Western Digital.

WHITE PAPER JANUARY 2024

Lossless Networking Technology Overview

Abstract

This document provides an overview of lossless Ethernet technology concepts and how they apply to NVMe-oF™ storage systems.

Notices

Western Digital® Technologies, Inc. or its affiliates' (collectively "Western Digital") general policy does not recommend the use of its products in life support applications wherein a failure or malfunction of the product may directly threaten life or injury. Per Western Digital Terms and Conditions of Sale, the user of Western Digital products in life support applications assumes all risk of such use and indemnifies Western Digital against all damages.

This document is for information use only and is subject to change without prior notice. Western Digital assumes no responsibility for any errors that may appear in this document, nor for incidental or consequential damages resulting from the furnishing, performance or use of this material.

Absent a written agreement signed by Western Digital or its authorized representative to the contrary, Western Digital explicitly disclaims any express and implied warranties and indemnities of any kind that may, or could, be associated with this document and related material, and any user of this document or related material agrees to such disclaimer as a precondition to receipt and usage hereof.

Each user of this document or any product referred to herein expressly waives all guaranties and warranties of any kind associated with this document any related materials or such product, whether expressed or implied, including without limitation, any implied warranty of merchantability or fitness for a particular purpose or non-infringement. Each user of this document or any product referred to herein also expressly agrees Western Digital shall not be liable for any incidental, punitive, indirect, special, or consequential damages, including without limitation physical injury or death, property damage, lost data, loss of profits or costs of procurement of substitute goods, technology, or services, arising out of or related to this document, any related materials or any product referred to herein, regardless of whether such damages are based on tort, warranty, contract, or any other legal theory, even if advised of the possibility of such damages.

This document and its contents, including diagrams, schematics, methodology, work product, and intellectual property rights described in, associated with, or implied by this document, are the sole and exclusive property of Western Digital. No intellectual property license, express or implied, is granted by Western Digital associated with the document recipient's receipt, access and/or use of this document or the products referred to herein; Western Digital retains all rights hereto.

Western Digital, the Western Digital logo, and OpenFlex are registered trademarks or trademarks of Western Digital Corporation or its affiliates in the US and/or other countries. The NVMe and NVMe-oF word marks are trademarks of NVM Express, Inc. All other marks are the property of their respective owners. Product specifications subject to change without notice. Pictures shown may vary from actual products. Not all products are available in all regions of the world.

Western Digital
5601 Great Oaks Parkway
San Jose, CA 95119

Table of Contents

Introduction

This paper describes the concept of Lossless Networking as it relates to the OpenFlex™ family of devices. NVMe™ over RoCE protocol enables remote access of NVMe storage devices over an Ethernet network. The underlying Ethernet network is expected to provide reliable NVMe command and data delivery. Left untuned, traditional Ethernet networks provide best-effort delivery, which is not well-suited for modern data storage traffic. RDMA over Converged Ethernet (RoCE) works best when configured to run over a lossless Ethernet network.

Lossless Networking

Lossless Networking is a category of networking technologies that attempt to greatly reduce (and in some cases completely eliminate) packet loss in IT communications infrastructure. A Lossless Network is achieved through using network devices that support lossless operations and careful configuration of all devices that are responsible for processing data over the network. This includes any storage devices, Network Interface Cards (NIC), and switches that are connected to each other in a datacenter fabric. A Lossless network is required to leverage the full performance potential of the OpenFlex platform.

Strictly speaking, a lossless Ethernet network can still drop packets in certain cases. However, the amount of packet drops is significantly lower than best-effort networks, providing consistent and high performance when carefully selecting and configuring network components.

Why is Lossless Networking required?

Traditional Ethernet is a "Best-Effort" networking protocol. This means that under load packets are dropped and it is up to the transport protocol to cope with that loss (usually via retransmission). This would make Ethernet natively lossy. With packet drops comes retransmissions which reduce the apparent performance of the network (Bandwidth and Packets Per Second) and increases latency on the network.

Why are packets dropped under load?

To understand this let's use an analogy. Let's consider a garden hose. If you connect this hose to your house and it has consistent pressure you would expect a consistent flow rate from the end of your hose. Now consider a firehose. If you connect this firehose to fire hydrant that has consistent pressure you would have and exponentially higher flow rate than that of the garden hose. So, what would happen if you connect the firehose to a garden hose. In the event that the flow direction is from the firehose to garden hose the much higher flow previously experience by the firehose would be severely limited by the garden hose. This is called oversubscription. For water and other fluid dynamics oversubscription presents itself as increased back pressure as well as higher sustained flow rate due to increased pressure (Water Jets). For networking however, pressure does not impact flow rate, back pressure presents in memory queues overflowing and packets being dropped. This is called congestion.

Oversubscription causes Congestion

Network oversubscription is not always easy to identify

In a perfect world, network devices would have a full 1:1 relationship with each other. If a computer has a 10 Gb connection and is connected directly to another computer with a 10 Gb connection, there is no oversubscription. This 1:1 relationship is unrealistic in practice. First not every device will have matching network speeds. Some devices may be connected with 1 Gb while others may be connected at 50 Gb. In the event that the 50 Gb device (Firehose) communicates with the 1 Gb device (Garden Hose) there is oversubscription. Secondly let's consider a network where all devices are connected at 10 Gb. If every device were communicating with every other device simultaneously then it may average out that there would be no oversubscription, but again this is not how it would happen in the real world. In the real world you may have five devices all talking to the same single device. In this case you have five 10 Gb devices talking to a single 10 Gb device. This is effectively a 5:1 oversubscription. Lastly not all networks are simple. With a single network switch, baring switch inefficiencies, the switch would introduce no oversubscription, but with a complex network architecture where multiple switches are involved there will almost certainly be oversubscription introduced by switch to switch networking links.

The Mechanisms of Lossless Networking

Lossless networking consists of three mechanisms:

- Traffic Marking – The means of which to identify certain traffic types.
- Traffic Shaping – The ability to set QoS limits and guarantees onto specific traffic types.
- Congestion/Flow Control – The ability to detect possible congestion and throttle back network pressure to decrease the possibility of packet drops.

Traffic Marking

Traffic Marking is a means of which to identify, categorize, and mark traffic types to allow QoS manipulations of traffic flows.

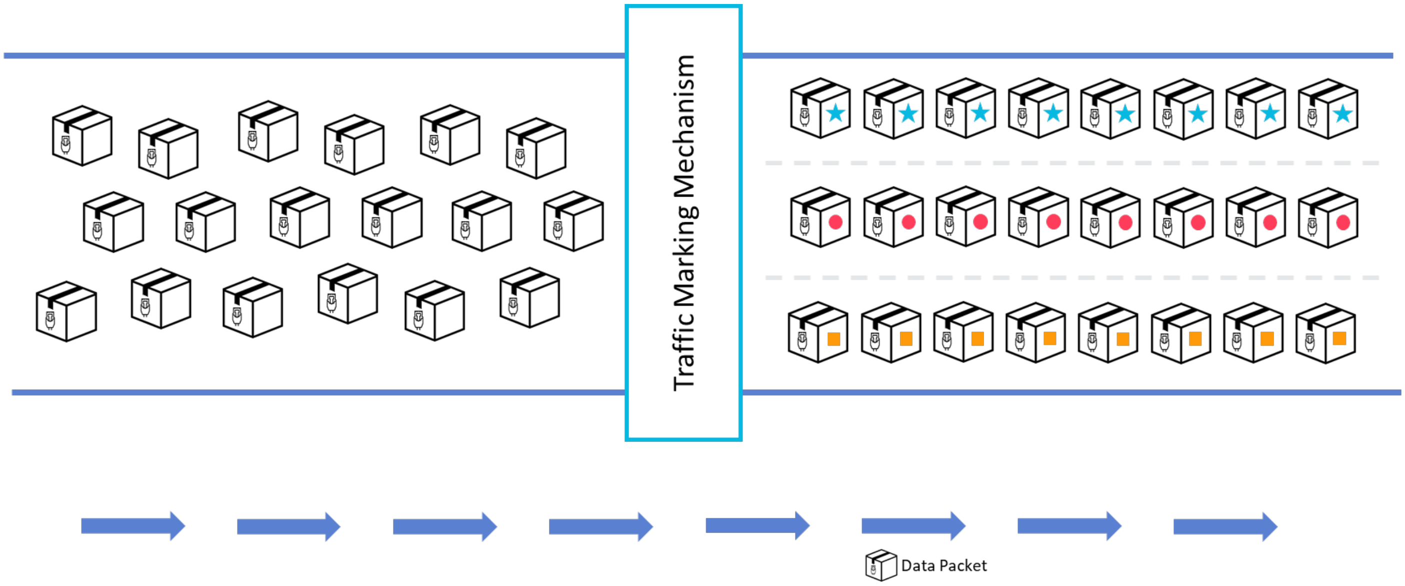


Figure 1 - Traffic Marking Mechanism

This mechanism is not new to the industry. One of the first forays into traffic marking was introduced by the IETF RFC 1349 which defined the Type of Service (ToS) Octet within the IP header. This RFC has since been replace by the IETF RFC 2474 which redefined the ToS Octet to the Differentiated Service (DiffServ) Octet. The redefinition modernized the ToS Octet and allowed for increased compatibility with protocols being standardized by the IEEE Data Center Bridging (DCB) task force.

The DiffServ Octet provides for two Lossless mechanisms. First it allows for a 6-Bit Differentiated Services Code Point (DSCP), which provides the packets Traffic Class (Similar to previous ToS) as well as the Drop Precedence. Secondly it provides the basic functionality for Explicit Congestion Notification (ECN). (Defined Later in this Guide)

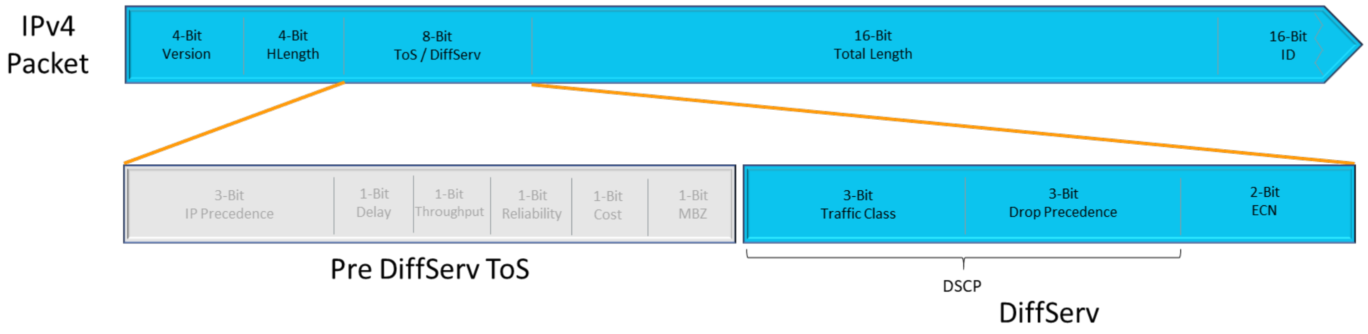


Figure 2 - IPv4 Packet Layout

DSCP can be set and read from anywhere in the network. Generally, we recommend setting the end points to properly configure the DSCP field on pertinent traffic before putting it onto the network.

DSCP is the only supported mechanism for traffic marking used by the OpenFlex devices.

Priority Code Point (PCP)

PCP is another traffic marking mechanism that was used in the early days of DCB. PCP as defined in IEEE 802.1p, a component of 802.1q, uses a 3-bit field in the VLAN tag to convey the Class of Service (CoS). This class of service is similar to the "Traffic Class" provided in DSCP or the "IP Precedence" provided by ToS. All provide 8 Network Priorities (0-7). As VLAN tags are a Layer 2 construct (Ethernet Mechanism) it makes the use of PCP on larger or more complex networks (Layer 3) difficult. Some switches will strip VLAN Tags and re-add them from packets as the enter and leave the switch.

To keep network implementation as simple as possible, PCP is not supported by OpenFlex devices.

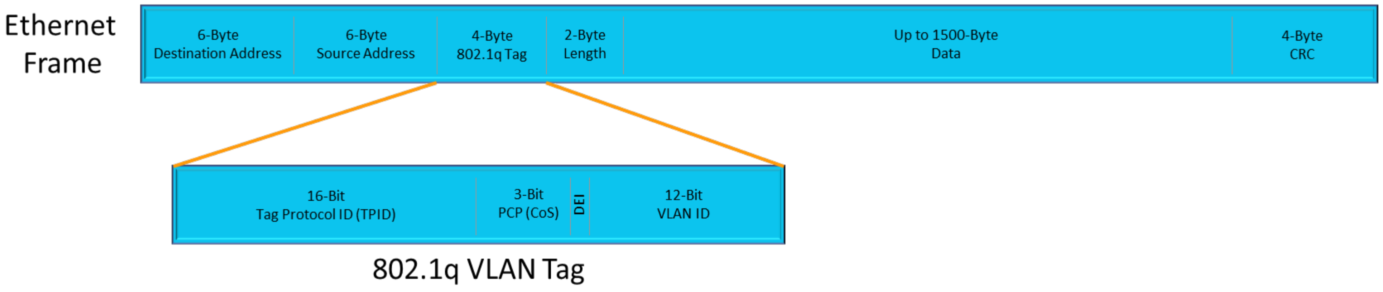


Figure 3 - PCP Field

Traffic Shaping

Traffic Shaping is the ability to set and enforce QoS metrics such as bandwidth limitation onto specific traffic classes. When the network is underload Traffic Shaping delays some packets in favor of other packets.

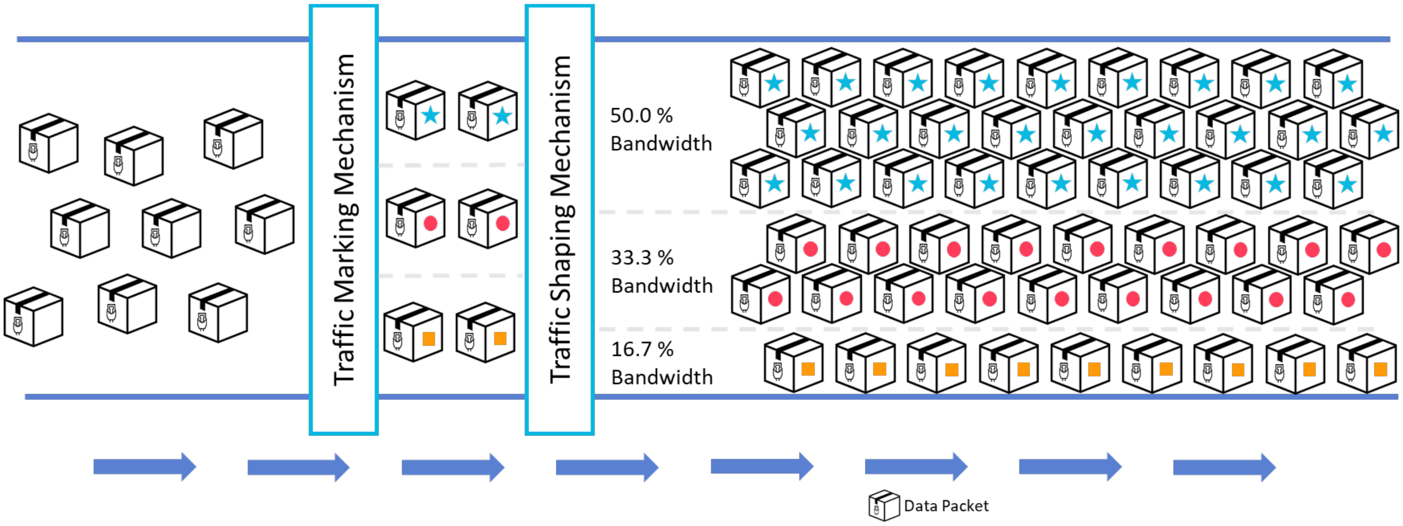


Figure 4 - Traffic Shaping Mechanism

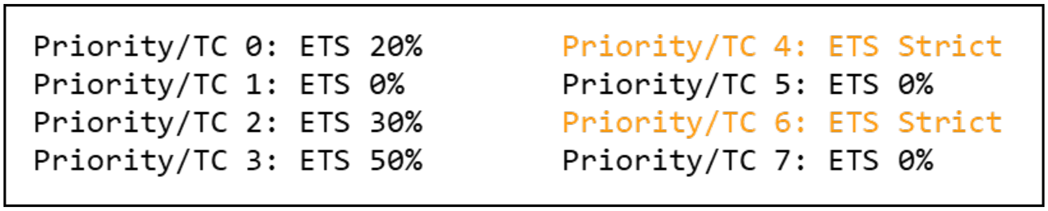
The IEEE 802.1Qaz the DCB taskforce defined Enhanced Transmission Selection (ETS). ETS allows for 8 Traffic Classes set to either Strict or Bandwidth Allocation.

With Strict, Any Traffic Class set to strict is guaranteed up to 100% of the bandwidth unless another Traffic Class of higher value is set to strict.

In example below: If TC 4 is using 60% of the BW and TC 6 rises above 40% of BW, TC 4 will be reduced in favor of TC 6.

With BW Allocation, if a Traffic Class is set to a bandwidth percentage it is guaranteed that percentage at a minimum. Minimum guarantees are only between BW allocated buffers. Strict always has priority over BW and can consume all bandwidth. If the network is not fully utilized any given TC is allowed to exceed its guaranteed BW minimum.

In example below: TC 3 is Guaranteed 50% of the bandwidth but can consume more if the network is not overly utilized.



```
Priority/TC 0: ETS 20%        Priority/TC 4: ETS Strict
Priority/TC 1: ETS 0%         Priority/TC 5: ETS 0%
Priority/TC 2: ETS 30%        Priority/TC 6: ETS Strict
Priority/TC 3: ETS 50%        Priority/TC 7: ETS 0%
```

Figure 5 - Traffic Class Policy Example

Congestion/Flow Control

Flow Control is a mechanism to temporarily stop or slow down traffic to reduce congestion (Packet Drops). Flow Control is also not an unknown concept. IEEE introduced 802.3x which defined Ethernet Flow Control (Global Pause). Global Pause is a simple mechanism that transmits a Pause frame to all transmitting neighboring ports contributing to the buffer overflow of a given port. As Global Pause is Ethernet mechanism (Layer 2) it only understands MAC address and direct neighbors. Global Pause does not rely on Traffic Marking as it simply pauses all traffic from the neighboring ports for a specified duration.

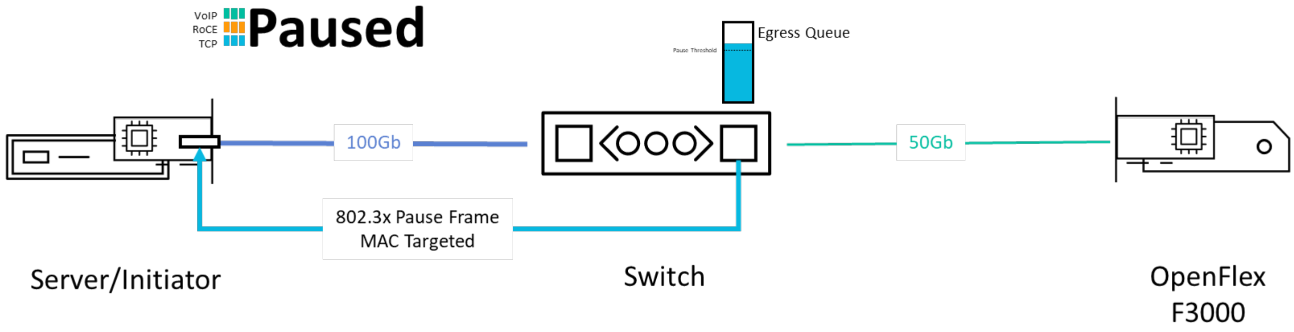


Figure 6 - Global Pause

Priority Flow Control (PFC)

The DCB Task Force introduced IEEE 802.1Qbb which defines Priority Flow Control (PFC) to extend Global Pause by adding a Priority or Traffic Class marking to the Pause Frame. This allows a PFC pause frame to specify an individual Traffic Class to pause while allowing other Traffic Classes to continue to flow without impedance.

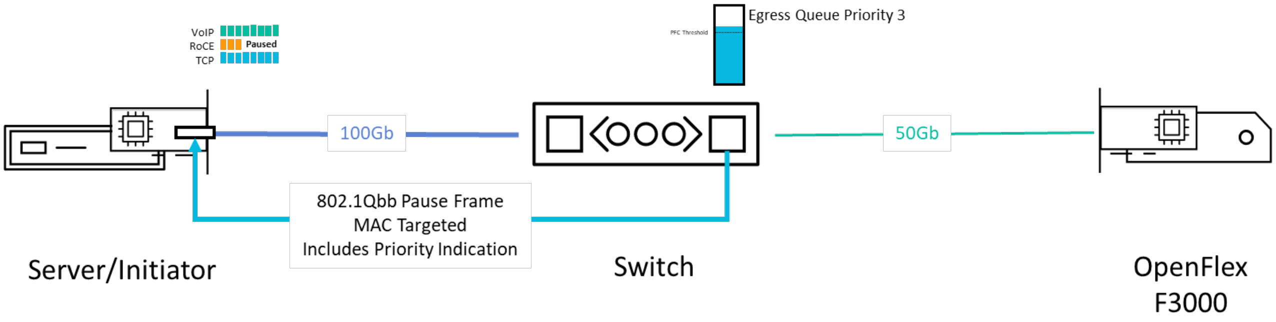


Figure 7 - Priority Flow Control

PFC uses priority-based pause frames for flow control, which is an essential component for lossless Ethernet. Although PFC operates on an Ethernet link between two adjacent Ethernet devices, a meaningful PFC setup for storage traffic should take a wider look at the network and at least configure the following:

- Fabric Storage Devices
- Network Switches
- Hosts (initiators), typically consumers of storage

As PFC is a Layer 2 mechanism, it has two drawbacks:

- Deficient Neighbor – In this concept, when there are two or more clients pulling data over the same traffic class from a target and one of the clients has substantially lower performance than the other, PFC could reduce the higher performance system down to match that of the lower performance system.

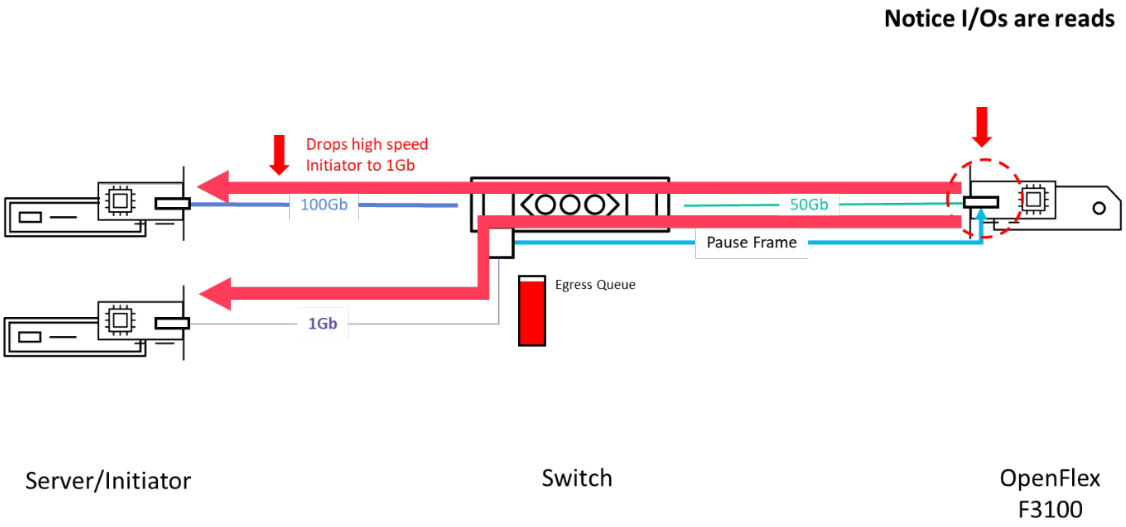


Figure 8 - Deficient Neighbor

- Congestion Sprawl – This concept involves a more complex network where two or more switches are being used. As PFC is a Layer 2 mechanism and only pauses traffic on neighboring ports, a switch to switch link that provides an artery for multiple streams of traffic within the same traffic class could be paused impacting the performance of third-party data streams unrelated to the congestion event. This can in tern cause more congestion sending out more pause frames branching through the network.

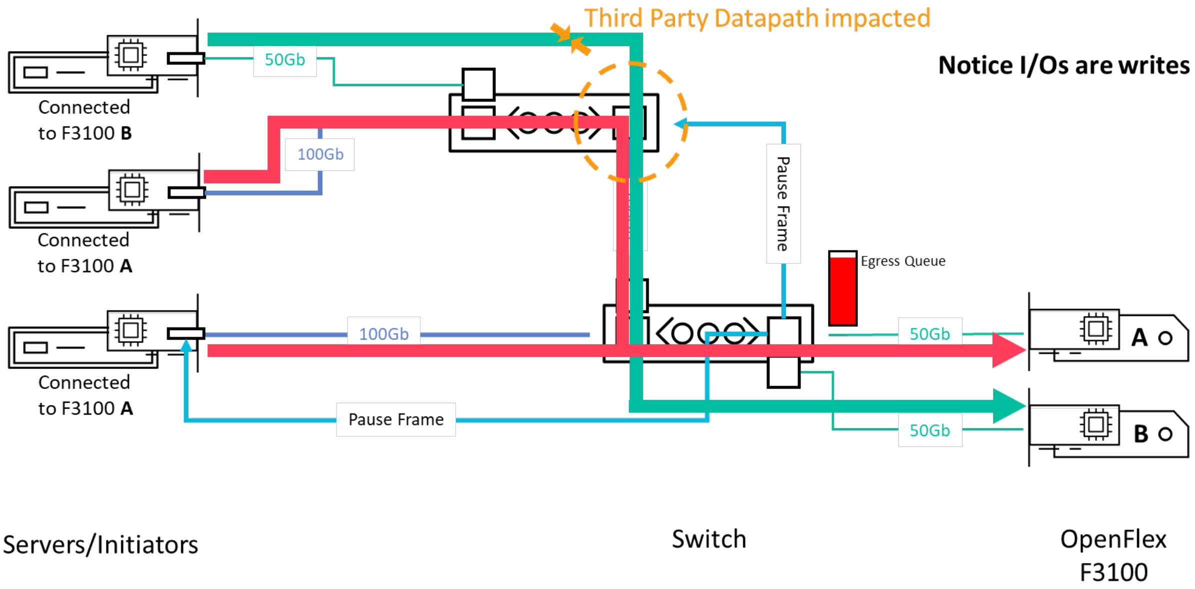
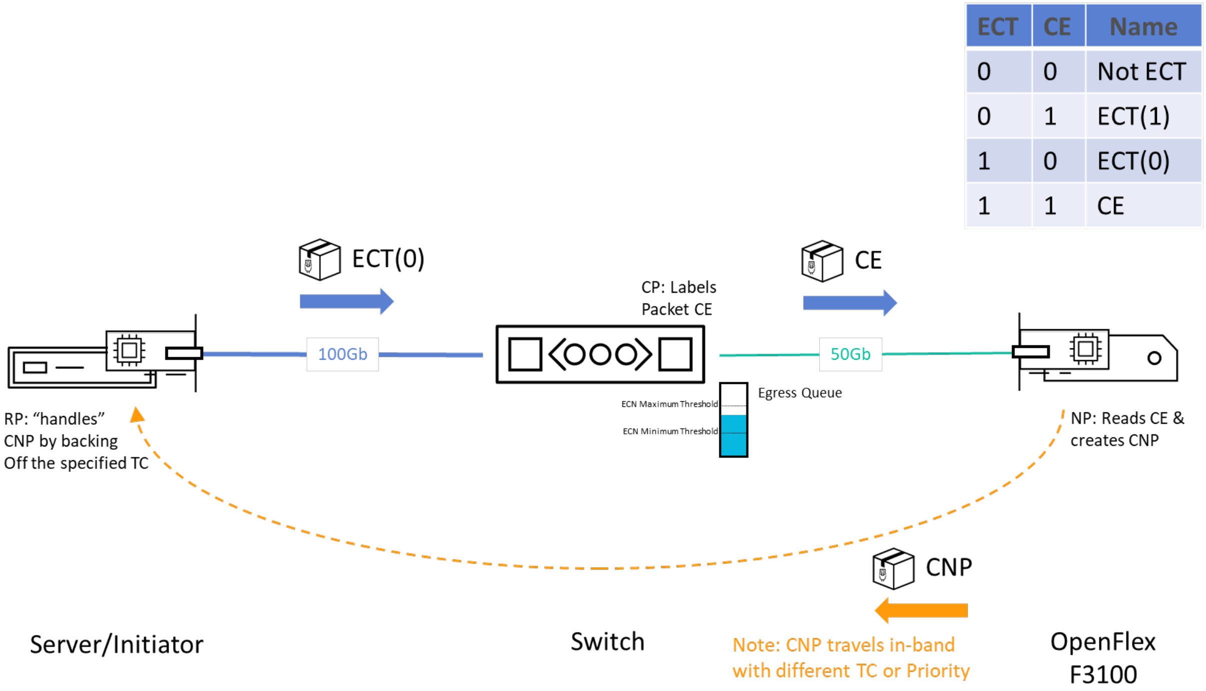


Figure 9 - Congestion Sprawl

Explicit Congestion Notification

Explicit Congestion Notification (ECN) is an extension to the Internet Protocol's (IP) RFC 7567 and is defined in RFC 3168. RFC 7567 recommends the use of Active Queue Management (AQM) to use algorithms such as Random Early Detection (RED) or Weighted Random Early Detection (WRED) to identify traffic better suited for loss to avoid costly retransmission of important traffic. ECN extends AQM to "mark" traffic instead of dropping traffic to allow for end to end notification and back-off control to avoid congestion. NVMe over Fabrics with RoCE uses ECN so that congestion can be reported in a feedback loop, thereby significantly reducing packet loss.

When ECN is properly configured the initiator sending traffic marks the packets as an "ECN Capable Transport" using the two least significant bits in the Differentiated Services (DiffServ) field of the IP header. (Defined above in traffic marking) When congestion is encountered, as identified by the AQM algorithms, the congestion point (CP) changes the two ECN bits in the DiffServ to indicate congestion encountered (CE). The packet continues its natural progression to its end point which strips off the CE and generates a Congestion Notification Packet (CNP). This makes the end point the notification point (NP). A CNP is a specialized packet that traverses the network in the opposite direction returning to the traffic streams origin to instruct it to slow down. This makes the originator of the traffic the reaction point (RP) as it reacts to the CNPs by slowing itself down for a period of time.



ECT	CE	Name
0	0	Not ECT
0	1	ECT(1)
1	0	ECT(0)
1	1	CE

ECT(0)
CE
CP: Labels Packet CE
100Gb
50Gb
ECN Maximum Threshold
ECN Minimum Threshold
Egress Queue
RP: "handles" CNP by backing Off the specified TC
NP: Reads CE & creates CNP
CNP
Server/Initiator
Switch
Note: CNP travels in-band with different TC or Priority
OpenFlex F3100

Figure 10 - Explicit Congestion Notification

As ECN is a Layer 3 mechanism, not all traffic in a traffic class may support ECN or be configured, so congestion can still occur. Therefore, for best Lossless configuration, ECN should be considered a compliment to PFC because ECN provides targeted notification, and because of this, it does not have the same issues as PFC related to "Deficient Neighbor" and "Congestion Sprawl". PFC acts as a catchall flow control when ECN isn't configured on all traffic flows of any given port. ECN operates end-to-end and requires configuration on:

- Fabric Storage Devices
- Network Switches
- Hosts (initiators)

Appendix

Contributors

Name	***Company***	***Title***
Jon Flynn	Western Digital	Sr. Technologist, Applications Engineering
Niall Macleod	Western Digital	Director, Applications Engineering
Barrett Edwards	Western Digital	Sr. Director, Field Engineering

References

Document Title
OpenFlex Data24 Deployment Considerations
OpenFlex F3200 Deployment Considerations
Configuring an Arista Switch for Lossless Networking
Configuring a Cisco Switch for Lossless Networking
Configuring a Lenovo Switch for Lossless Networking
Configuring a Mellanox Switch for Lossless Networking
Configuring a Switch Running Cumulus for Lossless Networking
Configuring a Broadcom Network Adapter for Lossless Networking
Configuring a Marvell Network Adapter for Lossless Networking
Configuring a Mellanox Network Adapter for Lossless Networking

Document Feedback

For feedback, questions, and suggestions for improvements to this document send an email to the Data Center Systems (DCS) Technical Marketing Engineering (TME) team distribution list at pdl-dcs-tm@wdc.com.

Version History

Version	Date	Notes
1.0	7 Apr 2021	Initial release

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com