A Riverbed Technology White Paper

HIGH-SPEED TCP FOR LONG FAT NETWORKS (LFNs)

Riverbed Technology, Inc.
501 Second Street, Suite 410 • San Francisco, CA 94107 • USA
Tel: +1 415 247 8800 • Fax: +1 415 247 8801 • www.riverbed.com

RIVERBED™

HIGH-SPEED TCP FOR LONG FAT NETWORKS (LFNS)

Introduction

The Transmission Control Protocol (TCP) is the cornerstone of intercommunications over Internet Protocol (IP) networks. TCP allows client-server applications to communicate reliably over the unreliable IP packet service and share network bandwidth across connections in a roughly "fair" fashion. While TCP's algorithms for congestion control have proven to be remarkably scalable across the wide range of conditions and link bandwidths in most networks, additional mechanisms are necessary to enhance and optimize TCP performance for use on the very high-speed WAN links that have recently become available to today's enterprises.

Riverbed has implemented Internet Engineering Task Force (IETF)-specified congestion control mechanisms in the Steelhead appliance that enable TCP performance to scale to hundreds of Mbps over significant latencies (>100ms RT). Riverbed customers with high-speed WAN links can now achieve full utilization of their investment in network bandwidth without losing or compromising any of the familiar and essential characteristics and benefits of TCP. This includes "safe" congestion control, even when high-speed TCP connections share WAN links with "normal" TCP connections.

TCP Limitations for High-Bandwidth Links

TCP implements reliable delivery between a sender and receiver by detecting when packets are lost in the network, and retransmitting the lost packets. In the early days of the Internet, this simple scheme turned out to have a big problem: since packets are generally lost when the network is overloaded, increasing the load by retransmitting lost packets lead to even more overload, ultimately leading to "congestion collapse". To fix this problem, control congestion for TCP was introduced in the mid-1980's by Van Jacobson, who at the time was a scientist at the Lawrence Berkeley National Laboratory. While ingeniously elegant and widely adopted by the Internet community, these congestion control algorithms were designed and validated at a time when wide-area networking speeds were much lower.

The same T1 and E1 Wide Area Network (WAN) links that were prevalent in the 1980's are still in wide use and TCP performs as suitably over these networks today as it did back then. But today, due to advances in long-distance fiber-optic technologies, network providers can also offer customers high-speed WAN links that are significantly faster. These "Long Fat Networks" (LFNs) offer WAN performance speeds in excess of 100Mbps, and they can span global distances resulting in significant round-trip latencies measured in the hundreds of milliseconds of delay.

Today, Jacobson's original congestion control algorithms from the 1980's are universally implemented in most workstations and servers, including Microsoft Windows and Unix. Although necessary to address network congestion issues, these algorithms have the unfortunate side-effect of limiting overall performance on high-speed WAN links. Many enterprises have attempted to run applications using TCP over LFNs, only to experience dismal and disappointing performance despite the abundance of WAN bandwidth.

The basic problem limiting TCP performance arises from how Jacobson's original "congestion avoidance" algorithm interacts with networks having large "bandwidth-delay products". Congestion avoidance increases the sender's TCP window (equivalently the sending computer's transmission rate) by only a single packet for each successful round-trip acknowledgement. When the TCP window is small, increasing it by a single packet is a reasonable thing to do. But when a window is very large (say hundreds of packets), then each additional round-trip acknowledgement adds just a miniscule increase to the sender's TCP window. In such a situation it takes an extraordinarily large number of round-trips to rebuild the TCP window in response to a single packet loss, and this translates into very sluggish TCP behavior.

For example, to sustain a throughput of 1 Gbps over a WAN with 100ms of round-trip latency, the sender's TCP window must be approximately 8000 packets in size. When a single packet is lost with a TCP window this large, the window will be cut in half to about 4000 packets. It then takes 4000 successful round-trips between the sender and receiver for the window to grow back to its original 8000 packets. At 100ms per round-trip, this adds up to 400 seconds, or more than 6 minutes of lossless high-speed data transfers in order to recover from a single packet loss.

The Riverbed High-Speed TCP Solution

In response to these TCP performance issues described above for high-speed WANs, Riverbed has implemented IETF-specified mechanisms (RFC 3649 and RFC 3742) for improving TCP performance in high-speed networks. These mechanisms deliver all of the familiar characteristics and benefits of TCP, but scale performance to hundreds of megabits per second for individual TCP

connections. The resulting "High-speed TCP" capability allows customers to maximize utilization of the WAN bandwidth received from their network provider.

Specifically, RFC 3649 addresses issues in the "congestion avoidance" algorithm by specifying an approach where each individual high-speed TCP connection manages its congestion window as if it were an aggregate of multiple TCP connections. This allows a TCP window to expand more appropriately in a high-speed WAN environment, as well as to reduce its TCP window less disruptively when congestion is encountered. IETF-specified algorithms describe how to scale the connection multiple as the TCP throughput speeds increase beyond 100Mbps. The end result of implementing RFC 3649 is preservation of TCP's safe and proven congestion control characteristics that facilitate bandwidth sharing among any number of applications utilizing the WAN, even with applications using "normal" TCP connections without high-speed capabilities.

A second IETF-specified adjustment to the TCP algorithms is described in RFC 3742, which addresses the management and expansion of each connection's TCP window when in the "slow start" phase. Rather than the old method involving exponential expansion of the TCP window from a small base, RFC 3742 provides a direct approach that expands the TCP window more expeditiously in the "slow start" phase. This approach accounts for the round-trip latency in the WAN and prevents TCP connections from stalling at low throughputs relative to overall available bandwidth for extended periods of time.

Riverbed's implementation of these IETF-specified approaches allows the Steelhead appliance to scale TCP performance into the hundreds of megabits per second over the WAN. Familiar TCP performance characteristics have been preserved. For example, there is no need to pre-determine available WAN bandwidth—the Steelhead uses TCP to self-adjust transmission throughput appropriately. Congestion control algorithms provide safe sharing of available WAN bandwidth among any number of applications, including those not using high-speed TCP. Overall, the Riverbed high-speed TCP behavior is remarkably similar to "normal" TCP, although performance levels are far greater.

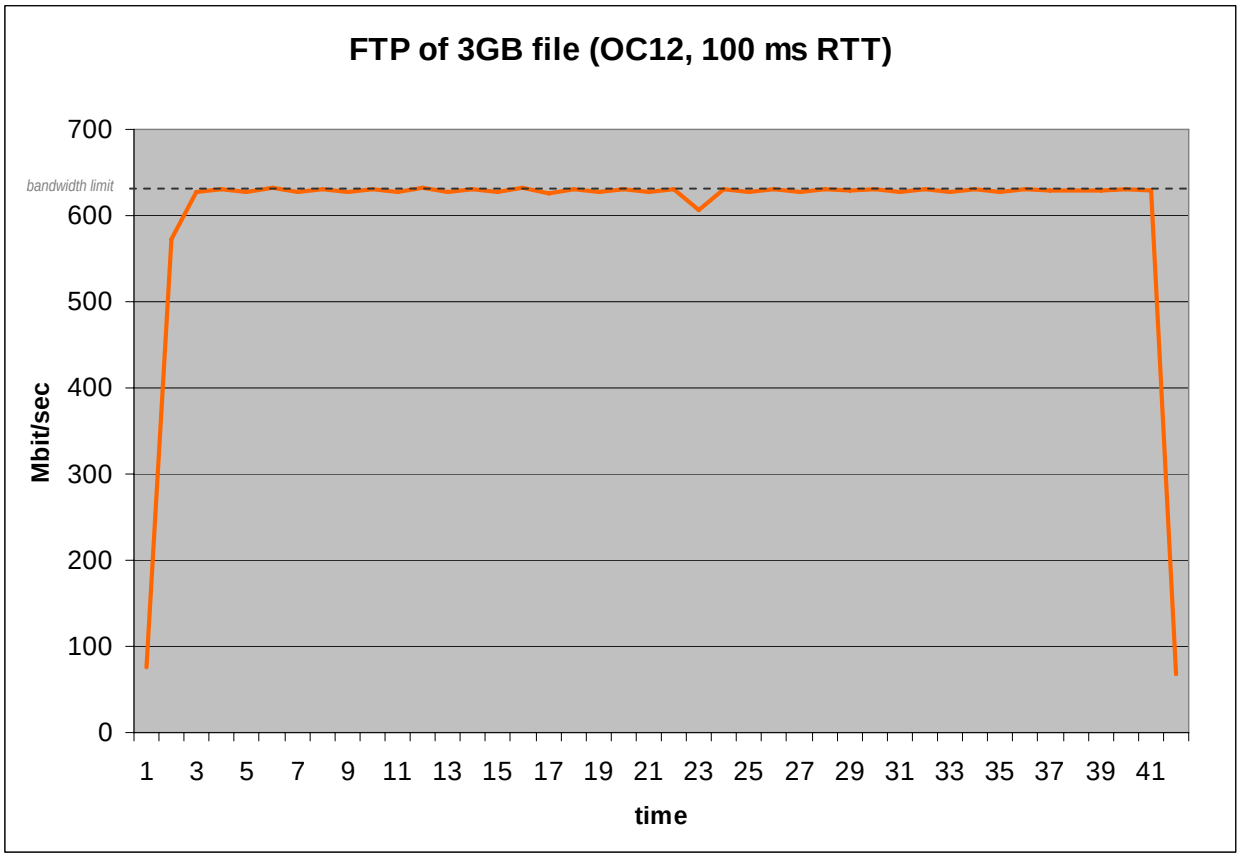

Figure 1 – A plot of high-speed TCP data transfer of a 3GB file across an OC-12 WAN (622Mbps) with 100ms RT latency. The Steelhead appliance is able to achieve full utilization of the link.

Continuous Improvement of TCP

Note that Riverbed's approach to adopt incremental improvements is consistent with TCP's historical evolution. TCP has sustained continuous improvements over the years by researchers adding clever enhancements without any change to the original TCP protocol architecture. For example, in 1990 the slow-start and congestion avoidance algorithms were improved with the "fast recovery" algorithm that allowed TCP to recover from multiple packet losses in a single round-trip. In 1993, "large windows" support was added so that the 16-bit window field would not fundamentally limit performance. Then in 1995, the "rate-halving" algorithm was introduced to reduce the burstiness of TCP. Most recently in 1998, "selective acknowledgements" (SACK) was added to improve the efficiency of retransmission from packet loss. All of these enhancements and improvements are now codified in IETF standards and are part of virtually every widely-deployed TCP stack, including those in Microsoft Windows and Linux. Riverbed's implementation of High-speed TCP is merely the latest in a long line of incremental improvements to TCP.

Caveats of Using Non-TCP Solutions

Since the early pioneering work of Van Jacobson in the 1980s, TCP congestion control has been the focal point of innumerable research efforts, both in academia as well as in industry. A great many attempts have been made toward replacing TCP with a radically different protocol and congestion control scheme, yet none of these attempts have proven more effective than the original TCP congestion control model that has been universally adopted by the Internet community.

Myths in the marketplace abound concerning TCP's issues in operating in high-speed, long-distance WAN links. Some vendors ignorantly proclaim that TCP/IP was never designed to transport large amounts of data over distance. They further believe that "new" technology, typically their own "special" proprietary approach, delivers the solution that uniquely addresses the issues involved with transferring large amounts of data over vast distances. We find this to be an awkward position because TCP was designed precisely for this purpose.

Note that other similar "new" approaches have been attempted many times in the past, and discarded by the Internet community as inferior to the tried-and-true TCP approach. Most approaches involve a sacrifice of any congestion control safeguards, and can result in "stealing" of bandwidth from those applications not fortunate to be supported with the proprietary technology. They are also vulnerable to "congestion collapse", a common occurrence for Internetworks in the early-to-mid 1980's prior to the advent of Van Jacobson's congestion control algorithms.

Summary

TCP has been the essential enabling technology for the Internet for the past 30 years. Over that period of time, the Transmission Control Protocol has evolved and improved to meet the needs of Internet users. Riverbed embraces TCP's rich heritage, and has implemented IETF-specified enhancements to TCP to facilitate its use over high-speed WAN links. Unlike non-TCP approaches that use proprietary approaches to boost performance over high-speed WANs, Riverbed's capabilities allow customers to experience the proven and familiar benefits that TCP has delivered since its inception.

• • •

For more information, visit www.riverbed.com or call +1-415-247-8800. In the US, call toll-free: 1-87-RIVERBED (1-877-483-7233)