OCP SUMMIT

March 20-21
2018
San Jose, CA

OPEN. FOR BUSINESS.

GEN-Z: HIGH-PERFORMANCE INTERCONNECT FOR THE DATA-CENTRIC FUTURE

Greg Casey/Senior Server Architect and Strategist/
Dell/EMC – Office of the CTO

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.

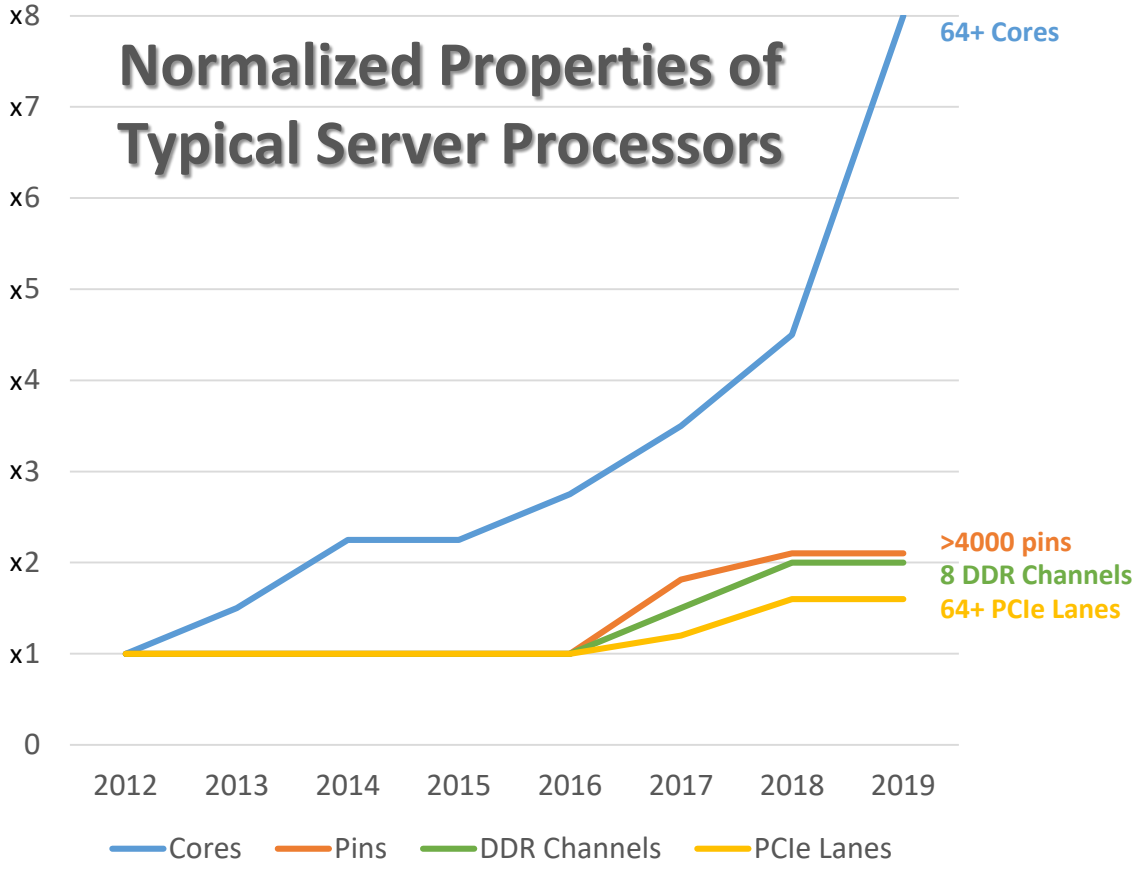
All material is subject to change at any time at the discretion of the Gen-Z Consortium

http://genzconsortium.org/

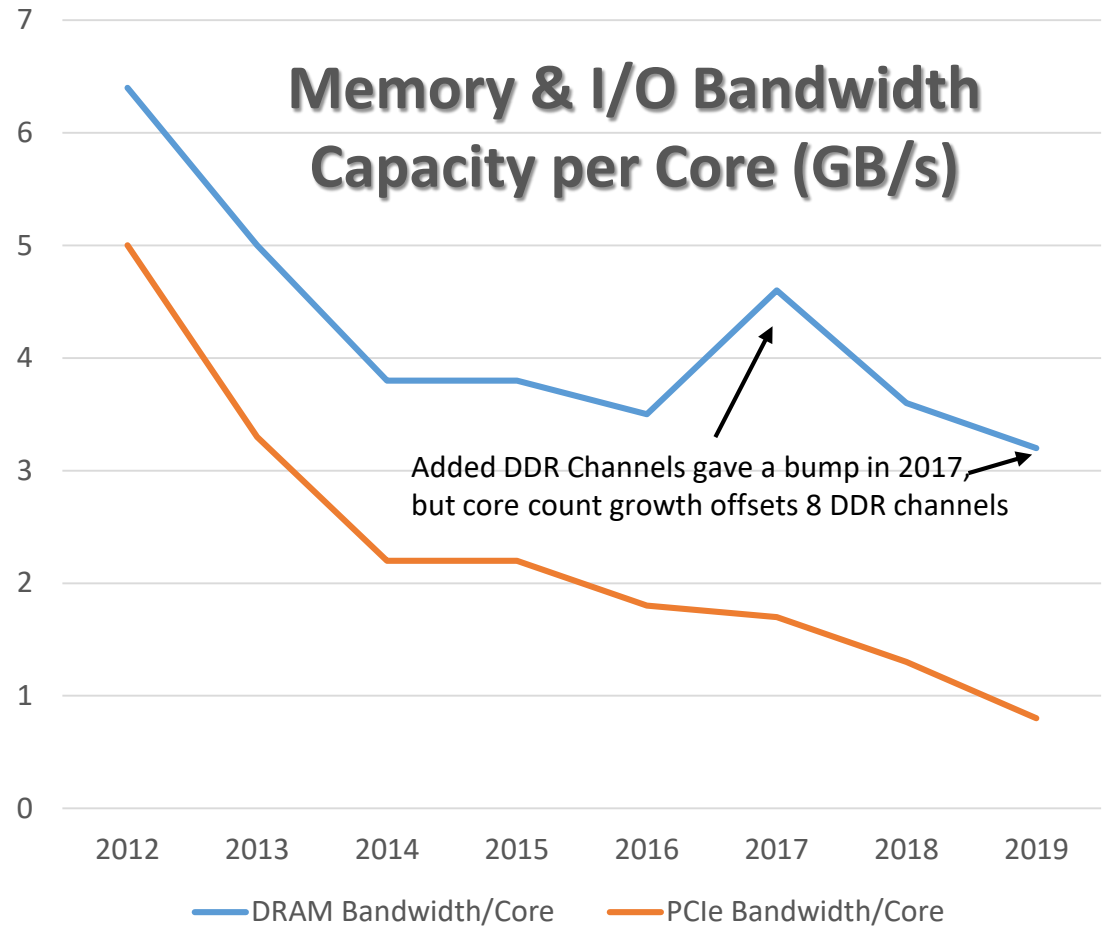
Server Customer Challenges

- Data Explosion
  - 50B+ IoT devices (2020)
  - 180ZB annually (2025)
- Digital Transformation
  - To be competitive, companies have to monetize their data
  - BI, ML, DL, AI Business Intelligence, Machine learning, Deep Learning, Artificial Intelligence: All ways to analyze the large amount of data that will be available and get to information that provide insights into decisions that need to be made. Companies that can get to these insights quickest will have the advantage.
  - Customers have more data and require more processing on that data
- Data Center Efficiency and Agility
  - Customers see the benefit of the promise of true composable via disaggregation infrastructure
  - The composable via disaggregation vision resonates with customers and they are asking for it

Compute-Memory Balance is Degrading

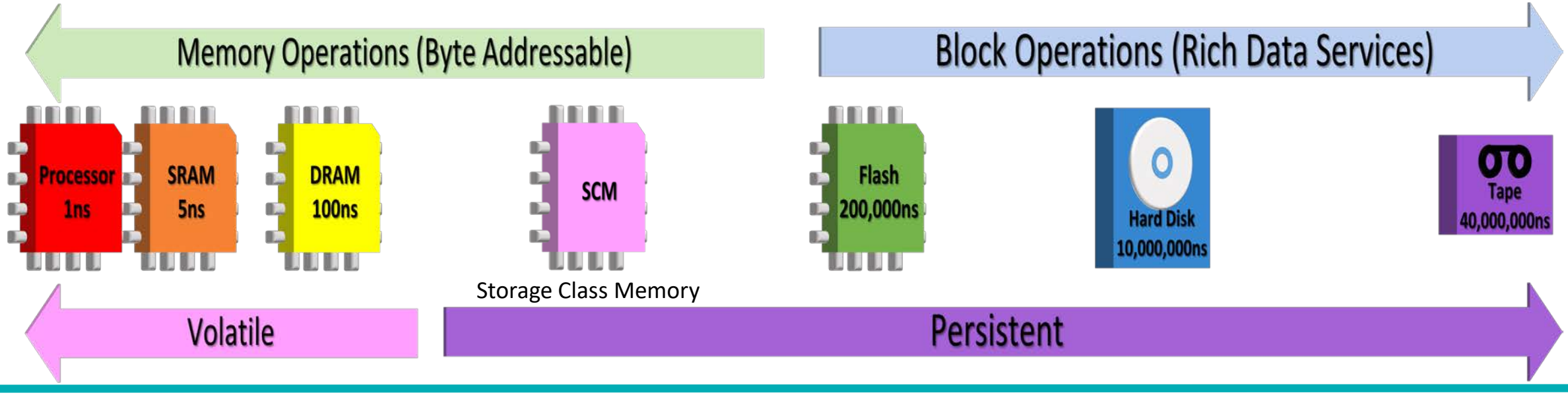


Processor memory and I/O technologies …

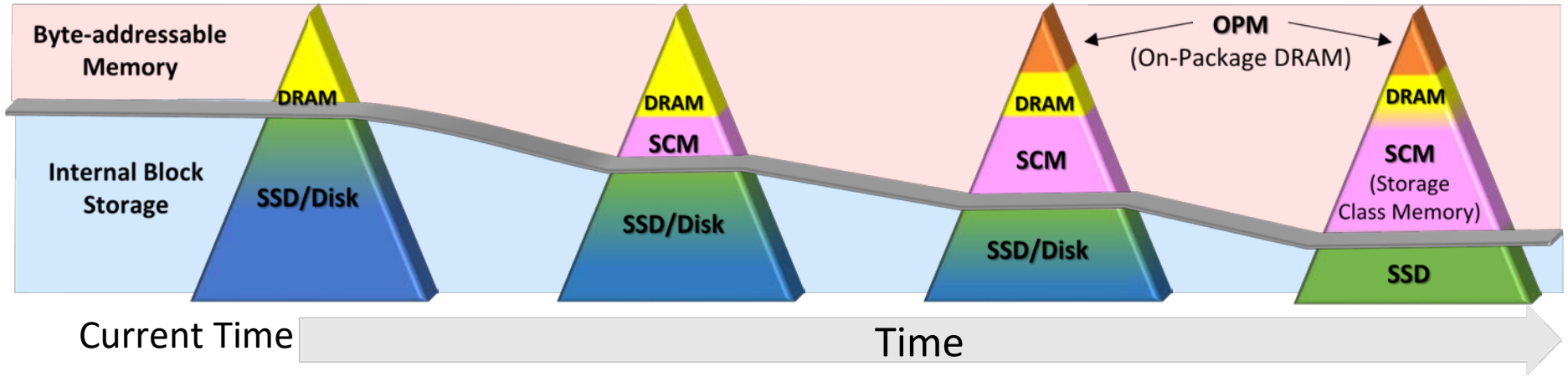


… are being stretched to their limits

Memory and Storage are Converging



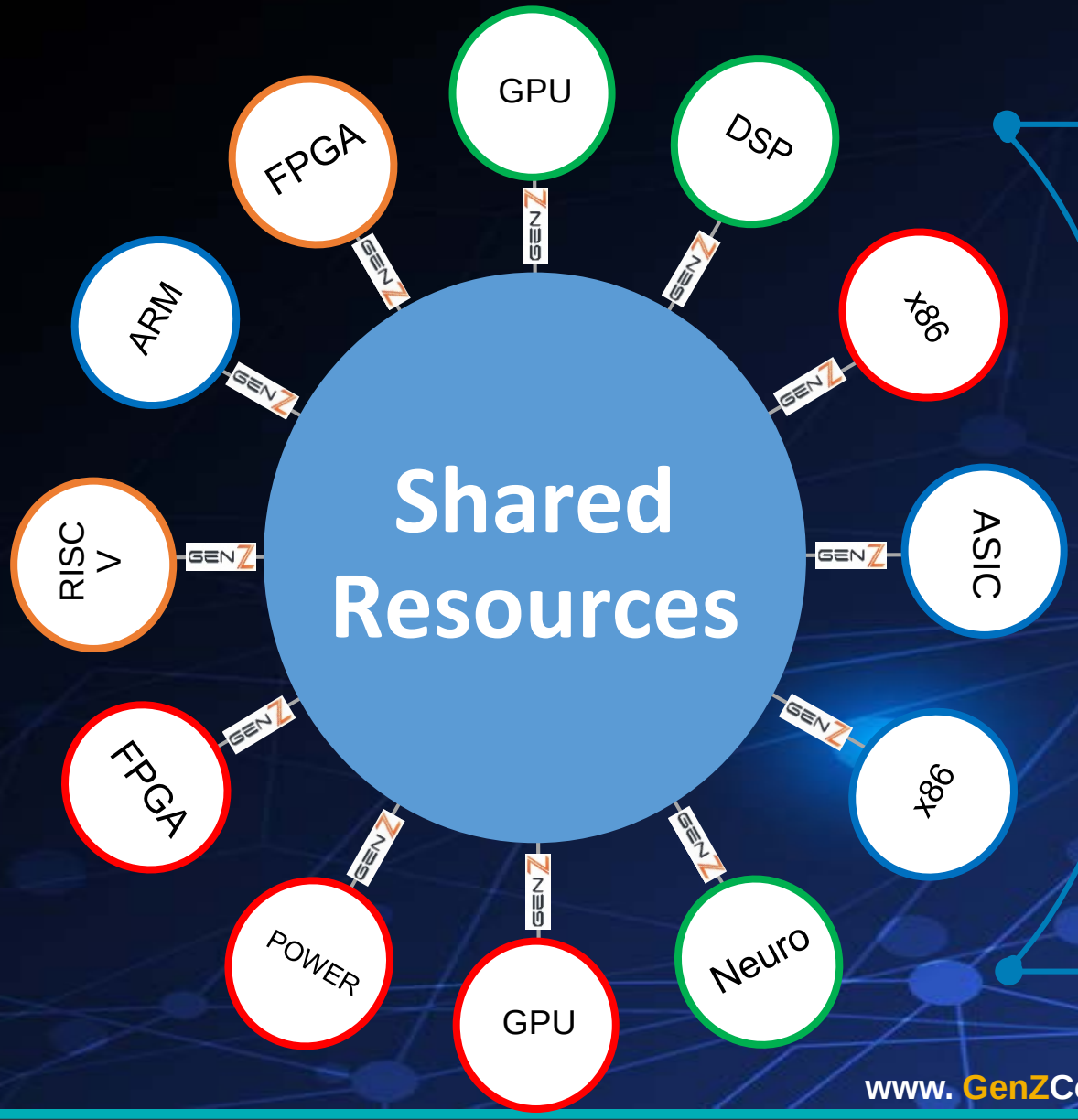
With memory/storage convergence, memory semantic operations become predominant (volatile & non-volatile)



A Memory Fabric is the Missing Piece

- Addresses the Rise of Microsecond Devices
  - Block IO paradigm adds to much SW latency, overshadowing the performance of SCM
  - New sub-microsecond devices cannot be software defined nor be accessed through a SW protocol stack
  - Memory semantics eliminates SW overhead while providing byte addressability
  - A memory fabric allows the addition of SCM without taking valuable memory channels from DRAM
- Enables a Memory Centric Architecture
  - Allows Accelerators, GPUs and FPGAs to directly access SCM
  - Supports copying of data between memory domains (node-to-node or CPU-to-accelerator), improving performance while reducing power consumption
  - Facilitates shared memory programming
- Required for Full Composability
  - Enables disaggregation of Accelerators, GPUs and FPGAs
  - Enables disaggregation of SCM
    - Customers want the benefit of SCM but don’t want it stranded when a server fails

Gen – Z: A new data access technology



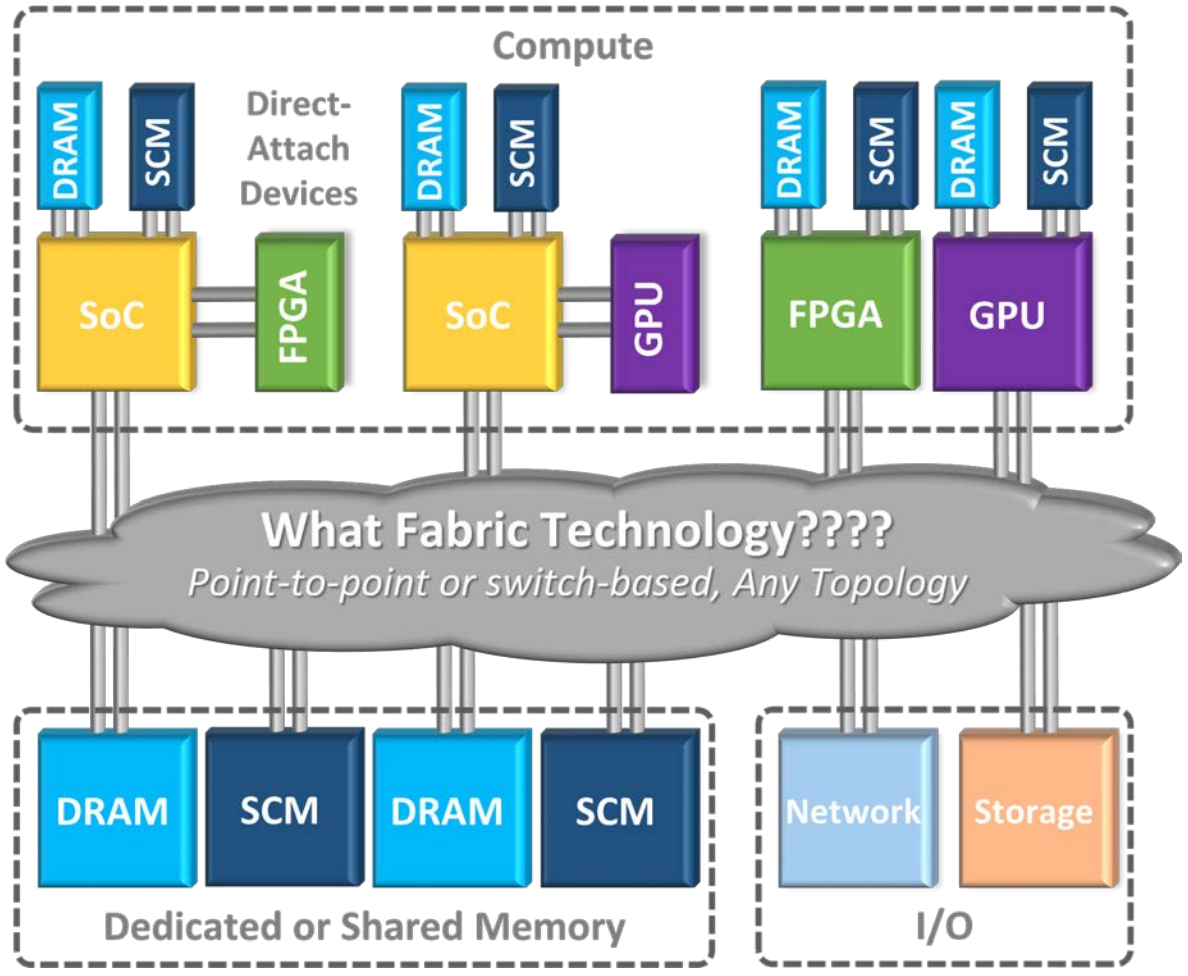
- MEDIA / COMPUTE AGNOSTIC
- INCREASED BANDWIDTH
- LOW LATENCY
- OPEN INNOVATION
- HIGH AVAILABILITY
- SECURITY

www. GenZConsortium.org

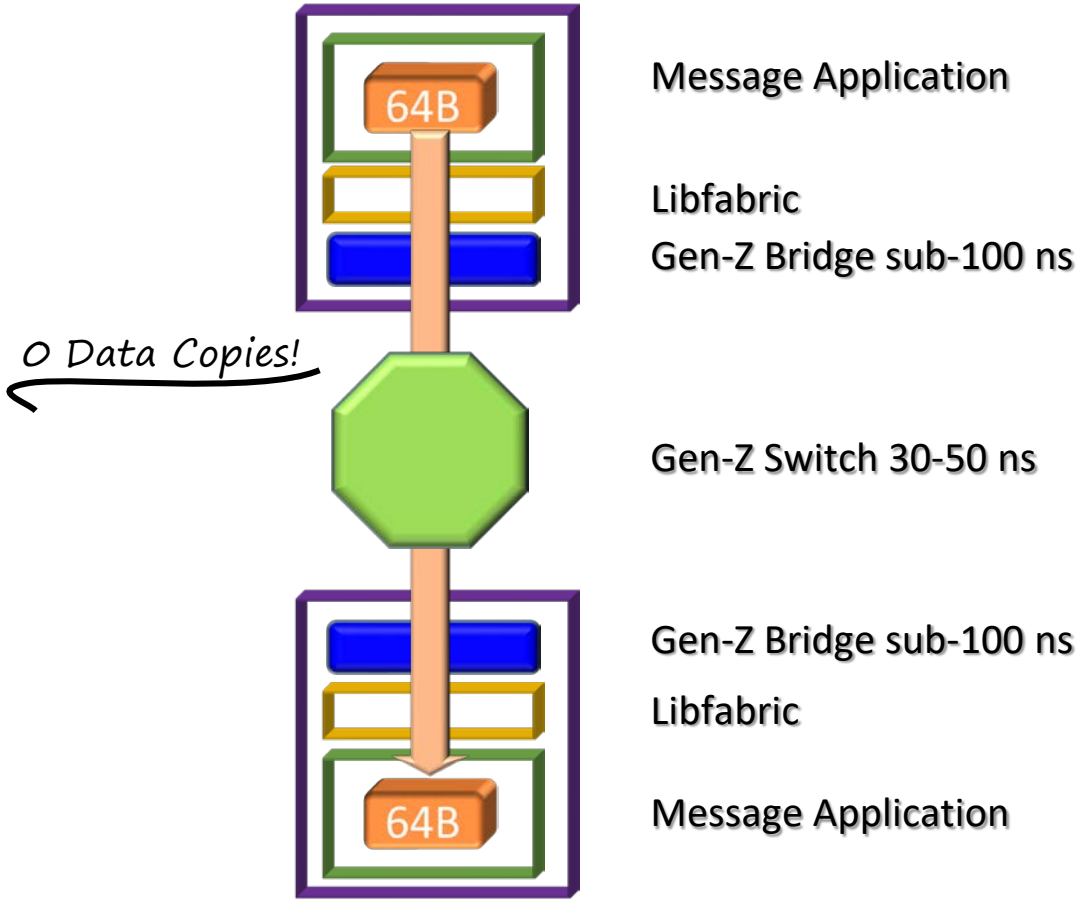
What's the Solution? *Gen-Z!*

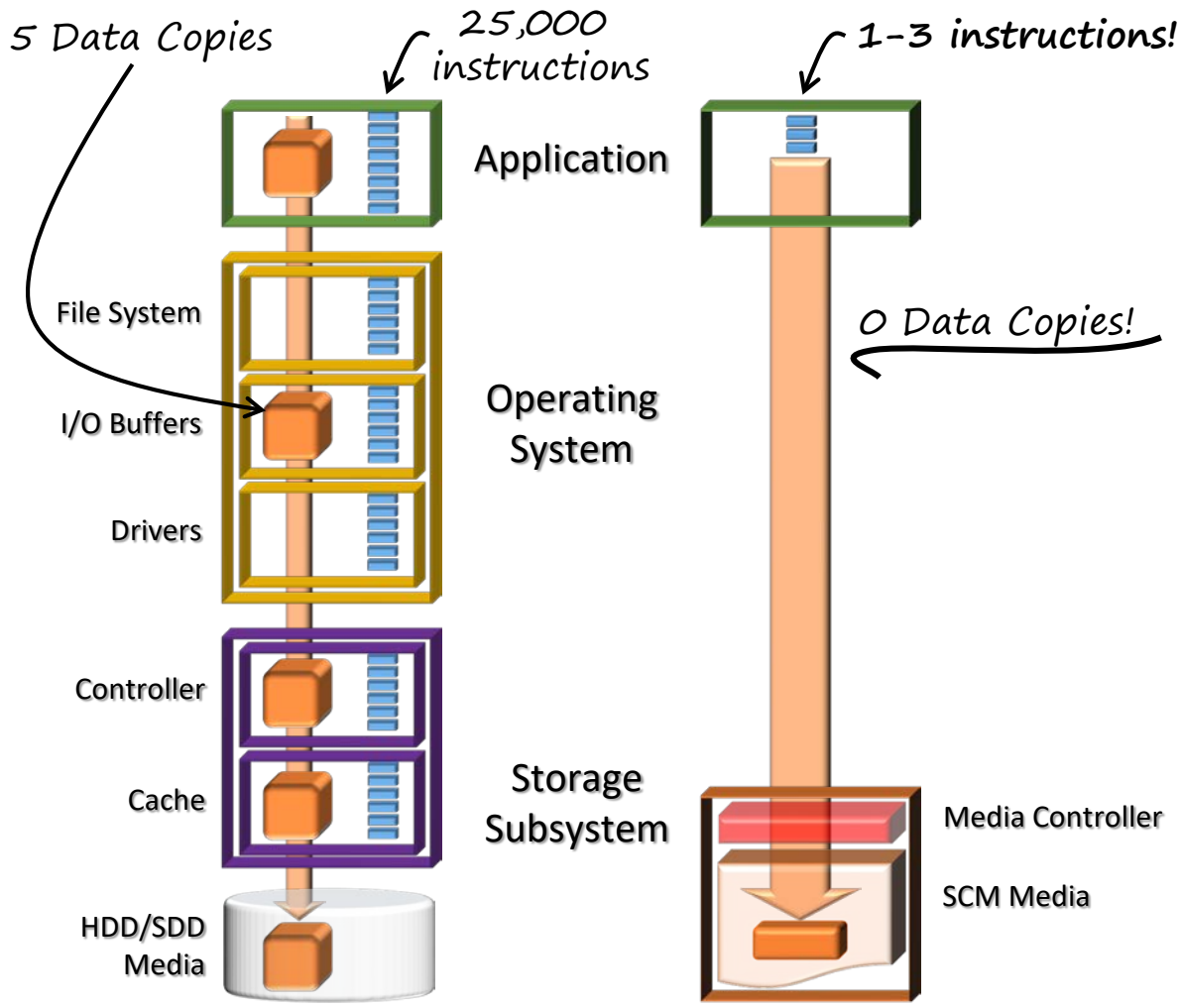
- High Performance
  - High Bandwidth, Low Latency, Scalable
  - Eliminates protocol translation cost / complexity / latency
  - Eliminates software complexity / overhead / latency
- Reliable
  - No stranded resources or single-point-of-failures
  - Transparently bypass path and component failure
  - Enables highly-resilient data (e.g., RAID / erasure codes)
- Secure
  - Provides strong hardware-enforced isolation and security
- Flexible
  - Multiple topologies, component types, etc.
  - Supports multiple use cases using simple to robust designs
  - Thorough yet easily extensible architecture
- Compatible
  - Use existing physical layers, unmodified OS support
- Economic
  - Lowers CAPEX / OPEX, unlocks / accelerates innovation

Gen-Z speaks the language of compute



High Performance: *Latency*





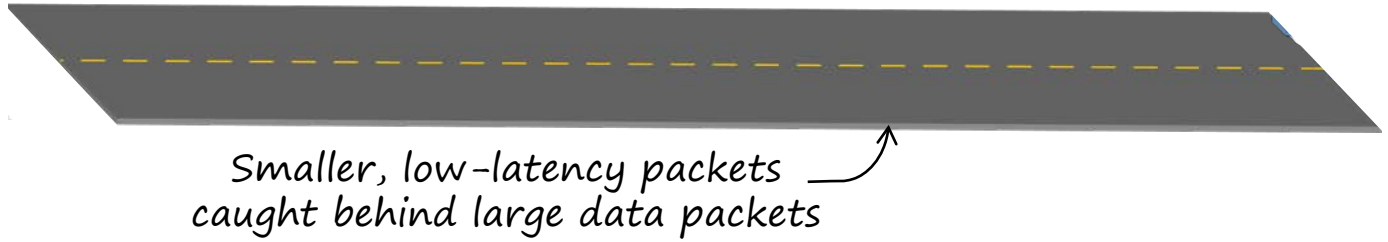
- Ultra-low-latency messaging (Sub-250 ns one-way latency)
- Load/Store byte-addressable access to DRAM or SCM
  - Sub-100 ns load-to-use latency (DRAM media)
- Reduced CPU utilization = ***More workloads per core per CPU***

High Performance: *Throughput*

Traditional Networks

- Modest traffic prioritization
- Modest multi-path support
- Challenging to mix bulk data and low latency

Traditional Network: 2 Lane Road

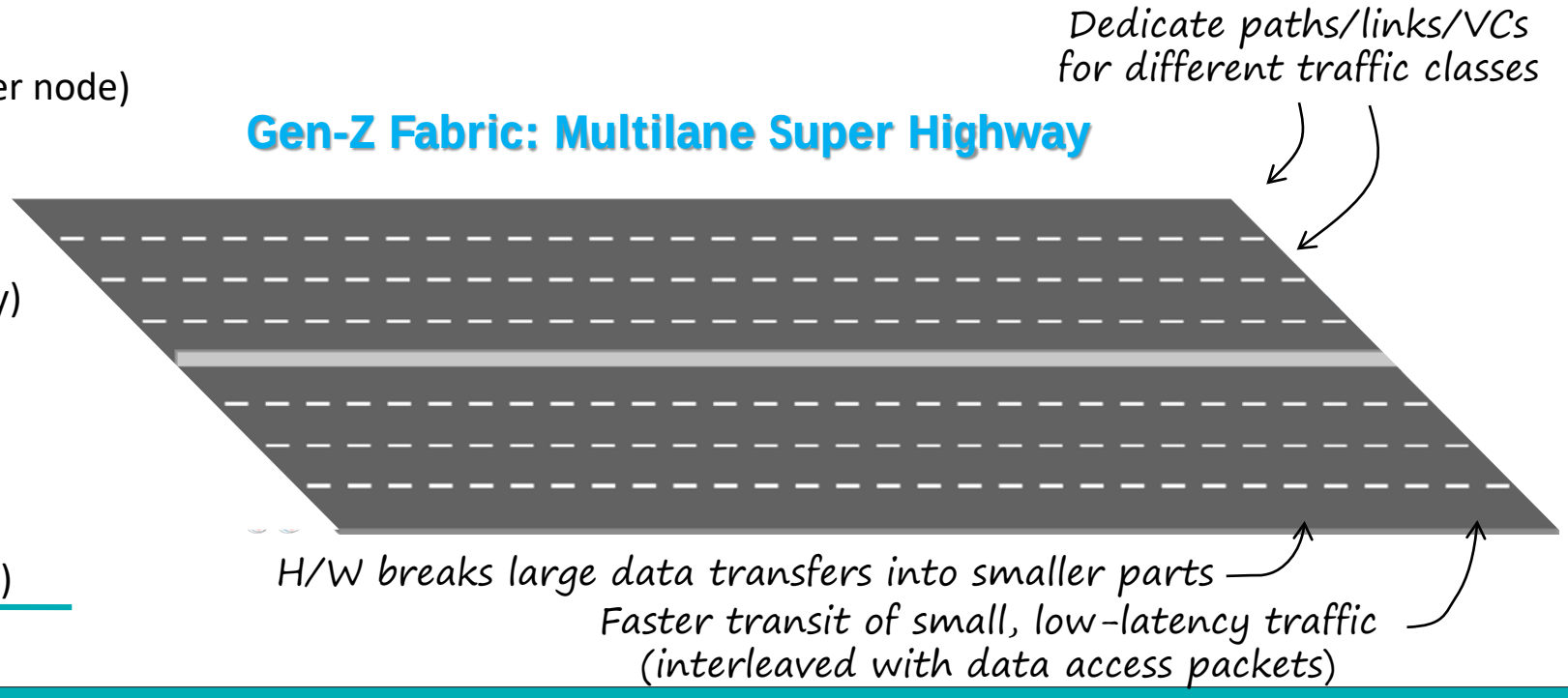


Gen-Z

+ Hardware-driven Multi-Link (e.g., 8 links per node)
+ Hardware-driven Multi-Path
+ 32 Virtual Channels (VCs) per link
+ 256 byte max payload size (90+% efficiency)
+ 30-50 ns switch latency
+ Adaptive / Dispersive routing
+ Dynamic congestion management
+ Universal protocol (no protocol translation)

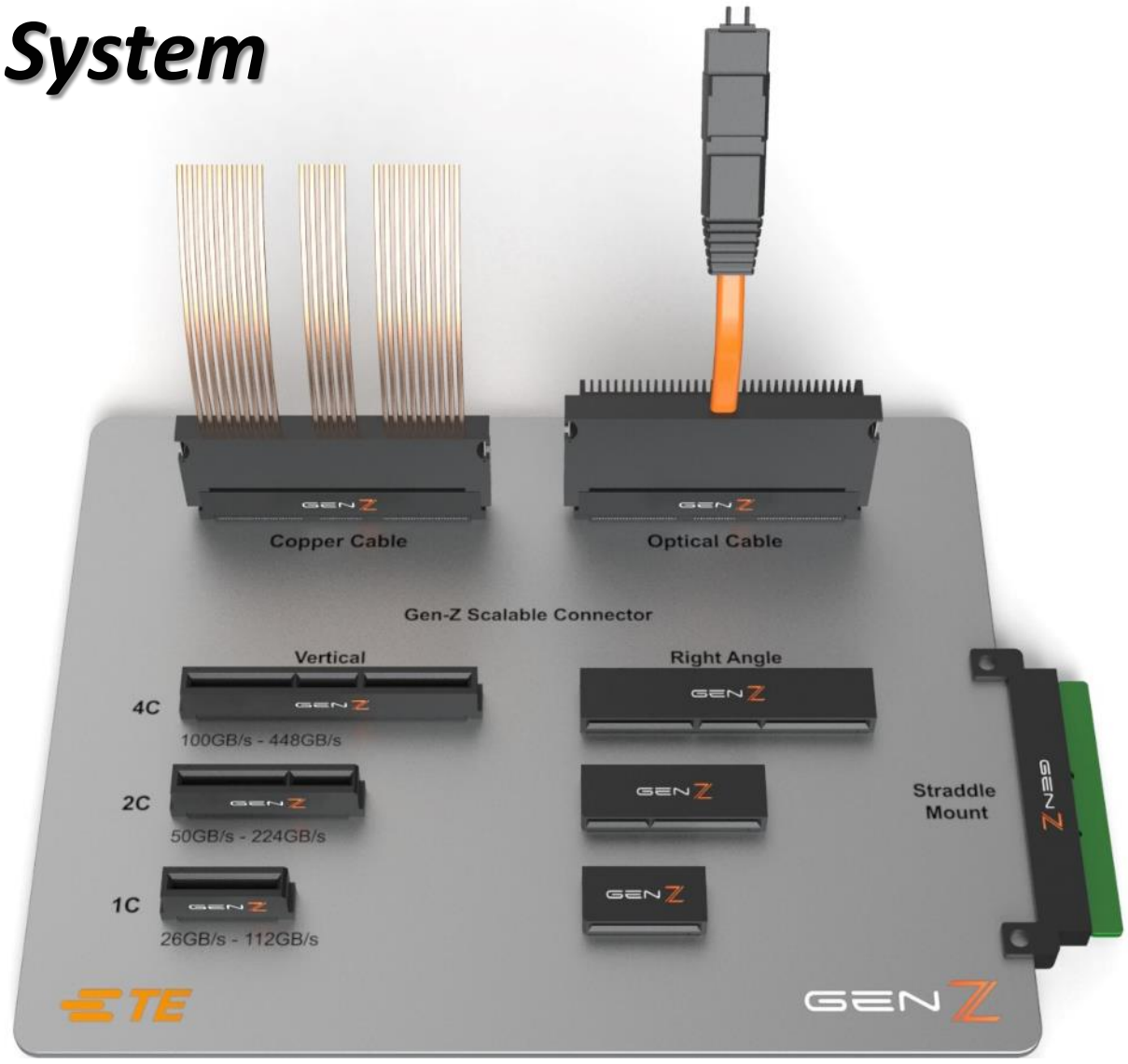
= Industry leading performance

Gen-Z Fabric: Multilane Super Highway



Flexible: *Universal Connector System*

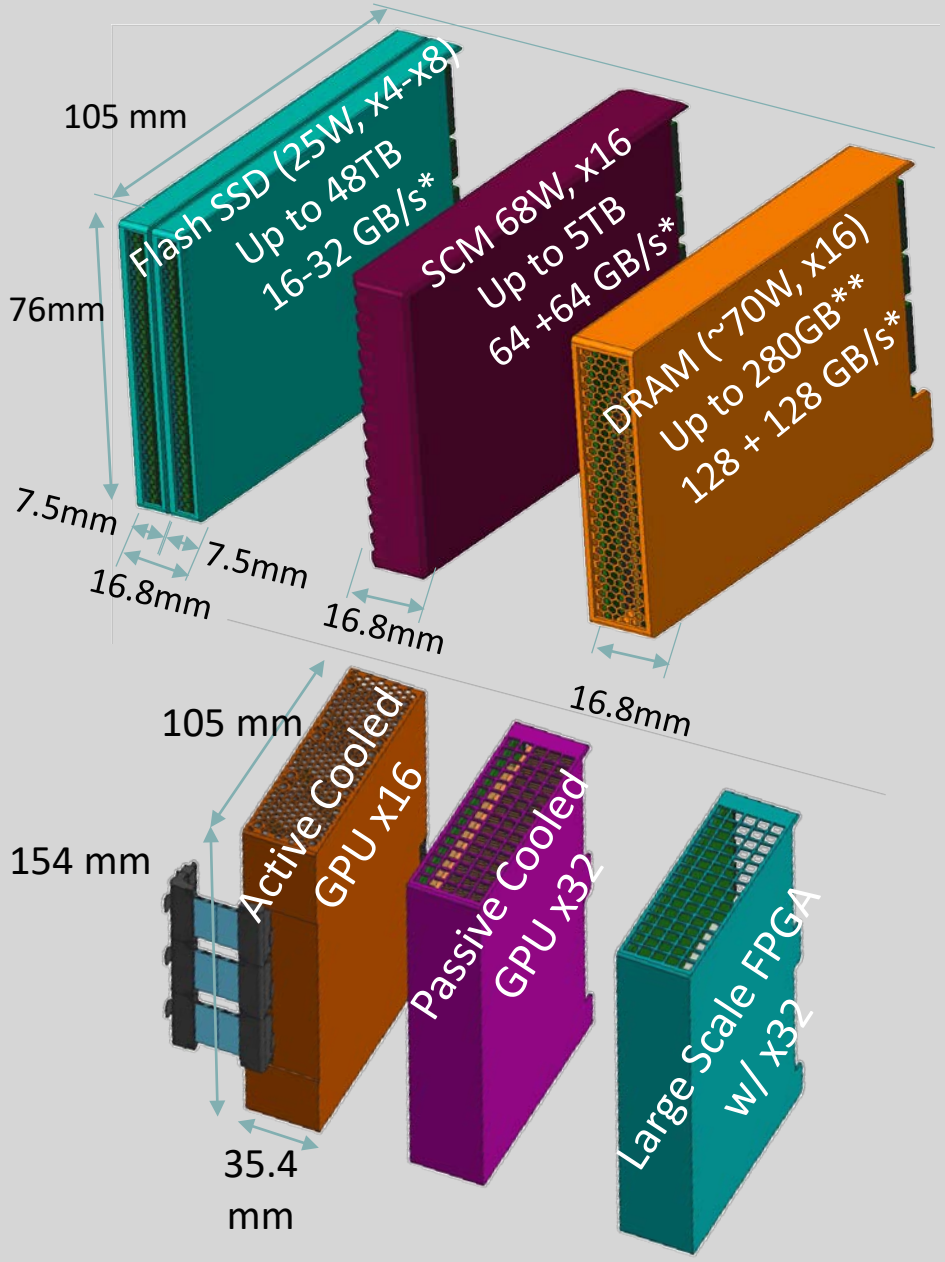
- Vertical, horizontal, right angle, straddle mount
- Same connectors for memory, I/O, storage, etc.
- Cabled solutions: for copper & optical
- Unified connector and compatible pinout across storage, memory and I/O applications in a server
- Eliminates “hard choices”
  - Universal connector eliminates industry fragmentation
    - Simplifies supply chain—drives volume and lowers cost
  - Any component, any slot, any time
    - Any mix of static and hot-plug form factors
  - Multi-connector option to provide added scalability
    - 80W incremental power
    - Incremental bandwidth
  - Supports internal and external cable applications
    - Enables modular system design
    - Enables system disaggregation
    - Eliminates expensive board materials
  - Multipath—can bifurcate connector into multiple links
    - Aggregate bandwidth, resiliency, no stranded resources
    - Support multiple topologies—point-to-point, daisy-chain, mesh
  - Supports multiple interconnect technologies—Gen-Z, PCIe, etc.
- OCP NIC 3.0 spec is using the SFF-TA-1002 Spec



Gen-Z members contributed mechanical & electrical specification to SNIA—see SFF-TA-1002
Gen-Z Scalable Connector specification (final version is publicly available) covers remaining functionality.

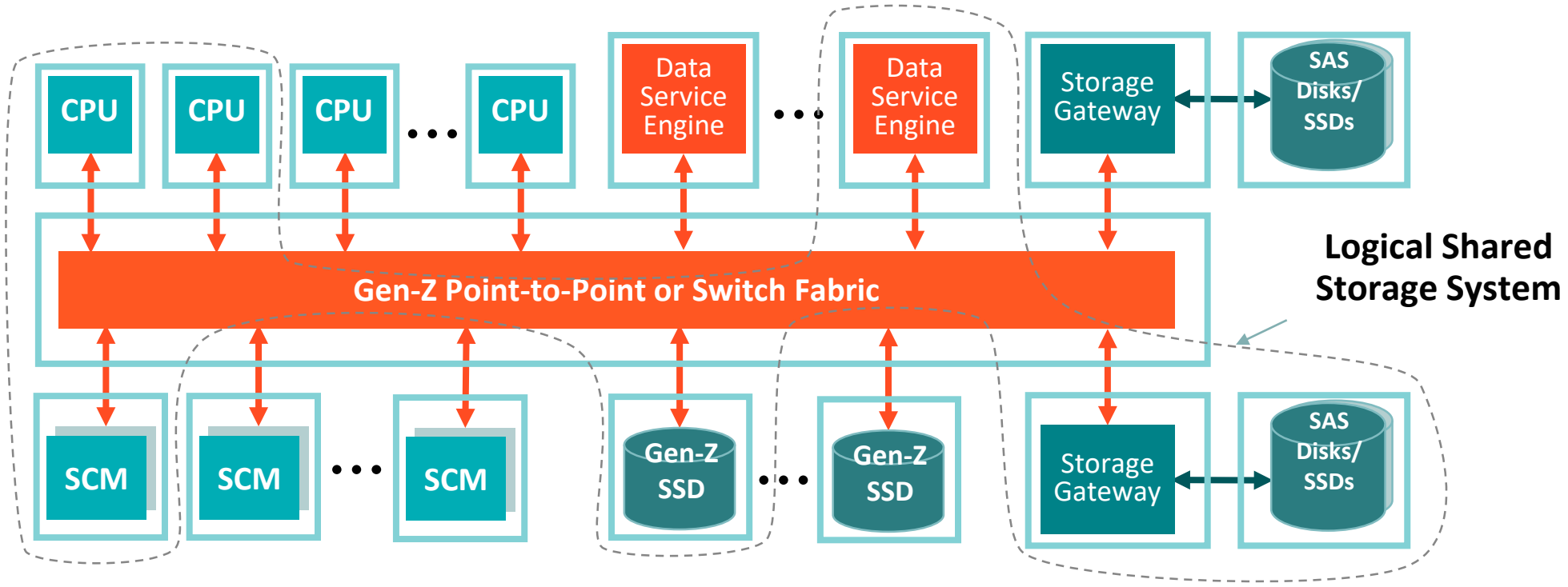
Flexible: *Scalable Form Factor*[1]

- Supports any component type
  - Flash, SCM, DRAM, NIC, GPU, FPGA, DSP, ASIC, etc. (these are conceptual)
- Supports multiple interconnect technologies—Gen-Z, PCIe, etc.
- Single and double-wide—scale in x-y-z directions
  - Increased media, power, performance, and thermal capacity
  - Double-wide can be inserted into pairwise single slots
- Supports 1C, 2C, and 4C scalable connectors
  - Larger modules can support multiple connectors—scale power & performance
- Scalable Form Factor Benefits:
  - Simplifies supply chain
  - Lower customer CAPEX / OPEX
  - Consistent customer experience
  - Increases solution and business agility @ lower dev cost
  - Eliminates Potential ESD Damage
    - Can safely move modules from failed / old to new enclosure
  - Eliminates SPOF or stranded resources
    - Multiple links per connector, multiple connectors per module
  - Scalable thermal plus improved airflow across components



[1] Draft specification publicly available—see www.genzconsortium.org
* Bandwidth calculated using 32 GT/s Signaling
** DRAM module provides 3.5x the highest-capacity DDR5 DIMM

Composable Storage



- Logical storage systems composed from components on Gen-Z fabrics
- Supports: Object storage, Key-Value Stores, Storage Arrays (small/large), etc.
- Supports Rich Data Service Accelerator components
  - RAID, dedup, replication, compression, thin provisioning, encryption, etc.

Key Gen-Z Attributes for Scale-out Solutions

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

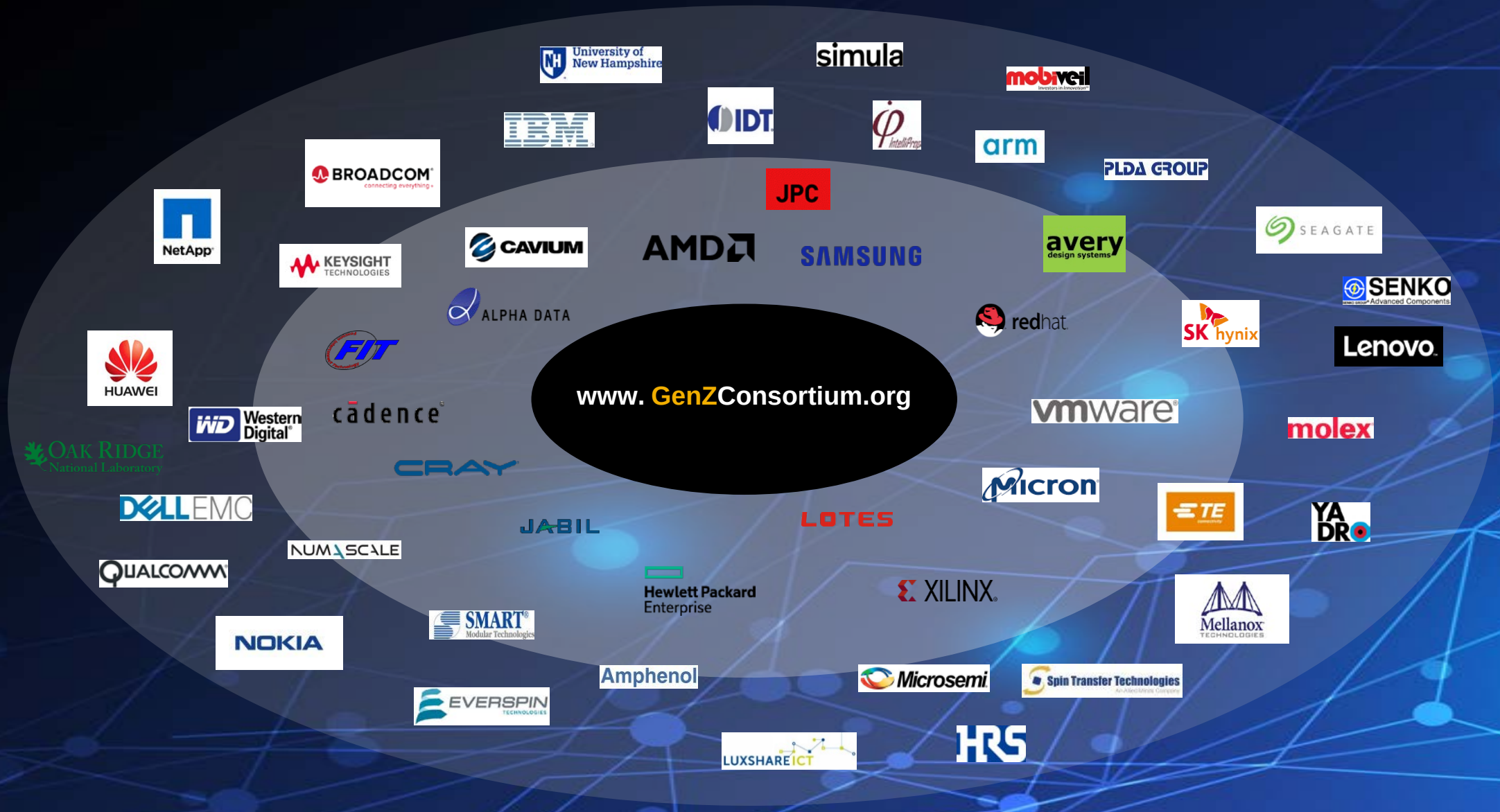
All planes can be co-packaged within a single switch

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.

A broad-based industry consortium



GEN Z

Gen-Z Background

- The Gen-Z Consortium launched in October of 2016 to create an open, industry standard for a high speed, low latency, scalable, memory centric fabric
- There are currently over 50 member companies covering all of the disciplines required to create an ecosystem based on this open standard
- Consortium members have released four 1.0 specifications
  - Core Specification 1.0 (released 2/13/18)
  - Scalable Connector Specification 1.0
  - SFF 8639 2.5-inch Specification 1.0
  - SFF 8639 2.5-inch Compact Specification 1.0
- Gen-Z showed a demo of multiple servers using multiple memory pools at SC'17
- Released and draft specifications are available on **www. GenZConsortium.org** for public review and comment

What's Next for Gen-Z?

- The Gen-Z consortium and its members will continue efforts to create Gen-Z IP and silicon in 2018
  - IntelliProp will release design IP in Q1 2018
  - Avery will release verification IP in Q1 2018
- Consortium workgroups continue developing our Compliance, Core, Mechanical, Phy, and Software & Management Specifications for public release
- Gen-Z will be showing demos at ISC'18, FMS'18, and SC'18
- Development systems targeted for late 2018
- Early Adopter systems targeted for late 2019 and early 2020
- Gen-Z is always interested in engaging new members

Get Involved!

To find out more about Gen-Z's implementation within OCP, please join the OCP HPC Community: http://www.opencompute.org/projects/high-performance-computing-hpc/

Or contact Gen-Z: https://genzconsortium.org

Thank You

OCP SUMMIT