Hewlett Packard Enterprise

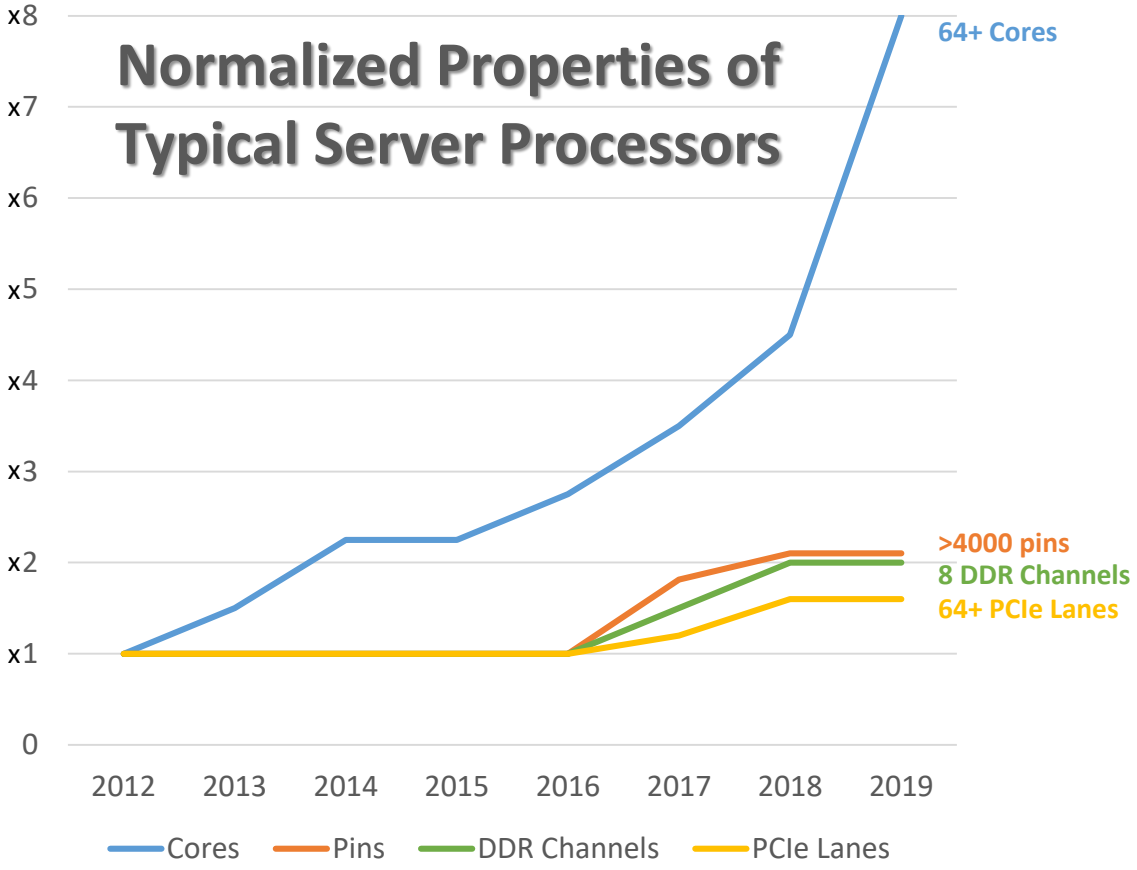
Integrating Gen-Z with RISC-V

Mohan Parthasarathy

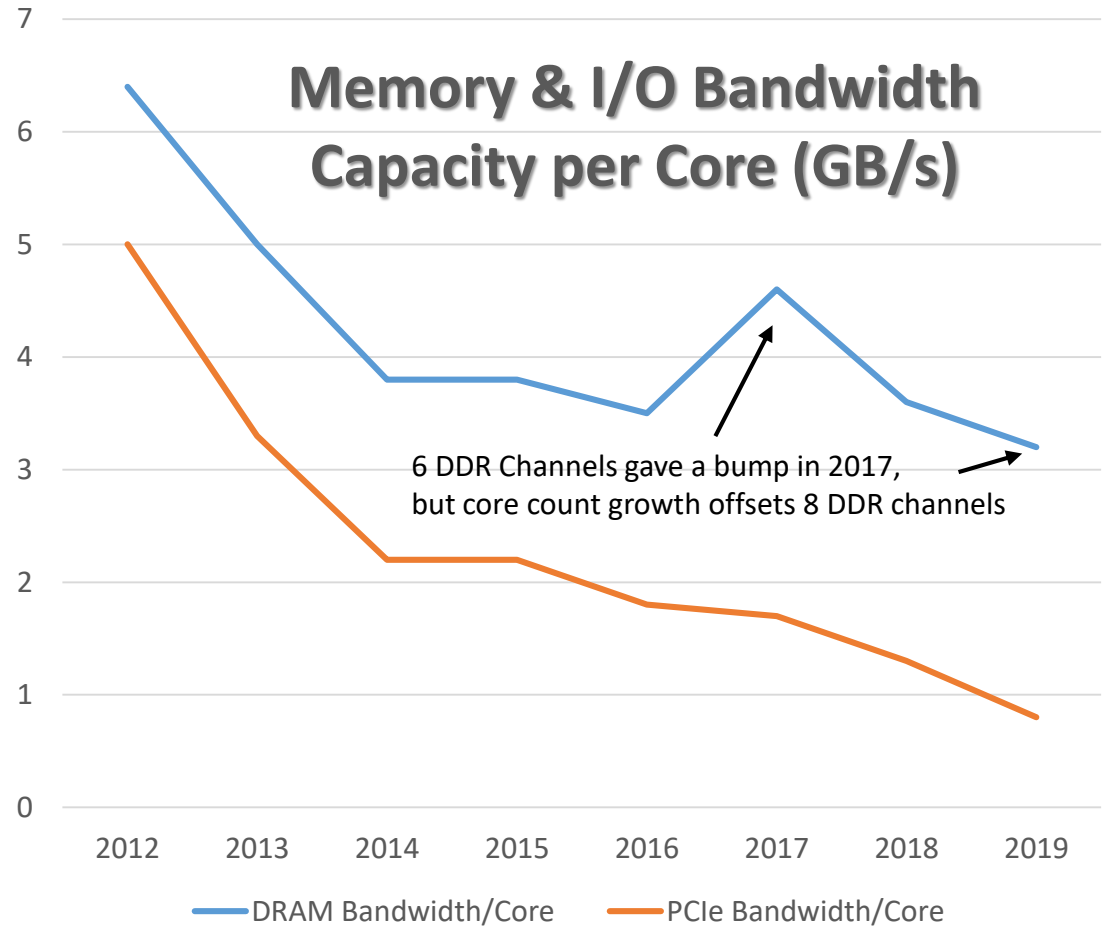
July 19th 2018

The Motivation for a Gen-Z Fabric

Compute-Memory Balance is Degrading



Processor memory and I/O technologies …

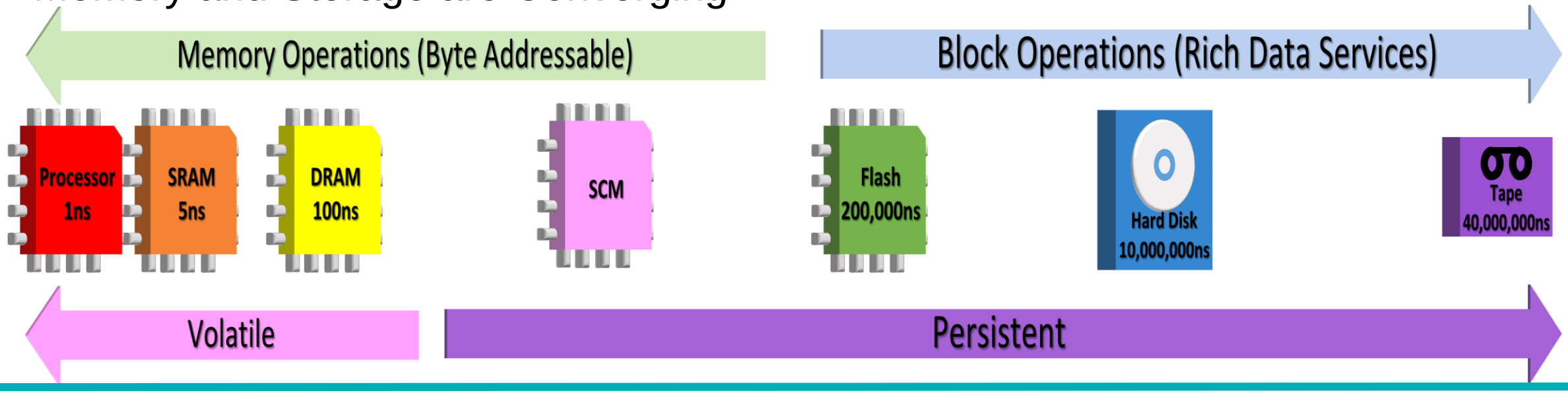


… are being stretched to their limits

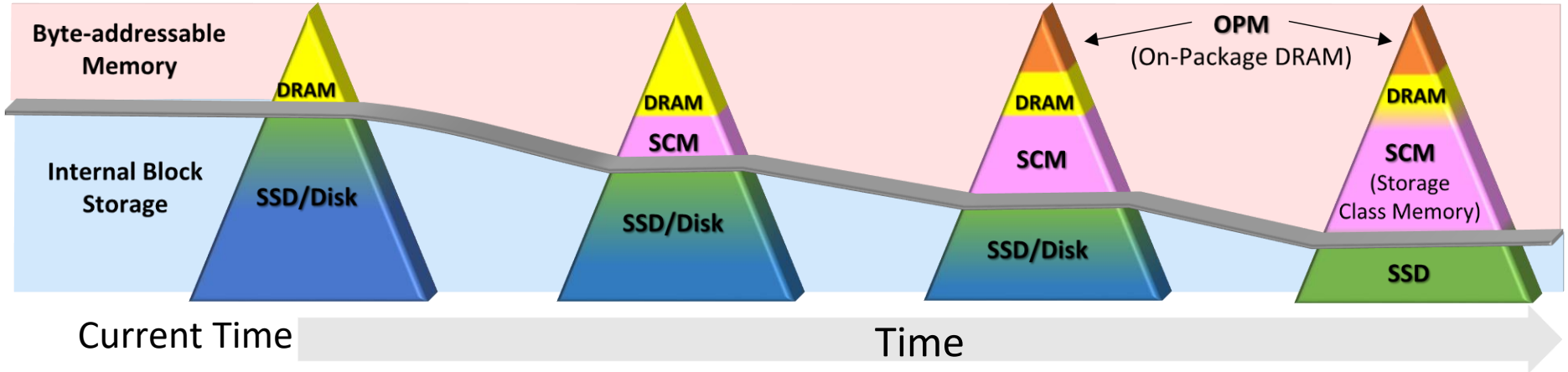
Hewlett Packard Enterprise

The Motivation for a Gen-Z Fabric

Memory and Storage are Converging



With memory/storage convergence, memory semantic operations become predominant (volatile & non-volatile)



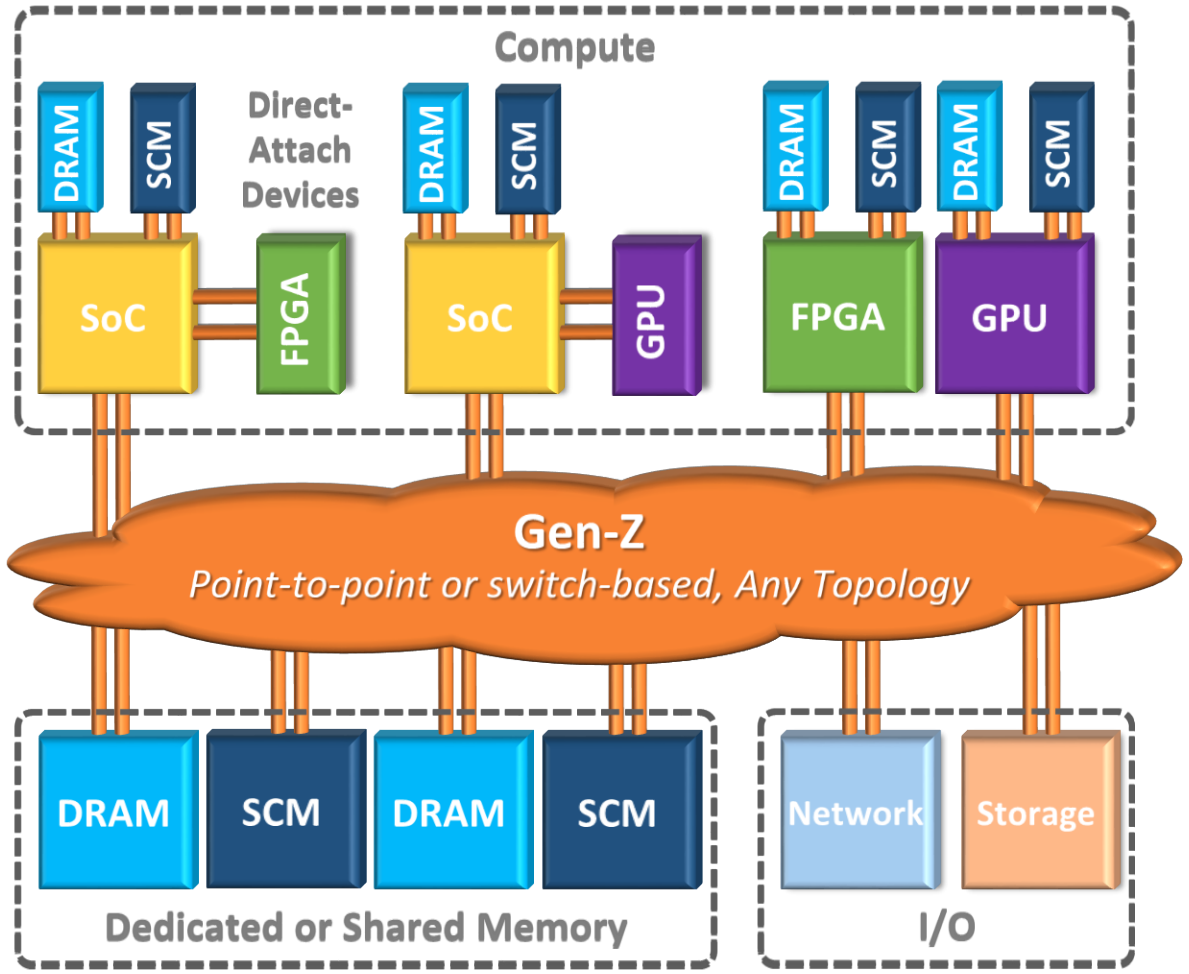
Byte-addressable persistent memory delivered using DRAM + Flash or SCM

Hewlett Packard Enterprise

The Solution : Scalable Memory Fabric – Gen-Z!

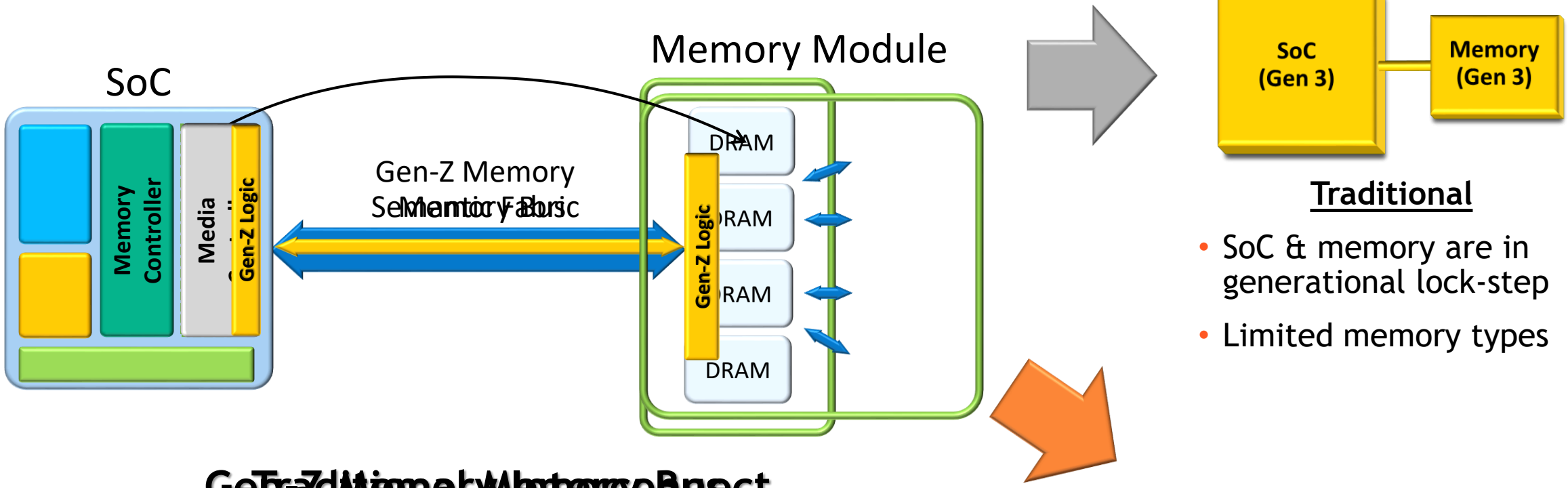
- ***High Performance***
  - Very high bandwidth (16 GT/s to 112 GT/s signaling), low latency
  - Delivers 32 GB/s to 400+ GB/s per memory module
- ***Reliable***
  - Flattens memory / storage hierarchy w/integrated resiliency, multipath, aggregation, etc.
  - No stranded resources or single-point-of-failures
- ***Secure***
  - Provides strong hardware-enforced isolation and security
- ***Flexible***
  - Multiple topologies, component types, etc.
  - Supports legacy and new high-capacity form factors. Multiple media types can be physically co-located.
  - Scales from co-packaged to single motherboard to rack-scale
- ***Compatible***
  - Use existing physical layers, unmodified OS support
- ***Futuristic***
  - Breaks processor-memory interlock enabling innovative solutions.
  - Built from the “ground up” to support persistent memory semantics

Gen-Z speaks the language of compute



Hewlett Packard Enterprise

So What's an Example of How Gen-Z Helps Us?



Traditional

- SoC & memory are in generational lock-step
- Limited memory types

Gen-Z Memory Interconnect

- Media specific logic integrated into memory module
- Tight coupling of SoC and memory technology evolution
- Limits the types of memory that can be supported

Gen-Z + DDR : High Bandwidth

Gen-Z Signaling Rate	Gen-Z	8 DDR 6400 Channels	Aggregate Memory Application Bandwidth
25 GT/s 64 Tx / Rx Lanes	320 GB/s	400 GB/s	720 GB/s
25 GT/s 128 Tx / Rx Lanes	640 GB/s	400 GB/s	1.04 TB/s
32 GT/s 64 Tx / Rx Lanes	400 GB/s	400 GB/s	800 GB/s
32 GT/s 128 Tx / Rx Lanes	800 GB/s	400 GB/s	1.2 TB/s
56 GT/s 64 Tx / Rx Lanes	700 GB/s	400 GB/s	1.1 TB/s
56 GT/s 128 Tx / Rx Lanes	1.4 TB/s	400 GB/s	1.8 TB/s
112 GT/s 64 Tx / Rx Lanes	1.4 TB/s	400 GB/s	1.8 TB/s
112 GT/s 128 Tx / Rx Lanes	2.8 TB/s	400 GB/s	3.2 TB/s

Key Gen-Z attributes for Scale-out Computing

Network addressing

16-bit subnet IDs +
12-bit component IDs +
[64-bit memory address]

A theoretical maximum of 2^{28} (~268M) components

Memory-semantic datagram packets independent of fabric scale

No performance degradation to communicate across subnets

Does not require multiple component IDs to support multipath

Flexible destination and packet relay tables to support nearly any routing topology

Advanced Operations

Multiple buffer put / get variations

Collectives + Collective Acceleration

Signaled writes / Write MSG (send) with Receive Tags and Embedded Read

Virtual Channels

32 VCs

Remove cyclic resource dependencies for routing deadlock avoidance.

Reduce head-of-line blocking and / or cross path blocking.

Segregate traffic classes for performance isolation.

VC remapping to support components with different number of VCs

Packet Injection / Relay

Robust congestion management with automatic packet injection rate

Common source node adaptive / dispersive packet injection and switch adaptive / dispersive packet relay

Traffic Classes

Set of VCs for user-defined purposes

Performance within a TC is not affected by other TCs, e.g., TCs separate:

- Latency Sensitive (e.g., SHMEM)
- Bandwidth Sensitive (e.g., check point)
- Noise Sensitive (e.g., collectives)
- High-priority Applications

Multi-plane Support

All planes can be co-packaged within a single switch

A single cable can be used to connect to all planes

A single interface can be drive all planes

Adaptive / dispersive routing enables load balancing, resiliency, etc.

Low-latency FEC

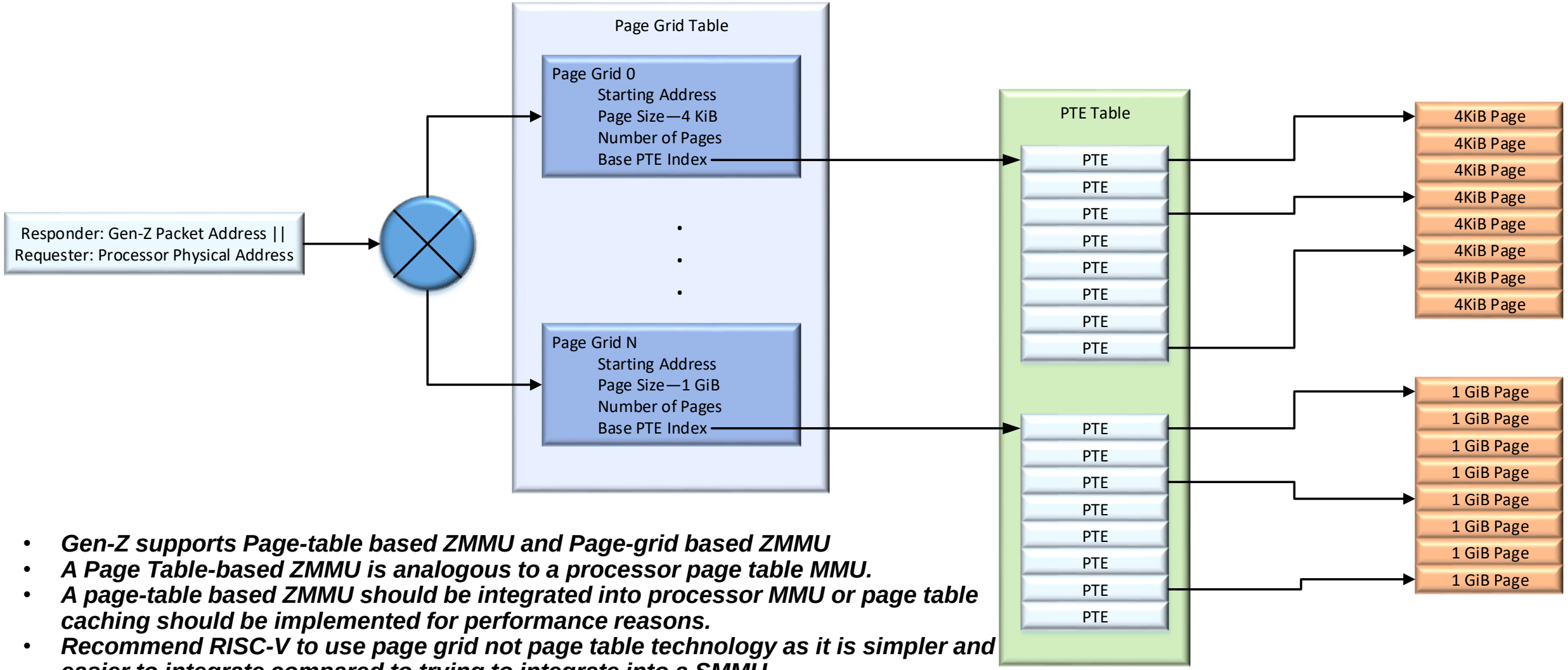
2 ns per link hop

Gen-Z : A RISC-V should take

What RISC-V should do to take advantage of Gen-Z

- Should support Gen-Z in addition to DDR natively to differentiate from processors supporting only DDR natively.
- Should support 52 bit physical addressing.
- Cache Management Challenges (Support for Memory consistency models across fabric)
- Should support a minimum of 64 Gen-Z lanes of the PCIe Phy at 32 GT/s.
- Should support a minimum of 1K outstanding transactions on the coherency interface to support inherent SCM media latencies.
- Memory Mapping using ZMMU for Gen-Z
- LPD Support for PCIe compatibility
- Implement Far Atomics
- Take advantage of Gen-Z Security Features
- Protection/Translation architecture for a secure fabric

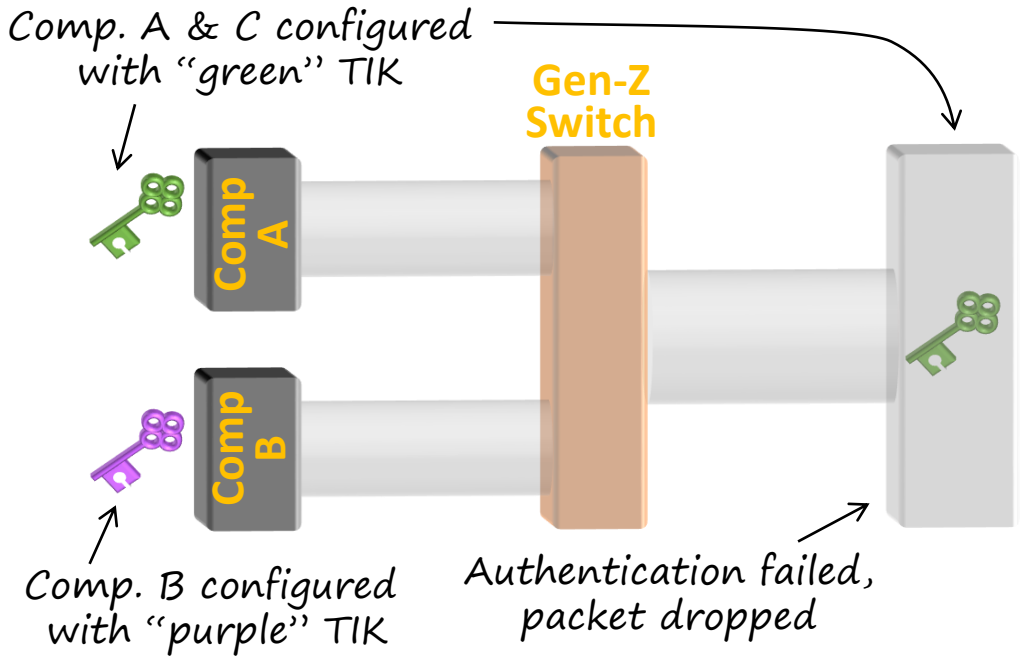
ZMMU Implementation : Page-Grid based ZMMU



- ***Gen-Z supports Page-table based ZMMU and Page-grid based ZMMU***
- ***A Page Table-based ZMMU is analogous to a processor page table MMU.***
- ***A page-table based ZMMU should be integrated into processor MMU or page table caching should be implemented for performance reasons.***
- ***Recommend RISC-V to use page grid not page table technology as it is simpler and easier to integrate compared to trying to integrate into a SMMU***

Gen-Z : Trust and Security

- Multiple Access Control Techniques including:
  - Access Keys (component group level access)
  - Access Request and Access Response (fine-grain component-level access)
  - R-Keys (page-level access control)
  - R-Key Domains (Requester R-Key filtering)
  - Switch Packet Filtering (control plane, leaf component)
  - Peer Component Authorization (whitelist)
  - Peer Nonce to detect rogue component insertion while in low-power state
- Packet authentication with Anti-replay Tags
  - Hashed Message Authentication Code (HMAC)
  - Uses transaction integrity key (TIK)
    - Only devices sharing TIK can communicate
- Packets dropped due to authentication violations
  - Configurable interface / component containment
- Violations reported to management
- Currently, developing new component authentication, page-level encryption, and data authentication



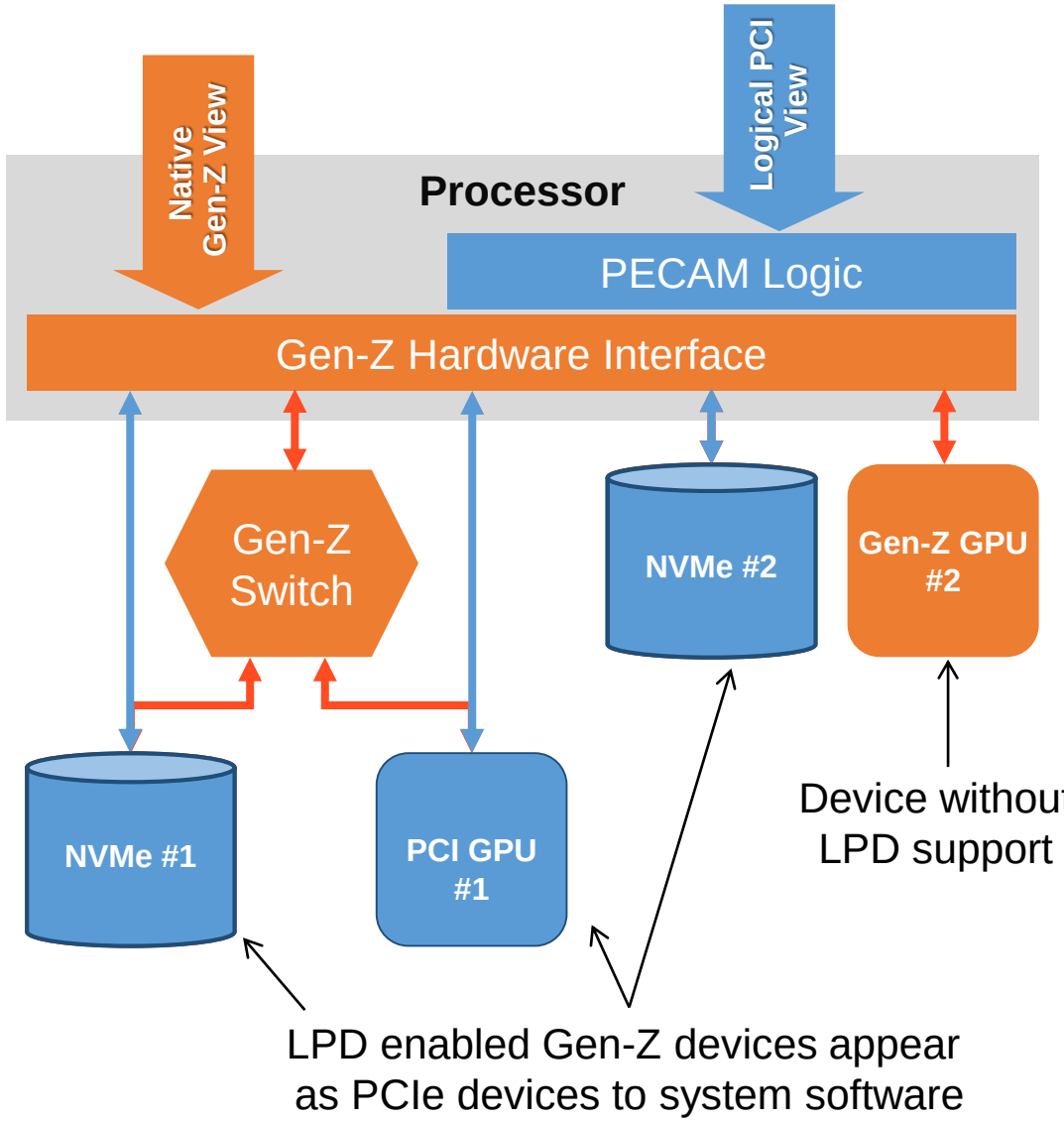
- ***RISC-V could add capabilities support to augment Gen-Z security features.***
- ***Recommend RISC-V add a trust zone like capability to build on what Gen-Z enables to provide more page level security (Who configures the ZMMU – do you trust the node?)***

Hewlett Packard Enterprise

Gen-Z : I/O with PCIe compatibility

Gen-Z Logical PCI Devices (LPDs)

- Gen-Z devices can be discovered/configured
  - Via standard PCI / PCIe system software
- LPDs can fully exploit Gen-Z Architecture
  - Low-latency switching
    - Gen-Z 30 ns vs. PCIe 130-150 ns translates to 200-240 ns savings per read operation
  - Memory-speed CPU-to-device communication
  - Security and fine-grain hardware-enforced isolation (any-to-any communication without compromise)
  - Supports all x86 / ARM / Power architecture Atomics
  - Simplified single and multi-host I/O virtualization and sharing capabilities
  - Multipath—aggregation / resiliency / robust topologies
  - PCIe 2.5-32 GT/s PHY and 25-112 GT/s 802.3 Electrical
  - Legacy plus New Gen-Z Scalable Connector and Scalable Form Factors
  - CPU-based data movers to enable new software paradigms
  - Scale-up and scale-out connectivity and performance
  - Simplified software—any mix coherent and non-coherent operations
  - And much more…



- ***RISC-V should support a PECAM through the Requestor ZMMU for LPD support***

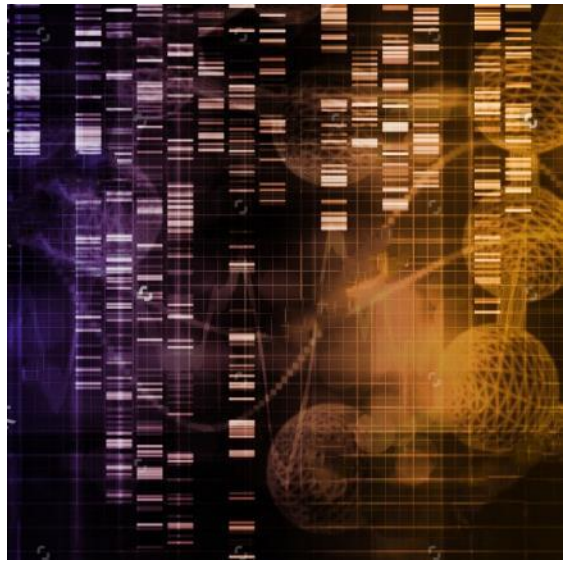
Hewlett Packard Enterprise

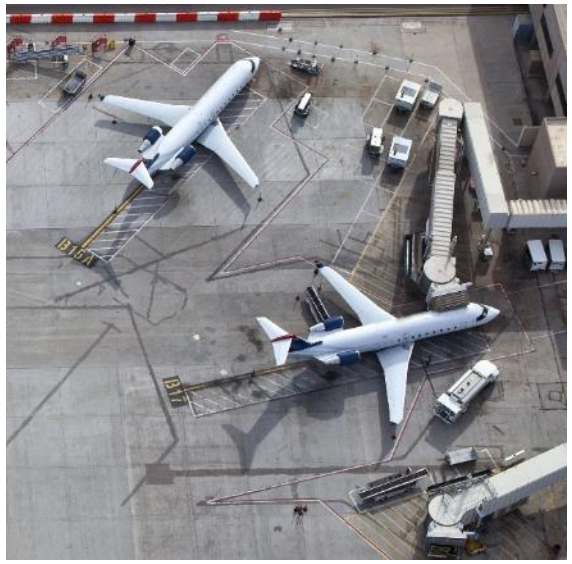
Gen-Z Transforms Performance

Modify existing frameworks

New algorithms

Completely rethink





In-memory analytics — **Similarity search** — **Large-scale graph inference** — **Financial models**

15x faster

20x faster

100x faster

8,000x faster

Hewlett Packard Enterprise

Gen-Z : Broad Industry and Component Support

GEN Z Consortium Members

Alpha Data	Jess-Link	Smart Modular
AMD	Keysight	Spin Transfer
Amphenol	Lenovo	Technologies
ARM	Lots	TE Connectivity
Avery Design	LUXSHARE-ICT	Teledyne
Broadcom	Mellanox	LeCroy
Cadence	Micron	Toshiba Memory
Cavium	Microsemi	Tyco Electronics
Cisco	Mobiveil	UNH
Cray	Molex	VMware
Dell EMC	NetApp	WDC
ETRI	Nokia	Xilinx
(Research)	Oak Ridge National	YADRO
Everspin	Labs	Yonsei
Foxconn	PLDA	University
Interconnect	Qualcomm	
Google	Red Hat	
HPR	Samsung	
Hirose Electric	Seagate	
Huawei	Senko	
IBM	Simula Research	
IDT	SK hynix	
IntelliProp		

Components

Intellectual Property

Connectors

Subsystems

Systems

Software

GEN Z Component Categories

CPUs

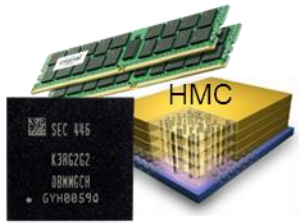

GPUs

Special Purpose Processors

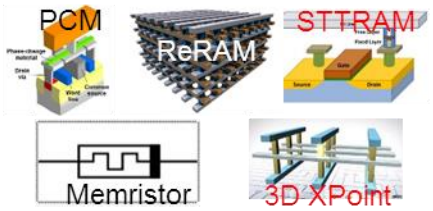
Storage Controllers

Switches/ Gateways

FPGAs



Volatile Memory



Storage Class Memory

Hewlett Packard Enterprise

Gen-Z Consortium Milestones

- Significant milestones over the past year
  - Multi-vendor Proof-of-Concept Demonstrated (FMS'17 / SC'17)
    - New demonstrations at HPE Discover / FMS'18 / SC'18
  - Multiple draft and final specifications publicly available (core architecture, multiple mechanicals, PHY, scalable connector including new high-power and cabling, etc.)
  - 40+ tutorials publicly available, YouTube channel, etc.
  - Expanded membership (including academic & government agencies)
- Key Upcoming Objectives
  - Expand Gen-Z security to support component authentication and page-level data encryption / authenticated
  - Deliver design guides covering:
    - DRAM / SCM, LPD, Storage, eNIC, and high-speed messaging
  - Complete ZSFF and PECFF mechanical form factor specifications
    - PECFF (September) and ZSFF (4Q2018)
  - Release Gen-Z PHY specification with support:
    - PCIe 16 GT/s and 802.3 electrical 25 GT/s (September)
    - 802.3 electrical 56 GT/s PAM 4 (4Q2018)
    - PCIe 32 GT/s and 802.3 electrical 112 GT/s PAM 4 (2Q2019)
  - Develop compliance testing

Hewlett Packard Enterprise

RISC-V : Building on Gen-Z Basics

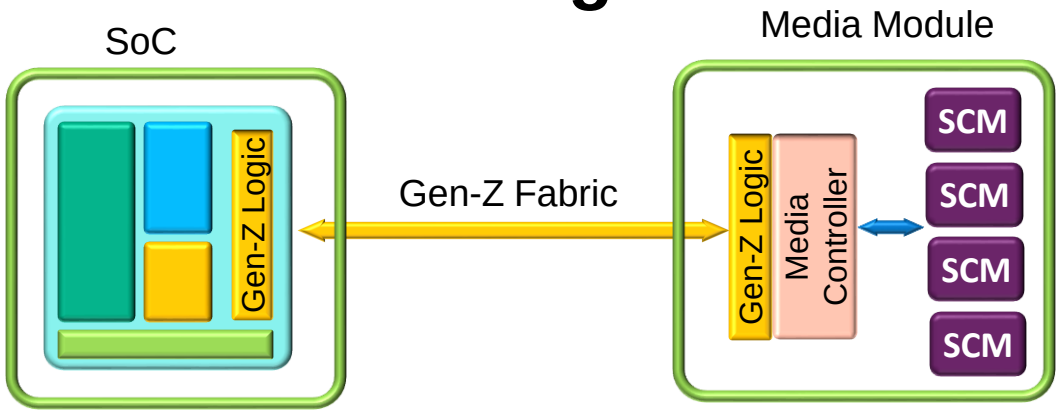


Figure 1 – Storage Class Memory

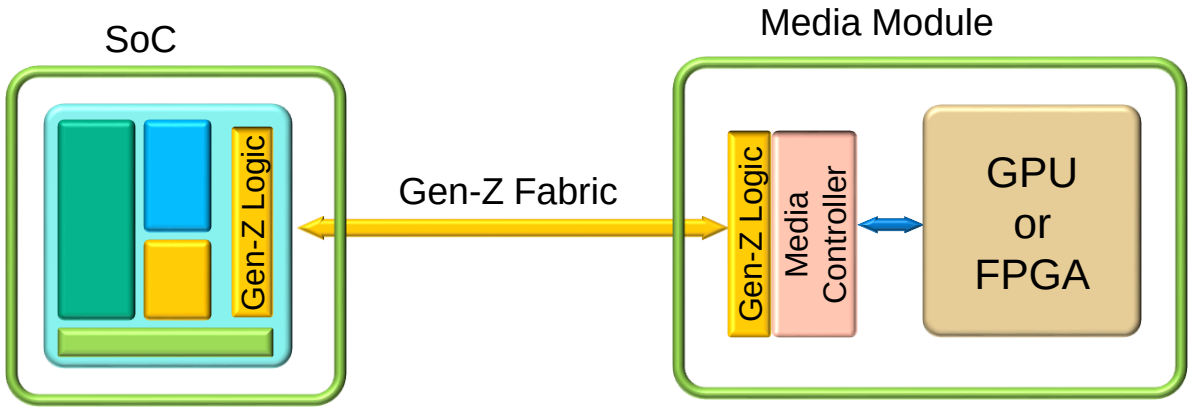


Figure 2 – GPU or FPGA

- Supports DRAM, Flash, Memristor, PCRAM, MRAM, 3D-Xpoint…**Universal Interconnect**
- Decouples CPU/memory design
- Enables independent innovation

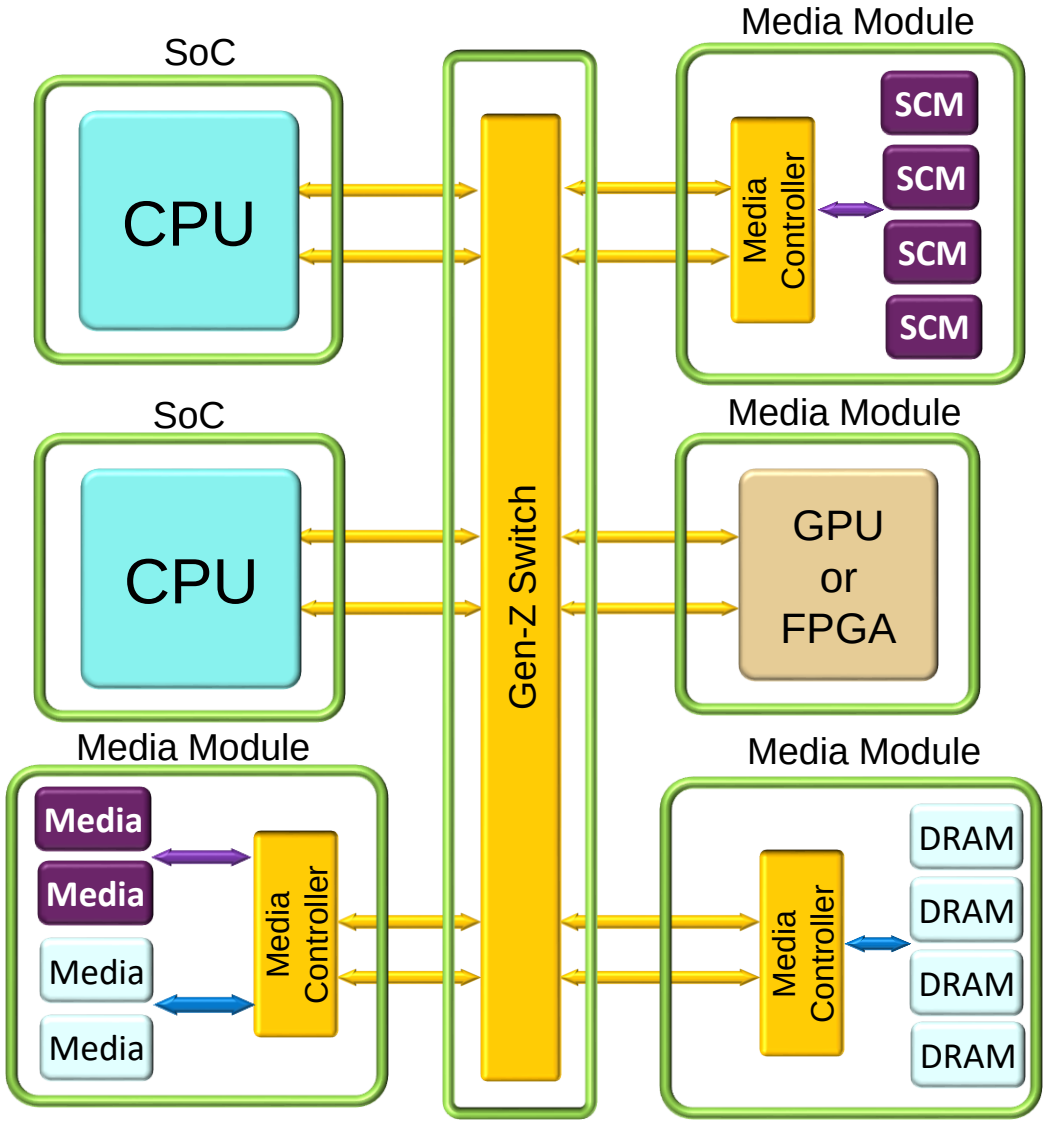


Figure 3 – Multiple resources enabled by Universal Interconnect

Hewlett Packard Enterprise

Hewlett Packard
Enterprise

Thank you

Contact information