Block Storage Overview

Introduction / Motivation

Though byte-addressable volatile and non-volatile memory will play a significant role in future infrastructure solutions, these technologies will not completely displace the need for block storage solutions: e.g., to support unmodified operating system and applications, to support archival and cold storage applications, and to support storage rich data services such as data replication, thin provisioning, data protection, etc.

Figure 1 illustrates a traditional local block storage solution today, where a processor is connected to a host bus adapter (HBA) via PCIe® fabrics made up of point-to-point links or PCIe switches. The HBA is connected to the drives via SAS/SATA fabrics made up of point-to-point or SAS / SATA switches. The HBA provides command processing and data movement operations between protocols used on the PCIe and native storage fabrics. Note for this example, the HBA could also represent a local RAID array controller that provides caching, RAS features via mirroring, striping, and data protection, and other data management services. Figure 2 illustrates the evolution of local storage using NVMe SSD drives that simplify and integrate HBA functionality into each drive, allowing them to be placed directly on the PCIe interconnect. This SSD specific solution improves performance, simplifies the system, and lowers cost. However NVMe solutions cannot scale to number of drives and capacities of traditional hard drives and their SAS / SATA interconnect.

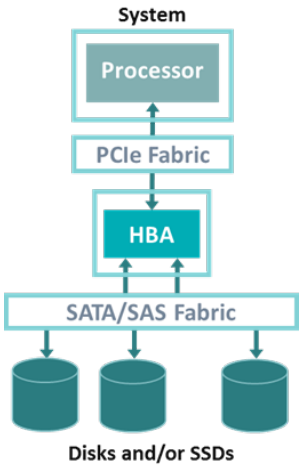

Figure 1: Traditional Local Disk/SSD

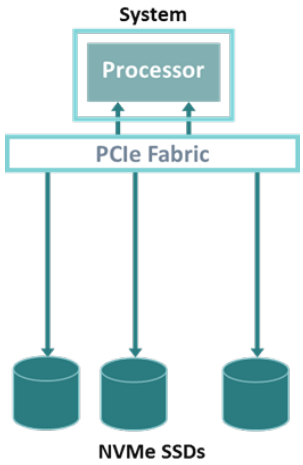

Figure 2: Local NVMe SSD

The configurations in Figures 1 and 2 can be extended to provide limited shared storage features where a subset of the drives can be allocated to small number of servers using SAS expanders or PCIe switches that provide simple component-level partitioning. However, shared storage is typically deployed using enterprise-class block storage systems, such as the dual-node shared storage system shown in Figure 3. Each node provides block storage services to any number of servers over fabrics such as Fibre Channel (FC) or Ethernet (e.g. for iSCSI or FC over Ethernet – FCoE). The nodes share a common set of storage drives across another back-end fabric that can be SAS / SATA or NVMe. At the heart of these nodes is the “Data Service Engine” which is typically a hardware-based accelerator that provides data protection and acceleration (RAID), thin provisioning, data replication, and other value add services. These systems are typically designed to scale in capacity and performance by adding nodes to the cluster and using a fabric (often proprietary) to enable the nodes to exchange data and status. Each node also has

its own storage class memory (SCM) for caching, processors / CPUs for management, and front and back-end HBAs for server and storage connectivity, respectively.

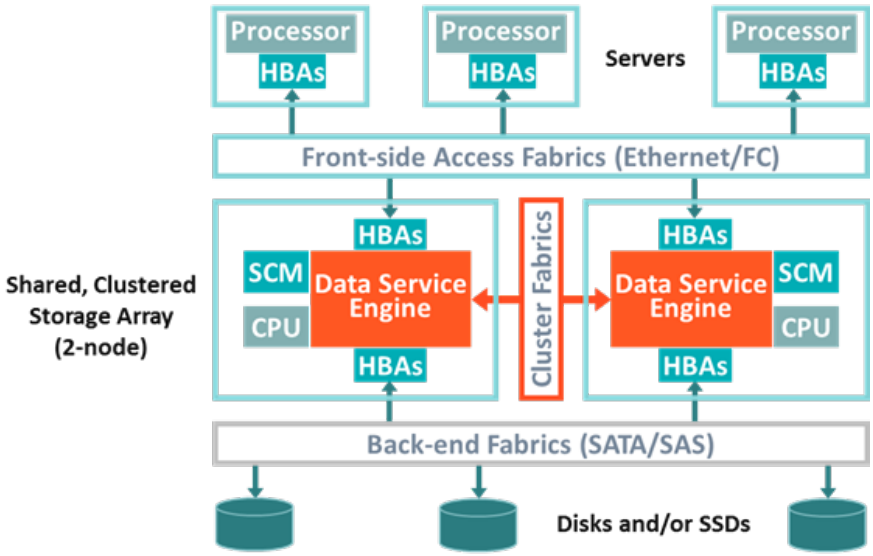

Figure 3: Traditional Shared Block Storage System

But these traditional local and shared block storage architectures face challenges, including:

- Each architecture uses a different combination of interconnects and technologies.
- Each architecture uses as many as three types of interconnects / fabrics (in addition to PCIe for HBA connectivity)
- Each architecture was never intended to support byte-addressable storage-class memory (SCM) , thus may be unable or ill-suited to provide a viable infrastructure for this new type of non-volatile media and application access methodology
- Shared storage architectures are typically implemented using a mix of custom, semi-custom, and industry standard components.

These challenges increase cost and complexity, limit solution composition and flexibility, slow innovation, and constrain industry and customer agility. Gen-Z is a new data access technology that addresses these challenges and more by specifying a common architecture and universal protocol that enables multiple component types – DRAM memory, SCM, block storage, SoCs, FPGAs, graphics, DSP, accelerators, I/O, etc. – to communicate with one another. This yields the following advantages:

- Systems utilizing universal internal slots and external bays can support interchangeable memory, storage, or I/O modules
- Common interconnect for both local and shared storage solutions lowers cost and accelerates innovation
- Facilitates storage systems that can support traditional block access as well a future byte-addressable access to SCM
- Reduces the complexity of shared storage systems by utilizing a common interconnect for multiple functions including: front-side access from compute systems, backend-end storage media connections, and cluster interconnect
- Cost and complexity reduction and increased resiliency achieved by eliminating HBAs

Flexible, High Bandwidth Serial Interfaces

In discrete component solutions, Gen-Z utilizes high-speed serial physical layer technology supporting 16, 25, 28, 56, and 112 GT/s rates on each serial signal. Gen-Z supports link interfaces that support 1-256 serial signals for transmit and receive (the most common link widths will be 1, 2, 4, 8, or 16 serial signals). Gen-Z interfaces support both copper electrical and optical connectivity between components. Gen-Z supports symmetric link interfaces where there are an equal number of transmit and receive serial signals. Gen-Z also supports asymmetric link interfaces where the number of transmit and receive serial signals is not the same. Asymmetric links enable solutions to tailor read and write bandwidths to application-specific needs, e.g., most applications require higher read bandwidth than write bandwidth.

Figure 4 illustrates single-link read and write bandwidths when operating in symmetric or asymmetric mode. Gen-Z also enables each link to be configured in symmetric or asymmetric link mode at link initialization to enable solutions to adapt to new application needs. This capability demonstrates that component and system providers can "dial in" the proper Gen-Z link configuration to match the bandwidth and capabilities of block storage components.

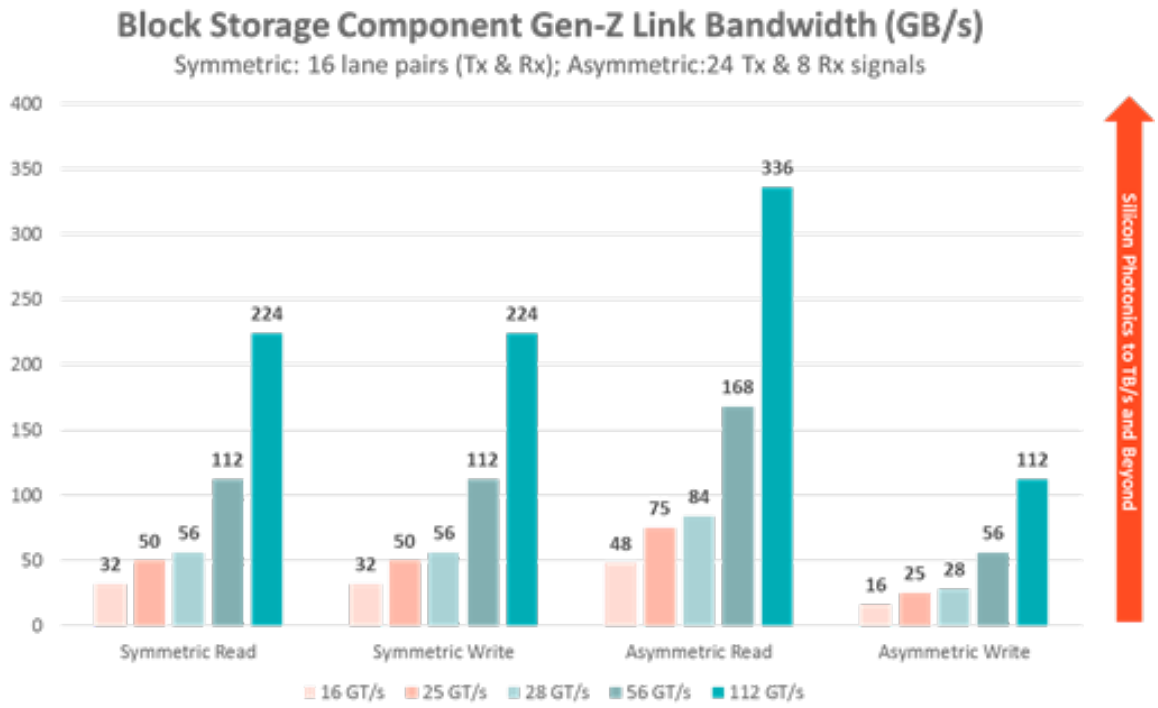

Figure 4: Examples of Gen-Z Bock Storage Component Link Bandwidth

Flexible Topologies with Enhanced Multipath

All Gen-Z enabled components can support multiple link interfaces. Multiple links increase aggregate bandwidth and resiliency, enable robust and flexible topologies, and – for storage systems – provides multipath support that exceeds existing multipath solutions. Figure 5 illustrates an example server system with a number of locally attached Gen-Z SSD or hard disk drive block accessed storage modules. Storage Module #1 utilizes four symmetric Gen-Z links with 4 lanes per link (for a total of 16 transmit and receive lanes) to connect to the Processor. This configuration provides the equivalent bandwidth of a single, wider x16 link, but also provides resiliency in the event of link and path failures. For example, if link A failed, then the Processor and Storage Module #1 will continue to use links B-D for operations. A key aspect of Gen-Z's resiliency features include link level and end-to-end retries and timers such that during link and path failures, the low-level Gen-Z logic automatically recovers operations without impacting upper layer component functionality. This provides a consistent resiliency feature across all component types and vendors. Though similar multipath and resiliency are available with shared storage solutions, they utilize complex, less responsive multipath and management software, whereas Gen-Z enables these capabilities to be implemented entirely in hardware, which simplifies and accelerates failure recovery. In addition, multipath solutions are unavailable in traditional local storage solutions.

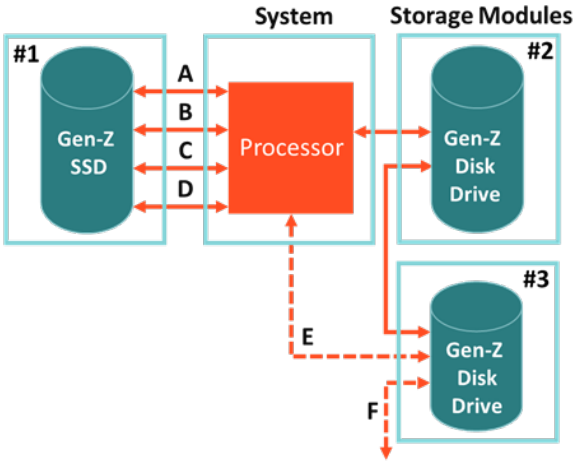

Figure 5: Gen-Z Enables Local Storage Subsystem Design Flexibility

Gen-Z's multipath, multi-link feature also enables flexible topologies for local storage solutions. Figure 5 illustrates a Gen-Z topology where Storage Modules #2 and #3 implement Gen-Z-enabled disk drives with two Gen-Z links and an integrated Gen-Z switch. The Processor is connected though a single Gen-Z link to Module #2, and Module #3 is connected to Module #2 though another single Gen-Z link. Additional drives can be chained as indicated by the optional link F in Figure 5. Module #2's integrated switch allows request and response packets from the Processor destined for Module #3 to flow through Module #2. This configuration provides high capacity without utilizing additional Gen-Z links on the Processor, and is ideal for hard disk drives that cannot individually maintain Gen-Z bandwidths. For added resiliency, multiple links can be used for every link. In addition, the last storage module in the chain can be connected back to the Processor to provide additional resiliency and bandwidth as demonstrated with optional link E in Figure 5.

Composable Shared Storage Systems

Gen-Z enables composable shared storage solutions. Instead of provisioning individual systems containing a fixed mix and number of component types, Gen-Z enables components to be organized into pools, e.g., a SoC pool, a memory pool, a block storage pool, etc. Interconnecting these pools using Gen-Z enables administrators to dynamically compose a solution with any mix and number of components to deliver a software-defined infrastructure. Figure 6 illustrates a composable shared storage system where one of the nodes (dashed line) is logically composed from components within the individual pools. From an operating system and application perspective, the composed system is indistinguishable from a fixed-component system.

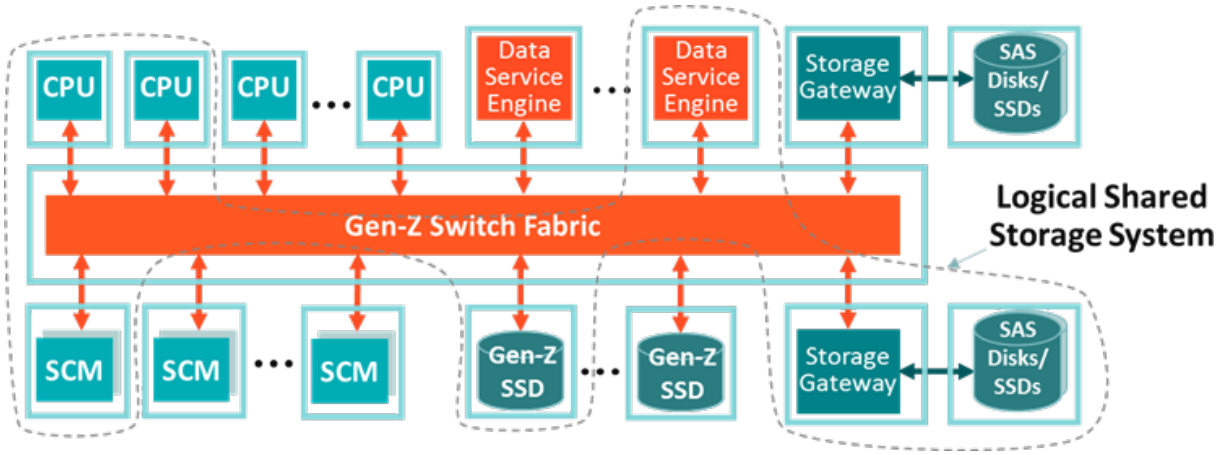

Figure 6: Gen-Z Enables Dynamic Logical Shared Storage Subsystem Composition

Block Storage Access Models

Gen-Z supports a storage Compatibility Mode, where Gen-Z-enabled block storage devices (e.g. Gen-Z SSDs or Disk Drives) can implement a protocol identical in function to an existing standard like NVMe, but using Gen-Z packet formats instead of the PCIe packet formats. This minimizes the changes necessary to support Gen-Z without changing the interpretation or execution of block storage commands. In addition, Gen-Z enables Gen-Z block storage devices to appear to the host system as an NVMe endpoint such that these devices can be discovered and used with unmodified operating systems. This feature is discussed further in the *Gen-Z Logical PCI Overview* document.

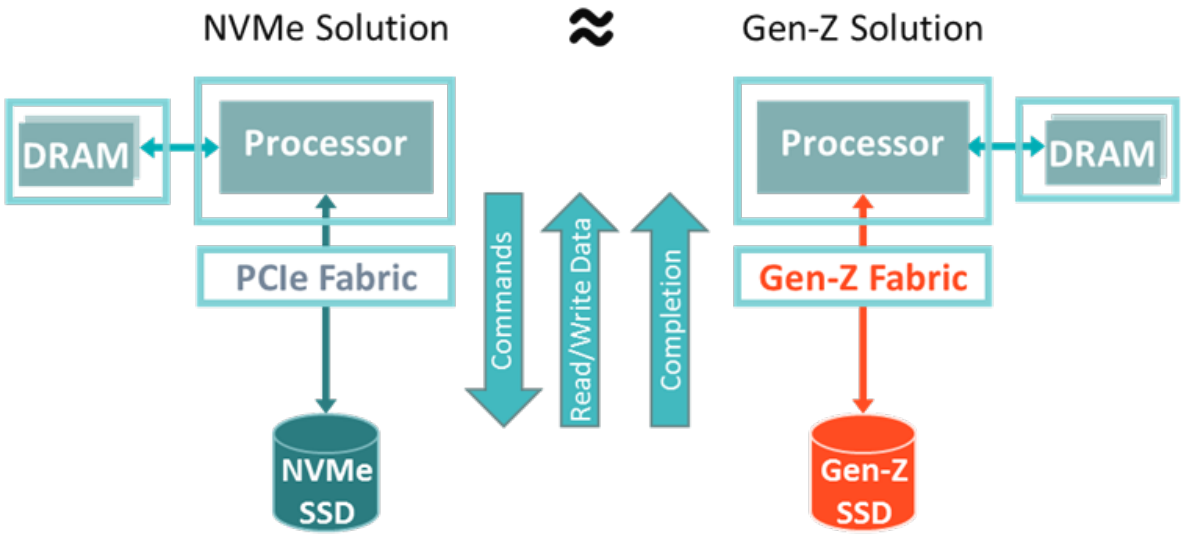

Figure 7: Compatibility Mode Example: NVMe Flow ≈ Gen-Z Flow

As software evolves to take further advantage of Gen-Z, storage access becomes even more streamlined. Figure 8 illustrates simplification using a Unified Mode of operation wherein data is moved between storage and memory using Gen-Z buffer operations. Buffer operations enable a components to 'Put' or 'Get' buffers up to 2^{32} bytes in size from memory or a storage device and directly place it any authorized location in the fabric. Buffer operations eliminate software-driven queue management and resources by using integrated data movers in the Gen-Z logic to perform the buffer operations on behalf of the requesting component, improving solution efficiency and performance. In addition to put and get operations, Gen-Z supports scatter / gather data management and a variety of buffer operation types beyond get / put, including dynamic buffer allocation / release, etc..

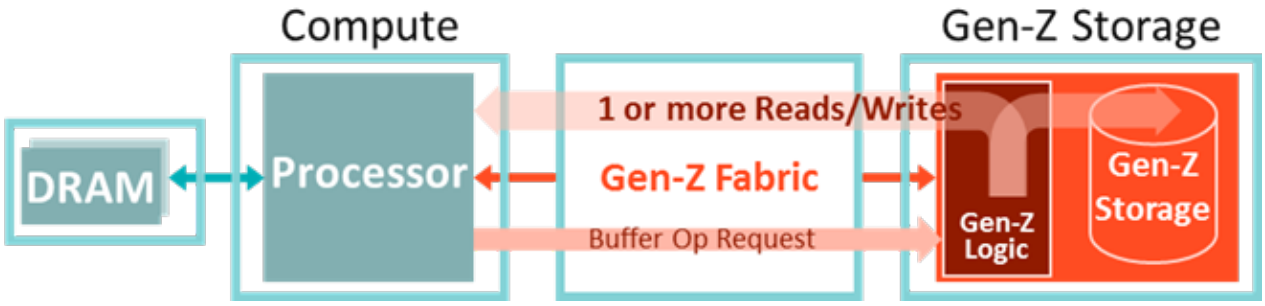

Figure 8: Unified Block Storage Access using Buffer Operations

Figure 9 illustrates how unified access applies to deployments that include Gen-Z fabric-attached SCM. The storage system is instructed to place data directly in fabric-attached SCM (e.g. copy it between application memory and SCM). Figure 9 also illustrates that buffer operation allow byte-addressable reads or writes to or from SCM without imposing any block structure.

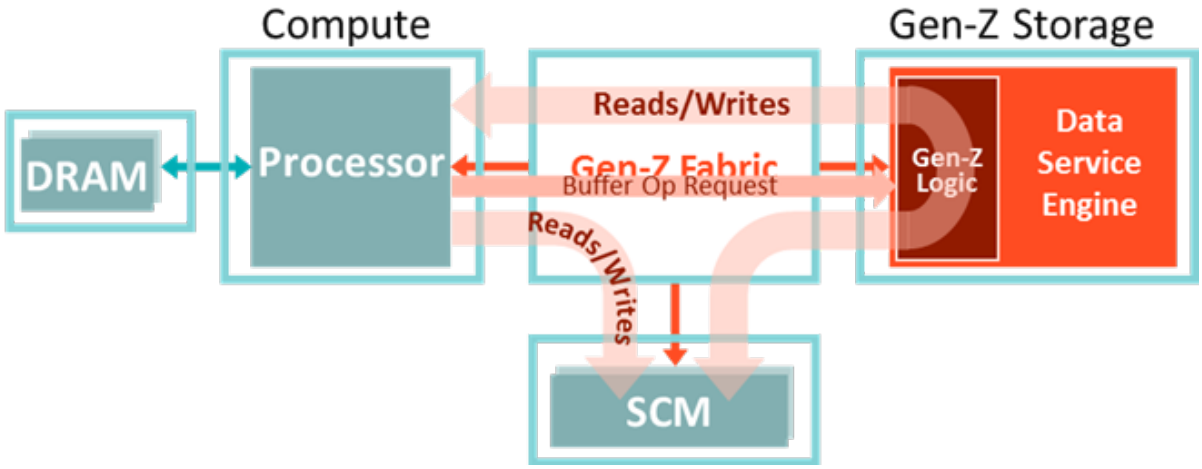

Figure 9: Unified Block Storage Access with Fabric Attached SCM

Security

The Gen-Z architecture supports a combination of hardware-enforced isolation techniques and full packet authentication to prevent errant or malicious components from communicating with unauthorized components or accessing unauthorized resources, including block storage components and their Gen-Z memory resources (e.g. SCM).

Summary

Gen-Z technology is designed to supplement platforms and infrastructure with capabilities to support both traditional block storage access mechanisms and next generation storage and memory-based technology that will be difficult or impossible to achieve with existing storage, block access protocols, interfaces, and interconnect technology. These capabilities include:

- A universal interface for compute, memory, storage, and I/O supports local and shared storage solution
  - o Provides greater solution flexibility with lower capital and operating costs
  - o Facilitates disaggregated storage systems enabling software-defined, composable storage infrastructure
- Offers a robust roadmap of standard serial interface technology for electrical and optical interconnect
  - o Selectable link widths, speeds, and modes enables bandwidth for today and well into the future
- Offers an enhanced multipath solution for better performance and more robust block storage solutions
  - o Provides bandwidth throughput (IOP/s), link and path failure resiliency, and flexible topology options
- Supports compatibility and unified block access mode of operation
  - o Compatibility mode easy implementation and support of unmodified operating systems and applications
  - o Unified mode for streamlined, highest performance storage operations
- Gen-Z offers security features that protect and isolate an infrastructure's valuable storage assets

DISCLAIMER

This document is provided 'as is' with no warranties whatsoever, including any warranty of merchantability, noninfringement, fitness for any particular purpose, or any warranty otherwise arising out of any proposal, specification, or sample. Gen-Z Consortium disclaims all liability for infringement of proprietary rights, relating to use of information in this document. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Gen-Z is a trademark or registered trademark of the Gen-Z Consortium.

All other product names are trademarks, registered trademarks, or servicemarks of their respective owners.