How to Configure SAN Solution and Avoid the Risk of Congestion

Objective

How to configure hosts and devices to use storage ports and minimize the risk of congestion and performance issues.

Environment

- All Storage, all devices and hosts.

Procedure

The "Hitachi Open Systems Connectivity Guide" recommendation states:

> "Note: Hitachi multipathing best practice is Single-Initiator Single-Targets configuration in which each HBA has only one path to the same LU. For example, you can provide four paths to each LU if you have four HBAs."

Therefore, follow these current best practices:

1. Recommended to configure each host HBA to actively and concurrently use only a single target storage port.
2. Recommended to match fabric negotiated speeds on each HBA and storage port.
3. Where multiple storage systems are used (for example, host mirroring) at least two HBA per storage system should be used (second HBA and path for resilience).
4. When using GAD with each HBA zoned to both primary (PVOL) and secondary (SVOL) sides use ALUA settings to set optimized and non-optimized paths.
5. There is also an additional Host Mode Option (HMO) 78 for setting non-preferred paths with GAD for HNAS or HDLM.

Additional Notes

All vendors of storage, host server and fabric switches are effected by buffer congestion due to sub-optimal configuration. **This is not only specific to Hitachi storage.**

Note that existing configurations can operate normally ignoring the above recommendations for extended periods but

may still be susceptible to the described buffer congestion. Impacts are not often noticed until after workload, device or infrastructure changes (for example, new servers, storage or switches).

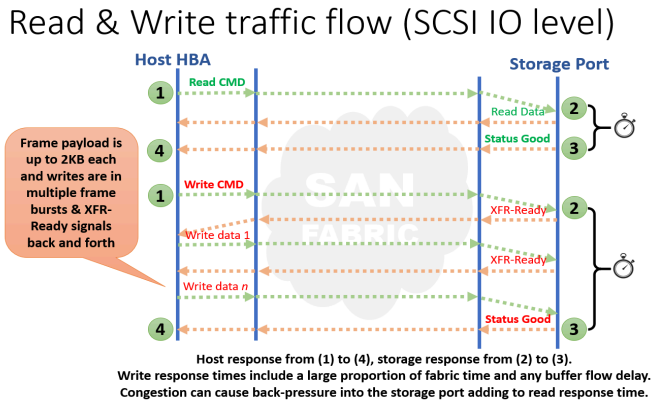
Some semi-intelligent host multi-path algorithms and throttling or QoS in the fabric switches or storage can assist with multiple targets and overcome buffer congestion issues with varying success. However, the issues described below cannot be eliminated entirely unless the solution is configured as described in the procedure above.

Host multi-path tuning and QoS facilities are outside the topic of this article and usually require consultancy assistance to POC test and implement. In effect these are tuning options and specific to every different implementation.

Deeper dive

The reasons for congestion when configurations are sub-optimal is a complex and deep technical topic. However, the notes below should assist in understanding.

When buffer congestion is occurring it is usually noticed firstly in higher than normal write response times. Usually when writes are sharing the same paths as large block and heavy reads (for example backup). To assist understanding, it is helpful to know how SCSI write operates and flows between devices differently than reads as below:



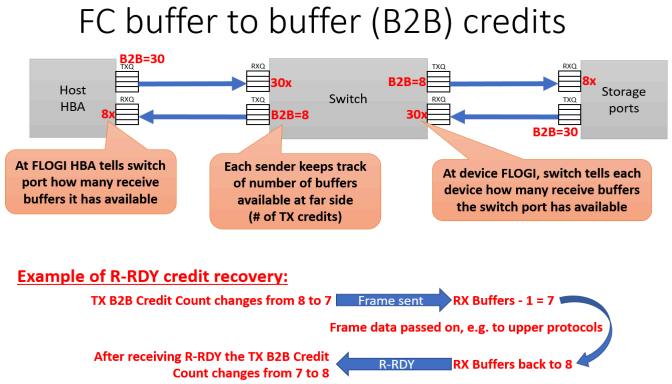
You can see above, after the SCSI write command (without any data payload) is received, the storage sends back an XFR-Ready signal to the host before the host sends any write data. The write operation contains at least four periods of SAN fabric time for the host response and at least two crossings of the fabric for the storage response. With large block writes multiple XFR-Ready and write data frame bursts can occur. Reads only cross the fabric twice from host perspective and once from the storage perspective (however, that does not exclude reads from suffering due to congestion as described later).

The flow control built into SCSI write protocol enable storage to pace the incoming data so that it does not suffer congestion at the lower FC protocol level. As reads do not have this flow control it is most common that congestion occurs in the direction of read frames and hosts are usually not able to pace incoming workloads.

It is also useful to have an understanding of how buffer credit flow operates at a fabric switch primitive FC level lower down the stack than the SCSI protocol and what causes buffer congestion at this level.

The diagram below shows how the sender keeps track of how many TX credits are available at the receiving end and

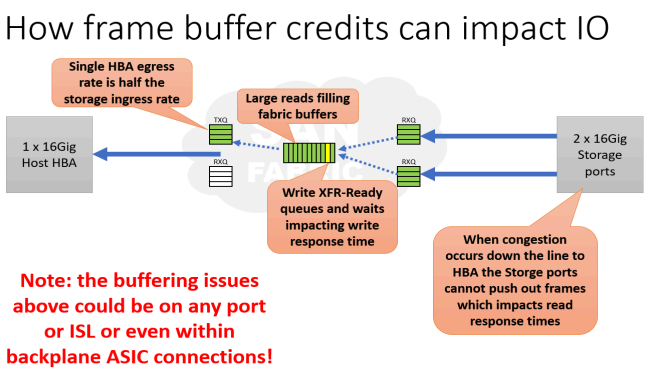
decrements the buffer to buffer counter. When the sender receives an R-RDY it increments the counter. However, once a counter reaches zero no more frames can be sent. If the device port is at credit zero for a length of time this is known as buffer starvation and most switches have counters named "buffer to buffer credit zero" for tracking these situations.



There are three main causes of buffer congestion:

- **Credit stalled device or slow drain.**
  - Receiver not sending credits (R-Rdy) back and senders Tx credits hit zero.
- **Lost Credits.**
  - Physical errors, credits and frames not sent reliably resulting in credit loss over time.
- **Overutilization or oversubscription.**
  - Device requesting more data than it can consume.
  - Speed mismatch, Fan-in mismatch.

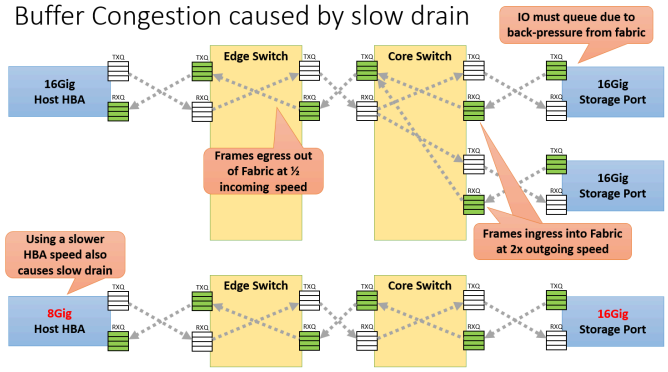
Below is an example of the most common sub-optimal configuration situation where a single HBA can be overwhelmed by the amount of frames put into the fabric by two storage ports:



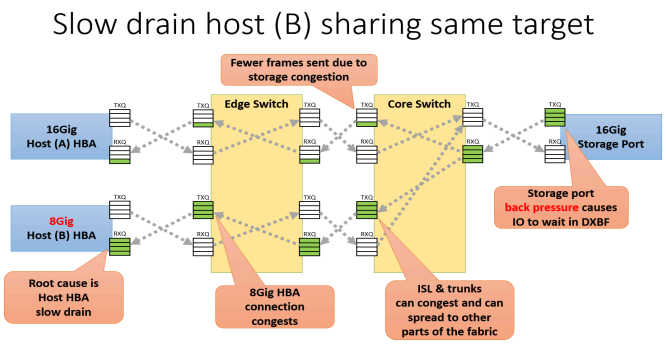
The example above shows buffer congestion but the impact and effect can be the same when experiencing bandwidth congestion too. In severe cases Hitachi arrays may also log **SSB=D555** diagnostics with reasons of "**Waiting for initial start**" for writes or "**Waiting for data transfer**" for reads or writes. In both cases it means that the storage is waiting on

data transfers to or from the external device.

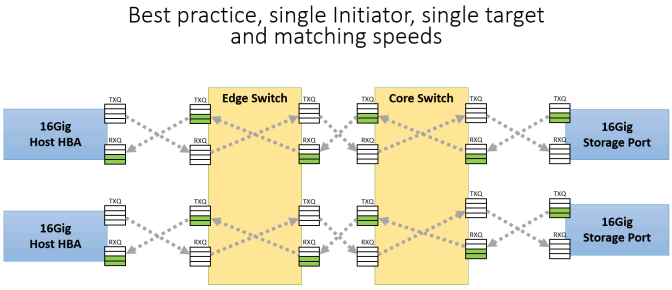
Below expands to show the impact of sub-optimal configuration across multiple hops and also provides an example of mismatched fabric speeds. When multiple switch hops or internal switch connections are in the paths the impact can be spread to multiple devices.



Below is another example this time showing a fabric victim Host(A) of **slow drain** in Host(B) causing **back-pressure** to the storage adapter. The storage adapter is unable to send out frames due to congestion further down the line

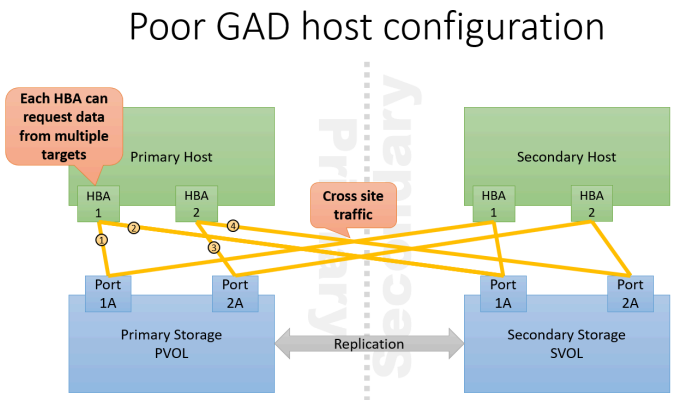


Below is an example of an optimal configuration which reduces the risk of congestion.

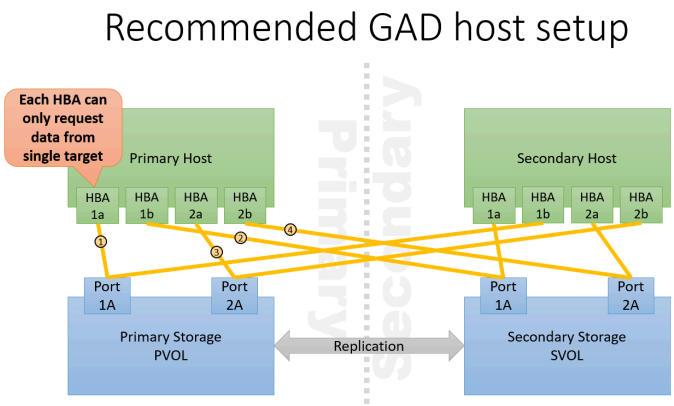


When implementing GAD the chances of congestion can be increased because a host is often implemented with the minimum of HBA but uses target ports from both primary (PVOL) and secondary (SVOL) storage systems as below. Sometimes the secondary host is not implemented but congestion can still occur due to multiple paths.

Assuming a round robin multi-path host configuration. Usually each group of four IO would cross paths (1) thru (4) in turn. Write IO on paths (1) & (3) would replicate from PVOL to SVOL and the IO on paths (2) & (4) would replicate SVOL to PVOL. When reads occur across all four paths, primary host HBA1 would receive frames from both storage ports 1A and HBA2 from both 2A ports increasing the risk of HBA buffer congestion.

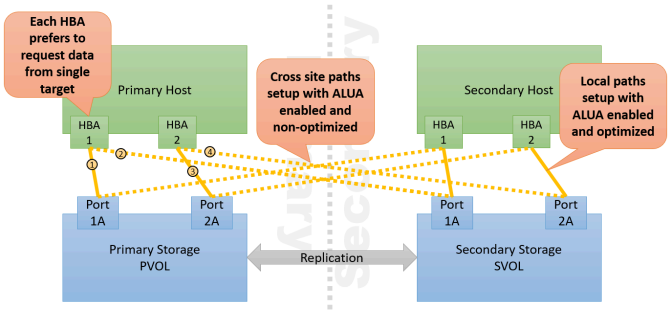


Below is the ideal way to setup hosts to use GAD but needs additional HBA adapters or HBA ports. Each path (1) thru (4) has a single target only.



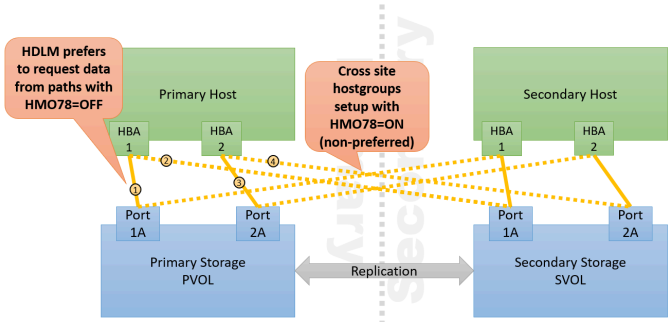
Using ALUA with GAD enables the host to avoid sending traffic down paths which have been marked as non-optimized. The example below shows that hostgroups defined on the secondary storage used by paths (2) & (4) are marked as non-optimized and the hostgroups on the primary storage paths (1) & (3) are optimized. Therefore, the non-optimized paths will only be used when the optimized path fails.

Recommended GAD ALUA setup

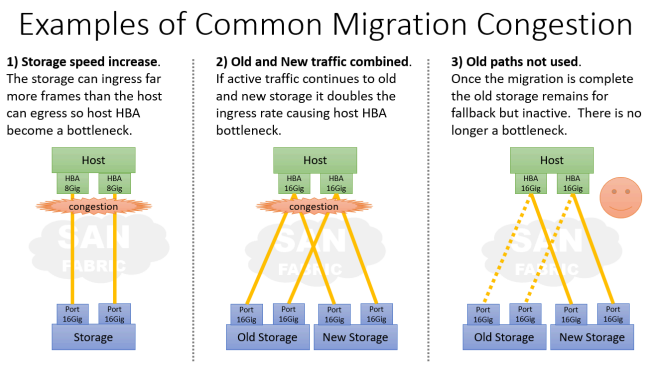


There is also an option for hosts without ALUA to use HDLM and Host Mode Option (HMO) 78 to set non-preferred paths. Note that HNAS can also use the functionality of HMO78 to set preferred (HMO78=OFF) and non-preferred paths (HMO78=ON):

Recommended GAD HDLM+HMO78
(Note HMO78 also sets non-preferred paths for HNAS)



Sometimes after upgrading to newer faster storage a customer can experience slower performance and this can be very puzzling, especially when the new storage is an improved and higher performing model. When migrating a legacy environment to any new storage, bottlenecks can move and performance issues introduced for a variety of reasons, sometimes due to congestion. This can be a common problem in particular when host hardware and switched fabrics are not upgraded. Below are some simple examples showing consideration of the host and storage. Of course ISL bandwidth and fan in/out ratios should also be considered but is beyond the scope of this article.



Other useful links:

[What Are Slow Draining Devices in Fiber Channel Environments]

Internal Notes

Please reach out to mark.butterworth@hitachivantara.com to discuss or improve this article.