Best Practices in Quality of Service

Strategic Technologies
April 12, 2002

3212 LeMone Industrial Blvd. Columbia, MO 65201
Phone: (573) 884-7200 ◆ Fax: (573) 884-6673
http://www.more.net ◆ info@more.net

Contents

Best Practices in QoS

Introduction

This paper discusses best practices in achieving the proper levels of Quality of Service and Traffic Management on most networks. Although some of the practices discussed here might also be used in general network design and operation, this paper should not be considered a network design primer. This paper assumes that the reader has a properly maintained, working computer network.

There are two basic QoS best practices that inform all other attempts to establish good Quality of Service measures. Both practices stem from a process improvement document that can be found at http://peak.lanl.gov:1500/pdfs/qp/pig1.pdf. The most basic best practice is simply using a "systematic incremental improvement methodology," as outlined in the process improvement document cited above. The steps to this type of methodology are: do the planned improvement, check the results, act on the results and then continue the cycle. Network designers are already familiar with this methodology; in network design terms, the cycle is plan, implement, verify and monitor.

- **Plan:** Make a plan outlining the QoS level you wish to support and how you will implement it. This point can be the starting point in the cycle if the current network attributes and QoS goals are known.
- **Implement:** Using the plan you created for the first step, implement the policy, hardware and software pieces required to achieve the proper QoS level.
- **Verify:** Once you've implemented your plan, verify that the implementation was completed and that it is yielding the correct results.
- **Monitor:** Watch the network for differences between the planned and actual QoS. Start at the planning stage again if conditions change. Monitoring the network is the starting point of establishing a QoS implementation if the current attributes of the network are not well known.

The second basic best practice is: It is better to proceed by slow, incremental changes rather than radical shifts. Many networks have been stopped cold and forced into a long, fitful recovery by attempts to do too much, too quickly. Most QoS deployments start with a working network and the need to establish priority treatment for a small subset of applications or users. The most successful strategy has been to limit changes to only what is currently needed, while preserving the correct operation of the network and reserving the ability to make further changes later.

Definition

"Best practice" is an industry term that can have different definitions. For the purpose of this document, "best practices" means a set of generally agreed upon strategic configurations, process or policies that maximize utility and minimize risk. Best practice strategies are derived from industry experts and reflect their experience. A strategy is chosen as a best practice only if it functions well over a wide variety of situations. Thus, a best practice might not be a locally optimal solution because of unique local conditions, recent advances or innovations. Best practices, however, are valuable because they are generally proven to increase the chances of success and decrease overall risk. Several good best practice definitions are available in Appendix A.

With the above definition in mind, QoS best practices refer to a configuration, process or policy that maximizes the assurance that traffic will be handled according to a designated priority. A QoS best practice may conflict with a general industry best practice. If so, network administrators should decide which best practice to follow based on local policies and their own testing and experience.

Background

Engineering a network to support a given quality of service standard is a highly specialized undertaking. There are no generally accepted "best" configurations for QoS. Current decisions depend on who controls the network, how the network is currently engineered, what traffic currently exists on the network and what QoS goals need to be met. Decisions should also take into account expected changes in the network's administration, engineering, traffic and goals.

QoS Attributes

The list of attributes considered to be QoS measures can change rapidly. When e-mail was the primary type of network traffic, networks needed to be able to maintain their stability. Store-and-forward applications like e-mail do not cope well with unstable address names; without stable address names, mail goes undelivered. Other QoS attributes, like packet loss and transmission delay, are of relatively little importance to e-mail applications. When IBM's Systems Network Architecture (SNA) became a priority, the QoS goals shifted. Suddenly, packet loss and transmission delay were key issues. Another example is FTP. Bulk transfer applications like FTP require high bandwidth for quality operation. Currently, the primary application that is driving QoS requirements is interactive voice and video.

The following QoS attributes form a vocabulary that covers most current QoS requirements. Individual networks and applications may need to maintain specific values in each of the attributes or some of the attributes. The next big critical network application may require some attribute that is not listed.

But generally, network planners should consider the attributes below in assuring network quality.

- **Network Availability:** The probability that any communication can occur. Generally given as a percent. Also known as uptime.
- **Available Bandwidth:** The transmission capability. Generally measured in megabits per second (Mbps). Also known as capacity.
- **Packet Loss:** Percent of the transmission that does not arrive correctly.
- **Latency:** The average time between transmission and reception. Generally measured in milliseconds (ms). Also called delay.
- **Jitter:** The variability in latency between parts of the transmission. Jitter can be measured various ways (e.g., the difference between the highest latency and the lowest latency, the standard deviation of the latency, the statistical probability of a given delay variance, etc.), so it is often given a qualitative value. Also called delay variation.
- **Sequence Error:** The probability that a packet will arrive out of sequence and thus must be buffered prior to reassembly. Generally given as a percent of out-of-sequence packets over the total number of packets.

Future QoS attributes are hard to predict. The most likely measures will be second order derivations of the above attributes. For instance, as applications incorporated defenses against jitter, the QoS emphasis may shift to variation in the jitter (or the variability of the variability in latency). Also, as more network applications incorporate multiple types of transmission (such as a separate timing signal associated with voice or video) then packet loss or sequence error variability by the packet size or type may become a relevant QoS attribute.

Plan

The best first step to solving a current QoS problem is to gain an understanding of your current network and your Quality of Service needs. With this fact in mind, you need to map your network before you begin planning your QoS strategy.

Mapping the Network

Map physical and logical connections including bandwidth available (and, if possible, bandwidth used), end devices in use and application flow.

Physical Connections

The information required to map physical connections includes wire types (i.e., coax, CAT 5, single or multimode fiber, wireless, etc.), communication mode (half or full-duplex Ethernet, Token Ring, ATM, frame relay, etc.) and connection devices (hubs, switches, bridges, routers, etc.). Wire type information should include the distance of the wire run and any available information on the quality of the wire. Communication mode information should include the serialization delay (see Serialization Delay, below). Connection device information should also include any known queuing or capacity limitations.

Delay in Physical Connections: Any physical connection introduces latency. Latency is composed of a propagation delay, which varies by the distance of the connection, and a switching delay for each network device encountered. For LAN links, this latency can be ignored due to short distances and high bandwidths. However, low-speed dial-up and WAN connections introduce measurable latency. For earth-bound connections, the latency is mostly a function of the bandwidth. Satellite connections, however, always involve significant latency.[1]

Heavily used networks introduce latency and jitter in the form of queuing delays. Queuing delays stem from the need for switching and routing devices to wait for a limited resource. Well-designed devices virtually eliminate queuing delays within the devices themselves but may still need to hold a packet during congestion on the physical link. This type of delay can be much worse where a shared-media link such as half-duplex Ethernet (or an Ethernet hub) is involved.[2] Network congestion also significantly increases packet loss and decreases available bandwidth.

Serialization Delay: No form of communication is completely multi-tasking. At some level any packet may have to wait for the packet in front of it to be

[1] See http://www.adec.edu/nsf/tcpip-performance.pdf for a discussion of satellite roundtrip delay.
[2] See http://www.cs.georgetown.edu/~torsneyw/cosc390/materials/Lecture_6_23Oct01.ppt for a good discussion on congestion and queuing delay.

transmitted before it can start transmission. This wait is called the Serialization delay. ATM has the lowest serialization delay because priority cells need only wait for 53 bytes to be transmitted ahead of them. Ignore delay involved in interior ATM devices, such as switches, provided that they recognize cell priority. However ATM edge devices may take packets and break them down into cells one at a time. Thus a serialization delay must be considered at ATM edges. Token Ring networks may have significant delays if large packet sizes are used. Serialization delay may be calculated by:

Size of packet/Transmission rate=Serialization delay

For example, a 1,500-byte packet transmitted at 256 Kbps will have a serialization delay of 47 ms (1,500 bytes*8 bits per byte/256 Kbps).

Note that high transmission rates have low delays. A general rule of thumb is that serialization delay is only significant in WAN links that are 768 Kbps or slower and are being used for real-time voice or video.

Loops: Physical loops in the network structure are valuable for fault recovery but complicate the mapping. Network devices compensate for loops by turning off one leg of the loop. Thus, where loops are present, it may be difficult to determine the actual physical path of transmission and any QoS decisions should be based on the longest physical path.

End devices: All end devices introduce latency in filling and forming packets. These types of latency are called packetization delay and pipelining delay. Unfortunately, in sensitive voice and video applications, there is also a significant amount of signal processing and compression algorithm delay from the sending device, introducing even more latency. The receiving device introduces a slight delay to compensate for jitter.[3] In addition, processing multiple interactive video connections into one unified conference introduces delays from the video multi-conference unit (MCU). Thus, one of the most delay sensitive applications, multipoint video classes/conferencing, tends to have the most inherent latency.

Logical Connections

Routing devices connect multiple physical/switched paths together. They allow the forwarding of transmissions between networks and connection redundancy while limiting broadcast traffic to cut down on congestion. Due to the nature of their operation they will introduce more latency and jitter than a switching device. In addition they are the most likely devices to introduce packet sequence errors. A proper network map will distinguish between routing and switching devices and routed paths vs. physical links.

[3] See http://www.cisco.com/warp/public/788/voip/delay-details.html or http://ftp.tiaonline.org/tr-41/Tr4112/Public/2000-02-Maui/TR4112_00-02-04_Delay%20Budget%20Planning,%20K.Patel,%20M.Knappe,%20Cisco.doc for a good discussion of calculating a delay budget for voice/video devices.

Application Flow

Proper planning requires knowing what your network is or will be used for.

Class characteristics of applications: In the last decade, the number and type of computer applications used over networks has skyrocketed. Each application needs certain QoS parameters to best function, and some applications cannot function without a minimum QoS level. It is not unusual for devices that are responsible for maintaining QoS standards to try to track 500 or more application or protocol types.

Real-time voice or video is the most sensitive to latency, jitter (including serialization delay) and sequence errors.[4] From the Workshop's research, interactive video requirements are typically: available bandwidth between 0.1 and 2 Mbps, much less than five percent packet loss, less than 400 ms latency and a probability of less than ten percent that jitter will exceed 20 ms. The Workshop's recommendations vary widely for other types of video. Sensitivity to sequence error is often noted for voice/video applications, but rarely quantified.

Traffic type	Bandwidth	Conversation style	Latency sensitivity	Jitter sensitivity	Loss sensitivity	Sequence sensitivity
Bulk transfer	10-100 Mbps	Periodic, two way	Low	None	Low	Very low
Transaction-based	<1 Mbps	Bursty, two way	Moderate	None	None	Low
Voice and Fax	8-64 Kbps	Bursty, two or more party	High	High	Low	Counts as loss
Voice and image	>383 Kbps	Bursty, two or more party	High	Moderate	Low	May count as loss
Video on Demand	Varies widely up to 1.5 Mbps	Bursty, multi-party	Low	Low	Low	Low

Application Identification

Various devices can identify applications.[5] Some routers and switches are being designed with QoS features to discriminate between applications. A

[4] More specific QoS recommendations were given in a July 2001 QoS Workshop presentation titled "QoS Requirement to Support Voice and Audio Applications." The presentation can be found online at http://www.ia.net/conferences/qos/DPrice.pdf.
[5] MOREnet released an evaluation of some of the devices in 2001. The paper can be found at http://www.more.net/technical/research/bandwidth/qos-ced/index.html.

good network map should note the transmission paths and end devices for any significant applications.

Application Quantity

The quantity of the application transmissions may also be significant. Some bulk transfer applications can substantially impact network congestion. Thus, schedule bandwidth-hogging backups of computing devices or large databases during times when network load is low or the backup can have a chosen QoS level. The QoS experienced by some applications may vary by the number of conversations that occur. A classic example is voice/video sessions. A single video connection may experience an acceptable QoS. As other video connections are added, each connecting may experience an acceptable QoS. Eventually, however, the addition of another video session will reduce the available bandwidth for all the sessions, and every session may fail to have unacceptable quality.

Planning Decisions

Decide Priority by Application / Device / Physical Segment

Once the network and its uses are mapped, make any necessary QoS planning decisions. What is an acceptable QoS standard for each application? What end devices need higher priority or better physical connections? Can things be done to reduce or eliminate unnecessary network traffic? Can new connections be added or bandwidth increased? Can the physical path be shortened between network elements to reduce congestion for other traffic or enhance its own QoS? For example, can centralized backup devices be located closer to the servers that use them? Can video devices be located so that there are no congested links between them?

Decide Protocols / Devices / Configurations

In some cases, multiple network protocols are available for a given application. For instance, interactive video can use a variety of the H.3xx protocols, and the minimum QoS is different for each protocol. Can an application be better served by using a different communication protocol? Does the network need a QoS-enforcing device? If so, how should the device be configured? Can current switches and routers be upgraded to assist with maintaining the QoS policy? Can switch, router or end devices be configured to enforce a QoS policy?

End devices are also incorporating more QoS abilities. The latest versions of Microsoft Windows can set an IP Precedence value for any network traffic it generates. Network administrators will need to consider whether to trust precedence settings generated by endpoints.

Where to Classify

If the internal LAN switches can enforce QoS policies, then decisions must be made about where to classify and mark the priority of packets. In order to use the QoS provisions of the MOREnet network, traffic must be marked by priority. Can the end devices be trusted to appropriately mark traffic? Can in-line devices make use of marking standards like 802.1p, IP Precedence or Differential Services? Should a separate QoS-enabled device be added to classify, mark and queue the traffic as close to the source as possible so that the maximum number of devices can use the information? Some switches, however, are capable of controlling the entire TCP/IP traffic stream even if they are not placed near the source of the traffic. These switches modify the TCP handshaking of the traffic.

Reduce Complexity

As much as possible, work to reduce the number of devices and links on the network in favor of a more sustainable QoS configuration. Is it possible to eliminate loops and maintain an acceptable network uptime? Is it possible to design the network in a hierarchical or modular fashion that increases stability and helps implement a consistent QoS?

If possible, keep physical connections consistent. Network interface cards (NICs) and switches can negotiate link speed and duplex, but allowing them to select line characteristics sometimes leads to problems. Many video problems are the result of connections where the switch and NIC did not agree on the duplex setting after negotiation. Also, broken connections take longer to restore, if negotiated. Consider establishing a policy that all low priority users are required to connect at 10 Mbps, half-duplex, and all servers connect at 100 Mbps, full duplex.

Decide on a Deployment Strategy

Various ways of implementing QoS have been elaborated upon in other MOREnet white papers.[6] The variety of networks and administrative policies is too complex to determine a single way of enforcing QoS policies. Here are a few things to consider.

Should your QoS policy limit some applications and/or users in the amount of bandwidth they can access? Should you apply rate-limiting strategies at the edge of the network or per network device? What queuing strategies are available for the network devices? Are policy-based routing, virtual private networks or priority queuing available for use? What QoS protocol should be used (IP Precedence, DiffServ, 802.1P, etc.)? Does the network need gateway devices to limit the number of sessions per application?

Are the physical links appropriate for the priority of the traffic? Anything connecting with a shared 10 Mbps Ethernet is not a priority. Switched 10 Mbps half duplex connections can support priority traffic for low volume

[6] MOREnet publishes its white papers at http://www.more.net/technical/research/.

applications only. Full duplex connections can support higher traffic volumes with less serialization and fewer collision delays. Fast Ethernet (100 Mbps) can support medium volumes of mixed priority traffic or high volumes of top priority traffic.

Consider the Cost

Is there a significant cost savings or benefit for some network designs? Will reducing unwanted network traffic save money? Is it worth the cost to handle priority traffic in a priority way? Are the planned changes affordable?

Implement/Verify

The basic implementation strategy is to simply follow the plan. Roll the plan out in phases. You may first wish to implement the technology and strategy to classify traffic. After classifying the traffic, implement the procedures and configurations to limit destructive applications. Third, establish priority handing of priority traffic. Continue to make relatively small, incremental changes. Pause to assess the implementation after each step. Is the network reacting as expected? Are the users reacting as expected? Each implementation step changes the network's condition and your user community. Be aware of any adaptations the users may make that limits or circumvents your QoS plan.

Review the plan by breaking it down into verifiable checklist items. With each item, list the network modifications to be made and the expected results.

Administrator Buy-in

It is critical that everyone who has a stake in the success of the network plan review and accept the package of changes before attempting to implement the plan.

Acceptable Use Policy (AUP)

MOREnet has guidelines for drafting an Acceptable Use Policy.[7] Some QoS problems are a result of user behavior and can be cured by establishing an AUP. For best results, you must publicize and enforce your AUP. The intellectual property provisions of most AUPs might have cut down on the impact of applications such as Napster, had they been known and enforced. AUPs can have an effect, even if not enforced. AUPs communicate standards for the reasonable use of the network is and can change user behavior in cases where the user was uninformed of the consequences of his/her actions.

Checklists

Review the checklist of planned features. Are all the physical attributes in place? Does the reality of the implementation reflect the benefits listed in the plan?

[7] See http://www.more.net/about/policies/aup.html for MOREnet's AUP.

Monitor

Monitoring a network is its own field of study. How the network can/should be monitored varies widely between networks and network administrators. While network monitoring is outside the scope of this paper, it is important to note what administrators should monitor for proper QoS operation.

Monitor the current behavior of applications

Applications display an evolution over time. Initially, Napster was a fringe application that had little impact on a network. When it became well known and widespread, its characteristic large, multiple-file downloads wreaked havoc on some networks. Filtering and firewalling reduced this burden. Napster advocates reacted by hiding or redirecting the nature of the traffic. Eventually, outside legal influences caused Napster usage to drop to levels that are not dangerous to most networks. This example highlights the need for network administrators to actively monitor their networks by application. Some QoS appliances and RMON probes can monitor by application. Look for devices that recognize and report on specific applications.

Monitor the development of new applications

Napster use identified a market segment of network users. As Napster has declined, other applications rose to meet the demands of this market and have taken over. Network administrators need to be able to identify and react to the development of new applications and uses of the network. Application recognition devices must be kept up to date and/or upgraded as conditions change. In addition, network administrators need to be flexible enough to identify traffic trends even if unable to pinpoint a specific application.

Monitor the development of new QoS tools

Quality of Service definitions, requirements and methodologies are changing. The tools to implement and measure QoS are also changing. Monitor new developments in QoS standards and processes as part of the monitoring strategy.

Repeat the Cycle

Computer networks are not static. Network traffic considerations are dynamic, the user community is dynamic, application requirements change and sometimes, the network design itself is dynamic. The final key to improving the quality of the network is to start again at the planning step when monitoring indicates that network conditions have changed, and the changes impact network performance. Even if there are no significant changes, it is often a good idea to periodically review QoS goals and the QoS implementation.

Conclusion

What is considered Quality of Service varies widely across networks. No simple rules apply to all user communications and network administrators. No single plan fits all physical networks. This paper has presented guides to considering how to implement QoS in your network. The key is to decide what QoS means in your environment and to use a comprehensive methodology (like the Plan, Implement, Verify, Monitor method presented here) to incrementally improve the quality of your network.

Appendix A: Best Practices Definitions

Whatis.com:
http://searchvb.techtarget.com/sDefinition/0,,sid8_gci498678,00.html

Legislative Analyst's Office:
http://www.lao.ca.gov/1298_info_technology_best_practices.html

MOST Clearing House Best Practices:
http://www.unesco.org/most/bphome.htm

Appendix B: References

Microsoft, Windows 2000 Server Best Practices,
http://www.microsoft.com/windows2000/en/server/help/default.asp?url=/windows2000/en/server/help/sag_qosbestpract.htm

Microsoft, Enable QoS on Multimedia and Mission-Critical Networked Applications,
http://msdn.microsoft.com/library/default.asp?url=/library/en-us/apcompat/appcomp_444z.asp

Sun, Quality of Service Best Practices,
http://www.eforceglobal.com/hpa_webinar/

Best new practices for network design,
http://www.nwfusion.com/best2000/best-design/best-design.html

Sun, http://playground.sun.com/mobile-ip/QoS-archive/msg00154.html

Cisco, Performance Management: Best Practices White Paper,
http://www.cisco.com/warp/public/126/perfmgmt.htm

Global Grid Forum, http://www.epm.ornl.gov/ghpn/straw_1.txt

Virginia Tech, IP QoS: A Progress Report,
http://rdweb.cns.vt.edu/public/slides/vit2000/qos-pr/qos-pr.ppt

CITI, Authenticated QoS,
http://www.citi.umich.edu/projects/qos/authenticated_qos.html

Euclid, http://www.euclid.com/newsAug20.html

IT Papers, Quality of Network Service White Papers,
http://itpapers.com/cgi/SearchIT.pl?scid=45&search=traffic%20management

Telecom Regulatory Authority of India, Consultation Paper,
http://www.trai.gov.in/qos_8oct.htm

Network World Fusion, Testing switch performance,
http://www.nwfusion.com/reviews/1019bg.html

Nick Farber, Setting TCP Window Size for Legacy LAN VoIP, http://www-ise.stanford.edu/~yiliang/netergy/MEETING/setwin.pdf

Usenix, Wide Area Network Ecology, http://www.usenix.org/publications/library/proceedings/lisa98/full_papers/meek/meek_html/meek.html

Cisco, Designing for LAN/WAN QoS, http://www.cisco.com/univercd/cc/td/doc/product/voice/ip_tele/solution/11405 (Good decision guide.)

Cisco, Planning the IP Telephony Network, http://www.cisco.com/univercd/cc/td/doc/product/voice/ip_tele/solution/3_planni.htm (Various rules of thumb.)

Los Alamos National Laboratory, Overview of Process Improvement, http://peak.lanl.gov:1500/pdfs/qp/pig1.pdf

Dave Price, QoS Requirements to Support Video and Audio Applications, http://www.ja.net/conferences/qos/DPrice.pdf (Very good presentation on jitter, latency and loss requirements.)

Owens, Congestion on Data Networks, http://www.cs.georgetown.edu/~torsneyw/cosc390/materials/Lecture_6_23Oct01.ppt (Rule of thumb: when links are 80 percent utilized, queuing increases rapidly.)

Cisco, Understanding Delay in Packet Voice Networks, http://www.cisco.com/warp/public/788/voip/delay-details.html

Cisco, VoIP end-to-end delay budget planning, http://ftp.tiaonline.org/tr-41/Tr4112/Public/2000-02-Maui/TR4112_00-02-04_Delay Budget Planning, K.Patel, M.Knappe, Cisco.doc