



[Home](#) [Archives](#) [Profile](#) [Subscribe](#)

[Main](#)

04/13/2017

Slow Drain due to Bandwidth mismatch and Large Block READs

Every once in a while, you come across information that makes you understand something in a completely new way.



Image credit: <https://farscapedevelopment.wordpress.com/2012/07/06/eureka/>

I had just such an experience a few weeks back, and Alan Rajapa and I have been working in the lab since then to try and come to terms with it.

The person who originally provided the test data is a gentleman by the name of Vijay and what he noticed is extremely insightful, relatively easy to understand, and since we've confirmed it in the lab, I can now state that I believe:

Any time an FC HBA or FCoE CNA is performing multiple sequential Large Block READs (> 128k block size) AND the HBA/CNA:Storage bandwidth ratio is less than 1 (e.g., 4G FC HBA and 8G storage, 10G FCoE CNA and 16G storage,



[Erik Smith](#)

[ARCHIVES](#)

[September 2022](#)

[August 2020](#)

[April 2019](#)

[December 2017](#)

[November 2017](#)

[July 2017](#)

[April 2017](#)

[September 2016](#)

[July 2016](#)

[February 2016](#)

SEARCH

[CATEGORIES](#)

[Subscribe to this blog's feed](#)

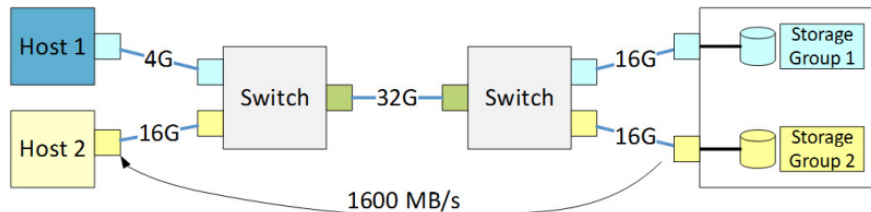
DISCLAIMER

This is not an official EMC blog. The opinions expressed here are my personal opinions. Content published here is not read or approved in advance by EMC and does not necessarily reflect the views and opinions of EMC nor does it constitute any official communication of EMC.

etc), the HBA/CNA will start behaving as a mild slow drain and cause congestion spreading.

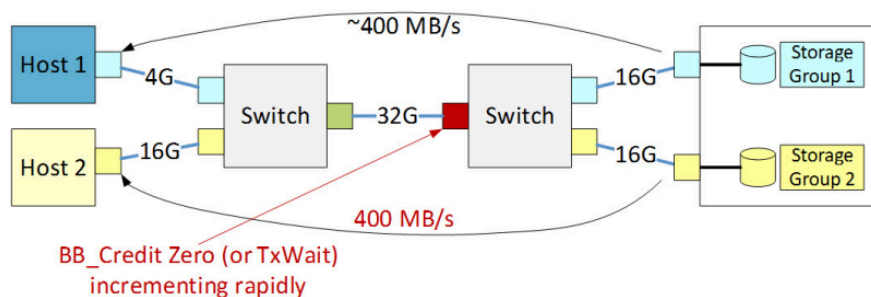
We've found the above rule holds for all HBA/CNA and OS combinations that we've tested.

So you may be wondering why this is important? The answer has to do with the impact that congestion spreading has on innocent flows. For example, consider the following topology:



Host 2 was performing Large Block (512k) sequential READs and was able to get up to about 1600MB/s since both it and the storage array interface it was accessing were both running at 16G. This is what we would expect and is completely normal.

However, as soon as we added Host 1 (connected at 4G) and started to perform the same Large Block Sequential READs to a 16G target, the throughput for Host 2 dropped to about 400MB/s! During this timeframe we noticed the `tim_txcrd_z` (Time spent at Zero BB_Credit) counter incrementing on the E_Port attached to the Brocade switch one hop away. We were then able to reproduce the same symptoms on Cisco switches and noticed that under the same conditions their `TxWait` counter would increment.



As we dug into the root cause of this behavior, we realized that even though the 4G adapter was running at 4Gb/s, it was requesting data from the array that would ordinarily result in a much higher rate. In fact, we believed, and then were subsequently able to prove, that the 4G adapter was requesting data at a rate that would have resulted in a data rate of 16Gb/s!! How did we prove this? We increased the link speed of Host 1 to 16G and then noticed that both adapters were receiving data at a rate of 1600 MB/s. We also confirmed with the adapter vendors (that we tested) that they do not take link speed or block size into account when transmitting READ requests. So, while in this configuration, there was effectively a bandwidth mismatch of about 1200 MB/s and it's interesting to note that Host 2's throughput dropped to the same as that of Host 1. This is typical for this kind of BW mismatch scenario.

So what should you do?

You need to monitor for oversubscription!

Which as it turns out is much easier said than done.

Here's the problem. From the switches point of view, if the 4G adapter (in above example) is running near line rate and returning R_RDY's regularly, then it's not doing anything wrong and you probably won't even see high bb_cred_z or TxWait at switch port connected to Host 1. As a result, in order to monitor for this kind of a condition, you need to monitor the F_Ports connected to the end devices (host and storage port) for high utilization (e.g., >= 95%) and also monitor the ISLs (E_Ports) for signs of congestion that would be similar to what we described in the [slow drain due to lost credit KB article 464246](#) (i.e., a ratio of "time spent a zero transmit credit": "frames transmitted" that is greater than .02)

If you'd like a bit more detail regarding how to do this in a Cisco environment, see the [Identifying Slow Drain with Connectrix MDS video](#).

In part 2 of this series we'll discuss how to prevent, detect and remediate slow drain due to oversubscription. If you'd like a preview of this information, join Alan and I at Dell EMC World for our "Configuring your SAN to support all Flash Arrays" session where we'll dive into the details of the following slide.

Congestion Spreading Avoidance			
	Prevention	Detection	Remediation
Short Term (today)	- BW Ratio - Modernize SAN - Fan-In/Out Ratio - BW Limits - Isolation	- Brocade MAPs - Cisco Port Monitor	- Port Fencing - Brocade SDDQ - Cisco Slow Drain Isolation –Virtual Link
Medium term	- BW Limit Enhancements - ViPR	ViPR	ViPR
Long term	Future Fibre Channel Enhancements (TBD)		

The session information is:

Tuesday May 9th, 8:30-9:30 AM

Thursday May 11th, 8:30-9:30 AM

You can also join us for an open discussion on this topic during our Birds-of-a-feather session which is:

Tuesday May 9th, 1:30-2:30 PM

Thanks for reading!

Posted at 05:27 PM | [Permalink](#)

 Reblog (0)

Comments

 You can follow this conversation by subscribing to the [comment feed](#) for this post.

The comments to this entry are closed.

[Brass Tacks](#) [Powered by TypePad](#)