

Monitoring your Servers and Services...

The Monitoring Servers:

All non-essential services must be turned off, except the monitoring program itself and outgoing mail if used for alerts. If mail is on, it must be secure allowing only connection from another mail server that it is allowed to use it to relay the alert. It should not respond with a Hello message to any machine other than the mail relay machine. Allowing others will subject the machine to discovery by spam bots. These will tax the CPU by creating a marginal mail DoS in an attempt to find was to deliver spam. This will cause false alarms and latency.

Logical ports for all non-essential applications, known and unknown, need to be blocked prior to reaching the server. Another device needs to perform this task. For mail alerts, the logical port block should only allow the relay machine by IP address. If the server's software firewall is expected to also perform this task, it can be over run and tax the CPU. This can cause false alarms and latency.

Monitoring through the net requires that it monitor all its' gateways to monitoring destinations. These need to be compared with the monitoring destination paths to determine if a latency problem is not one of local congestion at that specific gateway. Based on others using the gateways this can occur from time-to-time.

The Monitoring Servers' OS and Capabilities:

Windows is a known risk in monitoring. Of all the operating systems, it has the desire to phone home more than others without your knowledge. Be very sure that all services related to updates are turned off and remain off. Double check these setting after manually updating as Windows PCs are known for automatically changing these and opening ports during updates.

The number of monitor tasks needs to be matched with the server capabilities. The type and number of tasks can easily exceed the machines capability. Task intervals are important, check the CPU idle time as it relates the multiple tasks launching at the same time. The server must have plenty of RAM and CPU power. Memory page swaps with virtual drives can not be allowed at all when monitoring at millisecond levels. This will create false latency or alarms.

TCP/IP drivers themselves repeat code can create false latency for monitoring in various applications. A missed packet in the LAN results in a delay and repeat packet. Usually tasks that have large numbers of packets, like HTML page compares and load timing are effected by this with no indicators that a repeat took place.

When expecting to monitor for latency, its' kernel should be tuned to eliminate drivers and code that are not essential to monitoring. Simply turning off interfaces is not enough since the code is there and it scans motherboard devices for interrupts. Handling a false interrupt can cause false internal latency. Note: Windows can not be easily tuned, therefore it is recommended these should only be used to detect outage monitoring and not for latency monitoring.

When accessed by service personnel for upgrades, configuration changes, task additions, virus testing, disk exploration, etc. the monitor tasks can provide false latency and alarms. It is important to know when these events co-exist, if there was an alarm or large latency issue at the same time.

The Local Network:

It is important to evaluate every application on a network to determine there are no large tasks (i.e. network backups) that can run at full port speeds. This can create local congestion that can cause false alarms, latency, or lost TCP/IP packets. Both on the LAN and on the monitor machine.

QoS settings for packet delivery need to be tuned for monitors to have the highest priority throughout the LAN, through the firewall, and to the Internet gateway.

The Firewall:

While logical ports are blocked, it is also important to understand that other functions performed by the firewall can cause false alarms and latency issues. Firewalls and their services need to be matched to the circuit as well as to the applications. Here it is easy to understand that a firewall with a 100 meg/sec port usually does not pass data through it at the 100 meg/sec wire speed. Like many computers, consumer firewalls have a max throughput speed. The average consumer firewall has a 5 meg/sec throughput and is built this way to be cost effective for home and small office Internet services. Underpowered firewalls that have Stateful Packet Inspection (SPI) activated can slow all packets to a crawl.

The Network's Internet Gateways:

When monitoring applications located across the Internet, the monitoring machine should not be a hosted solution unless you are monitoring from multiple locations. This will enable you to compare multiple results and realize that when one claims high latency or outage and the others do not, the data or alarm are actually false and the problem is related to that specific monitoring location.

When monitoring across the Internet, it is important to understand that packets from all over the world converge with your packets on routers not managed by you or your ISP. This is based on BGP announcements that exchange route paths with each other. A broken path can cause a BGP flap where routers get updated by removing or changing route paths to specific routes of IP address destinations. One router tells another of the change, and the next tells another, etc. This time to update is typically 180 seconds (3 minutes) per router in the path.

You cannot monitor your server using a consumer grade, DSL or Cable, internet service. It is more likely that you are really monitoring your consumer Internet connection than you are your servers at a data center. (Need I explain why? Maybe, you should read all Fiber Internet Center's whitepapers.)

Paths Through the Internet:

Circuit paths can become congested. Although you may have QoS properly deployed in your local LAN and to your Internet gateway, it does not mean your packet has any priority over others when traversing Internet router hops. Therefore, only latency averages compared over a long period should be considered valid data.

The Routers in the Path:

Today, most modern routers have been tuned to help lessen the effects of a Denial of Service (DoS) attacks. They have Access Control Lists (ACLs) in hardware with the interface chips to prevent the need for the routers CPU to make decisions about unwanted packets. DoS attacks that pass the ACLs all take a CPU hit on the router. Even a small DoS attack can cause false latency to be reported. It is false since it has nothing to do with the LAN, Internet, or router configurations because it stops when the attack stopped.

As router attacks evolved, manufacturers began changing the designs to maintain performance during attacks. One of the most common processes abused by DoS was "ping". In most all advanced routers today, ping is performed in software and has the lowest priority. Delivery of non-ping packets through the router receives the highest priority. This means that latency times for ping tests (and trace routes) can vary a great deal. An average over tens of thousands of pings over periods of days is considered valid information when searching for low latency issues. This is why gamer web sites latency (often called lag times) are not compared with routers but to the game server itself. As the packet travels through the router and is not processed by the router. Multiple game

servers are tested at the same time from the gamer Network Operation Center (NOC). When compared against each other they can see if a single game machine reporting high lag times has an issue or if an internet path is the issue if all machines report high lag times.

The Applications:

When testing a web page for delivery, it is important to know everything about that web page. For example, web pages that run scripts to determine which files to deliver are open to a great deal of latency issues not related to the network. This is commonly referred to as Server Side Includes (SSI). A web machine with any SSI pages will need to make decisions which take CPU time. This means you are not just measuring the LAN or Internet performance, but you are also measuring your machines performance as well. This makes it impossible to test reliably for latency. While the HTML page that you are monitoring may not have SSI upon it, it should be remembered that the other pages that are in the servers delivery queue often do, and page requests are addressed in the order they are received. Your non-SSI page must wait for those that came first. You should know that all of the elements of the OS are tuned properly and that there is ample CPU access allocated for all levels of the applications being monitored.

Many applications are tuned to limit the maximum bandwidth allowed on the Ethernet port. Other levels of tuning need to be carefully considered, such as, the bit rate allowed per connection, the maximum number of connections the server can open, etc. These are often overlooked and misunderstood in both Apache and IIS web servers.