OpenCAPI, Gen-Z, CCIX: Technology Overview, Trends, and Alignments

BRAD BENTON| AUGUST 8, 2017

Agenda

AGENDA

- Overview and motivation for new system interconnects
- Technical overview of the new, proposed technologies
- Emerging trends
- Paths to convergence?

Overview and Motivation

NEWLY EMERGING BUS/INTERCONNECT STANDARDS

AMD

Three new bus/interconnect standards announced in 2016

- **CCIX:** Cache Coherent Interconnect for Accelerators
  - Formed May, 2016
  - Founding members include: AMD, ARM, Huawei, IBM, Mellanox, Xilinx
  - ccixconsortium.com

- **Gen-Z**
  - Formed August, 2016
  - Founding members include: AMD, ARM, Cray, Dell EMC, HPE, IBM, Mellanox, Micron, Xilinx
  - genzconsortium.org

- **OpenCAPI:** Open Coherent Accelerator Processor Interface
  - Formed September, 2016
  - Founding members include: AMD, Google, IBM, Mellanox, Micron
  - opencapi.org

NEWLY EMERGING BUS/INTERCONNECT STANDARDS

AMD

Motivations for these new standards

- Tighter coupling between processors and accelerators (GPUs, FPGAs, etc.)
  - unified, virtual memory address space
    - reduce data movement and avoid data copies to/from accelerators
    - enables sharing of pointer-based data structures w/o the need for deep copies
- Open, non-proprietary standards-based solutions
- Higher bandwidth solutions
  - 25Gbs and above vs. 16Gbs forPCIe-Gen4
- Better exploitation of new and emerging memory/storage technologies
  - NVMe
  - Storage class memory (SCM), persistent memory
  - HMC, HBM, etc.
- Ease of programming
  - accelerator operates in the address space of the process
  - support for atomic operations

NEWLY EMERGING BUS/INTERCONNECT STANDARDS

AMD

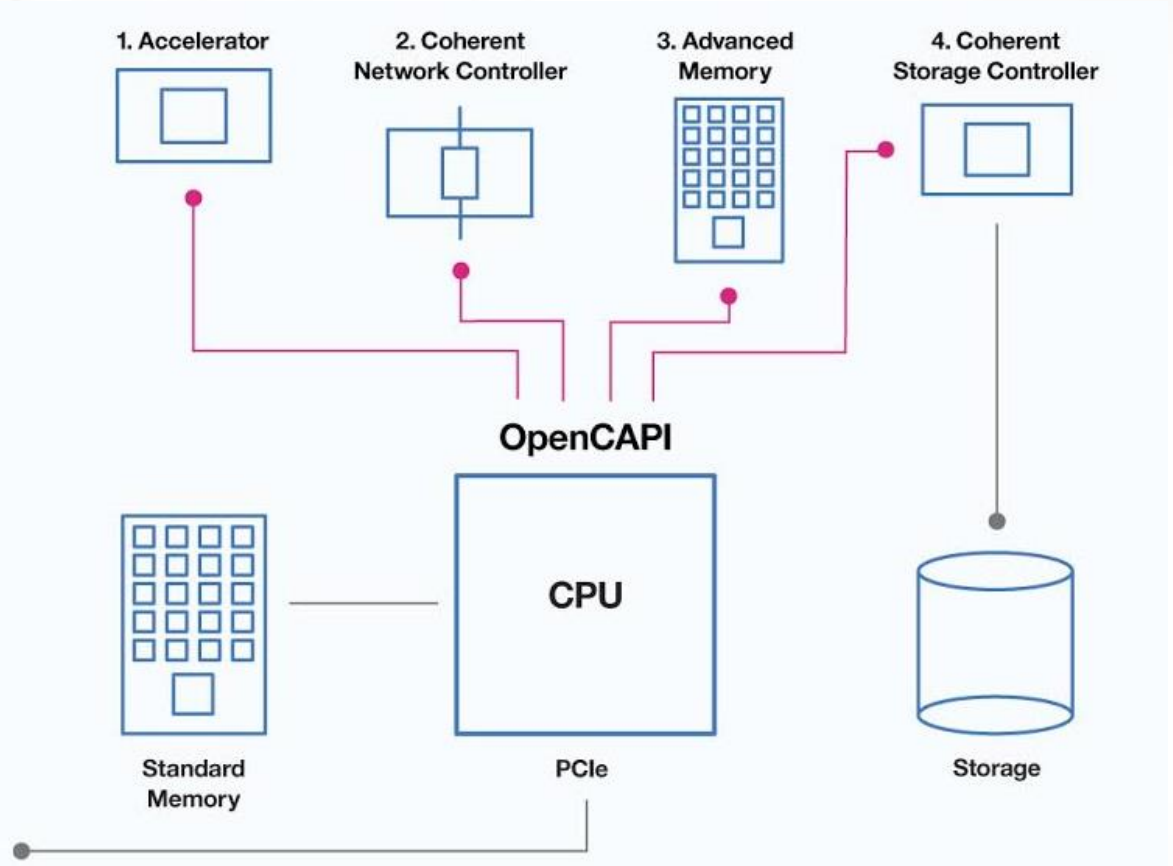
Why 3 different standards bodies?

- Different groups have been working to solve similar problems
- However, each approach has its differences
- Many companies involved with more than one consortium
- Possible shake outs/convergence as things move forward
- We've been here before: Future I/O + Next Generation I/O => InfiniBand
- But the landscape is different this time around

Technical Overview

OpenCAPI: OPEN COHERENT ACCELERATOR PROCESSOR INTERFACE

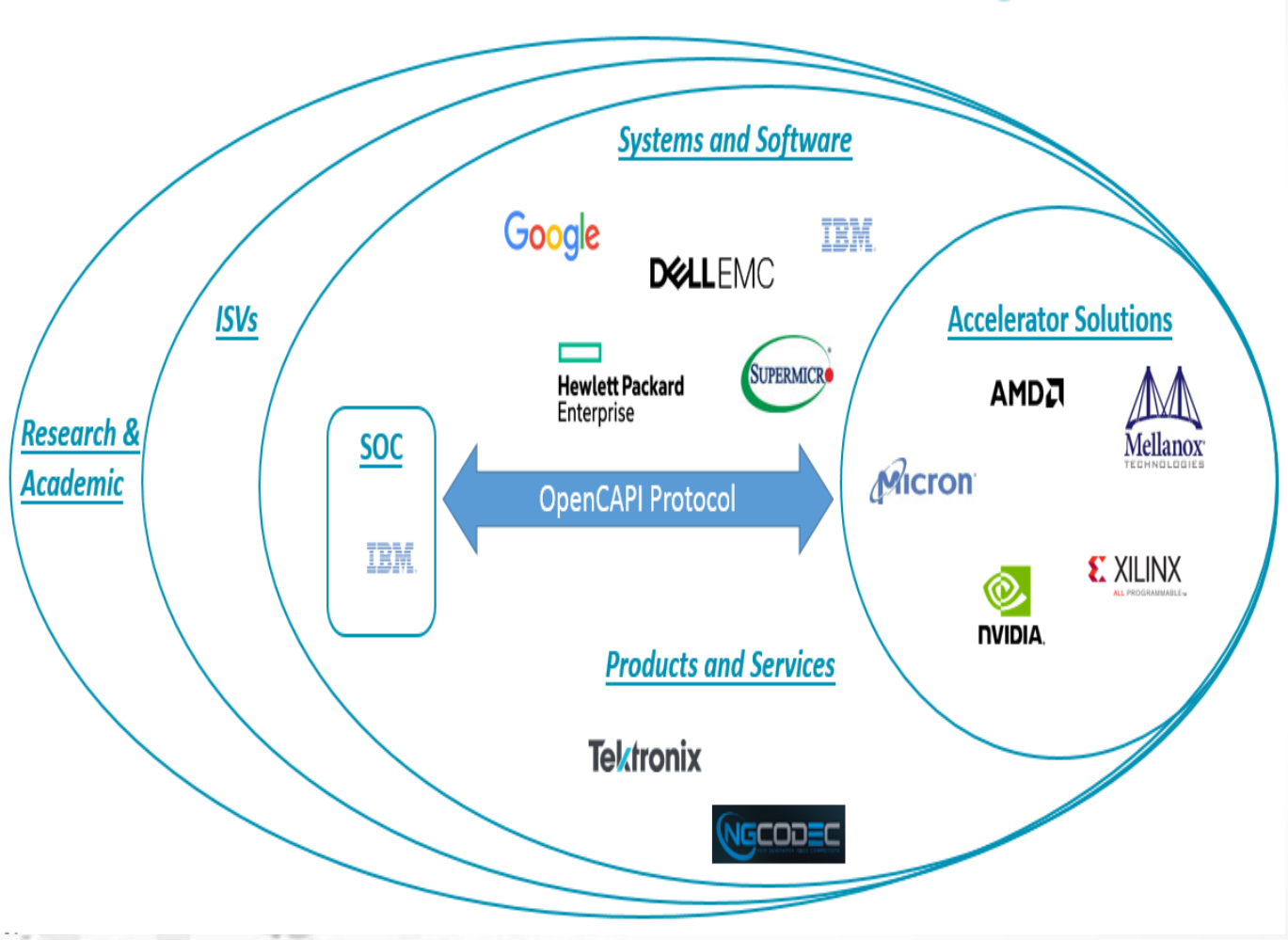
- Tightly coupled interface between processor, accelerators and memory
  - Operates on virtual addresses
  - virt-to-phys translation occurs on the host CPU
  - OpenCAPI 3.0:
    - Bandwidth: 25 Gps/lane x8
    - Coherent access to system memory from accelerator
  - OpenCAPI 4.0:
    - Support for caching on accelerators
    - Bandwidth:
      Support for additional link widths: x4, x8, x16, x32
- Use Cases
  - Coherent access from accelerator to system memory
  - Advanced memory technologies
  - Coherent storage controller
  - Agnostic to processor architecture



http://www.opencapi.org/

ATTRIBUTES

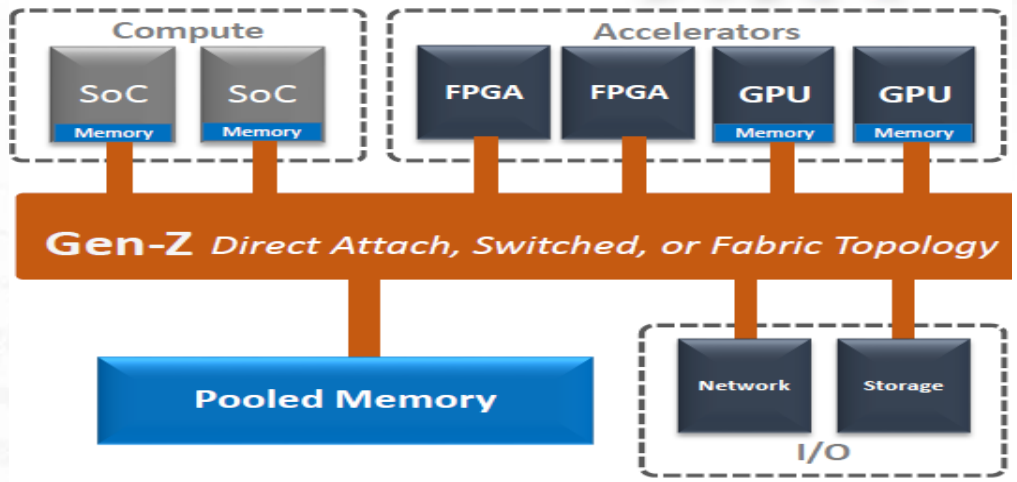
- Integrated into POWER9 (e.g., Zaius)
- Supports features of interest to NIC/FPGA vendors
  - Virtual address translation services
  - Aggregation of accelerator & system memory
- Communication with OpenCAPI-attached devices managed by vendor-specific drivers/libraries
- Accelerator holds virtual address; address translation managed by the host

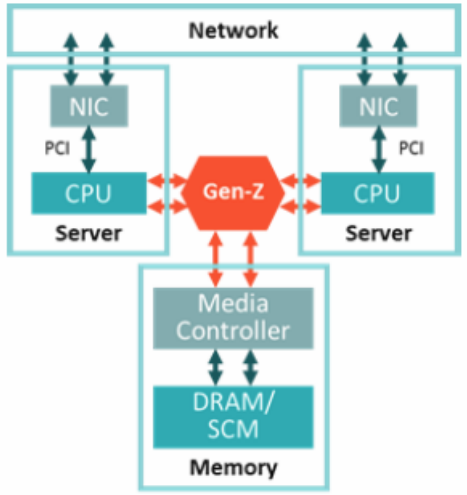


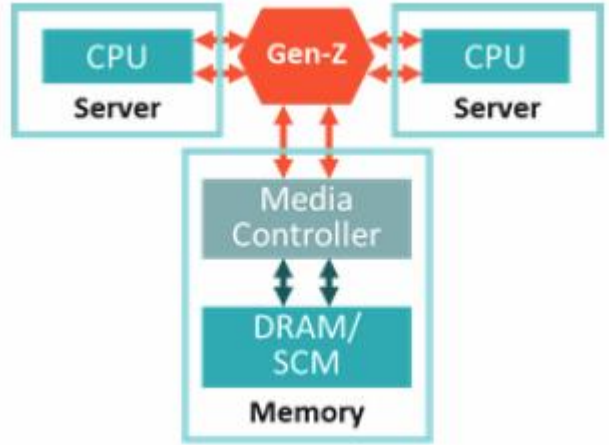
http://opencapi.org/wp-content/uploads/2016/11/OpenCAPI-Overview-SC16-vf.pptx

GEN-Z: MEMORY SEMANTIC FABRIC

- Scalable from component to cross-rack communications
  - Direct attach, switched, or fabric topologies
  - Bandwidth:
    - 32GB/s to 400+ GB/s
    - Support for intermediate speeds
  - Can gateway to other networks e.g., Enet, InfiniBand
  - Unify general data access as memory operations
    - byte addressable load/store
    - messaging (put/get)
    - IO (block memory)
- Use Cases
  - Component disaggregation
  - Persistent memory
  - Long haul/rack-to-rack interconnect
  - Rack-scale composability



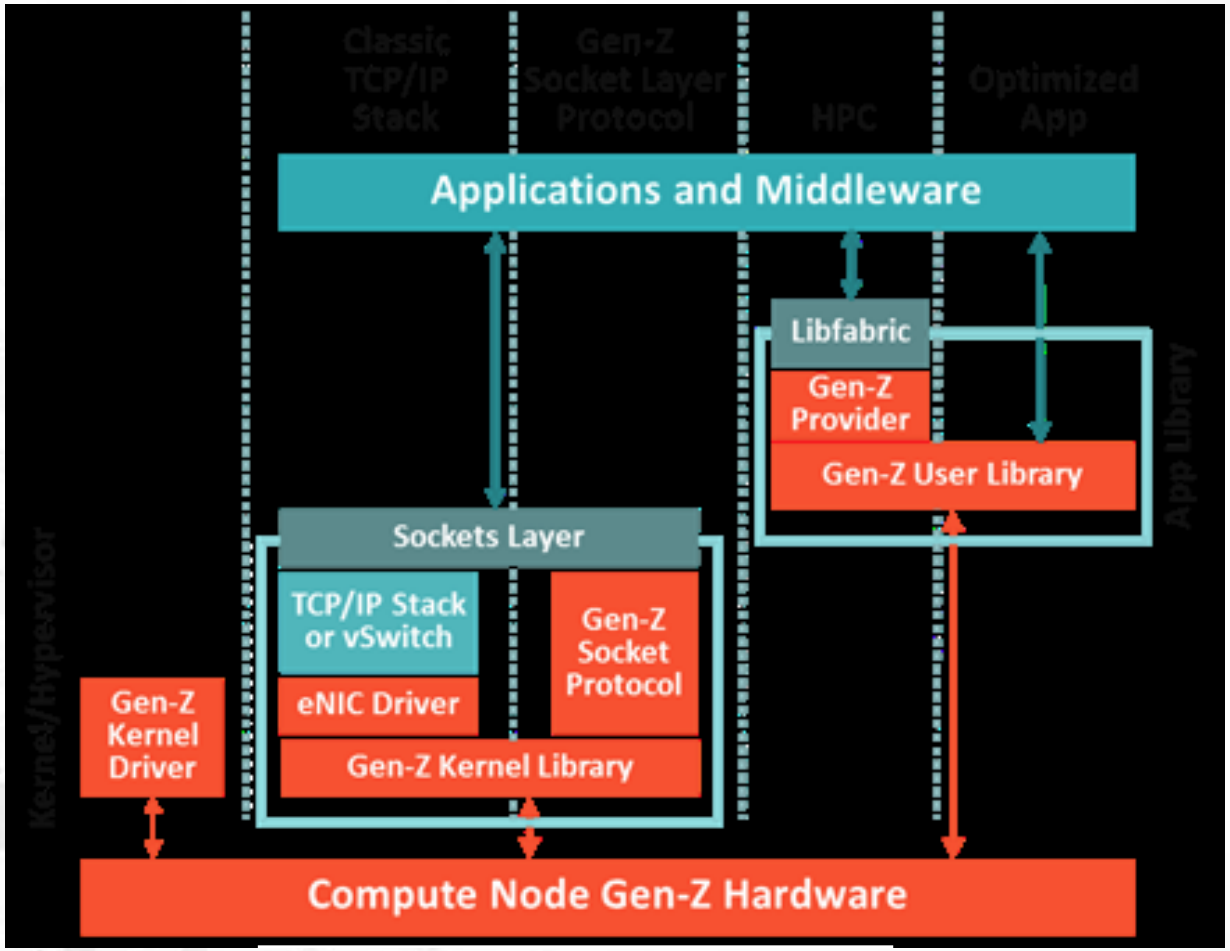




http://www.genzconsortium.org/

ATTRIBUTES

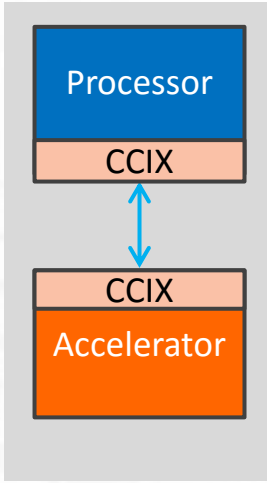
- Defines new mechanicals/connectors
- But will also integrate with existing mechanical form factors, connectors, and cables
- Multiple ways to move data:
  - Load/store
  - Messaging interfaces
  - Block I/O interfaces
- Gen-Z to provide libfabric integration for distributed messaging (MPI, OpenSHMEM, etc.)
- Scalability
  - Up to 4096 components/subnet
  - Up to 64K subnets

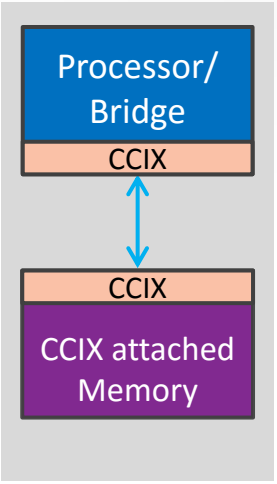


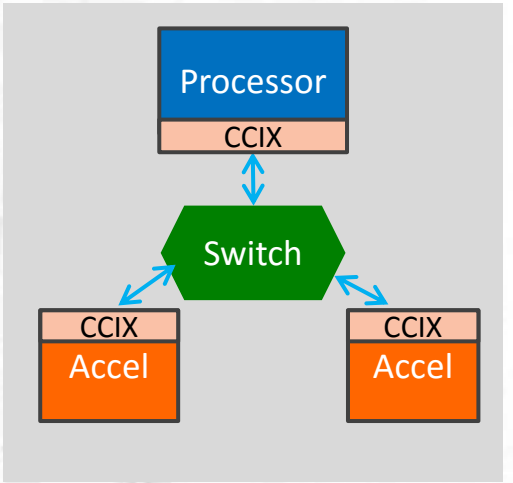
http://www.genzconsortium.org/

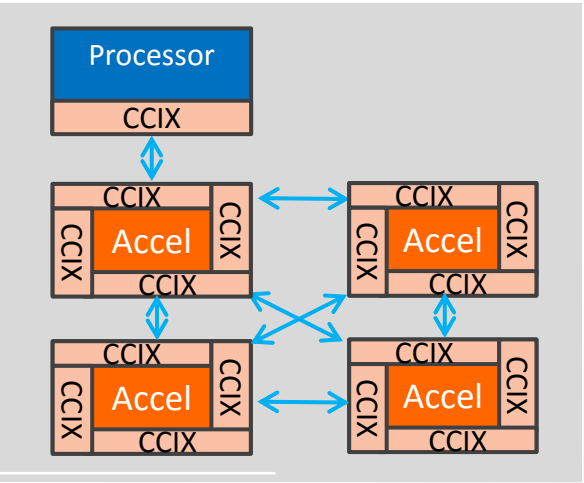
CCIX: CACHE COHERENT INTERCONNECT FOR ACCELERATORS

- Tightly coupled interface between processor, accelerators and memory
  - Bandwidth:
    - 16/20/25 Gps/lane
  - Hardware cache coherence enabled across the link
  - Driver-less and interrupt-less framework for data sharing
- Use Cases
  - Allows low-latency main memory expansion
  - Extend processor cache coherency to accelerators, network/storage adapters, etc.
  - Supports multiple ISAs over a single interconnect standard





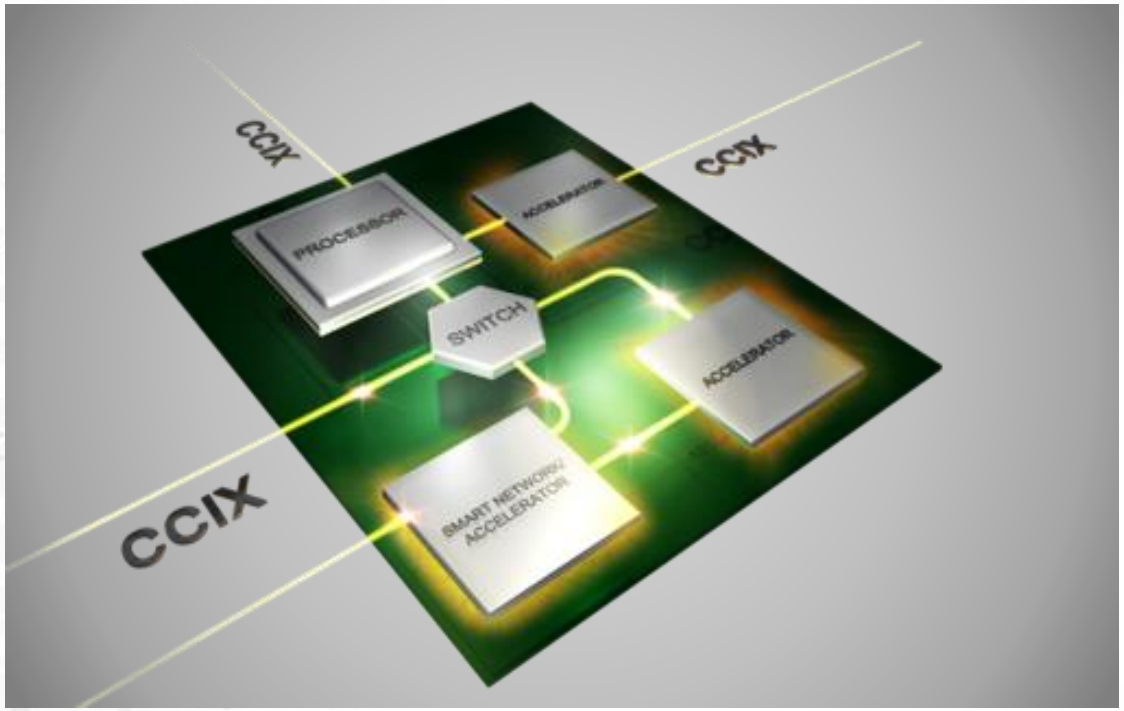




http://www.ccixconsortium.com/

ATTRIBUTES

- Link layer is a straightforward extension of PCIe
- In box/node system interconnect
- Preserves existing mechanicals, connectors, etc.
- Communication with CCIX-attached devices managed by vendor-specific drivers/libraries
- Address translation services via ATS/PRI



http://www.ccixconsortium.com/

COMPARISONS

AMD

Standard	Physical Layer	Topology	Unidirectional Bandwidth	Mechanicals	Coherence
CCIX	PCIe PHY	p2p and switched	32-50GB/s x16	PCIe	Full cache coherence between processors and accelerators
GenZ	IEEE 802.3 Short and Long Reach PHY	p2p and switched	Multiple Signaling Rates: 16, 25, 28, 56 GT/s Multiple link widths: 1 to 256 lanes	Supports existing PCIe mechanicals/form factors Will develop new, Gen-Z specific mechanicals/form factors	Does not specify cache coherent agent operations, but does specify protocols that support cache coherent agents
OpenCAPI 3.0	IEEE 802.3 Short Reach	p2p	25GB/s x8	In definition, see Zaius design for a possible approach	Coherent access to system memory
OpenCAPI 4.0	IEEE 802.3 Short Reach	p2p	Multiple link widths x4, x8, x16, x32 12.5, 25, 50, 100GB/s		Coherent access to memory Cache on accelerator

Trends

OPENCAPI MEMBERSHIP

AMD

Board Level

AMD
Google
IBM
Mellanox
Micron
NVIDIA
WesternDigital
Xilinx

Contributor Level

Amphenol
Microsemi
Molex*
Parade
Samsung
SK hynix*
TE Connectivity*
Tektronix
Toshiba

Observer/Academic Level

Achronix
Applied Materials
Dell EMC
ELI Beamlines
Everspin*
HPE
NGCodec
SmartDV*
SuperMicro
Synology
Univ. Cordoba

*New Members (since March, 2017)

GEN-Z MEMBERSHIP

AMD

General Member

Alpha Data
AMD
Amphenol Corporation
ARM
Avery Design*
Broadcom
Cadence*
Cavium
Cray
Dell EMC
Everspin Technologies*
FoxxConn
HPE
Huawei
IBM
Integrated Device Tech
IntelliProp
Jabil Circuit
Lenovo
Lotes
Luxchare*
Mellanox
Mentor Graphics*
Micron
Microsemi
Molex*
NetApp*
Nokia
Numascale*
PDLA Group
Red Hat
Samsung
Seagate
SK Hynix
SMART Modular Tech*
Spin Transfer
TE Connectivity*
Tyco Electronics*
Western Digital
Xilinx
Yadro

*New Members (since March, 2017)

CCIX MEMBERSHIP

AMD

Promoter Level

AMD
ARM
Huawei
Mellanox
Qualcomm
Xilinx

Contributor Level

Amphenol
Avery Design
Broadcom
Cadence
Cavium
Keysight
Lenovo*
Micron
Netspeed
Red Hat
Samsung*
SK hynix*
Synopsys
TE Connectivity

Adopter Level

Integrated Device Tech
INVECAS
Netronome*
Phytium Tech*
PLDA Group*
Shanghai Zhaoxin*
Silicon Labs*
SmartDV*

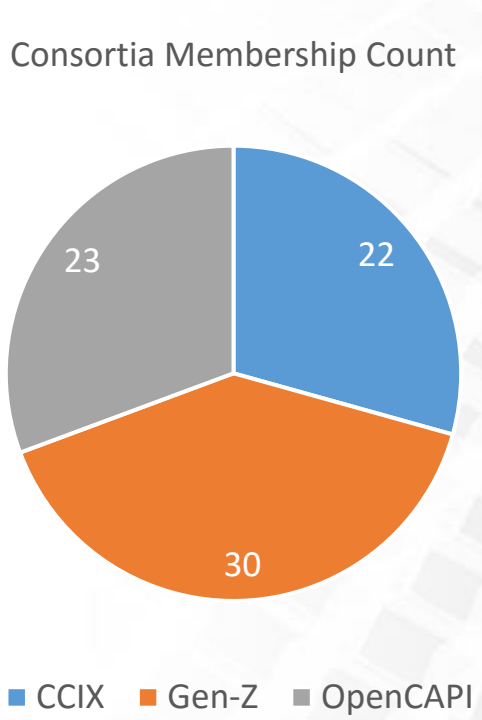
No Longer Members

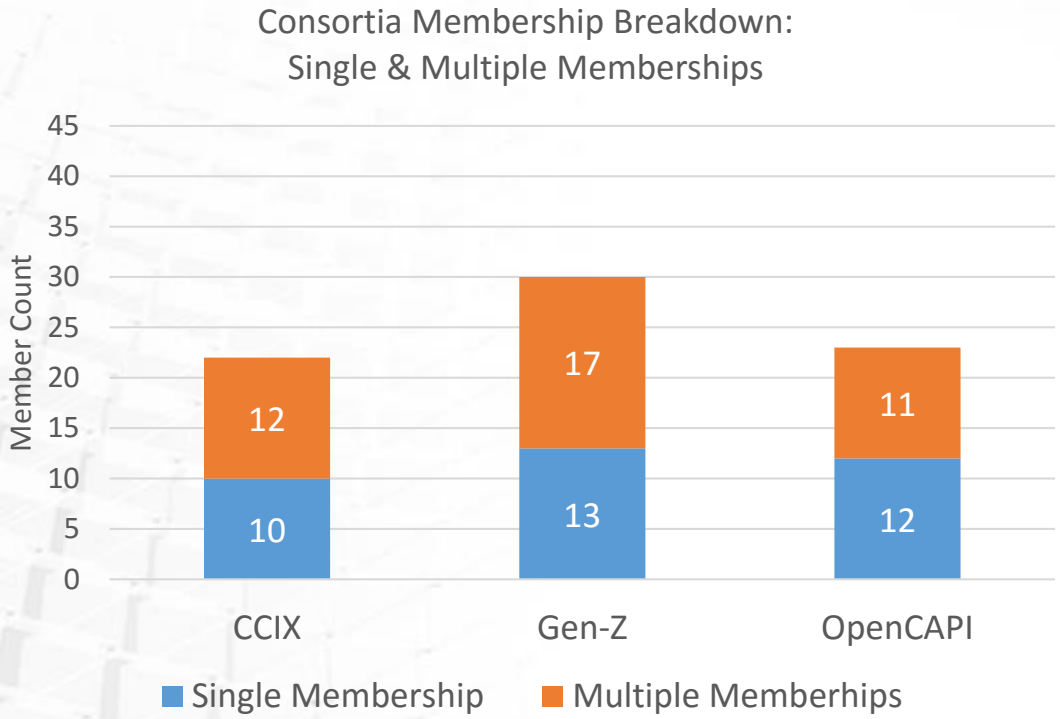
Arteris
Bull/Atos
IBM
Teledyne
TSMC

*New Members (since March, 2017)

CONSORTIA MEMBERSHIP DISTRIBUTION

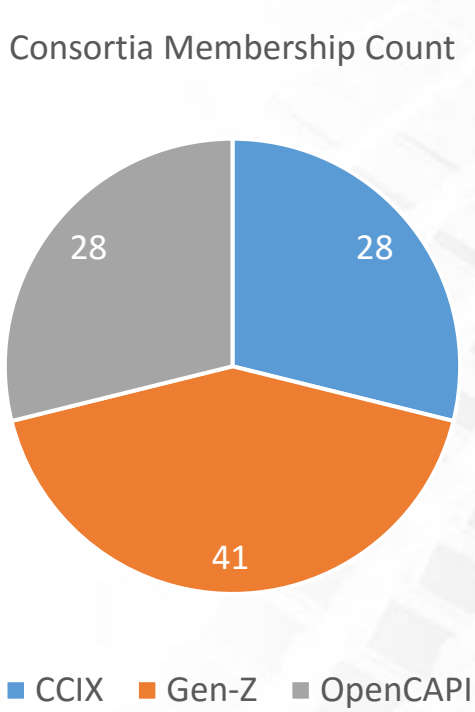
Membership Distribution: March, 2017; All Levels

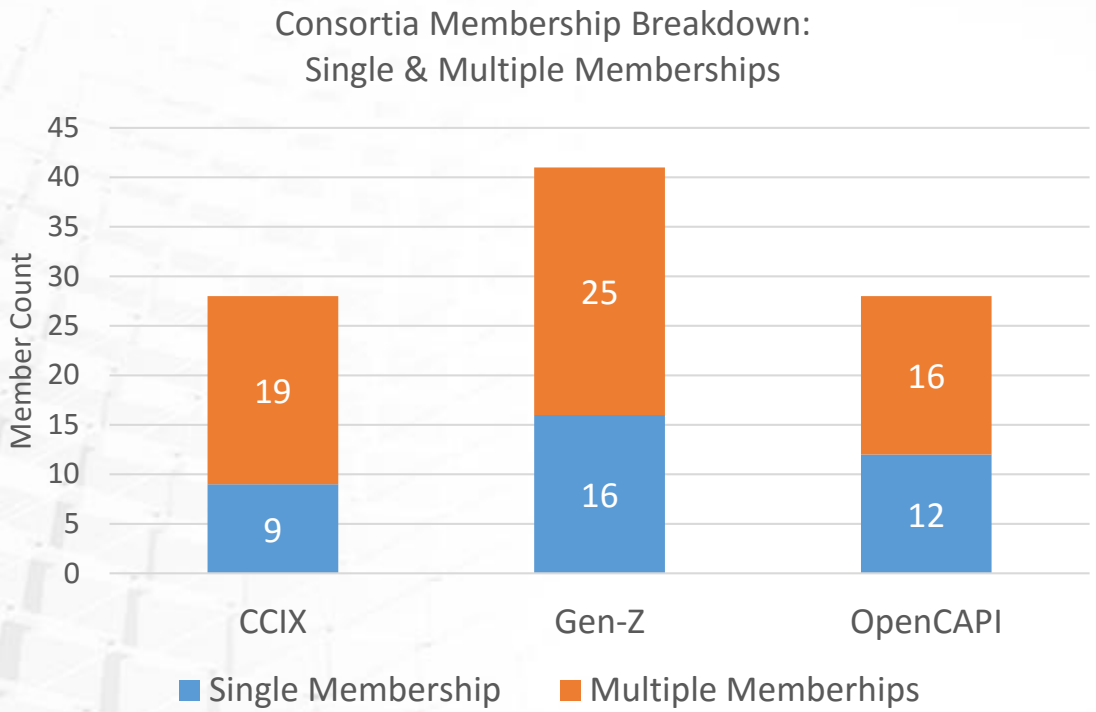




CONSORTIA MEMBERSHIP DISTRIBUTION

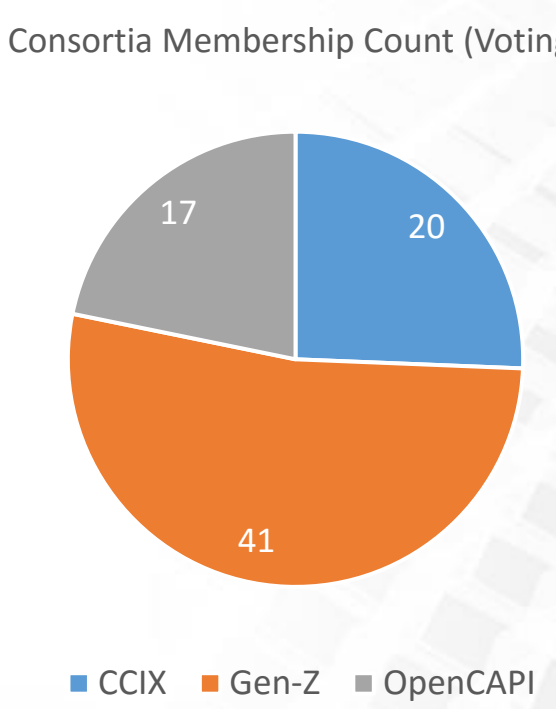
Membership Distribution: August, 2017; All Levels

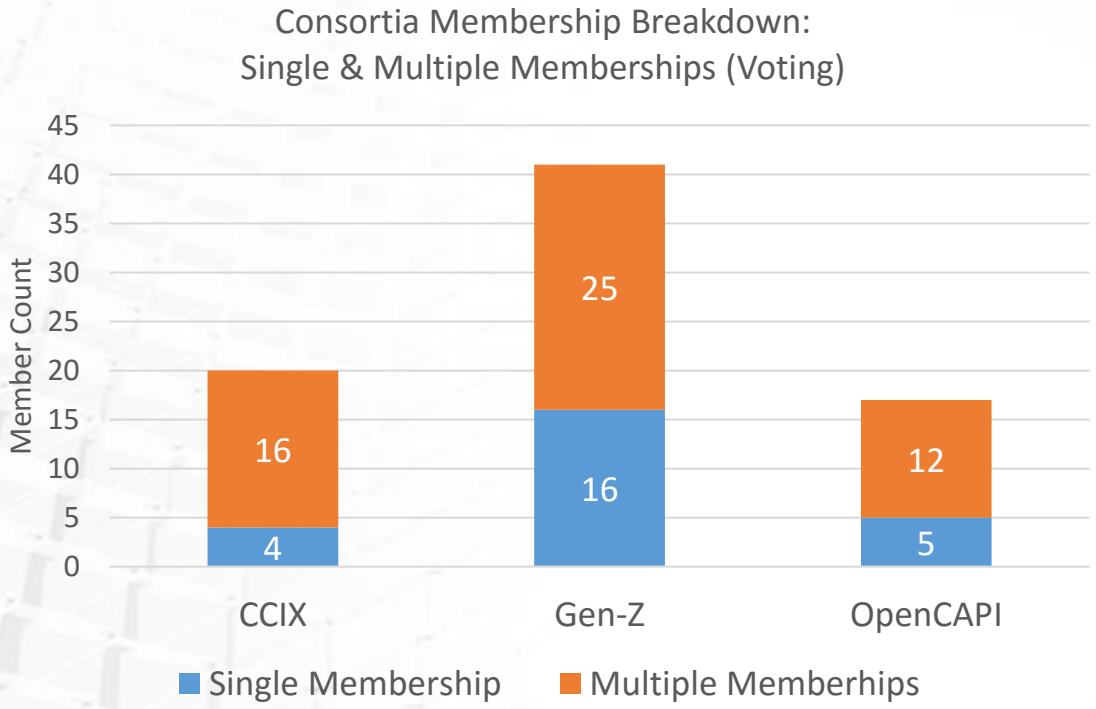




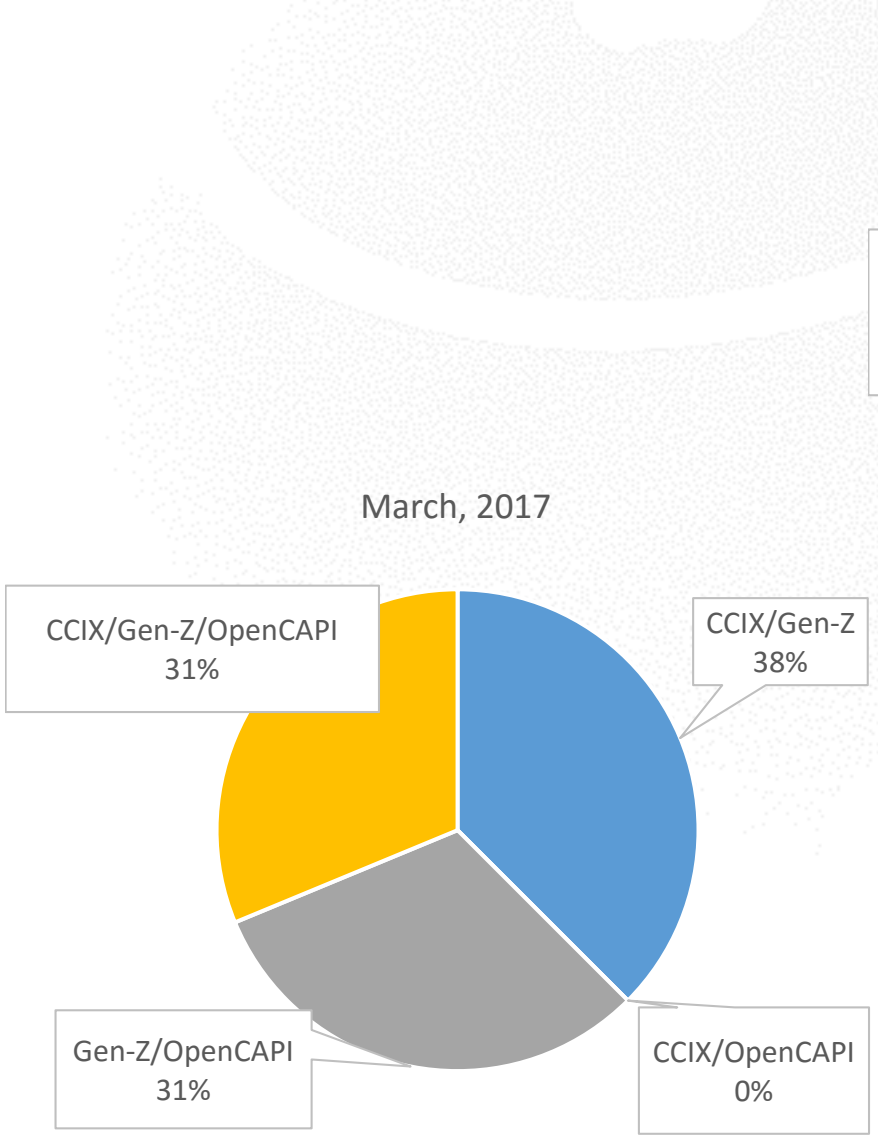
CONSORTIA MEMBERSHIP DISTRIBUTION

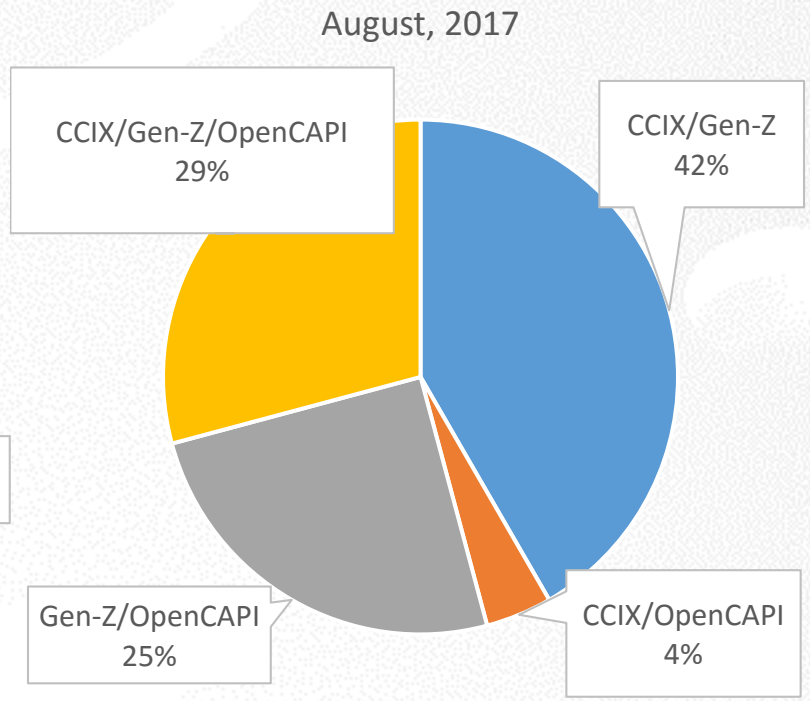
Membership Distribution: August, 2017; Voting Levels

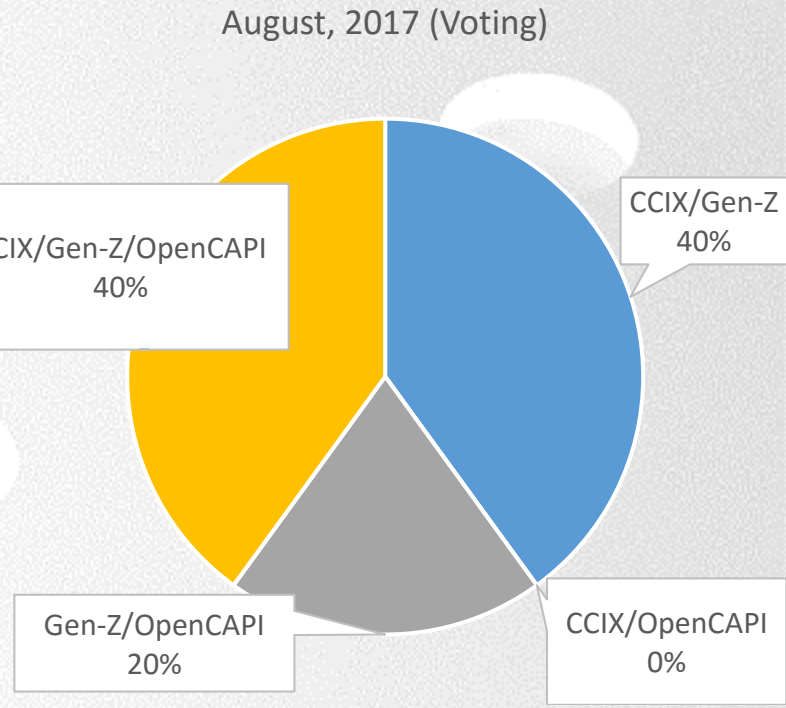




MULTIPLE MEMBERSHIP DISTRIBUTION

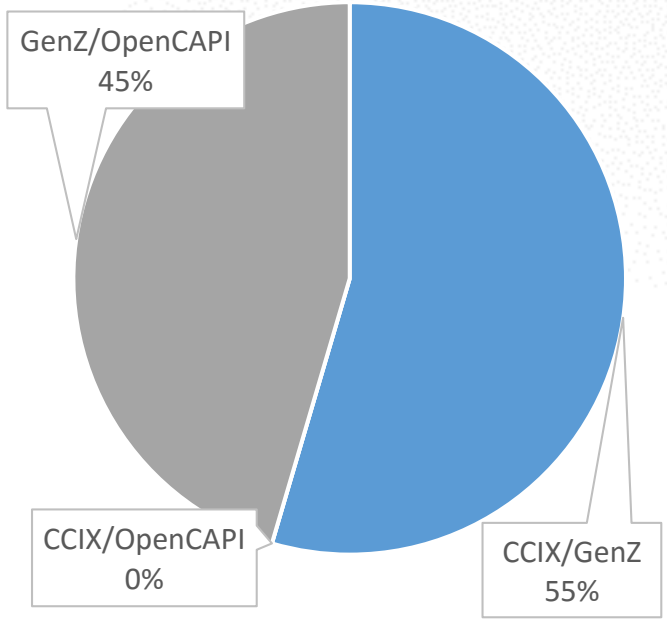




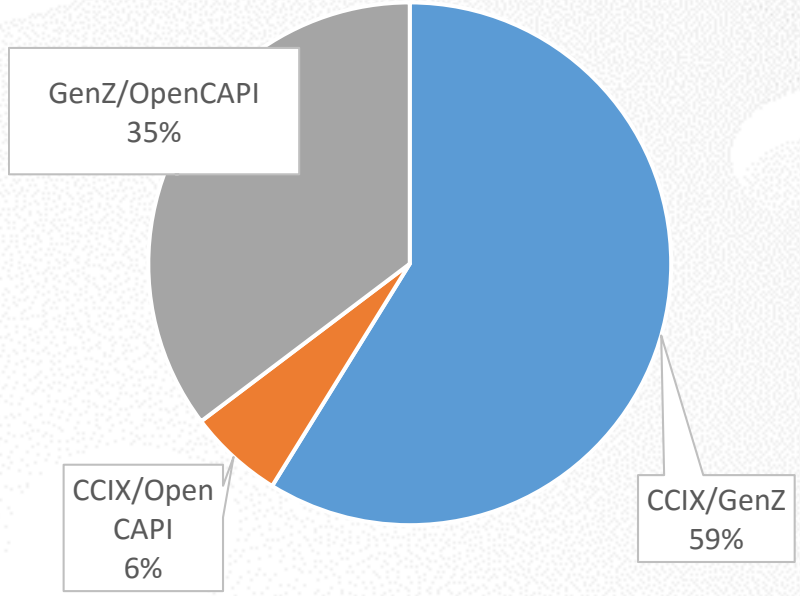


MEMBERSHIP PAIRINGS

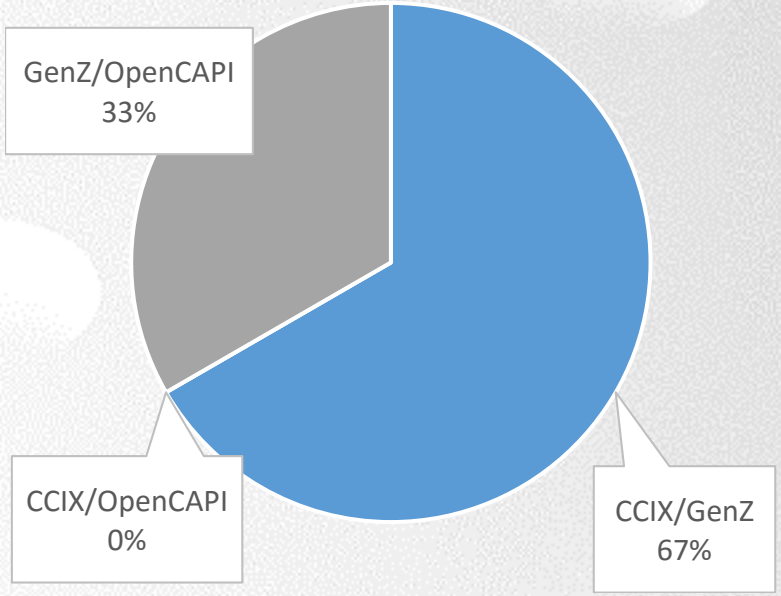
March, 2017
All Membership Levels



August, 2017
All Membership Levels



August, 2017
Voting Membership Levels



SOME ANALYSIS OF THE TRENDS

AMD

- Membership is in flux in all consortia
  - Many new memberships
    - At multiple levels, some at voting level, many at observer level
  - Some organizations have dropped memberships
  - Some organizations have added memberships
  - Each consortium has increased membership
  - More multiple memberships
    - Some straddling the fence (more triple memberships)?
    - or exposing alignments toward convergence?

- Not all membership changes yet visible
  - CCIX is over 1-year old, expect membership drops on annual boundaries
  - Gen-Z and OpenCAPI have yet to hit their annual boundaries
  - Current membership numbers for Gen-Z and OpenCAPI show additions, but we won't see companies that have left until fall

SOME ANALYSIS OF THE TRENDS

- CCIX
  - PCIe is a known quantity and strongly entrenched, particularly from electrical, connector standpoints
  - Move to CCIX provides a “natural” evolution: PCIe-Gen4 ESM + coherence
- Gen-Z
  - Highest membership, high multi-participation from both CCIX and OpenCAPI camps
  - Provides rack-level scale-out strategy
  - Potential for convergence strategies with both OpenCAPI and CCIX by which they manage short reach and use Gen-Z for long reach
- OpenCAPI/CCIX
  - These two technologies have a high degree of overlap
  - Built on two difference electrical technologies (802.3, PCIe)
  - Harder to see a convergence path

Paths to Convergence?

PATHS TO CONVERGENCE

- Vendors, integrators anxious for clarity
  - We've been here before: Future I/O + Next Generation I/O => InfiniBand
  - But the landscape is different this time around

- Reach
  - CCIX and OpenCAPI are short reach solutions
  - Gen-Z provides both short and long reach solutions
  - Gen-Z provides for rack-scale composability

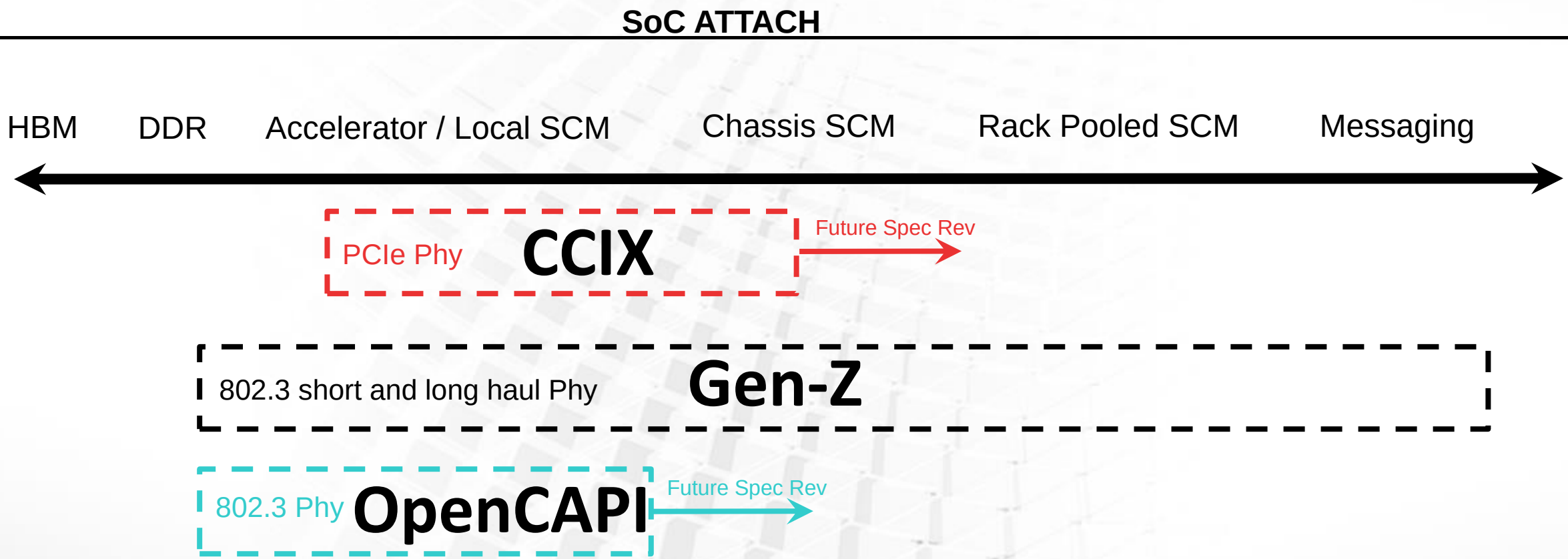
- Electricals
  - OpenCAPI and Gen-Z electricals are based on 802.3
  - CCIX is based on PCIe electricals
  - Gen-Z supports existing PCIe connectors & form factors

- CCIX & Gen-Z Connectivity has been suggested:

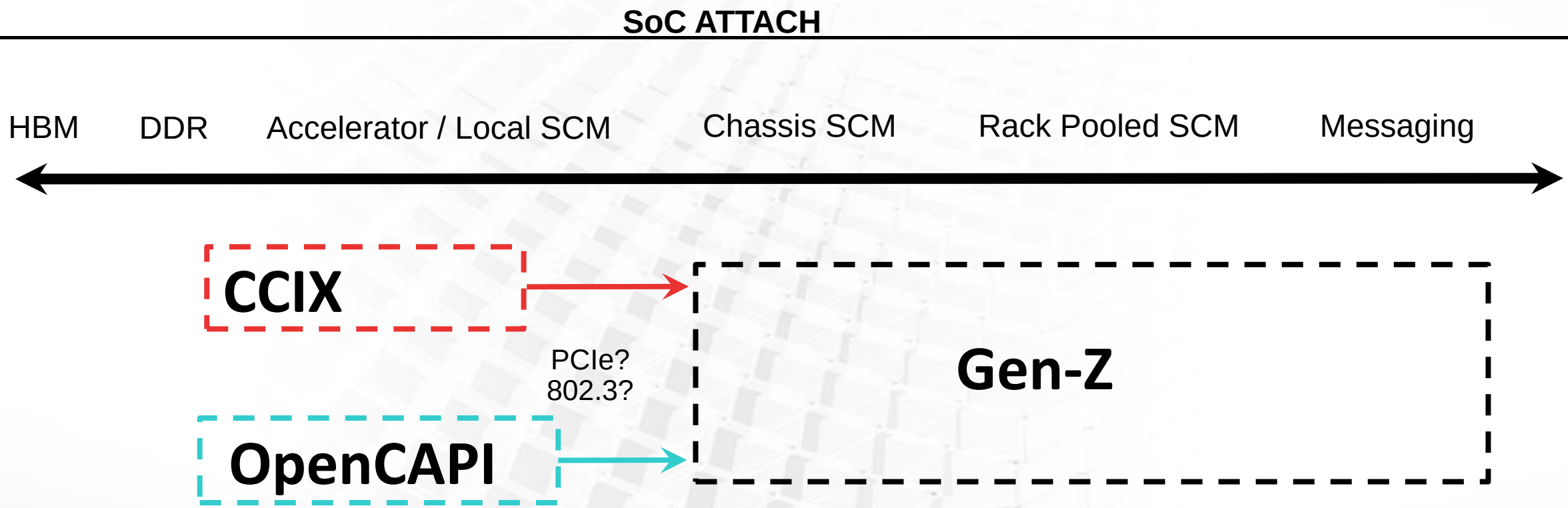
  GenZ is a new data access technology that enables memory operations to direct attach and disaggregated memory and storage. CCIX extends the processor's coherency domain to heterogeneous components. These heterogeneous "nodes" would then get access to the large and disaggregated storage and memory through the GenZ fabric.

  From: http://www.ccixconsortium.com/about-us.html
  Interestingly, this text was up on the site in March, 2017. But as of August, 2017, is no longer posted.

BUS REACH


SoC ATTACH
HBM
DDR
Accelerator / Local SCM
Chassis SCM
Rack Pooled SCM
Messaging
PCIe Phy
CCIX
Future Spec Rev
802.3 short and long haul Phy
Gen-Z
802.3 Phy
OpenCAPI
Future Spec Rev

POSSIBLE CONVERGENCE PATH



Questions, Thoughts, Discussion?

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION


AMD