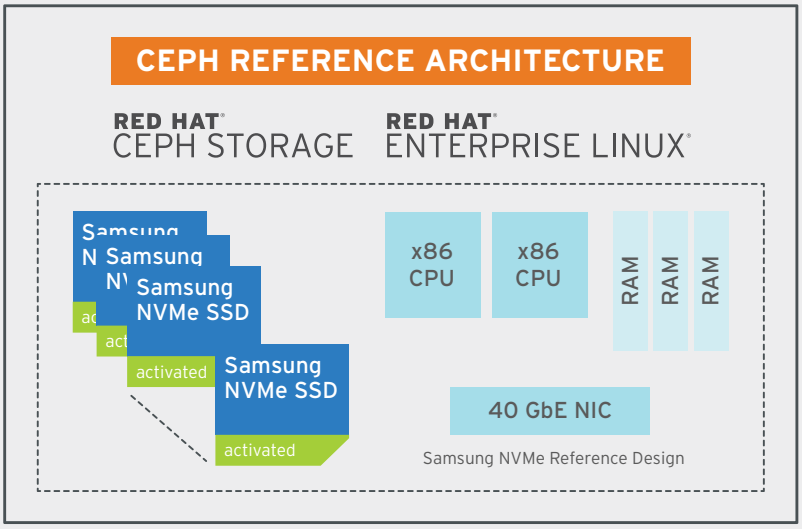

Figure 3. Red Hat Ceph Storage/NVMe reference architecture

Features of the joint reference architecture include:

- OpenStack integration
- S3 and Swift support using RESTful interfaces
- High performance:
  - 700K+ IOPS for a small (4KB) random IO across a three-node Ceph cluster
  - 30GB/s for a large (128KB) sequential IO across a three-node Ceph cluster
- Extensively tested by Red Hat and Samsung
- Uses widely available, standards-based hardware
- Striping and replication across nodes to enable data durability, high availability, and high performance
- Automatic rebalancing using a peer-to-peer architecture to deliver instant capacity and ensure data protection with minimal operational effort
- Minimal or no downtime cluster upgrades
- Less power consumption and higher reliability than similar capacity HDD configurations

The hardware supporting the reference architecture is the Samsung NVMe Reference System, a compact two rack-unit (2U) server with a high-performance, all-flash NVMe scale-out design. This dual-socket Xeon-based system supports up to 24 x 2.5" hot-pluggable Samsung NVMe SSDs, providing extremely high capacity and density. It uses 4x 40Gb/s networking with remote direct memory access (RDMA) for high-performance connectivity. It is based on PCIe Gen3 NVMe SSDs and offers the lowest latency in the industry with an optimized data path from the CPU to the SSDs. Each SSD slot in the system provides power and cooling for up to 25W to support current and future large-capacity SSDs, as well as SSDs with different performance levels. The maximum server capacity is 46, 92, or 153 TB, depending on the model of NVMe SSD drive deployed.

The balance of high-performance compute, networking, and storage designed into the Samsung NVMe Reference System means that performance scales more linearly, without tending to be over-provisioned along any one of the components. With Ceph's distributed cluster capabilities, enterprises can combine multiple Reference System servers to deliver a performance tier capable of handling hundreds of thousands of IOPS, alongside a scale-out capacity tier.

The Samsung NVMe Reference System is available for immediate purchase in fully supported, off-the-shelf configurations through StackVelocity (a business unit of Jabil Systems) as the Greyguard platform.

SUMMARY

Enterprises have come to know and trust Red Hat Ceph Storage as a reliable, scalable storage tier for capacity- and cost-constrained applications like media serving and cloud storage. As this trust has grown, many enterprises seek Ceph's benefits for their performance-oriented workloads as well. Historically, however, this has been challenging, principally due to performance limitations inherent in legacy drive interfaces. Samsung Enterprise NVMe SSDs, by using a modern, performance-oriented interface designed for current and future solid-state drives, enables these more demanding workloads.

To ensure that enterprises can easily adopt Red Hat Ceph Storage running on Samsung NVMe SSDs, the two companies have partnered to develop and validate a joint solution. A few of the many benefits of this solution include:

- **High performance.** Taking advantage of both the distributed nature of Ceph and the increased performance of Samsung NVMe SSDs, the joint solution offers significant performance gains.
- **Reduced cost.** Samsung NVMe SSDs offer greater density and power efficiency than legacy drives, allowing enterprises to reduce capital and operational costs. Furthermore, the lower failure rate of SSDs relative to HDDs allows replication levels to be reduced for some applications, further lowering storage costs.
- **Agility.** Enterprises using Ceph with Samsung NVMe SSDs now have an elastic layer for provisioning a converged pool of storage that is both fast enough to support IOPS- and throughput-constrained applications and affordable enough to support capacity-constrained applications.
- **Multiprotocol.** With support for object, file, and block access to the same underlying storage pool, Ceph allows enterprises to simultaneously support both cloud-native and traditional workloads in the same environment.

To validate the performance of the Red Hat and Samsung solution, and to demonstrate its suitability for high-performance workloads, the companies have produced a joint reference architecture that has been extensively tested and documented, the results of which are available in the technical publication, "High-performance cluster storage for IOPS-intensive workloads."[1]

While the Samsung NVMe Reference System is a product and well-supported option available to customers worldwide, NVMe SSDs are an industry standard, allowing customers to design their own software-defined storage systems to meet their unique use case requirements. For more information on the Red Hat and Samsung solution, please visit http://www.samsung.com/semiconductor/support/tools-utilities/All-Flash-Array-Reference-Design/

1 https://www.redhat.com/en/resources/ceph-samsung-reference-architecture

ABOUT RED HAT

Red Hat is the world's leading provider of open source software solutions, using a community-powered approach to provide reliable and high-performing cloud, Linux, middleware, storage, and virtualization technologies. Red Hat also offers award-winning support, training, and consulting services. As a connective hub in a global network of enterprises, partners, and open source communities, Red Hat helps create relevant, innovative technologies that liberate resources for growth and prepare customers for the future of IT.

NORTH AMERICA
1 888 REDHAT1

EUROPE, MIDDLE EAST, AND AFRICA
00800 7334 2835
europe@redhat.com

ASIA PACIFIC
+65 6490 4200
apac@redhat.com

LATIN AMERICA
+54 11 4329 7300
info-latam@redhat.com

facebook.com/redhatinc
@redhatnews
linkedin.com/company/red-hat

redhat.com
#f7626_0617