WHITEPAPER

August 2018

Mission Peak

High-performance shared NVMe-oF™ storage solution with Samsung NF1 SSDs

SAMSUNG AIC Mellanox TECHNOLOGIES E8 storage MEMORY SOLUTION

Contents

Executive summary

The rapid acceleration of data creation has put traditional storage architectures under increasing pressure because it lacks the scalability and flexibility required to fulfill future capacity and performance requirements. The data tsunami calls for a new innovative hardware and software ecosystem, which break the barriers of existing designs by offering unforeseen density and performance.

Industry leaders Samsung, AIC, Mellanox, E8 Storage and Memorysolution have partnered to develop the "Mission Peak" platform, which is a 1U server with up to 550TB of high-speed NF1 form factor NVMe SSDs, offering more than 50% higher capacity than traditional storage solutions within the same footprint. NVMe is a new protocol that has been designed solely for SSDs, and when combined with E8 Storage's innovative software, "Mission Peak" can offer up to 10 million IOPS of shared storage accessible via Mellanox Ethernet networking.

The "Mission Peak" solution is now available for PoCs in eShelter's Innovation Lab in Frankfurt, Germany. By partnering with one of the leading colocation providers, Samsung and its partners are able to provide a seamless "PoC as a service" offering, in which end-customers can login remotely to the PoC platform for testing and benchmarking. Compared to traditional PoCs, a remote PoC can reduce the time from months to just 1-2 weeks and as all hardware has been pre-installed and tested; there is no additional cost or workload for the end-user to perform the PoC.

Introducing the technologies powering **Mission Peak**

The need for high-density low-latency storage

Every second we create huge amounts of new data. The data visible to us, for example the photos and videos we take on our smartphones, are just the tip of the iceberg. Most of the data is generated by the ubiquitous sensors and algorithms that surround and track us all over the world.

According to a study by IDC, 20 zettabytes (ZB) of data was created globally in 2017. This is equal to 38 petabytes (PB) of data being created every second, a figure which is forecast to grow to over 300PB per second by 2025.

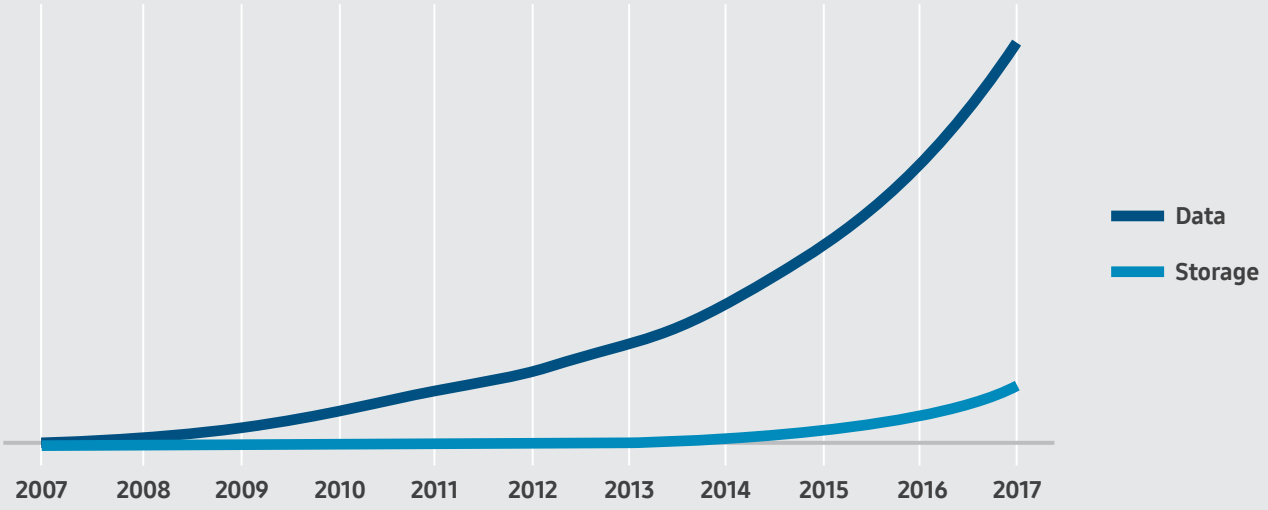


Figure 1. Data versus storage growth in the past decade

The exponential growth of data has increased pressure on data center and IT executives to ensure storage infrastructures can keep up with the future demand. However, historically data has grown at a much higher rate than storage density, which creates the question: how can traditional storage architecture support the quantum leap that data growth will be undergoing over the coming years?

The hard truth is that it cannot. This has led Samsung to partner with AIC, E8 Storage, Mellanox and Memorysolution to bring a disruptive "Mission Peak" solution to the market with previously unseen density and performance.

Built upon the new industry standard NF1 SSDs, Mission Peak packs in up to 550TB of usable SSD capacity in only 1U, which is over 50% more than traditional 2U storage appliances can support. And by using NVMe protocol and E8 Storage's revolutionary software, Mission Peak can support up to 10 million I/O operations per second (IOPS), which is up to 20 times more than what traditional all-flash arrays provide.

Unleashing the true performance of SSDs with NVMe™

HDDs primarily use serial ATA (SATA) and serial-attached SCSI (SAS) interfaces as the links between the drives and the host system. To enable compatibility with existing systems, SSDs have also been using SATA and SAS interfaces, but these were not designed for the high performance that SSDs offer. Hence SSDs have been severely limited by the interface, but non-volatile memory express (NVMe), a new interface designed from ground up for SSDs, is finally becoming mainstream and enabling end users to unleash the full potential of SSDs.

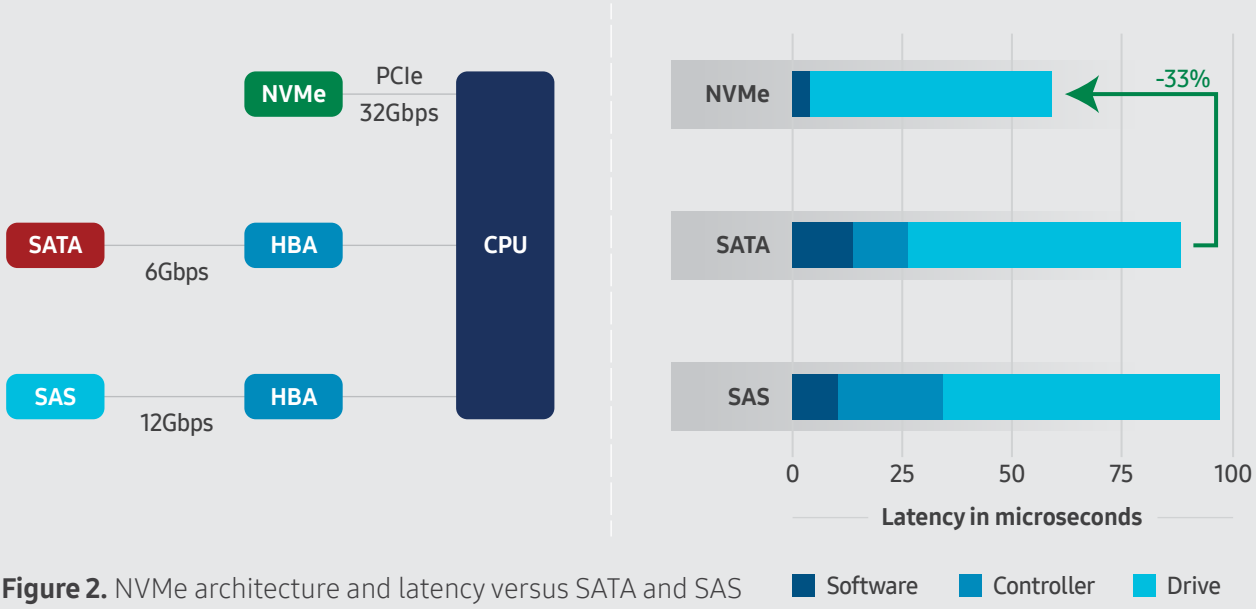


Figure 2. NVMe architecture and latency versus SATA and SAS

NVMe is not actually an electrical interface like SATA and SAS — it's a software interface built on top of peripheral component interconnect express (PCIe), which is the high-speed interface available directly on the central processing unit (CPU). While SATA and SAS drives require a separate drive controller, NVMe enables drives to connect directly to the CPU via PCIe, reducing the latency significantly.

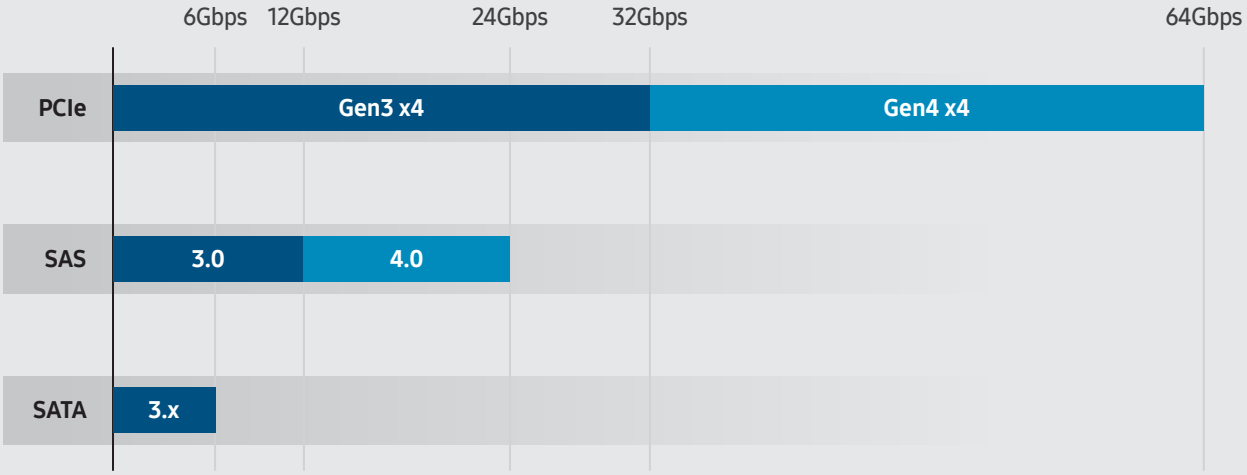


Figure 3. Comparison of SSD interface bandwidths

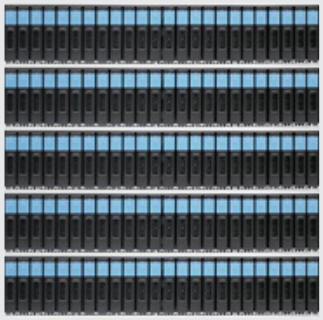
In addition to lower latency, NVMe enables much higher data transfer speeds. SATA interface has a bandwidth limit of 6Gbit/s, while SAS offers up to 12Gbit/s. As NVMe uses the PCIe interface, the available bandwidth scales with the number of PCIe lanes, with one PCIe 3.0 lane providing 8Gbit/s. Mainstream NVMe SSDs use four PCIe Gen3 lanes for bandwidth of up to 32Gbps. Over the next two years PCIe will move to Gen4, doubling the bandwidth to up to 64Gbps.

To use the full bandwidth of PCIe interface, NVMe supports up to 64,000 I/O queues, with each queue supporting up to 64,000 I/O commands. SATA and SAS interfaces only support one I/O queue with 32 and 254 commands respectively. The increased number of queues enable multiple applications and users to access NVMe SSDs in parallel for maximized throughput.

NVMe over Fabrics:
Shared NVMe storage at the latency of local storage

While NVMe SSDs are already deployed in data center and enterprise environments, the use cases have been limited to local storage such as cache or application drives because of the lack of optimized networking standards. The issue with local storage is that it's only available to the server it's attached to, which creates an "islands of storage" bottleneck where the SSDs are not used to their full capability in terms of capacity and performance.

Before
- 96 servers in a rack
- 96 * 2.2TB SSDs in servers



Local SSD utilization: 30%

After
- 92 servers in a rack
- 24 * 4TB SSDs in E8 Storage appliance

Central SSD utilization: 88%

Figure 4. Islands of storage and underutilization of capacity

The limits of local storage are the reason why SAN, DAS and NAS architectures have become mainstream in enterprise deployments as separating storage from compute allows for more flexibility. However, the existing architectures and protocols were designed for HDDs and don't offer the necessary latency and scalability for NVMe SSDs.

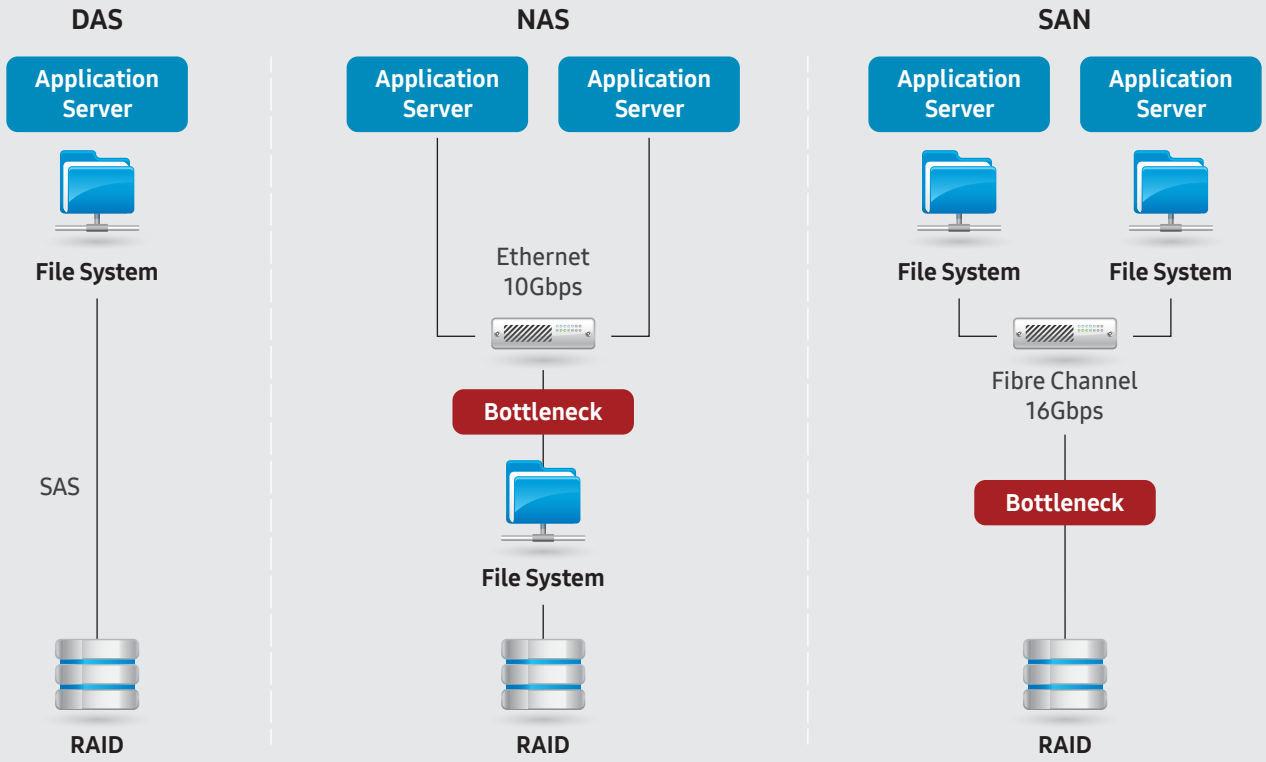


Figure 5. Comparison of traditional storage architectures and the bottlenecks

NVMe over Fabrics (NVMe-oF™) is a new innovative protocol that defines a standard for NVMe block storage that can run over a variety of existing storage networking fabrics such as Fibre Channel, Ethernet and InfiniBand. The NVMe-oF protocol has been designed by NVM Express, which also manages the NVMe standard, and its members include companies such as Samsung and Microsoft.

The key objective of the NVMe-oF standard is to provide a scalable low-latency protocol that enables NVMe SSDs to be used in any kind of storage architecture at their native latency, regardless of whether the drives are accessed locally or remotely over a network. This is achieved by removing the unnecessary overheads that exist in HDD-optimized legacy protocols such as SCSI and also eliminating the need for protocol translation.

Furthermore, NVMe-oF supports remote direct memory access (RDMA), which enables direct server-to-server data transfers for higher bandwidth and lower latency. Without RDMA, the initiator server needs to go through several kernel layers to read data in the target server, consuming CPU cycles and adding latency. So not only does RDMA reduce the latency, it also frees up CPU resources for the application, which can improve overall performance or even allow for lower cost CPU to be used for higher cost efficiency.

E8 Storage:
Unlocking NVMe performance with innovative software

As the first centralized NVMe storage solution in the industry, E8 Storage delivers high-performance, low-latency storage and simplified storage management without compromising on reliability, availability or scalability. The E8 Storage system scales up to 126 host servers per E8 Controller, each connected concurrently to shared storage. With support for shared read/write volumes, applications can be deployed with the performance acceleration of clustered parallel databases and file systems.

E8 Storage appliances provide up to 10 times the performance of other all-flash arrays, with consistently strong performance and low latency. Built on high-performance 100Gb/s Ethernet or InfiniBand networks, E8 Storage delivers near line-rate performance up to 10 million read IOPS and 40GB/s read throughput, with latency on par with SSDs hosted locally within a server.

The E8 architecture

E8 Storage has developed a revolutionary storage architecture that meets the demands of mission-critical customer applications. E8 Storage's patented software architecture was built from the ground up to leverage NVMe.

The key to performance and scalability is the separation of control and data path operations between the E8 Controller software and the host-side E8 Agents. The E8 Controller software provides centralized control and management, while the E8 Agents manage data path operations with direct access to shared storage volumes. With NVMe protocol supporting exponentially more connections per SSD, the E8 Agents use RDMA to give each host server a direct connection to the drives. This unique approach enables the E8 Agents to perform up to 90% of the data path operations between the host server and storage, ensuring that the controller is not a bottleneck to performance.

With simple connectivity via Ethernet or InfiniBand, new appliances can be added to the network seamlessly and new storage capacity provisioned to host servers without affecting host availability. E8 Agents can access storage volumes from multiple E8 Storage appliances, enabling capacity scaling into the petabyte range. With support for shared read/write volumes, applications can be deployed with the performance acceleration of clustered parallel databases and file systems.

The E8 Storage software architecture was designed for high availability, with data protection and disaster recovery features essential for today's data centers. The host-side E8 Agents operate independently of other agents in the E8 Storage system, performing data path operations for that host only. If any host server loses access to the E8 Controller, there is no impact to the availability of the data stored on the E8 Controller, and remaining host servers continue operations with no performance impact.

NF1:
Maximizing storage density with a new SSD form factor

Much like interfaces, the physical form factors of SSDs have followed the legacy of HDDs. The SSDs used in the enterprise today are primarily 2.5" to allow the same slot to be populated with an SSD or HDD, but this approach is not optimal for SSDs.

The form factor revolution started in the PC industry driven by the race towards ever-thinner laptops, which obviously couldn't be achieved with traditional 2.5" drives. mSATA was first announced in 2009, and in 2012 the now-mainstream M.2 form factor was released. Both mSATA and M.2 are card-like form factors rather than drives, enabling much thinner host system form factors by not requiring a case.

2.5"

+ Compatible with HDD infra
- Capacity Limited
- Thermals not optimal

M.2

+ More space efficient vs 2.5"
- Capacity not optimized
- Only Client-oriented feature

NF1

+ Optimized for 1U server
+ 2x capacity vs M.2
+ Support hot-plug/dual-port

Figure 6. SSD form factor evolution

While some server designs adopted M.2, it never gained popularity in the broader enterprise market because of a lack of specific features. To better address the enterprise market, Samsung has developed a new NF1 form factor, which improves the M.2 concept by adding hot-plugs and dual-ports along with higher capacity support. Using the industry-leading 512Gbit 64-layer V-NAND technology, Samsung has been able to develop a 16TB NF1 NVMe SSD.

Thanks to its thin, caseless design, the NF1 form factor allows up to 36 SSDs to be installed in a 1U chassis, enabling IT architects to create petabyte-class storage solutions in a slim 2U form factor. Compared to 2.5" and 3.5" SSD form factors, NF1 can reduce the storage space requirements in a data center by over 30%.

Mission Peak:
Building a high-density dual-socket 1U system around NF1

Mission Peak is a reference design server developed together by AIC and Samsung with the goal of delivering maximum SSD capacity in a 1U form factor. The design features a custom-made AIC backplane that enables up to 36 NF1 SSDs to be installed, which are front-loadable and hot-swappable.

The dual-socket configuration offers high and balanced PCIe bandwidth. Both CPUs are also accompanied by a PLX PCIe switch to ensure that all 36 SSDs have a full four-lane path to the CPU. With four slots for NICs, Mission Peak can support up to 300Gbps of network bandwidth using Mellanox ConnectX 5 series, which can enable nearly 10 million IOPS when using E8 Storage software.

Mission Peak reference design specifications

Form factor	1U dual-socket
Dimensions (W x D x H)	**mm:** 482.6 x 801.2 x 44 **inches:** 19 x 31.5 x 1.7
Motherboard	AIC server board "Lynx"
Processor	2x Intel Xeon scalable processor family (LGA-3647 socket)
Chipset	Intel C620
Memory	24x DIMM slots across 12 channels - DDR4-2666 (1DPC) or DDR4-2400 (2DPC)
Storage	36x NF1 front-loadable hot-swap bays 2x M.2 internal (PCIe Gen3 x4 or SATA 6Gbps)
Networking	2x add-in cards (PCIe Gen3 x) 2x OCP Mezzanine 2.0 cards (PCIe Gen3 x)

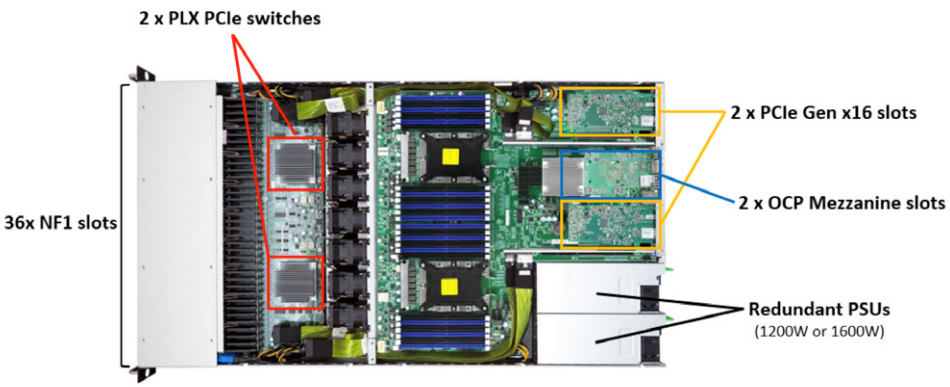


Figure 7. Mission Peak system design overview

PoC as a service: bringing a PoC to the modern era

How can I test Mission Peak?

The number-one goal for Samsung and its partners was to find a way to make testing Mission Peak happen as easy as possible for both the vendors and end customers. A traditional PoC can be a burden for all parties because vendors need to sample customers individually and it's up to the end customer to integrate all the hardware and software together and put the system into a data center for testing. Ultimately, it's time consuming and expensive, and that results in fewer PoCs and successful deployments as vendors and end customers cannot justify resources for every project.

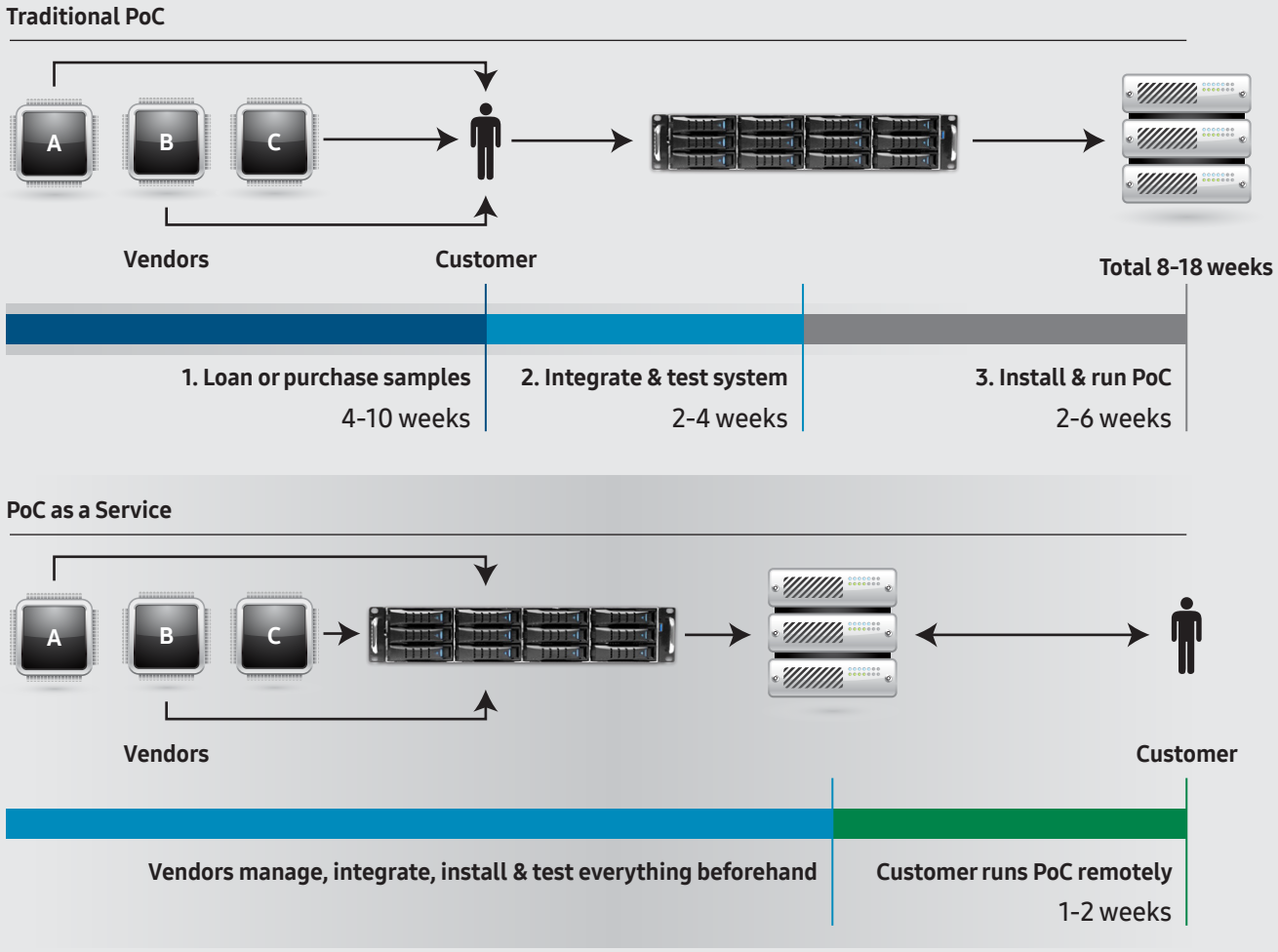


Figure 8. PoC acceleration with an "as a service" approach

To solve the problem, Samsung and its partners are working with eShelter, one of the leading colocation providers in central Europe, to install Mission Peak in eShelter's innovation lab in Frankfurt. The innovation lab is a room in a fully operating data center where vendors can install their latest technology for PoC purposes in a near-production environment, allowing for a realistic performance projection of a similar system in productive operation.

This approach enables customers to test Mission Peak remotely without investing in any hardware or data center space. In eShelter's experience, the innovation lab can cut down PoC time from an average of several months down to just 1–2 weeks because all hardware is pre-tested and configured. It's about time that PoCs became available as a service!

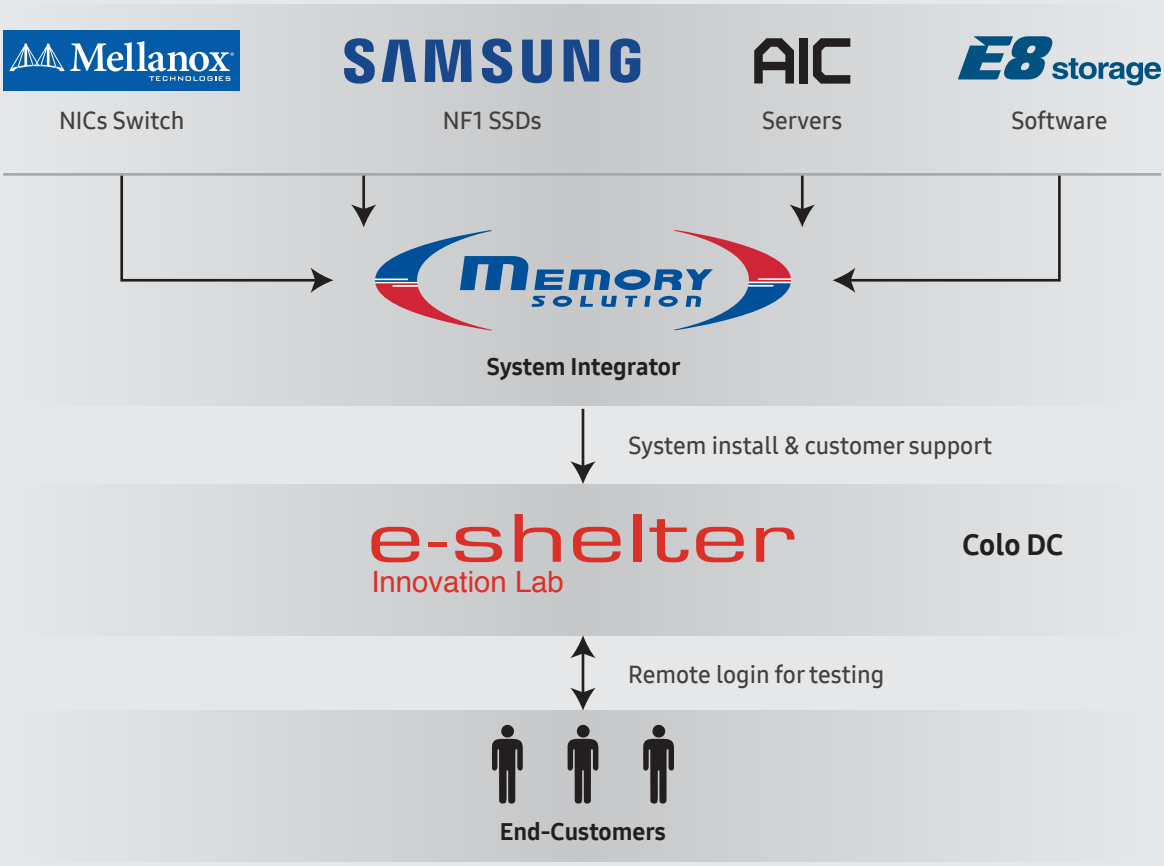
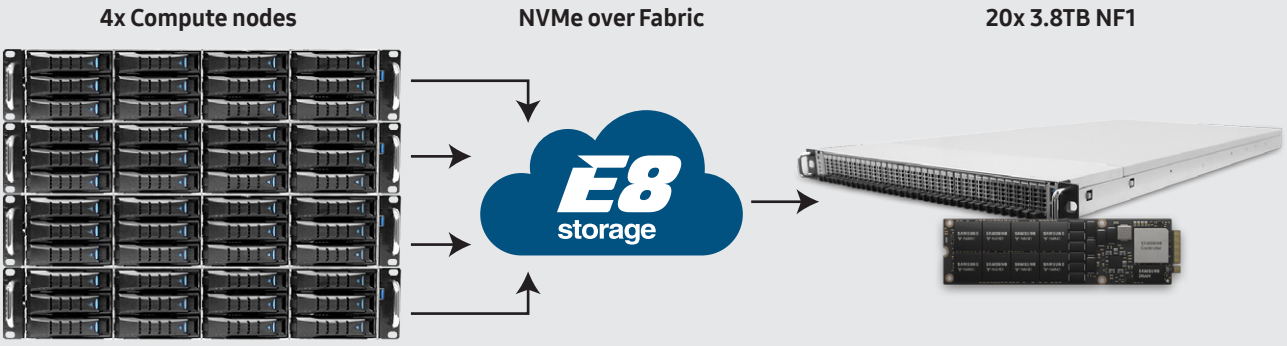


Figure 9. Greatly improved PoC availability for end-customers

Mission Peak PoC configuration and specifications

The PoC setup consists of four dual-socket compute nodes that are connected to one Mission Peak target over 100Gbps Ethernet. Each compute node as well as the Mission Peak system is equipped with two Intel Xeon scalable CPUs and 256GB of DDR4 2400MHz RDIMM memory. The compute nodes have a total of 16TB of storage provided by Toshiba 4TB SAS HDDs, whereas the Mission Peak system hold a massive 76.8TB of Samsung PM983 NF1 NVMe SSDs. With 36x 15.36TB NF1 SSDs installed, the Mission Peak system can support up to 552.96TB in 1U form factor.



Networking is enabled by Mellanox 100Gbps Ethernet using 16-port SN2100 switch and ConnectX 5 series NICs. The compute nodes are each equipped with one 100Gbps NIC, while the Mission Peak system has two 100Gbps NICs with maximum bandwidth of 25GB/s.

Mission Peak PoC Performance

4KB random read	2M IOPS
4KB random write	260K IOPS
64KB sequential read	18GB/s
64KB sequential write	13GB/s
Read latency (QD1)	70µs
Write latency (QD1)	36µs

* Performance was tested using similar setup. Final PoC performance figures will be available soon.

The performance of the PoC system is limited by the number of compute nodes. E8 Storage software supports the connection of up to 126 compute nodes to one target and thus with more compute and network bandwidth the solution could scale to up to 10M IOPS and 40GB/s of throughput.

Mission Peak PoC Specifications

	Compute (4x)	Mission Peak (1x)
Form factor	2U	1U
Chassis	AIC SB202-VG	AIC FB127-LX
Motherboard	AIC "Virgo"	AIC "Lynx"
Processor	2 x Intel Xeon Scalable Silver 4116	2 x Intel Xeon Scalable Gold 5120
Chipset	Intel C620	Intel C620
Memory	256GB DDR4-2400 RDIMM (8 x 32GB)	256GB DDR4-2400 RDIMM (16 x 16GB)
Storage	4 x Toshiba MG04SCA40EA 4TB SAS	20 x Samsung PM983 NF1 NVMe 3.84TB
Boot drive	2 x Samsung PM871b M.2 128GB	2 x Samsung PM871b M.2 128GB
Networking	1 x Mellanox ConnectX 5 100GE	2 x Mellanox ConnectX 5 100GE
Operating system	Linux	Linux
Software	E8 Storage Agent software	E8 Storage Controller software

About Samsung Electronics Co., Ltd.

Samsung inspires the world and shapes the future with transformative ideas and technologies. The company is redefining the worlds of TVs, smartphones, wearable devices, tablets, digital appliances, network systems, and memory, system LSI, foundry and LED solutions.

For the latest news, please visit the Samsung newsroom at **news.samsung.com**

For more information on Samsung memory products please visit: **www.samsung.com/memory**

About AIC Inc.

AIC is a leading provider of both standard off-the-shelf and OEM/ODM server and storage solutions. With expert in-house design, manufacturing and validation capabilities, AIC's products are highly flexible and configurable to any form factor, standard or custom. AIC leads the industry with over 20 years of experience in mechanical, electronic, system-level engineering as well as a dedication to innovation and customer satisfaction. Headquartered in Taiwan, AIC has offices and operations throughout the United States, Asia and Europe.

For more information, please visit: **www.aicipc.com**

About Mellanox

Mellanox Technologies (NASDAQ: MLNX) is a leading supplier of end-to-end InfiniBand and Ethernet smart interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high-performance computing, enterprise data centers, web 2.0, cloud, storage and financial services.

More information is available at: **www.mellanox.com**

About E8 Storage

E8 Storage is a pioneer in shared NVMe storage for data-intensive, high-performance applications that drive business revenue. E8 Storage's affordable, reliable and scalable solution is ideally suited for the most demanding low-latency workloads, including real-time analytics, financial and trading applications, genomics and large-scale file systems. Driven by the company's patented architecture, E8 Storage's high-performance NVMe over Fabrics-certified storage delivers record breaking performance at half the cost of existing storage products. When performance matters, enterprise data centers turn to E8 Storage for unprecedented storage performance, density and scale, without compromising on reliability and availability. Privately held, E8 Storage is based in Santa Clara with R&D in Tel Aviv, and channel partners throughout the US and Europe.

For more information, please visit **www.e8storage.com**, and follow us on Twitter **@E8Storage** and **LinkedIn**

Memorysolution GmbH

Established in 1997 with head offices in Breisach am Rhein, Memorysolution GmbH specializes in the flawless distribution of high-quality IT components for client, server and industrial applications. Ready-to-use server and IT solutions under the BTO brand Mustang® systems complete the range for specialist IT retailers, systems companies and businesses. Memorysolution has a subsidiary in Hong Kong and branches in Hamburg, Frankfurt, Braunschweig and Bremen. A total of 40 employees support 4,000 customers in 44 countries. The consistent success of the company is based on competent, personal points of contact for the innovative component range, excellent services and added value through IT and server solutions. Customers benefit from strong partnerships with such leading A-brand manufacturers as Samsung, Supermicro, Micron, SK Hynix, Nanya and Crucial.

More information: **www.memorysolution.de**