Cisco − Performance Management: Best Practices White Pap

Table of Contents

Performance Management: Best Practices White Paper

Introduction

Performance management is the practice of optimizing network service response time. It also entails managing the consistency and quality of individual and overall network services. The most important service is the need to measure the user/application response time. For most users, response time is the critical performance success factor. This variable will shape the perception of network success by both your users and application administrators.

Background Information

Capacity planning is the process of determining the likely future network resource requirements to prevent a performance or availability impact on business−critical applications. In the area of capacity planning, the network baseline (CPU, memory, buffers, in/out octets, etc.) can be shown to affect response time. Therefore, keep in mind that performance problems often correlate with capacity. In networks, this is typically bandwidth and data that has to wait in queues before it can be transmitted through the network. In voice applications, this wait time almost certainly impacts users because factors such as delay and jitter affect the quality of the voice call.

Another major issue complicating performance management is that although high network availability is becoming mission−critical for both large enterprise and service provider networks, the tendency is to seek short−term economic gains at the risk of incurring (often unforeseen) higher costs in the long run. During every budget cycle, network administrators and project implementation personnel face the dilemma of finding a balance between performance and fast implementation. Further, network administrators face challenges that include rapid product development to meet narrow market windows, complex technologies, business consolidation, competing markets, unscheduled downtime, lack of expertise, and often insufficient tools.

In light of these challenges, how does performance fit within the network management framework? The primary function of an ideal network management system is to optimize a network's operational capabilities.

Once you accept this as the ultimate goal for network management, then the focus of network management becomes keeping the network operating at peak performance. An ideal network management system has the following principle operations:

- It informs the operator of impending performance deterioration.
- It provides easy alternative routing and workarounds when performance deterioration or failure takes place.
- It provides the tools for pinpointing causes of performance deterioration or failure.
- It serves as the main station for network resiliency and survivability.
- It communicates performance in real time.

Using this definition for an ideal system, performance management becomes essential to network management. The following performance management issues are critical:

- User performance
- Application performance
- Capacity planning
- Proactive fault management

It is important to note that with newer applications like voice and video, performance is the key variable to success and if you can't achieve consistent performance, then the service will be considered of low value and fail. In other cases, users simply suffer from variable performance with intermittent application time−outs that degrade productivity and user satisfaction.

This document details the most critical performance management issues, including critical success factors, key performance indicators, and a high−level process map for performance management. It also discusses the concepts of availability, response time, accuracy, utilization, and capacity planning, including a short discussion on the role of proactive fault analysis within performance management and the ideal network management system.

Critical Success Factors

Critical success factors identify the requirements for implementation best practices. To qualify as a critical success factor, a process or procedure must improve availability or the absence of the procedure must decrease availability. The critical success factor should also be measurable so the organization can determine the extent of their success.

Note: For more detailed information, see Performance Management Indicators.

The critical success factors for performance management are:

- Gather a baseline for both network and application data.
- Perform a what−if analysis on your network and applications.
- Perform exception reporting for capacity issues.
- Determine the network management overhead for all proposed or potential network management services.
- Analyze the capacity information.
- Periodically review capacity information for both network and applications, as well as baseline and exception.
- Have upgrade or tuning procedures set up to handle capacity issues on both a reactive and longer−term basis.

Indicators for Performance Management

Performance indicators provide the mechanism by which an organization can measure critical success factors. Performance indicators for performance planning include:

- Document the network management business objectives. This could be a formal concept of operations for network management or a less formal statement of required features and objectives.
- Create detailed and measurable service level objectives.
- Provide documentation of the service level agreements with charts or graphs showing the success or failure in meeting these agreements over time.
- Collect a list of the variables for the baseline, including things like polling interval, network management overhead incurred, possible trigger thresholds, whether the variable is used as a trigger for a trap, and trending analysis used against each variable.
- Have a periodic meeting that reviews the analysis of the baseline and trends.
- Have a what−if analysis methodology documented. This should include modeling and verification where applicable.
- When thresholds are exceed, develop documentation on the methodology used for increasing network resources. One item to document is the time line required to put in additional WAN bandwidth and a cost table.

Performance Management Process Flow

The following steps provide a high−level process flow for performance management:

1. Developing a Network Management Concept of Operations
    1. Defining the Required Features: Services, Scalability and Availability Objectives
    2. Defining Availability and Network Management Objectives
    3. Defining Performance SLAs and Metrics
    4. Defining SLAs
2. Measuring Performance
    1. Gathering Network Baseline Data
    2. Measuring Availability
    3. Measuring Response Time
    4. Measuring Accuracy
    5. Measuring Utilization
    6. Capacity Planning
3. Performing a Proactive Fault Analysis
    1. Using Thresholds for Proactive Fault Management
    2. Network Management Implementation
    3. Network Operation Metrics

Developing a Network Management Concept of Operations

Before defining the detailed performance and capacity variables for a network, you must look at the overall concept of operation for network management within your organization. Defining this overall concept provides a business foundation upon which you can build precise definitions of the features desired in you network. Failing to develop an operational concept for network management can lead to a lack of goals or goals that are constantly shifting due to customer demands.

You normally produce the network management concept of operations as the first step in the system definition phase of the network management program. Its purpose is to describe the overall desired system characteristics from an operational standpoint. The use of this document is to coordinate the overall business (non−quantitative) goals of network operations, engineering, design, other business units, and the end users. The focus of this document is to form the long range operational planning activities for network management and operation. It will also provide guidance for the development of all subsequent definition documentation, such as service level agreements. This initial set of definitions obviously can not focus too narrowly on managing specific network problems but on those items that emphasize its importance to the overall organization and in relationship to the costs that must be managed as well. Some objectives are:

- Identify those characteristics essential to efficient use of the network infrastructure.
- Identify the services/applications that the network supports.
- Initiate end−to−end service management.
- Initiate performance−based metrics to improve overall service.
- Collect and distribute performance management information.
- Support strategic evaluation of the network with feedback from users.

In other words, the network management concept of operations should focus on the overall organizational goals and your philosophy to meet those goals. The primary ingredients consist of the higher level definitions of the mission, mission objectives, system goals, organizational involvement, and overall operational philosophy.

As a network manager, you are in the position of trying to unify often inconsistent performance expectations of your users. For instance, if the primary requirement for the network is the transfer of large files from one location to another, you will want to focus on high throughput and less on the response times of interactive users. Be careful not to limit your view of performance without considering a variety of issues. For instance, when testing a network, look at the load levels that are being used. The load is often based on very small packets and the throughput on very large packets. Either of these performance tests may produce a very positive picture, but depending on your network traffic load may not present a true picture of performance. Study your network's performance under as many possible workload conditions as possible and the performance documented.

Also, while many network management organizations have effective alarm techniques to notify technicians about a device failure, it is much more difficult to define and implement an assessment process for the end−to−end application performance. Therefore, while the network operations center (NOC) can respond quickly to a downed router or switch, the network conditions that may undermine network performance and affect user perception may easily go unnoticed until that perception becomes negative. However difficult, this second process can provide immense benefit to both the business organization and network management.

Finally, ensure that you are not making unrealistic expectations of your network performance. This usually comes from misunderstanding the details of networking protocols or the applications. Often times poor performance is not the fault of the network, but rather a result of poor application design. The only way to document and measure an application's performance is to have a baseline of the network's performance prior to installing the application.

Defining the Required Features: Services, Scalability and Availability Objectives

Defining the required features and/or services is the first step of performance management, ongoing capacity planning, and network design. This step requires an understanding of applications, basic traffic flows, user and site counts, and required network services. The first use of this information is to determine the criticality of the application to the organizational goals. You can also apply this information to create a knowledge base for use in the logical design in order to understand bandwidth, interface, connectivity, configuration, and

physical device requirements. This initial step enables your network architects to create a model of your network.

Creating solution scalability objectives help network engineers design networks that meet future growth requirements. This helps to ensure that proposed designs don't experience resource constraints during growth or extension of the network. Resource constraints can include:

- Overall traffic
- Volume
- Number of routes
- Number of virtual circuits
- Neighbor counts
- Broadcast domains
- Device throughput
- Media capacity

Network planners should determine the required life of the design, expected extensions or sites required through the life of the design, volume of new users, and expected traffic volume or change. This helps to ensure that the proposed solution will meet growth requirements over the projected life of the design.

When you do not investigate solution scalability, you may be forced to implement major reactive design changes. This can include adding hierarchy to the design, media upgrades, or hardware upgrades. In organizations that rely on fairly precise budget cycles for major hardware purchases, this can be a major inhibitor to overall success. In terms of availability, networks can begin experiencing unexpected resource limitations causing periods of non−availability and reactive measures.

Interoperability and interoperability testing can be critical to the success of new solution deployments. Interoperability can refer to different hardware vendors, or different topologies or solutions that must mesh together during or after a network implementation. Interoperability problems can include hardware signaling up through the protocol stack to routing or transport problems. Interoperability issues can occur before, during, or after migration of a network solution. Interoperability planning should include connectivity between different devices and topology issues that might occur during migrations.

Solution comparison is the practice of comparing different potential designs in relation to other solution requirement practices. This practice helps to ensure that the solution is the best fit for a particular environment and that personal bias doesn't drive the design process. Comparison can include different factors including cost, resiliency, availability, risk, interoperability, manageability, scalability, and performance. All of these can have a major effect on overall network availability once the design is implemented. You can also compare media, hierarchy, redundancy, routing protocols, and similar capabilities. Create a chart with factors on the X−axis and potential solutions on the Y−axis help to summarize solution comparisons. Detailed solution comparison in a lab environment also helps to objectively investigate new solutions and features in relation to the different comparison factors.

As part of the network management concept of operations, it is essential to define the goals for the network and supported services in a way that all users can understand. The activities that follow the development of the operational concept are greatly influenced by the quality of that document.

The standard performance goals are:

- Response time
- Utilization
- Throughput

- Capacity (maximum throughput rate)

While these measurements may be trivial for a simple LAN, they can be very difficult on a switched campus network or a multi−vendor enterprise network. With the use of a well thought out concept of operations plan, each of the performance goals is defined in a measurable way. For instance, the minimum response time for application "x" is 500 Ms or less during peak business hours. This defines the information to identify the variable, the way to measure it, and the period of the day on which the network management application should focus.

Defining Availability and Network Management Objectives

Availability objectives define the level of service or service level requirements for a network service. This helps ensure the solution meets end availability requirements. Define different classes of service for a particular organization and detail network requirements for each class that are appropriate to the availability requirement. Different areas of the network may also require different levels of availability. A higher availability objective may necessitate increased redundancy and support procedures. By defining an availability objective for a particular network service and measuring the availability, a network organization can understand components and service levels required to achieve projected SLAs.

Defining manageability objectives helps ensure that overall network management is not lacking management functionality. Setting manageability objectives requires an understanding of your support process and associated network management tools. Manageability objectives should include an understanding of how new solutions will fit into the existing support and tool model with references to any potential differences or new requirements. This is critical to network availability since the ability to support new solutions is paramount to deployment success and meeting availability targets.

Manageability objectives should uncover all important MIB or network tool information required to support a potential network, training required to support the new network service, staffing models for the new service and any other support requirements. Often times this information is not uncovered prior to deployment and overall availability suffers as a result of the lack of resources assigned to support the new network design.

Defining Performance SLAs and Metrics

Performance SLAs and metrics help define and measure the performance of new network solutions to ensure they meet performance requirements. The performance of the proposed solution might be measured using performance monitoring tools or with a simple ping across the proposed network infrastructure. The performance SLAs should include the average expected volume of traffic, peak volume of traffic, average response time, and maximum response time allowed. This information can then be used later in the solution validation section and ultimately helps determine the required performance and availability of the network.

Defining SLAs

An important aspect of network design is defining the service for users or customers. Enterprises call these service level agreements while service providers refer to it as service level management. Service level management typically includes definitions for problem types and severity and help desk responsibilities, including escalation path and time before escalation at each tier support level, time to start working on the problem, and time to close targets based on priority. Other important factors are what service will be provided in the area of capacity planning, proactive fault management, change management notification, thresholds, upgrade criteria, and hardware replacement.

When organizations don't define service levels up front, it becomes difficult to improve or gain resource requirements identified at a later date. It also becomes difficult to understand what resources to add to help

support the network. In many cases, these resources are only applied after problems are discovered.

Measuring Performance

Performance management is an umbrella term that incorporates the configuration and measurement of distinct performance areas. This section discusses six concepts of performance management:

- Gathering Network Baseline Data
- Availability
- Response time
- Accuracy
- Utilization
- Capacity Planning

Gathering Network Baseline Data

Most corporate intranets have sufficient bandwidth. However, without adequate data, you may not be able to rule out network congestion as a contributor to poor application performance. One of the clues for congestion or errors is if the poor performance is intermittent or time−of−day dependent. An example of this condition would be where performance is adequate late in the evening but very slow in the morning and during peak business hours.

Once you have defined the network management concept of operations and defined the needed implementation data, it is necessary to gather this data over time. This type of collection is the foundation for the network baseline.

Perform a baseline of the existing network prior to a new solution (application or IOS change) deployment and following the deployment to measure expectations set for the new solution. This will initially help determine if the solution meets performance and availability objectives and benchmark capacity. A typical router/switch baseline report would include capacity issues related to CPU, memory, buffer management, link/media utilization, and throughput. There are other types of baseline data that you may also include, depending on the defined objectives in the concept of operations. For instance, an availability baseline would demonstrate increased stability/availability of the network environment. Perform a baseline comparison between old and new environments to verify solution requirements.

Another specialized baseline is the application baseline. It can be valuable in trending application network requirements. This information can be used for billing and/or budgeting purposes in the upgrade cycle. Application baselines can also be important in the area of application availability in relation to preferred services or qualities of service per application. Application baseline information mainly consists of bandwidth used by applications per time period. Some network management applications can also baseline application performance. A breakdown of the traffic type (i.e., Telnet or FTP) may also be important for planning. In some organizations, more critical resource−constrained areas of the network are monitored for top talkers. The network administrators can use this information for budgeting, planning, or tuning the network. Tuning may consist of modifying quality of service or queuing parameters for the network service or application.

Measuring Availability

One of the primary metrics used by network managers is availability. Availability is the measure of time for which a network system or application is available to a user. From a network perspective, availability represents the reliability of the individual components in a network.

Measuring availability, for example, requires coordinating the help desk phone calls with the statistics collected from the managed devices. However, availability tools cannot determine all of the reasons for failure.

Network redundancy is another factor to consider when measuring availability. Loss of redundancy indicates service degradation rather than total network failure. The result may be slower response time and a loss of data due to dropped packets. It is also possible the results will show up in the other areas of performance measurement such as utilization and response time.

Finally, if you are being held responsible for delivering against an SLA, it is usually important to take into account scheduled outages. These outages could be the result of moves, adds, and changes, plant shutdowns, or other events that you may not want reported. This is not only a difficult task, but may also be manual.

Measuring Response Time

Network response time is the time required for traffic to travel between two points. Response times slower than normal, seen through a baseline comparison or exceeding a threshold, may indicate congestion or a network fault.

Response time is the best measure of your customer's network use and can help you gauge your network's effectiveness. No matter what the source of the slow response is, users get frustrated as a result of delayed traffic. In distributed networks, many factors affect the response time, including:

- Network congestion
- Less than desirable route to destination (or no route at all)
- Under−powered network devices
- Network faults such as a broadcast storm
- Noise or CRC errors

In networks employing QoS−related queuing, response time measurement is important to determine if the correct types of traffic are moving through the network as expected. For instance, when implementing voice traffic over IP networks, voice packets must be delivered on time and at a constant rate in order to maintain good voice quality. By generating traffic classified as voice traffic, you can measure the response time of the traffic as it appears to users.

Measuring response time can help resolve the battles between application servers and network managers. Network administrators are often presumed guilty when an application or server appears to be slow. The network administrator then must prove that the network is not the problem. Response time data collection provides an indisputable means for proving or disproving that the network is the source of application troubles.

Whenever possible, you should measure response time as it appears to users. A user perceives response as the time from when they press Enter or click on a button until the screen displays. This elapsed time includes the time required for each network device, the user workstation, and the destination server to process the traffic.

Unfortunately, measuring at this level is nearly impossible due to the number of users and lack of tools. Further, incorporating user and server response time provides little value when determining future network growth or troubleshooting network problems.

You can use the network devices and servers to measure response time. You can also use tools like ICMP to measure transactions, although it does not take into account any delays introduced into a system as the upper layers process it. This approach solves the problem of understanding how your network is performing.

At a simplistic level, you can measure response time by timing the response to pings from the network management station to key points in the network, such as a mainframe interface, end point of a service provider connection, or key user IP addresses. The problem with this method is it does not accurately reflect the user's perception of response time between their machine and the destination machine. It simply collects information and reports response time from the network management station's perspective. This method also masks response time issues on a hop−by−hop basis throughout the network.

An alternative to server−centric polling is distributing the effort closer to the source and destination you wish to simulate for measure. Use distributed network management pollers and implement Cisco IOS Service Assurance Agent (SAA) functionality. Enabling SAA on routers allows you to measure response time between a router and a destination device such as a server or another router. You can also specify a TCP or UDP port, thereby forcing traffic to be forwarded and directed in the same manner as the traffic it is simulating.

With the integration of voice, video, and data on multi−service networks, customers are implementing QoS prioritization in their network. Simple ICMP or UDP measurement will not accurately reflect response time since different applications will receive different priorities. Also, with tag switching, traffic routing may vary based on the application type contained in a specific packet. So an ICMP ping may receive different priorities in how each router handles it and may receive different, less efficient routes.

In this case, the only way to measure response time is to generate traffic that resembles the particular application or technology of interest. This forces the network devices to handle the traffic as they would for the real traffic. You may be able to achieve this level with SAA or through the use of third party application aware probes.

Measuring Accuracy

Accuracy is the measure of interface traffic that does not result in error and can be expressed in terms of a percentage that compares the success rate to total packet rate over a period of time. You must first measure the error rate. For instance, if two out of every 100 packets result in error, the error rate would be 2% and the accuracy rate would be 98%.

With earlier network technologies, especially in the wide area, a certain level of errors was acceptable. However, with high−speed networks and present day WAN services, transmission is considerably more accurate, and error rates are close to zero unless there is an actual problem. Some common causes of interface errors include:

- Out−of−specification wiring
- Electrical interference
- Faulty hardware or software

Use a decreased accuracy rate to trigger a closer investigation. You may discover that a particular interface is exhibiting problems and decide that the errors are acceptable. In this case, you should adjust the accuracy threshold for this interface in order to reflect where the error rate is unacceptable. The unacceptable error rate may have been reported in an earlier baseline.

The following variables are used in accuracy and error rate formulas:

Notation	Explanation
”ifInErrors	The delta (or difference) between two poll cycles of collecting the snmp **ifInErrors**

	object, which represents the count of inbound packets with an error.
”ifInUcastPkts	The delta between two poll cycles of collecting the snmp **ifInUcastPkts** object, which represents the count of inbound unicast packets.
”ifInNUcastPkts	The delta between the two poll cycles of collecting the snmp **ifInNUcastPkts** object, which represents the count of inbound non−unicast packets (multicast and broadcast).

The formula for error rate is usually expressed as a percentage:

Error Rate = (”ifInErrors) *100

−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−

(”ifInUcastPkts + (”ifInNUcastPkts)

Notice that outbound errors are not considered in the error rate and accuracy formulas. That is because a device should never knowingly place packets with errors on the network, and the outbound interface error rates should never increase. Hence, inbound traffic and errors are the only measures of interest for interface errors and accuracy.

The formula for accuracy takes the error rate and subtracts it from 100 (again, in the form of a percentage):

Accuracy = 100 − (”ifInErrors) *100

−−

(”ifInUcastPkts + (”ifInNUcastPkts)

These formulas reflect error and accuracy in terms of MIB II interface (RFC 2233) generic counters. The result is expressed in terms of a percentage that compares errors to total packets seen and sent. The resulting error rate is subtracted from 100, which produces the accuracy rate. An accuracy rate of 100% is perfect.

Since the MIB II variables are stored as counters, you must take two poll cycles and figure the difference between the two (hence the Delta used in the equation).

Measuring Utilization

Utilization measures the use of a particular resource over time. The measure is usually expressed in the form of a percentage in which the usage of a resource is compared with its maximum operational capacity. Through utilization measures, you can identify congestion (or potential congestion) throughout the network. You can also identify under−utilized resources.

Utilization is the principle measure to determine how full the network pipes (links) are. Measuring CPU, interface, queuing, and other system−related capacity measurements allows you to determine the extent to which network system resources are being consumed.

High utilization is not necessarily bad. Low utilization may indicate traffic flows in unexpected places. As lines become over−utilized, the effects can become significant. Over−utilization occurs when there is more traffic queued to pass over an interface than it can handle. Sudden jumps in resource utilization can indicate a fault condition.

As an interface becomes congested, the network device must either store the packet in a queue or discard it. If a router attempts to store a packet in a full queue, the packet will be dropped. Forwarding traffic from a fast interface to a slower interface can result in dropped packets. This is indicated in the formula **Q = u / (1−u)** where u = utilization and Q = average queue depth (assuming random traffic). So high utilization levels on links result in high average queue depths, which is predictable latency if you know the packet size. Some of the network−reporting vendors indicate that you can order up less bandwidth and pay less for your WAN. However, running WAN links at 95% utilization does have latency implications. Furthermore, as networks are migrated to VoIP, the network administrators may need to change their policies and become used to running WAN links at something more like 50% utilization.

When a packet gets dropped, the higher layer protocol may force a re−transmit of the packet. If lots of packets get dropped, excessive re−try traffic may result. This type of reaction can again result in backups on devices further down the line. You should consider setting varying degrees of thresholds.

The primary measure used for network utilization is interface utilization. Use the following formulas, depending on whether the connection you measure is half−duplex or full duplex. Shared LAN connections tend to be half−duplex mainly because contention detection requires that a device listen before transmitting. WAN connections are typically full duplex because the connection is point to point; both devices can transmit and receive at the same time since they know there is only one other device sharing the connection.

Since the MIB II variables are stored as counters, you must take two poll cycles and figure the difference between the two (hence the Delta used in the equation).

Notation	Explanation
”ifInOctets	The delta (or difference) between two poll cycles of collecting the snmp **ifInOctets** object, which represents the count of inbound octets of traffic.
”ifOutOctets	The delta between two poll cycles of collecting the snmp **ifOutOctets** object which represents the count of outbound octets of traffic.
ifSpeed	The speed of the interface as reported in the snmp **ifSpeed** object. Note that **ifSpeed** may not accurately reflect the speed of a WAN interface.

For half duplex media, use the following formula for interface utilization:

(”ifInOctets + ”ifOutOctets) * 8 * 100

−−

(number of seconds in ”) * ifSpeed

For full duplex media, calculating the utilization is trickier. For example, with a full T−1 serial connection, the line speed is 1.544 Mbps. This means that a T−1 interface can both receive and transmit 1.544 Mbps for a combined possible bandwidth of 3.088 Mbps.

When calculating interface bandwidth for full duplex connections, you could use the following formula in which you take the larger of the **in** and **out** values and generate a utilization percentage:

max(”ifInOctets, (”ifOutOctets) * 8 * 100

−−

(number of seconds in ”) * ifSpeed

However, this method hides the utilization of the direction that has the lesser value and provides less accurate results. A more accurate method would be to measure the input utilization and output utilization separately, such as:

Input Utilization = ”ifInOctets *8 * 100

−−

(number of seconds in ”) * ifSpeed

And

Output Utilization = ”ifOutOctets *8 * 100

−−

(number of seconds in ”) * ifSpeed

While these formulas are somewhat simplified, they do not take into consideration any overhead associated with a particular protocol. There are more precise formulas to handle the unique aspects of each protocol. As an example, RFC 1757 contains Ethernet utilization formulas that take into consideration packet overhead. However, the high availability team has found that for most cases, the general formulas presented here can be used reliably across both LAN and WAN interfaces.

Capacity Planning

As stated earlier, capacity planning is the process of determining the likely future network resource requirements to prevent a performance or availability impact on business−critical applications. Refer to the Capacity and Performance white paper for more detailed information on this topic.

Performing a Proactive Fault Analysis

Proactive fault analysis is essential to performance management. The same type of data that is collected for performance management can be used for proactive fault analysis. However, the timing and use of this data is different between proactive fault management and performance management.

Proactive fault management is the way that the ideal network management system can achieve the goals you determined. The relation to performance management is through the baseline and the data variables that you are using. Proactive fault management is the conceptual area that ties together fault, performance, and change management in an ideal, effective network management system by integrating customized events, an event correlation engine, trouble ticketing, and the statistical analysis of the baseline data.

Where polling for performance data is normally accomplished every 10, 15, or even 30 minutes, recognition of a fault condition must be at a much shorter time interval. One method of proactive fault management is through the use of RMON alarms and event groups. You can set thresholds on your devices that are not polled by external devices so the thresholds are much shorter. Another method, which will not be covered in this document, is through the use of a distributed management system that enables polling at a local level with aggregation of data at a manager of managers.

Using Thresholds for Proactive Fault Management

Thresholding is the process of defining points of interest in specific data streams and generating events when thresholds are triggered. Use your network performance data to set those thresholds.

There are several different types of thresholds, some of which are more applicable to certain types of data. Thresholds are only applicable to numeric data so convert any textual data into discrete numeric values. Even if you don't know all of the possible text strings for an object, you can still enumerate the "interesting" strings and assign all other strings to a set value.

There are two classes of thresholds for the two classes of numeric data: *continuous* and *discrete*. Continuous thresholds apply to continuous or time series data such as data stored in SNMP counters or gauges. Discrete thresholds apply to enumerated objects or any discrete numeric data. Boolean objects are enumerated values with two values, true or false. Discrete data can also be called event data because events mark the transition from one value to the next.

Continuous thresholds can trigger events when the time series object crosses the specified value of the threshold. The object's value crosses the line by either rising above the threshold or falling below it. It can also be useful to set separate rising and falling thresholds. This technique, known as a hysteresis mechanism, helps reduce the number of events generated from this class of data. The hysteresis mechanism works to reduce the volume of events generated by thresholds on rapidly varying time−series data. This mechanism can be used with any threshold technique on time−series data.

The volume of events is reduced by an alarm that is generated to track the value of an object. Rising and falling thresholds are assigned to this alarm. The alarm is only triggered when the rising threshold is crossed. Once this threshold is crossed, a rising alarm is not generated again until the falling threshold is crossed. And the same mechanism prevents falling thresholds from being generated until the rising threshold is crossed again. This mechanism can drastically reduce the volume of events without eliminating information required to determine if a fault exists.

Time series data can be represented either as counters, where each new data point is added to the sum of the previous data points, or as a gauge, where the data is represented as a rate over a time interval. There are two different forms of continuous thresholds applicable to each data type: *absolute continuous thresholds* and *relative continuous thresholds*. Use absolute continuous thresholds with gauges and relative continuous thresholds with counters.

In order to determine the threshold values for your network, you will need to take the following steps:

1. Select the objects.
2. Select the devices and interfaces.
3. Determine the threshold values for each object or object/interface type.
4. Determine the severity for the event generated by each threshold.

Determining what thresholds to use on which objects (and for which devices and interfaces) requires a fair amount of work. Fortunately, if you've been collecting a baseline of performance data, you have done a

significant amount of that work already. Also, NSA and the high availability service (HAS) program can make recommendations on setting objects and creating ranges. However, you must tailor these recommendations for your particular network.

As you have collected performance data for the network, the HAS program recommends that you group your interfaces by categories. This simplifies setting thresholds because you will probably need to determine thresholds for the media type of each category rather than for each device and object on that device. For example, you would want to set different thresholds for Ethernet and FDDI networks. It is commonly thought that you can run FDDI networks at closer to 100% utilization than you can a shared Ethernet segment. However, full−duplex Ethernet can be run much closer to 100% utilization because they are not subject to collisions. You'll probably want to set your thresholds for collisions very low for full−duplex links because you should never see a collision.

You can also consider the combination of the interface importance and the category/severity of the threshold type. Use these factors to set the priority of the event and, therefore, the importance of the event and its attention by the network operations staff.

The grouping and categorizing of network devices and interfaces cannot be over−emphasized. The more you are able to group and categorize, the easier you can integrate the threshold events into your network management platform. Use the baseline as the principle resource for this information. Refer to the Capacity and Performance Management white paper for more information.

Network Management Implementation

The organization should have an implemented network management system that is able to detect the defined threshold values and report on the values for specified time periods. Use a RMON network management system that can archive threshold messages in a log file for daily review or a more complete database solution that allows searches for threshold exceptions for a given parameter. The information should be available to the network operations staff and manager on a 7X24 basis. The network management implementation should include the ability to detect software/hardware crashes or tracebacks, interface reliability, CPU, link utilization, queue or buffer misses, broadcast volume, carrier transitions, and interface resets.

Network Operations Metrics

A final area of proactive fault management that overlaps with performance management is network operations metrics. These metrics provides valuable data for fault management process improvement. At a minimum, these metrics should include a breakdown of all problems that occurred during a given period. The breakdown should include information such as:

- Number of problems occurring by call priority
- Minimum, maximum, and average time to close in each priority
- Breakdown of problems by problem type (hardware, software crash, configuration, power, user error)
- Breakdown of time to close for each problem type
- Availability by availability group or SLA
- How often you met or missed SLA requirements

The help desk often has a reporting system with the ability to generate metrics or reports. Another means of gathering this data is the use of an availability monitoring tool. Overall metrics should be made available on a monthly basis. Process improvement based on the discussion should be implemented to improve missed service level agreement requirements or to improve the handling of certain problem types.

Performance Management Indicators

Performance indicators provide the mechanism by which an organization measures critical success factors.

Documenting the Network Management Business Objectives

This document could be a formal concept of operations for network management or a less formal statement of required features and objectives. However, it should definitely assist the network manager measure success.

This document is the organization's network management strategy and should coordinate the overall business (non−quantitative) goals of network operations, engineering, design, other business units, and the end users. This focus enables the organization to form the long range planning activities for network management and operation, which includes the budgeting process. It will also provide guidance for the acquisition of tools and the integration path required to accomplish the network management goals, such as SLAs.

This strategic document can not focus too narrowly on managing specific network problems, but on those items important to the overall organization, including budgetary issues. For example:

- Identify a comprehensive plan with achievable goals.
- Identify each business service/application requiring network support.
- Identify those performance−based metrics needed to measure service.
- Plan the collection and distribution of the performance metric data.
- Identify the support needed for network evaluation and user feedback.
- Have documented, detailed, and measurable service level objectives.

Documenting the Service Level Agreements

In order to properly document the SLAs, you must fully define the service level objective metrics. This documentation should be available to users for evaluation. It provides the feedback loop to ensure that the network management organization continues to measure the variables needed to maintain the service agreement level.

SLAs are "living" documents because the business environment and the network are dynamic by nature. What works today for measuring an SLA may become obsolete tomorrow. Only by instituting a feedback loop from users and acting on that information can network operations maintain the high availability numbers required by the organization.

Creating a List of Variables for the Baseline

This list would include such things as polling interval, network management overhead incurred, possible trigger thresholds, whether the variable is used as a trigger for a trap, and trending analysis used against each variable.

These variables are not limited to the metrics needed for the service level objectives mentioned above. At a minimum, they should include the following sets of variables: router health, switch health, routing information, technology−specific data, utilization, and delay. These variables would be polled periodically and stored in a database. Reports can then be generated against this data. These reports can assist the network management operations and planning staff in a variety of ways:

- Reactive issues can often be solved faster with a historical database.
- Performance reporting and capacity planning require this type of data.

- The service level objectives can be measured against it.

Reviewing the Baseline and Trends Analyses

Network management personnel should conduct meetings to periodically go through specific reports. This provides additional feedback, as well as a proactive approach to potential problems in the network.

These meetings should include both operational and planning personnel. This provides an opportunity for the planners to receive operational analysis of the baseline and trended data. It also puts the operational staff "in the loop" for some of the planning analysis.

Another type of item to include in these meetings is the service level objectives. As objective thresholds are approached, network management personnel can take actions to prevent missing an objective and, in some cases, this data can be used as a partial budgetary justification. The data can show where service level objectives are going to be breached if proper measures aren't taken. Also, because these objectives have been identified by business services and applications, they are easier to justify on a financial basis.

Conduct these reviews every two weeks and hold a more thorough analytical meeting every six to twelve weeks. These meetings allow you to address both short and long term issues.

Documenting a What−if Analysis Methodology

A what−if analysis involves modeling and verification of solutions. Before adding a new solution to the network (either a new application or a change in the Cisco IOS release), you should document some of the alternatives.

The documentation for this analysis includes the major questions, the methodology, data sets, and configuration files. The main point is that the what−if analysis is an experiment that someone else should be able to re−create with the information provided in the document.

Documenting the Methodology used for Increasing Network Performance

This documentation includes additional WAN bandwidth and a cost table for increasing the bandwidth for a particular type of link. This can increase the "organizational" memory by helping the organization realize how much time and money it costs to increase the bandwidth. Formal documentation allows performance and capacity experts to discover how and when to increase performance, as well as the time line and costs for such an endeavor.

Periodically review this documentation, perhaps as a part of the performance review quarterly, to ensure that it remains up to date.

Summary

The only way to achieve the goals of the ideal network management system is to actively integrate the components of performance management into the system. This would include the use of availability and response time metrics tied into a system of notification upon exceeding thresholds. It would have to include the use of a baseline for capacity planning that would have links to a heuristic model for provisioning and exception reporting. It could have a built−in modeling or simulation engine that enables the model to be updated in real time and provide a level of both planning and troubleshooting through software simulations.

While much of this system may seem an impossible ideal that could never be achieved, each of the components is currently available today. Further, the tools to integrate these components also exist in programs like MicroMuse. We should continue to work toward achieving this ideal as it more realistic today than ever.

Related Information

- **Other Related White Papers**
- **Capacity and Performance Management White Paper**
- **Technical Support − Cisco Systems**