Best Practices for Designing Highly Available Storage – Network for iSCSI

This is the 4th Part of a 6 part article, where I am talking about what are the things we should make sure while designing a highly available storage. Today I am going to talk about what it takes to design a robust network when we have a iSCSI Storage in place. Read my earlier posts.

Part 1: Best Practices for Designing Highly Available Storage – Host Perspective
Part 2: Best Practices for Designing Highly Available Storage – FC SAN Perspective
Part 3: Best Practices for Designing Highly Available Storage – iSCSI SAN Perspective

Avoiding iSCSI network congestion is our main concern for achieving iSCSI LAN performance. Network congestion is usually the result of an inappropriate network configuration or improper network settings. Network settings include IP overhead and protocol configuration of the network's elements.

For example a common problem is a switch in the data path into the storage system that is fragmenting frames. As a minimum, the following recommendations should be reviewed to ensure the best performance.

Simple network topology

Both bandwidth and throughput rates are subject to network conditions and latency.

It is common for network contentions, routing inefficiency, and errors in LAN and VLAN configuration to adversely affect iSCSI performance. It is important to profile and periodically monitor the network carrying iSCSI traffic to ensure the consistently high Ethernet network performance.

In general, the simplest network topologies offer the best performance. Minimize the length of cable runs, and the number of cables, while still maintaining physically separated redundant connections between hosts and the storage systems.

Avoid routing iSCSI traffic as this will introduce latency. Ideally the host and the iSCSI front end port are on the same subnet and there are no gateways defined on the iSCSI ports. If they are not on the same subnet, users should define static routes. However, multiple subnets is supported with VMware's Native Multipathing.

Latency can contribute substantially to iSCSI based storage system's performance. As the distance from the host to the storage system increases; a latency of about 1 millisecond per 200 kilometers (125 miles) is introduced. This latency has a noticeable effect on WANs supporting sequential I/O workloads.

For example, a 40 MB/s 64 KB single stream would average 25 MB/s over a 200 km distance.

Bandwidth balanced configuration

A balanced bandwidth iSCSI configuration is when the host iSCSI initiator's bandwidth is greater than or equal to the bandwidth of its connected storage system's ports. Generally, configure each host NIC or HBA port to only two storage system ports (one per SP). One storage system port should be configured as active, and the other to standby. This avoids oversubscribing a host's ports.

Network settings

Manually override auto-negotiation on the host NIC or HBA and network switches for the following settings. These settings improve flow control on the iSCSI network:

- Jumbo frames
- Pause frames

- TCP Delayed ACK

Jumbo frames

Using jumbo frames can improve iSCSI network bandwidth by up to 50 percent.

Jumbo frames can contain more iSCSI commands and a larger iSCSI payload than normal frames without fragmenting or with less fragmenting depending on the payload size. On a standard Ethernet network the frame size is 1500 bytes. Jumbo frames allow packets configurable up to 9000 bytes in length.

If using jumbo frames, all switches and routers in the paths to the storage system must support and be capable of handling and configured for jumbo frames. For example, if the host and the storage system's iSCSI ports can handle 4470-byte frames, but an intervening switch can only handle 4,000 bytes, then the host and the storage system's ports should be set to 4000 bytes.

Pause frames

Pause frames are an optional flow control feature that permits the host to temporarily stop all traffic from the storage system. Pause frames are intended to enable the host's NIC or HBA, and the switch, to control the transmit rate.

Due to the characteristic flow of iSCSI traffic, pause frames should be *disabled* on the iSCSI network used for storage. They may cause delay of traffic unrelated to specific host port to storage system links.

TCP Delayed ACK

On Microsoft Windows and ESXi based hosts, *TCP Delayed ACK* delays an acknowledgment for a received packet for the host.

TCP Delayed ACK should be *disabled* on the iSCSI network used for storage.

When enabled, an acknowledgment is delayed up to 0.5 seconds or until two packets are received. Storage applications may time out during this delay. A host sending an acknowledgment to a storage system after the maximum of 0.5 seconds is possible on a congested network. Because there was no communication between the host computer and the storage system during that 0.5 seconds, the host computer issues Inquiry commands to the storage system for all LUNs based on the delayed ACK. During periods of congestion and recovery of dropped packets, delayed ACK can slow down the recovery considerably, resulting in further performance degradation.