What is Network Latency and Why Does It Matter?

by O3b Networks

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

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.

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.

This paper will address the sources of delays to internet data and describe why O3b Networks satellite constellation has lower path length and latency than traditional geosynchronous satellites and in some cases fiber based networks.

What Causes Latency?

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. Four key causes of latency are: propagation delay, serialization, data protocols, routing and switching, and queuing and buffing.

Propagation Delay

Propagation delay is the primary source of latency. It is a function of how long it takes information to travel at the speed of light in the communications media from source to destination. In free space, the speed of light is approximately 3 x 105 km/sec. The speed of light is lower in other media such as copper wire or fiber optic cable. The amount of slowing caused by this type of transmission 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 through the atmosphere and space. The information is converted from electrical signals to radio signals by the transmitter and the antenna. Once these radio signals leave the antenna, they travel approximately at the speed of light for free space.

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 ms
The email sent using a fiber-optic link:	5458 / 209854.720 = 26.01 ms
The email sent using a radio link:	5458 / 299792.458 = 18.21 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.

Serialization Delay

Serialization is the conversion of bytes (8 bits) of data stored in a computer's memory into a serial bit stream to be transmitted over the communications media. 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 a lower transmission rates, but for most links this delay is a tiny fraction of the overall latency when compared to the other contributors.

Voice and video data streams generally use small packet sizes (~20 ms of data) to minimize the impact of serialization delay.

Data Protocols and Latency

Data communications protocols at various layers in the protocol stack use handshakes synchronize the transmitter and receiver, for the transmitter and receiver to update each other on link status and to correct for erros in transmission. These handshakes take time to propagate across the link and can add to the latency of the transmission of information from source to destination.

Routing and Switching Latencies

In IP networks such as the Internet, IP packets are forwarded from source to destination through a series of IP routers or switches that continuously update their decision about which next router is the best one to get the packet to its destination. A router or circuit outage or congestion on a link along the path can change the routing path which in turn can affect the latency.

High performance IP routers and switches add approximately 200 microseconds of latency to the link due to packet processing. If we assume that the average IP backbone router spacing is 800 km, the 200 microseconds of routing/switching delay is equivalent to the amount of latency induced by 40km of fiber; routing/switching latency contributes to 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 routing/switching delay has been accounted for. This can add up to an additional 20 ms of latency.

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 of the copper wires and the capabilities of the network interface card are also a factor.

Fiber-optic the transmission rates range from around 50 Mbps up 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 local area networks (LANs) and satellite links use modems (modulator/demodulator) to convert digital bits into an analog modulated waveform at the transmitter end of a link, and then at the receive end a demodulator will then convert the analog signal back into digital bits. The limiting factor in transmitting information over radio-based channels is the bandwidth of the channel that is available to a particular signal and the noise that is present that will corrupt the signal waveform.

Radio Channel Bandwidth and Noise

Signals transmitted using radio waves occupy radio spectrum. Radio spectrum is not an unlimited resource and 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.

Noise in the radio channel will perturb the analog signal waveform and can cause the demodulator at the receiver to change a digital one into a zero or vice versus. The effect of noise can be overcome by increasing the power level of the transmitted signal, or by adding a few extra error correcting bits to the data that is being transmitted. These error correcting bits help the receiver correct bit errors. However, the error correction bits increase the bandwidth that is required.

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 of what is going on at the Transport protocol layer. The Transport layer sits between the application-specific Application layer and the IP layer that is used for packet routing. The Transport layer performs a number of functions including sequencing the packets into their proper order.

There are two types of Transport layer protocols:

- connectionless
- connection-oriented

Connectionless data protocols simply push the data packets to the end destination without regard for packet loss or packet order. Connectionless protocols are often used for real-time communications such as voice or interactive video where retransmission of lost packets is not useful because they will arrive too late to be played out. In the Internet, the most common connectionless Transport layer protocol is the User Datagram Protocol (UDP).

Connection-oriented protocols are more complex. These protocols rely on the establishment of a "connection" which requires handshakes and interaction between the source and destination computers. Connection-oriented Transport protocols provide error free sequenced delivery of packets. The most common connection-oriented Transport protocol used in the Internet is the Transmission Control Protocol (TCP). If packets arrive out of order the TCP layer puts them back in order. If packets are missing or corrupted in transmission, TCP uses an acknowledgement (destination back to source feedback) mechanism to ask for the missing or corrupted packet to be retransmitted. The TCP protocol also measures the performance of the link in terms of available bandwidth, delay and error rate. Using a windowing technique, it adjusts the transmission rate of data packets to match the ability of the channel between source and destination to accept them and deliver them correctly. As the TCP protocol at the source and destination determine that the channel is operating efficiently with low delay, it will open the window to allow more packets to be outstanding (in flight between the source and destination and not yet acknowledged). This has the effect of increasing the transmission rate. As errors, congestion, or delay increases, the TCP protocol will close down the window to slow down the transmission rate. Because of the high delays in geosynchronous satellite links, the TCP protocol acts to reduce the overall throughput of the link. To compensate for this, satellite link equipment vendors have developed TCP spoofing mechanisms that fake the TCP handshake by terminating the TCP link locally to the transmitting site and try to convince the transmitting site that the link has lower delay than it really has. This spoofing works if the link is relatively uncongested and has a very low error rate, which is not necessarily the case for satellite links.

Satellite Link Latencies

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.

Altitude is key determinate in latency because the higher the satellite the greater the latency.

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 following table compares two types of satellites using a ground station in Lagos, Nigeria.

Latency Calculations – London Internet Exchange (LINX) to Lagos

	Gateway to Satellite km	Customer to Satellite km	Gateway to Satellite msec	Customer to Satellite msec	Total Sat RTT msec	Fiber to LINX km	Fibre RTT msec	Total Sat RTT msec
O3b w/8 Satellites	Spain GWY	Lagos						
	10126	8135	33.8	27.1	121.7	2144	10.7	132.5
Geosyndchronous Sat	UK Teleport	Lagos						
	38,469	35,847	128.2	119.5	495.4	135	0.7	496.1
Fiber	London Lagos km							Total Fiber RTT
	8,000							80.0

Satellite Latency Calculations

The O3b Networks MEO satellite using to a teleport in Spain will experience round trip time (RTT) of approximately 122 msec. If we add in the average Internet latency from the Spain teleport to the London Internet Exchange (LINX) (~10 msec.), we end up with an overall latency from Lagos to a European internet site of 132.5 msec. O3b's 132.5 msec. round trip latency is close to fiber's 80 msec. RTT

By comparison, the same Lagos customer site using a geosynchronous satellite to a UK Teleport and then on to LINX will have latencies of 496 msec.

To actually retrieve a web page from an Internet server at LINX, will take some additional time for the data request packets and their acknowledgments:

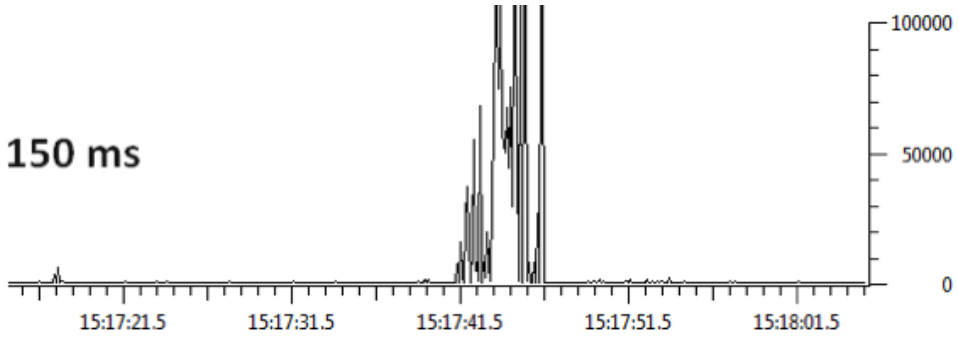
This data request time includes:

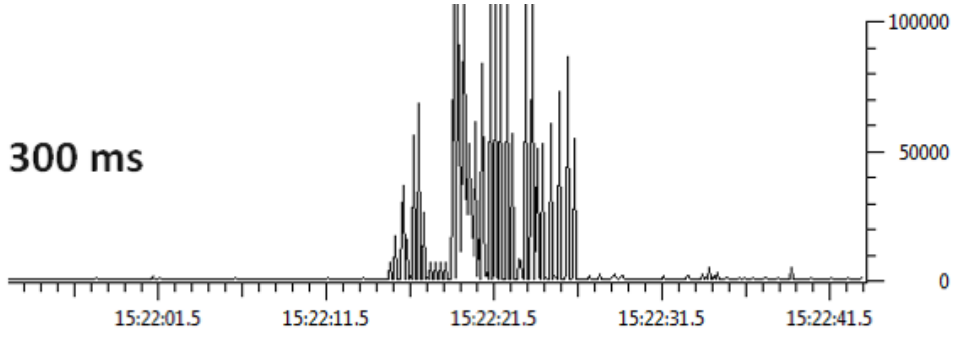
- The request packet from the user to the web server
- The web server acknowledging the request
- The web server pushing to requested data to the user
- The data arriving at the user's computer

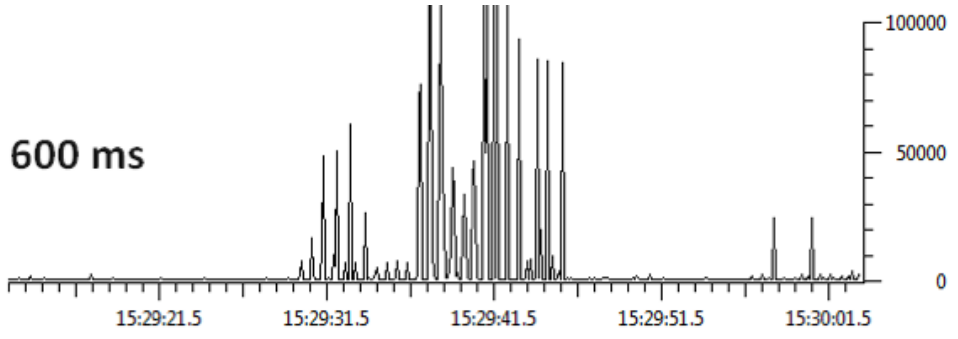
Geosynchronous satellite users must wait almost 750 msec. before they start getting data, whereas the lower latency 03b satellite link will receive it nearly 4x sooner.

Another problem with high latency links is the introduction of a large amount of inter-packet jitter due to the TCP protocol flow control operating less efficiently over longer delay paths. 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 re-load 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**

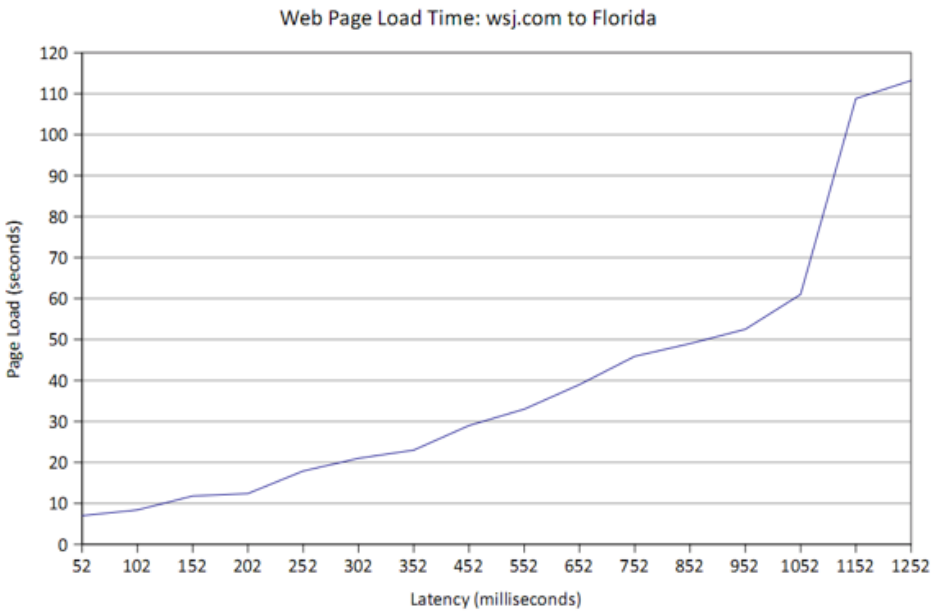
The spreading of network data over time reduces what's called the "Effective Bandwidth" of a link. Packets are still being transported at the same bit rate but due to latency it is taking much more time for all of the web-page packets to arrive.

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):

Web Page Load Time: wsj.com to Florida



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

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.

Applications

What applications can now be run over lower orbit satellites that were possible a few years ago?

Many applications, such as basic web site downloads, Internet video streaming that operated sluggishly (or were very slow to start up) over geosynchronous satellite links now have much faster response and work much better over O3b's low latency connections. Video conferencing and regular voice calls over geosynchronous satellites were very stilted and non-interactive. Conversations were "half-duplex" and one party had to wait for the other party to finish talking before they could start up. You could not break into a conversation and interrupt. Voice calls and video conferencing are much more natural and interactive over O3b's links. Interactive gaming players who were on the end of a geosynchronous satellite link were often killed before they could get a shot off. Now they can be full participants in the action.

Many business applications such as remote database access and interactive transaction processing contain software timers that do not adjust for long delays over geosynchronous satellites. Examples include many banking applications and airline reservation and scheduling applications. Now these applications will work for users that are in remote underserved locations.

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. O3b's network latency and throughput approximate that of fiber-based terrestrial networks and permit users in remote locations that are not served by fiber to run many applications that only work over fiber.

Appendix:

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.

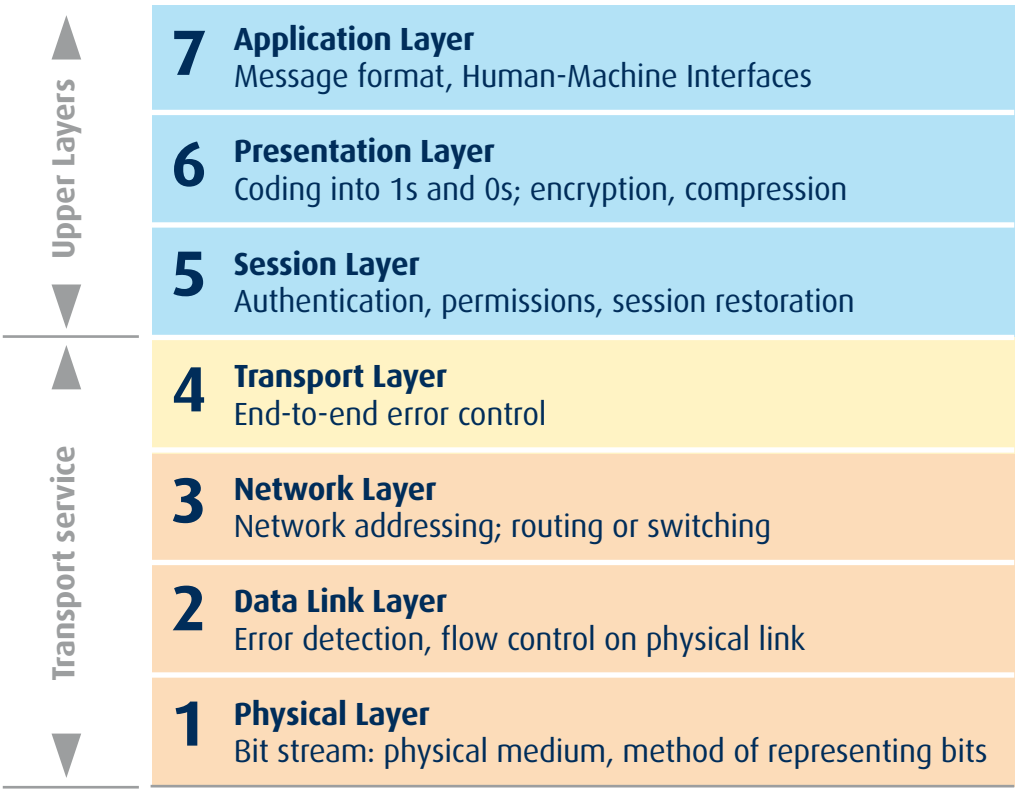
Early networking engineers anticipated the need be able to handle thousands to millions of users on one cohesive network, and thus the TCP/IP networking model was developed.

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:



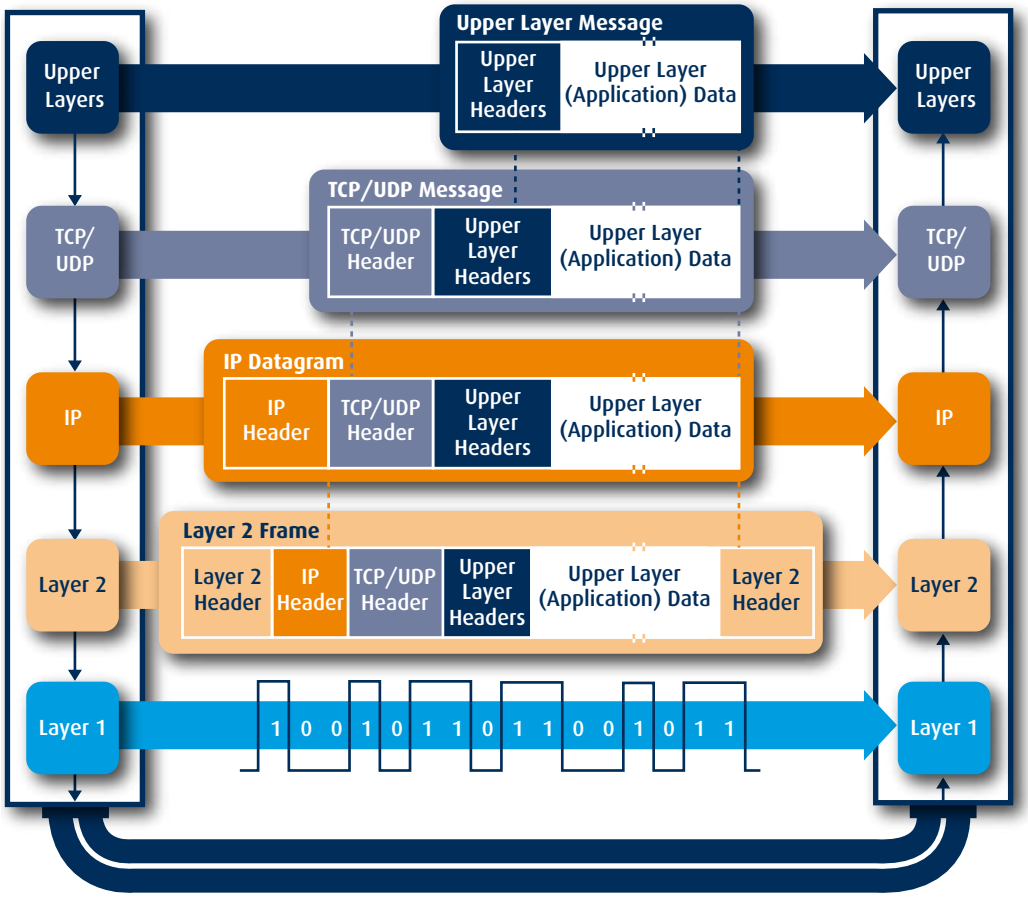
The OSI Model

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.