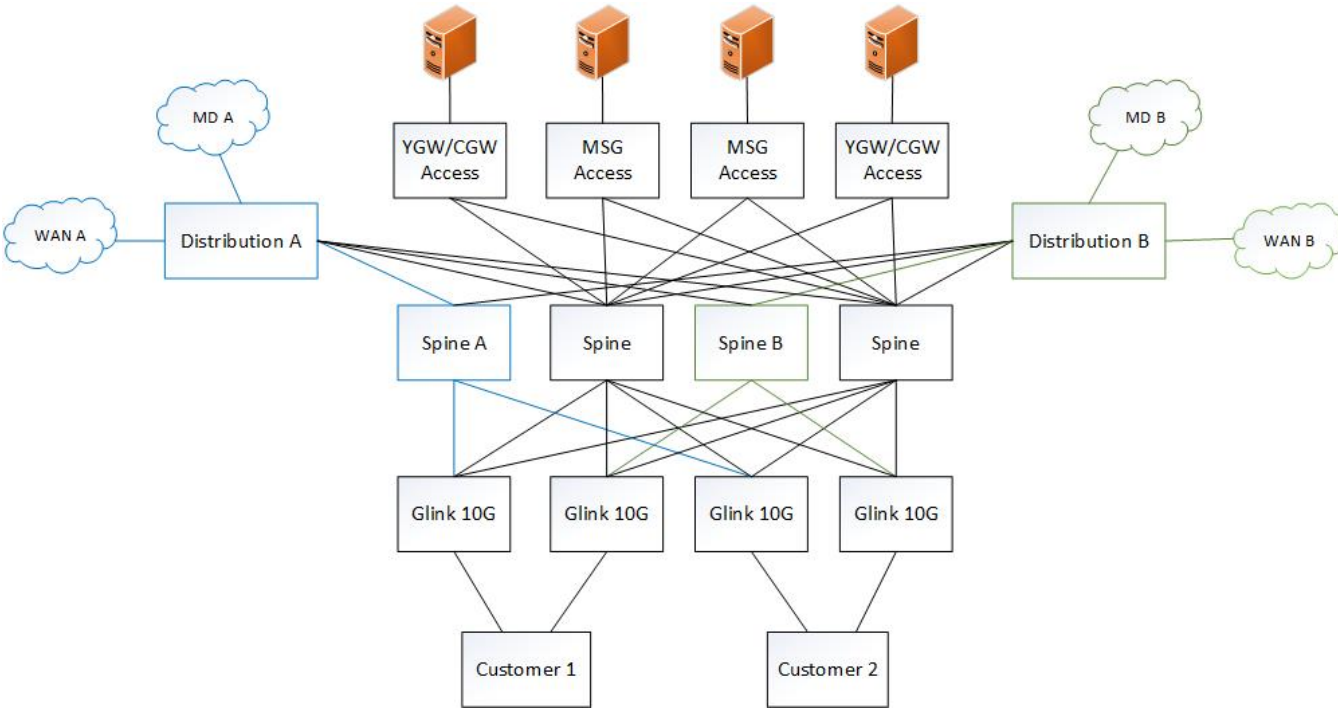
GLink Architecture

GLink architecture is a spine-and-leaf topology designed to provide determinism and Denial of Service (DoS) protections within the CME Globex order routing network.

Topology



Physical (L1)

From bottom of the diagram to top:

- GLink access switches (customer access)
  - Juniper QFX 3500s using 10 Gbps Ethernet
  - Each customer connects to an 'A' switch and a 'B' switch
    - All GLink switches are deployed in pairs
  - Each GLink switch is connected to three spines at 40 Gbps. 'A' feed Glink access switches do not connect to the 'B' feed spine, 'B' feed Glink access switches do not connect to the 'A' feed spine
  - Count:
    - 10 pairs of GLink switches today
  - Ports: The first 6 and last 6 ports of the switch have additional chips to support FCoE, and are therefore not used for connectivity due to the performance difference
- Spine switches
  - Juniper QFX 5100s using a mix of 40 Gbps and 10 Gbps (legacy connectivity to distribution/WAN)
  - Spine 'A' and Spine 'B' pass market data multicast traffic
  - Non multicast spines pass order entry (MSG/YGW/CGW) unicast traffic
  - Spine 'A' and Spine 'B' also pass non order-entry unicast traffic to services outside of Glink. This class of traffic does not affect receipt of market data across multiple customer switches.
  - Count: 4 switches
- MSG access switches
  - Juniper QFX 3500s using 10 Gbps Ethernet for gateway connectivity
  - Each MSG access switch is connected to the 2 non-multicast Spine switches at 40 Gbps
  - MSGs connect to only one gateway access switch at 10 Gbps. Fault tolerant pairs should be on separate switches
  - Count: 4 switches
- YGW access switches
  - Juniper QFX 3500s using 10 Gbps Ethernet for gateway connectivity
  - Each YGW access switch is connected to the 2 non-multicast Spine switches at 40 Gbps
  - Count: 2 switches
- WAN distributions (to the left and right of the spines)
  - Switches using 10 Gbps Ethernet connectivity
  - Each WAN distribution is connected to all four spines
  - Market data routes through this distribution layer into the 'A' and 'B' spines

- Count: 2 switches

Data Link (L2)

- 10 GbE interfaces are supported
- All customer connected interfaces have policing applied which reduces available bandwidth to 1 Gbps (covered in more detail below)
- We do not use VLANs within the GLink front end network
    - All nodes, including servers, have routable L3 addresses
    - We do not run Spanning Tree Protocol (STP) nor any variants of it
- Although we're using the Juniper QFX platform, we are not using QFabric, which is a proprietary Juniper technology
- All switching layers (customer, spine and gateway) operate in 'store-and-forward' mode
    - Store-and-forward mode means that, for any given switch, it must completely receive a datagram before it will transmit that datagram to another interface
        - Implication: If a switch starts receiving two separate datagrams at exactly the same time and both datagrams are destined to leave the same port, then the smaller of the two datagrams will leave the switch first

Network (L3)

Active-Standby Routing and Paths

Each MSG server is 'available' via 2 paths from the 2 non-multicast spine switches, in an active-standby manner.

When a session arrives at a customer access switch, the outbound path will be selected based on which of the 2 non-multicast spines is active for a given MGW access switch. The return path traffic (MSG to customer) would follow the same path in a symmetric fashion.

- Ordering/re-ordering of packets
    - Packets can be reordered within the spine layer provided that
        - two different sessions are used
        - the load from customer access to spine or within the individual spines are different
    - The ordering of packets from all sessions to MSG will be 'final' at the server access switching layer (egress traffic to the server - 'last hop' switch)
    - Latency difference between spines will vary depending upon load but is expected to stay within the hundreds of nanos range

- Traffic can be routed over the 'A' and 'B' simultaneously.

    - On 'A' advertise via BGP the summary route for the CME supplied address range.
    - On 'B' advertise via BGP a more specific route for the traffic you want returned on the 'B' interface.
    - If a more specific route is not advertised, then all traffic will return to the summary interface (asymmetric routing) - which will break Network Address Translation if it's in use.

- The path for a particular GLink connection to a specific market segment is the same for all messages

Performance

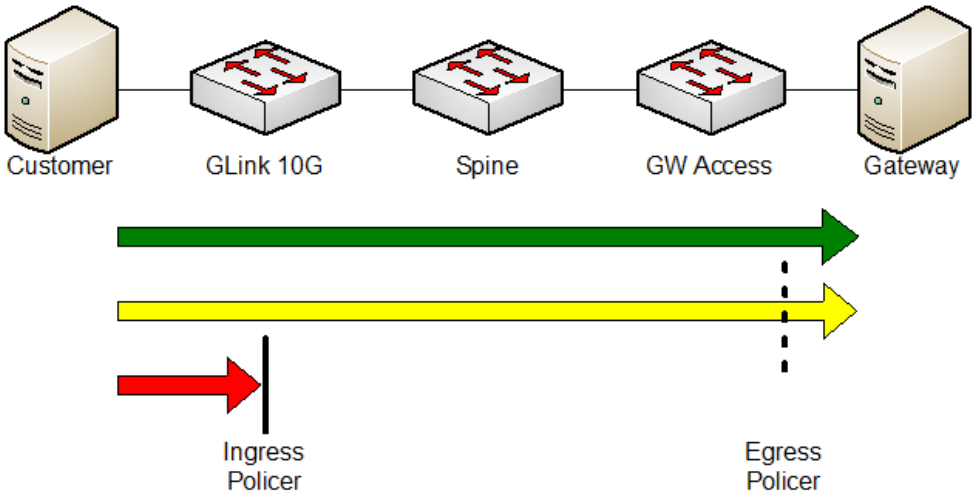
- Basic Behavior
    - Reduce network jitter by dispersing microbursts through the customer-side fabric
    - Queuing delays will be associated with the order routing gateways

- Pertinent Info
    - Nominal 1-way latency of 3 microseconds (slightly less than 1 us forwarding latency per switch)

    - Oversubscription - GLink customers per switch
        - 0.75:1 – All customers to single Spine (worst case)
        - 0.19:1 – All customers to all Spines (best case)

    - Oversubscription - Spines
        - 12:1 – All customers GLink switching to a single spine.
    - There will be differences in performance between any two switches based upon standard networking principles.
    - The QFX is built around the Broadcom Trident chip family and is a "Switch on Chip" (SOC), shared memory design. Queue depth monitoring is generally more prevalent with multi-stage switching designs where buffers are separate physical entities on a per port basis.

Policing Overview



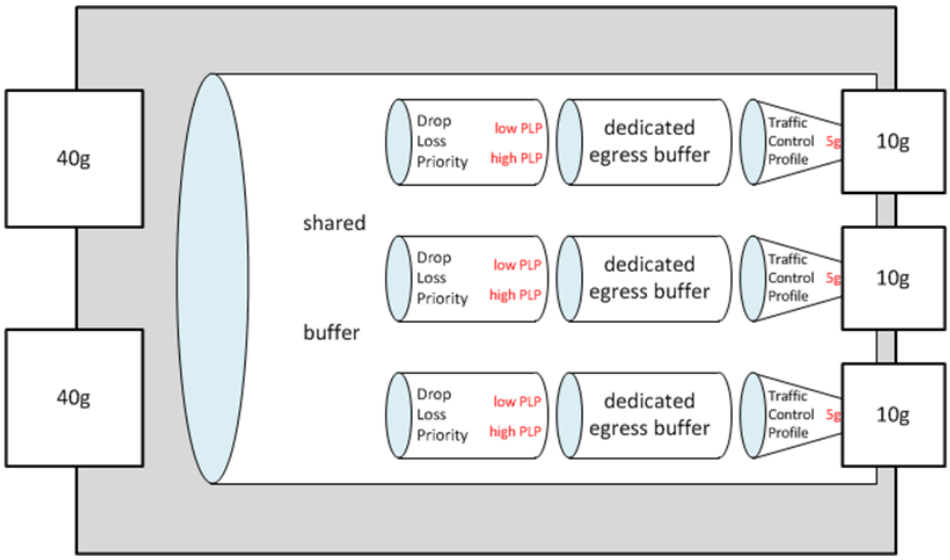
- At Ingress - per GLink:
  - Green: < 750 Mbps is allowed and marked ‘normal’ (AF11)
  - Yellow: 750 Mbps – 1 Gbps is allowed and marked ‘discard eligible’ (AF12)
  - Red: > 1 Gbps is silently dropped

- At Egress - per MSG:
  - Green: At X rate, all traffic permitted through to the gateways
  - Yellow: At > X, AF11 traffic will be allowed and AF12 traffic dropped

Ingress Policing

- General Mechanics
  - ‘Credit’ value similar to token bucket
  - Two-rate three-color marker (RFC 2698)
- Metering Calculation
  - Obtain current credits: CCURR = CPREV + (TDELTA * RCREDIT)
  - Check current against the limit: If CCURR > CLIMIT, then CCURR == CLIMIT
  - Calculate the eventual credits: CPOST = CCURR – PSIZE
  - Check whether packet will be policed
- Marking and Action
  - Committed Information Rate (CIR) = 750 M
    - Action: Mark (AF12)
- Committed Burst Size (CBS) = 500 K
- Peak Information Rate (PIR) = 1 G
  - Action: Drop
- Peak Burst Size (PBS) = 625 K
- Conforming traffic under CIR is marked as AF11

Egress Policing



- Switch Basics
  - QFX 3500 has 9 MB buffers split into dedicated and shared buffer space
  - Buffer is cell based with 208 bytes per cell
- General Mechanics
  - On ingress, map incoming traffic to forwarding classes
  - Forwarding classes are mapped to output queues
  - Generally, BW and packet drop characteristics mapped to output queues
    - First scheduling hierarchy
    - only one queue in our implementation so skipping the scheduling
    - This is where we make an AF11/AF12 decision
  - Output queues are mapped to forwarding class sets (priority groups)
    - Second scheduling hierarchy
  - Priority groups mapped to egress ports
- Implementation
  - Egress ports are being configured with a single queue to preserve packet ordering
  - Drop Loss Priority
    - Low Packet Loss Priority (PLP) – AF11, no drop profile defined
    - Med-High Packet Loss Priority (PLP) – AF12, drop profile defined
  - Traffic Control Profile determines the rate at which egress buffers are drained
    - TCP is separate from DLP