

The Data Replication Bottleneck: Overcoming Out of Order and Lost Packets across the WAN



By Jim Metzler
Jim@Kubernan.Com

Background and Goal

Many papers have been written on the effect that limited bandwidth and high latency have on application performance across the Wide Area Network (WAN). The purpose of this document is to address a less commonly understood WAN challenge that can also affect the performance of critical business applications – packet loss and reordering.

While packet loss and out of order packets are a nuisance for a network that supports typical data applications¹ like file transfer and email, it is a very serious problem when performing data replication and backup across the WAN. The former involves thousands of short-lived sessions made up of a small number of packets typically sent over low bandwidth connections; the latter involves continuous sessions with many packets sent over high capacity WAN links. Data applications can typically recover from lost or out of order packets by retransmitting the lost data. Performance might suffer, but the results are not catastrophic. Data replication applications, however, do not have the same luxury. If packets are lost, throughput can be decreased so significantly that the replication process cannot be completed in a reasonable timeframe - if at all.

This paper will discuss why packet loss and ordering is becoming a bigger problem in today's WANs, and what can be done to overcome these packet delivery challenges. To achieve this goal, this document will integrate theory and practice. In terms of theory, this document will include a widely accepted model that identifies the impact of packet loss on the throughput of a TCP stream. In terms of practice, this document will include input from two IT professionals who have a lot of experience with data replication. One of these professionals is a technical business consultant at a storage company. The other is a manager of network services for a company that does hosting in multiple data centers. These two professionals will be referred to in this document as The Consultant and The Manager respectively.

Key WAN Characteristics: Loss and Out of Order Packets

Historically data networks have been built using some form of a hub-and-spoke design and were based on technologies such as Frame Relay and ATM. Hub-and-spoke designs have been common in large part because that design reflected what had been the natural flow of traffic between a branch office and a headquarters site. However, that situation is changing as new traffic types now need to be supported. Today branch offices typically need access to multiple data centers, either for disaster recovery or for access to applications that are only hosted

¹ Throughout this brief, the phrase *typical data application* will refer to applications that involve inquiries and responses where moderate amounts of information are transferred for brief periods of time. Examples include file transfer, email, web and VoIP traffic. This is in contrast to data replication applications that transfer large amounts of information for a continuous period of time.

in one of the company's multiple data centers. In addition, the vast majority of companies have deployed VoIP and other peer-to-peer traffic that does not tend to follow a hub-and-spoke pattern.

Because it is easier to setup mesh environments using MPLS and IP VPN technologies, they are growing in popularity over traditional frame and ATM services. At the same time, service providers are offering MPLS and IP VPN services at relatively low price points. This is causing a significant increase in MPLS and IP VPN deployments across most enterprises.

While there are significant advantages to MPLS and IP VPN technologies, there are drawbacks, one of which being high levels of packet loss and out of order packets. This is due to routers being oversubscribed in a shared network, resulting in dropped or delayed packet delivery. This issue often goes undetected because the typical service level agreement for MPLS allows the service provider to lose a few tenths of a percent of the packets and still meet their commitment. In addition, packet loss is typically calculated as the arithmetic mean of loss measurements taken over a month and also taken over a large number of circuits. As such, the actual packet loss could be periodically quite high for multiple hours of the day on several circuits even though the service provider has still met their contractual requirements.

The Manager stated that the packet loss on a good MPLS network typically ranges from 0.05% to 0.1%, but that it can reach 5% on some MPLS networks. He added that he sees packet loss of 0.5% on the typical IPsec VPN. The Consultant said that packet loss on an MPLS network is usually low, but that in 10% of the situations that he has been involved in, packet loss reached upwards of 1% on a continuous basis.

Both The Manger and The Consultant agreed that out of order packets are a major issue for data replication, particularly on MPLS networks. In particular, if too many packets (i.e. typically more than 3) are received out of order, TCP or other higher level protocols will cause a re-transmission of packets. The Consultant stated that since out of order packets cause re-transmissions, it has the same affect on goodput² as does packet loss. He added that he often sees high levels of out of order packets in part because some service providers have implemented queuing algorithms that give priority to small packets and hence cause packets to be received out of order.

The Impact of Loss and Out of Order Packets

The affect of packet loss on TCP has been widely analyzed³. Mathis, et.al., provide a simple formula that provides insight into the maximum TCP throughput on a single session when there is packet loss. That formula is:

$$\text{Throughput} \leq (\text{MSS}/\text{RTT}) * (1 / \sqrt{p})$$

where:

MSS: maximum segment size

RTT: round trip time

p: packet loss rate.

The preceding equation shows that throughput decreases as either RTT or p increases. To exemplify the impact of packet loss, assume that MSS is 1,420 bytes, RTT is 100 ms., and p is 0.01%. Based on the formula, the maximum throughput is 1,420 Kbytes/second. If however, the loss were to increase to 0.1%, the maximum throughput drops to 449 Kbytes/second. Figure 1 depicts the impact that packet loss has on the throughput of a single TCP stream with a maximum segment size of 1,420 bytes and varying values of RTT.

² Goodput refers to the amount of data that is successfully transmitted. For example, if a thousand bit packet is transmitted ten times in a second before it is successfully received, the throughput is 10,000 bits/second and the goodput is 1,000 bits/second.

³ The macroscopic behavior of the TCP congestion avoidance algorithm by Mathis, Semke, Mahdavi & Ott in Computer Communication Review, 27(3), July 1997

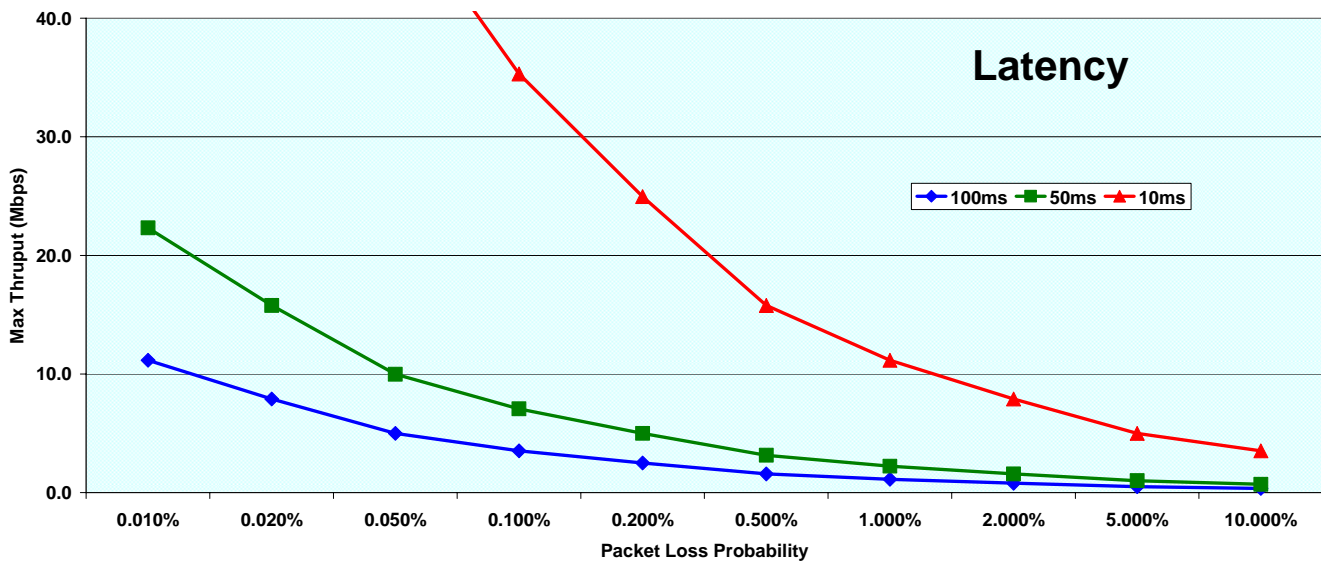


Figure 1: Impact of Packet Loss on Throughput

One conclusion that can be drawn from Figure 1 is that with a 1% packet loss and a round trip time of 50 ms or greater, the maximum throughput is roughly 3 megabits per second no matter how large the WAN link is. The Consultant stated that he thought Figure 1 overstated the TCP throughput and that in his experience if you have a WAN link with an RTT of 100 ms and packet loss of 1% “the throughput would be nil”.

Techniques for Coping with Loss and Out of Order Packets

The data in Figure 1 shows that while packet loss affects throughput for any TCP stream, it particularly affects throughput for high-speed streams, such as those associated with multi-media and data replication. As a result, numerous techniques, such as Forward Error Correction (FEC)⁴, have been developed to mitigate the impact of packet loss.

FEC has long been used at the physical level to ensure error free transmission with a minimum of re-transmissions. Recently many enterprises have begun to use FEC at the network layer to improve the performance of applications such as data replication. The basic premise of FEC is that an additional error recovery packet is transmitted for every ‘n’ packets that are sent. The additional packet enables the network equipment at the receiving end to reconstitute one of the ‘n’ lost packets and hence negates the actual packet loss. The ability of the equipment at the receiving end to reconstitute the lost packets depends on how many packets were lost and how many extra packets were transmitted. In the case in which one extra packet is carried for every ten normal packets (1:10 FEC), a 1% packet loss can be reduced to less than 0.09%. If one extra packet is carried for every five normal packets (1:5 FEC), a 1% packet loss can be reduced to less than 0.04%. To exemplify the impact of FEC, assume that the MSS is 1,420, RTT is 100 ms, and the packet loss is 0.1%. Transmitting a 10 Mbyte file without FEC would take a minimum of 22.3 seconds. Using a 1:10 FEC algorithm would reduce this to 2.1 seconds and a 1:5 FEC algorithm would reduce this to 1.4 seconds.

The example demonstrates the value of FEC in a TCP environment, although the technique applies equally well to an any application regardless of transport protocol. FEC, however, introduces overhead which itself can reduce throughput. What is needed is a FEC algorithm that adapts to packet loss. For example, if a WAN link is not experiencing packet loss, no extra packets should be transmitted. When loss is detected, the algorithm

⁴ RFC 2354, Options for Repair of Streaming Media, <http://www.rfc-archive.org/getrfc.php?rfc=2354>

should begin to carry extra packets and should increase the amount of extra packets as the amount of loss increases.

Packet Order Correction (POC) is a technique that helps to overcome out of order packets. It works by tagging packets as they are sent across the WAN so that they can be re-sequenced on the far end of a WAN link to avoid the re-transmissions that occur when packets arrive out of order. POC is performed in real-time and across all IP flows, regardless of transport protocol, making it an effective WAN optimization tool.

Packet re-ordering and FEC can both be performed in either the router or in a separate appliance. If the network is supporting just typical data applications, either approach will work. However, as pointed out by The Consultant, data replication applications are more demanding. He said that routers are designed to support typical data applications and that part of their design is to have buffers that fill up when a lot of data is being sent and then drain as either less data is transmitted or the TCP session ends. In a data replication application the data flow is constant and the session never terminates. As such, there is no chance for the router's buffers to drain and hence packet re-ordering and FEC are best implemented in a separate appliance. The Manager agreed with The Consultant and said that given the nature of data replication, packet re-ordering and FEC are best performed in a separate appliance.

Summary

Most IP networks are designed to support tens or even hundreds of applications. These applications have varying characteristics. Some applications, such as VoIP and email, require a modest amount of bandwidth for only a few minutes. Inquiry/response applications also require a modest amount of bandwidth for relatively brief periods of time. Even most file transfer applications only transfer data for a bounded amount of time.

Data replication and backup applications are quite different. They require moving massive amounts of information on high-capacity WAN links. In addition, at start-up the typical data replication application spikes to line speed and keeps transmitting at that rate indefinitely. As such, Wide Area Networks must be designed to support these unique requirements.

Unfortunately, as shown in this document, packet loss and out of order packets severely limits the maximum throughput that a single TCP session can support independent of the actual capacity of the WAN link. As pointed out by The Consultant, if you have a WAN link with an RTT of 100 ms and packet loss of 1% "the throughput would be virtually nil". According to The Manager, "you simply cannot get more than a few Mbps of throughput on a WAN with loss, regardless of how big the WAN link is."

Techniques such as packet re-ordering and adaptive FEC can significantly improve goodput across the WAN. However, given the demanding nature of data replication applications, it is usually not possible to support these techniques in a router. Instead, IT organizations that need to support data replication and backup applications should consider implementing a separate WAN optimization appliance that can effectively support these techniques. In many instances, this is just as important as addressing the bandwidth and latency challenges that are also inherent to enterprise WANs.

A Word from the Sponsor – Silver Peak

Silver Peak improves backup, replication and recovery between data centers and facilitates branch office server and storage centralization by improving application performance across the WAN. This is achieved using a variety of WAN optimization techniques, including disk based data reduction, compression, latency/loss mitigation and Quality of Service (QoS).

Silver Peak's award winning NX appliances bring unprecedented security and scalability to WAN acceleration, making them uniquely suited for the rigorous demands of high capacity WAN environments. Common IT initiatives facilitated by the Silver Peak solution include asynchronous data replication, network backup, disaster recovery, SQL transactions, file transfers, email, VoIP and video streaming.

Silver Peak is headquartered in Santa Clara, California. For more information, visit <http://www.silver-peak.com>.

About Kubernan™

Kubernan™, a joint venture of industry veterans Steven Taylor and Jim Metzler, is devoted to performing in-depth analysis and research in focused areas such as Metro Ethernet and MPLS, as well as in areas that cross the traditional functional boundaries of IT, such as Unified Communications and Application Delivery. Kubernan's focus is on providing actionable insight through custom research with a forward looking viewpoint. Through reports that examine industry dynamics from both a demand and a supply perspective, the firm educates the marketplace both on emerging trends and the role that IT products, services and processes play in responding to those trends.

Kubernan is the Greek root word for *helmsman* as well as the phrases to guide and to steer. As such, the name Kubernan reflects our mission of guiding the innovative development and usage of IT products and services.

Kubernan Briefs

Vol.1, Number 4

Published by Kubernan

www.Kubernan.com

Cofounders:

Jim Metzler

jim@kubernan.com

Steven Taylor

taylor@kubernan.com

Professional Opinions Disclaimer

All information presented and opinions expressed in this publication represent the current opinions of the author(s) based on professional judgment and best available information at the time of the presentation. Consequently, the information is subject to change, and no liability for advice presented is assumed. Ultimate responsibility for choice of appropriate solutions remains with the reader.

Copyright © 2007, Kubernan

For editorial and sponsorship information, contact Jim Metzler or Steven Taylor. Kubernan is an analyst and consulting joint venture of Steven Taylor and Jim Metzler.