Impact of Packet Loss on BC/DR Deployments

Introduction

Business Continuity and Disaster Recovery (BC/DR) are the business drivers for deploying wide area networking of storage. Whether recovering from a total loss caused by either a natural or a man-made disaster, or simply recovering from an operational mishap resulting in a partial loss of data, companies must be able to plan for and recover from any scenario in order to sustain their business.

In planning for wide area deployments, IP networks are the network of choice. They have traditionally served well in the arena of file-level access. There is an existing Ethernet knowledge base, many network management tools are available, and the technology is well understood. IP infrastructures are in place, and IP network bandwidth is available.

There are several uniquely different deployments of storage over IP (e.g. iSCSI, FCIP, iFCP, as well as native Ethernet), but they all share the same underlying IP technology. In addition, there are different methodologies for deploying storage BC/DR applications. Applications may replicate changes from a source volume to a destination volume as they occur, or a more traditional backup methodology may be deployed, in which volumes of data are periodically backed up in their entirety to a remote site.

The particular method being deployed dictates the performance requirements of the IP network, as each method generates a unique workload that is required to be completed within a particular timeframe. In the case of real-time replication, the demands on the IP network are such that changes made on source volumes must be replicated to the remote volumes without violating service level requirements for providing access to both the source and destination volumes. This means that the network bandwidth and replication methodology must be sufficient to be able to deliver the data to the destination volumes at a rate equal to the rate at which the data is being changed on the source volumes. In the case of volume backup, backups must be completed within predefined windows of time. This means that the network bandwidth must be sufficient to be able to deliver all of the data to the destination volumes at a rate such that it can be completed within that period of time.

There are three key impediments that determine the quality of IP data transport services over a wide area network.

- packet loss refers to the percentage of packets that are not delivered to their destination, and must be retransmitted
- latency refers to the round-trip-time (rtt) between the source and destination
- jitter refers to a variation in the arrival rate of packets due to latency fluctuations

These impediments cause performance degradation to occur for TCP/IP applications. This paper focuses on the causes of packet loss and its effect on TCP/IP application performance.

Causes of Packet Loss

Packet loss occurs when packets are damaged and discarded, or when the capacity of an intermediate network component is exceeded, which results in packets being discarded. Packets can be damaged as they move across a network, or as they traverse network components such as routers and switches. This type of damage is detected by a failure in "checksum" processing. The checksum is a mathematical sum of bits that is calculated by the sender and appended to each packet. The receiver also calculates the checksum and compares its calculated value to the value received with the packet. If the received and calculated checksums do not match, the receiver drops the packet. Regardless of the network topology, there is always a chance that some level of packet loss may occur due to checksum detected errors, especially as larger numbers of routers and switches are traversed, or over telephone service links covering the "last mile" of a network connection.

When the capacity of an intermediate network component is exceeded, congestion occurs at that component and packets will be discarded. For example, if packets arrive at a router at a rate faster than the router can store them or transmit them, some number of packets will be discarded by that router.

Service level policies can be in place, guaranteeing specific service levels to particular groups of applications. In order to meet these service levels, routers will intentionally discard packets from data streams that are outside these groups of applications.

Packet Loss Is Real

All IP service providers monitor their networks and report various statistics on average latency and packet loss rates. Round trip latency is expressed in terms of milliseconds, and represents the time it takes a packet to travel from source to destination and back. Latency includes both the transmission time on the physical media, and the time it takes a packet to travel through routers and switches on the network. Packet loss is expressed in terms of a percentage, which represents the average packet drop rate. For example, a 1% packet loss rate (which would be extremely severe) indicates that 1 of every 100 packets is being dropped. More typical packet loss rates are on the order of 0.1% or 0.01%, which indicate that 1 of every 1,000 or 1 of every 10,000 packets are dropped.

The numbers reported by service providers are just averages, usually collected in 15 minute time intervals throughout the course of a month. Most intervals are reported as clean (i.e. 0% packet loss), but yet the monthly average is typically reported in the range of 0.01% to 0.03%. This means that if 90% of the intervals report 0% packet loss, the remaining 10% of the intervals must average a 0.3% packet loss rate in order to achieve a 0.03% average for the month. If 95% of the intervals report 0% packet loss, the remaining 5% of the intervals must average a 0.6% packet loss rate in order to achieve a 0.03% average for the month. If 99% of the intervals report 0% packet loss, the remaining 1% of the intervals must average an incredible 3% packet loss rate in order to achieve a 0.03% average for the month. These high loss intervals represent accumulations of time ranging from 7 to 72 hours over the course of a month, during which time a very disruptive packet loss condition would be occurring.

Unlike IP applications such as e-mail or web serving, storage replication applications are company-critical, and involve sending large amounts of data over long periods of time. The chance of a replication running during one or more of these degraded time intervals sometime during the month is quite high, since replication applications tend to be characterized either by the use of long standing connections, or by connections that reoccur at predefined time intervals. In either case, storage replication applications use TCP/IP connections for long periods of time, either as one long-standing connection, or the sum of multiple shorter connections.

The amount of data that gets sent over the storage replication connections actually depends on the methodology of the replication being performed. If volumes are being resynchronized, or if a traditional backup is being performed, then all of the data from the source volumes is sent to the target volumes. If the replication supports a mechanism of recognizing and tracking changes to the source volumes, then just the changed data is sent from the source volumes to the target volumes. In this case, the amount of data that actually gets sent is dependent on the frequency and number of updates to the source volumes, as well as the amount of data that is changed. But in all cases, replicating data over a network that is experiencing significant packet loss may cause the replication application to experience significant delays, or possibly even to miss production windows.

TCP and Packet Loss

In order to satisfy required performance levels, it is crucial that the data transport used by storage replication applications over a wide area network can recover from packet loss conditions as efficiently as possible.

When using TCP/IP as the data transport, there are two indications that packet loss has occurred: either a timeout occurs on a segment before receiving an acknowledgement for that segment; or duplicate acknowledgements are received. Regardless of whether the packet loss was caused by intentional discard due to congestion, or by actual data corruption, TCP/IP interprets both of these indications as congestion existing somewhere

in the network between the source and the destination, and reduces its throughput capacity. However, if the packet loss was not really caused by congestion, this results in an unnecessary reduction in throughput.

TCP Windows

Two TCP/IP windows are used to control the amount of data that can be "in the air" before an acknowledgement is required: the receiver's advertised window, and the sender's congestion window. The receiver's advertised window provides flow control to the sender that is based on the receiver's ability to accept data. The sender's congestion window limits the number of segments that can be injected into the network. It provides flow control for the sender that is based on the sender's assessment of perceived network congestion. When TCP/IP sends data, the amount of data that is sent is limited to the minimum of the advertised window and congestion window sizes.

TCP/IP windows scale very well to network latency, especially if both sides of the TCP/IP connection support the window scaling feature, which provides support for large windows. Large windows are necessary to achieve high performance on high bandwidth and/or high latency networks. However, during periods of high packet loss, the TCP/IP recovery algorithms reduce the size of the windows, which prevents them from achieving the size needed to scale to the network latency.

When packet loss occurs, TCP/IP assumes it is due to congestion, so immediately reduces its transmission rate. It saves the current value of the window size, then immediately reduces the rate at which data can be sent by decreasing the window size to one-half of its current value, and limiting the rate at which the window size can grow. If the congestion was detected by the occurrence of a segment timeout, TCP/IP makes a further adjustment by reducing the size of the sender's congestion window to one segment.

When new data is acknowledged after congestion is detected, the congestion window is increased in a way that depends on whether TCP/IP is in slow start or congestion avoidance. It is in slow start if the congestion window is less than half of the window size that existed at the time the congestion was detected; otherwise it is in congestion avoidance.

Slow Start

In slow start, the window is opened exponentially by setting the size of the congestion window to one segment, then doubling the number of segments every time an acknowledgement is received. This results in sending one segment, then two, then four, etc. Slow start continues until the congestion window reaches half of the window size that existed at the time the congestion was detected. At that point, TCP/IP goes into congestion avoidance.

Congestion Avoidance

In congestion avoidance, the window is opened linearly by increasing the size of the congestion window by a maximum of one segment for each round trip time. This allows

the sender to slowly increase its transmission rate as it approaches the point where congestion had previously occurred.

When a TCP/IP connection is recovering from packet loss, it may take a long time for the window to be reopened to a sufficient size to satisfy performance requirements. If additional packet loss occurs during this recovery, TCP/IP again reduces the size of the window and limits the rate at which the window can grow. If the window is repeatedly smaller than the available network capacity, it will be impossible for TCP/IP to ever fully recover to the required performance levels.

Figure 1 illustrates the effect of various packet loss rates on an iSCSI TCP/IP connection at 4ms and 10ms round-trip times. As the packet loss rate increases, the overall throughput decreases due to the inability of the windows to grow. The slow start and congestion avoidance algorithms keep the windows small, which prevents the windows from being able to scale to cover the added latency.

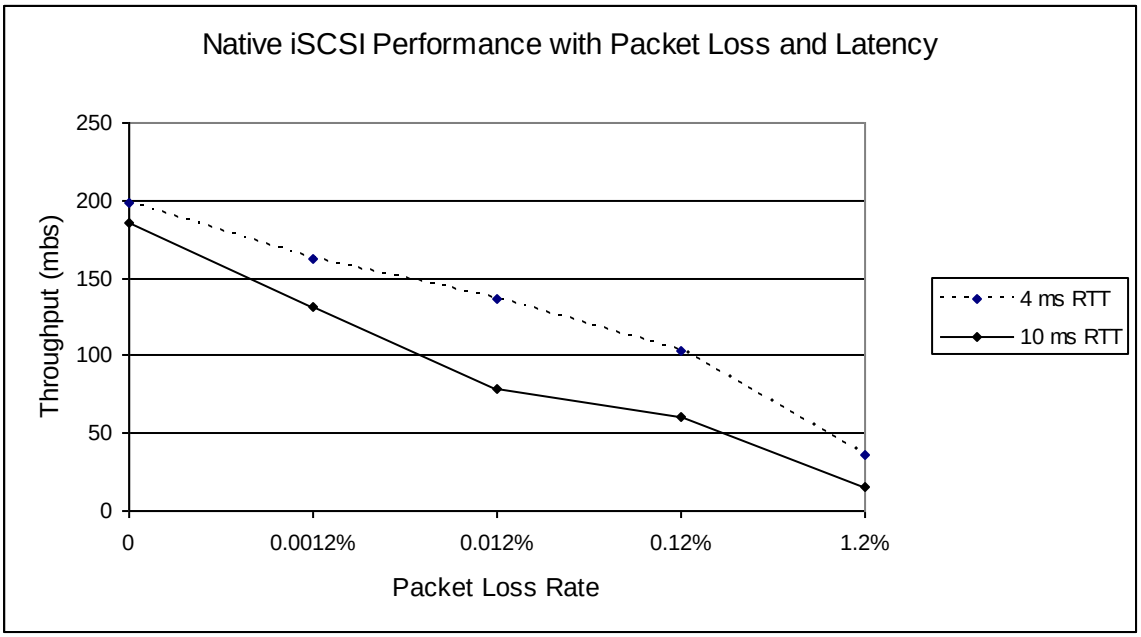

Figure 1. TCP/IP (iSCSI) Performance with packet loss and latency of 4 ms and 10 ms RTT.

Selective Acknowledgement

Selective Acknowledgement (SACK) has been deployed in many TCP/IP implementations. SACK provides more efficient recovery from packet loss by minimizing the number of packets that must be resent when packet loss occurs. However, SACK does not eliminate the rate reductions and window size adjustments that occur as a result of the congestion control algorithms.

High Bandwidth Networks

Packet loss has a more significant impact on high bandwidth networks, even over short distances. High bandwidth networks require large windows to fully utilize the network

capacity. However, the slow start and congestion avoidance algorithms may prevent large windows from being achieved when packet loss occurs.

Multiple TCP Connections

Some storage replication methodologies attempt to mitigate the packet loss problem by maintaining multiple TCP/IP connections. However, each TCP/IP data stream is responsible for its own bandwidth management, slow start recovery, and congestion avoidance. Each connection manages its own view of the link, and has no knowledge of the other connections. During periods of packet loss, each connection suffers and rescales independently of the other connections. Having too many TCP/IP connections coming and going may actually be counterproductive, since that may contribute to a congestion problem.

Figure 2 and Figure 3 illustrate examples of the performance impact on a storage replication application that uses four TCP/IP connections. In Figure 2, there is no additional network latency introduced, so the performance impact is completely caused by the packet loss. In Figure 3, a 10ms round trip time is introduced to show the additional impact on performance caused by packet loss when the reduction of window sizes further impacts the ability of the windows to scale to the network latency.

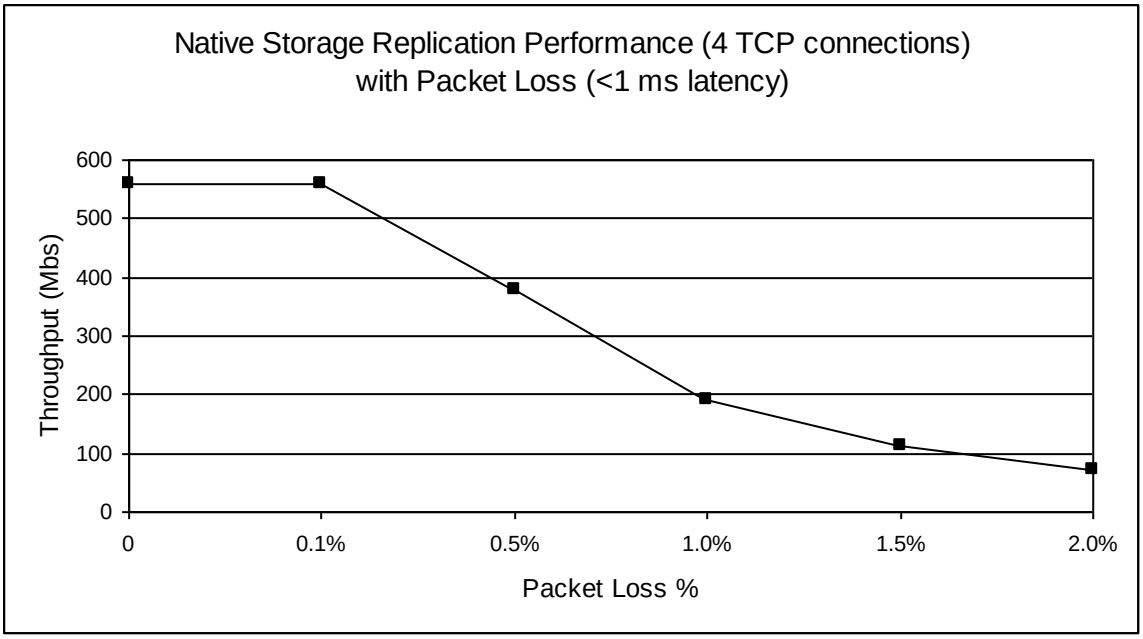

Figure 2. Storage replication performance of four TCP/IP connections with packet loss.

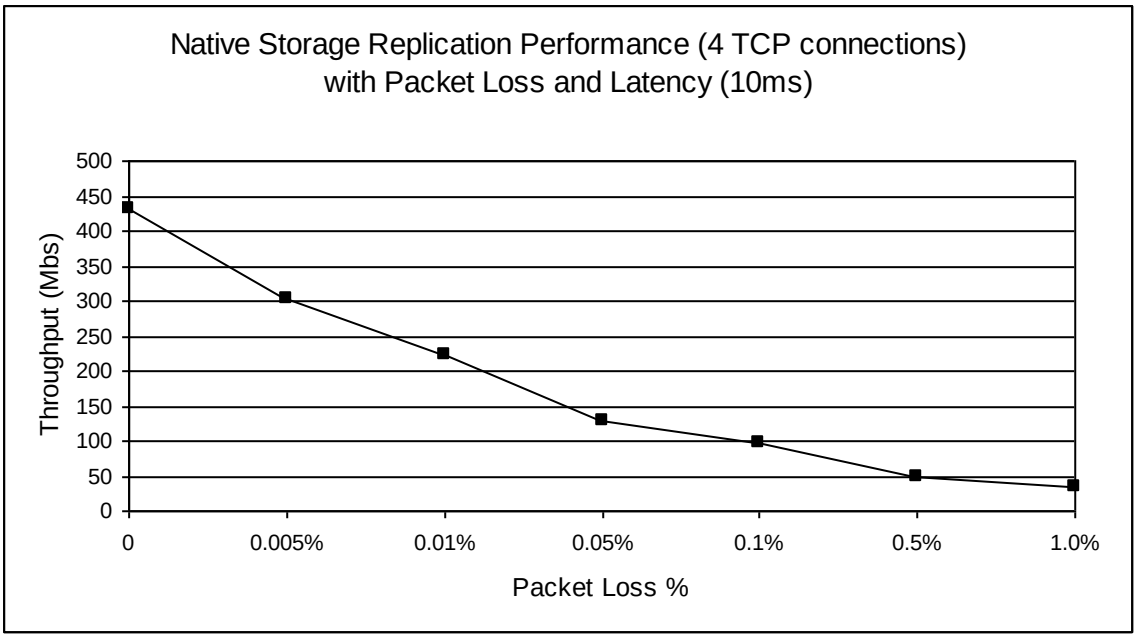

Figure 3. Storage replication performance of four TCP/IP connections with packet loss and latency.

HyperIP

HyperIP® is a virtual network appliance that enhances the performance of large data transfers over high bandwidth networks. Due to the critical requirements of storage replication, it is particularly well suited to working with storage replication applications. HyperIP provides a highly efficient data transport, especially in cases where packet loss, congestion and/or latency exist. A HyperIP configuration consists of an virtual appliance on both sides of a connection that connects to a virtual LAN switch. High Availability configurations consist of a pair of appliances on both sides of the connection. HyperIP optimizes TCP/IP connections by repackaging the TCP/IP data streams into a more efficient transport protocol that is used between the HyperIP appliances.

HyperIP dynamically calculates round-trip times, bandwidth capacity, and transmission rates, and uses that information to calculate the capacity of the network. This network capacity essentially becomes the window size used over the HyperIP connection, and is independent of any application TCP/IP window sizes.

HyperIP and Packet Loss

Just as with TCP/IP, HyperIP cannot distinguish between packet loss due to congestion and packet loss due to corruption. However, rather than making window adjustments when packet loss occurs, HyperIP uses rate-based congestion controls to limit throughput. HyperIP will subsequently attempt to increase the send rate, so that when the event causing the packet loss subsides, the receiver will be able to sustain a higher rate.

Figure 4 illustrates the performance of HyperIP for the same iSCSI test shown in Figure 1. In this case, the performance achieved by HyperIP is the same for both the 4 ms and

10 ms tests, and is higher than both of the native iSCSI tests, regardless of the packet loss rate.

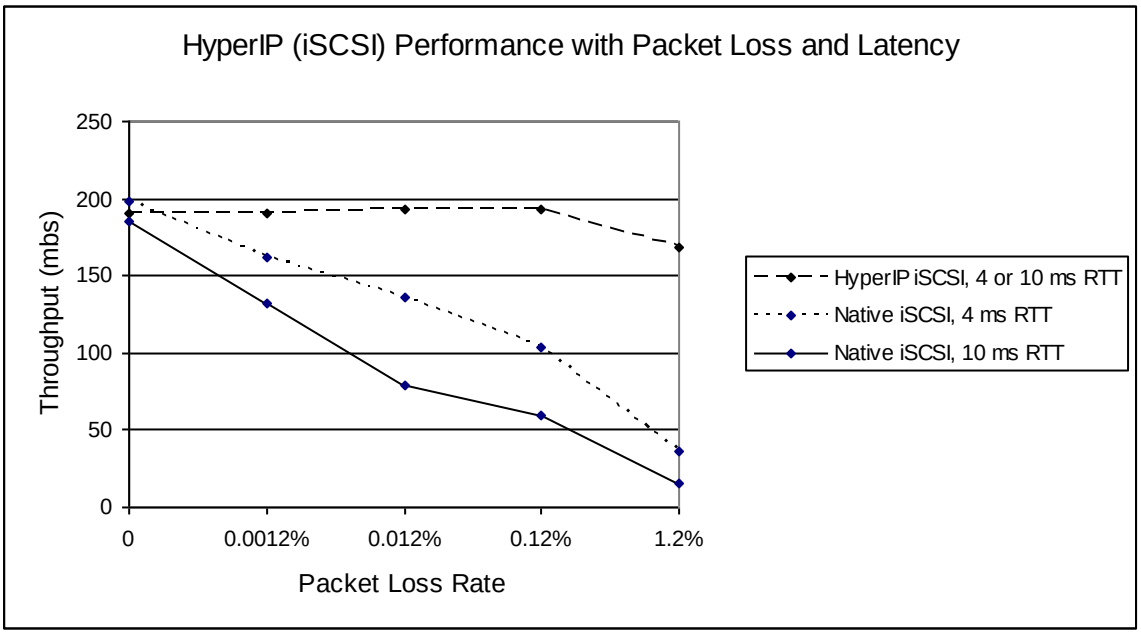

Figure 4. HyperIP iSCSI performance with packet loss and latency.

Figure 5 illustrates the performance benefit of HyperIP for the same storage replication test shown in Figure 2. This graph illustrates the value of HyperIP in mitigating the performance degradation caused by packet loss. As shown by this figure, HyperIP is able to sustain the throughput, even with the high packet loss rates.

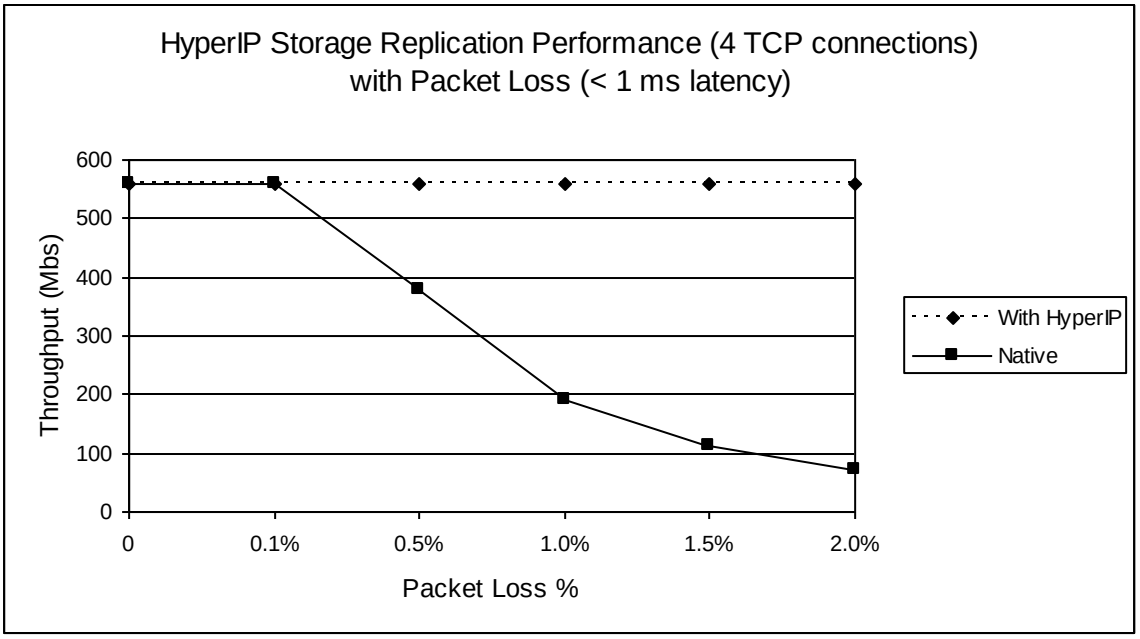

Figure 5. HyperIP storage replication performance with packet loss.

Figure 6 illustrates the performance benefit of HyperIP for the same storage replication test shown in Figure 3, when both packet loss and network latency are experienced. As shown by this figure, HyperIP is able to mitigate the performance degradation caused by packet loss by sustaining the same application throughput even when additional latency exists on the network.

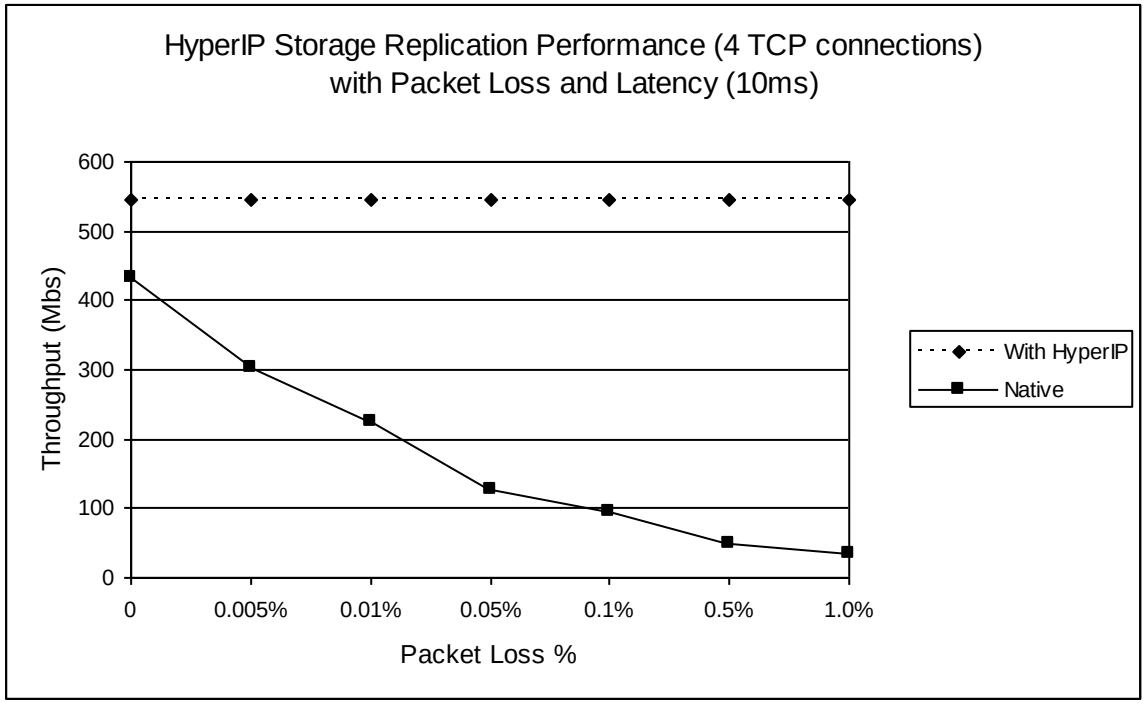

Figure 6. HyperIP storage replication performance with packet loss and latency.

Efficient Bandwidth Management

HyperIP manages multiple LAN TCP data streams, and aggregates them over a HyperIP connection. As new TCP application connections are started, HyperIP is able to accommodate the additional workload by inserting the new data into the HyperIP connection without creating congestion. As TCP applications are stopped, the additional bandwidth capacity is automatically reclaimed by HyperIP for sharing among the remaining connections.

Because of this aggregation capability, HyperIP provides an efficient mechanism for mitigating performance degradation caused by packet loss. When packet loss occurs, HyperIP makes adjustments on its rate and window size, rather than having multiple connections each making their own adjustments independently of each other. When HyperIP is the sole driver of the WAN, it has accurate knowledge of the WAN capacity and sustainable link rate.

Summary

HyperIP provides significant value to TCP/IP storage replication applications. HyperIP shields the TCP/IP applications from performance variations due to packet loss and latency, since the performance over the WAN is managed by HyperIP. Packet loss will not significantly impact HyperIP's ability to efficiently drive the replication data over the wide area network.

By leveraging existing network infrastructures, IP-based storage replication applications make BC/DR deployments affordable to a growing number of customers. However, there is a price to pay for this convenience. IP networks may already be in place, but deploying BC/DR IP replication applications places a whole new set of requirements on the IP infrastructure:

- Real-time replication deployments must be able to deliver replicated data to the BC/DR site at the rate at which the source data is being changed
- Traditional backup or snapshot deployments must ensure that backup windows are met

With these time-sensitive requirements, IP-based storage replication deployments cannot tolerate performance slowdowns caused by packet loss on the IP network. HyperIP mitigates these performance slowdowns by managing the replication data over the wide area network, and shielding the replication applications from these performance variations. Regardless of the replication application that is used, or the type of deployment, HyperIP maximizes the efficiency of transporting replication data over IP networks, and makes IP-based BC/DR deployments not only affordable, but also feasible.

Network Executive Software, Inc. (NetEx)
6450 Wedgwood Road North
Suite 103
Maple Grove, MN 55311
(763) 694-4300 or (888) 604-5573
www.netex.com