Dell PowerConnect and FTOS - Flow Control and Network Performance

Dell Engineering
September 2014

A Dell EMC Technical White Paper

Revisions

Version	Date	Authors
2.0	September 2014	Ed Blazek, Mike Matthews, Sulaimon Odunmbaku
1.0 Initial Release	April 2011	Dell PowerConnect Team

Table of contents

1 Introduction

Flow control is defined in Annex 31B “MAC Control PAUSE operation” of the IEEE 802.3 Standard [1].

Transmission of Annex 31B PAUSE frames may be useful when deployed at the edge of a network for certain specific situations, but is generally considered harmful in the network core due to possible poor network performance when flow control is enabled. Due to this, Annex 31B flow control is not considered a viable method for implementing lossless Ethernet in general network deployments. There are other flow control alternatives that implement lossless Ethernet in very limited deployments [2].

This paper discusses the various types of flow control, interactions of Annex 31B flow control and TCP, implementation of flow control on Dell PowerConnect and FTOS Enterprise switches, issues of Annex 31B, and possible deployment recommendations and alternatives in implementing Annex 31B flow control.

Note: Annex 31B, Ethernet Pause Frame and Link-level are used interchangeably within this paper as they refer to the same feature.

2 Types of Flow Control

2.1 Ethernet Pause Frame Flow Control

Annex 31B flow control allows the receiver on a point-to-point Ethernet link to pause the adjacent sender from transmitting, which prevents buffer overflow and packet loss. It operates by sending PAUSE frames addressed to the peer or to 01-80-C2-00-00-01 (a well-known multicast MAC address specifically used for flow control). In the frame is a timer quanta (in increments equal to the time it takes to transmit 512 bits) for which the sender is required to cease transmission. The station sending the pause frame may also send a frame with a 0 pause quanta value, indicating that the paused peer may resume transmission.

In reality, IEEE 802.3 Annex 31B flow control is a method of congestion control. It is understood that IEEE 802.3 Annex 31B flow control does not and cannot solve steady-state over- subscription [3]. Flow control temporarily increases the network device buffer by utilizing the buffer of a neighbor for a brief period of time. A consequence is that the maximum link capability is reduced, which exacerbates the very condition that Annex 31B flow control was intended to solve. In some cases, the maximum link capacity may be reduced to half of its original capacity. Network equipment manufacturers generally recommend that flow control only be used on access ports connected to end hosts [3], [6], [7]. This is because of the issues surrounding congestion spreading and the fact that nearly all switches today can forward at line rate speeds.

Note: In the Protocol Implementation Conformance Statement (PICS) Proforma section 31B.4.3 or IEEE 802.3, support for the transmission of PAUSE frames is optional.

Figure 1 illustrates how Annex 31B flow control works under congested conditions. In this example, the storage has a 10Gbps NIC and receives traffic from three servers, each with a 10Gbps NIC. This allows the servers to send traffic to the storage at three times the line rate that it can handle.

If flow control is not configured and the servers are transmitting at over 100% line rate of the storage NIC, traffic congestion will occur and packets will be dropped by the storage. However, if flow control is supported and enabled on all the devices, when the servers are transmitting at over 100% line rate, the storage will notify the switch via PAUSE frames to pause transmission until a subsequent “resume” PAUSE frame is sent to commence transmission. The switch starts to build up a queue in its buffer until a certain buffer threshold is reached. At this point, the switch sends PAUSE frames to the transmitting severs in order to avoid dropping frames. When space becomes available in the storage’s buffer, it sends another PAUSE frame to the switch to resume transmission. As the switch’s buffer empties to a certain point, the switch consequently sends another PAUSE frame to inform the servers to resume transmission. This cycle continues until transmission is completed.

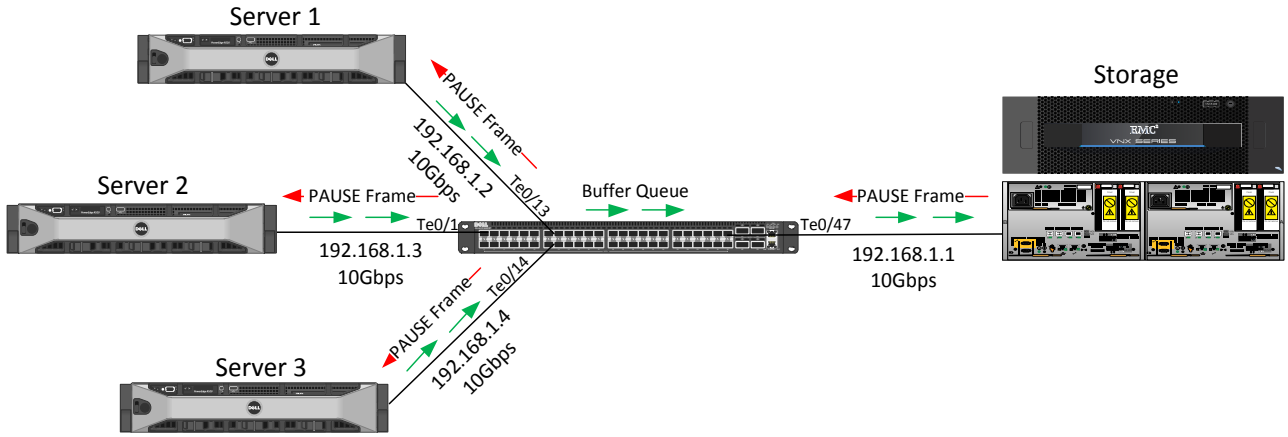

Figure 1 Annex 31B Flow Control

Note: PAUSE frames are a direct-link mechanism i.e. they are transmitted between directly connected devices, but indirectly between transmitting and receiving devices.

2.2 Priority-Based Flow Control

Priority-based flow control (PFC) is a part of the Data Center Bridging (DCB) protocol, which was developed to address the limitations of Annex 31B flow control. In a data center network, PFC manages large bursts of one traffic type in multiprotocol links so that it does not affect other traffic types and no frames are lost due to congestion.

When there is congestion on a queue for a specified priority, PFC sends PAUSE frames for the 802.1p priority traffic to the transmitting device(s), thereby ensuring that large amounts of queued LAN traffic do not inhibit the transmission of storage or server traffic. It also ensures that storage or server traffic does not experience high latency because of the congestion.

PFC enhances the existing Annex 31B PAUSE and 802.1p priority capabilities to enable flow control on 802.1p priorities (classes of service). Instead of pausing all traffic on a link (as in the case of Annex 31B), PFC pauses traffic according to the 802.1p priority set on a traffic type, thereby creating lossless flow for storage and server traffic while allowing for loss in the case of LAN traffic congestion on the same physical link [4].

Note: This paper is focused on Annex31B flow control. For additional information on Priority-Based Flow Control, refer to the Dell FTOS configuration guides located with the Dell Networking Switch Manuals.

3 Ethernet Pause Frame Flow Control and TCP

Higher layer protocols like Transmission Control Protocol (TCP) rely on packet loss as an indication to slow down the transmission (half the normal transmission rate) [8]. When implemented throughout a network, Annex 31B flow control makes the TCP retransmission algorithm redundant, but at the cost of lower network throughput. In addition, Annex 31B flow control interferes with the TCP RTT (Round Trip Time) measurement. Because of external head of line blocking and congestion spreading effects (see section 5.1 and 5.2), most network administrators prefer to use TCP retransmission instead of Annex 31B flow control as it allows utilization of the full network bandwidth.

3.1 TCP implementation of Flow Control

TCP uses the sliding window mechanism for flow control. TCP sliding window determines the number of unacknowledged packets (in Bytes) that a transmitting device can send to a receiving device. It determines this number (receive window size) using the size of the send buffer on the transmitting device and the size of the receive buffer on the receiving device.

When packets arrive at the receiving device, TCP stores the packets in the receive buffer and sends an acknowledgement which includes a new receive window size to the transmitting device. This new receive window size represents the available bytes in the receiving device's receive buffer. If the receive buffer is full, the new window size will be zero, and the transmitting device must hold transmission until buffer space is available on the receiving device's receive buffer. The available space in the receiving device's receive buffer depends on the read rate of the application utilizing the data. As the receiving application reads the data stored in the receive buffer, space will become available for new data, and the transmitting device can resume transmission.

In the event of missing packets or lack of acknowledgment of transmitted packets during the timeout (packet retransmission timer), the transmitting device's TCP stack assumes packet loss due to congestion at the receiving device and reduces the size of the sending window by half before any further packet transmission. The TCP stack will also assume packet loss if a duplicate acknowledgement is received for a sent packet. Usually, a packet loss is caused by network congestion or a transmission error.

3.2 Ethernet Pause Frame Flow Control interoperability with TCP Flow Control

When Annex 31B flow control is enabled, it functions independently from TCP sliding window flow control. Annex 31B and TCP both try to control the flow of data, causing a conflict, which results in reduced bandwidth utilization.

At the beginning of data transmission, TCP starts sending data at the normal transmission rate until the receiver is overloaded with data and Annex 31B flow control sends a PAUSE frame to the sender. The sender sees the PAUSE frame and stops sending temporarily. During this pause, TCP on the sender is unaware of the pause and continues sending data. This data becomes backed up in the sender's transmit buffer until the PAUSE frame expires, and data transmission resumes. If the sender's transmit buffer gets filled up before the PAUSE frames expires, the sender will start dropping its data. At this point TCP will recognize the packet loss, and will eventually reduce the transmission rate. Once the PAUSE frame expires, data transmission will

continue until the receiver is overloaded again. Because Annex 31B flow control and TCP flow control function independently, Annex 31B flow control obstructs the usually reliable TCP sliding window flow control mechanism.

Note: TCP flow control and Annex 31B flow control may use different buffer thresholds to detect oversubscription and to notify a sender to slow down transmission.

4 Industry Flow Control Implementations

It is apparent from Table 3 (see section 4.3) below that different vendors implement flow control in very different – and not necessarily compatible ways. What may be more interesting is that many network peripheral vendors, while recommending use of flow control, are not clear about the desired behavior of flow control in the network or with their devices, and thus may mislead network operators into deployments of flow control, which lead to lower network performance or high packet loss ratios or both.

4.1 Dell PowerConnect Ethernet Pause Frame Flow Control Implementation

This section examines the behavior of flow control on the Dell PowerConnect switches. Dell PowerConnect switches have a flow control feature, and if enabled, can receive and process PAUSE frames during link congestion. However, there are differences between the 62xx PowerConnect and later PowerConnect devices. PC62xx switches are based on StrataXGS-III Silicon, which has limited buffer space and therefore becomes congested sooner when confronted with burst traffic. Later PowerConnect devices such as the PC70xx/PC80xx series switches are based on StrataXGS-IV silicon with significantly more internal buffer space available to handle transient bursts. All PowerConnect devices use a shared-memory non-blocking architecture.

Table 1 shows the default configuration of flow control on the Dell PowerConnect switches.

Table 1 Dell PowerConnect Default Configuration of Flow Control

Model	Default Configuration				Change Configuration	
	Global		Interfaces		Global	Interfaces
	Rx	Tx	Rx	Tx		
PC2800	On	N/A	On	N/A	No	Yes
PC3500	Off	N/A	Off	N/A	No	Yes
PC5500	On	N/A	On	N/A	No	Yes
PC6200	On	N/A	On	N/A	Yes	No
PC7000	On	N/A	On	N/A	Yes	No
PC8000	On	N/A	On	N/A	Yes	No

StrataXGS-IV devices implement more aggressive memory allocation schemes intended to better tolerate bursty network behavior than StrataXGS-III devices. These allocation schemes include dynamic allocation of buffers and adjustment of limits based on real-time usage information.

PowerConnect Switches does receive and process PAUSE frames, but does NOT transmit PAUSE frames. Some PowerConnect switches allow flow control to be configured as `auto|on|off` while others only allow `on|off`. When flow control is configured to `auto`, the port auto-negotiates flow control settings with its connecting partner. Flow control configuration can be confirmed on a specific interface by using the `show interfaces` *`port-number`* command. To view the flow control configuration on all interfaces of a PowerConnect switch, enter the `show interface status` command. Interfaces showing flow control as *Active* are Up, and have flow control enabled while interfaces showing flow control as *Inactive* are down, but have flow control enabled.

Note:
Enabling the iSCSI optimization feature will automatically enable global flow control if it is not already enabled. On most PowerConnect Switches, by default, iSCSI optimization is globally disabled and flow control is globally enabled.

Interfaces set to half-duplex mode on Dell PowerConnect Switches do not use Annex 31B for flow control. Half-duplex ports uses the *Ingress Back Pressure* mechanism to ensure traffic flow control.

Further information about flow control on Dell PowerConnect Switches can be found in the User's Configuration Guide of each Dell PowerConnect Switch model.

4.2 Dell FTOS Ethernet Pause Frame Flow Control Implementation

This section focuses on examining flow control behavior on some of the Dell Networking Enterprise switches running the latest version of FTOS (version 9.5 (0.0).

As mentioned earlier, support for the transmission of PAUSE frames is optional. When flow control is enabled and configured, Dell FTOS can receive and transmit PAUSE frames when the connecting device becomes congested.

4.2.1 On Dell FTOS Switches

The Dell Networking Switch series running the Dell FTOS includes the E-Series, C-Series, S-Series, MXL 10/40GbE Switch IO Module, and the M I/O Aggregator. These switch platforms are ultra-low-latency 10/40GbE switches purpose-built for applications in high-performance data center and computing environments. This section details the behavior of the E-Series, C-Series, S-Series, and the MXL 10/40GbE Switch IO Module with respect to flow control.

Annex 31B flow control is supported on these FTOS platforms and globally disabled by default, and the default configuration on the interfaces is Rx off Tx off. To change the default configuration, use the `flowcontrol rx on tx off` interface configuration command.

Changes in the flow-control values may not be reflected automatically in `show interface` output. To display the change, apply the new flow control setting, perform a `shutdown` followed by a `no shutdown` command on the interface, and then check the `show interface` output again.

On the FTOS platforms, PAUSE frame threshold settings are supported. When an interface on a FTOS platform is set to transmit PAUSE frames, three thresholds can be set to closely define the controls. Annex 31B PAUSE frame can be triggered when either the flow control buffer threshold or flow control packet pointer threshold is reached. The thresholds are:

- Number of flow control packet pointers: from 1 – 2047 (default = 75)
- Flow control buffer threshold in KB: from 1 – 2013 (default = 49KB)
- Flow control discard threshold in KB: from 1 -2013 (default = 75 KB)

The pause is started when either the packet pointer or the buffer threshold is met. When the discard threshold is met, packets are dropped.

The pause ends when both the packet pointer and the buffer threshold fall below 50% of the threshold settings.

The discard threshold defines when the interface starts dropping the packet on the interface. This may be necessary when a connected device does not accept PAUSE frames sent by the FTOS device.

The discard threshold should be larger than the buffer threshold in order for the buffer to hold at least three packets.

To set the flow control thresholds, use the command `flowcontrol rx [off | on] to [off | on] [threshold {<1-2047> <1-2013> <1-2013>}]` on the interface.

Note:
When Annex 31B is enabled on a non-VLT interface, it must be enabled on all other non-VLT interfaces. Otherwise, the system may exhibit unpredictable behavior.

If `flowcontrol rx` is turned off, it is recommended that the system be rebooted.

4.2.2 On Dell M I/O Aggregator Modular Switch

The Dell M I/O Aggregator (IOA) is a zero-touch, layer-2 blade switch with two fixed 40 GB ports on the base module and supports two optional plug-in modules. The IOA operates in a Dell PowerEdge M1000e chassis, which can support up to 32 servers and 6 IOA blade switches. The IOA acts very similar to a pass-through module providing connectivity between network adapters internally and external upstream network devices. The IOA supports four operational modes: Standalone, Stacking, VLT and Programmable MUX (PMUX) mode. This section will address the behavior of Annex 31B flow control in each of the IOA's operational modes.

4.2.2.1 Standalone Mode

Standalone is the default mode for the IOA, it will boot to this mode after a factory default. This fully automated zero-touch mode allows the configuration of VLAN memberships.

In standalone mode, Data Center Bridging (DCB) is globally enabled by default. With DCB globally enabled, PFC is automatically enabled on the interfaces with no dot1p priorities configured. However, an auto-configured DCB-MAP policy is mapped to all the interfaces that disables PFC on the interfaces.

When an IOA powers on, it boots with auto-DCB-enable mode. In this mode, the IOA ports detect whether peer devices support Converged Enhanced Ethernet (CEE). This determines if PFC or Annex 31B flow control is enabled.

By default, the individual IOA interfaces come up with DCB disabled and Annex 31B flow control enabled to control data transmission between the IOA and other network devices.

If DCBx protocol packets are received (while DCB is globally enabled), interfaces automatically enable DCB and disable Annex 31B flow control The DCB-MAP and flow control configurations on the interfaces are

removed. If no DCBx TLVs are received on a DCB-enabled interface for 180 seconds, DCB is automatically disabled and Annex 31B flow control is re-enabled.

To configure the IOA so that the IOA and all its interfaces are DCB disabled and Annex 31B flow control enabled, use the `no dcb enable` command and then reload the IOA. The PFC buffer memory will automatically be freed. The default configuration of Annex 31B flow control in Standalone mode is RX on, TX off.

Note: In standalone mode, PFC and Annex 31B flow control cannot be enabled at the same time on an interface. To enable Annex 31B, first disable PFC on the interface using the `no pfc mode on` command.

4.2.2.2 Stacking Mode

The IOA stacking mode permits stacking of up to six IOA stack units into a single logical switch. The stack units can either be in the same or different M1000 chassis. Stacking mode is a low-touch mode where all configurations except VLAN membership is automated. VLAN must be manually configured. In this operational mode, the 40GbE base module ports are dedicated to stacking, and Annex 31B flow control behavior is the same as in standalone mode.

4.2.2.3 VLT Mode

The IOA VLT mode allows administrators to multi-home server interfaces to different IOA modules. Just like the stacking mode, it auto-configure all configuration except VLAN membership, which must be manually configured. In this mode, port 9 links of the IOA are dedicated to the VLT interconnect. Annex 31B flow control behavior in VLT mode is the same as in standalone mode.

4.2.2.4 Programmable MUX (PMUX) Mode

The IOA PMUX mode provides flexibility of operation with added configurability. This includes creating LAGs, configuring VLANs on uplinks and the server side, configuring DCB parameters, and so forth. When an IOA starts up in its default (standalone) mode, the mode can be changed to PMUX mode using the `stack-unit 0 iom-mode programmable-mux` command and reloading the IOA.

When an IOA boots in PMUX mode, DCB is enabled and Annex 31B flow control is disabled by default. This can be confirmed by running the `show dcb` command. With DCB enabled, `flow control Rx on/off Tx on/off` command will not be available on the interfaces. To enable flow control, disable DCB using the `no dcb enable` global configuration command, and reload the IOA. After rebooting, the IOA will have Annex 31B flow control enabled on all the interfaces.

Note: The default configuration of Annex 31B flow control on an IOA in PMUX mode is disabled with *RX off TX off*. When DCB is disabled and IOA reloaded, flow control is automatically enabled with *Rx on Tx off* on all interfaces.

Further information about flow control on IOA can be found in the Configuration Guide for the M I/O Aggregator Guide, located on the Dell Support website www.support.dell.com.

Table 2 shows the default configuration of flow control on the Dell FTOS Switch series.

Table 2 Dell FTOS Default Configuration of Flow Control

Model		Default Configuration				Change Configuration	
		Global		Interfaces		Global	Interfaces
		Rx	Tx	Rx	Tx		
S-Series		Off	Off	Off	Off	No	Yes
Z-Series		Off	Off	Off	Off	No	Yes
C-Series		Off	Off	Off	Off	No	Yes
E-Series		Off	Off	Off	Off	No	Yes
M IOA	Standalone Mode	On	Off	On	Off	No	Yes
	PMUX Mode	Off	Off	Off	Off	No	Yes
MXL IO		Off	Off	Off	Off	No	Yes

Note: On a FTOS Switch
Enabling iSCSI feature will automatically enable flow control on interface with PFC disabled. On most FTOS Switches, by default, iSCSI is globally disabled, flow control is globally disabled, and DCB is globally enabled.

Flow control is not supported on Interfaces set to half-duplex mode on Dell FTOS Switches.

For details on the flow control configuration and behavior on each of the Dell Networking switches running FTOS, refer to Dell FTOS configuration guides located with the Dell Networking Switch Manuals.

4.3 Other Vendors Flow Control Implementation

Flow control implementation differs among network peripheral vendors. Some network devices come with flow control enabled globally and on interfaces by default. This could cause serious network impairment especially when LAN and SAN traffic are passed across a converged Ethernet infrastructure. Table 3 shows other vendors Annex 31B flow control implementation and their default settings.

Table 3 Other Vendors Default Configuration of Flow Control

Device	Configurability	Tx-Pause	Rx-Pause
Cisco 2970	Interface	No	Yes
Cisco 3560	Interface	N/A	Yes
Cisco 3750	Interface	No	Yes
Cisco 4500	Interface	No	Yes
Cisco 6500	Interface	Desired (Negotiate)	Yes
Cisco Nexus 9000	Interface	Not Supported	Not Supported
Cisco Nexus 7000	Interface	Desired (Negotiate)	Desired (Negotiate)
Cisco Nexus 5000	Interface	No	No
J and SRX Series	Interface	Yes	Yes
HP ProCurve 9300	Interface	Yes (Global threshold)	Yes
HP ProCurve 2400m/4000m	Interface	No	No
HP FlexFabric 12900	Interface	No	No
HP FlexFabric 7900	Interface	No	No

For more information about network peripheral vendors flow control implementation and default configuration on their respective network products, visit the website of the device manufacturer.

5 Ethernet Pause Frame Flow Control Implementation Issues and Recommendations

In this section, Annex 31B flow control issues and possible alternatives to global deployment of symmetric flow control are discussed. While considering viable options of implementing flow control, it is very important to understand the following:

- Flow control is not intended to solve the problem of steady-state congested links or networks.
- Flow control is not intended to address poor network capacity.
- Flow control is not intended to provide end-to-end flow control.

If Annex 31B flow control is to be utilized, it must be implemented in a consistent manner across the network with an understanding of the specific implementation on each network device. Ad hoc implementations of flow control are likely to cause significant network impairments, including high packet loss ratios and significantly degraded network throughput. If Annex 31B flow control will be globally implemented, two major concerns are external head of line blocking and congestion spreading.

5.1 External Head of Line Blocking

Consider the network diagram below (Figure 2). In this example, S1 is transmitting at 100% line rate to Storage and S2 is periodically transmitting bursts of traffic to Storage at 10% of line rate. S2 and S3 are transmitting at 90% of line rate to each other. This results in a periodic 10% oversubscription of the link from the Switch to Storage.

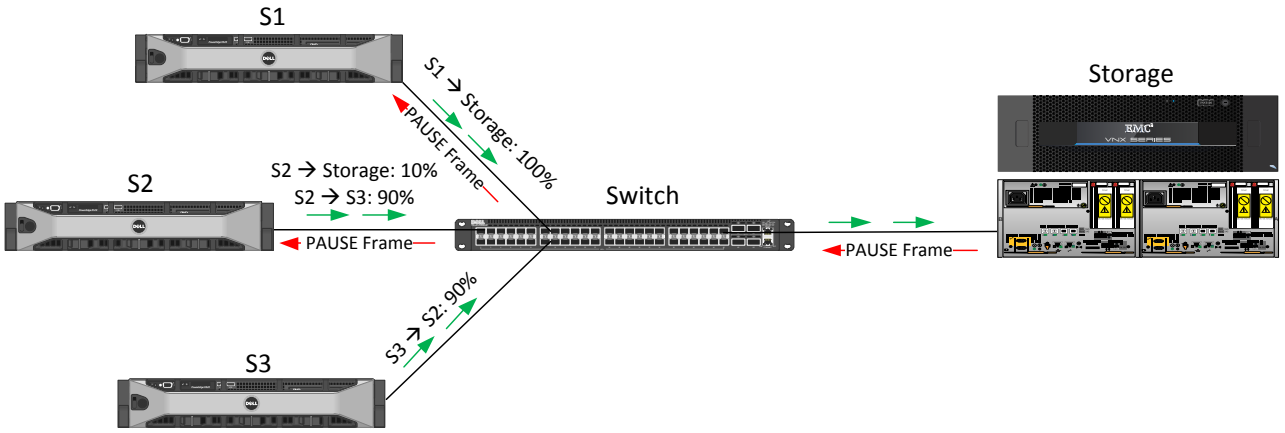

Figure 2 External Head of Line Blocking

Due to the oversubscription of the link between the Switch and Storage, the Switch will send pause frames to all ports attempting to send packets to Storage. In this example, the Switch will send pause frames to both S1 and S2, which has the undesirable effect of blocking the packets S2 is transmitting to S3. This is known as external head of line blocking.

Head of line blocking is undesirable as it not fair (traffic from S2 to S3 is blocked) and it is wasteful of network resources (the overall S1 and S2 link utilization is reduced).

5.2 Congestion Spreading

Consider the network diagram below (Figure 3). In this example, S1 is transmitting at 100% line rate to Storage via Switch 2. S2 and S3 are transmitting at 45% of line rate to S4 via Switch 1. S2 and S3 are also each periodically transmitting bursts of traffic at 5% of line rate to Storage via Switch 1 and Switch 2. This situation results in a periodic 10% oversubscription of the link from Switch 2 to Storage.

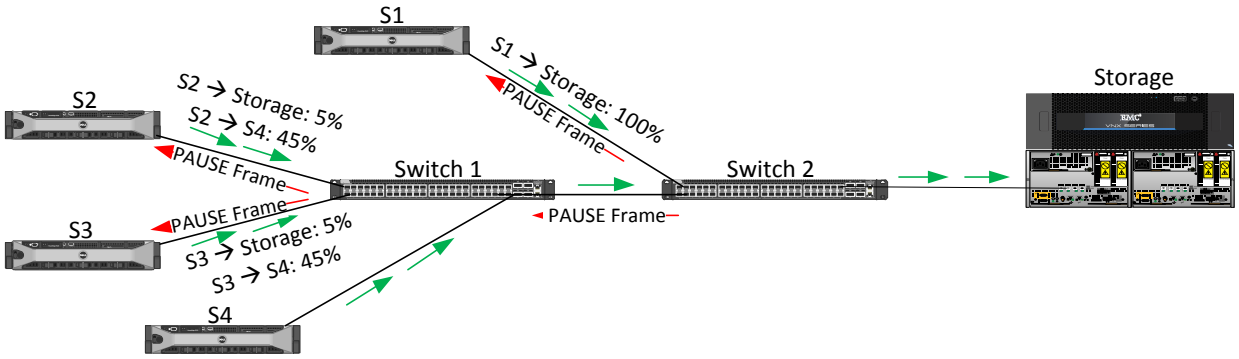

Figure 3 Congestion Spreading

Due to the oversubscription of the link between the Switch 2 and Storage, Switch 2 will send pause frames to all ports attempting to send packets to Storage. In this example, Switch 2 will send pause frames to both S1 and Switch 1. This in turn may cause PAUSE frames to be sent to S2 and S3 as Switch 1 becomes congested due to the lowered throughput on the link from Switch 1 to Switch 2. This has the undesirable effect of blocking packets sent from S2 and S3 to S4. This is known as congestion spreading [9].

5.3 Recommendations

There are multiple options for configuring flow control. Table 4 shows these basic flow control configuration options along with their effect on network performance. Some network administrators and engineers may think that flow control should be completely disabled on all switches and end devices because of the complexity in implementing it, and issues that may arise if not properly implemented. As a rule of thumb, it is generally safe to configure flow control using the configurations with *No* in the *Network Issue* column in Table 4.

Table 4 Basic Flow Control Configuration Options.

Connectivity		Flow Control Rx/Tx ←→ Rx/Tx	Network Issue	Comment
Server/Storage ←→ Switch		On/Off ←→ On/Off	No	
		On/On ←→ On/On	Yes	Latency
		On/On ←→ On/Off	No	
Switch ←→ FTOS Switch	Trunk Connection/ VLT	On/Off ←→ On/Off	No	
		On/On ←→ On/On	Yes	Latency, LACP flap
		On/On ←→ On/Off	Yes	Latency, LACP flap
Switch ←→ PowerConnect Switch	Trunk Connection	On/On ←→ On	Yes	Latency, LACP flap
		On/Off ←→ On	No	

Below are possible alternatives to deploying global symmetric flow control in a network.

5.3.1 Asymmetric Flow Control

Deploy asymmetric (Rx only) flow control throughout the network. Should a device, which implements symmetric flow control be deployed in the network, the directly attached devices will operate in a compatible manner. This alternative will allow the network to be utilized at maximum capacity, although with a potentially higher packet loss ratio. Deploying sufficient network capacity can lessen the packet loss to near zero. Monitoring network traffic flows by periodically polling the switch from a network management system will assist in planning network capacity enhancements and in understanding network traffic flow.

5.3.2 Asymmetric with Symmetric Flow Control

Deploy asymmetric (Rx only) flow control throughout the network in conjunction with symmetric (Rx and Tx) flow control utilized for directly attached hosts. This deployment pattern assists in protecting the network from any host or group of hosts that could affect network operation by sending long or large busts of traffic. This alternative will allow the network to be utilized at close to maximum capacity, since interior links operate at full capacity and only exterior links are flow controlled. With sufficient network capacity deployed, packet loss can be limited to a very small fraction of total traffic. Monitoring network traffic flows by periodically polling the switch from a network management system will assist in planning network capacity enhancements and in understanding network traffic flow.

5.3.3 Flow Control on iSCSI Ethernet Network

Enable flow control for iSCSI Ethernet networks. Enabling flow control in a well-designed and high performance iSCSI Ethernet network is very important. During congestion at the target, retransmission of dropped packets will result in high latency and degradation of I/O performance. This problem can be eliminated by flow control, which will allow the target to process its backlog so it can later resume accepting packets. Enabling flow control at the target will drastically reduce the overhead caused by TCP/IP packet re-transmission. Increase buffer size and enable asymmetric flow control on the switch to receive but not transmit pause frames.

5.3.4 Disable Flow Control on Network Core

Disable flow control at the core of the network. Enabling flow control at the core can cause congestion spreading. A constantly congested link is most likely caused by a problem with the current implementation of the network. This problem can be solved by reducing the load across the link, redesigning the network, or implementing proper host-to-host QoS. The best way to prevent or control potential congestion at the core of a network is to implement CoS/QoS. Prioritizing the packets at this part of the network provides better traffic control than pausing packet transmission regardless of the importance.

5.3.5 Global Disable of Flow Control

Disable flow control globally on the network. An alternative to flow control is QoS. Implementing global QoS and moving traffic sensitive to jitter and delay variance (e.g. VoIP) to the head of the queue for transmission while allowing other less sensitive or less important traffic to be buffered or dropped can provide better utilization of network capacity, but with potential high packet loss ratio of unimportant traffic. At least, external head of line blocking and congestion spreading will be totally avoided.

6 Ethernet Pause Frame Flow Control on Stacking and VLT Interfaces

Stacking and VLT on Dell Networking Enterprise switch platforms operate over Ethernet ports configured to act as stacking ports or VLT peer ports using proprietary protocols to transport Ethernet frames with low latency. In general, stacking and VLT ports have higher bandwidth limits in order to reduce congestion issues and mitigate the need for flow control on the stacking and VLT links.

6.1 Dell PowerConnect and FTOS Stacking

Dell PowerConnect and FTOS switches do not support flow control over stacking links and do not have a feedback mechanism to control packet ingress from the egress ports located on the other stack members. The stacking ports themselves are egress ports with fixed limits (combined original speed of each ports making up the stack link). This leads to a situation where multiple ingress links may forward traffic to a stacking link in excess of the egress limits. This will lead to internal packet discards as the output queue exceeds the configured thresholds. On the other hand, the disabling of ingress limits when flow control is disabled, coupled with the excellent memory management, low latency of Dell PowerConnect and FTOS devices can often lead to higher throughput with minimal loss in stacking environments if an appropriate design to limit oversubscription is in place.

6.2 Dell FTOS Virtual Link Trunking

Dell FTOS does support flow control on VLT interfaces on all Dell FTOS Switch platforms. In a VLT domain, flow control is supported on the VLT physical interfaces, VLT port-channels, and the VLTi link. Flow control configuration on VLT physical interfaces is the same as the configuration on non-VLT physical interfaces.

7 Summary

Many factors come into play when considering whether to use flow control in a network. These include packet loss prevention, network congestion prevention, improved iSCSI performance, external head of line blocking and congestion spreading.

It is recommended that Annex 31B flow control only be implemented at the network edge. Flow control should be disabled in the core. In cases where it is absolutely needed, only asymmetric flow control should be implemented in the core to ensure maximum network throughput. Appropriate network design must be performed to ensure that interior network links are not over-subscribed. Flow control is supported on interfaces in a VLT domain, and must be configured in the same manner as non-VLT interfaces. Operators should be aware of the limitations of utilizing flow control in stacking solutions and take steps to mitigate any issues that may be encountered.

If Annex 31B flow control is needed, Dell PowerConnect and FTOS switches conform to all relevant IEEE standards concerning Annex 31B flow control.

A References

1. IEEE 802.3 Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements
   http://standards.ieee.org/about/get/802/802.3.html

2. IEEE 802.1Qbb – Priority-based Flow Control
   http://www.ieee802.org/1/pages/802.1bb.html

3. Dell FTOS Configuration Guides
   http://www.force10networks.com/CSPortal20/KnowledgeBase/Documentation.aspx

4. Dell 9.5(0.0) PowerEdge Configuration Guide for the M I/O Aggregator
   http://www.dell.com/support/home/us/en/19/product-support/product/poweredge-m-io-aggregator/manuals#./manuals?&_suid=140629681032707892607446123341

5. Best Practices for Catalyst 4500/4000, 5500/5000 and 6500/600 Series Switches Running CatOS Configuration and Management
   http://www.cisco.com/c/en/us/support/docs/switches/catalyst-4500-series-switches/13414-103.html

6. Network World - Vendors on Flow Control
   http://archive.today/mDZlm

7. Virtual Threads – Beware Ethernet Flow Control
   http://virtualthreads.blogspot.com/2006/02/beware-ethernet-flow-control.html

8. TCP/IP Illustrated – Volume 1 – The Protocols, Stevens, Chapter 21

9. Congestion Control for Switched Ethernet, McAlpine
   http://www.cercs.gatech.edu/hpidc2005/presentations/GaryMcAlpine.pdf

Support and Feedback

Contacting Technical Support

Support Contact Information	Web: http://Support.Dell.com/
	Telephone: USA: 1-800-945-3355

Feedback for this document

We encourage readers of this publication to provide feedback on the quality and usefulness of this deployment guide by sending an email to Dell_Networking_Solutions@Dell.com