

Latency Variance in Wide Area Transport

Andy Helland

andyh@lightsand.com

A brief comparison of the latency variance associated with different options in Wide Area Transport.



December 2003

Introduction

When interconnecting remote data centers, most people assume that routed IP is the only option that is available to them for wide area transport. In this white paper, we will briefly review all of the available options and discuss the variance in latency associated with each of them. This latency variance may be a significant issue for applications such as synchronous mirroring or distributed file systems. In a synchronous mirror, *both* storage systems (local and remote) must acknowledge the SCSI WRITE command before the computer can move on. Additional latency (say from a transport system that has variable latency) will therefore slow down the WRITE process (in an unpredictable fashion). In systems that employ distributed file locks, additional latency will slow the system down because a cluster node that is expecting a quick authorization to access a file will have to wait for a longer time to receive that access. Every increase in latency slows the system down.

In a separate white paper, we address the impact of latency and packet loss on the *macroscopic throughput* of distributed storage systems¹. Here, our emphasis is on the variability of latency and its impact on the performance of higher-layer operations such as synchronous mirroring and distributed file systems.

SANs, LANs, MANs, and WANs

First, let's define some terms regarding the distance of data movement. In this context, we will define local area networks (LANs) as those in which distances are generally short and for which it is easy for the enterprise to add additional bandwidth. Typically, this would include "inside the data center" but it could also include the "campus environment" as well. Distances are typically less than 1-2 Km.

For the Metropolitan Area Network (MAN), the distances are limited to 50-75 Km and typically, a commercial carrier is used to provide the connectivity.

For Wide Area Networks (WANs), a carrier is always involved and the distances can be regional, national, or international (75 Km up thousands of Km).

Options for Moving Data

Within the SAN or LAN, it is relatively easy to add bandwidth. Assuming that the switching fabric is capable of supporting the additional bandwidth, one need only add more GbE or FC cables to increase the throughput. If the LAN extends to a campus, the fiber can be trenched from building to building by the enterprise and installed relatively easily. The Enterprise is free to pull additional cables or use any protocol that is appropriate.

As the geography of the fabric grows to metro size (MAN), it becomes increasingly difficult to interconnect separate sites. Site-to-site communication is limited to dedicated

¹ Please refer to "*The Economics of Large-Scale Data Transfer*". This paper is available for download from the LightSand website.

fiber or services provided by commercial carriers. Dedicated fiber can range in cost from hundreds of thousands of dollars per mile up to a million dollars per mile. Although it can be expensive, it is still an option for the large enterprise. If commercial carriers are involved, the mode of delivery will typically be:

- a) Point-to-point SONET (well established)
- b) Routed IP over SONET (well established)
- c) Layer 2 Gigabit Ethernet (emerging)
- d) Leased lambda service (emerging)

These services are all delivered to the enterprise using duplex fiber optic cables.

Finally, as the distance between data centers grows to the Wide Area, we see the available options for interconnecting sites reduced even further. By the time we have moved out of the MAN into the WAN, the only available options are (a) and (b). These two transport options have decidedly different characteristics with respect to their variations in latency, as we will discuss below.

Latency Variance

When a SONET link is established between two sites, a dedicated channel is committed to the link. Because the link is stable and there are no routers involved, the latency is deterministic and stable. If it were to be drawn as a histogram, it would appear as a narrow spike. This is shown in red in the diagram below. The area under the curve is 100% because this represents a probability distribution of latency variance.

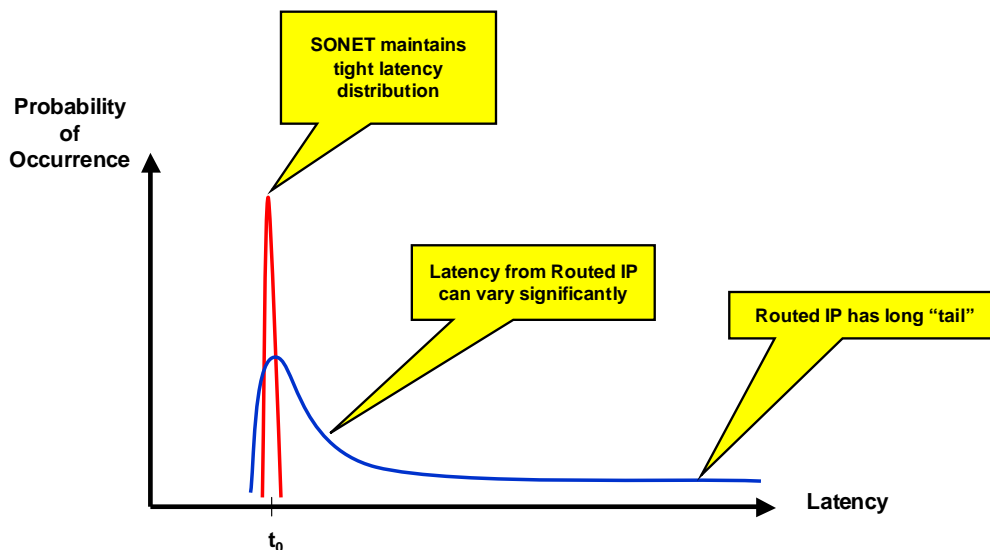


Figure 1. Notional Probability Distribution of Latency Variance

On the other hand, routed IP networks are comprised of numerous routers that are multiply interconnected. Since the networks have many options for delivering data from one site to another, routing control protocols such as RIP, OSPF and BGP are used to

create a path from one location to another. These routing protocols constantly adjust for changes in routing paths that result from equipment failures, link failures, or simply the addition of new equipment into the network. The actual path taken by IP data from one site to another will be largely the same but will still exhibit much variation based on routing changes that are constantly occurring outside the control of the IP data itself. Because of this, the probability distribution for routed IP networks is better approximated by the blue curve in Figure 1.

As with the SONET latency distribution, the arrival time is largely centered on the flight time through the network but the distribution is much fatter and exhibits a long tail towards higher latency delivery. Another significant component to the spread of the distribution curve is packet loss. Routers are designed to drop data to protect themselves when large surges of data occur. From the perspective of the user, dropped IP packets will be re-sent through the network using the TCP protocol. However, since these packets must be re-sent, they will arrive at the application with a significant increase in latency. This affects the packets that have been dropped as well as all subsequent packets that have been received but cannot yet be used. TCP adds a sequence number to all packets and ensures that they are delivered *in order*. As with the SONET curve (red), the area under this curve is still 100% because this represents a probability distribution. Note that this curve is notional only to illustrate the general nature of the distribution.

Logical Paths vs. Physical Paths

Latency is a critical factor in deploying distributed storage and computing systems. Despite this fact, it remains largely ignored by the networking industry. Routed IP networks will typically derive the shortest *logical* path but not always the shortest *physical* path. Since latency is not factored into the routing protocols, there is no penalty for taking a longer path from one site to another. Routing protocols will minimize the number of hops between routers and gravitate towards higher bandwidth links without any paying any attention towards latency. While this may be optimal for managing the aggregate flow of data through the network, very often this produces results that *run directly against the best interest* of individual users. Let's see how this might happen.

Figure 2 is a traceroute taken from LightSand's Corporate Headquarters in Milpitas, California to the website for the City of Sacramento (www.cityofsacramento.org). Traceroute lists all of the routers that are used in an IP path and estimates the roundtrip latency to each router along the way. Including the gateway router at LightSand, notice that 15 routers are involved in the communication path from Milpitas to Sacramento. Notice that the latency has increased to a steady 20 mS by the time we have reached the final destination. Although the resolution of the traceroute is only 10 mS, this is a clear indication the latency is somewhere between 15 mS and 24 mS.

```

C:\Documents and Settings\andyh.LIGHTSAND>tracert cityofsacramento.org

Tracing route to cityofsacramento.org [206.170.172.2]
over a maximum of 30 hops:

  1  <10 ms    10 ms    <10 ms    milpitas-router.lightsand.com [10.10.0.1]
  2  <10 ms    <10 ms    <10 ms    64.214.104.161
  3  <10 ms    10 ms    <10 ms    s11-1-1-3-0.ar1.SFO1.gblx.net [64.214.96.77]
  4   10 ms    <10 ms    10 ms    pos2-2-155M.cr2.SFO1.gblx.net [67.17.72.142]
  5  <10 ms    10 ms    10 ms    so1-1-0-2488M.ar1.SJC2.gblx.net [67.17.64.65]
  6  <10 ms    10 ms    10 ms    208.50.13.94
  7  <10 ms    10 ms    10 ms    144.232.20.58
  8  <10 ms    10 ms    10 ms    sl-gw18-sj-14-0.sprintlink.net [144.232.3.26]
  9  <10 ms    10 ms    10 ms    sl-swb-87-0.sprintlink.net [144.228.44.42]
 10  <10 ms    <10 ms    10 ms    bb1-p4-0.sntc01.sbcglobal.net [151.164.188.206]
 11  10 ms    20 ms    10 ms    bb1-p14-0.scrm01.sbcglobal.net [151.164.188.122]
 12  10 ms    30 ms    20 ms    ded3-g1-3-0.scrm01.pbi.net [64.171.152.236]
 13  20 ms    10 ms    20 ms    VIP-City-of-Sacramento-1052433.cust-rtr.pacbell.net
[206.13.19.142]
 14  20 ms    20 ms    20 ms    ppp-207-105-160-5.dialup.anhm01.pacbell.net
[207.105.160.5]
 15  20 ms    10 ms    20 ms    capital.sacto.org [206.170.172.2]

Trace complete.

C:\Documents and Settings\andyh.LIGHTSAND>

```

Figure 2. Traceroute from Milpitas to Sacramento

Let's examine this trace a little more closely. No less than three carriers are involved in providing connectivity between Milpitas and Sacramento. LightSand's Internet Service Provider (ISP) is Global Crossing. This is represented by "gblx" on line 3, 4, and 5. Sprint is also involved (line 8 and 9). Finally SBC/Pacbell touch our data (line 10, 11, 12, 13, and 14). Even more significant is the path that was chosen to get from San Jose to Sacramento. Note that the data immediately goes to San Francisco (line 3 and 4) and then returns back to San Jose (line 5) before actually leaving the bay area towards Sacramento! Clearly, this routing choice was not based on minimizing the inter-site latency between Milpitas and Sacramento. As chosen, the IP path took an unnecessary roundtrip to San Francisco (150 Km as the crow flies).

We have seen that a routed IP network has chosen a path of approximately 15 – 24 mS to route data from Milpitas to Sacramento. Let's examine what the inter-site latency might have been had we created a dedicated SONET path between the two sites. The "crow fly" distance from Milpitas to Sacramento is approximately 131 Km. However, fiber optic cables will never take the "crow fly" distance. Instead, they will follow major right-of-way paths such as railroads, freeways, and rivers. Also, the cables must first be terminated in the local carrier POP (point-of-presence) facilities and then enter the SONET backbone. All of this will add distance to the "crow fly" length of the cable. We have observed additional distances of approximately 20% to 30%. For planning purposes, let's add 50% to the "crow fly" distance. 131 Km + 50% = 196 Km. Let's call it 200 Km. The speed of light in glass is 5 μS/Km. Therefore, we should expect to see a one-way latency of approximately 1 mS (2 mS round trip). If we add the contribution of the gateway equipment, the expected roundtrip latency is still only 2.1 mS. Furthermore, this latency is stable since it is not routed.

Impact of Latency on File I/O Operations

Let's look at the impact of latency on file I/O operations. If we separate a disk system from its server, the addition of latency on the link will have a direct impact on the maximum number of file I/O operations that can be performed.

In the scenario described above, we had two possible links that might connect Milpitas and Sacramento. The first was a routed IP link and the second was a dedicated SONET channel. The routed IP link was measured to have a minimum roundtrip latency of 15 mS (minimum). The actual latency may be considerably higher but for comparison purposes, let's start with the lowest value. The second link was a dedicated SONET channel whose latency we estimated to be 2.1 mS.

Let's look at a SCSI READ operation. The READ operation starts with a request from the Initiator and ends with the data returning from the Target (two trips through the link). Assuming there is no additional latency from the disk controller, the SCSI READ rate will be on the order of the reciprocal of the roundtrip time. For a routed IP network, that translates to $1/(15 \text{ mS})$ or approximately 67 READ operations per second. In the case of the dedicated SONET interconnection, the latency is considerably smaller (2.1 mS). The maximum file I/O rate could therefore be estimated at 476 I/O operations per second. These values are summarized in the table below.

	Roundtrip Time (RTT)	SCSI READ Operations per Second
Routed IP (best)	15 mS	67
Routed IP (typical)	20 mS	50
SONET	2.1 mS	467

Table 1. SCSI READ Operation per Second. SONET versus Routed IP.

We have assumed that the impact of latency can be characterized by the reciprocal of the roundtrip time. In fact, it may be much worse than that. A SCSI WRITE operation requires four trips through the link to execute the command.

Start WRITE →
← Ready
Here's Data →
← Ready

Thus the throughput rate will be on the order of $1/(2 * \text{RTT})$. The estimated WRITE performance of a distributed storage system is shown below in Table 2.

	Roundtrip Time (RTT)	SCSI WRITE Operations per Second
Routed IP (best)	15 mS	33
Routed IP (typical)	20 mS	25
SONET	2.1 mS	238

Table 2. SCSIWRITE Operations per Second. SONET versus Routed IP.

There are steps that can be taken to mitigate the impact of high latency links such as the use of SCSI Command Queuing, etc. These steps allow the server to create multiple outstanding requests at the same time. Nonetheless, the impact of latency on any particular SCSI operation is very significant. Please see *Bandwidth vs. Latency in SAN Extension* for a more detailed discussion of this effect².

Conclusions

We have examined the causes of latency when moving data across the WAN. Latency comes from the actual length of fiber in the ground, unusual routing paths chosen by IP networks and also packet loss and subsequent retransmission. We have illustrated the differences in behavior that can be seen with routed IP networks when compared with dedicated SONET channels and illustrated why dedicated SONET channels offer significantly improved performance. Not only is the latency dramatically lower with the use of dedicated SONET channels, it is also stable. Both of these factors work together to offer the user significantly higher performance.

There are many applications for which routed IP transport is a great solution for WAN connectivity. However, when performance is extremely important, a holistic view of bandwidth, latency, and packet loss should be taken. Fully acknowledged protocols such as SCSI are particularly sensitive to increased latency. Dedicated SONET links for site-to-site connectivity of Fibre Channel and IP offer a cost-effective and very high performance method of creating distributed computing and distributed storage.

² *Bandwidth vs. Latency in SAN Extension* was published in InfoStor Magazine in December, 2001. It is available for download from www.infostor.com or from LightSand.