



What is Network Latency and Why Does It Matter?

By
O3b Networks, Ltd.

November 11, 2008

This paper is presented by O3b Networks to provide clarity and understanding of a commonly misunderstood facet of data communications known as latency. It is our goal that the reader takes away a clear understanding of latency and gains insight into the significant advantage that O3b Networks brings to the satellite communications market for the developing world.

Contents

Executive Summary.....	4
What is latency?.....	4
Networking 101.....	4
What Causes Latency?	7
What is Propagation Delay?.....	9
Transmission Rate and Bandwidth	9
How does TCP know a link is operating poorly and what can it do about it?	12
A Fictional Data Download.....	13
The real world	14
Satellite Link Latencies.....	18
Summary	20

Executive Summary

Internet data is packaged and transported in small pieces of data. The flow of these small pieces of data directly affects a user's internet experience. When data packets arrive in a smooth and timely manner the user sees a continuous flow of data; if data packets arrive with large and variable delays between packets the user's experience is degraded.

This paper will address the sources of delays to internet data and describes how the users of the O3b Satellite constellation will be observe significantly less impact from these delays when compared to other satellite-based systems.

What is latency?

Definition of Latency from Wikipedia:

Latency is a *time delay* between the moment something is initiated, and the moment one of its effects begins or becomes detectable. The word derives from the fact that during the period of latency the effects of an action are latent, meaning "potential" or "not yet observed".

Most people understand that it takes time for web pages to load and for emails to get from your outbox to the destination inbox and yes, this is a form of latency. But in order to understand why this happens we need to think about latency at a lower level:

Latency is a time delay imparted by each element involved in the transmission of data.

The remainder of this paper will discuss the specific mechanisms involved and how latency limits the performance of a data communications link.

Networking 101

It's important for the reader to understand the very basic elements of networking to properly grasp the latency issue.

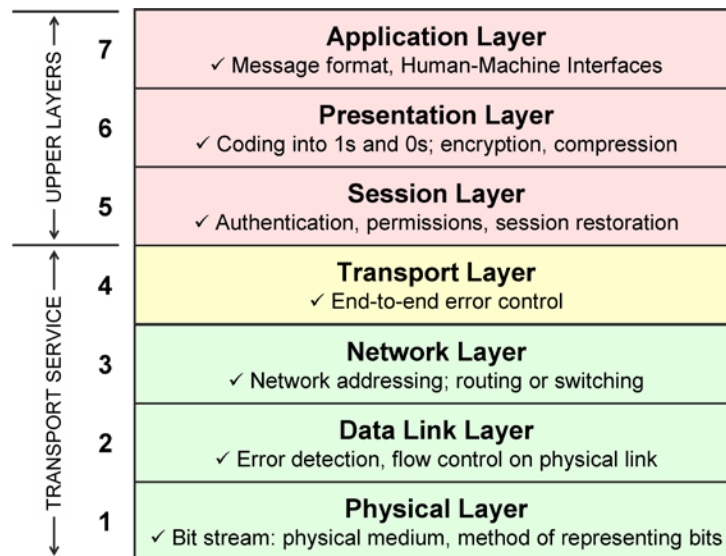
Early networking engineers anticipated the need be able to handle thousands to millions of users on one cohesive network, and thus the TCP/IP networking model was developed.

The key design feature of the TCP/IP networking model is the concept of encapsulation which is the idea of taking data and wrapping it in a common container for shipping. The container that was developed is called the IP Datagram, also known as an "IP Packet".

The IP Packet is a very simple thing: a header, followed by data.

The Header contains information used for routing the packet to the destination. The data can be any information which needs to be transported such as a snippet of streaming music or a portion of email traffic. The exact construct of the data portion of an IP Packet is defined by the data protocol that is being carried. Data protocols will be discussed later.

To understand exactly where latency occurs, it's valuable to know how this most basic unit of networking data is built and transported. For this we turn to the OSI Model:



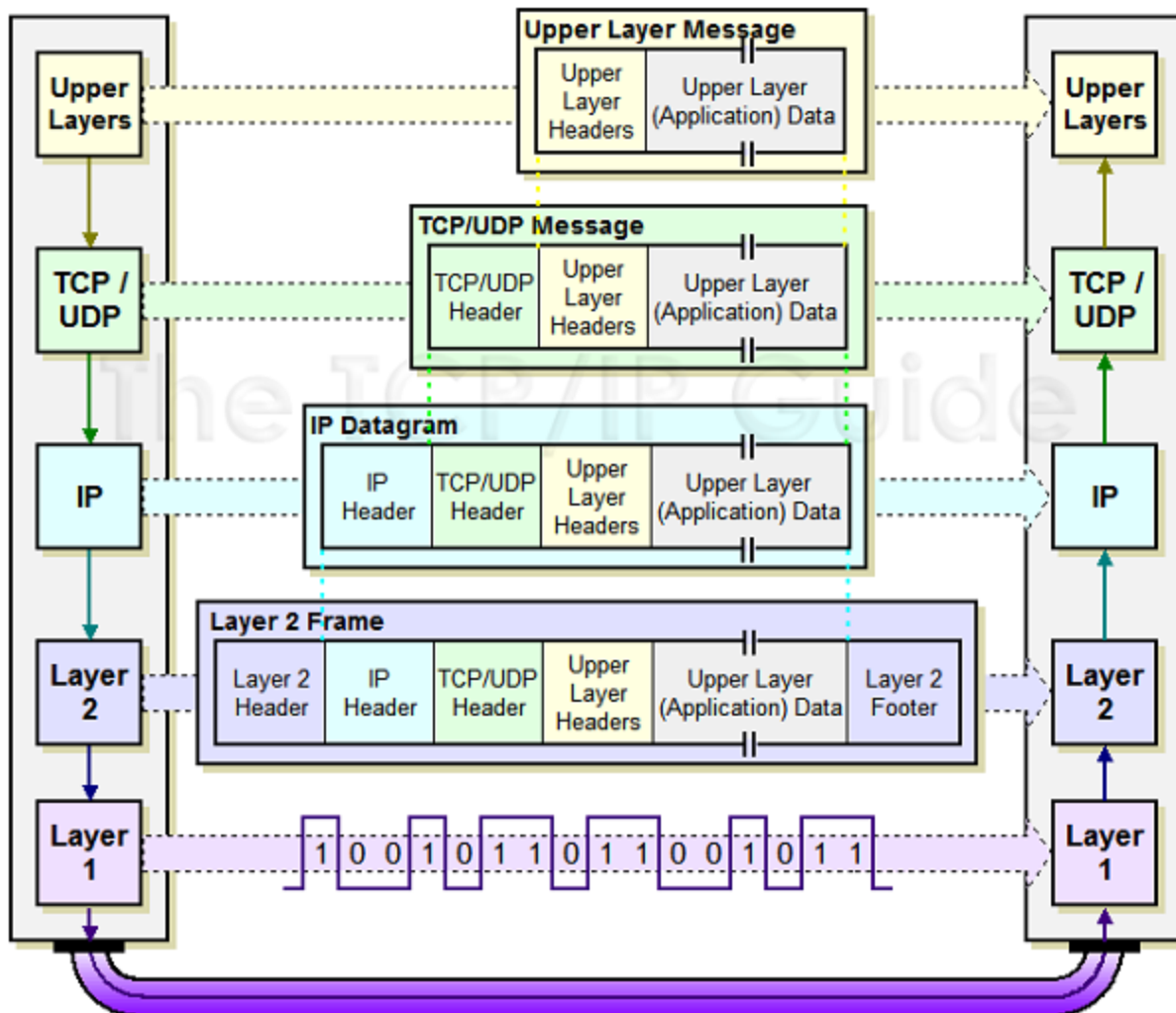
The OSI Model

The OSI model was created to describe the process of turning your application data into something that can be transported on the internet.

The upper layers of the OSI model describe things that happen within the applications that are running on the computer. These applications are web browsers, email programs, etc.

The lower layers are where information to and from applications are turned into data for transport on a network. This is where data encapsulation occurs and our basic networking data element – the IP Datagram or “packet” is built.

The following diagram shows the encapsulation process in what's known as the TCP/IP Stack. The precise workings of the TCP/IP stack can be different between various computer operating systems. These differences may seem trivial as long as the protocols are implemented properly but when seeking the absolute highest levels of performance it's important to know that the network stack implementation can be a significant cause of networking performance variability.



The TCP/IP Stack

The transport of network data is a three step process:

1. Data from a source application is passed down through the stack. During this process the application data is wrapped into IP Datagrams which are commonly called “packets”. Packets are then transmitted by the sending computer in the network
2. Packets are passed along the network (purple line) until they reach the destination computer
3. Packets are received from the network by the destination computer and are passed up through the stack. During this process the application data is extracted and the passed along to the destination application.

The additional encapsulation at Layer 2 is called framing. This is the stage where the IP Datagram is turned into bits which are appropriate for a particular type of network.

Layer 1 is the physical network medium connection. This layer handles the conversion of the layer 2 bits into electrical, optical, or radio signals that can be transported. The network interface, often called the NIC or Network Interface Card, can be fiber-optic, copper wire, or a wireless radio interface.

What Causes Latency?

As described above there are many logical, electrical, and physical elements involved in computer networking. The OSI model identifies each of these elements with regard to specific functionality and delays, another name for latency, occur at every stage of the process.

Application Layer Latency

Layer 7, 6, 5 are the upper “application layers”. Regardless of the speed of the processor or the efficiency of the software it takes a finite amount of time to manipulate and present data. Whether the application is a web page showing the latest news or a live camera shot showing a traffic jam, there are many ways in which an application can be affected by latency. One common source of application latency is the need to read and write data to a disk. There are also hardware limitations which affect application performance such as the amount of memory.

Serialization Latencies

The encapsulation of data which occurs at the Transport Layers (1 through 4) is called serialization. Serialization takes a finite amount of time and is calculated as follows:

Serialization delay = packet size in bits / transmission rate in bits per second

For example:

- Serialization of a 1500 byte packet used on a 56K modem link will take 214 milliseconds
- Serialization of the same 1500 byte packet on a 100 Mbps LAN will take 120 microseconds

Serialization can represent a significant delay on links that operate at lower transmission rates, but for most links this delay is a tiny fraction of the overall latency when compared to the other contributors.

Data Protocols and Latency

Protocols which operate in the transport layers range from simple to advanced. Many transport protocols use information obtained from the timing of the data flow to ascertain link

performance and attempt to adapt to specific conditions present on the network. This is a key issue and will be addressed later in this paper.

Routing and Switching Latencies

For a network to do its job packets have to be passed from Point A to Point B. This would be simple if the internet was just two computers and two locations, but this is certainly not the case. In IP networks such as the Internet, IP packets are forwarded from source to destination through a series of IP routers or switches that are interconnected by links such as circuits. The IP routers use the destination address in the IP header to determine the next router in the path from source to destination. The IP routers utilize routing algorithms to continuously update their decision about which router is the best one to get the packet to its destination. A router or circuit going down or congestion along the path can change the routing.

Managing the wealth of internet traffic induces delays (latencies) caused by the routing and switching process. This refers to the amount of processing time for a router or switch to receive a packet, process it and transmit it on its way.

Modern IP hardware interfaces have delays on the order of a few nanoseconds and are generally negligible when compared to network propagation delays which are discussed in the next section. High performance IP routers and switches each add approximately 200us (microseconds) of latency to the link due to processing and forwarding. If we assume that the average IP backbone router spacing is 800 km, the 200us of routing/switching delay is equivalent to the amount of latency induced by 40km of fiber. Therefore routing/switching latency tends to contribute only 5% of the end to end delay for the average internet link.

Queuing and Buffer Management

Another issue which occurs within the transport layers is called “queuing latency”. This refers to the amount of time an IP packet spends sitting in a queue awaiting transmission due to over-utilization of the outgoing link after the switching delay has been completed.

While over-utilization of high-speed Internet backbone links tends to be rare, instances of congestion are common on lower speed access circuits. When congestion occurs routers use sophisticated data queue management algorithms such as WRED (Weighted Random Early Detection) to minimize data loss. Congestion can cause queuing delays to become infinite since packets are dropped when router buffers become full. Queuing algorithms use a variety of packet management schemes to ensure queuing latency is minimized; best common practice WRED configurations typically bound queuing latency at 20ms.

What is Propagation Delay?

Propagation delay is a phenomenon in which the physical properties of the medium cause the transmitted information to slow down. The amount of slowing caused by the medium is called the velocity factor (VF).

Most people are surprised to learn that copper wire and fiber-optic cables have similar velocity factors. Fiber optic cables typically measure around 70% of the speed of light whereas copper cable varies from 40% to 80% depending on the construct. Coaxial cable is commonly used and many types have a VF of 66%.

Satellite communication links use electromagnetic waves to propagate information. The information is converted from electrical signals to radio signals using a modem (modulator/demodulator). Once these radio signals leave the antenna, they travel at the speed of light.

Let's calculate how long it will take an email to travel from New York to London assuming that we are the only user on a private communications channel.

Ignoring the actual routes taken by undersea cables due to the ocean's floor, let's assume the path from New York to London is the great circle distance of 5458 km.

Propagation delay = distance/speed:

The email sent using a copper link:	$5458 / 197863.022 = 23.58 \text{ ms}$
The email sent using a fiber-optic link:	$5458 / 209854.720 = 26.01 \text{ ms}$
The email sent using a radio link:	$5458 / 299792.458 = 18.21 \text{ ms}$

These are the latencies caused only by propagation delays in the transmission medium. If you were the only one sending one single data bit and you had unlimited bandwidth available, the speed of the packet would still be delayed by the propagation delay.

This delay happens without regard for the amount of data being transmitted, the transmission rate, the protocol being used or any link impairment.

Transmission Rate and Bandwidth

Transmission Rate is a term used to describe the number of bits which can be extracted from the medium. Transmission rate is commonly measured as the number of bits measured over a period of one second.

The "maximum transmission rate" describes the fundamental limitation of a network medium:

If the medium is a copper Local Area Network, maximum transmission rates are commonly 10, 100, or 1000 Megabits per second. These rates are primarily limited by the properties and construction of the copper wires but the capabilities of the network interface card are also a factor. Even the most inexpensive LAN cables can handle transmission rates of 100 Mbps, but if the NIC only supports 10 Mbps then the link will be rate limited to 10 Mbps.

Fiber-optic the transmission rates range from around 50 Mbps up to 10 Gbps. While 10 Gbps is the highest existing standard, work is being done to extend fiber transmission rates to 100 Gbps. Unlike copper networks, the primary factor limiting fiber-optic transmission rates is the electronics which operates at each end of the fiber.

Wireless LANs and satellite links use a modem (modulator/demodulator) to convert the bits into a modulated waveform, and then on the other end a demodulator will then convert the signal back into bits. The limiting factor in radio-based links is the fact the signal which carries the data must occupy a limited bandwidth when compared to wire or fiber links.

Radio Bandwidth

Signals transmitted using radio waves occupy radio spectrum. Radio spectrum is not an unlimited resource and as such must be shared. To prevent radio interference between users the use of radio spectrum is controlled by nearly every government on the planet. The amount of radio spectrum occupied by any given radio signal is called its bandwidth.

The nature of radio spectrum use is beyond this paper but it's important to understand that generally the occupied radio spectrum of a modem signal will increase with the data rate:

- Higher modem data rates cause the modem to occupy more radio bandwidth
- Lower modem data rates will let the modem occupy less radio bandwidth

Since radio spectrum is a limited resource, the occupied radio bandwidth is an important limiting factor in wireless and satellite links.

Data Bandwidth

In data transmission, the data bandwidth is synonymous to the transmission rate being used. Bandwidth is important because it defines the maximum capacity of a data link.

- A 10 Mbps copper LAN cannot sustain traffic flowing at a higher rate than 10 megabits every second.
- a satellite link using modems operating at a 600 Mbps rate cannot flow any more than 600 megabits every second.

It's very important to understand that data bandwidth is a maximum data flow obtainable over a given transportation segment over a given period of time.

Latency and TCP/IP

The final detail required for understanding why latency is important is an understanding what is going on in the transport layer.

Recall that the transport layer is the process of encapsulating application data into IP Packets suitable for transport. The protocol used in "data" portion of the IP packet defines the type of data exchange that is taking place.

There are two types of network data exchanges:

- connectionless
- connection based

Connectionless data exchange is where data is simply pushed to the destination without regard for its well-being. A connectionless packet traverses the internet bound for a destination computer but if anything happens to it along the way the sending and receiving computer are none-the-wiser. Transporting data in this manner seems risky but depending on the application it's generally not detrimental if a few packets get lost along the way. This is a common method of transporting streaming music, and video, and Voice over IP traffic.

The transport protocol commonly used for connectionless traffic is called User Datagram Protocol or UDP. This protocol is a very popular because there is no overhead or connection management – the data just sent along. There is no retransmission of lost packets because having them arrive late is not useful for voice or video that is being played out in real-time by the receiving computer.

Connection based data exchanges are far more complicated. These data exchanges rely on the establishment of a "connection" which manages every packet which is transmitted. The reason for this is that both the sending and receiving computer applications are very interested in ensuring the integrity of every piece of data being exchanged.

The transport protocol commonly used for connect-based traffic is called the Transmission Control Protocol, or TCP. TCP provides error free sequenced delivery of packets. If packets arrive out of order the TCP layer puts them back in order. If packets are missing, TCP asks for retransmission. The TCP protocol makes the Internet reliable.

To support the additional connection management features, TCP packets contain additional information in the header as well as many different packet "types". TCP packet types are used

throughout the establishment of the connection and are the key to providing control of the connection.

TCP connections use a client/server model to describe the sender and receiver of data. The use of the term “server” does not necessarily mean a computer performing server duties, rather it’s a computer that is listening for TCP connections.

TCP connections have three phases:

1. establish the connection
2. send the data
3. close the collection

Phase 1: Establishing the connection requires 3 packets:

- the client sends a connect request SYN (synchronize) packet to the server
- the server replies with a SYN-ACK (synchronize acknowledge) packet
- the client confirms the receipt of the SYN-ACK by sending back an ACK (acknowledge)

Phase 2: Once the link is established, the data transfer can start.

During the TCP data exchanges, ACK (acknowledge) and NACK (negative acknowledge) packet types are used to tell the sender that packets have been properly received. If a packet not received or it contains a bit error, the transmission of the exact same packet is repeated.

Phase 3: Upon completion of the data transport session, the connection will be closed by the following 3 packet exchange:

- the closing initiator sends a FIN (finish) packet
- the other side of the link replies with a ACK
- the closing initiator send a combination FIN/ACK to end the connection

How does TCP know a link is operating poorly and what can it do about it?

To protect the integrity of the data, TCP packets have several features:

- Sequence Numbers
- Timestamps
- Flow Control
- Congestion Control
- Checksums

All of these features are used to guarantee the integrity of the data. They are also used by TCP to determine the quality of the link and to tune the flow of data to maximize the use of the available bandwidth.

An example of this behavior is the way TCP responds to congestion control. The TCP congestion control process uses timers to examine the data flow and subsequent ACK/NACK responses. When TCP detects that ACK/NACKs are taking longer than normal to respond, TCP assumes that the link is being congested somewhere and will slow down the release of packets using flow control. This vital step helps reduce the impact of congestion at routing and switching buffers as well as receiving computer data processing limitations.

By throttling the output of data when congestion is sensed, the TCP congestion control mechanism plays a key role in the flow of TCP/IP traffic.

Congestion control is generally a valid response to a lethargic network since slow response does often indicate that a portion of the link has a data bottleneck.

A Fictional Data Download

Let's examine the details of downloading a digital image which is 4 megabytes in size. For calculation purposes we need to use bits so our 4 megabytes (4 MB) is actually 32 megabits (32 Mb).

Our downlink will use TCP/IP but we cannot simply create a 32 Mb packet to transport our image; such a large packet would be a very cumbersome to deal with. Internet traffic is made up of packets of variable sizes, but the Maximum Transport Unit (MTU) is generally only around 1500 bytes. Packets larger than 1500 bytes are considered "Jumbo" packets but handling these is not yet commonplace for many parts of the internet. To ensure our image file makes it, we'll stick with 1500 byte packets.

Our 1500 byte TCP packet has a header which is required for transportation but not useful image data. The size of the header can vary in length from 20 bytes to 60 bytes depending on TCP packet options. If we assume that our header is a full 60 bytes in length, this leaves only 1440 bytes for our image file data.

Based on the amount of data payload available, the maximum amount of image file data that can be transported is 11520 bits per packet. Even this number can be a little lower depending on upper layer formatting of the data, but for this exercise we'll assume that the entire TCP payload is useful image data.

Our total file size divided by this maximum packet size tells us that transporting our entire image file will take 2777.7 packets. The last packet would normally become shortened instead of our full MTU but to keep the math easy we'll round up to 2778.

Our assumptions:

- We are downloading the file from a local computer using a 10 Mbps LAN
- There is no other traffic on that 10 Mbps LAN – we get the whole pipe
- The link is operating perfectly, no need to repeat any data during this transmission
- There is no appreciable latency. This is a copper LAN and a relatively a short cable run which is not adding any appreciable propagation delay

Activity	# packets x packet length	Total bytes	Total Bits (bytes X 8)
Connection setup	3 x 60 bytes	180	1440
Transmit the file	2778 x 1500 bytes	4,167,000	33,336,000
ACK packets	2778 x 60 bytes	166,680	1,333,440
Connection close	3 x 60 bytes	180	1440
Totals	5562 packets	4,334,040	34,672,320 bits

Notice that even for a link that is operating perfectly, the transfer of our 32 Mb image will actually take 34.6 Mb of data. The overhead of TCP added 7.8% to the total amount of data which needs to be transported. This overhead is an absolute worst case since we assumed all packets need to be acknowledged.

Now that we have the number of bits transmitted, we need to calculate how long that should take.

This is simple: $34.6 \text{ Mb} / 10 \text{ Mbps} = 3.46 \text{ seconds}$. Using our nice clean, short networking link we should be able to transport our image file in just over 3 seconds.

The real world

We've made 4 assumptions in the above analysis which do not mimic the real world internet at all.

Assumption 1: The file you want is available on a local computer.
 Reality 1: The file is more likely to be some physical distance away. This is not necessarily a problem, but it means we need to traverse a much more complicated network path to get to our data.

Assumption 2: We get the whole 10 Mbps to ourselves.

Reality 2: This is feasible only for the local LAN. It is a certainty that once your little 1500 byte TCP/IP packet reaches the internet backbone, it will be joined with millions of other packets working their way through the internet. Your image file packets are going to be mixed in with other traffic such as emails, streaming music files, etc. You simply don't get the internet all to yourself.

Assumption 3: The link is operating perfectly.

Reality 3: Internet traffic is routed through an extremely complex collection of hardware which is scattered all over the earth. The reality is that sometimes a fiber or copper cable is cut or is mistakenly disconnected. A piece of networking equipment such as a router or a switch can break, leaving some other path to pick up and route the extra traffic. When this happens, internet traffic can start to fill up queues and bottlenecks occur. As mentioned earlier, queuing delays can become significant when the network is operating through a bottleneck.

Assumption 4: There is no appreciable latency present in our network.

Reality 4: The reality is that all of the earlier discussed sources of latency are genuine factors in real-world networks. The impact of latency starts to become noticeable when the latency is significantly longer than the transmission time for the data.

The previous example discussed the data transmission rate in terms of the number of bits per second. To understanding how the user is exposed to the effects of latency, we need to convert transmission rate into its measure of time.

Bit transmission time = $1/(\text{bits per second})$

Type of Link Time required to transmit 1 bit Time for one 1500 byte packet

14.4 Kbps telephone modem	69 microseconds	823 milliseconds
1 Mbps LAN	1 microsecond	12 milliseconds
10 Mbps LAN	100 nanoseconds	1.2 milliseconds
600 Mbps Satellite channel	1.6 nanoseconds	19.2 microseconds

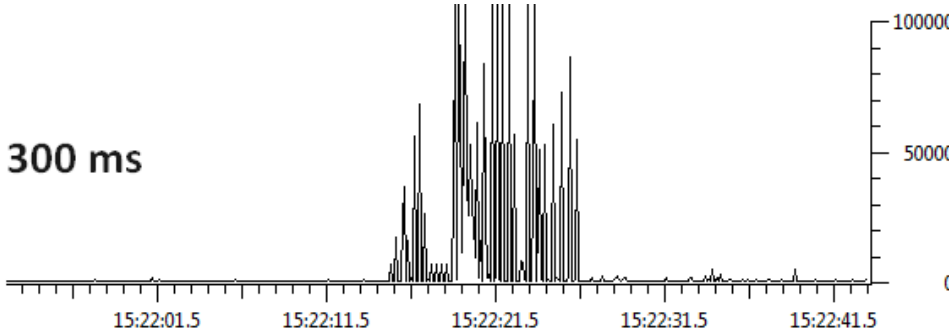
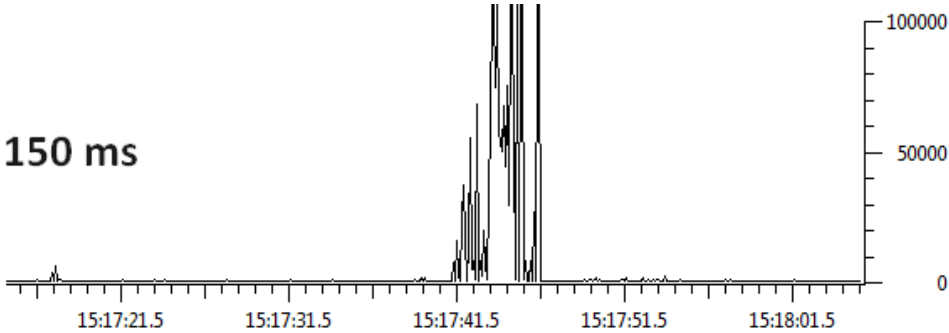
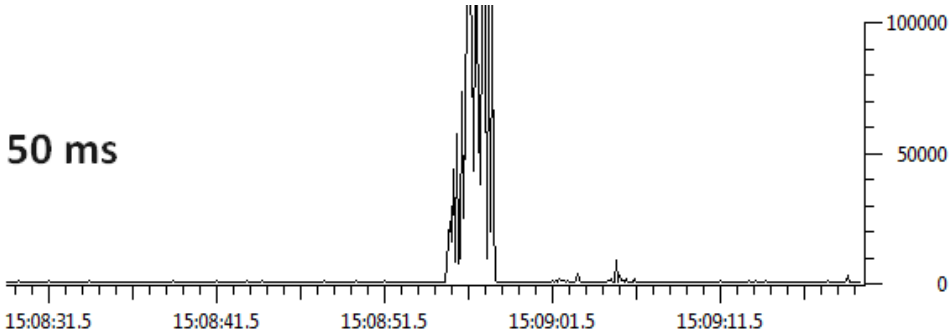
It's easy to see that slower transmission rates take longer to transport packets of data.

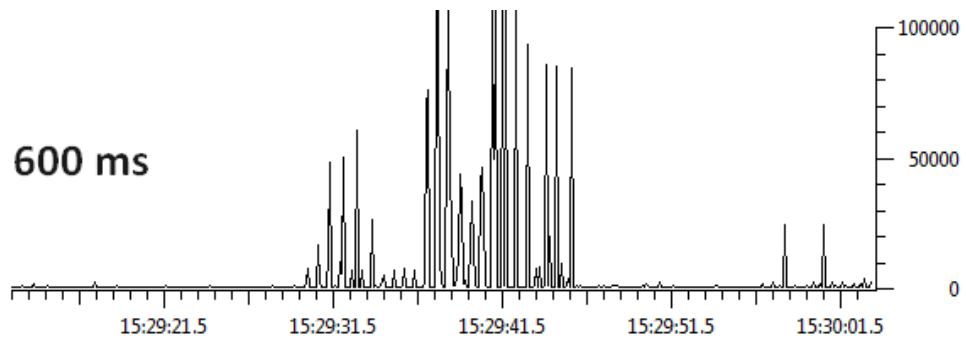
If the latency on a network is the same as the bit transmission rate, then the impact is very low since the IP packets can still be streamed very close to each other.

If the latency on a network is several times longer than the bit rate, the impact will become much more noticeable because the latency spreads the entire data TCP/IP data exchange session over time.

The following plots were made using a TCP/IP packet capture utility. These plots show the packet bit rate on the y-axis and time of day on the x-axis. The data being transmitted was the un-cached web-page reload of the content from the CNN web page (<http://www.cnn.com>).

The only condition changed during was the delay between packets – the transmission rate remained the same.





The added network latency and it affect on the flow of TCP data spread the web page load over time.

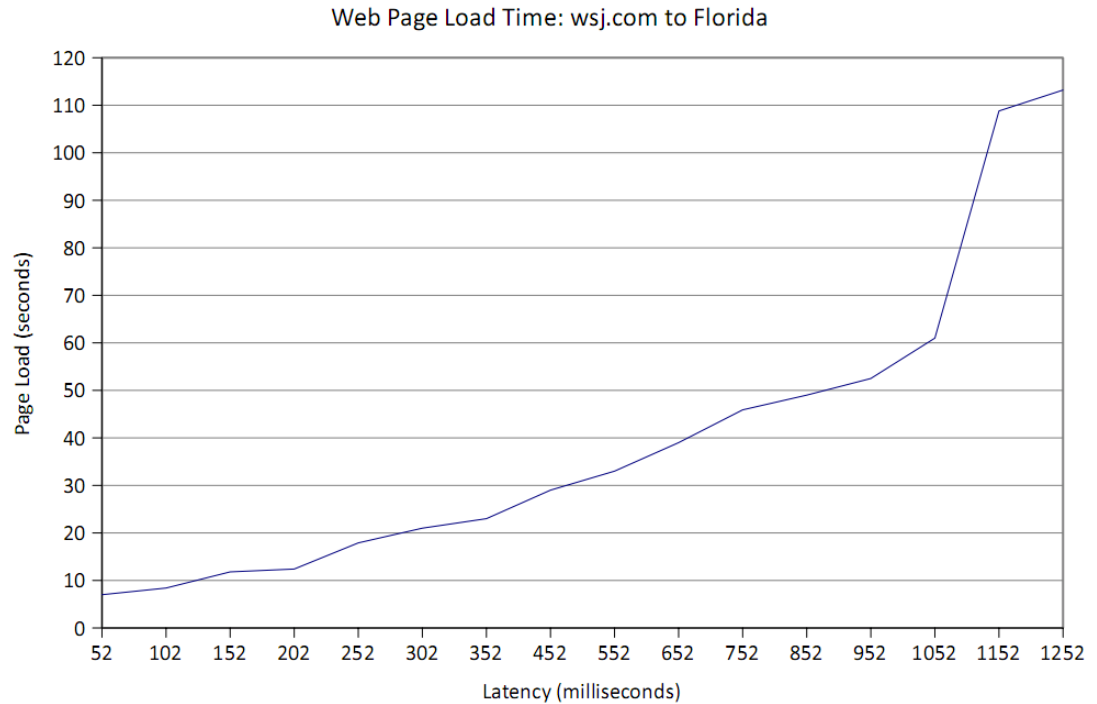
- The 50 ms latency link took 3 seconds
- The 150 ms latency link took 5 seconds
- The 300 ms latency link took 11 seconds
- The 600 ms latency link tool 17 seconds

The spreading of network data over time reduces what’s called the “Effective Bandwidth” of a link. Packets are still being transported at the same bit rate but due to latency it is taking much more time for all of the web-page packets to arrive.

It’s this “spreading over time” behavior of high latency networks which becomes noticeable to the user and creates the impression that a link is not operating at a high speed.

O3b recently conducted another demonstration of real-world effects of latency using the time to load a web page. This is a very common activity and clearly shows users that latency directly affects the way a user obtains data from the internet.

The following plots show the effects of latency on the time to load the Wall Street Journal web page (<http://www.wsj.com>):



It is clear that web page load times dramatically increase when latency increases. In this case doubling the latency nearly doubled the web page load times.

Satellite Link Latencies

Now that we know the effects of latency on real-world traffic, we'll discuss the latency differences in two satellite technologies.

Satellite links can introduce larger latencies than most terrestrial networking segments due to long distances from the ground stations to the satellite. The following table shows the latency caused by propagation delays from two types of satellite configurations, 1) the O3b Networks MEO orbit constellation at an altitude of 8063 Kilometers, 2) a geosynchronous satellite at 35,786 Kilometers.

Latency Calculations - Lagos, Nigeria to the Internet								
	Teleport to Satellite (km)	Customer to Satellite (km)	TP to Sat (ms)	Sat to Cust (ms)	Sat RTT (ms)	Internet RTT (ms)	Total RTT (ms)	Data Request Cycle (ms)
O3b Networks								
8 satellites	Almeria	Lagos						
AOS	10427	9019	35	30	130	60	190	285
Max Elev.	10126	8135	34	27	122	60	182	273
LOS	10986	9019	37	30	133	60	193	290
Geosynchronous	Italy	Lagos						
	37923	35847	126	120	492	60	552	828

Satellite Latency Calculations

It is important to understand that for satellites which operate as a bent-pipe, the propagation delays are doubled since the signal has to travel both up to the satellite and back down to the earth before it reaches the next segment of the network.

The table shows that a ground station in Lagos, Nigeria using an O3b Networks MEO satellite to connect to a teleport in Almeria, Spain will experience round trip time (RTT) ranging from 122 to 133 milliseconds. If we add in the average internet latency from the Almeria teleport to most internet destinations in Europe (60 ms), we end up with an overall latency from Lagos to a European internet site of 183 to 193 milliseconds. This range of latency is caused by the change in distance to the ground sites relative to the moving O3b Satellites. The AOS is the Acquisition of Signal, and the LOS is the point at which the O3b system will perform the handover to the next rising satellite. The Maximum Elevation is the point at which the satellite is closest to the customer ground station which explains why this point has the lowest latency.

By comparison, the same Lagos customer site using a geosynchronous satellite to a European internet site using will have to see latencies of 552 milliseconds.

The last column in the chart shows the time required to make a data request and to start receiving the requested data.

This data request time includes:

- The request packet from the user to the web server
- The web server acknowledging the request
- The web server pushing to requested data to the user
- The data arriving at the user's computer

TCP also includes a returned ACK packet from the user to the web server but this time is not counted in the Data Request Cycle.

Geosynchronous satellite users must wait almost 1 second before they start getting data, whereas the lower latency O3b satellite link will receive it nearly 3x sooner.

When looking at the basic latency numbers, it's easy to see that the O3b Satellite constellation will offer users a noticeably better internet experience with more immediate feedback and quicker access to data.

Summary

We have described the structure of IP-based packet switched networks, the functions of the various protocol layers, and the causes of latency in packet switched data networks, such as the Internet. Latency and overall throughput is dominated by two factors, the length of the route that the packets have to take between sender and receiver and the interaction between the TCP reliability and congestion control protocols and this path length. O3b Networks satellite constellation in a much lower MEO orbit has significantly lower path length and therefore significantly lower latency than traditional geosynchronous satellites. Therefore O3b's network latency and throughput approximate and in some cases exceed that of fiber based terrestrial networks.