

## DATA COMMUNICATIONS MANAGEMENT

# THE MYTH OF THE “WELL-MANAGED” NETWORK

Andy Helland

## INSIDE

Routed IP: Built for Scalability — but at What Price?; Why Do Routed IP Networks Drop Data?; Carrier-Based IP Networks; Dedicated Enterprise IP Networks; Reliable Transport over SONET (RToS); Flow Control Provides Loss-Free Congestion Control; Cost Basis for Service; Moving Data from New York to Boston

## ROUTED IP: BUILT FOR SCALABILITY — BUT AT WHAT PRICE?

Given the tremendous success that IP routing has enjoyed in the development of the Internet, it is natural to assume that it is an ideal medium for interconnecting high-performance data centers. Nothing could be further from the truth. IP routing protocols, including the random early discard (RED) algorithms of routers, were optimized to service the needs of millions of users who each wished to move a relatively small amount of data. IP routing was not optimized to provide high throughput for a small number of users. This becomes clear when a detailed performance analysis of TCP/IP routing is conducted.

One of the most often-cited research reports on the performance of TCP/IP is “The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm.”<sup>2</sup> In this seminal report, the behavior of the TCP in the face of latency (the time delay between the sender and the receiver) and packet loss rate is carefully analyzed. The result of this analysis is a relatively straightforward equation that explains the relationship between bandwidth,

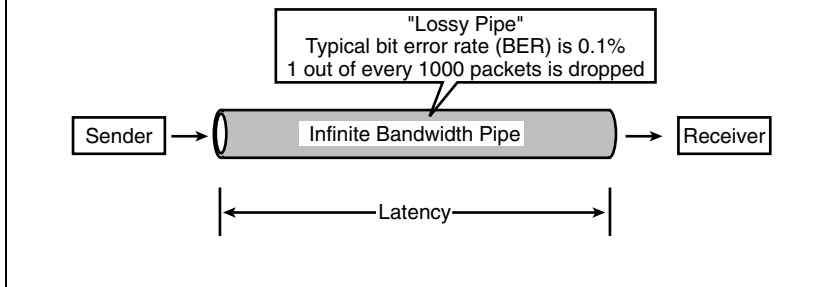
### PAYOFF IDEA

There is an ever-increasing need to provide high-performance interconnect between geographically distributed data centers. Much of this is driven by the need to maintain synchronous or nearly synchronous file systems at multiple sites. This is done to provide protection against catastrophic loss, centralize backup, or just cache data close to its intended users. This article discusses *the myth of the well-managed network*. It explains why the public carriers will never provide it and why it is not cost effective for the private enterprise to build it when compared with other options. It also discusses an emerging class of WAN connectivity (Reliable Transport over SONET — RToS) that provides point-to-point *reliable* transportation of IP or Fibre Channel traffic. For users who need to move large volumes of data between a limited number of sites, this is not only the highest performance interconnect methodology, but also the most cost effective.<sup>1</sup>

---

**EXHIBIT 1** — Interconnect Pipe with Infinite Bandwidth

---



packet size (referred to as maximum segment size), round-trip time, and packet loss rate.

$$BW \leq \frac{1.31 \infty MSS}{RTT \sqrt{PkLoss}}$$

- BW = bandwidth (bps)
- MSS = maximum segment size (bits)
- RTT = round-trip time (seconds)
- PkLoss = packet loss rate

As one models the effects of real-world packet loss rate, one reaches the inescapable conclusion that throughput (the total amount of data that is moved from point A to Point B in the WAN) is much more dependent on latency and packet loss rate than on the native bandwidth of the pipe. To demonstrate this point, take a look at an interconnect pipe with *infinite* bandwidth that connects a sender and a receiver as shown in [Exhibit 1](#).

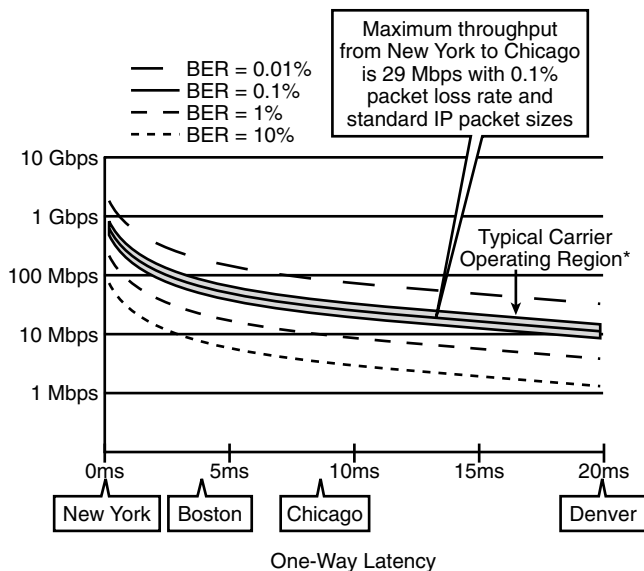
To understand the impact of packet loss on TCP/IP throughput, one can model a dedicated pipe with infinite bandwidth, non-zero latency, and a typical packet loss rate.

Even if one has a pipe of infinite bandwidth, the fact that one must stop and retransmit data every time a packet is dropped quickly becomes the dominant factor that limits throughput. Furthermore, while retransmitting the lost data, all of the successfully transmitted data that came after the lost data must wait until the lost data has been retransmitted to prevent misordering.

Based on this analysis of TCP performance, one can plot a family of curves that describes the behavior of TCP as a function of latency and packet loss rate. This graph is shown in [Exhibit 2](#). Each curve shows the expected actual throughput when operating over an infinitely fast connection with a specific error rate. The maximum length of an IP packet has been assumed to be 1500 bytes. This is the standard MTU (maximum transfer unit) for Ethernet. Note that as the error

---

**EXHIBIT 2** — Behavior of TCP as a Function of Latency and Packet Loss Rate



\*Network statistics obtained from [www.matrix.net](http://www.matrix.net).

rate increases, the throughput decreases. Increasing the distance between the sender and the receiver also decreases the throughput.

In a typical lossy IP network, throughput is limited more by latency and loss rate than by the bandwidth of the pipe.

The limiting factor in throughput is based on the need to retransmit dropped data and the amount of distance over which the data must be sent. To calibrate latency in terms of distance, it has assumed that the sender is in New York and the receiver is located in Boston, Chicago, or Denver. Packet loss rate is a critical parameter in this analysis. The average packet loss rate for the Internet at large<sup>3</sup> is between 1 percent and 2 percent. Service level agreements (SLAs) with major carriers typically guarantee only 0.5 percent packet loss. For the purpose of this analysis, we will assume a packet loss rate of 0.1 percent (five times *better* than premium SLA). With a packet loss rate of 0.1 percent, data that is sent between New York and Chicago can only achieve a maximum throughput of approximately 29 Mbps (even if the actual pipe that interconnects the two sites has infinite bandwidth).

This performance limitation of TCP/IP over lossy networks is nothing new to the supercomputing industry. For years, they have been trying to move large amounts of data across routed IP networks with little success. Los Alamos National Labs, for example, has resorted to sending data via tapes and Federal

---

Express to their partner facility at Sandia National Labs. The real-world throughput limitations of TCP/IP over lossy networks are so extreme that it is faster for them to send data via truck than to use their OC-3 IP wide area access.<sup>4</sup> Clearly, there are problems with transporting large amounts of data using a lossy network. But why is packet loss such an inherent part of a routed IP network? Take a look at the options for building a routed IP network to better understand why packet loss is such an inevitable part of a routed network.

### **WHY DO ROUTED IP NETWORKS DROP DATA?**

The enterprise seeking to interconnect distant data centers has two choices in obtaining IP WAN connectivity. First, it can contract with a major carrier (such as AT&T, Sprint, MCI, or Qwest) to provide routed IP connectivity. The second choice is to build and maintain a private network — ensuring that packet loss is kept to manageable levels that are always under the control of the enterprise. This article examines each of these options and explores why one inherently limits performance and the other is wasteful of bandwidth (and money). We start with the carrier-provided model.

### **CARRIER-BASED IP NETWORKS**

As discussed above, public carriers typically operate their networks at a fairly high packet loss rate (between 0.1 percent and 1 percent). Why is this so? Why would the carriers operate their networks with such a high loss rate and thwart the desires of large-scale users? The answer is simple once one examines the economics of the situation. Carriers are paid a flat rate by their customers for access to the routed IP network. They want the SONET pipes that power their networks to be as full as possible. To make sure that this happens, they oversell the bandwidth. The difficulty lies in the fact that there is great variation in the amount of traffic that passes through an IP network. Carriers sell IP service based on peak bandwidth. However, they manage the load of their network based on the average load that they see.

Major airlines operate the same way. They regularly oversell the number of seats on a given flight. Statistically, they know that some passengers will cancel at the last moment (without providing notice). In the event that all of the passengers that actually purchased tickets show up, they are forced to “bump” a passenger. Just like in the case of IP carriers, most customers never know that this took place. Only a few are inconvenienced.

This over-subscription business model works well if there are many small users whose requirements cannot be predicted in advance. However, if a single user attempts to move large amounts of data (passengers), this model begins to fail rapidly. Now return to the airplane analogy. Suppose there is an entire football team that must be moved from one city to another. What if the airline decides to “bump” one person (say, the quarterback)? The rest of the team will arrive at the destination at the correct time. However, because the team cannot function until the quarterback is safely delivered, the entire team is forced to delay its game until

---

---

he arrives. As the size of the team increases, the chances of being disrupted also increase. The same effect occurs with large-scale data transfer.

When single users attempt to transfer large amounts of data across the Internet, their chances of losing data increase as well. If the carriers drop one out of a thousand packets and the users are small, 999 customers will see great performance and one customer will have difficulty. On the other hand, if a single user is moving large amounts of data, every time he sends a thousand packets, he will get hit. TCP windowing will force him to reduce his bandwidth and he will constantly experience frustration (retransmission and bandwidth reduction).

Quite simply, today's Internet is designed to let the carriers manage the aggregate behavior of many small users who need to route data to any destination. It is not well suited to the transfer of large amounts of data between dedicated sites.

### **DEDICATED ENTERPRISE IP NETWORKS**

Now examine the traditional alternative to a carrier-provided WAN infrastructure. Instead of obtaining routed IP service from a major carrier, the private enterprise can build and maintain a dedicated WAN for IP connectivity. For the enterprise to maintain a private WAN, it must first obtain a dedicated link between its data centers or campus sites. This link will inevitably be a dedicated SONET (synchronous optical network) channel. SONET channels are used throughout the United States as the universal method for carrying IP, ATM, and voice. The main reason for the enterprise to build and maintain a dedicated IP network is to minimize the level of packet loss across the WAN. This is the so-called "well-managed network." The problem with the well-managed network is that it must be so over-provisioned that a large percentage of available bandwidth is wasted. Why is this so?

Routed IP networks are designed to allow for many traffic sources to send their data whenever they wish. Because the instantaneous traffic cannot be predicted, it is common for multiple sources to generate substantial amounts of data at the same time. When this occurs, the network's routers are overwhelmed and must drop data. As discussed above, the packet loss and latency of the WAN conspire to seriously degrade throughput. To counter this and minimize packet loss, the network must be significantly over-provisioned. There is a general "rule of thumb" in network design that a routed network must be operated at 50 percent capacity to ensure that packet loss is kept to a manageable level. If the average operating level is 50 percent, then the amount of time that the peak capacity is exceeded (more than 100 percent) is kept relatively small.<sup>5</sup>

The good news is that one can maintain predictable quality of service (QoS; minimal packet loss). The bad news is that one wastes 50 percent of the bandwidth to do so. As the bandwidth of the private WAN increases, the amount of wasted bandwidth also increases. For example, if a private enterprise were to interconnect two sites with an OC-12 SONET link (622 Mbps) running IP services, half of the capacity (approximately 300 Mbps) would be wasted in order to ensure that the packet loss is kept to a manageable level. As will be seen below,

---

---

this represents a significant cost to the enterprise because the monthly recurring cost of SONET must still be paid although only half of the capacity is being used.

Before comparing the economic differences between these options, we introduce a new class of WAN connectivity — Reliable Transport over SONET (RToS). This new form of WAN connectivity still allows for IP services to be interconnected between remote sites but it does so using coupled flow control instead of packet loss to manage congestion. This allows users to maintain very high levels of performance over the WAN.

### **RELIABLE TRANSPORT OVER SONET (RToS)**

Point-to-point connectivity used to be impractical. That is one of the main reasons that routed IP was invented in the first place. Who could imagine providing point-to-point connectivity between every possible site in the world? Routed IP emerged as a wildly successful model for worldwide scalability and connectivity. Indeed, no other method of providing multi-point connectivity is likely to emerge as a replacement for routed IP. However, an important paradigm shift has occurred as the Internet exploded into our lives. As the Internet grew, so did the underlying “plumbing” that carried the IP data — SONET. At last, it is affordable and easy for the enterprise to establish site-to-site connectivity using dedicated SONET circuits. Right now, there are more than 150,000 SONET/SDH rings deployed (SONET is used in the United States while SDH is its close cousin that is deployed in the rest of the world). These rings provide broadband connectivity between every major metropolitan area. The challenge remains to couple IP (Gigabit Ethernet) and Fibre Channel (FC) onto dedicated SONET circuits while preserving the flow control mechanisms that allow edge devices to cope with congestion. This new type of wide area networking (it is not really new, just newly rediscovered) does not drop data as a method of managing congestion. It can achieve levels of performance in the WAN that have never been possible using routed IP. Take a closer look at how this works, starting with a quick overview of SONET.

SONET (synchronous optical network) is a network of fiber-optic cables that carries data and voice throughout the United States. Data is passed through dedicated channels at rates from 155 Mbps up to 10 Gbps. Data paths are not changed rapidly like they are with routed IP networks. Rather, they are provisioned and remain very static. Each dedicated channel has a guaranteed fixed bandwidth for the data. A good way to think about SONET would be to imagine a dedicated pipe with water flowing from one location to another. Every drop of water that enters the pipe on one side will emerge from the other side of the pipe. The water is never lost along the way and the water always emerges in the same order that it entered the pipe. The same is true for SONET channels. All the data that enters the pipe will emerge from the far side of the pipe and it will exit in the same order that it entered the pipe.

---

---

SONET provides the universal infrastructure used by every carrier in the United States to transport data and voice. ATM and IP all ride on top of SONET (including the core of the Internet). Intuitively, it makes sense to use the lowest layer of transport available that can accomplish a particular goal. This eliminates unnecessary overhead and streamlines the transport process. Because it is so universally deployed, SONET-based signals are readily passed from carrier to carrier to move across country.

Another benefit of SONET-based networks is that they are generally deployed in redundant rings. In the event that there is a failure (e.g., due to a backhoe digging up a cable), SONET add-drop multiplexers (ADMs) will automatically reroute the traffic within 50 milliseconds. This level of reliability is one of the key reasons that the telecom and datacom infrastructure of the world has worked so well for so many years. Because of its ready availability and ability to transit from carrier to carrier, SONET is the ideal mechanism to carry high-performance FC and IP data over distance.

### **FLOW CONTROL PROVIDES LOSS-FREE CONGESTION CONTROL**

All networking systems need some method of conveying congestion between the edge devices that are attached to the network. In routed IP networks, congestion in the network is resolved by dropping data. Reliable Transport, on the other hand, makes use of credit-based flow control that does not drop data. Credit-based flow control was originally made popular by the Fibre Channel industry. Take a look at how it works and how it is being adapted for the WAN.

Credit buffering is a critical element in the success of Fibre Channel-based storage systems. Credit buffering accomplishes two things — both related to the concepts of congestion and flow control. First, it prevents the switches in the fabric from becoming congested and dropping data. The receiving switch controls the pace of traffic as it moves from the sender to the receiver by issuing credits back to the sender. If the receiving switch is congested, it will not issue credits to the sender and the sender will slow down the pace of traffic. Because the receiving switch controls the pace, it can never be overwhelmed by traffic. This is a crucial difference between Fibre Channel networks and routed IP networks. At any instant, an IP router can be suddenly overwhelmed with traffic. If this occurs, it will be forced to drop data to protect itself.

The second attribute of Fibre Channel credit buffering is that it allows the actual storage devices to communicate their levels of congestion to each other. Consider the example of a server that is suddenly faced with a large number of interrupts and is temporarily unable to service the needs of an in-process SCSI transfer from a disk. In this case, it will slow down the pace of credits that it issues to the Fibre Channel switch to which it is connected. By reducing the pace of credits that are issued, the server communicates back-pressure to the Fibre Channel fabric. That back-pressure will propagate through the fabric and ultimately

---

---

cause the transmitting disk to slow its pace of traffic. The important theme here is that congestion is transmitted back through the network *without dropping data*.

These attributes of Fibre Channel are extremely important when Fibre Channel is extended outside the data center and over the WAN. Once distance (latency) is added to the system, proper credit buffering becomes an even more crucial aspect of the system design. The WAN side of the extension gateway must use a credit-based system that has been optimized for the distance and bandwidth of the link extension. Otherwise, the system will be unable to sustain the desired data rates over long distances.

For this analysis, we look at transporting IP and FC data over an OC-12 (622 Mbps) SONET link. Once the SONET overhead is removed, the actual data rate will be approximately 580 Mbps. In a well-designed RToS gateway, this will be a readily achievable throughput over the WAN.

As an example, take a look at the economics of moving data from New York to Boston using the three methods discussed thus far:

1. Carrier-based routed IP
2. The “well-managed” network
3. Reliable Transport over SONET

We will estimate the cost of service and the achievable throughput for each option and then calculate the cost efficiency of the link in dollars per megabit per second (\$ per Mbps).

---

**EXHIBIT 3 — Monthly Recurring Costs for Carrier-Based IP Service**

---

| <b>Service</b> | <b>Monthly Recurring Cost (\$)</b> | <b>\$ per Mbps (line rate)</b> | <b>Source</b>   |
|----------------|------------------------------------|--------------------------------|-----------------|
| T1             | 1,200                              | 777                            | Published price |
| T3             | 28,000                             | 626                            | Published price |
| OC-3           | 49,000                             | 316                            | Published price |
| OC-12          | 98,000                             | 158                            | Estimated price |

**COST BASIS FOR SERVICE**

As a basis for estimating the price of carrier-based IP connectivity, we used the Web site [www.broadband-Internet-provider.com/research-information.htm](http://www.broadband-Internet-provider.com/research-information.htm). This is a broadband Internet service provider (ISP) and publishes average pricing information for T1, T3, and OC-3 service. While it does not publish rates for OC-12 Internet service, one can estimate OC-12 pricing from the numbers for T1, T3, and OC-3, as shown in Exhibit 3.

---

---

Because there is a clear trend of decreasing cost as the data rate increases, we will reduce the price per megabit per second by another factor of two in estimating the monthly recurring cost of OC-12 service.  $\$316/2 = \$158$  per Mbps per month for OC-12 Internet service (estimate). Multiplying this number by the 622 Mbps line rate yields an estimated monthly recurring cost of \$98,000 for OC-12 service (four times faster than OC-3 and twice the price).

Note that carrier-based IP access is not distance sensitive in its pricing. These estimates are based on providing IP access to the customer premises. However, to interconnect two different sites would require that each site obtain carrier-based IP services.

Unlike the pricing of IP service, the monthly recurring cost of SONET service *is* distance sensitive. Because the customer is contracting for a dedicated point-to-point link at a committed rate, the monthly recurring charge will be a function of the distance between the sites as well as the bandwidth. SONET price models are also complicated by differing rate structures that depend on the distance from the enterprise to the carrier's POP (point-of-presence) as well as the distance between the POPs. SONET links that move across telephone LATAs (local access and transport areas) may also be subject to additional charges. We used a pricing model that is representative of that used by the major carriers and assumed that the enterprise was located ten miles from a carrier's POP in each of the target cities. For OC-12 SONET service between New York and Boston, we estimated a monthly recurring cost of \$39,000.<sup>6</sup>

### **MOVING DATA FROM NEW YORK TO BOSTON**

The distance between New York and Boston is 190 miles (306 kilometers). We can assume a 50 percent overhead (due to cable path variations). Knowing the speed of light in fiber (five microseconds per kilometer), we can estimate the inter-site latency (one-way) as 2.3 milliseconds (ms).

Using the equation referenced above, we can estimate the bandwidth that a single user can actually achieve between New York and Boston. For the purposes of this analysis, we assume that the round-trip-time (RTT) is 4.6 ms, the packet loss rate is 0.1 percent, and the maximum data segment size is 1500 bytes. Note that the 1500-byte limit is based on the MTU (maximum transfer unit) limitations of Ethernet. Because all of this data originates at a server that is connected using Ethernet, this MTU is reflected throughout the system. For this combination of RTT, loss ratio, and MTU, the maximum single-user bandwidth is 108 Mbps. This limitation applies regardless of the how much faster the basic Internet access speed is. (That is, the limitation stems from the packet loss rate and retransmission — not the maximum bandwidth of the IP service.)

To interconnect two sites using the IP WAN of a major carrier, *each site* would require broadband Internet access. This will be factored into our cost model below. The actual costs for Internet service will thus be twice the monthly recurring costs shown above. Once service is obtained, each site is free to transmit as much data as desired (up to the data rate of the ISP access) although the effective

---

throughput for large-scale transactions will be limited to 108 Mbps due to packet loss and latency.

In the case of the “well-managed” private network, we will assume negligible packet loss (and therefore no performance limitation due to TCP/IP). However, the “well-managed” network will only provide 290 Mbps (half of the available 580 Mbps).

As discussed above, the Reliable Transport over SONET (RToS) model will deliver the full available bandwidth of the SONET link (580 Mbps). Exhibit 4

**EXHIBIT 4 — Cost Effectiveness of Data Transfer over the WAN**  
(Single-User, NYC to Boston)

|                                      | <b>Line Rate (Mbps)</b> | <b>Effective Single-User Data Rate (Mbps)</b> | <b>Time to Move 1 TByte of Data (hours)</b> | <b>Monthly Recurring Cost of Service (\$)</b> | <b>\$ per Mbps (Single-User Data Rate)</b> |
|--------------------------------------|-------------------------|---|---|---|--|
| OC-12 carrier-provided IP service    | 622                     | 108   | 23  | 196,000                                       | 1815                                       |
| OC-12 “well-managed” private network | 622                     | 290   | 8.4   | 39,000  | 134  |
| OC-12 Reliable Transport over SONET  | 622                     | 580   | 4.2   | 39,000  | 67   |

shows how these different options compare. Note the difference in the amount of time that it would take a single user to transmit a terabyte of data using OC-12 carrier-provided IP access (23 hours) and an OC-12 “well managed” IP network (8.4 hours). Compare this with the time required for RToS (4.2 hours). Note also the differences in cost of service (on a dollar per megabit per second basis). Carrier-provided IP access (\$1815 per Mbps) is an order of magnitude greater than the “well managed” network (\$134 per Mbps). However, the reliable transport model is better than the “well managed” IP network by another factor of two. This makes sense because the “well managed” IP network had to throw out half the available bandwidth in order to tame packet loss.

**CONCLUSIONS**

This article examined three options for interconnecting data centers that are geographically separated. These three options include routed IP services provided by a carrier, routed IP services managed by the enterprise, and Reliable Transport over SONET.

---

As seen, routed IP networks provide amazing flexibility, connectivity, and scalability but they do so at the cost of performance. Users who wish to move large amounts of data through carrier-based routed IP networks are continuously thwarted in their attempts to do so. Because users purchase IP service based on peak capacity but carriers provision their networks based on average capacity, there will always be an economic incentive for the carriers to over-commit the capacity of their network. This means that packet loss is here to stay. As long as there is packet loss in the WAN, there will be performance issues. Quite simply, anything that gets dropped must be retransmitted. The farther apart the two sites are from each other, the longer it will take to retransmit the data. Performance degradation (additional latency) inevitably follows.

The conventional alternative to the packet loss issues of carrier-based IP service is to create a private IP network in which the enterprise is able to regulate the traffic and therefore to keep packet loss to a minimum. As seen, packet loss can be minimized but half of the network's bandwidth must be sacrificed to accomplish this. As the SONET channels of the enterprise become faster and faster, the amount of bandwidth that is wasted increases as well.

On the other hand, Reliable Transport over SONET provides a high-performance and cost-effective method of transferring large amounts of data across the WAN. The secret to high performance really is not such a secret — do not drop data. If you have not dropped the data in the first place, there is no need to waste the precious bandwidth and time of the system by retransmitting it.

While routed IP will always remain the king of scalability, reliable transport will always provide the highest levels of performance and the most cost-effective method way of transporting large amounts of data across the WAN. In the end, as the requirements for large-scale data transfer increase, economics and practicality will dictate the solutions chosen by the end user. Reliable Transport over SONET provides performance that dramatically exceeds anything available via routed IP (carrier based or “well managed”).

## Notes

1. This article compares carrier-based IP service with private-line IP service and reliable transport of IP and Fibre Channel. Here, we emphasize the economic cost of creating the “well-managed network.” A more thorough analysis of service costs for carrier-based IP versus reliable transport between three different cities can be found in *The Economics of Large Scale Data Transfer*, by A. Helland, LightSand Communications, 2002.
  2. “The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm,” Mathis, Semke, Mahdavi, and Ott, *Computer Communication Review*, a publication of ACM SIGCOMM, Vol. 27, No. 3, 1997.
  3. Packet loss rates were obtained from [www.internetweather.com](http://www.internetweather.com), an independent source of IP network performance measurements statistics.
  4. [www.unix.mcs.anl.gov/discovery/wufeng.htm](http://www.unix.mcs.anl.gov/discovery/wufeng.htm).
  5. “Delay Bounds in a Network with Aggregate Scheduling,” draft version 4, 2/29/00, Charny, Cisco Systems.
  6. New pricing information for SONET was recently discovered that indicates that the POP-to-POP pricing for the major metropolitan areas may actually be substantially lower than the estimates used in this article. See [www.telegeography.com](http://www.telegeography.com). The abundance of available fiber and the current economy have produced a significant downward pressure on the price of SONET service between major metropolitan areas.
-

---

---

Andy Helland ([andyh@lightsand.com](mailto:andyh@lightsand.com)) is director of product management at LightSand Communications ([www.lightsand.com](http://www.lightsand.com)) in Milpitas, California. LightSand manufactures a SONET gateway that is optimized for moving large amounts of data over the WAN without packet loss. Helland is a member of the ANSI T11 standards group that specifies Fibre Channel and Fibre Channel extensions. He is also one of the contributing authors of the IETF standard for carrying Fibre Channel over IP (FCIP).

---