AN OVERVIEW OF QUALITY OF SERVICE COMPUTER NETWORK

Mrs. Amandeep Kaur, Assistant Professor,
Department of Computer Application,
Apeejay Institute of Management,
Ramamandi, Jalandhar-144001, Punjab, India.

Abstract

This paper highlights some of the basic concepts of QoS. The major research areas of Quality of Service Computer Networks are highlighted. The paper also compares some of the current QoS Routing techniques.

Keywords: GoS; QoS; QoS Routing.

1. Introduction

The network Quality of Service (QoS) is a relatively new term, which is defined as: “The capability to control traffic-handling mechanisms in the network such that the network meets the service needs of certain applications and users subject to network policies”. To provide the capabilities of measure and control required by either definition, QoS networks must have mechanisms to control the allocation of resources among applications and users.

The notion of QoS came up as a response to the new demands imposed on the network performance by modern applications, especially multimedia real-time applications. Those applications made it necessary to set limitations on what can be defined as an acceptable time delay when routing information over a network. Those time demands are classified into three main categories. The first is the subjective human needs for interactive computing such as chatting sessions and other interactive web applications. The second is the automated tasks under time constraints such as the automated once-per-day backups during a limited pre-assigned time period. The third category is the need of some applications for a transmission rate with limited jitter along with a temporal ordering of the transmitted packets. This is the case when streaming multimedia over a network. The transmission rate is needed to keep the transmitted material meaningful and perceptible while the preserved temporal order is needed for synchronization.

The temporal requirements presented above are intrinsic to QoS that some references define QoS in terms of those requirements. Webster’s New World Dictionary of Computer Terms defines QoS to be “the guaranteed data transfer rate”. The word “guaranteed” is of special importance since QoS can only be implemented through guarantees on the limits of some network parameters as will be explained below.

It is important here to note that although QoS became an issue only in the past few years, but the idea of QoS had been envisioned earlier before new applications mandated the use of QoS. In the initial IP specification, a Type of Service (ToS) byte is reserved in the IP header to facilitate QoS. Until the late 1980s, almost all IP implementations ignored the ToS byte since the need for QoS was not yet obvious.

Comparison of GoS and QoS

It is not an easy task to find the GoS(Grade-of-Service) standards needed to support a certain QoS. This is due to the fact that the GoS and QoS concepts have different viewpoints. While the QoS views the situation from the customer's point of view, the GoS takes the network point of view.

Reference configurations

In order to obtain an overview of the network under consideration, it is often useful to produce a so-called reference configuration. This consists of one or more simplified drawing(s) of the path a call (or connection) can take in the network including appropriate reference points, where the interfaces between entities are defined.

Consider a telephone network with terminals, subscriber switches and transit switches. In the example we ignore the signalling network. Suppose the call can be routed in one of three ways:

1. terminal → subscriber switch→ terminal

This is drawn as a reference configuration shown in Fig. 1.21.

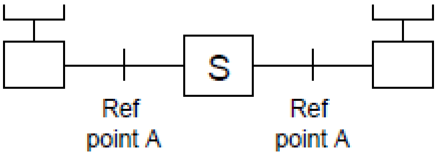

Figure 1.21: *Reference configuration for case 1.*

2. terminal → subscriber switch→ transit switch→ subscriber switch → terminal

This is drawn as a reference configuration shown in Fig. 1.22.

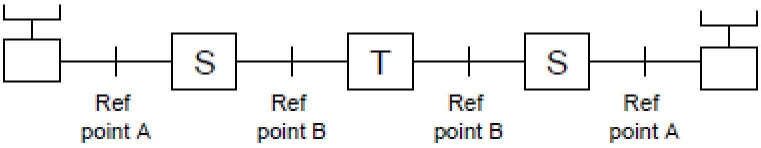

Figure 1.22: *Reference configuration for case 2.*

3.terminal→subscriber switch→transit switch→transit switch→subscriber switch→terminal

This is drawn as a reference configuration shown in Fig. 1.23.

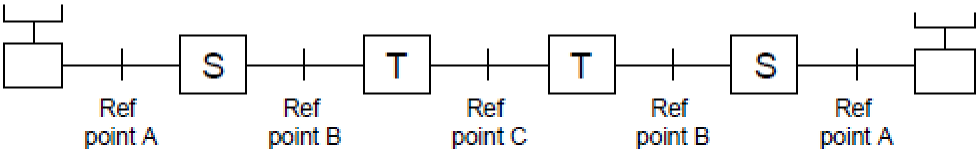

Figure 1.23: *Reference configuration for case 3.*

Based on a given set of QoS requirements, a set of GoS parameters are selected and defined on an end-to-end basis within the network boundary, for each major service category provided by a network. The selected GoS parameters are specified in such a way that the GoS can be derived at well-defined reference points, i.e. traffic significant points. This is to allow the partitioning of end-to-end GoS objectives to obtain the GoS objectives for each network stage or component, on the basis of some well-defined reference connections.

2. QoS Performance Measures

In order to provide QoS, some quantitative measures of what constitutes QoS must be defined. As mentioned above, QoS is quantitatively defined in terms of guarantees or bounds on certain network performance parameters. The most common performance parameters are the bandwidth, packet delay and jitter, and packet loss.

2.1 *Bandwidth*: The term bandwidth defines the transmission capacity of an electronic line. Theoretically, it describes the range of possible transmission rates, or frequencies. In practice, it describes the size of the pipe that an application program needs in order to communicate over the network .The significance of a channel bandwidth is that it determines the channel capacity, which is the maximum information rate that can be transmitted. The relationship between channel capacity and information transmission rate was set in the Information Theory of Claude Shannon in the 1940s.

According Shannon's information theory, if information rate is R and channel capacity is C, then, it is always possible to find a technique to transmit information with arbitrarily low probability of error provided $R \leq C$ and, conversely, it is not possible to find such a technique if $R > C$.

2.2 ***Packet Delay and Jitter*****:** The delay, also known as latency, consists of three different types, namely, serialization delay, propagation delay, and switching delay. Serialization delay, also called transmission delay, is the time it takes a device to synchronize a packet on a specified output rate. This transmission delay is a function of the bandwidth and the packet size. For example, a packet with size 64 bytes would take 171 μs when sent at the rate of 3Mbps. The same packet would take 26 ms when sent at the rate of 19.2 kbps.

Propagation delay is the time it takes a bit to travel from a transmitter to a receiver. Physics set upper limits on the speed of such a bit, making at best a fraction of the speed of light. Hence, propagation delay is a function of the distance traveled and the link medium.

Switching delay is the time lag between receiving a packet and starting to retransmit it. The switching delay is a function of the device speed.

In addition to those three types of delay, other delays also contribute to the overall performance of the network. Depending on the traffic, network condition, and the nature of the information being transmitted, different packets will experience different delays. The term packet jitter refers to this variation in packet delay. When the network is congested, queues will build up at the routers and start affecting the end-to-end delays.

Queuing delay may be negligible when the network is fast and not experiencing congestion. However, when the network is congested, the queuing delay grows and becomes significant. The number of clients in a queue is a random variable and its distribution depends on r, the ratio of arrival rate to service rate. The probability p of having n clients in the queue is computed as follows:

$$P(n)=(1-r)*r^n \qquad (1)$$

The queuing delay is a function of the number of packets in the queue and the service time for each queue. When the service rate is μ, the average queuing delay aqd can be computed as follows:

$$aqd= 1/(1-r)\ \mu \qquad (2)$$

2.3 ***Packet Loss*****:** Packet loss is another important QoS performance measure. Some applications may not function properly, or may not function at all, if the packet loss exceeded a specified number, or rate. For example, when streaming video frames, after certain number of lost frames, the video streaming may become useless. This number may be zero in certain cases. Therefore, certain guarantees on the number of rate of lost packets may be required by certain applications for QoS to be considered. Packet loss can occur because of packet drops at congestion points when the number of packets arriving significantly exceeds the size of the queue. Corrupt packets on the transmission wire can also cause packet loss.

3. QoS Levels

Even after realizing the inevitable need to QoS networking, there are still many applications that do not require any QoS. Moreover, those applications that do require QoS differ in the degrees of priorities and guarantees that they require to implement QoS. Therefore, on one extreme we have tasks that do not require any guarantees. On the other extreme, we have tasks that require absolute guarantees that may not be compromised. In between those two extremes there are numerous levels of QoS. However, those levels of QoS have been grouped into three main categories: best effort service, soft QoS, and hard QoS.

3.1 ***Best effort service*****:** The level of best effort service provides no guarantees at all. It represents the first extreme mentioned above. It cannot really be considered a QoS. Many network applications work very well with best effort service. An example of such applications is the File Transfer Protocol (FTP). No guarantees or performance measures are assumed for FTP. The only criterion is whether the transfer was completed successfully or not.

3.2 ***Soft QoS*****:** Soft QoS is also known as differentiated service. In this QoS level, no absolute guarantees are given. Rather, different priorities are assigned to different tasks. Hence, applications are grouped into different classes of priorities. Many application traffics work very well with this policy when absolute guarantees are not needed. For example, network control traffic should always be given higher priority over other data communications to ensure the availability of, at least, the basic connectivity and functionality at all times.

3.3 ***Hard QoS*****:** Hard QoS is also called guaranteed service. It represents the level of QoS for applications that require absolute guarantees on the minimum needed resources of the network in order to function correctly, or in order to function at all. Prior network resource reservation over a path is usually performed to enable the network to provide, or deny, the required guarantee. Applications that require Hard QoS include multimedia applications, where streaming audio and/or video data is done in real-time.

It is important to note here that the only level the Internet is currently able to provide is the first level. Although the IP protocol supports QoS but still the Internet does not offer any of the other two levels of QoS. However, other networks such as ATM networks do support QoS .

4. QoS Routing

The emergence of QoS networking created many challenges to network developers of many fields. It is true that physicists and engineers contributed significantly to the development of faster networks. On the horizon now are optical networks that use Wavelength Division Multiplexing (WDM) technology to increase the capacity of optical transmission systems by transmitting multiple wavelengths over a single fiber, reaching transmission rates on the order of terabits per second. However, as the physical capabilities of the networks grow, the demands by new applications to exploit those capabilities also grow. This necessitates the need for computer scientists to constantly develop algorithms and solutions to provide the needed exploitation.

Many factors complicate the problem of QoS routing. One of those factors is the diversity of the requirements and guarantees of different distributed computing applications running simultaneously. This problem expands to include applications with zero constraints requirements, making it necessary to develop routing mechanisms that handle all the three levels of QoS presented above. The other major factor is the impossibility of maintaining accurate network state information in a large dynamically changing network. This latter factor will be a very important theme in the current dissertation.

To maintain network state information, each node in the network needs to maintain its local state. Then all local states can be combined to form the global state information. Typically, a node maintains the network global state information using one of two algorithms: link-state algorithm and distance-vector algorithm. This is done by using the chosen algorithm to exchange the local states between all nodes in the network periodically. The resulting global state information cannot be accurate due to many factors that will be discussed later. Dealing with this uncertainty about the network global state is one of the main problems this dissertation is trying to solve.

4.1 Routing Classification

QoS routing algorithms can be classified according to the cardinality of the destination of the searched path into two main categories: unicast routing algorithms and multicast routing algorithms.

4.1.1 *Unicast Routing*

In a unicast routing algorithm, the problem is to find the best feasible path from a source node to a destination node, satisfying a pre-designated set of constraints.

4.1.2 *Multicast Routing*

In multicast routing, the problem is to find the best feasible tree that covers a source node and a set of destination nodes, satisfying a pre-designated set of constraints. The nature of the problem necessitates the use of algorithms from one category or the other with no trade-off between the two.

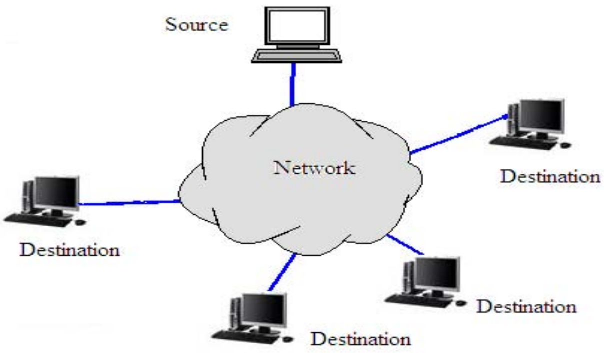

Figure 1.24 An example multicast routing network

Another way of classifying QoS routing algorithms is according to the path search and deployment strategy. Accordingly, there are three routing strategies: source routing, distributed routing, and hierarchical routing. These three strategies can be used interchangeably in many cases with the involvement of trade-offs between the advantages and disadvantages of each strategy.

4.1.3 *Source Routing*

In the source routing, the feasible path is computed locally at the source node, which is expected to have its own maintenance mechanism of the network global state information. The main advantage of source routing is the localized storage of the network state information and the centralized computation of the path. The local maintenance of the global state enables the source node to compute the path locally as well. The computational complexity in this case is much smaller than that of the distributed computing. This, in turn, makes source routing algorithms much easier to design and implement. In addition, it guarantees loop-free routing. However, source routing suffers from two major problems. The first is the inaccuracy of state information. The degree of precision of the global state at each node is directly proportional to the frequency of updates. Nevertheless, the updating frequency is also directly proportional to the updating overhead, which is inversely proportional to the availability of resources for the actual network activities. This inevitable imprecision in the global state information may result in the failure of finding an existing feasible.

4.1.4 *Distributed Routing*

The second routing strategy, distributed routing, depends on using distributed computing to compute the path. The computation is done by exchanging control messages and the global state information stored locally at each node. However, some distributed routing algorithms do not require the maintenance of global state at all .The main advantage of distributed routing is the distributed computation of the path, which enables shorter response time and better scalability. The shorter response time and higher scalability are achieved at the expense of higher network traffic due to more message exchanging. Furthermore, distributed routing cannot be loop-free, especially when global states at different nodes are inconsistent.

4.1.5 *Hierarchical Routing*

In the last routing strategy, namely, hierarchical routing, nodes are grouped into clusters, which are further grouped into a higher level clusters. This recursive clustering continues to build up forming a multi-level hierarchy. Instead of maintaining global state information at each node, the aggregated state is maintained, where an elected node in each cluster maintains the global state of the nodes in the cluster in which it is local to in addition to the aggregated states of the other clusters. The use of partial global states maintained by logical nodes enhances the scalability of the hierarchical routing significantly over other routing schemes. In addition, the overall traffic in the network does not get so intense as it does in distributed routing.

Thus, hierarchical routing combines advantages of both source and distributed routing. The only noticeable problem with hierarchical routing, which is not a trivial one, is that the aggregation of the network states introduces additional imprecision.

5. Conclusion

The paper concludes that the performance of routing algorithms that are not designed specifically to take imprecision into account degrades significantly as the imprecision grows. Most QoS routing algorithms available today do not take this uncertainty into account. Instead, they assume it does not exist, regardless of the inherent nature of this uncertainty. However, research has been done to evaluate the impact of neglecting this uncertainty on the performance of different routing algorithms. In addition, few routing algorithms have been proposed with the main objective of handling the intrinsic imprecision and reducing its effect.

6. **References**

[1] S. Chen and K. Nahrstedt, “An Overview of Quality of Service Routing for Next-Generation High-Speed Networks: Problems and Solutions,”.

[2] X. Yuan and W. Zheng, “A Comparative Study of Quality of Service Routing Schemes That Tolerate Imprecise State Information,” Florida State University Computer Science Department, Technical Report. [Online].Available: http://websrv.cs.fsu.edu/research/reports/TR-010704.pdf.

[3] “Quality of Service Based Routing: A Performance Perspective”, ACM SIGCOMM, 1998.

[4] “Predictive routing to enhance QoS for stream-based flows sharing excess bandwidth” Xun Su, Gustavo de Veciana.

[5] Chao Peng, Hong Shen, "New Algorithms For Fault-Tolerant QoS Routing",submitted to the International Conference on Dependable Systems and Networks(DSN-2006).