Improving Network Cost and Performance with Advanced Compression Techniques

By Peter Sevcik and John Bartlett
February 2003

Half of the typical enterprise telecommunications budget is spent in carrier services, according to the Gartner Group. Furthermore, data now represents 60 percent of that budget. The demand for data traffic keeps growing [1] faster than general price reductions in the marketplace. In particular, companies with international offices feel the overall effect of increasing telecommunications costs.

Enterprises must try to get as much value as possible out of their networks. The current strategy employed by most enterprises is to buy enough capacity to support an office and then delay upgrading that capacity until employees complain loudly that performance is poor. At this point, the enterprise typically doubles the capacity to improve conditions. This scenario plays out with many variations but generally forces users to live under oscillating conditions of satisfactory and unsatisfactory performance.

There are many approaches and techniques that propose to improve upon this default strategy of capacity planning. This report focuses on the new advances in compression technology that provide a dramatic improvement to the cost of supporting remote offices while improving the performance of mission-critical applications.

Office Applications Keep Evolving

Each enterprise is different but the applications they operate can often be grouped into very similar categories. This analysis is based upon the following generally representative mix of applications and their associated usage trends.

Mission-Critical

Most business processes operate using one of the many vendor-unique application platforms that are based upon a client-server model. This is the enterprise mission - critical application that requires the best performance in order to support the business process. In this study, we have modeled a generic client-server application based upon several popular application platforms.

There is a significant migration of these business process applications from the vendor-specific platform to a more open platform using Web-based techniques. Even the application vendors are migrating their offerings from the client-server model to the open Web services model. These applications often employ XML, SOAP, and other emerging standards. The effect is that the traffic load they produce changes over time [1].

This application evolution to a Web-model is causing the payload for the average task to grow from 20,000 Bytes to 200,000 Bytes based upon 2002 measurements. Furthermore, the payload and activity level of a typical user is also growing over time.

Electronic Mail

Every enterprise uses electronic mail. In fact, this is often the fastest growing source of traffic as users learn to rely upon email for document exchange. Although an important source of traffic, users typically do not demand fast response time from the application because most of the work is performed in the background. A typical email session produces 180,000 Bytes of traffic but this load is doubling every 12 to 24 months.

NetForecast Report 5065

©2003 NetForecast, Inc.

Web Access
Almost all enterprises give their users access to the public Web for both business and personal use. However, most users are forced to access the Web through their corporate network and then out through a single entry point into the Internet. This entry point has a firewall and may have several additional layers of security and management. Although least important in corporate priority, it is often a significant traffic load that has a different and evolving application profile (Payload and Turn count) as described in [2].

Advanced Compression Technology

Clearly, compressing the content of enterprise applications can alleviate the problem of unbridled traffic growth. There are two general approaches to compression: packet layer and application layer. Packet layer compression is transparent to any type of application and therefore provides some value to any application. However, by their very nature, packet layer approaches do not know anything about the content or context of the transactions.

Application-level compression available as software or hardware near the server has two major advantages. First, it can compress larger sections of payload. Second, these techniques can operate the TCP protocol more efficiently and reduce the number of TCP turns. More advanced implementations, modeled here, also add caching to further reduce the need to transfer payloads. The overall effect on the traffic load generated by the above application mix is dramatic, as shown in Figure 1.

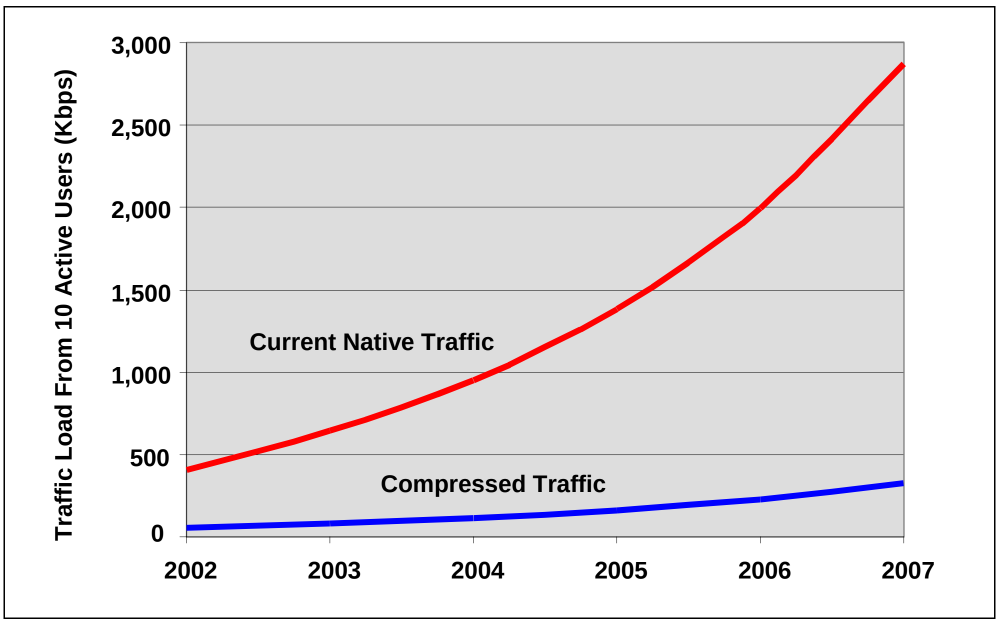

Figure 1 – Traffic Load Offered by 10 Active Users

Enterprise Office Planning Scenarios

The data in Figure 1 needs to be placed in the context of real enterprise scenarios. How much bandwidth savings are realistic over time? What happens to the application response time as seen by the remote user? In order to answer these questions, we have defined two typical office scenarios. In both cases we assume that the enterprise is operating a private network where bandwidth is explicitly purchased from a data center to field offices via dedicated circuits or virtual circuits (frame relay or ATM). These circuits are then upgraded in capacity over the planning period as shown in Figure 2.

Dallas Regional Office
This is a company with headquarters in New York and a regional office in Dallas, Texas. The office has a daytime staff of 1,500 employees of which 300 are active in the afternoon. In 2002, a 45 Mbps DS3 circuit supports the site. The users experience an average task response time of 5 seconds on the corporate mission-critical application.

Tokyo Field Office
The same company also operates a remote field office in Tokyo, Japan with 50 employees, of which 10 are active in their busy period. These employees operate the same application mix and are connected to the New York data center by a 1.5 Mbps DS1 circuit. The same mission-critical application presents an average response time of 12 seconds. The users have become accustomed to tolerating this slower performance.

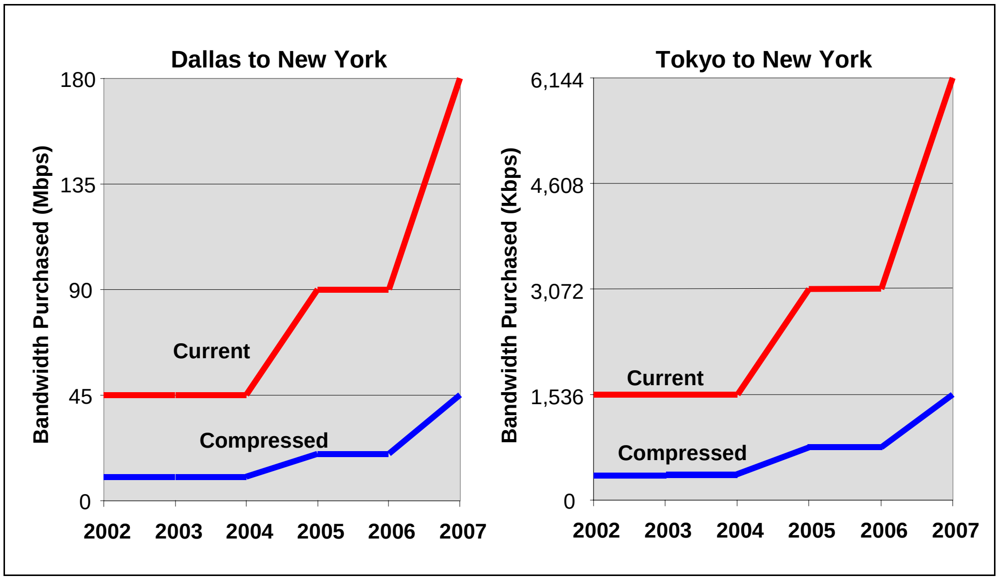

Figure 2 – Bandwidth Purchased Over Time

Comparing the Strategies
Both offices are expected to keep their staffing level over the planning period. Despite the fixed staff number, the applications grow in traffic demand. Also, the same active users become more facile with the applications, using them more heavily when active. The result is the need to add bandwidth every two years, as shown in Figure 2.

The alternative approach of using advanced compression permits both offices to buy only one-quarter of the bandwidth in each of the years. This is a significant savings in expenditure of recurring monthly costs regardless of how the bandwidth is purchased or the carrier being used. These cost savings will quickly pay for the cost of compression appliances, creating an attractive return on investment.

The difference between the current approach and the use of advanced compression is even more striking in the response times of the mission-critical application over time. Figure 3 clearly shows the effect of the typical scenario of letting performance suffer and then improving the performance with a capacity upgrade. The seesaw effect is less dramatic in Tokyo because of the damping effect of higher latency.

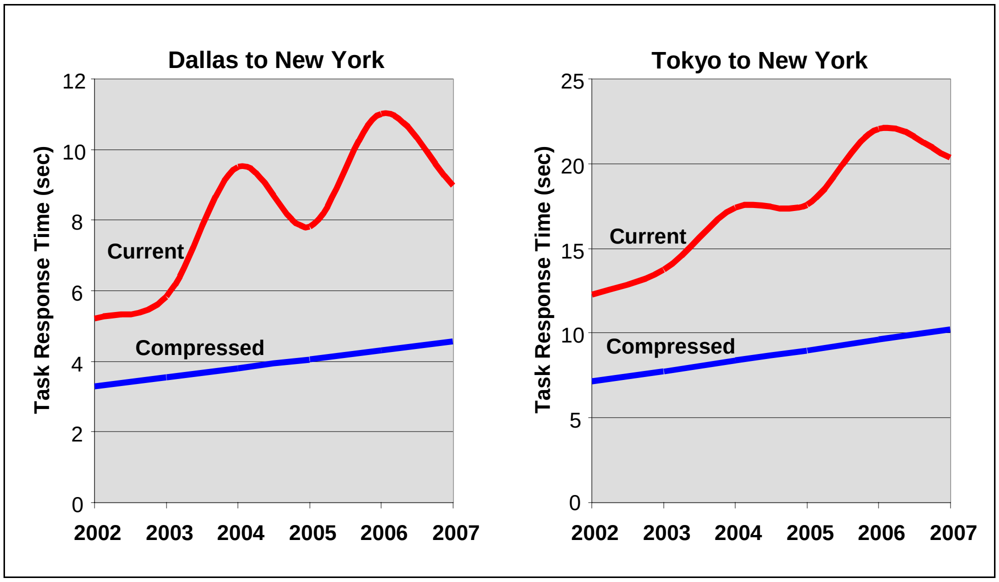

Figure 3 – Mission-Critical Application Task Response Time

Figure 3 shows that compression removes the oscillations in performance and always provides much faster response times. Although the compressed response time also climbs as the application profile changes over time, it is generally twice as fast as the current scenarios. These faster response times are achievable despite purchasing only one-quarter of the bandwidth relative to the typical default plan.

Summary

The benefits of lowering cost while supplying better performance are compelling. Furthermore, the value of compression is increasing as applications evolve to more data-rich and therefore more compressible content. Similar results were found in an Internet e-commerce application [3]. These results show what is achievable given a good product design. Enterprises should investigate the specific approach of each compression vendor under consideration. Since strategic use of compression technology is likely to benefit many enterprise networks, it should be a technology that is seriously considered.

References

1. Bartlett and Sevcik, "Impact of Web-Based Applications on Enterprise Traffic," NetForecast Report 5063, October 2002.
2. Sevcik and Bartlett, "Understanding Web Performance," NetForecast Report 5055, October 2001.
3. Sevcik, "Accelerating e-Commerce," *Business Communications Review*, November 2002.

Peter Sevcik is President of NetForecast and a leading authority on network technology. He specializes in performance analysis and holds the patent on application response time prediction. He can be reached at peter@netforecast.com.

John Bartlett is Vice President of NetForecast, and has 24 years of experience designing at the chip, board, system, and network levels, with a focus on performance. He can be reached at john@netforecast.com. ¤

NetForecast develops customized analytic models to determine the business value of new technologies.

www.netforecast.com

NetForecast and the curve-on-grid logo are registered trademarks of NetForecast, Inc.