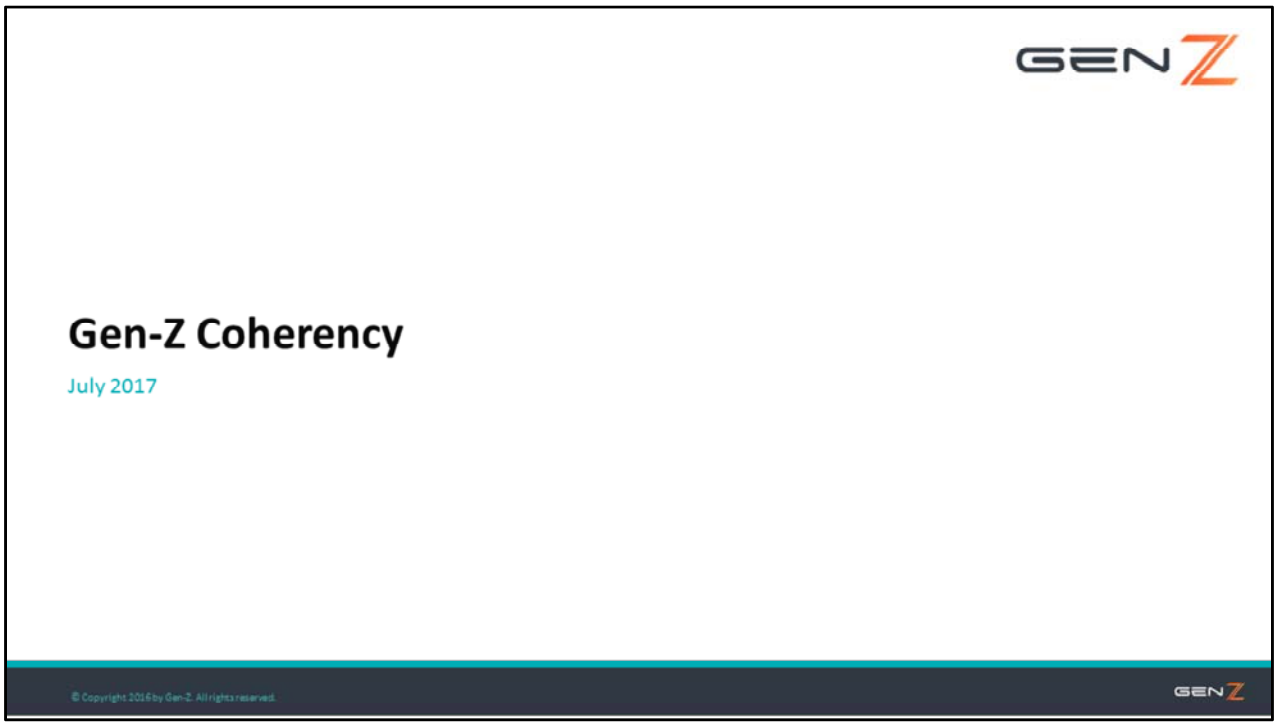


This presentation covers Gen-Z coherency operations and semantics.

Disclaimer

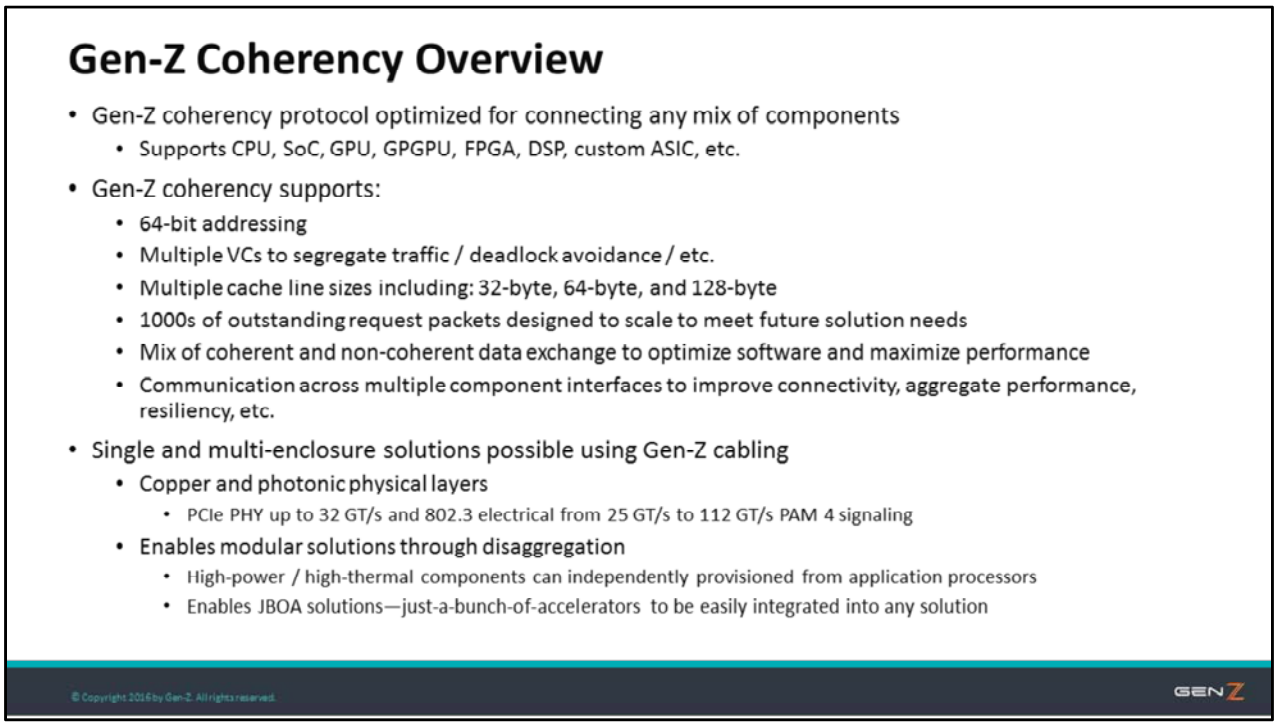
This document is provided 'as is' with no warranties whatsoever, including any warranty of merchantability, noninfringement, fitness for any particular purpose, or any warranty otherwise arising out of any proposal, specification, or sample. Gen-Z Consortium disclaims all liability for infringement of 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.

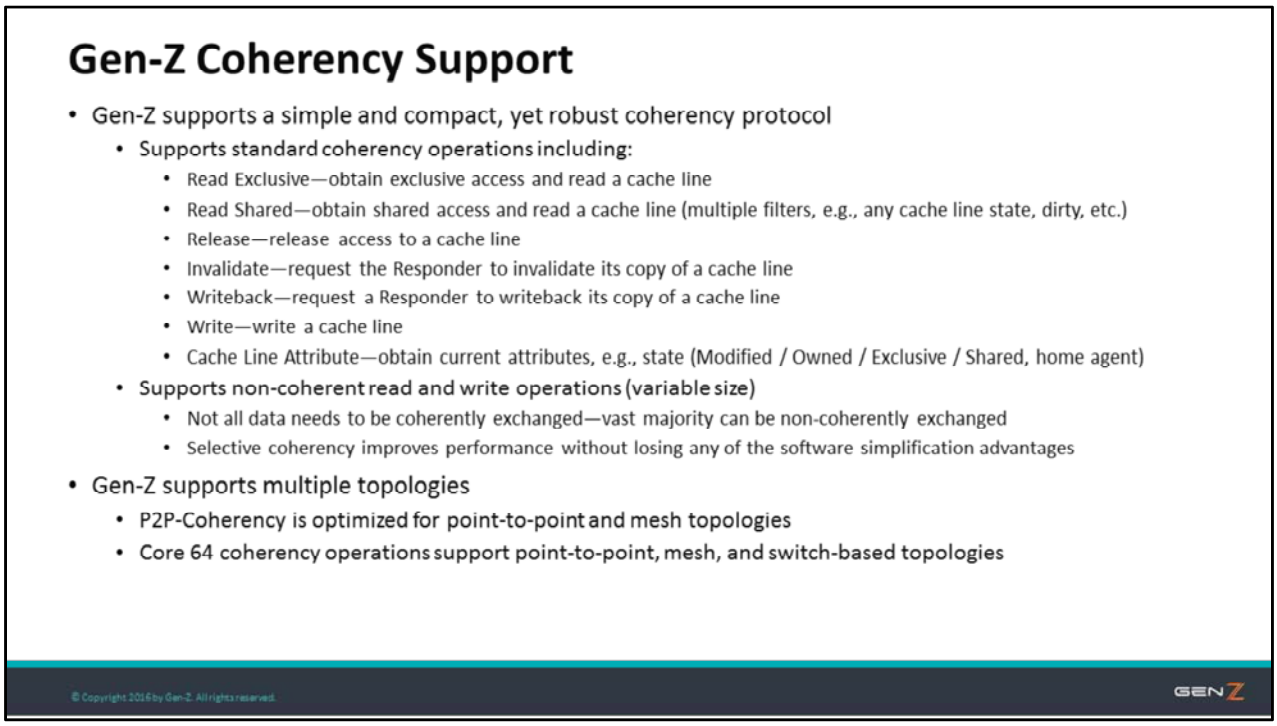
All material is subject to change at any time at the discretion of the Gen-Z Consortium

http://genzconsortium.org/



Gen-Z specifies a complete coherency protocol optimized for connecting any mix of components.

The primary attributes of Gen-Z Coherency are listed above. Gen-Z coherency can be used in point-to-point, mesh, or switch topologies. It can operate over a variety of copper and photonic physical layers and signaling rates. Single and multi-enclosure solutions are possible. Multi-enclosure solutions enable high-power and high-thermal components to be disaggregated from the application processors. For example, multi-enclosure solutions enable accelerator providers to interoperate with any platform provider and to quickly and easily provision and customize the accelerator enclosure to meet solution-specific needs. This will enable Just-a-Bunch-of-Accelerator solutions (similar to JBODs) to be constructed, unlocking solution innovation and increasing customer agility.



Gen-Z coherency supports the typical coherency operations required to support a variety of coherency schemes.

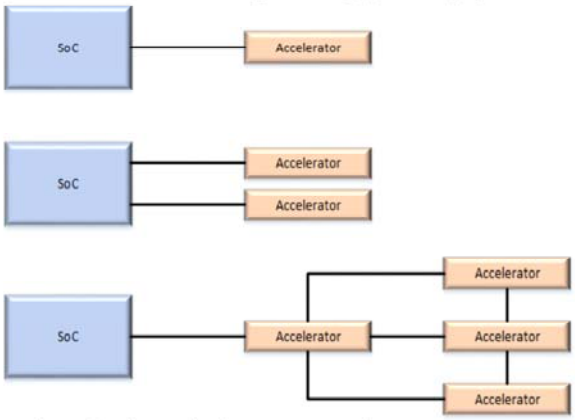
Gen-Z coherency also supports mixing coherent and non-coherent traffic and data sizes when communicating between components. In many cases, this can yield a significant improvement in protocol efficiency (more efficient data transfer) and in application performance (majority of the data exchanged between components does not require coherency semantics to be applied). For example, coherency could be applied to control operations such to application software pointers being directly passed to an accelerator via a coherent write and all subsequent data access being performed using non-coherent read and write operations.

The P2P-Coherency OpClass is optimized for use in point-to-point and mesh topologies. Components such as accelerators and SoCs can provision multiple Gen-Z links per component to provide switch-free connectivity between multiple components. Though Gen-Z switch latency should be much lower than alternative technologies (e.g., a Gen-Z switch latency should be 30-50 ns depending upon switch radix which is significantly lower than alternatives at 100-150 ns), many developers want to minimize latency as much as possible. Further, high-volume solutions that require coherency will consist of a small number of components (e.g., just a SoC and an accelerator), so point-to-point and meshing

are sufficient to meet these solution needs.

Core 64 OpClass supports point-to-point, mesh, and switch-based topologies. It supports the same coherency operations as supported with P2P-Core. It is intended for larger scale solutions or solutions containing a mix of coherent and non-coherent components, e.g., a memory-centric architecture where all components need to access shared memory or shared compute resources and only a subset need to communicate using coherency operations.

Point-to-Point / Meshed Topology Support

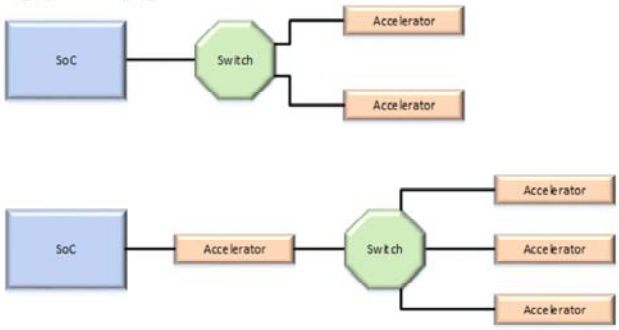


- Range of point-to-point and meshed topologies supported
- Optimized for SoC-to-accelerator or accelerator-to-accelerator communications
- Can intermix coherent and non-coherent data movement

GEN Z

P2P-Coherency can be used to connect components in a variety of point-to-point and meshed topologies. Components can be provisioned within a single enclosure or in multi-enclosure solutions attached through copper or optical cable solutions. Gen-Z architecture enables a component to support up to 4096 interfaces and links, thus provides more than enough connectivity to meet any solution's needs. This enables P2P-Coherency components to provision as many links as necessary to provide any level of connectivity.

Switch Topology Support

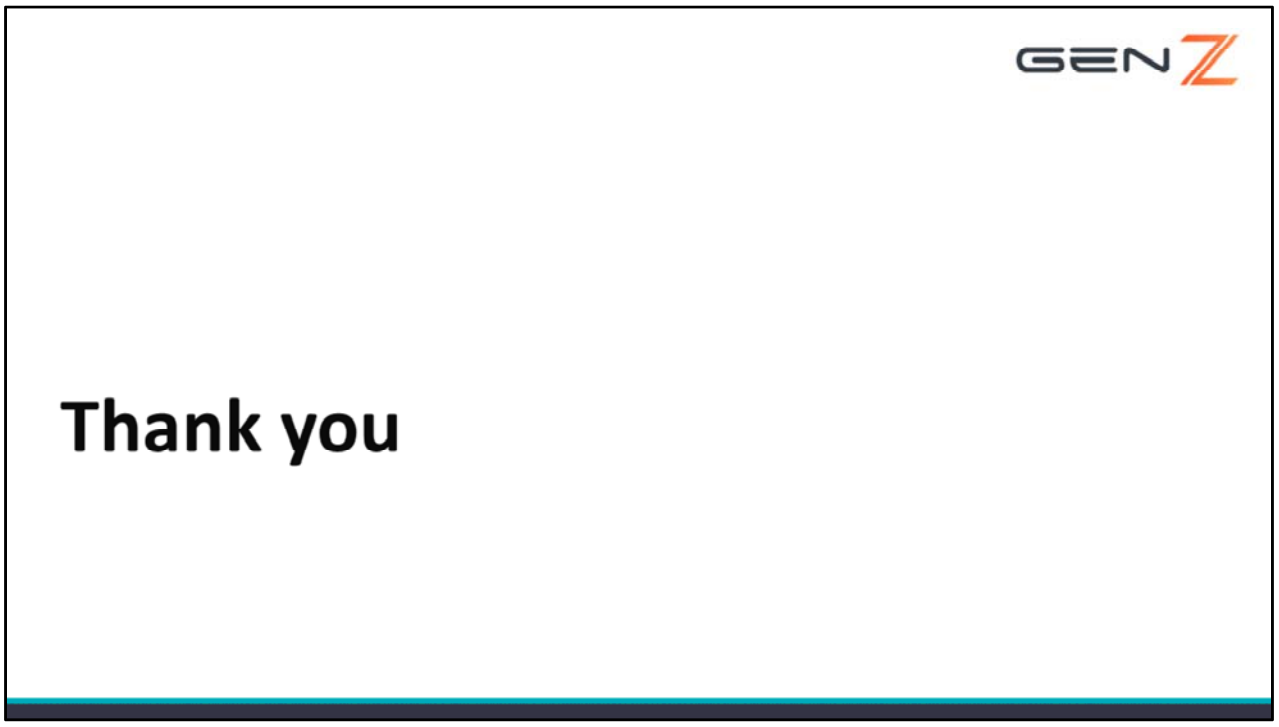


- Range of switch topologies supported
- Core 64 OpClass used for switch-based topologies
- Can intermix coherent and non-coherent operations and OpClasses
  - Can support Logical PCI Devices (LPDs) that use mix of coherent and non-coherent operations
- Gen-Z switch latency will vary by radix
  - Small radix switches (< 12 interfaces) should deliver 10-30 ns latency
  - Large radix switches (60 interfaces) should deliver 30-50 ns latency

GEN Z

Core 64 OpClass supports the same coherency operations as well as variable-length non-coherent read and write operations as P2P-coherency. Selective use of coherent operations as well as the ability to intermix any component types provides maximum solution composition flexibility without compromising performance. Core 64 can be used in single and multi-enclosure configurations to scale and meet any solution's needs.

Gen-Z switch latency is a key point of differentiation. Switch latency in alternative technologies range from 120-150 ns. A Read Exclusive-Read Response operation through a single switch will incur 20-60 ns of switch-specific latency compared to 240-300 ns of switch-specific latency using an alternative technology.



This concludes this presentation. Thank you.