1. Introduction

Gen-Z is a new data access technology that will allow business and technology leaders, to overcome current challenges with the existing computer architecture and provide future opportunities for innovative high performance, low latency solutions that will be open, efficient, simple and cost effective.

2. Background

Years of evolution have allowed layers of software to accumulate upon maturing interfaces that now consume a significant number of CPU cycles and increasing amounts of memory bus bandwidth. The dependence upon the steady increase in processor performance have hidden the latency penalties of evolutionary code design. In an effort to keep up with the demands of CPU performance, designers are vigorously working to revise, extend, and update system architectures for only modest gains in performance. These challenges, coupled with emerging technologies, offer an opportunity to reassess future infrastructure requirements and direction.

3. A better way to access data

Modern computer systems have been built around the paradigm that memory is fast and networks and storage are slow. The language that each speaks has never been compatible and requires layers of software to translate memory commands into storage and network commands and vice versa. All data communication within a computer system spends significant amounts of time traversing and translating from each bus, fabric or network to access a processor's memory.

Over the past three decades, the industry has demonstrated the advantages of memory-semantic communications. Memory-semantic communication moves data between memories located on different components with minimal overhead and latency.

The computer industry has a unique opportunity to eliminate existing system bottlenecks and dramatically improve efficiency and performance by unifying the communication paths and simplifying software by using load/store or memory-semantics everywhere.

4. Why a memory-semantic fabric

The memory bus has one purpose—to provide data to the processor as quickly and as efficiently as possible. The memory bus is specifically designed as a high-bandwidth, low-latency interface using a very simple language to provide maximum application performance. Applications run their very best when run in memory.

The language that the CPU speaks is called load/store. The load/store architecture allows for data transfers from memory to the CPU called a Load, or from the CPU to memory, called a Store. Using Memory-semantics or the load/store access model has proven to be the fastest and most effective way for a CPU to communicate with the memory subsystem. It also gives the CPU access to large pools of memory all the way down to a single byte. This is called byte addressability, and it allows the CPU fined-grained access to efficiently access data. As new persistent memory technologies increase in

popularity, byte addressability extended to the entire system will be a key ingredient in realizing optimal system performance.

Using memory-semantics to extend byte addressability to the entire system will require an architectural change that meets performance and workload requirements while promoting a heterogeneous compute environment with a diversity of media choices. This new architecture will keep pace with future technology trends and eliminate current architectural bottlenecks efficiently and economically.

Imagine a system where all devices are peers to one another and speak exactly the same language through simplified, high performance, low latency communication paths without incurring the translation penalties of current bus architectures.

5. What is Gen-Z?

Gen-Z is a high performance, low latency, memory-semantic fabric that can be used to communicate to every device in the system. By unifying the communication paths and simplifying software through load/store memory-semantics, Gen-Z creates an ecosystem where a variety of high performance solutions that can all work together.

Figure 1a shows a System-on-Chip (SoC) or CPU that connects to a typical DRAM interface over a memory bus. As shown in Figure 1b, Gen-Z makes a key architectural change by decoupling the compute from the media and placing the media-specific functionality where it belongs, with the media. This important change allows for every compute entity in the system to be media agnostic and disaggregated.

As new media trends emerge, the Gen-Z architecture easily comprehends different media types and each device or resource can now scale independently from any other resource in the system—from small installations at remote sites to racks of differing amounts of compute, memory, or storage. See Figure 2 below.
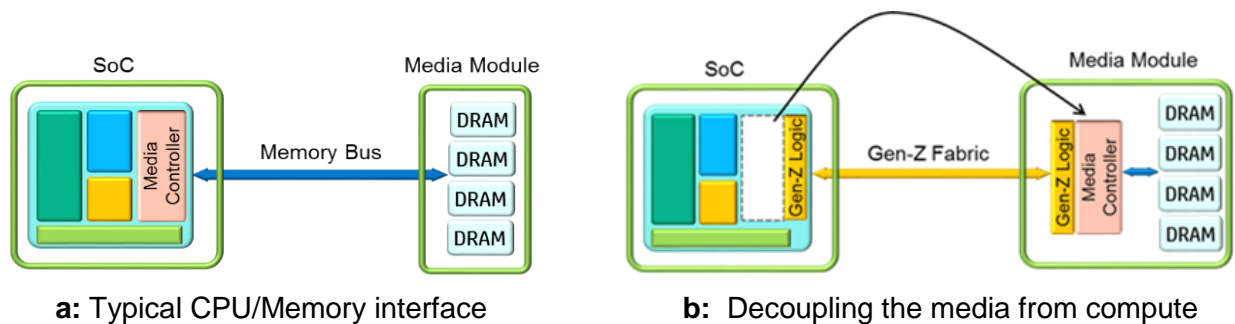

a: Typical CPU/Memory interface   **b:** Decoupling the media from compute

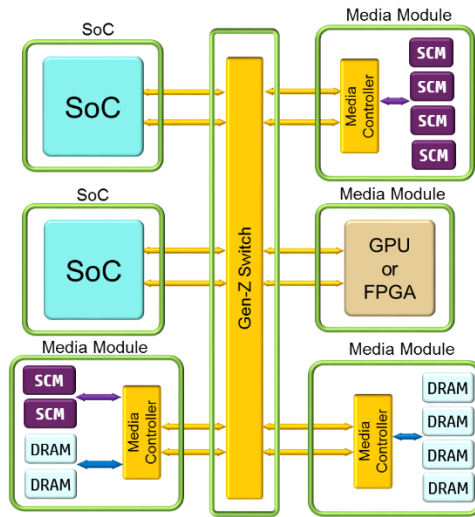Figure 1. Existing SoC with DRAM and an SoC with a Gen-Z Media Controller

Figure 2. Gen-Z architecture illustrating a diversity of media devices

6.   Why Gen-Z?

Current computer architectures are being pulled in many directions due to workload differentiation, new technologies, demands for higher bus speeds, and proposals for new extensions. Every communication path and device in the system is being overwhelmed by an unlimited supply of data that is being transferred, stored and processed. Current hardware can't scale to match the capabilities of new low latency storage class memory (SCM) or take advantage of increasing network communication speeds. Compounded with layers of legacy software and current system bottlenecks has driven the search for new ways to access, manage and store data.

Gen-Z is a high speed memory semantic fabric that not only simplifies every interface but also gives you building blocks to create high performance low latency solutions where every device in the system are peers to one another and speak the same memory-semantic language. Customers will be free to scale with a diversity of compute, memory, and storage resources independently and exactly tailored to every workload and environment while ensuring interoperability and business agility.

Customers who understand the value of data driven decisions that require real-time access to data will be able to turn data into actionable information and make quicker decisions on a scale not possible in today's compute environment.

The vision for a system where all devices are peers to one another and speak exactly the same language through simplified, high performance, low latency communication paths without incurring the translation penalties of current bus architectures. This is the functionality that Gen-Z can deliver.

7.   Key Gen-Z Tenets

Gen-Z delivers compelling customer and industry value by adhering to key business and technical tenets.

7.1. Open

The Gen-Z Consortium was created as a new industry standards body focused on developing an open ecosystem.   The consortium's mission is to deliver a suite of specifications and IP enablement

technology to enable Gen-Z to be integrated into any solution. Gen-Z Consortium membership is open to all. To ensure wide availability, the Gen-Z standard will be published and available free of charge, with no constraints on re-use. The Gen-Z consortium creates a fair, competitive market for implementations of the standard and supports multiple market segments and a diversity of solutions.

5 The founding board members consist of: AMD, ARM, Broadcom, Cray, Dell EMC, Hewlett Packard Enterprise (HPE), Huawei, IDT, Micron, Samsung, SK Hynix and Xilinx.

For more information, visit: genzconsortium.org

7.2. High Performance

The Gen-Z architecture is focused on delivering high efficiency, high bandwidth, and low-latency.

10 ### High Efficiency

As computer hardware has rapidly increased in performance, requisite software changes have not kept pace. Figure 4 below shows that the contribution of software to overall latency is the bulk of the overall latency budget for increasingly faster and faster devices.

With simplified high-speed hardware, the required Gen-Z software interfaces can dramatically improve
15 system performance by removing complexity, overhead, and overall system latency inherent in existing legacy protocols.
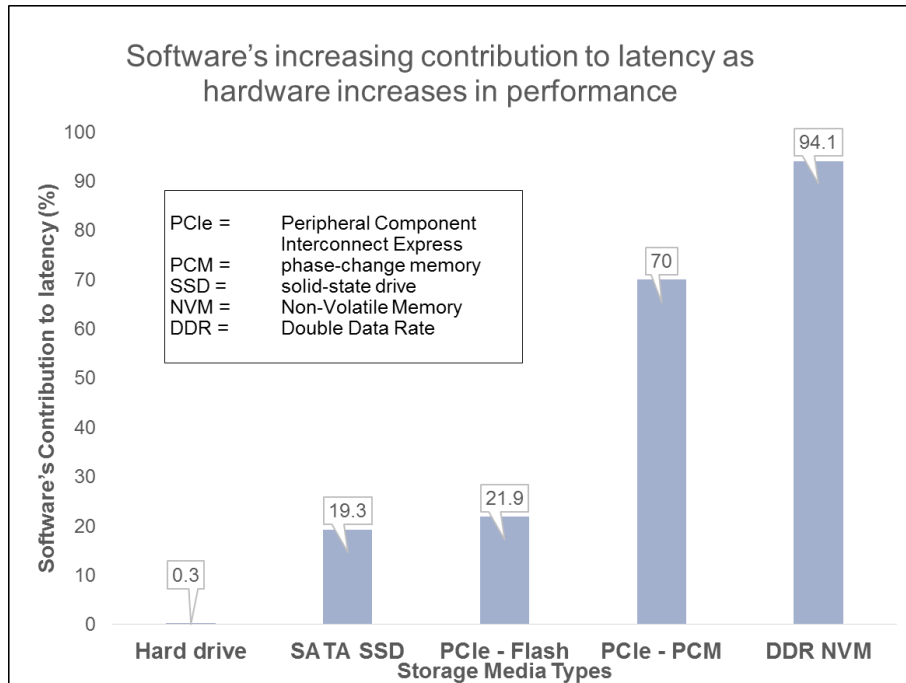
Figure 3 Software's increasing contribution to latency as hardware increases in performance.[1]

High Bandwidth

Figure 4 illustrates not only the bandwidth potential but also Gen-Z's flexibility. A symmetric communication path consists of equal number of read lanes vs. write lanes. Asymmetric communication paths dedicate more lanes to the read path than the write path or vice versa. Much of today's traffic patterns favor reads over writes (many applications issue three times more reads than writes). One of the strengths of Gen-Z is to support both symmetric and asymmetric configurations. Gen-Z support of multiple signaling rates (today's 16, 25, 28, 56 GT/s and tomorrow's 112 GT/s) illustrates how Gen-Z not only will keep pace with the industry's ever growing need for speed but also allows for Gen-Z communication paths to be tuned to specific workload traffic patterns.

---

[1] Reference: Refactor, Reduce, Recycle: Restructuring the I/O Stack for the Future of Storage. Steven Swanson and Adrian M. Caulfield, University of California, San Diego, Published by the IEEE Computer Society, August 2013, p. 51-59
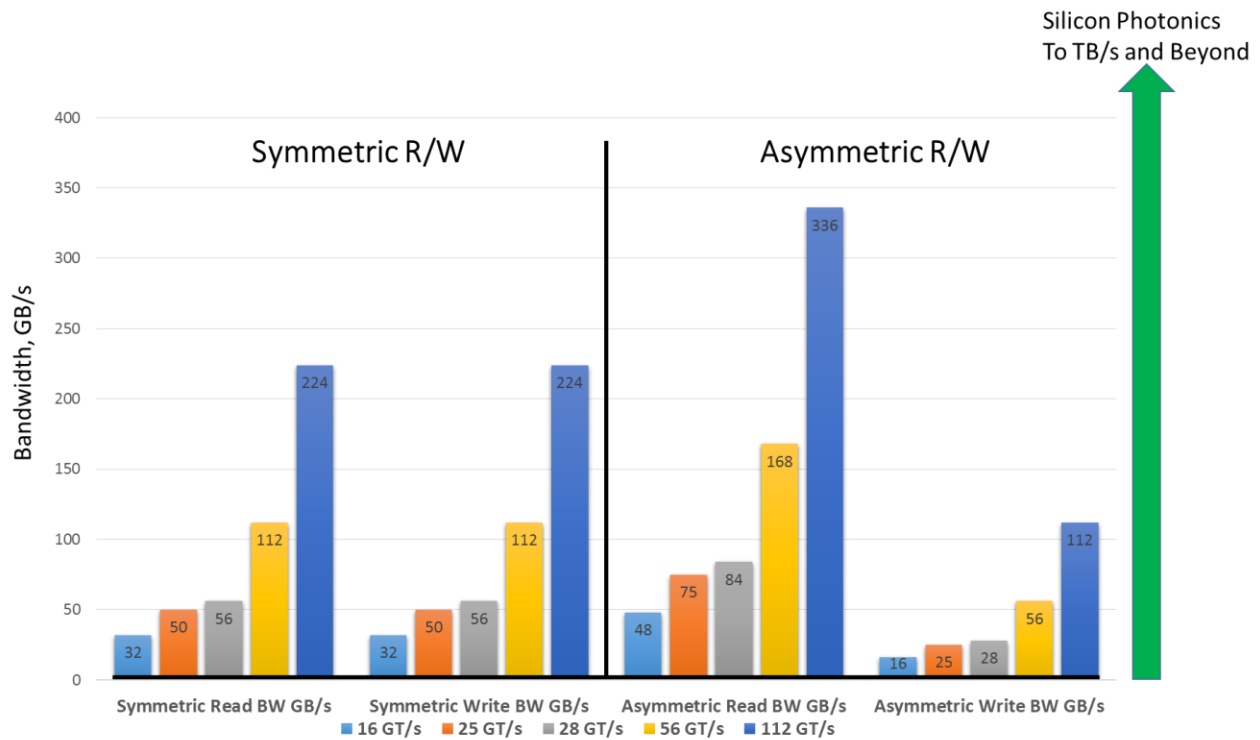
Figure 4 **–** Gen-Z enables High Bandwidth Solutions from today's data rates to future optical bandwidths

Low Latency

As illustrated in Figure 3, Gen-Z offers significant performance advantages at the application level due to a reduced software stack. A traditional server's storage and/or network stack is heavily layered that adds significant amounts of latency to every transaction. Gen-Z operates with a lightweight software interface that performs memory reads and writes directly to the hardware. Gen-Z is designed specifically for low-latency use cases such as Storage Class Memory (SCM), DRAM, Managed DRAM and high-speed messaging.

7.3. Flexibility

Gen-Z's peer-to-peer architecture works with a diverse set of devices such as processors, memory, storage, FPGAs, DSPs, I/O, etc. to create a variety of possibilities including device-level composability for software-defined architectures. Device-level composability is not constrained by platform sheet metal boundaries that can strand unused resources. Virtual machines and Containers can encapsulate and consume each resource in the required amounts without the inefficiencies inherit with current composable methods.

Other possibilities include shared memory solutions, scale-out memory pools, data-centric computing, low-latency point-point access, and switched topologies.

7.4. Compatibility

Gen-Z offers an easy transition from current applications and hardware platforms to high performance architectures of the future. Its simple design will be integrated with systems that incorporate current technologies and form factors without being disruptive to applications and data center operations.

Software compatibility means Gen-Z components and existing solution stacks can be transparently supported by unmodified operating systems and application middleware. For example:

- Gen-Z memory and SCM can be accessed by applications exactly as DRAM is currently accessed
- Gen-Z I/O components can be viewed as PCIe devices
- Gen-Z block storage can be mapped and viewed like any other PCIe block storage device
- A software NIC (commonly called a vNIC) can operate as they do today across any given Gen-Z topology without modification to the IP network stack

As components and solution stacks evolve to take advantage of advanced Gen-Z capabilities and topologies, software changes will be required. For example, rack-scale composable solutions will require new software to manage and coordinate shared memory pools.

Hardware compatibility means Gen-Z hardware components, modules, and devices can also be incorporated into any solution using a wide range of existing mechanical form factors, connectors and cables that allows Gen-Z to be integrated into unmodified platforms and racks.

7.5. Secure

Gen-Z supports a robust hardware-enforced isolation and security framework to help protect customer solutions from cyber threats.   Components and resources can be isolated to prevent unauthorized access or destructive intrusions.

Data plane and control plane communications can be fully authenticated to prevent tampering and anti-replay attacks.

Data payloads can be encrypted using industry or customer-driven algorithms and policies to prevent errant or malicious components from communicating with unauthorized components or accessing unauthorized resources.

7.6. Economic

Gen-Z simplifies the hardware interface by requiring all solutions to support basic read and write operations. This simplification enables Gen-Z hardware and software IP blocks to be leveraged and potentially reused in many designs. Both electrical and optical industry standard cables and connectors can be leveraged as line rates continue to increase. Leverage and reuse leads to lower development and support costs, increased product and solution agility, and efficient device composability.

In addition, the efficiency and bandwidth gains enabled by Gen-Z reduce the amount of power required to move data, enable new data-centric solutions to minimize data movement, and increase solution performance.

8.   Conclusion

Gen-Z offers a unique opportunity for the computer industry to re-invent itself by forging a path in a new direction, with new possibilities in an open and collaborative design environment that allows increased participation within the industry to be more innovative, creative and autonomous.

As more companies participate in the growing Gen-Z ecosystem, the rate of innovation will naturally increase. New designs and products will emerge that address a diversity of workloads and customer challenges. A flexible memory-semantic fabric gives companies building blocks to construct flexible disaggregated, low latency and high performance solutions that scale from the smallest of instances to rack-scale and beyond.