This presentation covers Gen-Z packet relay, specifically, switch-based packet relay.
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/
TR packet relay is covered in the Transparent Router tutorial.

This presentation is focused on switch-based packet relay. Explicit OpClass packets contain a DCID used to relay packets within a single subnet. Explicit OpClass packets can contain a DSID used to relay packets between subnets.

Using Gen-Z control structures, Gen-Z is capable of supporting nearly any routing topology: fat-tree, 2D / 3D torus, butterfly, hyper-X, etc.
Packet relay is configured through a set of tables that are accessed using packet-specific fields. As you review these tables, be aware that these tables have been designed such that they can support multiple routing algorithms without requiring routing algorithm-specific custom hardware. This is critical to fully enable a software-defined ecosystem, and unlock innovation.

These tables will be described in detail in the following slides.
LPRT / MPRT tables use the same table format. Each table consists of a set of rows.
• The LPRT is directly accessed using the packet’s DCID.
• The MPRT is directly accessed using the packet’s DSID.

Each row contains set of one or more route entries (enables multipath / adaptive routing). Each route entry contains:
• Valid bit indicates if the route is valid. Routes can become valid / invalid based on connectivity to the destination through a given route.
• Hop Count indicates the number of link hops to reach the destination through a given interface. This can be the actual or relative number of hops. This provides the switch a sense of distance to reach a destination through a given egress interface.
• VC Action Table entry is used to identify the set of VCs that can be used if a given route is taken. This is used to enable VC remapping.
• Egress Interface Identifier indicates the egress interface is used to relay the packet.
The VC Mask is used by VC remapping logic to remap the packet’s VC. The VC Mask is configured by management software to reflect the set of enabled VCs on the egress interface associated with a given route. A switch applies implementation-specific logic to select a VC, e.g., a switch might consider the number of available flow-control credits, VC arbitration state, etc.

Each row contains a threshold value. The threshold is used by some routing algorithms as an input to select the egress interface.

* Requester VCAT contains N rows that are indexed by the LPR / MPRT route entry’s VCA field
  * Number of rows equals the maximum number of supported VCs on any switch interface
  * Each row contains two fields:
    * VCM—VC Mask is a bit mask that indicates which VCs may be used to reach the destination
    * TH—Threshold is a hop-count threshold comparison value to use when evaluating a LPR / MPRT route entry and using the hop count as an input into egress interface selection
As with the unicast tables, the single-subnet and multi-subnet multicast tables share the same format. Each table contains a set of rows that identify the egress interfaces and associated VC action to take if a multicast packet is to be replicated and transmitted on the interface for a given multicast group.

- Single-subnet multicast uses the packet’s MGID field to directly index the MCPRT.
- Multi-subnet multicast applies a component-specific function to the GMGID to index the MSMCPRT.

```plaintext
<table>
<thead>
<tr>
<th>MGID</th>
<th>Interface N</th>
<th>Interface 1</th>
<th>Interface 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>MVCA, V</td>
<td>MVCA, V</td>
<td>MVCA, V</td>
</tr>
<tr>
<td>1</td>
<td>MVCA, V</td>
<td>MVCA, V</td>
<td>MVCA, V</td>
</tr>
<tr>
<td>K</td>
<td>MVCA, V</td>
<td>MVCA, V</td>
<td>MVCA, V</td>
</tr>
</tbody>
</table>
```

As with the unicast tables, the single-subnet and multi-subnet multicast tables share the same format. Each table contains a set of rows that identify the egress interfaces and associated VC action to take if a multicast packet is to be replicated and transmitted on a given interface. The V (valid) bit indicates if the egress interface is used to transmit multicast packets for the corresponding multicast group. The MVCA enables VC remapping to be performed based on the packet’s VC.
Similar to the VCAT, the MVCAT is used to identify the VC to use with a given egress interface. If multiple VCs are configured, then the switch applies an implementation-specific policy to select a VC. The policy can be the same as used for unicast VC remapping.
Single-Subnet – Single-Route Packet Relay

- The switch accesses the ingress interface’s LPRT, and uses the packet’s DCID to directly index and locate the corresponding LPRT route entry row.
- Since the switch supports a single route entry, the following applies:
  - MHC, HC, and VCA fields shall be ignored.
  - VC remapping shall not be performed.
  - If V == 1b, then the packet is relayed to the EI (egress interface), else it is handled as a UP error.
Multi-Subnet – Single Route Packet Relay

- In parallel, the switch accesses the ingress interface’s LPRT and MPRT.
  - The switch uses the packet’s DCID field to directly index and locate the corresponding LPRT route entry row.
  - If the GC field is present in the packet and GC = 1b, then the switch uses the packet’s DSID field to directly index and locate the corresponding MPRT route entry row.

- Since the switch supports a single route entry, the following applies to the LPRT and MPRT route entry rows:
  - MIIIC, IIC, and VCA fields shall be ignored.
  - VC remapping shall not be performed.
  - If V = 1b, then the route entry can be used to determine the egress interface, else the route entry is ignored.

- If GC field is present and GC = 1b and if the local SID does not equal the destination SID, then the MPRT route entry is used to relay the packet, else the LPRT route entry is used. If V = 0b, then the packet is handled as a UP error.
  - This step may be performed prior to accessing the LPRT or MPRT, in which case only one table is accessed to determine the route and egress interface.
Single-Route VC Remapping

- In parallel, the switch accesses the ingress interface’s LPR and MPRT.
  - The switch uses the packet’s DCID field to directly index and locate the corresponding LPR route entry row.
  - If the GC field is present in the packet and GC = 1b, then the switch uses the packet’s LSD field to directly index and locate the corresponding MPRT route entry row.
- Since the switch supports a single route entry (Max Routes = 1), the following applies to the LPR and MPRT route entry rows:
  - If V = 1b, then the route entry can be used to determine the egress interface, else the route entry is ignored.
  - The switch uses the VCA route entry to access the VCAT table and derive a VCM (VC mask).
- If GC field is present and GC = 1b and if the local SID does not equal the destination SID, then the MPRT route entry is used to relay the packet, else the LPR route entry is used. If V = 0b, then the packet is treated as a UP error.
  - This step may be performed prior to accessing the LPR or MPRT, in which case only one table is accessed to determine the route and egress interface.
- Prior to relaying the packet, the switch uses the VCM to perform VC remapping:
  - If the packet’s VC is a valid output VC (Bit_VC = 1b), then the VC does not need to be remapped.
  - If (Bit_VC = 0b), then any VC where Bit_VC = 1b may be used to remap the packet’s VC field.
  - If there are multiple valid output VC, then output VC selection may be random, adaptive (e.g., based on congestion information), etc.
Single-Subnet Multipath Routing

• In parallel, the switch accesses the ingress interface’s LPRT and MPRT.
  • The switch uses the packet’s GCID field to directly index and locate the corresponding LPRT route entry row.
  • If the GC field is present in the packet and GC = 1b, then the switch uses the packet’s DSID field to directly index and locate the corresponding MPRT route entry row.

• Since the switch supports multiple route entries, the following applies to the LPRT and MPRT route entry rows:
  • The switch uses random or adaptive route selection with no regard to the remaining hops to the destination subnet or component, hence, the MHC and HC fields shall be ignored.
  • If V = 1b, then the route entry can be used to determine the egress interface, else the route entry is ignored.
  • The switch uses the VCA route entry to access the VCAT table and derive a VCM (VC mask).

• If GC field is present and GC = 1b and if the local SID does not equal the destination SID, then the switch applies the route selection algorithm to the set of MPRT route entries, else the LPRT route entries are used. If all route entries have V = 0b, then the packet is handled as a UP error.

• Prior to relaying the packet, the switch uses the VCM to perform VC remapping.
Thank you

This concludes this presentation. Thank you.