Network Capacity Methodology

Network capacity is the volume of individual application transactions (simultaneous) that a network is able to support within pre defined response time service levels. This document will describe many of the key elements that effect network capacity and network performance.

In many ways, network performance is the gauge that quantifies network capacity. Network performance is measured primarily by four factors, network speed (bandwidth), throughput, reliability, and latency. From the user's perspective, network performance is best measured in response time – or the time required for individual application transactions to complete. This is not always an accurate reading on merely network performance as response time can be dramatically affected by back-end system response.

Looking strictly at the network, a capacity methodology must look at network resource availability, application requirements, and historical usage trends and evaluate the overall performance impact of these trends based on the available capacity. It must also provide a baseline for evaluating the impact of network changes (new applications, new technology, network change, etc.).

To predict the capacity of the network you must begin by documenting the following key elements regarding network resources:

- Network architecture – Identify every subnet on the network and how it connects to all the other subnets in the Enterprise. This includes:
    - Topology diagram
    - LAN media
    - User density on each LAN
    - Router locations (logical in the network)
    - WAN design (frame relay, leased line, ISDN, etc.)
    - Circuit speeds
    - Routing protocols
    - Route tables (paths actually used by routers to send/receive data)

- Clients – Information regarding client location, platform, and applications used:
    - Where are the clients located?
    - How many at each location?
    - What client OS is used?
    - How many clients use each application?
    - How often do the clients use each application?
    - Number of clients for simultaneous application access from each location?

- Network Baseline – Historical information regarding network performance on specific network links including:
    - Utilization
    - Latency across specific links
    - End-to-end latency

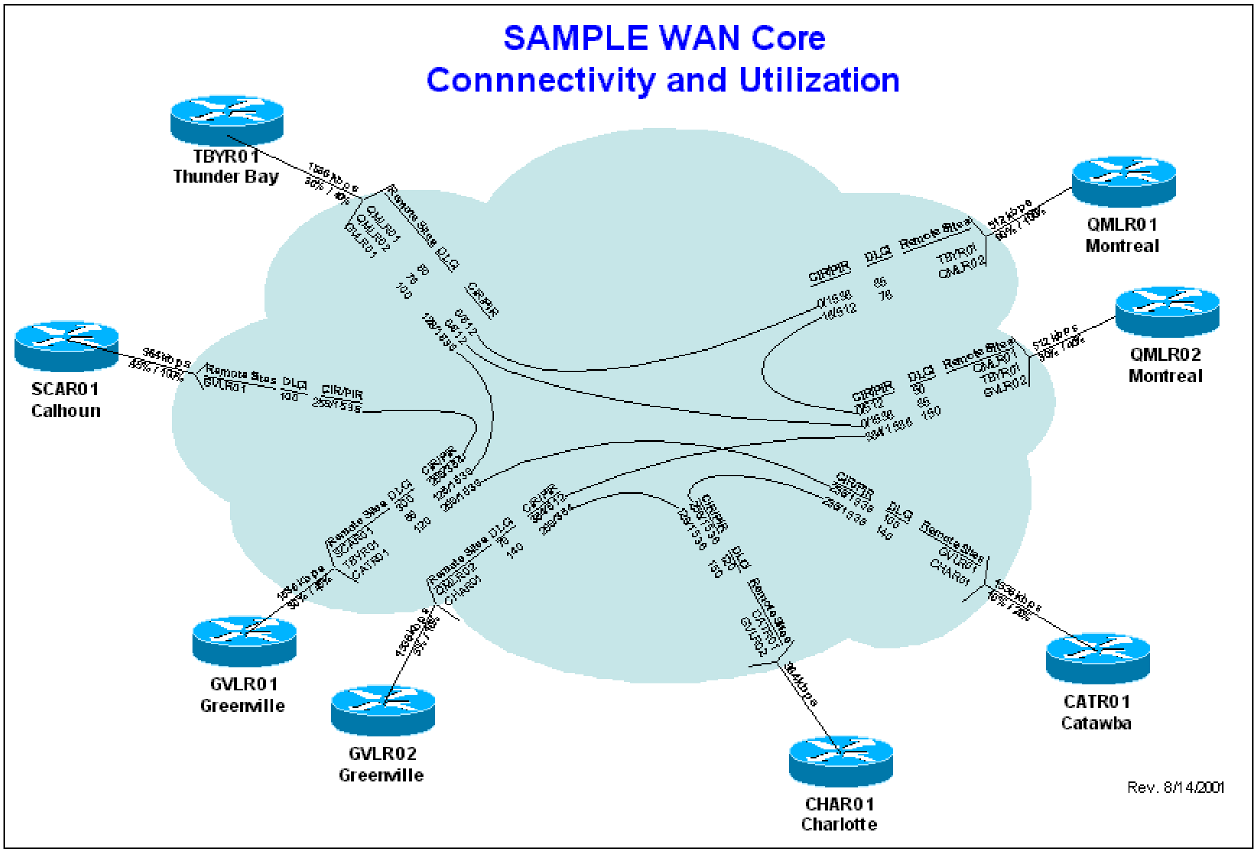
- Application Servers – What type of servers and gateways are used in the network for each application (Windows PDC and BDCs, DNS, Exchange, Notes, midrange systems, proxy, special applications).

- Application network baseline – A functional description of application software, how it is/will be used on the network, and the transmission characteristics of the application. Critical information will include the following:
    - Application server platform
    - Server location(s)
    - Client locations (number on each LAN)
    - Application data transmission detail for discrete application transactions.

- Application response time requirements – predefined requirements for application transactions to complete.

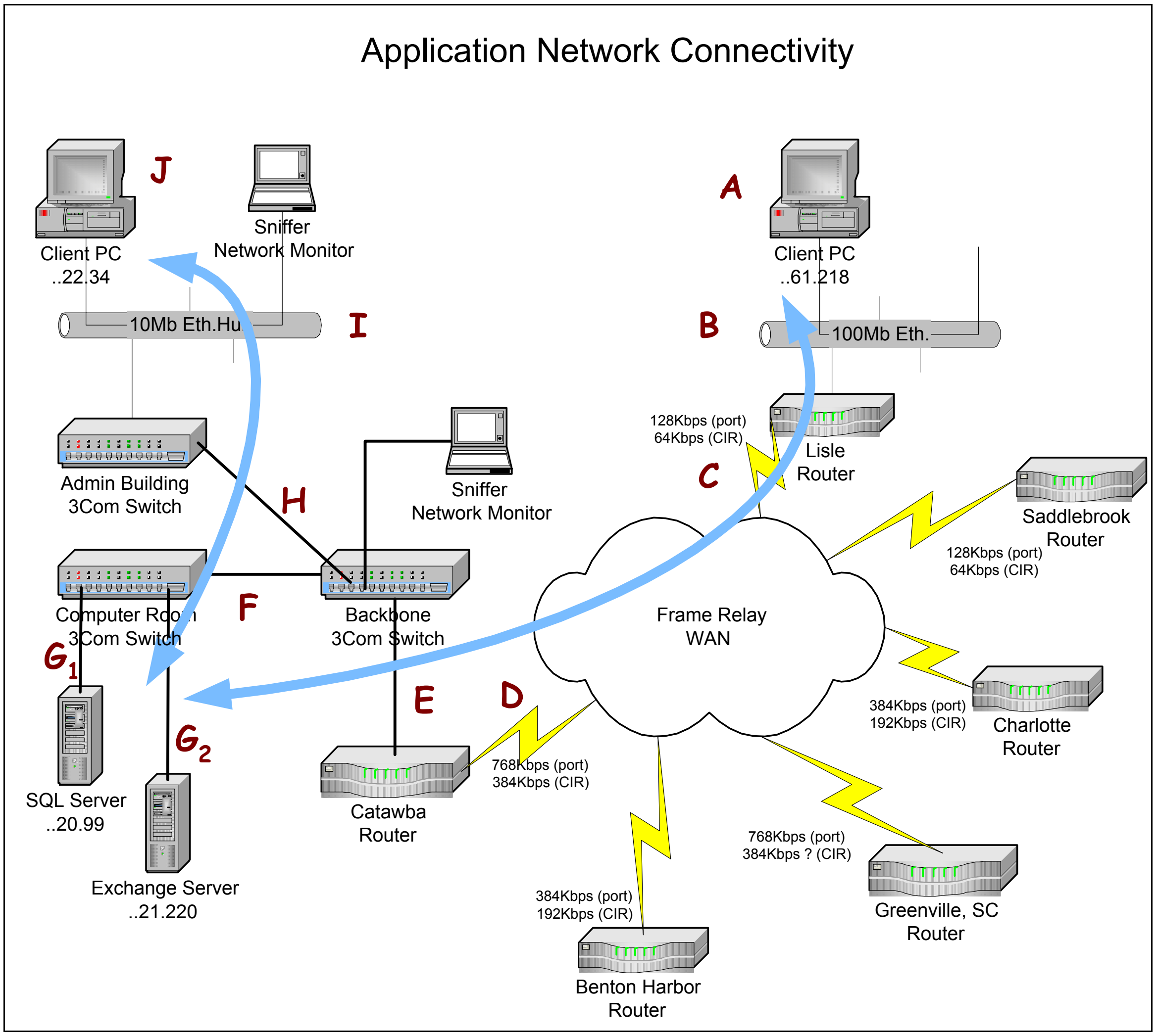
The science of network capacity is the way in which the above criteria are defined, collected, verified, and documented. The art of capacity planning is the way in which an experienced network planner uses this information to predict the network resources required to maintain network performance expectations.

Once the network capacity data described above has been accumulated, the easiest way to begin is with logical network and application diagrams. The following is an example of a simple WAN diagram:



Rev. 8/14/2001

On an application diagram each application server must be placed in its logical position as related to network clients, routers and circuits between the routers. An example of an application diagram for a manufacturing company Sales Application follows:

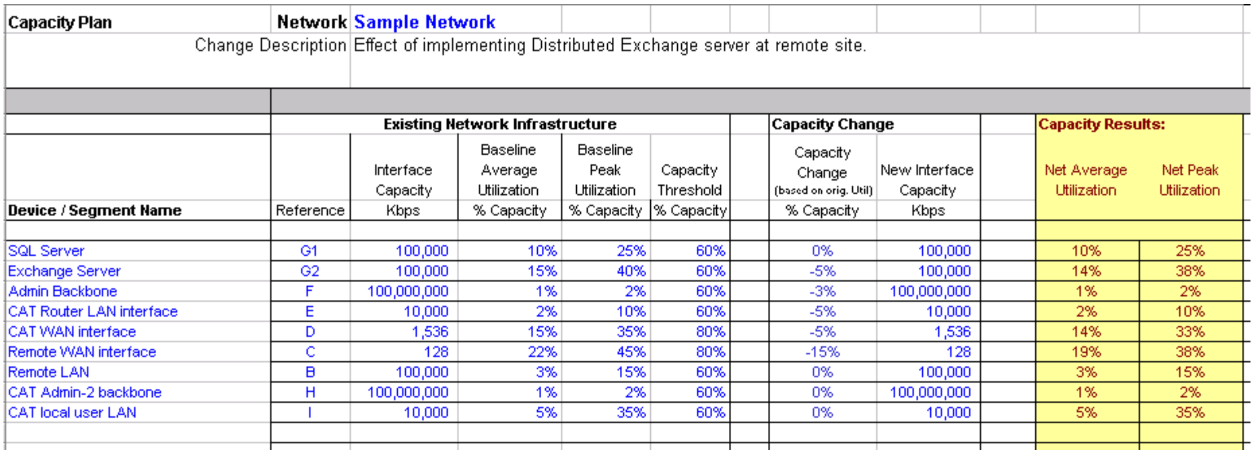


This diagram shows application connectivity from two client locations – one local and one remote. The data transmission routing is visually straight forward, however, should always be verified with a trace route or from the routing tables. The bandwidth of each WAN circuit and LAN subnet has been identified.

The next step in capacity planning is to document the average and sustained peak utilization, latency, and round-trip-time for each link in the application data flow - referring to the above diagram, segments B, C, D, F, G, and H.

Many commercial tools are available to model the network and evaluate capacity – both current capacity used as well as the capacity and performance impact of network changes and new applications. These tools require detailed and precise network topology background and utilization information. Network changes can then be modeled and reports generated to indicate the impact of the changes.

Short of using a commercial network modeling product, the most simple and straightforward method to document capacity change is to build a table (spreadsheet or database) where each connection point to be evaluated is identified along with the current capacity and the anticipated capacity changes.   The following table represents a simple version of this type of table:



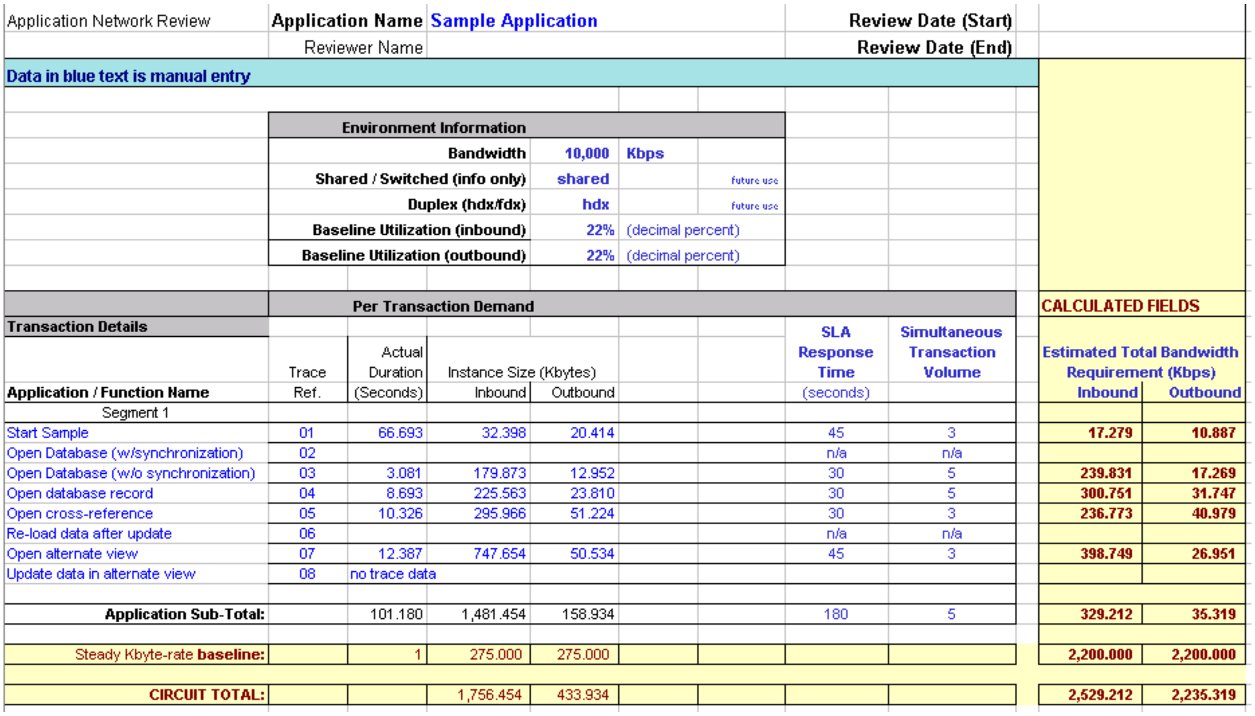
Capacity Plan	**Network**	**Sample Network**									
	Change Description	Effect of implementing Distributed Exchange server at remote site.									
	Existing Network Infrastructure						**Capacity Change**			**Capacity Results:**	
Device / Segment Name	Reference	Interface Capacity Kbps	Baseline Average Utilization % Capacity	Baseline Peak Utilization % Capacity	Capacity Threshold % Capacity		Capacity Change (based on orig. Util) % Capacity	New Interface Capacity Kbps		Net Average Utilization	Net Peak Utilization
SQL Server	G1	100,000	10%	25%	60%		0%	100,000		10%	25%
Exchange Server	G2	100,000	15%	40%	60%		-5%	100,000		14%	38%
Admin Backbone	F	100,000,000	1%	2%	60%		-3%	100,000,000		1%	2%
CAT Router LAN interface	E	10,000	2%	10%	60%		-5%	10,000		2%	10%
CAT WAN interface	D	1,536	15%	35%	80%		-5%	1,536		14%	33%
Remote WAN interface	C	128	22%	45%	80%		-15%	128		19%	38%
Remote LAN	B	100,000	3%	15%	60%		0%	100,000		3%	15%
CAT Admin-2 backbone	H	100,000,000	1%	2%	60%		0%	100,000,000		1%	2%
CAT local user LAN	I	10,000	5%	35%	60%		0%	10,000		5%	35%

The source for the existing network infrastructure data must be based on measured network utilization. Once the current network environment and capacity is documented. Changes in the infrastructure, resources, or applications can be overlaid on this information. Capacity change figures can be derived from a mixture of real and estimated values based on a number of factors including:

- Vendor information
- Actual circuit speed changes
- Lab testing
- Client changes – location and number
- Server changes – location, subnet, distribution

Not all application vendors provide the detail required to evaluate their software against network capacity.   VeriSign offers an Application Network Review (ANR) process in which we evaluate the above factors in conjunction with existing network infrastructure to predict the impact new applications or technology changes will have on various points in the network. A new feature in the ANR is the ability to incorporate network baseline information to predict capacity and response time limits.

The following table is a single segment sample of the ANR worksheet:



Application Network Review	Application Name	Sample Application	Review Date (Start)
	Reviewer Name		Review Date (End)

Data in blue text is manual entry

Environment Information			
Bandwidth	10,000	Kbps	
Shared / Switched (info only)	shared		future use
Duplex (hdx/fdx)	hdx		future use
Baseline Utilization (inbound)	22%	(decimal percent)	
Baseline Utilization (outbound)	22%	(decimal percent)	

Transaction Details	Per Transaction Demand						CALCULATED FIELDS	
Application / Function Name	Trace Ref.	Actual Duration (Seconds)	Instance Size (Kbytes) Inbound	Outbound	SLA Response Time (seconds)	Simultaneous Transaction Volume	Estimated Total Bandwidth Requirement (Kbps) Inbound	Outbound
Segment 1								
Start Sample	01	66.693	32.398	20.414	45	3	17.279	10.887
Open Database (w/synchronization)	02				n/a	n/a		
Open Database (w/o synchronization)	03	3.081	179.873	12.952	30	5	239.831	17.269
Open database record	04	8.693	225.563	23.810	30	5	300.751	31.747
Open cross-reference	05	10.326	295.966	51.224	30	3	236.773	40.979
Re-load data after update	06				n/a	n/a		
Open alternate view	07	12.387	747.654	50.534	45	3	398.749	26.951
Update data in alternate view	08	no trace data						
Application Sub-Total:		101.180	1,481.454	158.934	180	5	329.212	35.319
Steady Kbyte-rate **baseline:**		1	275.000	275.000			2,200.000	2,200.000
CIRCUIT TOTAL:			1,756.454	433.934			2,529.212	2,235.319

In the end, a network capacity methodology requires an ongoing process incorporating network monitoring, application distribution and usage management, network change management, application performance monitoring, and service level verification. With this information, network planners can monitor the performance of the existing network, evaluate the impact of network change, and make appropriate adjustments based on performance trends.

Network Analysis Glossary

The following glossary provides a description of network performance terms used in deliverables along with a description of the various graphs and charts, and “best practices” used in the analysis.

Application Response

This report provides a breakdown of the most used IP protocols (SNMP, HTTP, DNS, etc.) typically based on volume (total transactions) and average response time (latency). The Applications report is useful for spotting trends, usage statistics and spotting unusual network delay issues.

The measured application network response time is the roundtrip elapsed time between a packet leaving a device, reaching the destination device where it is processed and returning to the source device. Response times will vary based upon bandwidth availability, application requirements, routing protocols, topology, destination device response/performance, etc.

DLCI (CIR) Utilization

Frame Relay circuit utilization broken out by Data-Link Connection Identifiers (DLCIs). A DLCI Utilization report is useful for locating overused Frame Relay connections. It displays the percentage of the Committed Information Rate (CIR) used or the amount of time the CIR was exceeded (on the y-axis) for each DLCI (on the x-axis). Each PVC on a frame relay network should be able to sustain 100% CIR utilization without affecting throughput or response time. As a rule, the frame relay network should be designed to support 100% CIR utilization on all PVCs simultaneously. This is not to the recommended starting point for customers, however, it is the starting “commitment” from the carrier. 100% CIR utilization means that the customer is maximizing the use of their investment. Beyond this, burst capacity should be designed into each physical port to accommodate occasional traffic spikes. Sometimes it is necessary to oversubscribe a given circuit. This can happen in many ways. Basically this is when the total inbound or outbound CIR on all PVCs exceeds the capacity of the physical port. While this is not recommended, it can be a useful design option depending on the way the carrier bills for the circuit, with low overall utilization and bursty traffic. The balance between the number of PVCs on a single circuit, the total CIR, network over subscription, burst capacity and performance must be constantly monitored to ensure quality of service.

The graphs used in the WAN Performance Analysis are a mixture of DLCI utilization based on CIR and DLCI utilization based on total port speed. The preference is to report DLCI utilization based on percent of CIR. This provides a simple way to gauge how well the CIR value is provisioned.

When viewing utilization graphs, an apparent “ceiling” may be observed. Regardless of the utilization level at which this ceiling appears, it can indicate a bottleneck somewhere in the network. If this ceiling is at a low utilization level, it probably indicates a bottleneck in the carrier network or at the destination end of a lower bandwidth frame relay PVC. In a Frame Relay network, these ceiling's are usually accompanied by increased congestion reporting.

DLCI Congestion

The DLCI congestion report identifies Backwards Explicit Congestion Notifications (BECN) and Forward Explicit Congestion Notifications (FECN) that occurred during period of analysis. FECNs and BECNs are indicated by a bit in the frame relay header that indicates that congestion may be present in the network for traffic traveling in the direction opposite to the direction of the frame in which the bit is set.

While congestion notification alone does not cause or indicate retransmissions or data loss, persistent congestion does indicate constraints in the network to support the volume of traffic being transmitted. Persistent congestion can also cause data to be buffered in the routers and in the carrier network thus adding to transmission delay (latency).

IP Protocols

IP (Internet Protocol) is a network protocol used to uniquely identify hosts and transport data across an internetwork. Higher layer protocols are used to exchange data between IP and applications. Using the OSI model, higher layer protocols include TCP, UDP, NetBIOS, SMTP, SNMP, etc. At the top of the OSI model, IP applications are used in conjunction with other applications or stand-alone to exchange data between devices. The following table lists many of the most common IP based protocols.

DHCP
DNS
FTP
HTTP
HTTPS
ICMP
NetBIOS
NTP
Ping
POP3
SMTP
SNMP
TCP
Telnet
TFTP
UDP

Line Utilization

Line Utilization is the amount of used throughput capacity on a given network medium. On a WAN circuit, sustained utilization of more than 40 percent of the theoretical limit is not recommended. Greater than 60 percent utilization can result in a rapidly increasing number of dropped packets, retransmissions, and slow response times. Utilization above 80% is possible for short periods of time, however, due to windowing and acknowledgements, the TCP protocol is unable to sustain utilization in excess of 80% for more than a few minutes at a time. UDP is capable of sustaining utilization over 80%, however, this only further reduces TCP throughput.

Depending on the sampling period, sustained utilization over 40 percent usually indicates spikes at double that rate. The shorter the sampling period the more accurate the results. A period of 5 minutes or less typically relates true utilization spikes. Periods of 15 minutes and longer tend to reflect trends averaged through the period and hide traffic spikes.

When viewing utilization graphs, an apparent "ceiling" may be observed. Regardless of the utilization level at which this ceiling appears, it can indicate a bottleneck somewhere in the network. If this ceiling is at a low utilization level, it probably indicates a bottleneck in the carrier network or at the destination end of a lower bandwidth frame relay PVC. In a Frame Relay network, these ceiling's are usually accompanied by increased congestion reporting.

Network Performance

Network performance is primarily measured by four factors, network speed (bandwidth), throughput, reliability, and latency. From the user's perspective, network performance is best measured in the time required for a particular application task to complete. Strictly from the network, performance is best measured in throughput and latency.

The following list identifies many of the terms and factors that contribute to network performance.

- **Client Performance** and resource availability (CPU, disk, memory, etc.).
- **Server Performance** and resource availability (CPU, disk, memory).
- **Application Protocol** – Beyond TCP, UPD, IPX, DecNet, etc., higher layer protocols are used to transfer data to/from applications. Examples of these are SNMP, FTP, SMTP, NetBIOS, etc. Some protocols require more client overhead than others. Some applications communicate directly through core protocols (e.g. browser software accesses HTTP directly) while other applications "wrap" application data with multiple protocols (e.g. Microsoft - NetBIOS wrapped with TCP/IP) for network transmission.
- **Application performance** - how well an application is written, data interface, resource usage, etc.
- Network **Transmission Protocol** – TCP, UDP, SPX, SNA, NetBIOS, etc. The transmission protocol can make a significant difference in the performance of network transactions. For example, UDP can sustain higher network throughput but leaves transmission reliability up to the application, thus potentially slowing the application or overall response for large data retransmissions (transmissions more than 2x TCP window size). By contrast, TCP (Transmission Control Protocol) supports out of sequence reassembly, windowing, transmission acknowledgements and session control, however, TCP throughput is limited by the windowing technique used.
- **Protocol stack** – host settings for buffers, simultaneous sessions, number of active connections, etc. all affect overall performance.
- **Network topology** –Performance factors include connectivity between devices, number or routers, routing policies, route tables, route definitions, switch connectivity.
- **Bandwidth** – bit-rate capacity to transmit/receive data at a given location on the network. Note LAN media can be half or full duplex, WAN circuits are full duplex.

- **Throughput** – the actual (measured) end-to-end connection bit-rate capacity for a given connection across the network. Typically includes time to set up connection, data transmission, acknowledgements, protocol and application overhead, end-station latency, etc.
- **Reliability** – A measure of network availability and quality of data transmission.
- **Propagation Delay** – The time required to transmit (or pass a frame) through a distinct part of the network (first bit to last bit). In other words, how long it takes a given transmission point (router, frame relay switch, etc) to transmit a frame into the network.
- **Latency** – The end-to-end transmission time of a single frame. Can be measured first-out to first-in or first-out to last-in. Sometimes reported in round-trip-time.
- **Round Trip Time** – The total time for a frame to be transmitted between two devices and a response received by the sender. Depending on the protocol used, round-trip-time may also include time for the destination to process and "turn-around" or acknowledge the frame.
- **Congestion** – Network Congestion can cause forced slow-downs on the network. On frame relay, FECNs and BECNs are used to request that the transmitting network device (i.e. router) slow down. On shared Ethernet (hubs and coax), high utilization and congestion will cause data transmission delays.
- **Retransmission** – Certain protocols (like TCP) use acknowledgements to confirm the successful delivery of data packets. Retransmissions are duplicate data sent across the network when the original data was not acknowledged within a predefined length of time. If the time-out value for these acknowledgements is set too short, retransmissions can occur needlessly compounding network utilization and congestion.
- **Broadcasts** – broadcasts are frames of data sent to a special address (ex. 255.255.255.255 for an IP broadcast) recognized by all devices on the network (or subnet). Broadcasts are typically forwarded to all devices on the subnet of origin. Routers typically do not forward broadcasts to other subnets. All devices receiving a broadcast frame must accept that frame and evaluate if the data in the broadcast is for an application or process running on that device. Excessive broadcasts can degrade network, host and application performance by consuming valuable bandwidth, CPU processing time and protocol stack resources.

Top Conversations

The Top Conversations report lets you view the most "talkative" host pairs on the network. Alternatively, a set of devices and conversation statistics between those devices may also be reported. This report displays network traffic on the y-axis, and the hosts involved in the network conversations on the x-axis.

Top Hosts

The Top Hosts report shows the distribution of traffic to and from hosts on the network. The report displays each of the top hosts on the x-axis and data measurement in total bytes on the y-axis. This report is useful for examining the busiest hosts on the network.

Top IP Protocols

The Top IP Protocols report provides a break-down of the most frequently used high-level IP application protocols (SNMP, HTTP, FTP, ICMP, etc.). The report displays each of the top protocols on the x-axis and data measurement in bytes on the y-axis. The Top Protocols report is useful for spotting trends, usage statistics and identifying unusual network traffic.

In the Top IP Protocols report, unidentified protocols appear labeled as “Other”. These consist of IP traffic that do not use standard – well-known ports or is not identified by the monitoring software with an upper layer port. With further analysis, identifying source/destination addresses or looking at payload data may deduce these protocols.

Top MAC Protocols

The Top MAC Protocols report provides a break-down of the most frequently used high-level application protocols based on the MAC packet type. This report is useful for spotting trends, usage statistics and identifying unusual network traffic and protocols.