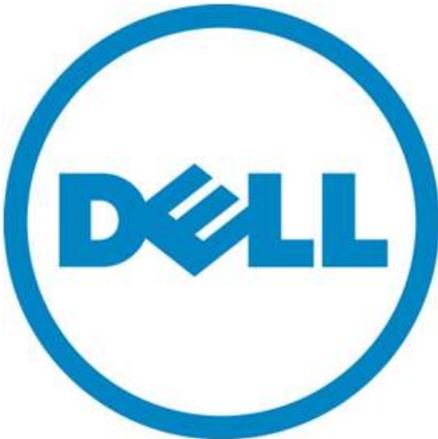


# Flow Control and Network Performance

---

A Dell Technical White Paper

Dell PowerConnect Team



THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2011 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

*Dell*, the *DELL* logo, and the *DELL* badge, *PowerConnect*, and *PowerVault* are trademarks of Dell Inc. *Symantec* and the *SYMANTEC* logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the US and other countries. *Microsoft*, *Windows*, *Windows Server*, and *Active Directory* are either trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell Inc. disclaims any proprietary interest in trademarks and trade names other than its own.

April 2011

## Contents

A Dell Technical White Paper .....	
Dell PowerConnect Team .....	
Introduction .....	2
External Head of Line Blocking .....	2
Congestion Spreading .....	3
Annex 31B Flow Control Operation .....	4
Annex 31B Flow Control and TCP .....	4
Industry Flow Control Implementations .....	4
Flow Control Alternatives .....	5
Alternative 1 .....	5
Alternative 2 .....	5
PowerConnect Flow Control Implementation .....	5
PowerConnect Stacking .....	6
Summary .....	6
References .....	7

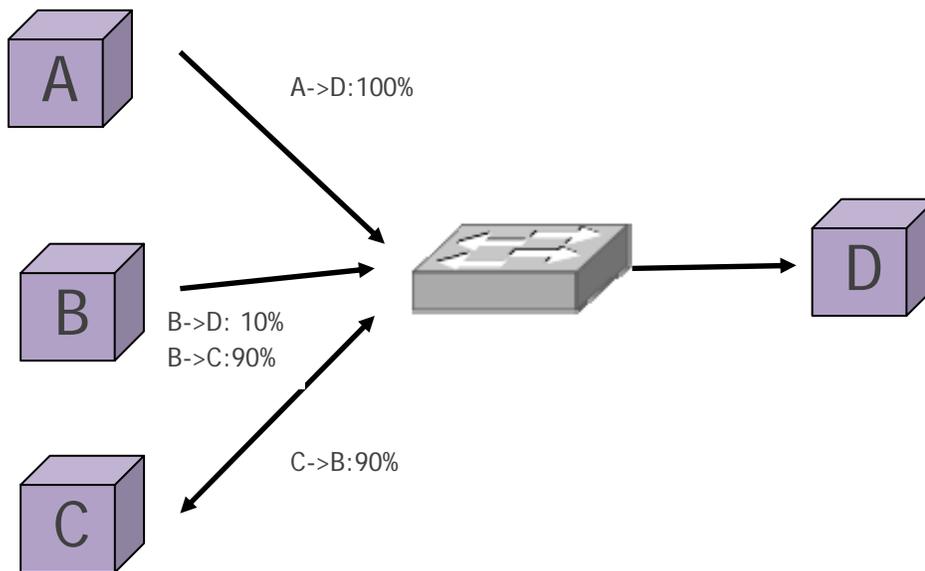
## Introduction

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 the effects of external head of line blocking and congestion spreading [3]. Due to these effects, Annex 31B flow control is not considered as 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 [10]. In this paper, we discuss head of line blocking and congestion spreading, Annex 31B flow control operation and its interactions with TCP, and then discuss the Dell PowerConnect implementation of flow control along with some possible deployment alternatives in implementing flow control.

### *External Head of Line Blocking*

Consider the network diagram below. In this example, host A is transmitting at 100% line rate to host D and host B is periodically transmitting bursts of traffic to host D at 10% of line rate. Hosts B & C are transmitting at 90% of line rate to each other. This results in a periodic 10% oversubscription of the link from the switch to D.

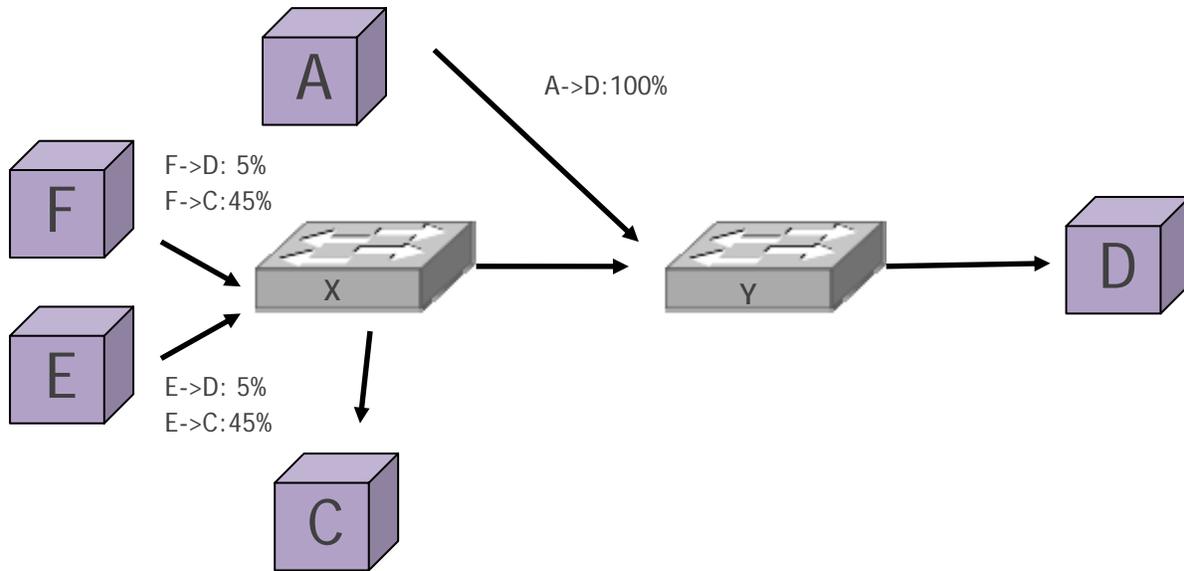


Due to the oversubscription of the link between the switch and host D, the switch will send pause frames to all ports attempting to send packets to host D. In this example the switch will send pause frames to both host A and host B which has the undesirable effect of blocking the packets host B is transmitting to host C. This is known as external head of line blocking.

It is readily understandable that head of line blocking is undesirable as it not fair (traffic from B to C is blocked) and it is wasteful of network resources (overall A and B link utilization is reduced).

### *Congestion Spreading*

Consider the network diagram below. In this example, host A is transmitting at 100% line rate to host D via switch Y. Hosts E and F are transmitting at 45% of line rate to host C via switches X and Y. Hosts E and F are also each periodically transmitting bursts of traffic at 5% of line rate to host D via switches X and Y. This situation results in a periodic 10% oversubscription of the link from switch Y to host D.



Due to the oversubscription of the link between the switch and host D, switch Y will send pause frames to all ports attempting to send packets to host D. In this example switch Y will send pause frames to both host A and switch X. This in turn may cause PAUSE frames to be sent to hosts E and F as switch X becomes congested due to the lowered throughput on the link from switch X to switch Y. This has the undesirable effect of blocking packets sent from hosts E and F to host C. This is known as congestion spreading [8].

## Annex 31B Flow Control Operation

Annex 31B flow control operates by sending a link local frame addressed to the peer or to a well-known MAC address specifically used for flow control. In the frame is a timer quanta (in increments of 512 bit times) for which the transmitter 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 [4]. It is generally well understood that IEEE 802.3 Annex 31B flow control does not and cannot solve steady-state over-subscription [3]. The effect of flow control is to temporarily increase the network device buffer by utilizing the buffer of the 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 a fraction of its original capacity. Network equipment manufacturers generally recommend that flow control only be used on access ports connected to end hosts [2,3,4]. This is because of the issues surrounding congestion spreading and the fact that nearly all switches today can forward at line rate speeds.

## Annex 31B Flow Control and TCP

Higher layer protocols like TCP rely on packet loss as an indication to transmit more slowly (halves the transmission rate) [7]. 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, flow control interferes with the TCP RTT measurement. Because of congestion spreading and head of line blocking effects, most network administrators prefer to use TCP retransmission over Annex 31B flow control as it allows utilization of the full network bandwidth.

## Industry Flow Control Implementations

Device	Configurability	Tx Pause	Rx Pause
Cisco 3750	Interface	No	Yes
Cisco 6500	Interface	Yes (default off)	Yes
Cisco 2970	Interface	No	Yes
J and SRX Series	Interface	Yes	Yes
HP ProCurve	9300	Yes (global threshold)	Yes
HP ProCurve	2400m/4000m	No	No

It should be readily apparent from the above table 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 [5,9].

## Flow Control Alternatives

In this section, we discuss possible alternatives to global deployment of symmetric flow control or no flow control which may be considered by network operators. What is most important is to understand is that flow control, if it is to be utilized, must be implemented in a consistent manner across the network and 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.

### Alternative 1

Asymmetric (rx only) flow control is deployed throughout the network. Should a device be deployed in the network which implements symmetric flow control, 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. Deployment of sufficient network capacity can lessen packet loss effects 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.

### Alternative 2

Asymmetric (rx only) flow control is deployed throughout the network in conjunctions with symmetric flow control utilized for directly attached hosts. This deployment pattern assists in protecting the network from any host or group of hosts from affecting network operation by sending long or large bursts 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.

## PowerConnect Flow Control Implementation

In the PICS Proforma section 31B.4.3 or IEEE 802.3, support for transmission of PAUSE frames is optional. When flow control is enabled, PowerConnect devices transmit PAUSE frames when congested, however, there are differences between the 62xx PowerConnect and later PowerConnect devices. PC 62xx switches are based on StratXGS-III silicon, which has limited buffer space and therefore becomes congested sooner when confronted with bursty 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.

An StrataXGS-IV PowerConnect switch configured with flow control enabled uses ingress back pressure to support lossless egress buffering. Ingress back pressure enables per switch management of an oversubscribed port through the use of PAUSE flow control at the source port. The destination module keeps counts for all incoming packets according to their source port. When those counters exceed the configured limit, a message is sent with flow control information to the offending source port(s). In this mode, the ingress limits are lowered and the egress limits are raised.

A StrataXGS-IV based PowerConnect switch configured with flow control disabled does not use the back pressure mechanism to control congestion. Instead, the egress limits are lowered and ingress limits are disabled. In this case, the switch will discard frames earlier than with flow control enabled.

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 Stacking

PowerConnect stacking operates over the stacking ports using a proprietary protocol to transport Ethernet frames with low latency. In general, stacking ports have higher bandwidth limits in order to reduce congestion issues and mitigate the need for flow control on the stacking links. However, PowerConnect stacking does not implement flow control over stacking links and does not have a feedback mechanism to control packet ingress from the egress ports located on other stack members. The stacking port itself is an egress port with fixed limits. 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 more aggressive memory management of StrataXGS-IV devices can often lead to higher throughput with minimal loss in stacking environments if an appropriate design to limit oversubscription is in place.

## Summary

Many factors come into play in considering whether to use flow control in a network. We have listed many of the factors to consider as well as recommendations from multiple switch vendors. We have also listed the differences between industry implementations of flow control. PowerConnect switches conform to all relevant IEEE standards with regards to Annex 31B flow control. Broadcom's recommendation mirrors that of the other large networking equipment vendors [2,4]. Flow control (transmission of PAUSE frames) should only be implemented at the network edge, when needed, and asymmetric flow control should be implemented in the core to interoperate with other vendor implementations and assure maximum possible network throughput. Appropriate network design must be performed to ensure that interior network links are not over-subscribed. Operators should be aware of the limitations of utilizing flow control in stacking solutions and take steps to mitigate issues that may be encountered.

## 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. Best Practices for Catalyst 4500/400, 55090/5000 and 6500/600 Series Switches Running CatOS Configuration and Management ([http://www.cisco.com/en/US/products/hw/switches/ps663/products\\_tech\\_note09186a0080094713.shtml](http://www.cisco.com/en/US/products/hw/switches/ps663/products_tech_note09186a0080094713.shtml))
3. Network World - Vendors on Flow Control (<http://www.networkworld.com/netresources/0913flow2.html>)
4. Flow Control ([http://en.wikipedia.org/wiki/Flow\\_control](http://en.wikipedia.org/wiki/Flow_control))
5. Virtual Threads - Beware Ethernet Flow Control (<http://virtualthreads.blogspot.com/2006/02/beware-ethernet-flow-control.html>)
6. Informatica - Ethernet Flow Control ([http://www.informatica.com/downloads/1568\\_high\\_perf\\_messaging\\_wp/Topics-in-High-Performance-Messaging.htm](http://www.informatica.com/downloads/1568_high_perf_messaging_wp/Topics-in-High-Performance-Messaging.htm))
7. TCP/IP Illustrated - Volume 1 - The Protocols, Stevens, Chapter 21
8. Congestion Control for Switched Ethernet, McAlpine (<http://www.cercs.gatech.edu/hpidc2005/presentations/GaryMcAlpine.pdf>)
9. A guide to building an iSCSI based SAN solution with Dell™ EqualLogic™ PS Series Arrays ([http://search.yahoo.com/r/\\_ylt=A0oGdWnw4LZN3g8AKFxXNyoA;\\_ylu=X3oDMTE1ZjJxZ2JvBHNIYwNzcgRwb3MDMQRjb2xvA3NrMQR2dGika1NNRTAyNV8xNjM-/SIG=15r0h89nm/EXP=1303852368/\\*\\*http%3a//en.community.dell.com/cfs-filessystemfile.ashx/\\_\\_\\_key/CommunityServer.Discussions.Components.Files/866/8863.Dell\\_5F00\\_EqualLogic\\_5F00\\_Configuration\\_5F00\\_Guide.pdf](http://search.yahoo.com/r/_ylt=A0oGdWnw4LZN3g8AKFxXNyoA;_ylu=X3oDMTE1ZjJxZ2JvBHNIYwNzcgRwb3MDMQRjb2xvA3NrMQR2dGika1NNRTAyNV8xNjM-/SIG=15r0h89nm/EXP=1303852368/**http%3a//en.community.dell.com/cfs-filessystemfile.ashx/___key/CommunityServer.Discussions.Components.Files/866/8863.Dell_5F00_EqualLogic_5F00_Configuration_5F00_Guide.pdf))
10. IEEE 802.1Qbb - Priority-based Flow Control (<http://www.ieee802.org/1/pages/802.1bb.html>)