A Generic Quality of Service Management Model for Network-aware Applications

Pattarasinee Bhattarakosol, Wijak Srisujjalertwaja

Department of Mathematics, Faculty of Science
Chulalongkorn University, Thailand
Bhattara@chula.ac.th, wss@cs.science.cmu.ac.th

Abstract

A network-aware application is an application that can adapt itself to the changing network environments. According to the Quality of Service (QoS) concept, QoS management is executed to deliver service license agreement (SLA) between client and server. By merging of these two concepts, this paper proposes a generic QoS management model as a framework of a Wireless Internet system for network-aware applications. An end-to-end QoS management functions on server-level and client-level are provided to maximize overall users' satisfaction and individual user's satisfaction, respectively. A QoS management factor, end-to-end transmission delay, is experimented and analyzed in this case study at application-level to see its implication to the QoS management functions.

1. Introduction

The Wireless Internet infrastructure is an interested environment due to its service limitations versus its usage growth. There are many service limitations of the Wireless Internet; such as low bandwidth, high resource variation, and intermittent connection, while the usage growth rate is rapidly high [4] [5]. The quality of service becomes an important need of all Wireless Internet users from their providers according to QoS concept [3] [7] [10]. Although there are many methods have been applied to improve QoS system, none of them can fully fulfill customers' satisfaction.

The Wireless Internet architecture can be separated into two parts: a Web server on wireline network, and clients on wireless networks as shown in Figure 1. These clients connect to the server through the Internet. Currently, there are many wireless communication standards with different data rates. This study concerns only application level or the end-to-end communication manner, while leaves the other communication parts; the Internet infrastructure, as a black box. Similar to other general Web servers, the server in this system supports three QoS classes, which are interactive, background, and streaming. Each service class has its own requirements which normally depend on timeliness, accuracy, capacity, and security.

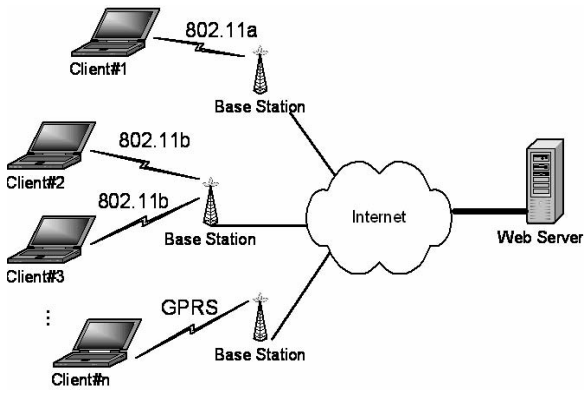

Figure 1. Wireless Internet

In traditional communication system, the performance of each service class is relied on the allocated communication resources; which are static to the system changes during run time. In order to improve the service performance, the dynamic resource management mechanism is required to manage and ensure resource availability of the system. Therefore, the resource management mechanism is a heart of the QoS.

This paper proposes a generic QoS management model for network-aware applications in section 2; also defines QoS management cycle and QoS management functions in section 2.1 and 2.2, respectively. End-to-end transmission delay at the application-level is experimented and analyzed as a sample of user requirements in section 3.

2. A generic QoS management model

Network-aware application [2] will adapt itself, whenever a system factor is exceeded the committed threshold in SLA, in order to maintain the network communication status. In this paper, the function that performs the adaptive process in the network-aware application is called the QoS management functions.

The QoS management functions are installed on both sites: the client site, and the server site. At the client site, the QoS management function is applied to maximize individual user's satisfaction, while the management function at the server site attempts to maximize the overall users' satisfaction.

QoS specification is concerned with capturing application-level QoS requirements and management policies [1]. The QoS requirements can be considered in two different levels: the user level, and the network level. At the user level, users generally concern with various issues such as availability, performance, accuracy, and reliability. On the other hand, the values of throughput, delay, jitter, response time, and loss rate are considered at the network level.

This paper selects to focus the end-to-end transmission delay as a case study, because it's one of the important communication factors of users' requirement. The system provides the following QoS policies to control and react to the system's events:

- Every user has single role, they have equivalent priority to get the services.
- New user admission and renegotiation process must regard to the other users who occupied the system resources.

2.1. QoS management cycle

In this study, a communication session viewpoint is considered as shown in QoS management cycle in Figure 2. When a client requests a service, the server creates a service agent to classify request based on the request's parameters and the categorized knowledge. Then network parameters' threshold generated from the classification process as a proper SLA. Next stage, the service agent tries to negotiate network resources with resource manager, who manages the network resources in the system. The output of the negotiation process is an assigned SLA. During transmission, if there is any problem, the client can send a request to activate the server, and then renegotiation process is invoked to adapt the transmission to the changing event. When the transmission is terminated, either success or failure of the user' request, the system will release the allocated network resources and end the service agent. Successful request will be kept as historical data. The maintenance process will analyze historical data into the categorized knowledge, as a back-end process.

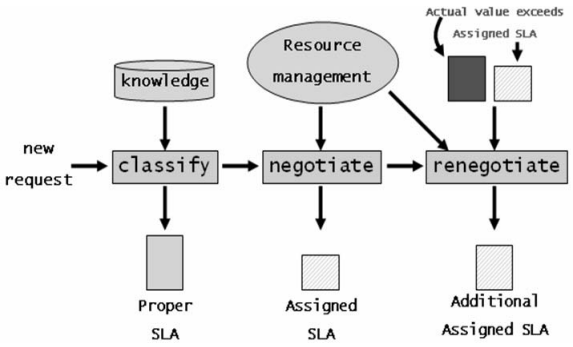

Figure 2. QoS management cycle

2.2. QoS management functions

The QoS management functions are the processes to maximize overall and individual users' satisfaction, according to SLA. The SLA is an agreement between the server and a client. Hence, the service that a client received must be maintained in the range of the agreed threshold in SLA. The QoS management functions must be concerned in both server-level and client-level.

- *The server-level QoS management function*
  A related definition of the server-level QoS management function is defined as following:

  Definition 1 *Provided Resource Capacity*

  Provided Resource Capacity, abbreviates $feature$, is the allocated capacity of resource element for a session, such as transmission rate, packet size.

  The server-level QoS management function can be defined as an optimization function, where the objective is to maximize the overall satisfaction of the users that currently connect to the particular server. The optimization function of the server-level QoS management function can be modeled as following:

$$\max \ satisfaction = \sum_{i=1}^{n} S(i)$$

  subject to:

- $\forall S(i) \geq SLA_{min}(i)$ for $i = 1$ to n,
- $\forall(j)(\sum_{i=1}^{n} feature(i,j) \leq overallFeature(j))$ for $j = 1$ to p

where

- n be the number of the sessions that currently connected to the specified server,
- $S(i)$ be a satisfaction value of the session(i),
- $SLA_{min}(i)$ be a minimum satisfaction value of the session(i),
- p be the number of the features in the system,
- $feature(i,j)$ be a resource feature(j)'s capacity which is occupied by the session(i),
- $overallFeature(j)$ be an overall resource feature(j)'s capacity which provided by the system.

- *The client-level QoS management function*
Related definitions of the client-level QoS management function are defined as following:

Definition 2 *SLA Impact Factor*

SLA impact factor, abbreviates $factor$, is a value impacts to satisfaction level, such as delay, jitter, loss rate, and throughput. The factor can be categorized into two types:

- Positive factor. If this factor value increases, the system performance will be increased, such as throughput.
- Negative factor. If this factor value increases, the system performance will be decreased, such as delay and loss rate.

Definition 3 *Acceptance Level*

Acceptance level of a factor, $acceptanceLevel()$, is a calculating function that calculates the accept value of a session's factor. This function returns a continuous value in range [0, 1], such that 0 is unaccepted value and 1 is excellent accepted value.

Furthermore, for any communication session k, the QoS management function can also be defined as an optimization model as following:

$$\max S(k) = \frac{\sum_{j=1}^{m} acceptanceLevel(factor(k,j))}{m}$$

subject to : $\forall factor(k,j)$ such that

- if factor(j) is a positive factor, $factor(k,j) \geq factor_{min}(k,j)$,
- if factor(j) is a negative factor, $factor(k,j) \leq factor_{max}(k,j)$,

where

- $S(k)$ be the user's satisfaction value of the client who owns the session(k),
- $factor(k,j)$ be an impact factor(j) of the provided services for the session(k)
- m be the number of concerning factors,
- $factor_{max}(k,j)$ be the maximum value that a user; session(k), can accept for any provided service factor(j),
- $factor_{min}(k,j)$ be the minimum value that a user; session(k), can accept for any provided service factor(j).

The objective of this function is to maximize individual user's satisfaction, the session(k). The constraints of this function are the maximum or minimum thresholds of all provided service factors, all session must hold these values, and these values must not exceed the capacity that the server can be provided.

3. The QoS management functions: a case study

At the network-level, there are many evaluation factors, such as jitter, delay, and loss rate. One key indicator of the user's satisfaction depends on the transmission delay, which related to the transmission rate and packet length as shown in section 3.1. This key indicator is used as a case study of the QoS management functions in this paper. The experimental detail to support the transmission rate adaptation is explained in section 3.2. Applications of the server-level and the client-level QoS management functions are explained by replacing of features and factors that affected by the transmission rate adaptation. The server site and client site QoS management functions are presented in section 3.3 and 3.4, respectively. The related definitions are defined as following:

Definition 4 *Maximum Transmission Delay*

The maximum transmission delay value is the maximum transmission delay that user can accept, according to SLA. This value depends on service class that the user requests, denoted by D_{max}.

Definition 5 *Transmission Delay*

Transmission delay, defined in [6], is amount of time required to put all of the packet bits into the link.

$$D(i) = L(i)/R(i)$$

where

- $D(i)$ be a transmission delay of session(i),
- $L(i)$ be a length of a packet of session(i),
- $R(i)$ be a transmission rate of session(i).

3.1. Transmission rate adaptation

In this paper, the end-to-end delay, which considers the delay from source to destination, are proposed. Assume the network between the source and destination host is a black box. Furthermore, every node in the network has high performance computing, and the network between the two hosts is uncongestion, therefore processing delays and queuing delays are negligible, while propagation delay is constant. Only the transmission rate out of the source host, R bits/sec, is considered. The transmission delay is the amount of time required to push or transmit all of the packet bits into the link [6]. The transmission delay is

$$D = L/R$$

where D be the transmission delay, L be the packet length, R be the transmission rate.

If L or R is changed, then D will effected.

3.2. Experimental results

This study uses ns-2 [8] as a tool for Wireless Internet simulation. This study observes average delay time (D, transmission delay) by varying bandwidth (R, transmission rate) and packet length (L). By passing 40,000 data packets, the experimental results are shown in Figure 3 and 4.

From figure 3, increasing the bandwidth (R) will decrease the transmission delay (D). From figure 4, decreasing the packet length (L) will increase the transmission delay (R). Both experiment fixed one sliding window and distance to 50. Bandwidth testing sets packet length to 128, while packet length testing sets bandwidth to 1M.

From the experimental result, transmission rate (R) is selected to put in server site QoS management function, while transmission delay (D) is putted in client site QoS management function, in section 3.3 and 3.4, respectively.

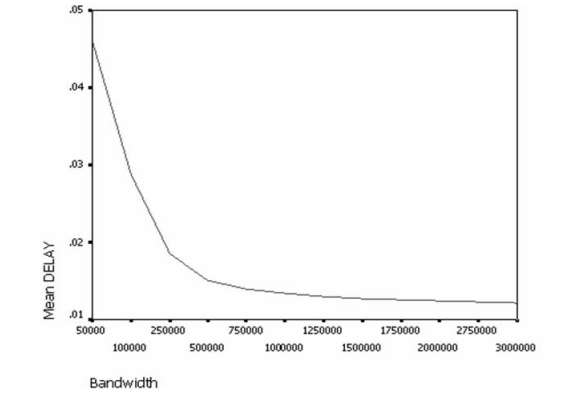

Figure 3. Average delay time and bandwidth variation

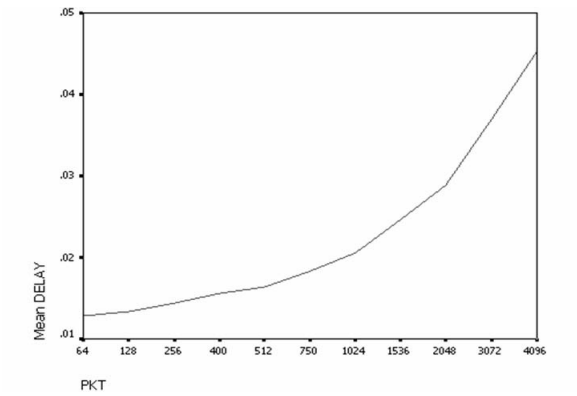

Figure 4. Average delay time and packet size variation

3.3. Server site QoS management functions

When client sends a request to the server, a request classification process is invoked. In this stage, the maximum transmission delay which is suitable for the request service class is calculated. Then system starts the connection. For this case study, the server-level QoS management function will concern only on the transmission rate (R) as a feature that impacts to the transmission delay (D), where the other features are left in term of $otherFeature$. The function is defined as following:

$$\max\ satisfaction = \sum_{i=1}^{n} S(i)$$

subject to:

- $\forall S(i) \geq SLA_{min}(i)\ for\ i = 1\ to\ n$,
- $\sum_{i=1}^{n} R(i) \leq overallR$,

- $\sum_{i=1}^{n} otherFeature(i,j) \leq overallOtherFeature(j)$,

where

- n be the number of the sessions,
- $R(i)$ be a transmission rate that system provides for the session(i),
- $overallR$ be the overall transmission rate of the system,
- $otherFeature = \bigcup_{j=1}^{m} feature(j) - R$,
- m be the number of the system features,
- $overallOtherFeature$ be the overall of feature(j), excluding the transmission rate.

3.4. Client site QoS management function

The client-level QoS management function at the client site can be defined as, for any session k, the function of individual user. This function can be modeled in term of transmission delay as following:

$$\max S(k) = \frac{\frac{D_{max}(k)-D(k)}{D_{max}(k)} + \sum_{j=1}^{m-1} otherFactor(k,j)}{m}$$

subject to: $D(k) \leq D_{max}(k)$,

- if $otherFactor(k,j)$ is a positive factor, $otherFactor(k,j) \geq factor_{min}(k,j)$, and
- if $otherFactor(k,j)$ is a negative factor, $otherFactor(k,j) \leq factor_{max}(k,j)$,

where

- m be the number of factors,
- $D(k)$ be the transmission delay of session(k)
- $otherFactor = \bigcup_{j=1}^{m} factor(j) - D$.

At the client site, the QoS management function is applied to maximize individual user's satisfaction. During transmission, user agent monitors whether an average transmission delay, $D(k)$, is more than the maximum transmission delay, $D_{max}(k)$, or not. If the average delay is more than the maximum delay, the user agent sends the request to invoke the renegotiation process at the server site. The simple moving average technique in [9] is used to find the average delay.

4. Conclusions

In this paper, a generic QoS management model for network-aware applications is proposed. A QoS management model is explained in QoS management cycle and QoS management functions, under the end-to-end transmission delay constraint. QoS management functions in server-level and client-level are stated to maximize overall users' satisfaction and to maximize individual user's satisfaction, respectively. Transmission delay is used as a key factor for the QoS management functions case study. The experimental results indicate the relation of packet length and transmission rate to the transmission delay. The QoS management functions, both server site and client site, are developed using the transmission delay and transmission rate, as a framework for improving QoS of network-aware applications.

References

[1] C. Aurrecoechea, A. T. Cambell, and L. Hauw. A survey of qos architectures. *ACM/Springer Verlag Multimedia Systems Journal, Special Issue on QoS Architecture*, 6(3):138–151, May 1998.

[2] J. Bolliger. *A Framework for Network-aware Applications*. PhD thesis, Swiss Federal Institute of Technology Zurich, 2000.

[3] C. D. and S. M. A survey of quality of service in mobile computing environments. *IEEE Online Communication Surveys*, 2(2), 1999.

[4] K. Flinders. *IDC Predicts Wireless Internet Growth*. VNU Network, VNU Business Publications, http://www.vnunet.com/news/1126661/, November 2001.

[5] Ipsos-Insight. *Wireless Internet Access Climbs Nearly 30% in 2004*. Ipsos-Insight Technology & Communications Practice, http://www.ipsos-na.com/news/pressrelease.cfm?id=2598/, March 2005.

[6] J. F. Kurose and K. W. Ross. *Computer Networking, A Top-Down Approach Featuring the Internet*. Pearson Education, Inc., U.S., second edition, 2003.

[7] Nortal Networks. *Benefits of Quality of Service (QoS) in 3G Wireless Internet*, September 2001.

[8] The network simulator – ns(version 2). ns homepage, http://www.isi.edu/nsnam/ns/.

[9] W. G. Sullivan and W. W. Claycombe. *Fundamentals of Forecasting*. Reston Publishing Company, Inc., Reston, Virginia, 1977.

[10] V. Witana and A. Richards. A qos framework for heterogeneous environments. In *DSTC Symposium*, July 1997.