Understanding CoS Virtual Output Queues (VOQs) on QFX10000 Switches

Link to original: http://www.juniper.net/documentation/en_US/junos15.1/topics/concept/cos-qfx-series-voq-understanding.html

The traditional method of forwarding traffic through a switch is based on buffering ingress traffic in input queues on ingress interfaces, forwarding the traffic across the switch fabric to output queues on egress interfaces, and then buffering traffic again on the output queues before transmitting the traffic to the next hop. The traditional method of queueing packets on an ingress port is storing traffic destined for different egress ports in the same input queue (buffer).

During periods of congestion, the switch might drop packets at the egress port, so the switch might spend resources transporting traffic across the switch fabric to an egress port, only to drop that traffic instead of forwarding it. And because input queues store traffic destined for different egress ports, congestion on one egress port could affect traffic on a different egress port, a condition called *head-of-line blocking (HOLB)*.

Virtual output queue (VOQ) architecture takes a different approach:

- Instead of separate physical buffers for input and output queues, the switch uses the physical buffers on the ingress pipeline of each Packet Forwarding Engine (PFE) chip to store traffic for every egress port. Every output queue on an egress port has buffer storage space on every ingress pipeline on all of the PFE chips on the switch. The mapping of ingress pipeline storage space to output queues is 1-to-1, so each output queue receives buffer space on each ingress pipeline.
- Instead of one input queue containing traffic destined for multiple different output queues (a one-to-many mapping), each output queue has a dedicated VOQ comprised of the input buffers on each packet forwarding chip that are dedicated to that output queue (a 1-to-1 mapping). This architecture prevents communication between any two ports from affecting another port.
- Instead of storing traffic on a physical output queue until it can be forwarded, a VOQ does not transmit traffic from the ingress port across the fabric to the egress port until the egress port has the resources to forward the traffic.

A VOQ is a collection of input queues (buffers) that receive and store traffic destined for one output queue on one egress port. Each output queue on each egress port has its own dedicated VOQ, which consists of all of the input queues that are sending traffic to that output queue.

VOQ Architecture

A VOQ represents the ingress buffering for a particular output queue. A unique buffer ID identifies each output queue on a PFE chip. Each of the six PFE chips uses the same unique buffer ID for a particular output queue. The traffic stored using a particular buffer ID on the six PFE chips comprises the traffic destined for one particular output queue on one port, and is the VOQ for that output queue.

A switch that has 72 egress ports with 8 output queues on each port, has 576 VOQs on each PFE chip `(72 x 8 = 576)`. Because the switch has six PFE chips, the switch has a total of 3,456 VOQs `(576 x 6 = 3,456)`.

A VOQ is distributed across all of the PFE chips that are actively sending traffic to that output queue. Each output queue is the sum of the total buffers assigned to that output queue (by its

unique buffer ID) across all of the PFE chips. So the output queue itself is virtual, not physical, although the output queue is comprised of physical input queues.

Round-Trip Time Buffering

Although there is no output queue buffering during periods of congestion (no long-term storage), there is a small physical output queue buffer on egress line cards to accommodate the round-trip time for traffic to traverse the switch fabric from ingress to egress. The round-trip time consists of the time it takes the ingress port to request egress port resources, receive a grant from the egress port for resources, and transmit the data across the switch fabric.

That means if a packet is not dropped at the switch ingress, and the switch forwards the packet across the fabric to the egress port, the packet will not be dropped and will be forwarded to the next hop. All packet drops take place in the ingress pipeline.

The switch has 4 GB of external DRAM to use as a delay bandwidth buffer (DBB). The DBB provides storage for ingress ports until the ports can forward traffic to egress ports.

Requesting and Granting Egress Port Bandwidth

When packets arrive at an ingress port, the ingress pipeline stores the packet in the ingress queue with the unique buffer ID of the destination output queue. The switch makes the buffering decision after performing the packet lookup. If the packet belongs to a class for which the maximum traffic threshold has been exceeded, the packet might not be buffered and might be dropped. To transport packets across the switch fabric to egress ports:

1. The ingress line card PFE request scheduler sends a request to the egress line card PFE grant scheduler to notify the egress PFE that data is available for transmission.
2. When there is available egress bandwidth, the egress line card grant scheduler responds by sending a bandwidth grant to the ingress line card PFE.
3. The ingress line card PFE receives the grant from the egress line card PFE, and transmits the data to the egress line card.

Ingress packets remain in the VOQ on the ingress port input queues until the output queue is ready to accept and forward more traffic.

Under most conditions, the switch fabric is fast enough to be transparent to egress class-of-service (CoS) policies, so the process of forwarding traffic from the ingress pipeline, across the switch fabric, to egress ports, does not affect the configured CoS policies for the traffic. The fabric only affects CoS policy if there is a fabric failure or if there is an issue of port fairness.

When a packet ingresses and egresses the same PFE chip (local switching), the packet does not traverse the switch fabric. However, the switch uses the same request and grant mechanism to receive egress bandwidth as packets that cross the fabric, so locally switched packets and packets that arrive at a PFE chip after crossing the switch fabric are treated fairly when the traffic is contending for the same output queue.

VOQ Advantages

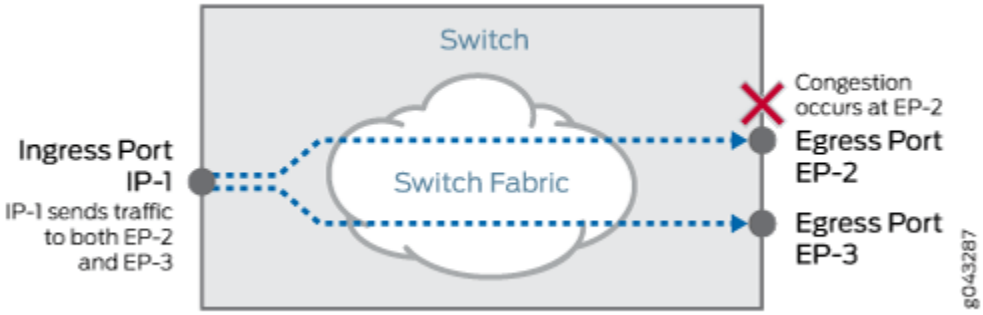
VOQ architecture provides two major advantages:

Eliminate Head-of-Line Blocking

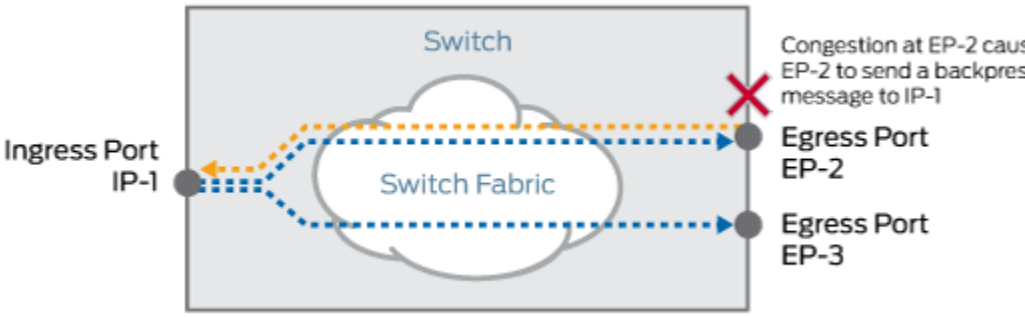
VOQ architecture eliminates head-of-line blocking (HOLB) issues. On non-VOQ switches, HOLB occurs when congestion at an egress port affects a different egress port that is not congested. HOLB occurs when the congested port and the uncongested port share the same input queue on an ingress interface.

An example of a HOLB scenario is a switch that has streams of traffic entering one ingress port (IP-1) that are destined for two different egress ports (EP-2 and EP-3):

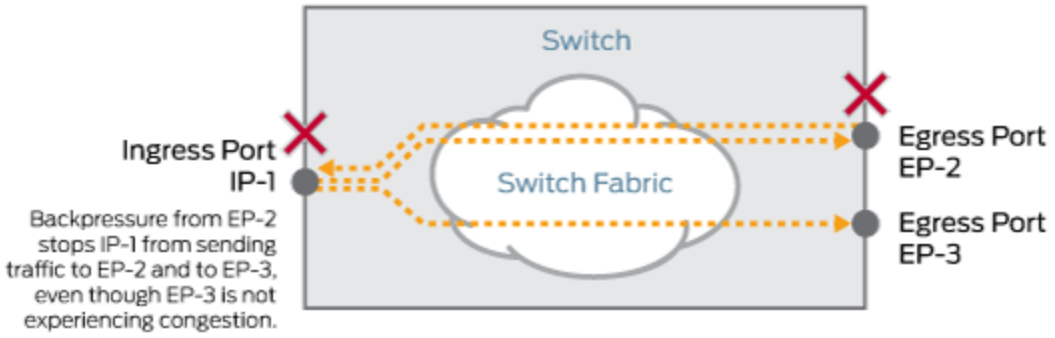
1. Congestion occurs on egress port EP-2. There is no congestion on egress port EP-3, as shown in Figure 1.
   Figure 1: Congestion Occurs on EP-2



2. Egress port EP-2 sends a backpressure signal to ingress port IP-1, as shown in Figure 2.
   Figure 2: EP-2 Backpressures IP-1

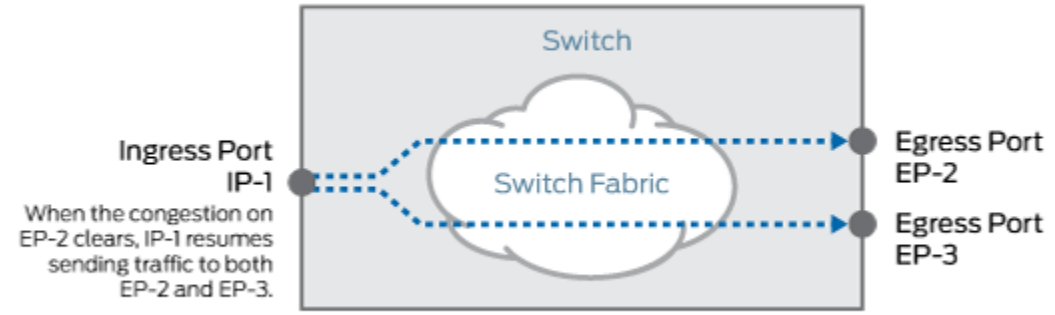


3. The backpressure signal causes the ingress port IP-1 to stop sending traffic and to buffer traffic until it receives a signal to resume sending, as shown in Figure 3. Traffic that arrives at ingress port IP-1 destined for uncongested egress port EP-3 is buffered along with the traffic destined for congested port EP-2, instead of being forwarded to port EP-3.
   Figure 3: Backpressure from EP-2 Causes IP-1 to Buffer Traffic Instead of Sending Traffic, Affecting EP-3



4. Ingress port IP-1 transmits traffic to uncongested egress port EP-3 only when egress port EP-2 clears enough to allow ingress port IP-1 to resume sending traffic, as shown in Figure 4.

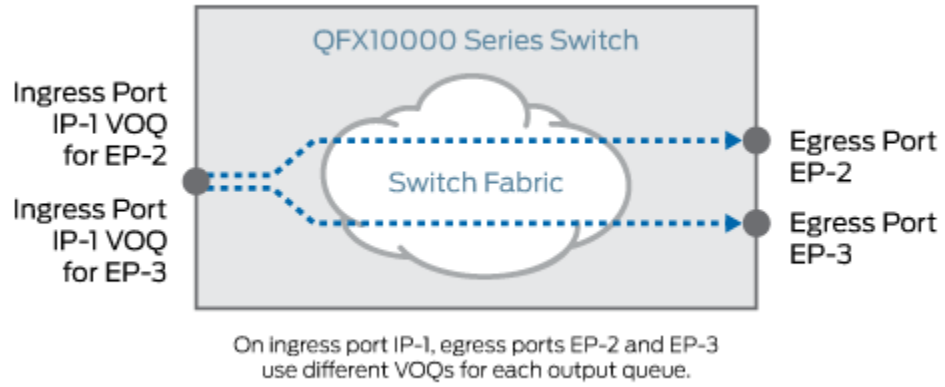
Figure 4: Congestion on EP-2 Clears, Allowing IP-1 to Resume Sending Traffic to Both Egress Ports



In this way, congested egress port EP-2 negatively affects uncongested egress port EP-3, because both egress ports share the same input queue on ingress port IP-1.

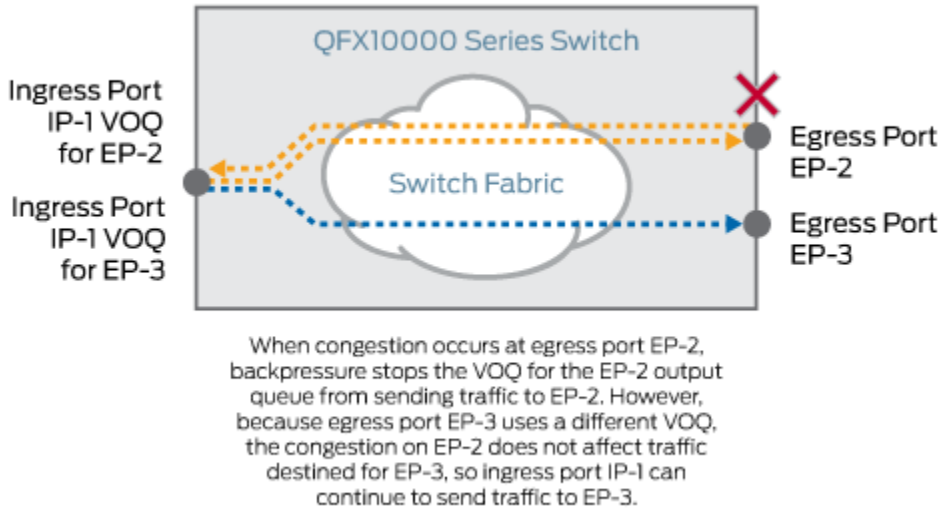
VOQ architecture avoids HOLB by creating a different dedicated virtual queue for each output queue on each interface, as shown in Figure 5.

Figure 5: Each Egress Port Has a Separate Virtual Output Queue on IP-1



Because different egress queues do not share the same input queue, a congested egress queue on one port cannot affect an egress queue on a different port, as shown in Figure 6. (For the same reason, a congested egress queue on one port cannot affect another egress queue on the same port—each output queue has its own dedicated virtual output queue composed of ingress interface input queues.)

Figure 6: Congestion on EP-2 Does Not Affect Uncongested Port EP-3



Performing queue buffering at the ingress interface ensures that the switch only sends traffic across the fabric to an egress queue if that egress queue is ready to receive that traffic. If the egress queue is not ready to receive traffic, the traffic remains buffered at the ingress interface.

Increase Fabric Efficiency and Utilization

Traditional output queue architecture has some inherent inefficiencies that VOQ architecture addresses.

- Packet buffering—Traditional queueing architecture buffers each packet twice in long-term DRAM storage, once at the ingress interface and once at the egress interface. VOQ architecture buffers each packet only once in long-term DRAM storage, at the ingress interface. The switch fabric is fast enough to be transparent to egress CoS policies, so instead of buffering packets a second time at the egress interface, the switch can forward traffic at a rate that does not require deep egress buffers, without affecting the configured egress CoS policies (scheduling).
- Consumption of resources—Traditional queueing architecture sends packets from the ingress interface input queue (buffer), across the switch fabric, to the egress interface output queue (buffer). At the egress interface, packets might be dropped, even though the switch has expended resources transporting the packets across the fabric and storing them in the egress queue. VOQ architecture does not send packets across the fabric to the egress interface until the egress interface is ready to transmit the traffic. This increases system utilization because no resources are wasted transporting and storing packets that are dropped later.

Independent of VOQ architecture, the Juniper Networks switching architecture also provides better fabric utilization because the switch converts packets into cells. Cells have a predictable size, which enables the switch to spray the cells evenly across the fabric links and more fully utilize the fabric links. Packets vary greatly in size, and packet size is not predictable. Packet-based fabrics can deliver no better than 65-70 percent utilization because of the variation and unpredictability of packet sizes. Juniper Networks' cell-based fabrics can deliver a fabric utilization rate of almost 95 percent because of the predictability of and control over cell size.

Modified: 2016-04-29