High-Performance SAN Extension Across Managed IP Infrastructures

Executive Summary

SAN, UDP, TCP, IP, MEP, WAN, MAN, SCSI. One of these acronyms can be dropped from this list without compromising Gigabit per second connectivity between islands of Fibre Channel networked storage. Which acronym it is might surprise you.

The unique nature of the communication between one Storage Area Network (SAN) and another favors the use of User Datagram Protocol (UDP) over Transmission Control Protocol (TCP). In fact, with today's technology, using UDP as a layer-4 Internet Protocol (IP) transport is the only way to sustain Gigabit per second connectivity across Metro- and Wide-Area Network distances.

The key to this innovative idea lies in the combined characteristics of the long-distance network infrastructure, and the communication between SANs. Offering carrier-grade Ethernet connectivity, managed IP networks eliminate the need for TCP congestion control features, originally intended to provide robustness in unreliable networks. Furthering the redundancy of TCP, the Small Computer Systems Interface (SCSI) protocol, present in every transaction between networked storage devices, offers error-recovery logic that provides integrity to UDP transmissions.

As the SAN is extended over the Metro-Area Network (MAN) and Wide-Area Network (WAN), so too the advantages of networked storage are spread across the organization. Freed from isolation, the SAN islands benefit from shared access to remote resources, like tape drives and tape silos, and the availability of applications previously impossible, synchronous mirroring of data between devices separated by thousands of miles, for example.

SAN Valley Systems provide the mechanism for the successful and efficient transmission of SAN traffic across MAN and WAN distances. The SAN Valley SL1000 IP-SAN Gateway utilizes UDP in transmitting encapsulated Fibre Channel traffic between islands of networked storage. Offering the best IP protocol choice for wire-speed SAN connectivity, UDP enables Gigabit per second extension of the SAN across managed IP networks.

Introduction

According to the latest data from IDC, by 2004 projected Fibre Channel ports in use by SANs will exceed 6.7 million. Of these, 20% will be used to extend SAN connectivity across the MAN or WAN, creating a $2B market for SAN extension technology.

The growing need for high-speed sharing of geographically isolated storage resources is driving the extension of the SAN infrastructure across the Metro- and Wide-Area Network. Cost savings, achieved through consolidating backup and recovery operations, have been cited by the Washington DC-based Working Council For CIOs as a primary benefit from SAN implementations. Broadening SAN connectivity, allowing greater access to remote tape drives and tape vaults, extends cost reduction benefits throughout the organization.

The heightened awareness of business continuance planning, whether for disaster recovery or in anticipation of outages from routine maintenance, is focusing attention on synchronous and asynchronous mirroring applications. Preparing for a disaster event, like an earthquake, flood or terrorist attack, requires that data facilities hosting mirrors be separated by hundreds, if not thousands, of miles. The success of such long-distance mirroring is dependant on Gigabit per second transmissions across the extended storage network.

Connecting the SAN islands across thousands of miles presents a physical challenge. Fibre Channel limits the distance between nodes in a SAN to 10 kilometers, about 6 miles. This limitation is forcing the consideration of other network technologies to provide long-distance connectivity. IP networking, free from such distance constraints, offers a solution.

Connecting SANs Islands over IP Networks

The growth of corporate IP-based applications is giving rise to a rapid build-out of high-bandwidth managed Ethernet networks. Delivering Gigabit per second connectivity, across tremendous distances, Metro Ethernet Providers (MEPs) like Yipes, Cogent and Telseon, now offer optical, carrier-grade networks connecting multiple access points.

Just as airlines over book flights to ensure fully utilized airplane capacity, hoping all ticketed passengers don't show up at once, networks can be over-subscribed, with "bumped" data resulting in packet-loss errors. Unlike the overcrowded, and unpredictable, public networks that host the Internet, private, managed Ethernet networks provide enough capacity for network vendors to offer congestion-free guarantees. This all but eliminates the errors arising from network congestion.

The MEPs also employ advanced traffic management techniques and operate fully redundant network architectures, delivering availability approaching 99.999%. These congestion-free, and reliability, guarantees are frequently codified in stringent Service Level Agreements (SLAs), holding the provider financially responsible for any lapse in service delivery.

The wide availability of high-quality, managed Gigabit Ethernet networks solves the problem of physically connecting SANs across MAN and WAN distances, shifting attention to the next bottleneck: the transport layer protocol. Historically TCP and UDP, the two choices for layer-4 IP transport, have either been considered too cumbersome or too unreliable for block-level SAN traffic. But this view needs to be reevaluated in the light of the Gigabit connectivity offered by the MEPs.

Transmission Control Protocol

TCP is a robust, highly resilient transport standard, and the preferred medium of exchange when communicating over IP networks. With built-in flow control and retransmission logic, TCP is ideal for the congested and error prone public networks that make up the Internet. The price of such ubiquity is performance. Each feature that adds robustness also increases latency, slowing down network throughput.

TCP windowing offers an example of a feature that adds overhead, but is redundant in the congestion-free world of managed IP networks. Designed to relieve network congestion by regulating the flow of traffic, window size dictates the number of data packets a sender transmits before requiring acknowledgment from the receiver. Window size varies during transmission, depending on the stability of the network. If packet-loss errors are detected, window size decreases, slowly increasing again as the network proves to be stable. The intermittent degrading and increasing of network throughput, caused by windowing, is a characteristic of TCP networks.

Windowing is both a strength and weakness of TCP. In unpredictable networks, windowing provides essential flow control. However, in the controlled environment of managed IP networks, where congestion is all but eliminated, windowing is redundant, adding processing overhead to each data packet.

Evaluating IP Transports For Storage Traffic

Despite its notorious CPU-sapping and latency-generating properties, TCP remains the ubiquitous first choice in IP layer-4 protocols. Although efforts are under-way to provide TCP Offload Engines (TOEs) that will cast the protocol stack in silicon, dramatically improving performance, these TOE technologies are new, unproven, and as yet, unavailable on the market.

As an alternative, UDP offers a streamlined, connection-less protocol capable of traveling unhindered across IP networks at wire-speed. Often regarded as the poor relation to TCP, because it lacks error-recovery and flow-control characteristics seen as necessary to guarantee delivery of data over unpredictable networks, UDP is, in fact, widely used in performance-critical applications such as Voice-Over-IP and streaming video. Long the mainstay of the mid-range community, UDP also provides the primary transport for UNIX Network File System (NFS) traffic.

SCSI Error Handling

Regardless of the transport protocol used to deliver storage traffic across an IP network, the SCSI protocol is the last stop before the data reaches the application. Positioned at layer-7, the application layer, in the Open Systems Interconnect (OSI) network model, SCSI is the standard mechanism for communicating block-level storage commands to a networked storage device.

SCSI offers built-in error checking and recovery features that overlap with the error-recovery of TCP, creating unnecessary redundancy in the path of storage traffic. Using UDP as the transport layer, network errors are detected by the SCSI driver, which then directs the storage device to retry the Input/Output (I/O) request (see Figure 1). In effect, this boosts the integrity of UDP, ensuring errors do not get through to the application, and leveling the playing field for layer-4 protocol choice.

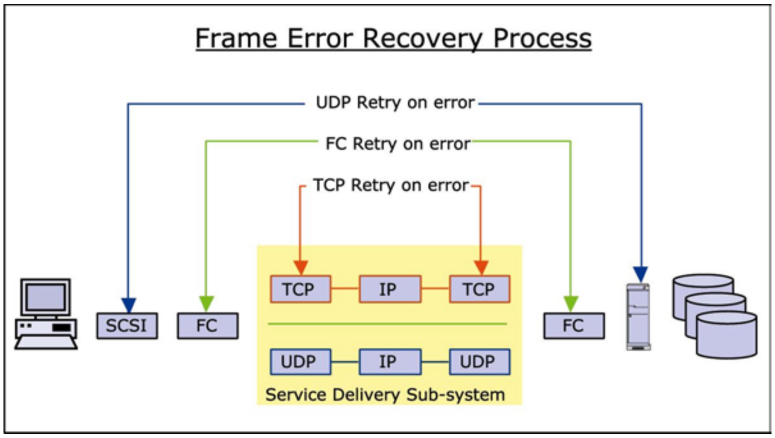

Figure 1. Comparing error recovery in FC, TCP and UDP.

Impact of Errors on Application Throughput

The Gigabit-level connections of managed IP-networks provide the same performance, capacity, and quality of service guarantees as the dedicated enterprise Fibre Channel networks used for storage networking. In fact, with a Bit Error Rate (BER) of 10^{-12}, corresponding to one error per Terabit, Gigabit Ethernet and the Fibre Channel Protocol have identical formal error rate specifications: BER being the primary determinant of packet-loss errors in a congestion-free network.

Despite the low occurrence of transmission errors in managed Ethernet networks, a BER of 10^{-12} equates to one error every 1000 seconds when transmitting data at Gigabit per second rates. TCP and UDP treat the recovery from errors differently, with significant consequences for throughput of storage traffic. Demonstrating this difference can best be achieved by focusing on a typical application scenario.

Synchronous Mirroring – A Case Study

Remote mirroring applications are one of the primary drivers of SAN extension technology. The ability to synchronously, or asynchronously, duplicate I/O operations in a remote location gives corporate business systems a high-degree of disaster tolerance. If anything happens to the primary data storage facility, applications can fail-over to a geographically separated mirror copy of the data, providing continuous availability to business users in the event of a disaster.

A typical storage system will be configured with multiple logical units (LUNs), the number of LUNs being a function of total available disk capacity and LUN size. As business applications generate writes on the storage system, each LUN spins off a separate I/O process (IOP) to perform the write operations. In a mirrored environment, with two systems of identically configured devices, every write operation on the primary LUN is matched by a corresponding write to the remote LUN.

TCP Error Handling

TCP manages the detection of network errors, and the retransmission of data, in the protocol stack. As packet loss errors occur, TCP ACK transmissions from the receiver indicate to the sender that a resend of a data packet is necessary. The data transmissions in synchronous mirroring represent multiple flows, as each LUN conducting write operations sends data to its mirror. TCP, however, transmits the traffic as a single stream. When an error occurs, the entire transmission is blocked, halting all LUN writes, pending receipt of the correct sequence of data packets.

UDP Error Handling

UDP has no method for detecting transmission errors at the protocol stack layer. Data packets are streamed directly from sender to receiver with line errors passed up the stack to the application, in this case the SCSI driver. The SCSI protocol monitors for errors, and retries the IOP in the event of a problem. Because error handling happens at the level of LUN-IOP communication, only a small portion of the total stream is blocked, waiting for the correct data to be retransmitted. A remote mirroring operation with hundreds of LUNs will have the write operations of only one LUN blocked, with the remaining LUNs continuing to successfully mirror writes.

Simulating Network Throughput

Figure 2 describes a model[1] simulating application performance of UDP and TCP, and compares maximum achievable network throughput in the presence of errors. The x-axis represents Packet Error Rate (PER) which can be computed from BER, assuming typical packet sizes based on Ethernet (1500 byte) and Fibre Channel (2178 byte) frames.

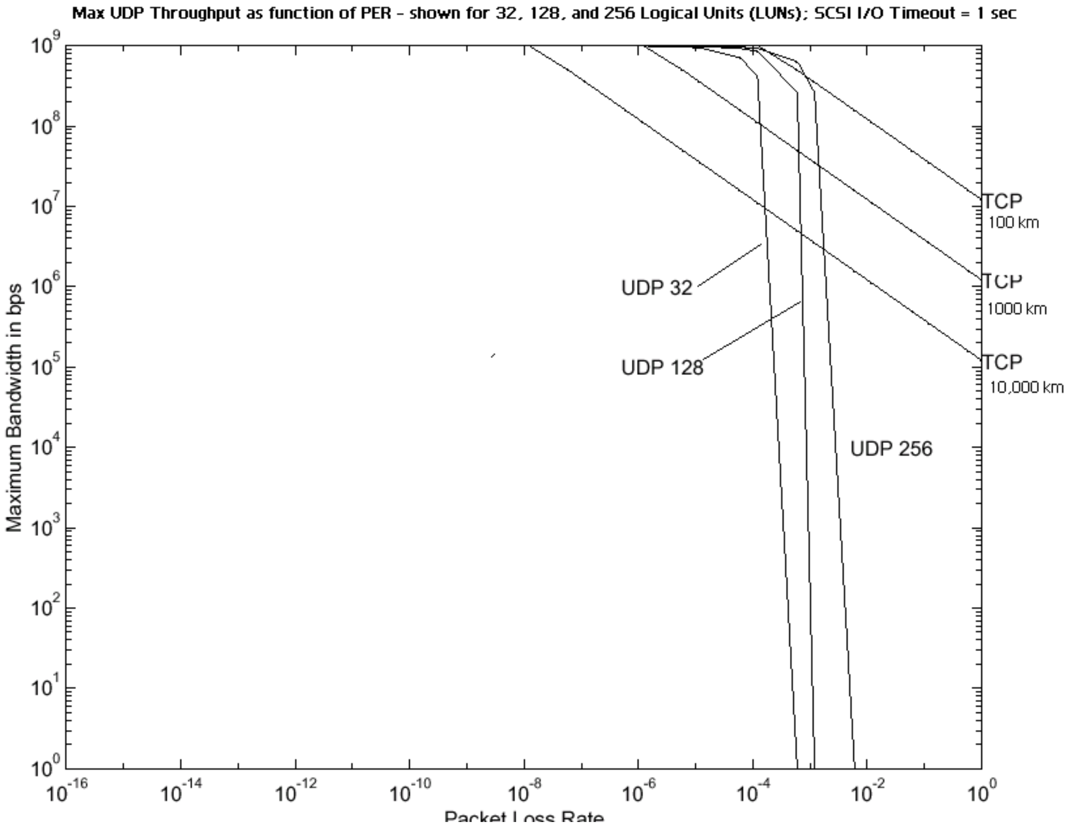

Figure 2. Software simulation of application performance comparing TCP and UDP.

In the simulation, UDP performance is modeled to cover a range from 32 to 256 LUNs. It is easy to see that for each of these cases, the achievable throughput does not drop until PER drops to 10e-5. This is well below the PER of 10e-8 expected for Metro-Ethernet services over optical networks. The models used in the analysis assume that TCP is capable of reaching theoretical maximums. In real-world implementations, TCP cannot reach these maximums.

The effect of the difference between UDP and TCP transports on network throughput is significant. At packet-loss error rates well above those anticipated for managed Ethernet networks, UDP is able to maintain wire-speed transmissions, matching or exceeding TCP performance.

[1] 1. "The Macroscopic Behavior of TCP Congestion Avoidance Algorithm" by Matthew Mathis et al, originally published "Computer Communications Review" ACM SIGCOMM volume 27 number 3 July 3, 1997.

SAN Valley Systems and UDP

With a pared-down protocol stack, UDP offers the most effective means of transporting storage data at wire-speeds, across MAN and WAN distances. By eliminating the overlap between TCP and SCSI error checking, UDP is able to streamline the connection between Fibre Channel storage devices, without compromising data integrity.

The successful implementation of corporate disaster recovery and business continuance applications, connecting SANs across hundreds of miles, is dependant on reliable, wire-speed communication. The wide availability of carrier-grade Gigabit Ethernet connectivity, combined with the in-built integrity of the SCSI protocol, pave the way for UDP as the optimum choice of layer-4 IP transport protocol.

Recognizing the opportunity to leverage UDP in providing Gigabit speed data transport between SANs, SAN Valley Systems has developed the SL1000 IP-SAN Gateway. Encapsulating Fibre Channel storage traffic in hardware, the SL1000 utilizes UDP to transport data across MAN and WAN distances with absolute assurance of data integrity. The SAN Valley Systems SL1000 makes the geographical location of SANs irrelevant, extending the value of storage networking across the entire enterprise.