QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL
Qual. Reliab. Engng. Int. 2006; **22**:429–444
Published online 28 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/qre.720

Research

QoS Model of a Router with Feedback Control

Zhibin Yang, Nong Ye*,† and Ying-Cheng Lai
Arizona State University, Tempe, AZ 85287-5906, U.S.A.

The Internet has evolved into a shared, integrated platform of a broad range of applications with different Quality-of-Service (QoS) requirements. Routers are an important part of the Internet and play a critical role in assuring QoS. A router is usually placed between two networks to receive data packets from one network and then transmit those data packets to another network if necessary. Data packets are the actual units of data traveling on computer networks. A data packet has two parts: header and data. The data carries messages, such as e-mail text, from computer applications. The header carries information that is required to control and manage the transmission of the data packet on computer networks. Existing approaches for providing QoS involve prediction or estimation for traffic characterization to determine parameters required of static traffic admission control. However, prediction or estimation inaccuracy in traffic characterization can result in inappropriate parameter settings for static admission control and, in turn, compromise QoS or resource utilization. This study presents a QoS model of a router with feedback control that monitors the state of resource usage and adaptively adjusts parameters of traffic admission control to overcome prediction or estimation inaccuracy and achieve a balance between QoS and resource utilization. The QoS model of a router with feedback control is simulated to test its performance on QoS and resource utilization in both heavy and light traffic conditions. The performance of the QoS model of a router with feedback control is also compared with that of two basic QoS models of a router with static admission control using admission control parameters resulting from over- and under-characterization of traffic, respectively. The simulation results show that the QoS model of a router with feedback control achieves a better balance between QoS and resource utilization than the basic QoS models with over- and under-characterizations of traffic in the heavy traffic condition. This study also shows that the three models of routers demonstrate similar QoS performances and resource utilization in the light traffic condition. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Quality of Service (QoS); router; feedback control; Internet

*Correspondence to: Nong Ye, Professor and Director, Information and Systems Assurance Laboratory, Arizona State University, Box 875906, Tempe, AZ 85287-5906, U.S.A.
†E-mail: nongye@asu.edu

Contract/grant sponsor: Air Force Research Laboratory—Rome; contract/grant number: F30602-01-1-0510

Received 9 July 2003
Revised 26 February 2004

INTRODUCTION

The Internet has evolved from a traditional data transmission network to a shared, integrated platform to carry a broad range of applications with varying traffic characteristics and different Quality-of-Service (QoS) requirements, including WWW, e-mail, IP telephony, and so on. Some applications come with no hard time constraints and QoS requirements. Others, such as audio and video applications and IP telephony, are time-dependent, placing strict QoS requirements on the carrier network. The diversity of the growing population of network applications requires the Internet to be capable of delivering QoS. Routers are an important part of the Internet to support networking and data communication, and thus play an essential role in providing QoS on the Internet. This paper focuses on the QoS model of a router.

One definition of QoS is 'the ability to differentiate between traffic or service types so that the network can treat one or more classes of traffic differently than other types' (Huston[1]). According to this definition, QoS is rooted in the ability to provide differentiated services with regards to different service requirements. In general, QoS requires the differentiation of services for different classes of network traffic with different QoS priorities and the assurance of certain performance measures such as delay, packet loss, throughput, and so on[2–10].

A router receives data packets from source addresses on the Internet at the input port(s) and sends out these data packets to destination addresses on the Internet through the output port(s). Because a router has a limited bandwidth of data transmission, the router typically uses a buffer or queue to keep incoming data packets when the bandwidth is busy transmitting other data packets. The queue has a limited capacity of storing data. The data packet at the front of the queue is taken out first for receiving services of data transmission through the bandwidth. If a data packet arrives at a router but the queue of the router is full, the data packet is dropped by the router.

Most routers on the Internet operate based on the best-effort model with the First-In-First-Out (FIFO) scheduling rule to determine the order of serving data packets or sorting data packets in the queue[5]. In this best-effort model, a data packet that arrives at the router first is put at the front of the queue and, thus, is taken out of the queue first for services. However, this best-effort model does not provide QoS. The FIFO scheduling rule provides services to data packets based on their arrival times without considering their QoS priorities.

There are considerable research efforts on QoS. QoS can be achieved on either per-flow basis or per-aggregate basis. Integrated Service (InteServ)[4,8,10] is a per-flow-based QoS architecture. Flow is defined as 'a distinguishable stream of related datagrams that results from a single user activity and requires the same QoS' (Braden *et al.*[4]). An end-to-end bandwidth reservation is required to guarantee the bandwidth to an individual flow. InteServ defines a QoS model which is made up of a predictive service, best-effort service and link-sharing service. A reference framework is proposed for the implementation of InteServ[4]. This framework incorporates packet scheduling, packet classification, admission control, and path reservation according to traffic prediction and estimation. Per-flow-based service differentiation provides a fine granularity to isolate flows from each other, thus achieving a firm end-to-end service guarantee. However, the flow-based QoS model has the scalability problem, especially in large-scale computer networks, where there are millions of flows and the management overhead becomes extremely high.

Differentiated Service (DiffServ)[3,7] provides QoS on a per-aggregate basis. DiffServ divides the network into domains. At the edge of a domain, traffic is classified into aggregates, marked and policed in accordance to given administrative policies. The core routers sitting inside the domain provide per-hop behavior (PHB) corresponding to the traffic aggregate. Compared with InteServ, DiffServ needs no end-to-end path reservation, but provides a weaker QoS guarantee than that of the per-flow-based approach.

Due to the varying nature of network traffic, the characterization of traffic in the InteServ model for bandwidth reservation and in the DiffServ model for traffic classification presents a considerable challenge. Kurose[6] discussed four approaches of providing a QoS guarantee. These approaches either prevent changes in traffic characteristics through tight traffic control or tolerate changes by taking into account changes (e.g. in peak rate) through approximation, bounding and observation. The approach of tight traffic control shapes and conditions traffic with a non-work-conserving queuing discipline, with which the output link may be idle even when there are packets waiting for service. To maintain the consistent traffic characterization, the approach of tight traffic

control may intentionally block the arriving flow while allowing the output link to be idle, causing the potential low utilization of the output bandwidth. The approximate approach characterizes traffic using a simple method such as an 'on/off' method that is not applicable to all traffic, and provides the QoS on a conservative basis. The bounding approach provides either a deterministic or a statistical bound on traffic. The most challenging issue in the bounding approach is 'the ability to bound the maximum length of each queue's busy period for a given set of traffic specifications' (Kurose[6]). The observation approach characterizes traffic using prior knowledge about certain types of traffic, which can be obtained either on-line or off-line. Overall, all of these approaches call for traffic characterization, which becomes the basis for providing QoS. Hence, the guarantee of QoS depends on the quality of traffic characterization. In all of the approaches except the approach of tight traffic control, traffic characterization is either estimated or predicted, leading to inaccuracy. When the resource is allocated according to a pre-determined traffic characterization, there is always a possibility that the actual incoming traffic either over-uses or under-uses the allocated resource. For example, when the queuing buffer as a resource is over-used, the queue length increases, packet transmission delay increases and packet loss may occur, compromising the QoS guarantee. When the resource is under-used, a waste of the service capacity occurs.

To overcome prediction or estimation inaccuracy in traffic characterization and achieve a balance between QoS and resource utilization, we introduce feedback control to monitor the state of resource usage and adjust the admission control of incoming traffic adaptively according to the state of resource usage. Next, we describe the QoS model of a router with feedback control. Then, we present the simulation experiments to test this QoS model of a router. Finally, we discuss the testing results and conclude the paper.

QoS MODEL OF A ROUTER WITH FEEDBACK CONTROL

The QoS model of a router with feedback control is built on a basic QoS model of a router with the capability of service differentiation. In this section, we first describe the basic QoS model of a router. We then introduce the QoS model of a router with feedback control.

Basic QoS model of a router

The basic QoS model of a router is designed to provide service differentiation and static resource allocation according to concepts in the Two-bit Differentiated Services Architecture for the Internet[7]. Figure 1 shows the basic QoS model of a router. This model considers two classes of service: high-priority service and low-priority service. In this study, we assume that data packets arrive with the mark of their service class. That is, we do not include packet classification and marking in the router model. The basic QoS model aims at providing service differentiation between two classes of data packets to assure QoS to high-priority data packets and providing the best-effort service to low-priority data packets only after satisfying the QoS need of high-priority data packets.

There are two input ports and one output port in the router model. At each input port, the admission control mechanism uses the token bucket model[7] to condition high-priority data packets; that is, to control the admission of high-priority data packets into the router. Admission control is applied to high-priority data packets for assuring the availability of the resources in the router to provide the premium service to admitted high-priority data packets. There is no admission control for low-priority data packets because: (1) low-priority data packets are served only when there are no high-priority data packets waiting for the resources in the router; and (2) QoS of low-priority data packets is not of concern and is provided on the best-effort basis.

In the token bucket model, admission control is determined by two parameters: token rate r and bucket depth p. Token rate r dictates the long-term rate of admitted traffic, and bucket depth p defines the maximum burst amount of admitted traffic. Any incoming data packets, which make the token rate and the bucket depth of admitted traffic exceed r and p, respectively, are not admitted into the router and are dropped by the admission control mechanism. The admission control mechanism in the basic QoS model of a router uses a fixed token rate, r, and a fixed bucket depth, p.

There are two queuing buffers, a high-priority queuing buffer and a low-priority queuing buffer, for the output port of the router to keep admitted data packets before their transmission through the output port.

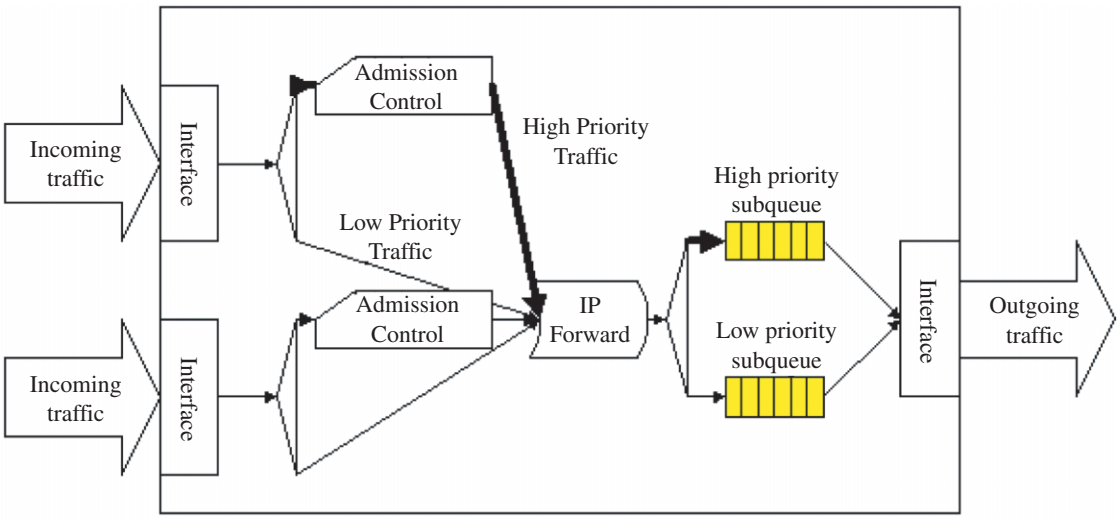

Figure 1. Basic QoS model of a router

The sizes of the two queuing buffers are pre-determined and fixed based on traffic characterization and bandwidth allocation between two classes of data packets. Admitted high-priority data packets are placed into the high-priority queuing buffer to form a queue, and incoming low-priority data packets are placed into the low-priority queuing buffer to form a queue. The capacity of the high-priority queuing buffer is usually small in order to bound the delay of high-priority data packets. The output port first serves data packets in the high-priority queue as long as the queue is not empty. The output port serves data packets in the low-priority queue only when the high-priority queue is empty. Hence, the two separate queuing buffers together provide the mechanism in the basic QoS model of a router to enforce service differentiation between two classes of data packets.

For each queue, the FIFO queuing discipline is applied to determining the order of serving data packets in the queue. If there is not enough space in each queuing buffer to take in a data packet, the data packet is dropped.

QoS model of a router with feedback control

In the basic QoS model of a router, the token rate—a parameter determining the flow rate of admitted high-priority data packets—is pre-determined and fixed according to traffic characterization, router bandwidth capacity and bandwidth allocation between two classes of service. An admitted data packet of high-priority is dropped if there is not enough space in the high-priority queuing buffer to accommodate the data packet. Due to prediction or estimation inaccuracy of traffic characterization, the amount of admitted high-priority data packets may exceed the size of the high-priority queuing buffer, resulting in packets being dropped and transmission delay, which in turn compromises the QoS of high-priority data packets.

To overcome this problem, we add feedback control to the basic QoS model of a router as shown in Figure 2. Note that the feedback control loop is applied to the high-priority queuing buffer only because QoS of low-priority data packets is not of concern and is provided on the best-effort basis. The QoS model of a router monitors the usage state of the high-priority queuing buffer and adjusts the token rate adaptively according to the usage state of the high-priority queuing buffer to prevent packet drop, transmission delay and, thus, compromised QoS. The usage state of the high-priority queuing buffer is measured by the total length of data packets (queue length) in the high-priority queuing buffer relative to the size or capacity of the high-priority queuing buffer. If at any given time we keep the total length of data packets in the high-priority queuing buffer within a certain level (e.g. less than the size of the high-priority queuing buffer by at least the maximum length of one data packet), packet drop and loss will not occur. By setting an upper bound of the queue length,

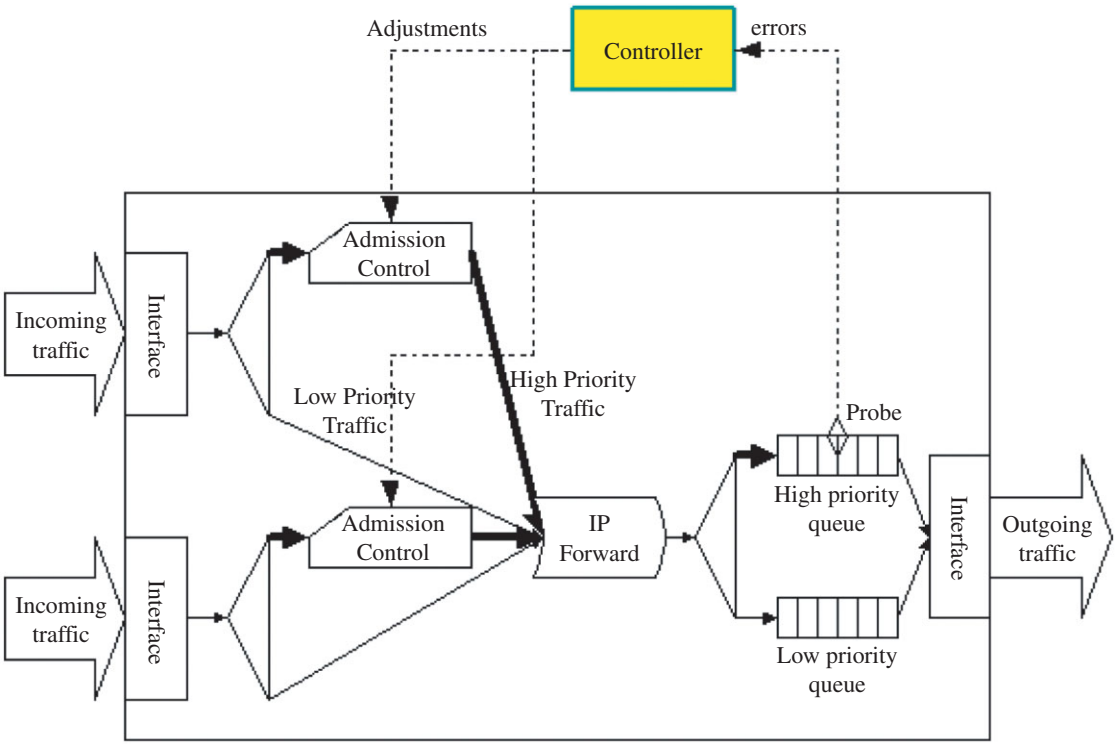

Figure 2. QoS model of a router with feedback control

packet drop and loss can be prevented and transmission delay can be bounded since transmission delay is correlated with the queue length.

The feedback control loop is made up of the probe for monitoring the queue length as the state of the router, the controller and the adaptive admission control. The queue length is constantly monitored and compared with the upper bound. The controller calculates the error or difference between the queue length and the upper bound, and computes the adjustment amount of the token rate. The admission control implements the dynamic adjustment of the token rate.

The error, $e(t)$, at a given time t is calculated as follows:

$$e(t) = l(t) - S(t) \tag{1}$$

where $l(t)$ is the actual queue length and $S(t)$ is the upper bound of the queue length. A positive value of $e(t)$ indicates the over-utilization of the queuing buffer, whereas a negative value of $e(t)$ indicates the under-utilization of the queuing buffer.

The controller uses the basic Proportional–Integral–Differential (PID) control[11] to take the error term, $e(t)$, and calculates the adjustment, $\mu(t)$, as follows:

$$\mu(t) = K_{\mathrm{p}} e(t) + K_{\mathrm{i}} \int_0^t e(t) + K_{\mathrm{d}} \frac{\mathrm{d}e(t)}{\mathrm{d}t} \tag{2}$$

where K_{p}, K_{i} and K_{d} are non-negative constants representing the proportional gain, integral gain and differential gain, respectively. According to formula (2), a negative value of e indicating the under-utilization of the queuing buffer results in a negative value of u, whereas a positive value of e indicating the over-utilization of the queuing buffer results in a positive value of u. Since data packets come from two input ports, the adjustment, u, is split up between the two input ports in proportion to the actual rates of incoming high-priority traffic at

the two input ports. The input port contributing mostly to the increase of the queue length receives the largest adjustment to its token rate for admission control. For the two input ports, the total amount of adjustment, u, is split up between the two input ports as follows:

$$\mu = \mu_x + \mu_y \tag{3}$$

$$\frac{\mu_x}{\mu_y} = \frac{X^*}{Y^*} \tag{4}$$

where μ_x and μ_y are the adjustments allocated for the two input ports, respectively, and X^* and Y^* are the actual rates of incoming high-priority traffic at the two input ports, respectively.

The admission control for each input port implements the dynamic adjustment of the token rate as follows:

$$r_i' = r_i - \mu_i \tag{5}$$

where r_i and r_i' stand for the token rate of input port i at the current time and the next time, respectively. That is, the token rate is increased to allow the admission of more data packets when the queuing buffer is under-utilized with a negative value of μ_y, whereas the token rate is decreased to slow down the rate of admitted data packets when the queuing buffer is over-utilized with a positive value of μ_y. Overall, the feedback control loop aims at maintaining the queue length around the level of the upper bound.

SIMULATION AND EXPERIMENTS

To examine the performance of the QoS model of a router with feedback control in comparison with the performance of the basic QoS model of a router, both models are simulated using OPNET[12] Modeler®. Two versions of the basic QoS model of a router are simulated to emulate over prediction or estimation of traffic characterization (over-model) and under prediction or estimation of traffic characterization (under-model), respectively. The basic QoS under-model sets the token rates for the two input ports over the peak rates of the incoming traffic, whereas the basic QoS over-model sets the token rates under the peak rates of the incoming traffic. In this section, we first describe the simulation of the QoS model of a router with feedback control, the basic QoS over-model and the basic QoS under-model. We then present the simulation experiments to test the performance of these router models.

Simulation

Figure 3 shows the simulation of the QoS model with feedback control. In the simulation, there are two input ports, ports 0 and 1, and one output port, with an IP forwarder module simulating the function of forwarding data packets from the input ports to the output port. Each input port is associated with three traffic sources. A priority-based queuing system is modeled at the output port. The queuing system is made up of a high-priority queuing buffer and a low-priority queuing buffer, each with a limited capacity and the FIFO queuing discipline. A packet sink (labeled 'egress' in Figure 3) is connected with the queuing module to collect the output packets.

Each traffic source generates a stream of data packets, each packet with the mark of its service class, high priority or low priority. The Type-of-Service (ToS) field of the IP header in each data packet is used to mark the service class. The ToS value of 7 (or 111 in binary) represents high-priority traffic, and the ToS value of 0 (or 000 in binary) represents low-priority traffic. For each traffic source, we assume a random Poisson arrival process of data packets. That is, inter-arrival time of data packets has an exponential distribution. For each traffic source, we assume that the size of data packets has a normal distribution. The expected rate of data packets generated from each traffic source, with units of bits per second, can be estimated by the ratio of the mean packet size to the mean inter-arrival time.

Theoretically, the probe in the feedback control loop monitors the queue length continuously. As data packets arrive and leave at the high-priority queuing buffer frequently at discrete time points, the queue length

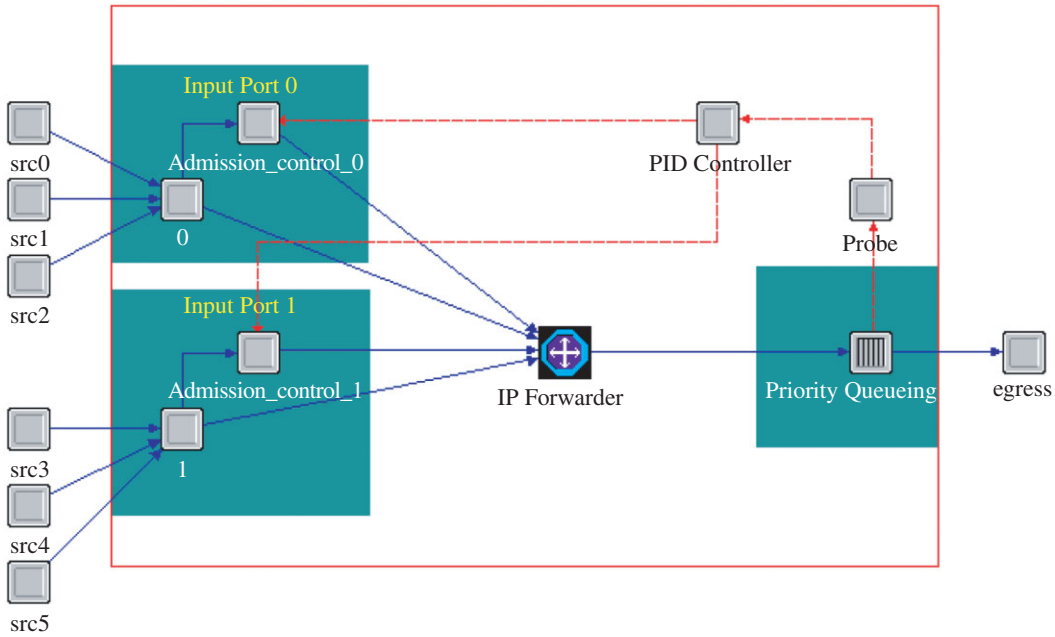

Figure 3. Simulation of the QoS model of a router with feedback control

changes frequently at these discrete time points. The frequent change of the state and consequently the frequent adjustment of the token rate may lead to instability. Hence, instead of getting the state—the queue length—continuously, the queue length is sampled at a certain interval. The maximum value of queue length values in a given interval is taken as the sample value of that interval because the target level is set using the upper bound of the queue length.

For the adaptive admission control, the token rate starts with an initial value of R. Formula (2) is made discrete as follows:

$$\mu = K_{\mathrm{p}} e_k + K_{\mathrm{i}} e_k (T_k - T_{k-1}) + K_{\mathrm{d}} (e_k - 2e_{k-1} + e_{k-2})/(T_k - T_{k-1}) \tag{6}$$

In the above formula, e_k represents the error at time T_k for the kth sampling interval. Formulas (3), (4) and (5) are used to adaptively adjust the token rate for each input port throughout the simulation. The simulation of the basic QoS models uses a fixed token rate throughout the simulation period rather than an adaptively adjusted token rate as in the QoS model with feedback control.

Experiments

We test the performance of the router models under two traffic conditions: heavy traffic and light traffic. Tables I and II give the description of the heavy traffic condition and the light traffic condition, respectively.

The heavy traffic condition introduces an overwhelming amount of high-priority traffic by letting the rate of incoming high-priority traffic exceed the bandwidth capacity of the output port. All the six traffic sources generate data packets whose sizes are determined according to a normal distribution with a mean of 10 000 bits and a variance of 2000 bits. Table I gives the arrival rates of data packets at six traffic sources. Sources 0, 1, 3 and 4 generate high-priority data packets, whereas sources 2 and 5 generate low-priority data packets. Each input port generates high-priority data packets at an average rate of 350 000 bits per second (b/s) and low-priority data packets at an average rate of 150 000 b/s. Hence, the average rate of high-priority data packets arriving at the router from both input ports is 700 000 b/s, which exceeds the bandwidth capacity of the output port at 640 000 b/s.

Table I. Description of heavy traffic condition

Source	Priority	Inter-arrival time probability distribution	Mean (s)	Rate of generated traffic (b/s)
0	High	Exponential	0.040 00	250 000
1	High	Exponential	0.100 00	100 000
2	Low	Exponential	0.066 67	150 000
3	High	Exponential	0.040 00	250 000
4	High	Exponential	0.100 00	100 000
5	Low	Exponential	0.066 67	150 000

Table II. Description of light traffic condition

Source	Priority	Inter-arrival time probability distribution	Mean (s)	Rate of generated traffic (b/s)
0	High	Exponential	0.133 33	75 000
1	High	Exponential	0.133 33	75 000
2	Low	Exponential	0.066 67	150 000
3	High	Exponential	0.133 33	75 000
4	High	Exponential	0.133 33	75 000
5	Low	Exponential	0.066 67	150 000

Table III. The configurations of the basic QoS models and the QoS model with feedback control

	Basic QoS over-model	Basic QoS under-model	QoS model with feedback control
Bandwidth of output port	640 000 b/s	640 000 b/s	640 000 b/s
High-priority queue capacity	100 000 bits	100 000 bits	100 000 bits
Upper bound of queue length	—	—	80 000 bits
Low priority queue capacity	450 000 b/s	450 000 b/s	450 000 b/s
Token rate (input ports 0, 1)	450 000 b/s	250 000 b/s	400 000 b/s (initial value)
Bucket depth (input port 0, 1)	100 000 bits	100 000 bits	100 000 bits
Proportional gain (K_p)	—	—	1.0
Integral gain (K_i)	—	—	0.2
Derivative gain (K_d)	—	—	0.2
Sampling interval	—	—	2 s

Table II gives the description of the light traffic condition. Each input port generates high-priority traffic at a rate of 150 000 b/s and low-priority traffic at a rate of 150 000 b/s. The rate of the total high-priority traffic arriving at the router from both input ports is 300 000 b/s, which is lower than the bandwidth capacity of the output link at 640 000 b/s.

Table III gives the simulation configurations of the basic QoS models and the QoS model with feedback control. In the simulation of the basic QoS over-model, the token rates at both input ports are set to 450 000 b/s, allowing most of the traffic to enter the router. The simulation of the basic QoS under-model sets the token rates at both input ports to 250 000 b/s, which is lower than the average rate of incoming traffic. All other settings of the basic QoS over-model and the basic QoS under-model are exactly the same.

The simulation settings of the QoS model with feedback control (see Table III) are the same as those of the basic QoS models except the variable token rates and additional parameters used in the feedback control loop. The proportional, integral and derivative gains K_p, K_i and K_d used in the PID control of the QoS model with feedback control are selected empirically through three sets of preliminary simulation runs. The selection of these parameters is based on the convergence and oscillation of the token rate. A quick convergence with a modest oscillation is preferable. These preliminary simulation runs are carried out under the heavy traffic

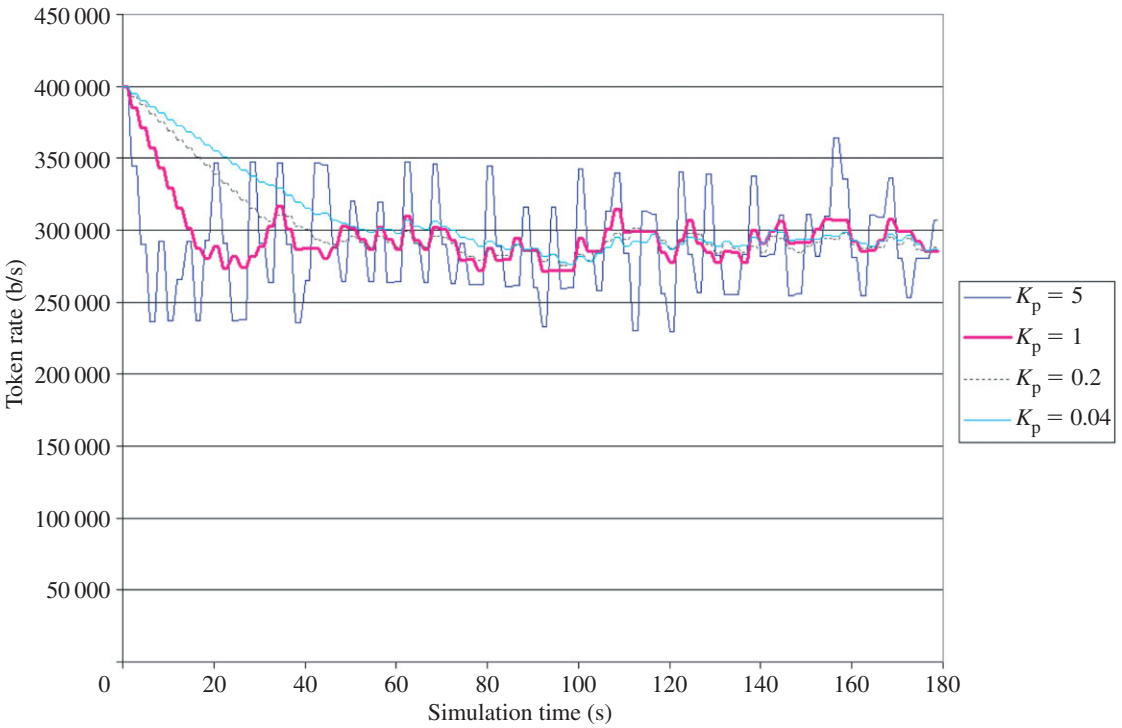

Figure 4. The plot of the token rates for different values of proportional gain

condition using the settings shown in Table III except that different values of K_p, K_i and K_d are tried in the preliminary simulation runs.

A set of four simulation runs is conducted to select K_p, with K_p set to 5, 1, 0.2 and 0.04, and K_i and K_d both set to 0.2. After examining the plot of the token rate as shown in Figure 4, we note that when K_p is equal to 1, the token rate converges quickly to a stable level with a modest oscillation. Hence, we set the value of K_p to 1. Another set of four simulation runs is conducted to determine the integral gain, K_i. K_i is set to 5, 1, 0.2, and 0.04, with K_p fixed at 1 and K_d fixed at 0.2. After examining the plot of the token rate as shown in Figure 5, we note that when K_i is equal to 0.2, the token rate converges quickly and exhibits a modest oscillation. Hence, we set the value of integral gain to 0.2. The derivative gain, K_d, is determined similarly. In a set of simulation runs, K_d is set to 5, 1, 0.2, and 0.04, with K_p equal to 1 and K_i equal to 0.2. After examining the plot of the token rate as shown in Figure 6, we note that when K_d is equal to 0.2, the token rate exhibits a fast convergence and a modest oscillation. Hence, we set the value of derivative gain to 0.2.

The upper bound of the queue length is also determined through a set of preliminary simulation runs. Three simulation runs are conducted with the upper bound set to 90 000, 80 000 and 70 000 bits, respectively, and other parameters using the settings in Table III. The selection of the upper bound for the queue length is based on how it affects the packet loss and throughput of high-priority traffic. The packet loss for the upper bounds of 90 000, 80 000 and 70 000 bits are 232, 107 and 43 packets, respectively. The throughput is plotted in Figure 7. From Figure 7 we observe that there is a trade-off between packet loss and throughput. When the upper bound of the queue length approaches the queue size or capacity of 100 000 bits, both packet loss and throughput increase. When the upper bound is set to 80 000 bits, both packet loss and throughput are moderate. Hence, we take 80 000 bits as the upper bound of the queue length.

After the parameters of the simulation are determined, the simulation experiments of performance testing are conducted with a total of six simulation runs for different combinations of the QoS models and the traffic conditions as shown in Table IV. Each simulation run lasts for 180 s.

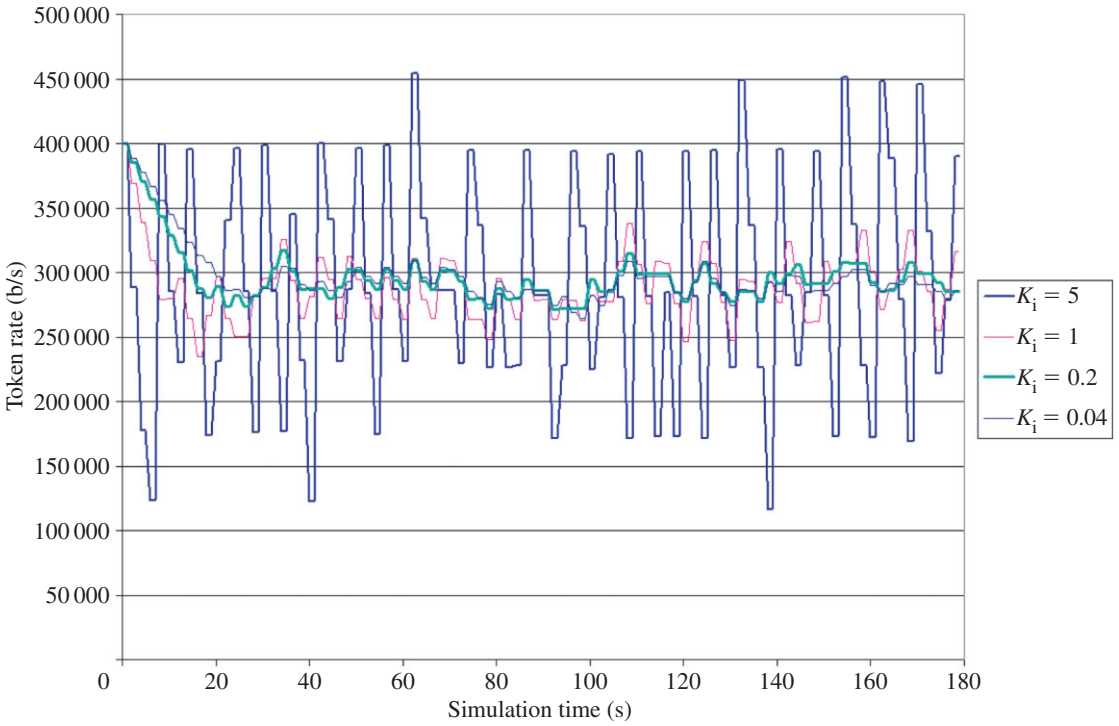

Figure 5. The plot of the token rate for different values of integral gain

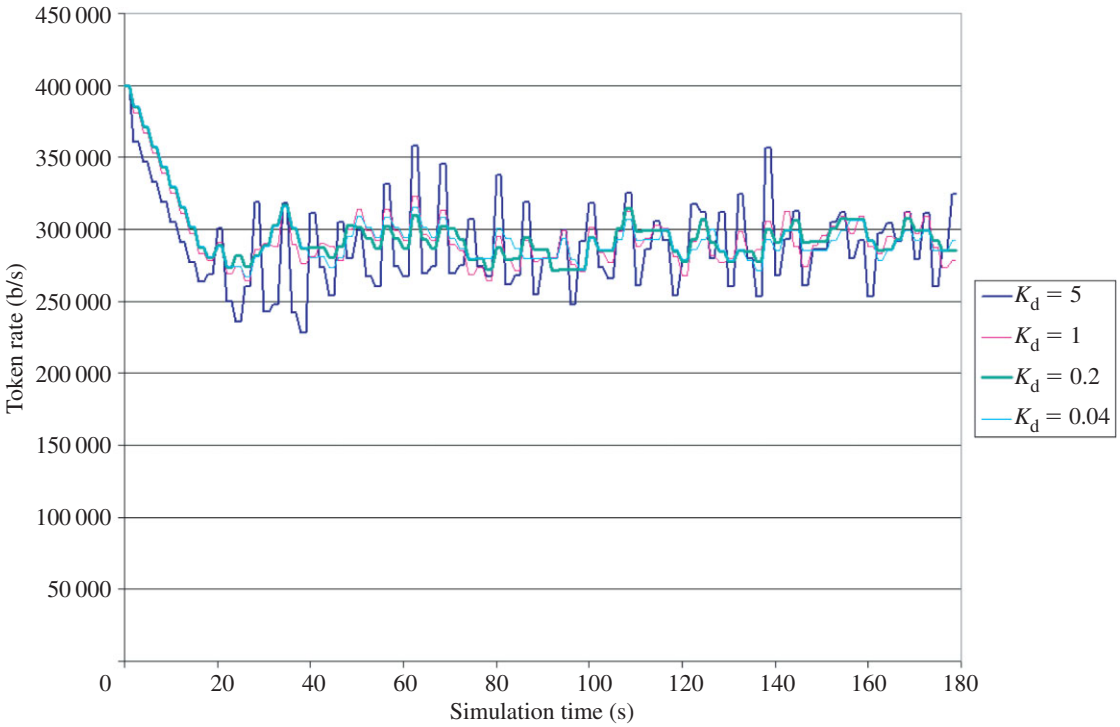

Figure 6. The plot of the token rate with different values of derivative gain

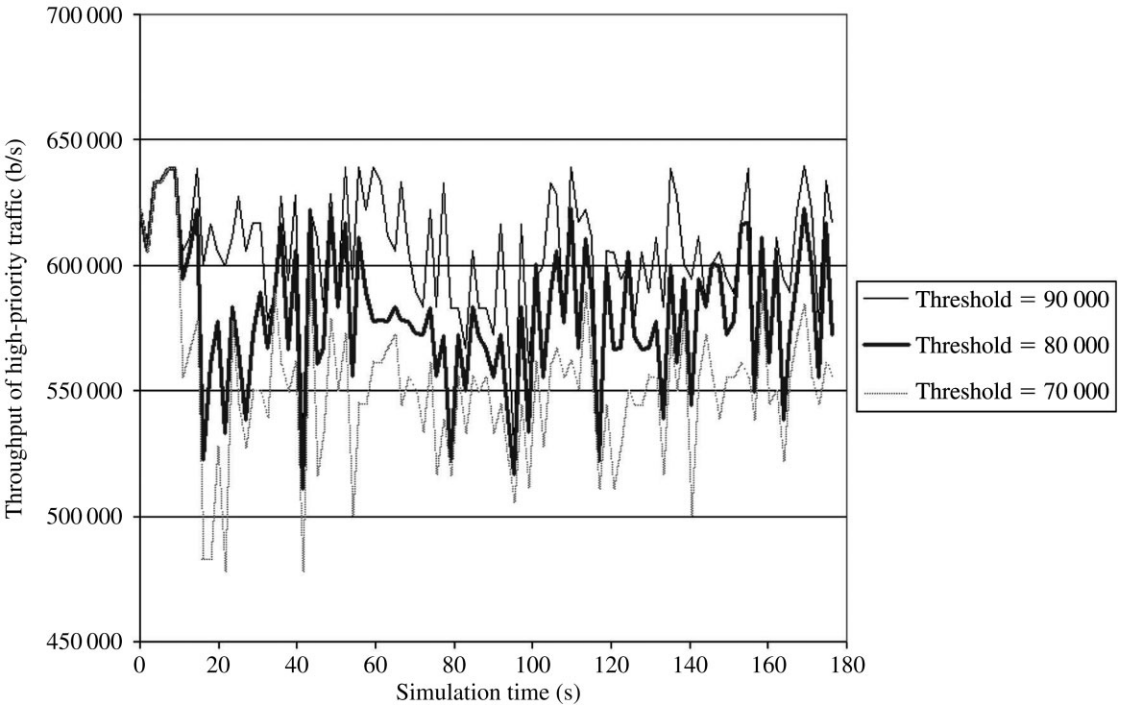

Figure 7. Throughput of high-priority traffic with different upper bounds (thresholds) of the queue length

Table IV. Simulation runs in the experiments

Simulation run	QoS model	Traffic condition
1	Feedback control	Heavy
2	Feedback control	Light
3	Basic over-model	Heavy
4	Basic over-model	Light
5	Basic under-model	Heavy
6	Basic under-model	Light

Performance measures

In general, QoS has three attributes to measure the output performance of a process: timeliness, precision and accuracy[13]. Timeliness measures the time taken to produce the output of the process. Precision measures the amount or quantity of the produced output. Accuracy measures the correctness of the produced output, usually relating to the content of the output. Since the router works at the network layer of TCP/IP and does not deal with the content and its correctness of a data packet, accuracy of transmitting a data packet is not collected during the simulation of the QoS models of routers in this study. Only timeliness and precision of the process of transmitting a data packet through a router are measured during the simulation of the QoS models. Since we are not concerned with QoS of low-priority data packets, we only measure timeliness and precision of the packet transmission process for high-priority data packets.

We use time-in-system to measure the timeliness of the packet transmission process. Time-in-system is the duration from the time of a packet entering the queue to the time of the packet being taken out of the queue. Note that time-in-system is different from packet delay in the router. Packet delay includes time-in-system, transmission time and other processing times in the router. Since transmission time depends on the size of a data

packet and, thus, varies with data packets, we use time-in-system which does not depend on the size of a data packet to measure timelines of serving a data packet. We use packet loss rate and throughput to measure the precision of the packet transmission process. We let the OPNET Modeler collect the time-in-system of each data packet in the high-priority queuing buffer in the 'bucket mode'. In this mode, sampling intervals are used. For the total simulation time of 180 s for each simulation run in this study, OPNET uses 100 sampling intervals with 1.8 s for each interval. For each interval, OPNET averages time-in-system values of all high-priority packets leaving the queue and presents this average as the time-in-system value for that interval. For the entire simulation duration of 180 s, an accumulative count of the lost packets due to packet drop at the queuing buffer is calculated. At the end of each simulation run, the packet loss rate is calculated as the ratio of the lost packet count to the number of admitted packets. In addition to time-in-system and packet loss rate, we also collect the traffic throughput to examine the bandwidth utilization. Throughput is defined by the rate (in b/s) of traffic leaving the router. Throughput values are also collected in the 'bucket mode' with the 1.8 s sampling interval.

RESULTS AND DISCUSSIONS

The QoS model of a router with feedback control is compared with the basic QoS over-model and the basic QoS under-model with regard to time-in-system, packet loss rate and throughput, all measured on high-priority traffic.

For packet loss rate in the heavy traffic condition, the QoS model with feedback control loses a total of 107 packets, accounting for 1% of all admitted high-priority data packets. The basic QoS over-model loses a total of 1299 packets, accounting for 10.3% of admitted high-priority data packets, because the token rate is set over the peak rate of incoming traffic and, thus, allows most traffic entering the queue. The basic QoS under-model router loses no packets because the token rate is set under the peak rate of the incoming traffic and, thus, allows less traffic entering the queue. In the light traffic condition, none of the three QoS models have packet loss.

Figure 8 shows the results of a time-in-system in the heavy traffic condition. The time-in-system of the QoS model with feedback control is smaller than that of the basic QoS over-model and larger than that of the basic QoS under-model. Figure 9 shows the results of a time-in-system in the light traffic condition. The time-in-system of the QoS model with feedback control is slightly higher than those of the basic QoS models.

Figure 10 shows throughput of the QoS models of a router in the heavy traffic condition. The QoS model of a router with feedback control exhibits a better throughput and, thus, a better bandwidth utilization than the basic QoS under-model, but a slightly worse throughput than the basic QoS over-model. The basic QoS over-model almost achieves the full utilization of the bandwidth capacity. In the light traffic condition, the three QoS models achieve very similar throughput performance as shown in Figure 11.

Overall, the over-characterization of traffic in the basic QoS over-model yields more lost packets, longer time-in-system, but better throughput and bandwidth utilization, whereas the under-characterization of traffic in the basic QoS under-model yields no packet loss, shorter time-in-system, but worse throughput and bandwidth utilization. The QoS model of feedback control achieves a better balance between the time-in-system, packet loss and throughput through monitoring the queue length and adaptively adjusting the admission control, in comparison with the basic QoS over-model and the basic QoS under-model. In the light traffic condition, the three QoS models demonstrate similar performances with regard to time-in-system, packet loss and throughput.

CONCLUSIONS

In the heavy traffic condition, the QoS model of a router with feedback control improves both the time-in-system and packet loss through the monitoring of the queue length and the adaptive admission control in comparison to the basic QoS over-model with static admission control using high admission rate. The QoS model of a router with feedback control also improves throughput and bandwidth utilization in the heavy traffic condition in comparison to the basic QoS under-model with static admission control using low admission rate. Obviously, there is a tradeoff between resource utilization and the possibility of packet loss and increased

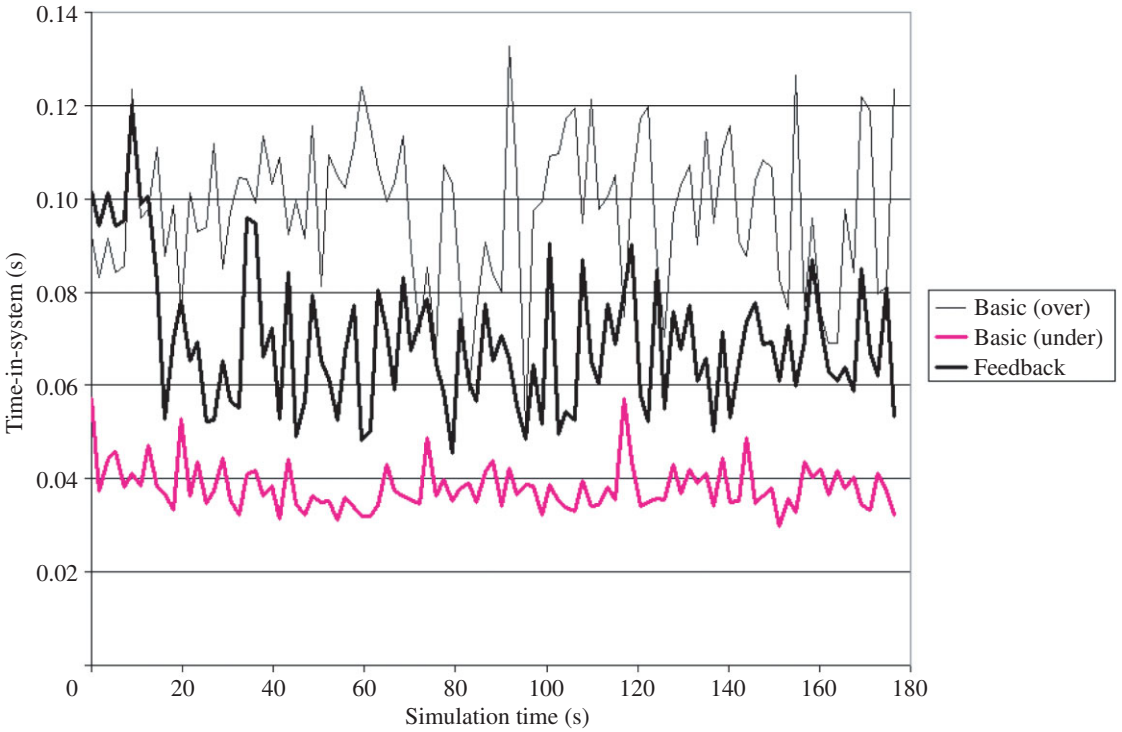

Figure 8. Time-in-system of the QoS models of router in the heavy traffic condition

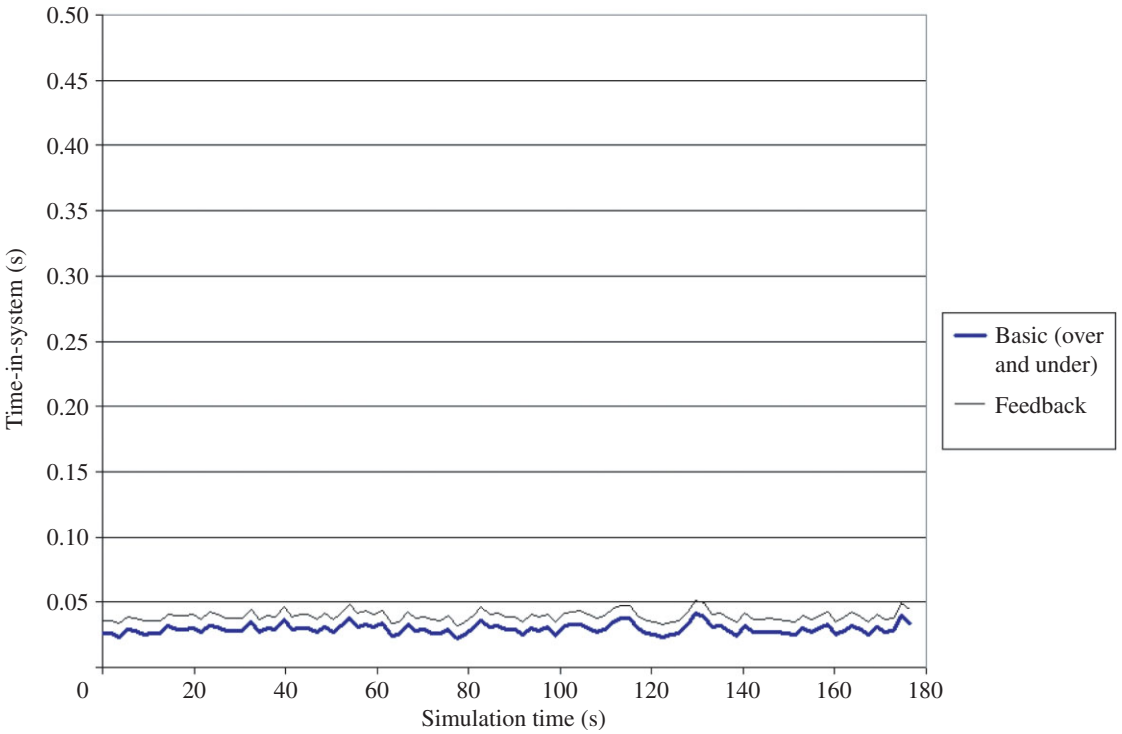

Figure 9. Time-in-system of the QoS models of router in the light traffic condition

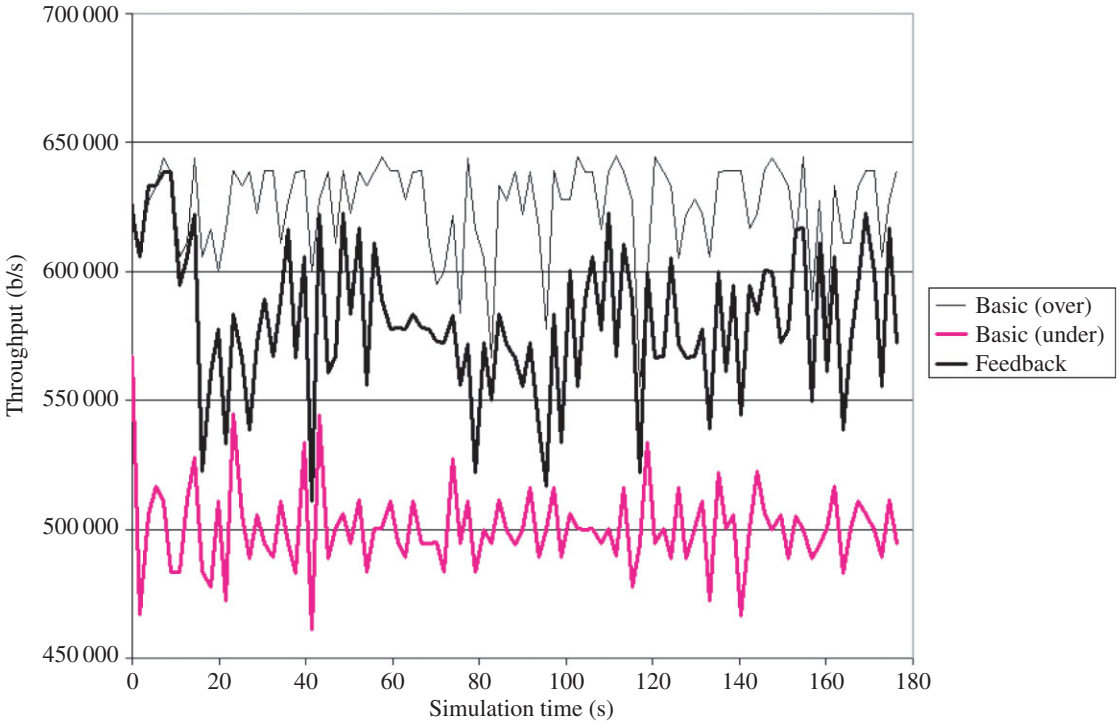

Figure 10. Throughput of the QoS models of router in the heavy traffic condition

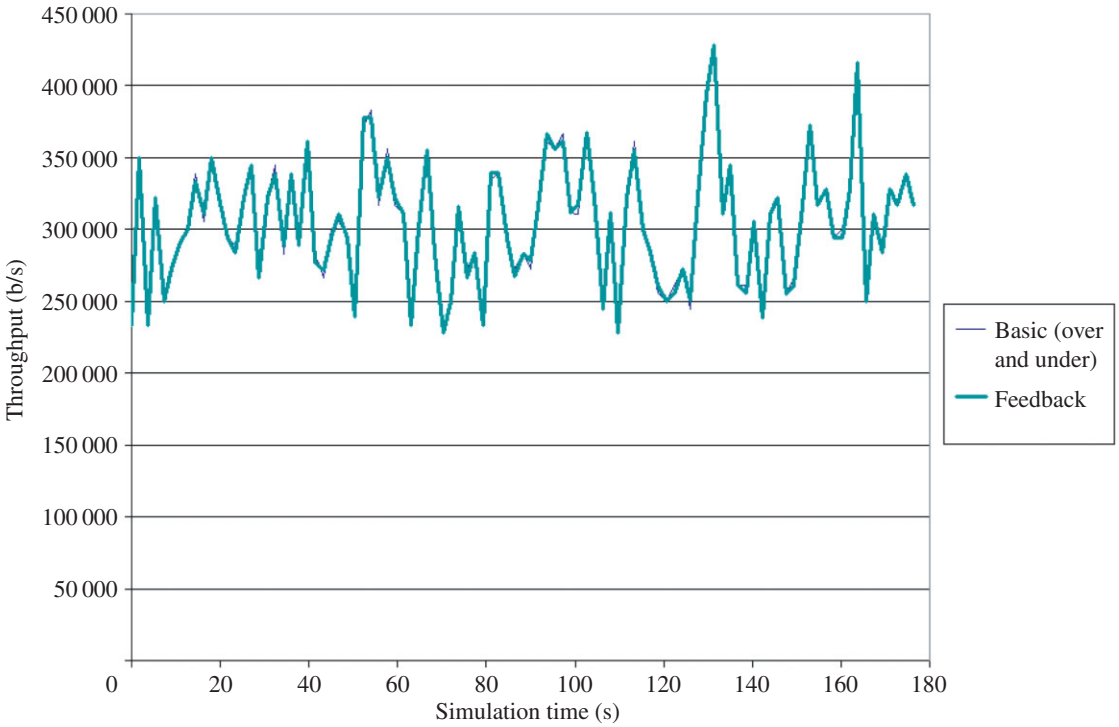

Figure 11. Throughput of the QoS models of router in the light traffic condition

time-in-system. With static admission control, the more the bandwidth is utilized, the greater the possibility of losing packets and the longer the time-in-system. Adaptive admission control balances the need for high resource utilization and QoS assurance with regard to timeliness and precision while maintaining reasonable resource utilization. Moreover, it is challenging or impractical to get an accurate prediction or estimation for traffic characterization. Feedback control with adaptive admission control allows us to overcome prediction or estimation inaccuracy in traffic characterizations and have a practical solution to providing QoS while maintaining reasonable resource utilization in the router. This study also shows that in the light traffic condition the different QoS models investigated in this study demonstrate similar performance of timeliness, precision and resource utilization.

Acknowledgements

This work is sponsored by the Air Force Research Laboratory—Rome (AFRL-Rome) under grant number F30602-01-1-0510. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of AFRL-Rome or the U.S. Government.

REFERENCES

1. Huston G. *Internet Performance Survival Guide*. Wiley: New York, 2000; 9.
2. Almquist P. Type of service in the Internet Protocol suite. *Request for Comments 1349*, Internet Engineering Task Force, July 1992. Available at: http://www.ietf.org/rfc.html.
3. Blake S, Black D, Carlson M, Davies E, Wang Z, Weiss W. An architecture for differentiated service. *Request for Comments (Informational) 2475*, Internet Engineering Task Force, December 1998. Available at: http://www.ietf.org/rfc.html.
4. Braden R, Clark D, Shenker S. Integrated services in the Internet architecture: An overview. *Request For Comments (Informational) 1633*, Internet Engineering Task Force, June 1994. Available at: http://www.ietf.org/rfc.html.
5. Gevros P, Crowcorft J, Kirstein P, Bhatti S. Congestion control mechanisms and the best effort service model. *IEEE Network* 2001; **15**(3):16–26.
6. Kurose J. Open issues and challenges in providing Quality-of-Service guarantees in high-speed networks. *ACM SIGCOMM Computer Communication Review* 1993; **23**(1):6–15.
7. Nicols K, Jacobson V, Zhang L. A two-bit differentiated services architecture for the Internet. *Request for Comments 2638*, Internet Engineering Task Force, November 1997.
8. Parekh AK, Gallager PG. A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Transactions on Networking* 1993; **1**(3):344–357.
9. Sabata AB, Chatterjee Davis SM, Sydir JJ, Lawrence TF. Taxonomy of QoS specifications. *Proceedings of 3rd Workshop on Object-Oriented Real-Time Dependable Systems*, 1997; 100–107. Available at: http://www.ietf.org/rfc.html.
10. Zhang H. Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE* 1995; **83**(10):1374–1396.
11. Harrison HL, Bollinger JG. *Introduction to Automatic Controls* (2nd edn). Harper and Row: New York, 1969; 159–182.
12. OPNET Technologies, Inc. *Modeler Tutorial, Part Number D00142*, OPNET Technologies, Inc., 2001.
13. Ye N. QoS-centric stateful resource management in information systems. *Information Systems Frontiers* 2002; **4**(2):149–160.

Authors' biographies

Zhibin Yang graduated from the Department of Industrial Engineering at Arizona State University with a MS degree.

Nong Ye is a Professor of Industrial Engineering and an Affiliated Professor of Computer Science and Engineering at Arizona State University. She holds a PhD degree in Industrial Engineering from Purdue

University, a MS degree in Computer Science from the Chinese Academy of Sciences, and a BS degree in Computer Science from Peking University. Her research interest is in information and systems assurance. She is an Associate Editor for *IEEE Transactions on Reliability*, and *IEEE Transactions on Systems, Man, and Cybernetics—Part A*, *Information, Knowledge, Systems Management*. She is a senior member of the Institute of Industrial Engineers and a senior member of IEEE.

Ying-Cheng Lai received BS and MS degrees in Optical Engineering from Zhejiang University in 1982 and 1985, and MS and PhD degrees in Physics from the University of Maryland at College Park in 1989 and 1992, respectively. He joined the University of Kansas in 1994 as an Assistant Professor of Physics and Mathematics and became Associate Professor in 1998. In August 1999, he came to Arizona State University as an Associate Professor in the Departments of Mathematics and Electrical Engineering. He was promoted to full Professor of Mathematics and of Electrical Engineering in August 2001.