NetQoS
Performance Experts

WHITEPAPER

Network Diagnostics Provide Requisite Visibility for Managing Network Performance:

The result of insufficient visibility can have unexpected and costly consequences. This paper seeks to broadly identify some of the key diagnostic information that's necessary to provide proper visibility into network performance.

Network Diagnostics Provide Requisite Visibility for Managing Network Performance

Visibility. The concept is so important that it stands alone as a sentence in any document, text, or presentation on network performance.

Without visibility, the last words captured on the black box flight recorder are likely to be "What's a mountain goat doing way up here in this cloud bank?[1]"

Imagine attempting to manage network performance across an enterprise without knowing who is using the network, when they are using the network, and without knowing if the routers and switches are running at their limits, or are failing intermittently.

The results of lacking necessary and sufficient visibility can have unexpected, if not tragic, consequences. This paper seeks to broadly identify some of the key diagnostic information that's necessary to provide proper visibility into network performance.

The main areas of visibility required to effectively manage a network for performance may be divided into three major categories:

End-to-end Response Time: On a per-application basis for business critical applications–the ability to understand the performance the user is experiencing, as well as to identify and isolate the source of a problem to the network, server, or application

Traffic Flow Data: The traffic utilization of network links (for WAN, MAN, and LAN) from SNMP and flow analysis from NetFlow or IPFIX data

Device Performance Information: The status and utilization of network devices themselves (especially routers, switches, and firewalls) via SNMP

In this article, we look at ways to improve visibility in these three categories to provide necessary and sufficient visibility into the performance your end users are experiencing as well as the performance of your network. It will also provide relevant diagnostic information to help you identify, manage, verify, and solve root cause issues within your network.

[1] Apologies to Gary Larson.

End-to-End Response Time

While availability can be visualized as Red or Green, performance is the million shades of grey in between. How do you define performance? Is it a 500 milliseconds response time or a five milliseconds response time? The answer is its all relative and that's why it is so critical to have visibility into historical performance to enable you to identify when performance strays significantly from normal bounds.

Once a performance problem is identified, the next hurdle to overcome is lack of visibility into the root cause of the problem or even where the problem occurred—within the application itself, the server that's hosting the application, or the network. That's why it often takes up to 20 highly paid IT professionals many hours or even days to diagnose performance problems, much less repair them.

To determine if an application performance problem is caused by the server or the way the application was written, metrics such as the following should be monitored and trended:

» Server processing and response times (is the server slow to respond to requests owing to hardware issues or resource constraints?)
» Server CPU, memory, and I/O utilization
» Counts of the number of application round trips required to service data requests (how many packets were used to send the data?)
» Data transfer time (how long did it take to fulfill the request?)
» Data response sizes (how much data was requested for each request and how much was sent in response to a request?)
» Number of concurrent connections (too many connections may overtax server resources)
» Number of refused connections (indicates likely problems with the server)
» Server throughput
» Traffic volume to/from the server (server may be overtaxed)

Key server subsystems should also be monitored to determine if they are overtaxed by tasks at hand. For example, CPU utilization remains over 80% constantly when averaged over 5 minute intervals may indicate insufficient processing power. A lack of memory availability may indicate the server lacks sufficient memory to handle all processes, or may suggest that an application process is not managing memory correctly or is leaking memory.

Traffic Flow Data

To determine if an application performance problem is caused by the network, engineers and managers need to know three things about the performance of critical links:

» Latency–how much time does it take for traffic to pass down the link?
» Utilization and protocol information–what is using network links, when, and for how long?
» Packet Drop–how much traffic is lost or delayed because of overflows in the queues?

Latency

Network latency, or how long it takes for information to travel from one location to another, is the basis for almost all network performance issues. Network latency in modern networks has three major constituent variables:

- » Latency due to distance. How long it takes a data packet to travel from Point A to Point B is determined by the laws of physics and the speed of light. Light traveling down a fiber cable travels at about 5.5 microseconds per kilometer. For the 4,800 km distance between New York and Los Angeles, the round trip is about 53ms. (We may call it "enterprise" computing, but we still don't have "warp drive.")

- » Latency due to serialization delay. Networks send serialized data–one bit at a time. The speed at which such data is transmitted is typically called the bandwidth for the circuit. For example, it takes a 56 Kbps link approximately 214ms to send a 1500 byte packet, while it takes a 1536 Kbps (DS-1) link approximately 8ms to send the same packet.

- » Latency due to queue delay. Networks can only send one packet at a time–one bit at a time. Other packets wait their turns for transmission in a network queue. How long packets wait depends upon how many packets are in front of them (the queue depth) and how fast the queue can be emptied (the bandwidth for the link). Higher bandwidth and/or lower link utilization reduce queue delay.

Because network latency affects all applications and their users–gathering, baselining, and monitoring changes in latency with respect to the baseline is essential when it comes to managing network performance effectively.

Utilization

WAN links are a valuable, yet expensive, resource within enterprises. To manage WAN links efficiently, utilization data is needed that provides insight as to what applications are using the link, when those applications are active, and who is using those applications. With such information, network "slowdowns" typically attributed to over-utilization can be traced back to actual applications and their users. This provides IT and business managers with the visibility and power to control application flows across WAN connections and helps ensure available bandwidth is used efficiently for business applications.

By only measuring raw throughput without examining who is using the data and why may cause some simple and inexpensive remedies to network problems to be missed. For example, engineers and managers armed with this information can see if the root cause for a performance problem between two remote offices is FTP traffic between workstations in each of the offices. If so, further investigation may show that two users are transferring media files between each other's desktops. This can be resolved easily by reminding users of the organization's network use policies, and an expensive WAN upgrade avoided.

Packet Drop

Packet drop typically occurs once network queues fill up. When a network queue becomes full, there is no longer any memory available to hold additional inbound or outbound packets. Thus, any additional packets will be discarded by the router or switch.

Packet drop most notably affects the quality of Voice over IP (VoIP) calls and video conferences because of the loss of key audio and video information items, and reduces the throughput for critical business applications as TCP throttles back in response to dropped packets. During periods of significant packet drop, voice can become inaudible, video unwatchable, and application throughput can be reduced to a tiny trickle compared to its normal rates. Because packet drop can have such a drastic effect on end user experience–it should be monitored and trended throughout the network for each interface, and when Layer 3 QoS is implemented–by queue for each interface on which QoS is implemented.

Device Performance Information

Devices can be overworked due to traffic utilization, device configuration, and/or hardware and software errors. When devices become overworked, they can drop critical application traffic or stop responding and forwarding such traffic altogether. The following entities should be monitored to determine the overall health of a network device in terms of performance capabilities:

» **CPU utilization**
» **Memory utilization (main memory, buffers, interface)**
» **Backplane and interface utilization**
» **Device errors**

CPU

CPU utilization should be monitored and trended to ensure it remains within accepted bounds for optimal performance while providing enough room to handle atypical events that may occur within the network (such as an outbreak of a virus, or a major change in routing or switching tables). For best practices, CPU utilization should not exceed the following values when averaged over 5 minute intervals:

» Core routers: 50% utilization
» Distribution routers: 60% utilization
» Access routers: 70% utilization
» Core Ethernet switches: 40% utilization
» Distribution Ethernet switches: 55% utilization
» Access Ethernet switches: 70% utilization
» Firewalls: 40% utilization
» VPN concentrators: 40% utilization
» VoIP gateways: 50% utilization

Changes in CPU utilization can be due to changes in device throughput or changes in device configuration. For example, STP enhancements may have been enabled on Ethernet switches that consume significant amounts of CPU clock cycles. When CPU utilization reaches critical values–significant amounts of data can be dropped by the network device and it may become unresponsive, creating a fail-over or service outage.

Memory

Memory utilization should also be monitored and trended to ensure that enough memory is available in free memory pools and available for allocation to key processes. For example, increased memory allocation to the buffer pool may be required for a specific process or interface due to the amount of traffic currently being forwarded by the router or switch. If insufficient memory exists, packets may be dropped due to the inability of the router or switch to buffer (place into memory) all the packets being forwarded to the device. Because memory utilization is dependent upon hardware resources and software processes, you should consult with your equipment manufacturer to determine the optimum reference points for your network.

When unexpected decreases in available memory occur, review your device configuration to determine if inappropriate or unauthorized configuration changes have been made. Changes such as the application of ACLs, changes in routing mechanisms, or the installation of a new operating system version may affect the performance and stability of the network device. If no configuration changes have occurred, open up a trouble case with your vendor to determine if a memory leak condition exists for your platform and OS version.

Backplane and Interface Utilization

Over-utilization of interface and backplane resources can lead to packet drop, route flapping, reduction of data throughput, and device instability.

For any given interface, utilization will typically be asymmetrical because client application traffic to the data center tends to be either control traffic or small requests, while server data to the client office tends to be large responses consisting of multiple packets (which can number in the tens, hundreds, or thousands). For this reason, traffic utilization in both directions should be monitored separately so that congestion in either direction may be detected and corrected.

Some router models limit the number of "high speed" interfaces that can be installed in a modular chassis due to backplane/bus speed limitations. Consult your equipment manufacturer to determine what, if any, limit exists for the network devices in your network.

If you have redundant links in your network that implement a load-sharing technology between the two links, you should ensure that utilization statistics for both the primary and secondary links do not exceed 40% utilization when averaged over 5-minute time intervals, and 30% when averaged over 15-minute intervals (without packet loss or an increase in network latency for either utilization point). Such values will typically permit fail-over to a single circuit without significant loss of performance due to excessive queue delay.

Device Errors

Another significant contributor to network performance issues concerns the existence of errors in networking equipment due to hardware/software malfunction or configuration errors. For example, a large number of framing errors (FCS, alignment, runts) on an Ethernet interface may indicate either a problem with the physical cabling or a problem with a duplex mismatch due to an incorrect configuration on the switch port (e.g., one side is set to "auto" while the other is manually set to "full duplex"). Such issues will affect the performance of the host attached to that port. In addition to interface errors, the status of major system components, like the CPU, memory, power supplies, and system processes, should be monitored for errors. For example, changes in STP topology may indicate a problem within the network, and topology changes may cause time sensitive applications to fail.

Conclusion

Information from the sources mentioned above can provide the necessary details to illuminate key areas within your IT infrastructure that affect performance and the end user experience, and show trends over time. Such trends can then be used to proactively upgrade network links, servers, and applications ***before*** a performance problem impacts the end user experience.

By identifying, monitoring, and trending relevant diagnostic information proper visibility can be gained into the network. Engineers and IT managers can then be more proactive in managing performance by identifying problems and performing root cause analysis quickly, and mitigate the risk of future performance issues by planning for anticipated capacity growth and performance needs.

About NetQoS

NetQoS is the fastest growing network performance management products and services provider. NetQoS has enabled hundreds of the world's largest organizations to take a Performance First approach to network management—the new vanguard in ensuring optimal application delivery across the WAN. By focusing on the performance of key applications running over the network and identifying where there is opportunity for improvement, IT organizations can make more informed infrastructure investments and resolve problems that impact the business. Today, NetQoS is the only vendor that can provide global visibility for the world's largest enterprises into all key metrics necessary to take a Performance First management approach. More information is available at **www.netqos.com**.

NetQoS, Inc.
e. info@netqos.com
p. 512.407.9443
t. 877.835.9575
f. 512.407.8629
www.netqos.com

WP rev1 20061024