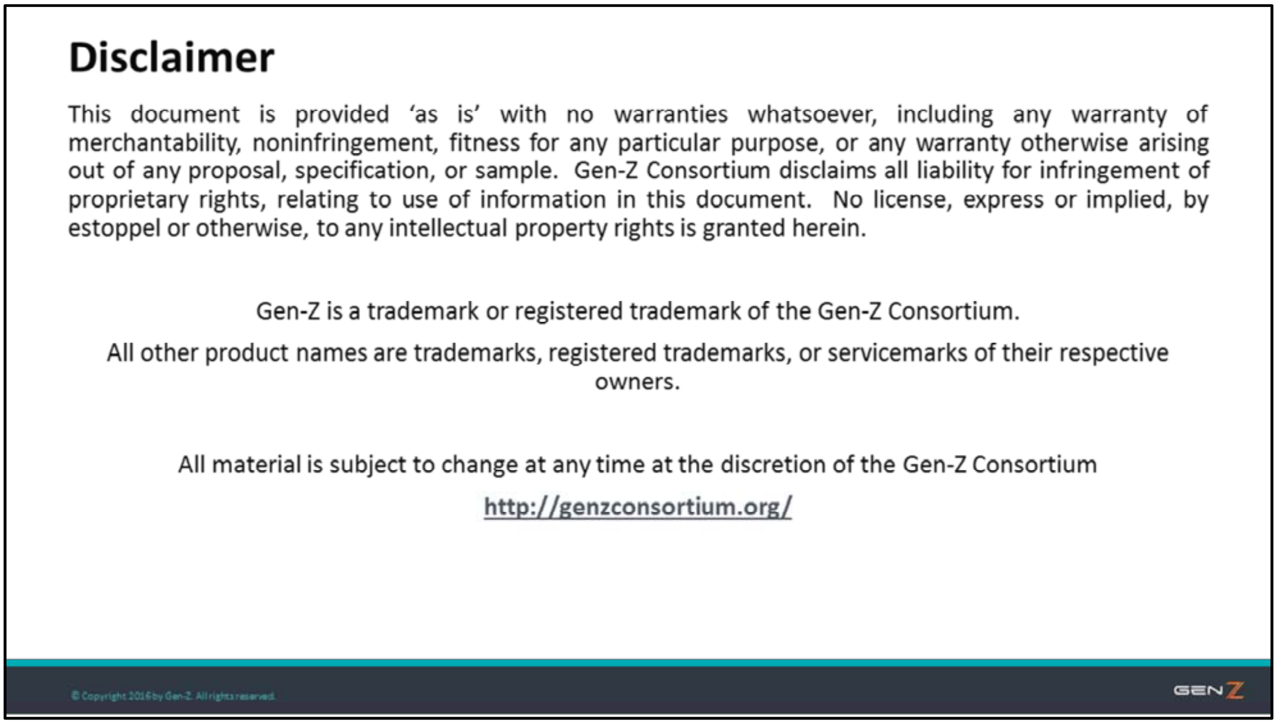
Gen-Z Unicast and Multicast Operations

July 2017

This presentation provides a high-level overview of Gen-Z unicast and multicast operations.



Disclaimer

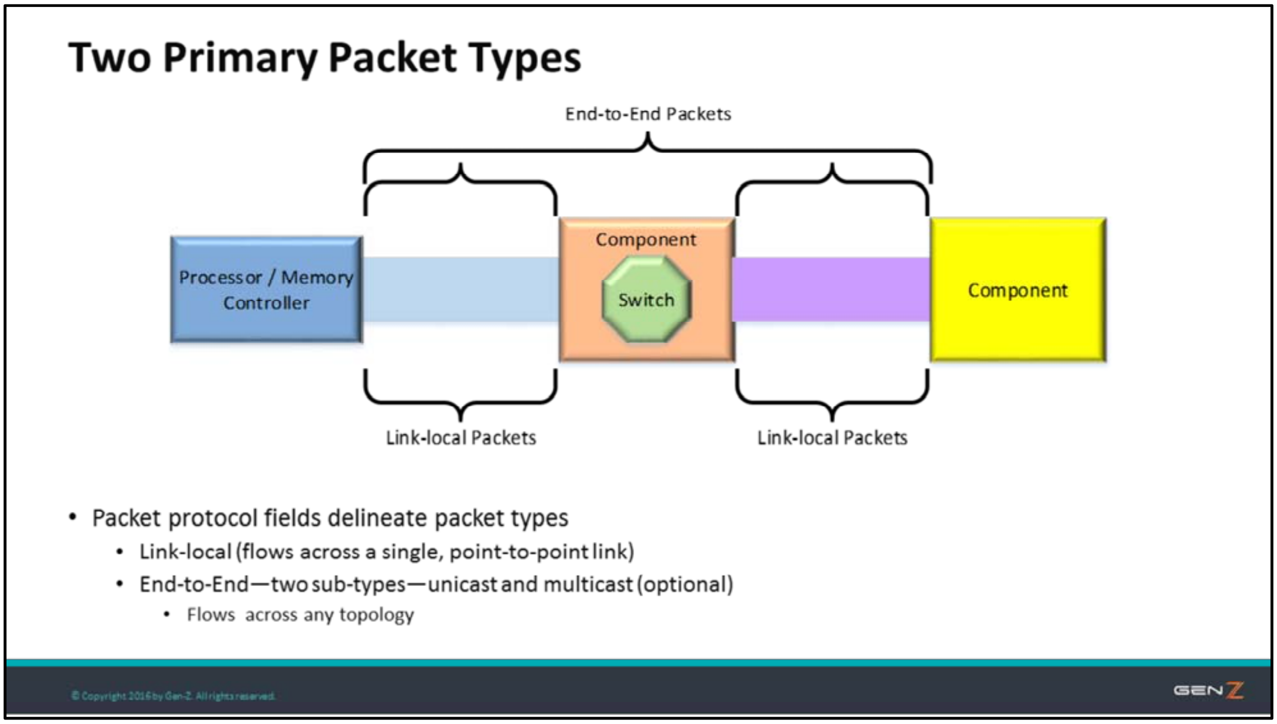
This document is provided 'as is' with no warranties whatsoever, including any warranty of merchantability, noninfringement, fitness for any particular purpose, or any warranty otherwise arising out of any proposal, specification, or sample. Gen-Z Consortium disclaims all liability for infringement of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

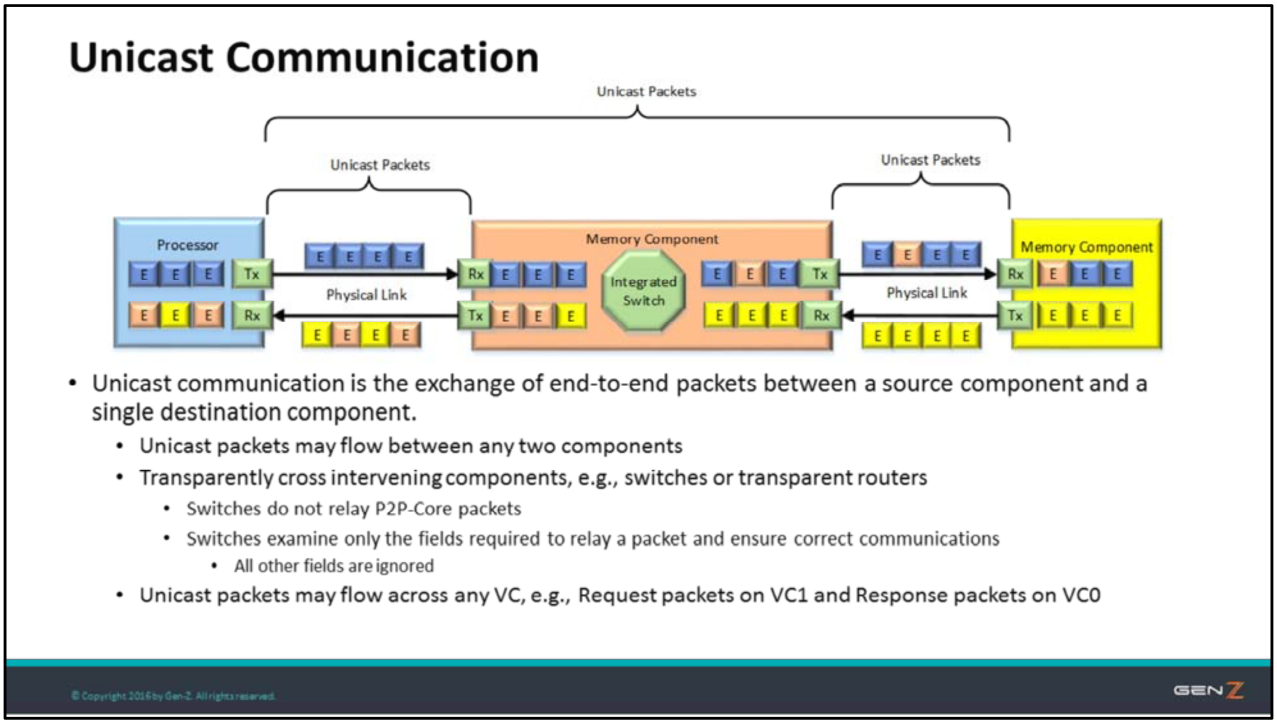


Two Primary Packet Types

- Packet protocol fields delineate packet types
  - Link-local (flows across a single, point-to-point link)
  - End-to-End—two sub-types—unicast and multicast (optional)
    - Flows across any topology

Gen-Z specifies two packet types:

- Link-local packets that are exchanged between two interfaces across a single link. Link-local packets are used to manage the link, e.g., exchange flow-control credits, perform link-level reliability, etc.
- End-to-end packets that are exchanged between at least two components. End-to-end packets flow across any topology—co-packaged, point-to-point, daisy-chain, mesh, switch-based, TR, etc.
  - End-to-end packets can be unicast packets. A unicast packet flows between a single source component and a single destination component.
  - End-to-end packets can be multicast packets. A multicast packet flows between a single source component and multiple destination components.
  - End-to-end packets can be exchanged within a single subnet or across multiple subnets (explicit switch-based packet relay or transparently relayed by a Transparent Router (TR)).



Unicast Communication

- Unicast communication is the exchange of end-to-end packets between a source component and a single destination component.
  - Unicast packets may flow between any two components
  - Transparently cross intervening components, e.g., switches or transparent routers
    - Switches do not relay P2P-Core packets
    - Switches examine only the fields required to relay a packet and ensure correct communications
      - All other fields are ignored
  - Unicast packets may flow across any VC, e.g., Request packets on VC1 and Response packets on VC0

End-to-end unicast packets can be exchanged across any topology. Further, they can be optimized for specific topologies, e.g., point-to-point.

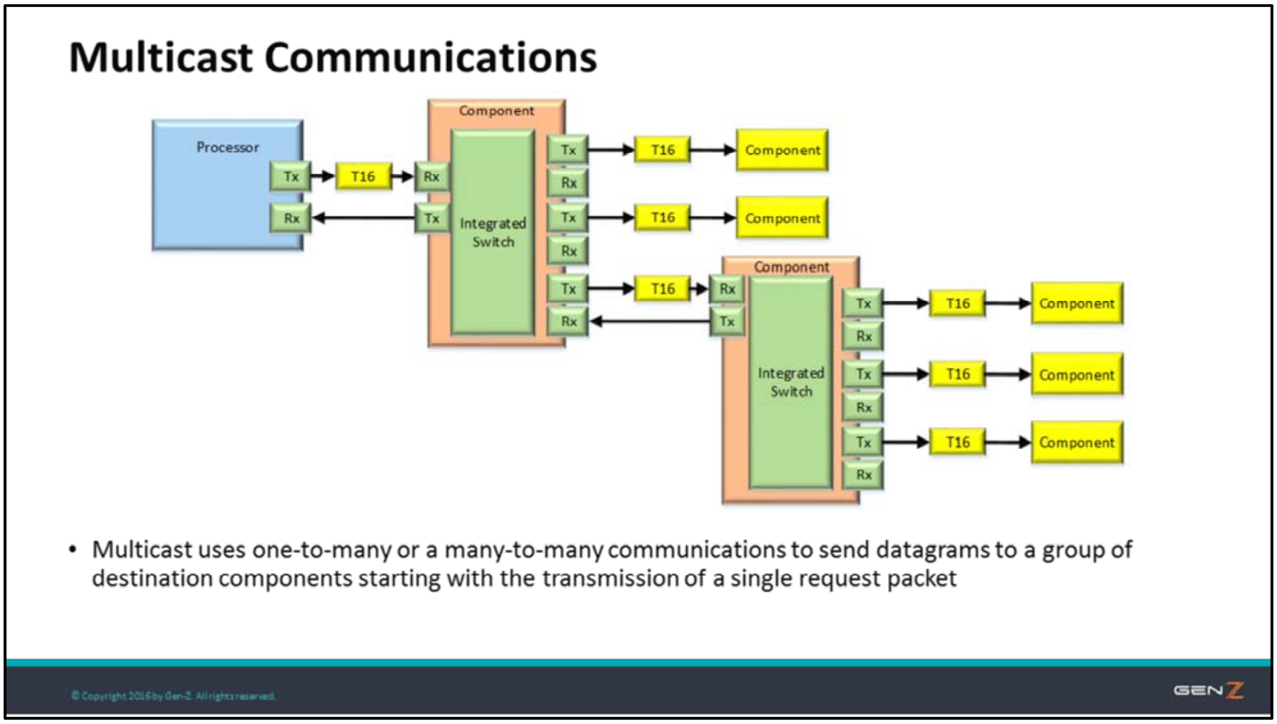
End-to-end packets from different components or from different VCs can be interleaved across any link.

End-to-end packets can be interleaved with link-local packets across a given link.

If a P2P-Core end-to-end packet, then the receiving component examines the encoded Tag field to determine whether the packet is destined to it or to the next hop in a daisy-chain topology.

If a P2P-Coherency or P2P-Vendor-defined, then the receiving component is the destination.

If an explicit OpClass packet, then a switch examines the DCID / DSID, VC, and OCL , or the MGID / GMGID, VC, and OCL to determine how to relay the packet.

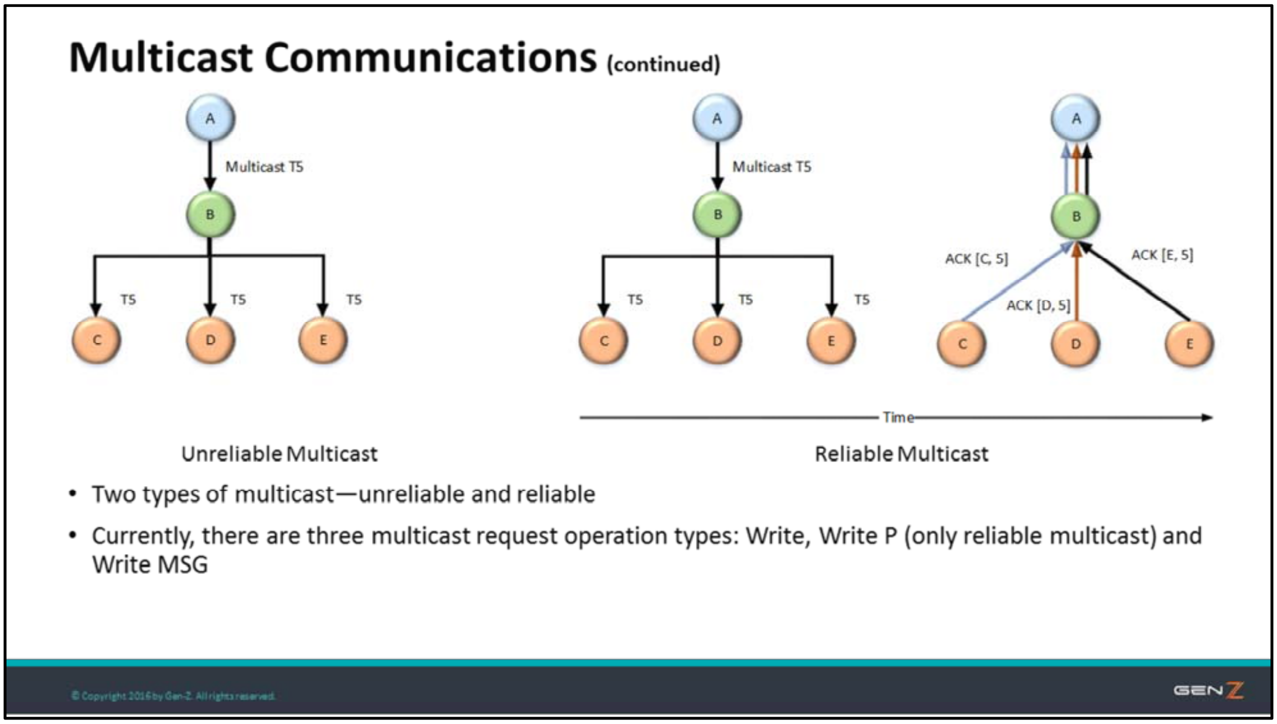


Multicast Communications

- Multicast uses one-to-many or a many-to-many communications to send datagrams to a group of destination components starting with the transmission of a single request packet

GEN Z

Multicast packets reduce fabric load by transmitting a single request packet and replicating it as it progresses throughout a switch topology such that each component participating in a multicast group receives a copy. Each switch that participates in a multicast group examines the packet’s OCL field to determine if it is the Multicast OpClass. If OCL == Multicast, then the switch treats the same bits that contain a unicast DCID as a multicast group identifier.



Multicast Communications (continued)

Unreliable Multicast

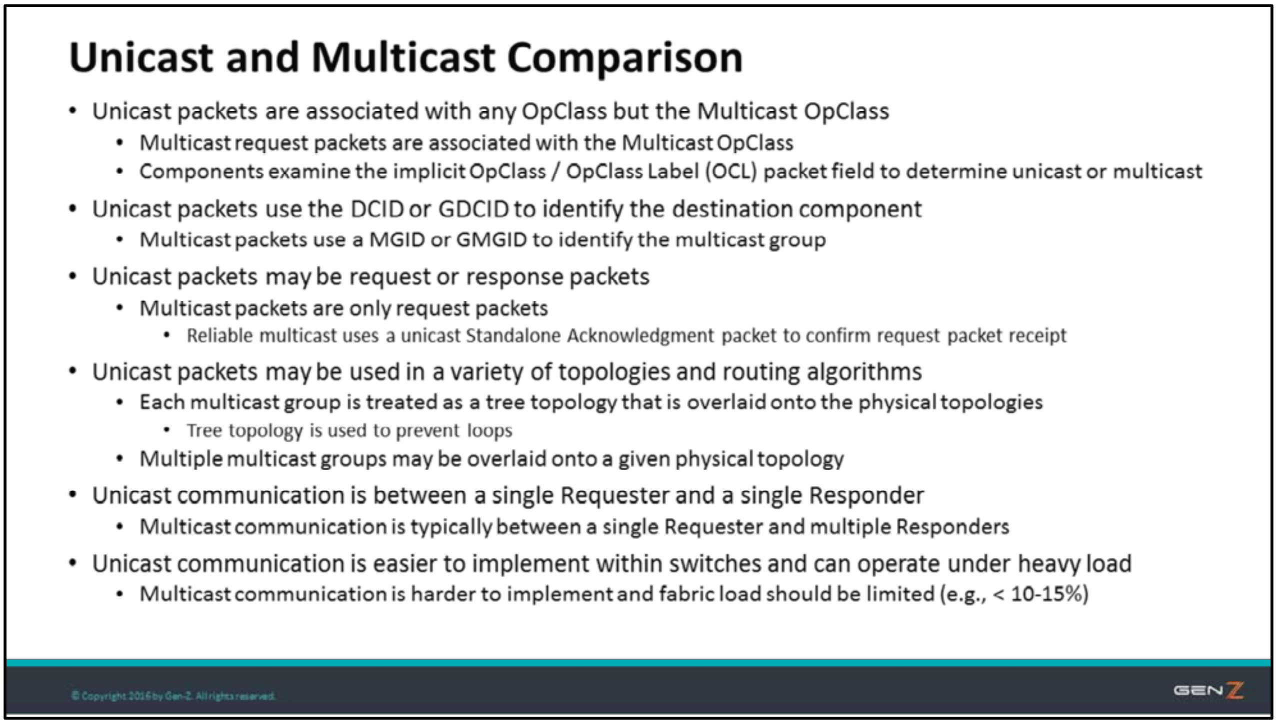
Reliable Multicast

- Two types of multicast—unreliable and reliable
- Currently, there are three multicast request operation types: Write, Write P (only reliable multicast) and Write MSG

GEN Z

Gen-Z supports two types of multicast—unreliable and reliable.

- An unreliable multicast group does not guarantee end-to-end reliability. A Requester is unaware of packet discards due to a transient error, component failure, packet deadline expiration, etc. A Requester does not know which components are participating in an unreliable multicast group; this enables components to join or leave the multicast group without disrupting communications. Unreliable multicast is the most common form of multicast. It is used extensively throughout the industry, e.g., IP multicast is used to stream content, interact with management services, etc.
- A reliable multicast group supports Reliable Delivery, i.e., packets are received in the order transmitted and acknowledged by each participating Responder. To ensure reliability, a Requester must know all participating Responders in order to track acknowledgments for each outstanding request. Reliable multicast can be used to synchronize components, reliably replicate data across multiple components, etc.



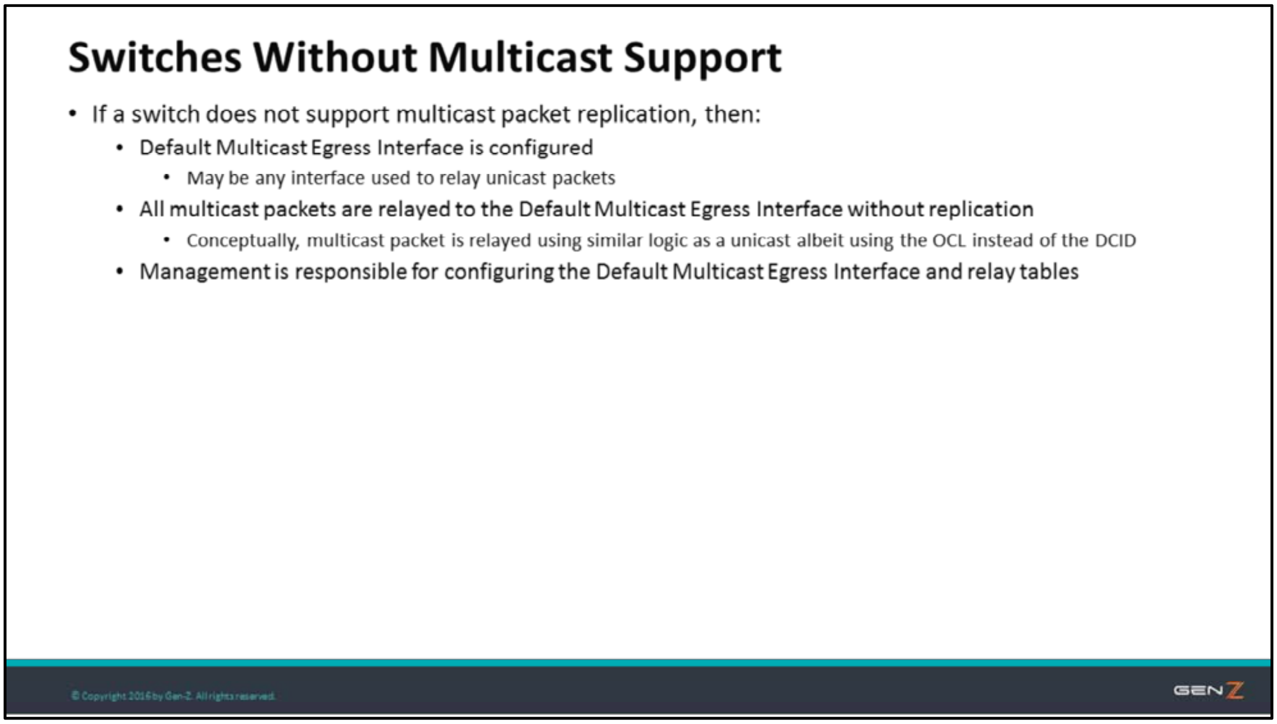
Unicast and Multicast Comparison

- Unicast packets are associated with any OpClass but the Multicast OpClass
  - Multicast request packets are associated with the Multicast OpClass
  - Components examine the implicit OpClass / OpClass Label (OCL) packet field to determine unicast or multicast
- Unicast packets use the DCID or GDCID to identify the destination component
  - Multicast packets use a MGID or GMGID to identify the multicast group
- Unicast packets may be request or response packets
  - Multicast packets are only request packets
    - Reliable multicast uses a unicast Standalone Acknowledgment packet to confirm request packet receipt
- Unicast packets may be used in a variety of topologies and routing algorithms
  - Each multicast group is treated as a tree topology that is overlaid onto the physical topologies
    - Tree topology is used to prevent loops
  - Multiple multicast groups may be overlaid onto a given physical topology
- Unicast communication is between a single Requester and a single Responder
  - Multicast communication is typically between a single Requester and multiple Responders
- Unicast communication is easier to implement within switches and can operate under heavy load
  - Multicast communication is harder to implement and fabric load should be limited (e.g., < 10-15%)

A multicast packet is an explicit OpClass packet with OCL = Multicast.

A unicast packet is a point-to-point-optimized or an explicit OpClass packet with OCL ≠ Multicast.

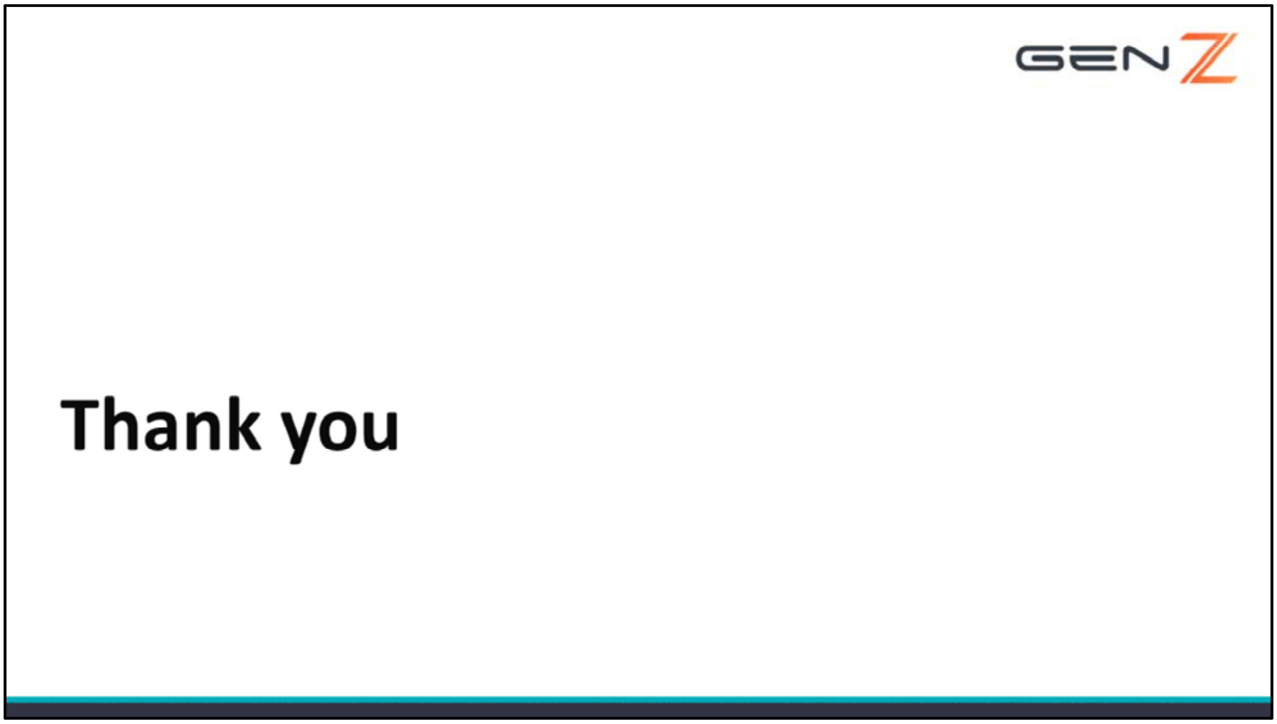
Unicast packets use CID / SID to identify the peer component. Multicast packets use MGID / GMGID to identify the multicast group.



Switches Without Multicast Support

- If a switch does not support multicast packet replication, then:
  - Default Multicast Egress Interface is configured
    - May be any interface used to relay unicast packets
  - All multicast packets are relayed to the Default Multicast Egress Interface without replication
    - Conceptually, multicast packet is relayed using similar logic as a unicast albeit using the OCL instead of the DCID
  - Management is responsible for configuring the Default Multicast Egress Interface and relay tables

All switches are required to support the Default Multicast Egress Interface. If multicast is not supported or is not enabled, then management configures the Default Multicast Egress Interface to relay multicast packets through a single egress interface. This enables any switch to support multicast applications even if it does not support multicast packet replication. For example, numerous in-band management services rely on multicast communications to simplify topology management, service location, support existing higher-level management services (e.g., DHCP, DNS, etc.), and so forth. Similarly, the Default Multicast Egress Interface enables simple switch components (e.g., an expansion switch) or components with integrated switches (e.g., a SoC or a memory component) to support multicast services if attached to a discrete switch (e.g., a top-of-rack switch) that supports multicast or to a multicast appliance that performs packet replication on behalf of a multicast group and transmits unicast packets to each participating Responder (e.g., a network gateway could also acts as a proxy for multicast groups).



Thank you

This concludes this presentation. Thank you.