Redefining Quality of Service for the New World Network: A Briefing on a New Alternative

S.P. Hanks
Enron Communications, Inc.

There is a trend which may be observed by witnessing a significant increase of articles in the trade press, and even in the popular and business press, discussing the coming convergence of all voice, video, and data traffic into transport across a network of common architecture, a network of IP. There is great popular sentiment around the inevitablity of this Convergence (and it is spoken of almost reverentially, where that capital C can be distinctly heard).

Still, from within the traditional world of time-division multiplexed telephony, dissenting voices can be heard. The loudest of these voices demand that attention be paid to the "fact" that "IP is unsuitable because it can't provide guaranteed quality of service."

In this brief, I will outline

- what is meant by "QoS" as it is commonly used, and
- a way in which an IP network -- even an IP network fully connected to the greater Internet -- can in fact deliver TDM grade quality of service without using Internet QoS mechanisms.

Basic Definitions

Before launching into a discussion of quality of service, it will be very useful to ensure that we understand a common reference model. Let us start with the assumption underlying all quality of service discussions: that for a given digital bit path, it is possible to define the physical characteristics for all data transmitted on that path. Provided that these characteristics are consistent with the laws of physics, for a given network it is also possible to specify the maximum permitted deviation from these characteristics over some amount of time.

When discussing the various aspects of quality of service, this briefing will utilize the following terms that describe the parameters of these physically measurable characteristics:

service availability, the percentage of time for which a network service is scheduled to be in operation. For instance, a network with 96% availability potentially might have service outage windows scheduled for an hour per day. That is, outages might be allowed to occur during these scheduled windows, but that they would not necessarily be required to occur and would in fact only occur if some element of network operation required the outage (e.g. a software update for the switches);

service reliability, the percentage of time for which a network service was actually available over some period of time. For instance, if a network could not be used for 4 hours in a month, it would have a reliability of 99.5%;

latency, the interval between the transmission of the first bit of a packet and the receipt of that same first bit of a packet in one direction between two reference points;

jitter, the average over time in variation of latency for all packets between two reference points;

packet loss rate, the percentage of packets discarded during transfer through a network;

restoration time, the amount of time from the occurrence of a service interrupting event until the service has been restored in a manner which satisfies the desired quality of service constraints;

bandwidth, the number of bits per second which are transmitted between two reference points;

throughput, the amount of bandwidth which is effective, that is actual bandwidth discounting overhead and re-transmission;

service level agreement (SLA), a business contract in which a network operator agrees that specific services to a specific client will meet specific physically measured characteristics over certain periods of time, and typically including financial penalties in the event that the service does not so perform.

Quality of Service in the World of TDM and ATM

In a world where all transmission happens across time division multiplexed networks, it is very easy to assure quality of service: bandwidth is permanently allocated between two points, and once allocated for one purpose may not be used for any other purpose until explicitly re-allocated. In a TDM world, it is not possible for bandwidth that is allocated but unused by in one circuit to be used by any other circuit. There is no variation once initial parameters are set; quality of service issues are governed by the laws of physics.

There are standards, set by ANSI and the ITU, which describe the manner in which a correctly functioning TDM circuit will operate at various speeds in the Synchronous Digital Hierarchy (SDH) -- DS-0, DS-1, DS-3, OC-3, OC-12, etc. In theory, any digital circuit over a particular path is exactly equivalent to any other digital circuit over the same path, at least in terms of its physical characteristics.

In such a world, there are only two classes of service: working correctly, and broken. So, one might ask, what exactly is all the fuss about over quality of service?

The whole "quality of service" concept came into being as frame and cell relay networks began coming into deployment in carrier networks. In direct contrast to the older circuit switched time division multiplexed technology, these newer technologies were statistically multiplexed. That is, instead of a "hard" allocation of a fixed quantity of bandwidth for a particular circuit, they used a more flexible approach in which bandwidth was delegated per application in a "soft" manner -- it was available if needed, but was equally available to other applications if it were not.

All quality of service issues arise from two conflicting states:

1. the statistical multiplexing effect of networks not built exclusively from TDM elements means that it is possible to oversubscribe the network, dramatically improving the economics of the network, and

2. the very nature of over-subscription means that it is possible (and to varying degrees, likely) that at some point in time, an application which has been "promised" bandwidth will find that it does not in fact have access to that bandwidth (i.e. the network is congested)

Thus, quality of service is fundamentally an issue about economic trade-offs. On the one hand, cell or packet switched networks make it possible to operate a network with much greater economic efficiency since there is the opportunity to never "waste" bandwidth. On the other hand,

if the network is operated in a totally under-subscribed manner to prevent all congestion and hence alleviate all quality of service issues, the economics may not work out to be nearly so favorable as originally hoped. In an over-subscribed network, the customer may enjoy significant pricing discounts relative to the facilities-based pricing model of TDM networks, but they do so having made the fundamental economic trade-off between price and guarantee of service.

Recapping, if there is no congestion in the network, if every packet always gets delivered in the shortest amount of time, then there are no "quality of service" issues. The issues only arise in the situation where there is congestion or failures of network links or routing elements. And in normal network operation, congestion is only a factor when economic or technological pressures require that lower amounts of bandwidth be used than might otherwise be required.

Quality of Service in the Internet

The Internet brought with it the notion that packets could cheaply and easily be sent from any location to any other location, at least most of the time. It wasn't perfect, however -- there were things that the Internet didn't deliver. Among the missing elements are mechanisms with which to deliver any kind of service guarantees.

The Internet industry is now focusing on two architectures developed by the Internet Engineering Task Force (IETF) - the integrated services architecture (Int-Serv), which features soft states and end-to-end signalling, and the differentiated services architecture (Diff-Serv), which features class flows and code points contained in the IP header's differentiated services field. Each system has its own advantages and disadvantages.

The Int-Serv initiative is more flexible as it permits each applications to negotiate its own quality of service, by including service classes and traffic parameters similar to those in ATM. The key protocol involved is RSVP, designed to allow the sender to request a certain set of traffic-handling characteristics for a given connection. This however places a great deal of reliance on the network to handle end-to-end signaling and to deal with maintaining state and the other messy issues of "call setup."

The Diff-Serv framework handles flow aggregates and minimizes signaling, thus avoiding the complexity of per-flow soft states at each node. Diff-Serv is more scalable than Int-Serv because of this, and will likely be applied most commonly in backbone networks for both service providers and enterprises. However, it is much less flexible than Int-Serv and you may attempt to use the network only to discover that the "flavor" that you want is not available.

IP routers have been capable of supporting multiple service classes for a very long time. So even though the "ideal" QoS system would require new hardware and a lot of new software, it's more realistic to create a solution based on today's available technology with only small changes to the software

Both the Int-Serv and Diff-Serv initiatives have an underlying world view that there absolutely will be congestion in the network. They then attempt to specify various technical measures to be taken to decide which packets will be transmitted in what order with the implicit assumption that some of these packets may be dropped completely. But what happens if there really isn't congestion in the network?

Techniques for "Fixing" QoS

If this briefing were to take a network hardware technology view of the world, it would now be time to focus on issues integral to the routers and switches. This would include a whole suite of new acronyms and buzzwords such as: RED, WRED, CAR, Class-of-Service bits, IP precedence bits,

contract policing, WFQ, CBQ, DRR and the like. While it is necessary to understand what some of these mechanisms do, it is much more important to understand what they do not do for you.

No hardware mechanism can translate a network capability into an SLA, nor translate an SLA into programming or engineering required to enable a network to ensure that an SLA will be met. This is fundamentally a business issue, and must be understood as a business issue. Network design, engineering, and operations must be consistent with the SLAs offered. This includes particularly decisions regarding over-subscription levels, "dynamic headroom" in the network, path restoration mechanisms, etc.

No hardware mechanism can force itself to be uniformly implemented among all hardware vendors, whether switch and router or end-system such as desktop computers and servers.

No hardware mechanism can force itself to be uniformly used in the same way by all carriers. This also is a business issue, not a technical issue. If it is to be possible to have two carriers deliver the same quality of service for a customer who crosses both of their networks, there must first be a business agreement to do this, and only secondarily is it possible to worry about transfer info between carrier technical implementations.

What routers and switches can do is pretty straightforward:

- They can forward packets and apply traffic conditioning at wire speeds.
- They can police contract agreements by rigorously checking input traffic for contract compliance and marking or dropping out-of-compliance packets
- They can gather statistics per interface on congestion and throughput, possibly even by flow or by class of service
- They can assist in policy management by making available interfaces to allow network engineers to have less likelihood of mis-provisioning a service and by automatic consistency checks across entire network paths.

Everything else is largely up to the carrier engineering and operation team.

Quality of Service in the Enron Intelligent Network

With a clear understanding of the technical limitations of specific equipment based technology solutions, Enron Communications Inc. elected to pursue a very different approach is to cure congestion outside the router. In the Enron Intelligent Network, the business decision has been made to avoid congestion with forward scheduling of bandwidth.

If it is possible to know the state of the entire network at some time T, then given that state plus all of the known committed reservations for, it is possible to predict the state exact state of the network at some future time Delta T. This mechanism can be compared to a railway track scheduling algorithm or a pipeline scheduling algorithm.

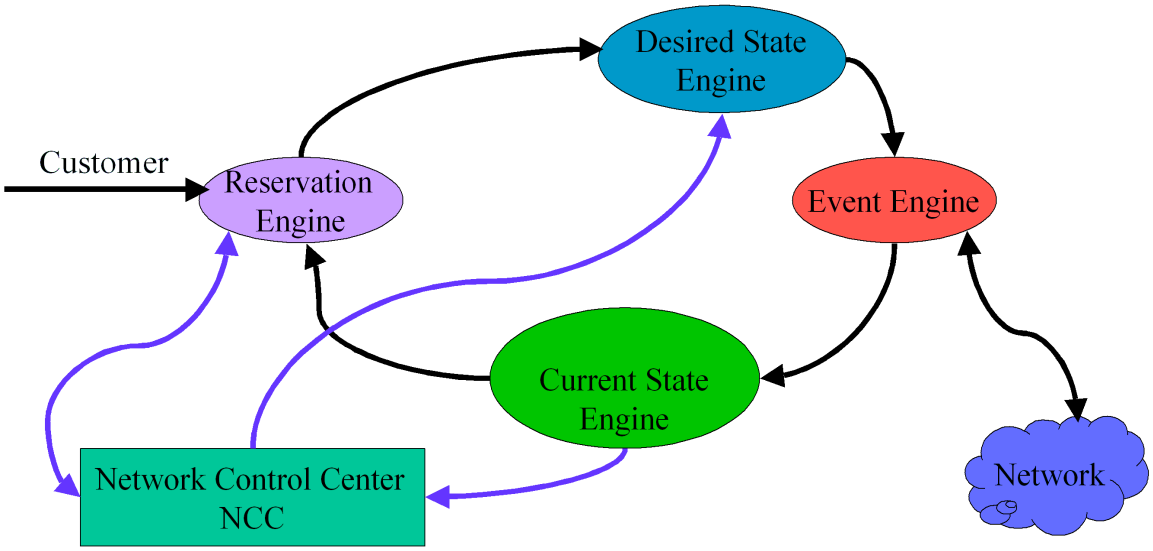
Further, if a history has been kept for all time before T of how the bandwidth has actually been used relative to the amount reserved, it is possible to calculate a likely statistical over-commitment rate that will not create congestion. This makes it possible to "fine tune" the manner in which reservations are accepted based on business case sensitivities as to the financial risk due to any event that would cause an SLA default

It is useful to note that this technique can work across hardware platforms, even across carrier network boundaries -- if the basic information is available to do determine the network state.

This technique is the core of the Enron Intelligent Network.

Initially, these techniques are employed by the Enron Intelligent Network applications. Over time, Enron Communications Inc. will make available the application programming interfaces (APIs) and software development kits (SDKs) for allowing any application to become enabled for the EIN. This will be done in a manner consistent with industry practice for open standards.

What the moving parts actually do



The Enron Intelligent Network control system consists of several components which are indicated in the state diagram above and described in the text below.

Reservation Engine – the Reservation Engine (RE) is the primary point of user or application interaction with the network. It is a software process that accepts requests via a network port to reserve bandwidth through the network. These requests are in the form of control messages.

The key function of the RE component is to deal with policy issues -- it provides the "should we accept this reservation" aspect of the decision to commit or not. It must deal authentication of the entity making the request, the application of policy and access control databases to the requestor, and the interaction with other software agents in the network.

The RE takes the user input and negotiates as the user's agent via messages with the Desired State Engine. This may include re-negotiation with the user of what might be acceptable alternatives.

Desired State Engine – the Desired State Engine (DSE) is a software process that maintains projected future state maps for the network based on current state after applying all approved future requests for bandwidth. It performs admission control based on whether or not capacity is available in the network along the path. This may include capacity for restoration, for duplicate paths, or for other special features.

The DSE receives state information from the Current State Engine and maintains its own database of reservations which it has accepted. It is from this understanding of current state and future reservations pending that the DSE is able to project the state of every link in the network for every level of service.

It is also the case that the DSE might have an understanding of technical capabilities to re-provision network capacity to change the underlying topology of the network. That is, part of determining the desired future state might be contingent on the DSE being able to "make new bandwidth" either by activating new paths, by re-allocating underused paths, or possibly by arranging to utilize capacity in other carrier networks. With the advent of high-capacity optical switching fabrics and the potential to bring new high-speed optical paths into service this is a significant advantage over any previous system.

In addition to reservation requests from the RE, the DSE may engage in negotiation in the style of "no, you can't have 500 Mb/s bandwidth for an hour at 4:00 PM, but you can have 250 Mb/s for two hours then, or you can have the 500 Mb/s at 6:15" with the RE and ultimately the entity placing the reservation. It also must be familiar with the use of multicast if appropriate and with concepts such as "must be completed by a certain time".

Finally, the DSE maintains an archival history of the network state that it predicted, history which can be correlated against whatever the actual state was determined to be at that time.

Current State Engine – the Current State Engine (CSE) is a software process that receives element and path state information updates from the Event Engine. It then interpolates this data into useful formats, and as a result is able to provide a current view into the state of the network. This is similar to the "network state map" displayed graphically by in many network management system products, but it includes other information, such as which data flows traverse which physical links, which SLAs would be impacted by any given network element or path failure, total usage on a path versus total reserved, etc.

The CSE also maintains an archival history of past state that can be correlated against the history archive from the DSE. By comparing the past performance as predicted against the past performance as actually recorded it is possible to tune the heuristic algorithms in the scheduling engines to bring error in network service availability asymptotically towards zero. It might also be possible to utilize adaptive learning techniques to effect such algorithmic truing up on a more real-time basis.

Event Engine – the Event Engine (EE) is a software process that provides the communications path with the actual transmission network elements. As such, it has two "sides". On its network facing side, it will send messages to reconfigure network elements and receive messages from the network elements regarding state of the network, fault conditions, and similar related matters. On the control system facing side, it will receive requests for modifications to the network control state from the DSE, and relay state information to the CSE.

By providing a software abstraction layer distinct from the underlying communications hardware, it is possible to provide easily generalized services, making the control system software more simple than might otherwise be the case. For instance, the control elements of the DSE neither know nor care how the EE is transmitting its requests for dynamic network reconfiguration, just that they are happening. Similarly, the CSE doesn't care if the EE is utilizing SNMP, TMN, or simple command-line interfaces to obtain status information as long as it is conveyed back according to the parameters of the application programming interface.

InterAgent$_{tm}$ -- underlying these software components is the patented InterAgent messaging software. The features of InterAgent allow the processes to communicate using event based messaging and in a distributed cooperating processes model. It handles issues of persistence, security, fault tolerance and overall system stability. Having an open standards model, it makes it possible to communicate with control elements from all major computing platforms.

How this is different from the status quo

Network Management

In the status quo, the roles of network configuration management and of state determination are both performed by engineers in direct contact with network elements using software tools hideously mislabeled as “Network Management Systems”. In a typical NMS, the software actually sends and receives information to and from each element of the network on a periodic basis to determine its state. The asynchronous events that occur in the network are communicated via “traps” which are communicated via the SNMP protocol to a centralized monitoring station.

The problems in this approach are generally well understood, and can be summarized by noting that this approach does not scale well with increased numbers of engineers or network elements, or with increased geographic size or complexity of the network.

The EIN approach provides a software repository which can exist as a distributed process while providing a single coherent view of the network state. Using security and authentication measures, it can also provide different views to different entities, making it possible for network operations to have one view while each customer might have different views.

Bandwidth Reservation

There has been significant work done for a very long time in the area of bandwidth reservation. However, almost all of this work has been done with the notion of *present* reservation as opposed to *future* reservation. Even protocols such as RSVP and other protocols under development inside vendors have no real ability to manage bandwidth allocation forward in time. In fact, all such approaches in wide deployment or public development focus instead on relative priority and the notion that it is perfectly acceptable to drop or delay packets to accommodate the “express priority” packets that must come through.

In a world where some applications absolutely depend on lossless transmission, the ability to accurately forecast availability for such service is an absolute requirement. While there will always be a market for available bandwidth services and best effort services, the true premium services will always require absolute guarantees of bandwidth availability.

Other Aspects

This model is a dramatic shift from all that has gone before in the way of network control and modeling. It provides a clear mechanism by which a carrier organization can project network bandwidth availability on any path, to provision “alternate path routing” with guaranteed network bandwidth, and to have user applications outside of the network control system provision their own bandwidth on demand.

Note as well that this describes the rest of the interaction between the NCC and the network as well: the NCC engineers and technicians view state as presented by the CSE to determine the state of the network. They use the DSE to communicate their desires for changes to the network, with some sort of authentication scheme giving their requests priority over others, but in a manner which preserves a correct future view of bandwidth availability. There is no “engineer telneting to a router and typing ‘conf t’” aspect to the operation of the network. There is no central network element manager doing SNMP “get” commands in a polling fashion. Many of the "operator error" class errors are no longer possible in this environment.

We do not underestimate the magnitude of the work involved in this undertaking, but rather anticipate that we will proceed in co-development with our vendor partners.

.

A Sample Walk Through the Work Flow

Let us look at three different cases:

Customer Using a Web Interface

A customer wishes to reserve bandwidth for a future time. As an example, let's assume that they wish to be able to transfer a file for arrival before 4:30 PM three days hence. They would note the source location, the destination location, the size of the file and that they wish it to be delivered at the specified time.

This request formed via a web interface, and parsed by a program on a web server, then framed and transmitted to the RE via the InterAgent messaging software. The RE then first validates the authenticity of the message, and confirms that the entity placing the reservation is authorized to do so (that they are under their credit limit, that the end points are authorized to accept connections, etc). The RE then engages the DSE by message which attempts to confirm the reservation.

The DSE determines the possible paths between the source and destination locations, and determines the available bandwidth on those paths around the desired time. Since this example has a "deliver by" constraint, it takes the size of the file and the available bandwidth and determines if there is enough bandwidth for long enough to enable delivery to meet the constraint. If the bandwidth is available, it then reserves specific bandwidth, by noting the anticipated usage and times. This act of reservation "subtracts" the bandwidth from the available bandwidth at the indicated time, which prevents over-subscription of the network.

Having communicated with the RE and gained confirmation from the customer, the DSE then drops back into a waiting state. At the appropriate time, the DSE "wakes up" and communicates with the EE to enable the pathways along the appropriate routes at the appropriate designated bandwidth and priorities. The EE marshals messaging into the network elements to accomplish this reconfiguration, and then polls to verify that the reconfiguration has indeed taken place. It then signals the DSE that the reservation has been honored, signals the CSE that there has been a state change, and awaits further instructions.

In the meantime, as the appointed time approaches, the customer application "wakes up" and commences transferring data. That the path previously did not exist, and that it will cease to exist after it is used for this purpose are both invisible to the application. All it "knows" is that a reservation was made, and that the reservation was apparently honored as the capacity that is required has become available.

Also during this time, the EE will have also been polling the network elements and taking asynchronous messages from them, and will be further updating the CSE. On inspection, the CSE will show a data flow enabled between the origin and destination, and that it is carrying a certain load. Unlike the DSE, which models an ideal state, the CSE will show actual usage information, so if a user reserves network bandwidth but does not fully utilize it, this may be noted and used for further refinement of the available bandwidth model.

At the conclusion of the allotted period, the path will be torn down. If no error or exception events which could affected the process were logged, a successful termination will be recorded and also transmitted to the user via the RE or the user agent designated at the time the reservation was made.

If the user attempts to continue transmission after the path has been torn down, the data receives a "destination unreachable" error at the point at which it is attempting to enter the network, and is discarded rather than allowing it to flood the network and possibly impair some other service level agreement. Less strict admission policies would also possibly be available under program control.

Customer Using a “Smart Application”

In this case, let us assume that the customer is connected to an ePowered distribution partner, using an application that contains the ePowered API and InterAgent messaging software. When the customer then invokes the application, the application will determine what bandwidth it requires and connect directly with the RE via InterAgent, and negotiate the required bandwidth. From this contact forward, the process is identical with that in which the user requests reservation via a web interface, with the exception of termination.

If there is a network fault that results in denial of adequate bandwidth, the CSE will have to message back to the application so that it may report the abnormal termination appropriately. If there are no network faults and the application is able to fully complete its use of the network, then an appropriate successful termination must also be messaged back.

Customer Using a Human Interface

In the case that the customer requires a bandwidth reservation to be made but has no computing platform appropriate for the task, assistance may be requested from the Enron Network Control Center (NCC). The customer would call via telephone and request the reservation. The NCC technician in this case would interact via a command console with the RE to place the reservation. The internal flow would be identical to the customer using the web interface, with event reporting going back to the NCC instead of the customer interface.

Conclusion

In the quest for a true quality of service solution, it is possible to get some help from router and switch vendors. However, it is possible to get even more help from the technology underlying the Enron Intelligent Network. By moving to a scheduled bandwidth model, risk of congestion is greatly mitigated while adequate control over the network is maintained to operate at very good rates of efficiency.

However, the ultimate end goal is that users still be able to achieve the goal of acceptable quality of service without having to think too much about the underlying technical issues. The EIN model allows them instead to think in terms of rational business models that recognize that all QoS issues are really issues of financial risk mitigation.

Glossary

ADSL (Asymmetric Digital Subscriber Line) - A variation of Digital Subscriber Line, optimized for asymetric data flow (i.e data flow in which more data flows one way than the other). Ideal for Internet connections where data volumes are much greater from server to client such as Web browsing.

ARPANET (Advanced Research Projects Agency Network) - A Department of Defense wide area network that was first operational in 1969. Tying together systems in universities, government, and business, it was used for networking research and was a central backbone for the development of the Internet.

Asynchronous Communication - The opposite of synchronous or, literally, not synchronous. This is a common method of communication for computers in which information is sent at irregular intervals. Communication is indicated by a start bit followed by a data element and ended with a stop bit. Due to the overhead of start and stop bits, asynchronous communication is slower than other more expensive methods of communication.

ATM (Asynchronous Transfer Mode) - A high-speed transmission technology that can dynamically allocate bandwidth. ATM is a connection-oriented switching and asynchronous multiplexing technique that transports fixed-size packets (called cells). ATM has been selected by the ITU as the basis for the future of broadband networking.

Backbone - A high-speed line between two or more networks.

Bandwidth - The amount of data that can be sent through a given communications medium in a given time interval. Bandwidth is measured in Hertz (analog) or Bits Per Second (digital).

Broadband - A transmission medium that is capable of carrying multiple signals. Broadband achieves this by supporting a wide range of frequencies and dividing the total capacity of the medium into multiple, independent channels, with each channel operating on a specific range of frequencies.

DWDM (Dense Wave Division Multiplexing) - a technique for increasing the capacity of optical fiber by first assigning incoming optical signals to specific frequencies (wavelength) within a designated frequency band and then multiplexing the resulting signals out onto one fiber. Since light waves of different lengths do not interfere with each other, multiple light signals can be sent over the same optical fiber without error, allowing multiple applications to share the same fiber simultaneously.

Ethernet - A local area network that connects computers and devices. Operates over twisted-pair or coaxial cable at speeds up to 10 Mbps. Like so many other things that the computer industry takes for granted. The Ethernet specification came from Xerox's Palo Alto Research Center. Currently Ethernet is the most widely used network access method.

Frame - Generally, a packet of data that contains the header and trailer information required by the physical medium. Usually a frame will also contain control information for addressing and error checking. A frame is a basic logical unit of data transmission.

Frame Relay - A form of packet switching that uses smaller packets and requires less error checking. Frame Relay handles high-speed bursty traffic over wide area networks well.

Hub - The center of the star in a network based on a star topology, or the point where multiple circuits on a network are connected. A hub allows for centralized wiring management and easy troubleshooting of failed network segments.

internet (small "i") - An internet is literally a network of networks which are individually under separate administrative control.

Internet (big "I") - The Internet is the largest network in the world; its roots can be traced back to ARPANET. The TCP/IP protocol suite is central to its operation. Physically, a collection of packet switching networks interconnected by routers along with TCP/IP protocols that allow them to function logically as a single, large, virtual network.

IP (Internet Protocol) - A connectionless protocol that allows a packet to travel across multiple networks on its way to its destination. IP is the network layer of the TCP/IP suite.

ISP (Internet Service Provider) - An organization that provides access to Internet services such as e-mail, World Wide Web browsing, and Internet Relay Chat groups.

ITU - The International Telecommunications Union.

Leased Line - A phone line that is rented for exclusive 24-hour, 7 days a week use from one location to another.

Packet - The unit of data and additional information required for transmitting to the correct network note. Packets are broken into frames for transmission across a medium.

Router - This is a device that interconnects different access methods and protocols. Routers act like bridges forwarding traffic between networks but have greater functionality. They are used to build wide area networks.

T1 - A point-to-point digital communications link that has a capacity of 1.544Mbps made up of 24 64,000bps channels.

T3 - A point-to-point digital communications link that has capacity of 44.736Mbps and is made up of 28 T-1 lines.

TCP (Transmission Control Protocol) - A connection, stream-oriented, end-to-end protocol developed for use on ARPANET. TCP is the most common transport layer protocol used on Ethernet and the Internet. TCP provides the reliable, full duplex, stream service on which many application protocols depend. TCP allows a process on one machine to send a stream of data to a process on another. TCP is connection oriented in the sense that before transmitting data, participants must establish a connection. All data travels in TCP segments, which each travels across the Internet in an IP datagram.

TCP/IP (Transmission Control Protocol/Internet Protocol) - A protocol suite developed by the U.S. Department of Defense to link dissimilar computers across different kinds of networks. TCP/IP is the transport protocol employed by the Internet and is commonly used on Ethernet networks.