1998 International Computer Symposium
Workshop on Computer Networks, Internet, and Multimedia
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

GUARANTEED VERSUS CONTROLLED LOAD: IMPLICATIONS FOR SERVICE SUBSCRIBERS AND PROVIDERS IN RSVP NETWORKS

Ying-Dar Lin and Chih-Yu Chen
Department of Computer and Information Science
National Chiao Tung University
Hsinchu, Taiwan
Tel: +886-3-5731899
Fax: +886-3-5721490
Email: ydlin@cis.nctu.edu.tw

Abstract— **IETF Integrated Service Working Group has specified two service classes: Guaranteed Quality (GQ) service and Controlled Load (CL) service. What concerns service subscribers and providers most is the cost of these two services and their performance. For service subscribers, the question is which application deserves which service. For service providers, the question is how to charge their users reasonably to obtain the maximum revenue and what kinds of mechanisms can achieve better resource utilization. In this paper, we try to answer the above questions under conservative and well-performed admission control schemes, respectively. Simulation results based on the common models of traffic, service, policing, and scheduling are presented. When the traffic burstiness increases, the cost difference between GQ and CL increases significantly but the average performances do not have much difference. Thus, subscribers are suggested to use the CL service when the traffic burstiness is high and the delay bound is not critical, and vice versa. For providers, a well-performed admission control scheme is important, especially when the traffic burstiness is high, in limiting the cost difference between GQ and CL. It is observed that, with well-performed admission control, the cost difference can be reduced from 20 times to 1.41 times and 8 times to 1.14 times for bursty and less-bursty traffic, respectively.**

keywords: Integrated Service, Guaranteed Quality, Controlled Load, RSVP, admission control, service provisioning

I. INTRODUCTION

In today's Internet, only a single class of service, i.e. best-effort service, is supported. With best-effort service, every network element does its best to transmit packets but not much is guaranteed. However, the best-effort service may not meet the requirements of some applications. Thus, quality-of-service (QoS) enabled Internet should be developed. There are two important elements for QoS: (1) a signaling protocol and (2) an efficient scheduling algorithm. A signaling protocol provides a mechanism for applications and network elements to exchange the information and requirement of QoS, and an efficient scheduling algorithm is necessary for network elements to enforce the requirement when transmitting packets. The Resource Reservation Setup Protocol Working Group (rsvp) and the Integrated Service Working Group (intserv) formed by Internet Engineering Task Force (IETF) has put a resource reservation protocol, RSVP [1], the standard signaling protocol and defined two services, Guaranteed Quality (GQ) service [2] and Controlled Load (CL) service [3] for individual packet flows.

The GQ service guarantees that packets will arrive within the desired delivery time and will not be discarded due to buffer overflow. This service is intended for applications which have firm delay bounds of transmitting packets from sources to destinations. It is worth noting that the GQ service concerns only the *largest* packet delay and ensures no packet loss. It does not attempt to guarantee the jitter and the average delay. The GQ service is invoked by specifying the traffic (*Tspec*) and the desired requirement (*Rspec*) to the network element. The CL service is invoked by specifying the traffic parameters (*Tspec*) to the network element. There is no *Rspec* specified for the CL service.

Tspec has 5 parameters [2], [3]: (1) token bucket rate, r, (2) token bucket depth, b, (3) traffic peak rate, p, (4) minimum policed unit, m, and (5) maximum packet size, M. The parameters of a token bucket describe the traffic and are also used in the policing module. m is given to facilitate resource allocation and policing. Any packets with size smaller than m are treated as having size m. This parameter is given to allow reasonable estimation of the per-packet resources needed to process a packet flow. M is the largest packet size that the source may have. *Rspec* has a pair of parameters [2]: a desired reservation rate, R, and a slack term, S. R is calculated, by clients, according to the desired delay bound. The slack term, S, also given by clients, tolerates the network elements to reduce their resource reservation for this flow. If any network element utilizes the slack term, it should update the *Rspec* to keep the delay bound constraint. The specification emphasizes that the buffer size is not specified in the *Rspec* because the network element is expected to derive the required buffer size to ensure no packet loss. A network element implementing the GQ service exports two error

terms, C and D, which represent how the implementation deviates from the ideal model. The error term C is rate-dependent. An example is the extra time spent on breaking a packet into several ATM cells. The term D is a rate-independent, per-element error term. A typical example is, in a slotted network, the maximum waiting time from once a packet is ready to be transmitted until it catches a slot. Considering the error terms, a network element supporting the GQ service should ensure the maximum delay $Delay_{max} = \frac{b}{R} + \frac{C}{R} + D$ [2].

In summary, a network element supporting the GQ service can provide firm guarantee on maximum queueing delay and zero packet loss. It is noticeable that the GQ service gives applications considerable control over their delays by tuning the *Rspec*. The CL service provides a virtually unloaded condition to applications even the real condition might be heavy load.

For customers, what concerns them most is how much they should pay to subscribe a service and its performance. Does it pay to use the GQ service rather than the CL service? What kinds of applications deserve the GQ service? For service providers, the cost to provide a service determines their billing policies and revenues. What kinds of mechanisms for admission control and packet scheduling can obtain better resource utilization and, in turn, cheaper service provisioning? What are the reasonable billing policies for the GQ and CL services in order to attract the subscribers for both of them? Previous studies [4] have shown that the link utilization can be, for example, 7% or 74%, respectively, when the link has admitted as many GQ or CL flows as it can. That means the price for the GQ service can be, in this example, ten times more expensive than the CL service's. Subscribers would be unwilling to use the GQ service if the CL service can offer a tolerable performance.

In this paper, we try to answer the questions raised above and investigate the possibilities of having cost-effective provisioning of both GQ and CL services. In section2, we describe the video, voice, and data traffic models and their corresponding *Tspec* and *Rspec* parameters. These are used to drive our simulation study. Models of policers and schedulers are defined in section 3. A standard token bucket and a cyclic implementation of weighted fair queueing (WFQ) [5] are adopted. We present our numerical results, for homogeneous and heterogeneous traffic-service scenarios, and give implications, for service subscribers and providers, in section 4.

II. Characterizing Traffic and Services

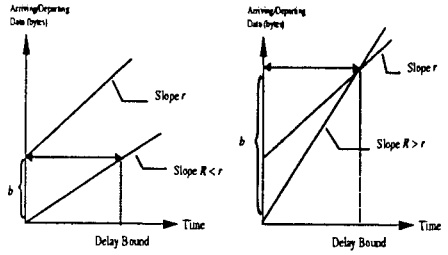
To measure the cost of the GQ and CL services, traffic models and their service parameters should be defined first. Three types of traffic models, video, voice and data, are discussed here. Tailored according to the traffic models, the *Tspec* for each model can be determined. In addition, *Rspec* should be specified if the GQ service is used. The most important parameter in *Rspec* is the request bandwidth, R. We do not use the slack term, S. As mentioned, R can be determined by the required delay bound and the token bucket depth parameter, b, in *Tspec*. In general, the delay bound can be adjusted according to the user's requirements. However, if the traffic flow traverses through IEEE 802 LANs, the draft of *Integrated Services over IEEE 802.1D/802.1p Networks* [6] provides only two classes of GQ service with bounds of 100ms and 10ms. The delay bound of 10ms is adopted in this study.

However, deriving R under the fixed delay bound without considering the average rate r may lead to a mistake shown in Figure 1(a). It is obviously necessary that R should be larger than r. However, if the token depth, b, is too small, the derived R might not meet this constraint. One solution is to reduce the delay bound. However, there are two drawbacks: (1) it will constrain the network elements too much and (2) the request may experience a higher rejection probability by the admission control. Hence, another solution is that the application provides a larger b in order to derive a reasonable request bandwidth R, as illustrated in Figure 1(b). In Table 1, we define *Tspec* and *Rspec* for video, voice and data traffic models. The rationale is explained below.

A. Video Traffic Parameters

One of the most commonly used video compression standards is the MPEG standard defined by ISO Motion Picture Expert Group [7]. Traditionally, an MPEG stream is viewed as a variable bit rate (VBR) traffic with a high peak rate relative to its average rate. There are some previous efforts to model the MPEG streams [8], [9]. These models provide an approximation to the real MPEG traffic streams and they help to design the mechanisms to handle the VBR traffic. However, some other efforts try to "smooth" the MPEG streams to be constant bit rate (CBR) traffic [10]. The main issue of the MPEG smoothing is to defer some packets in the stream without violating the deadline of each frame. It then becomes much easier, for network elements, to handle the CBR traffic, i.e. the smoothed stream, than the VBR traffic, i.e. the original stream. Hence, it is reasonable and convenient for us to use the CBR model to simulate the video traffic. According to this model, the token rate r, i.e. the average traffic rate, can be determined as 1.5 Mbps. The peak rate, p, is also defined as 1.5 Mbps. Two other parameters in *Tspec* are about the packet size. Following the specification of the MPEG standard [7], a fixed packet size of 188 bytes is determined. Hence, m and M are both defined as 188 bytes.

As mentioned above, b can not be determined without considering its impact on R. If b is defined just as the packet size to indicate that there is no bursty period, it results in an unreasonably low R ($(188 * 8bits)/(10ms) = 0.15Mbps$). In this case, R should be pre-defined as a rate a little higher than r, say



(a) R<r (b) R>r

Fig. 1. (a) Mis-derivation of R and (b) enlargement of b

Parameters / Traffic	r (bits/s)	b (bytes)	p (bits/s)	m (bytes)	M (bytes)	R (bits/s)
Video	1.5M	2000	1.5M	188	188	1.6M
Voice	25.6k	640	64k	64	64	512k
Data	1M	10000	13M	500	1500	8M

Table 1. *Tspec* and *Rspec* for video, voice, and data traffic models

1.6 Mbps, and b is reversely derived as 2000 bytes ($(1.6Mbps) * (10ms) = 2000bytes$).

B. Voice Traffic Parameters

The most commonly used coding standard for voice is pulse code modulation (PCM) [11] finalized in 1972. It samples the original analog voice source 8000 times per second and produces an 8-bit digital signal for each sample. The bit rate of PCM is $8000(sample/s) * 8(bit/sample) = 64kb/s$. In addition to the codec, the model of the speech activity is needed. In general, the speech activity can be modeled by a two-state source. One state represents the silent period where the source produces zero bit rate and another one represents the talk period where the source transmits at the peak rate. This two state model is named ON-OFF model accordingly. Both the silent and talk spurt durations are exponentially distributed with mean λ and α, respectively. The active ratio is defined as $\frac{\alpha}{\lambda+\alpha}$. Statistically, the active ratio is 0.4 [12], and the mean lengths of silence and talk spurt periods are 0.6 and 0.4 seconds, respectively. With the PCM codec of bit rate 64 kbps and silent removal, but without compression, the average rate is $64kbps * 0.4 = 25.6kbps$. The peak rate, which occurs during the talk spurt period, is 64 kbps. Fixed packet size of 64 bytes is considered here instead of forming a packet for each sample containing only 1 byte. This prevents a large overhead and also meets the minimum packet size requirement of Ethernet, the most popular LAN technology. R is 512 kbps ($640(bytes)/10(ms) = 512kb/s$) when a token bucket depth of 640 bytes is used.

C. Data Traffic Parameters

The characteristics of various data traffic is quite different. Hence, it is difficult to use a general model for all types of data traffic. However, in general, the Poisson model for packet arrivals is used for the convenience of analysis. Another important parameter is the packet size. An approximation is the Gaussian distribution, also known as the normal distribution. Combining the exponential inter-arrival time and the Gaussian distribution packet size, a model for data traffic is made. The same model is also used in [13].

With this model, an average rate of 1 Mbps is assumed. The peak rate, 13 Mbps, is determined by simulating the behavior of this model and capturing the maximum instant traffic rate. The packet size has a mean of 1000 bytes, a minimum of 500 bytes, and a maximum of 1500 bytes. The token bucket depth of 10 times of the average packet size, i.e. 10000 bytes, is defined. The request bandwidth, R, is calculated to be 8 Mbps ($10000(bytes)/10(ms) = 8Mb/s$).

III. Policing and Scheduling Packet Flows

A. IP vs. ATM

One of the major differences between IP and ATM is whether the packet size is fixed or not. In the ATM environment, the network elements can handle traffic in the unit of 53-byte "cell". But the network elements in the IP environment should face the variable length packets. This difference has impacts on the QoS control modules. Taking the output link scheduler as an example, ATM switches can allocate the time to each flow in the unit of the transmission time of a single cell. However, in the IP environment, if the scheduler allocates a period to a flow, it may happen that the allocated period terminates in the middle of packet transmission, as shown in Figure 2. There are two strategies for the scheduler to handle this case: credit or debt. The credit strategy means that the last packet will not be transmitted if its finish time exceeds the end of the allocated period. But the un-used time, ct, should be returned to the flow in the future allocation periods. On the other hand, the debt strategy allows the packet to be transmitted but the over-used time, dt, should be deducted in the future allocation periods. According to our simulation results, there is almost no difference between these two strategies. Hence, it is fine to choose any strategy, and the choice is necessary if a framing, or cyclic, scheduling algorithm is adopted in the IP network.

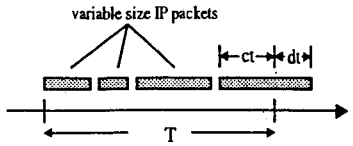


Fig. 2. An example of credit-debt strategy

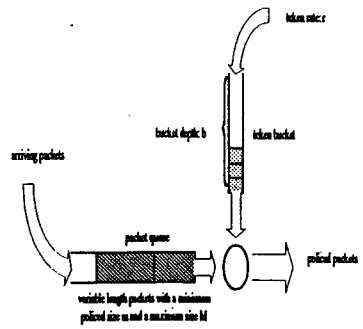


Fig. 3. A standard token bucket

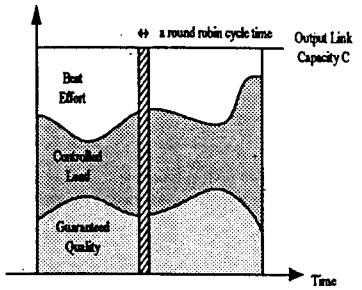


Fig. 4. Bandwidth sharing between different QoS classes

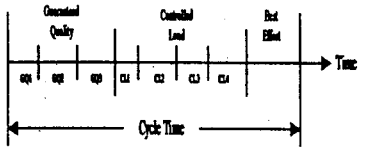


Fig. 5. Bandwidth sharing between flows of various classes within a cycle

B. Policing Model

Policing is necessary for both GQ and CL services. A policing mechanism enforces the flow to follow the *Tspec* it declares. Because *Tspec* uses a token bucket to describe the traffic, it is natural to use a token bucket as the policing mechanism. Figure 3 shows a standard token bucket with a token bucket and a packet queue. Two parameters of a token bucket are r and b. Because the packets are counted in the unit of *bytes*, the token rate and bucket depth are also measured in the unit of bytes per second and bytes, respectively. The token bucket stores tokens with a constant rate r (i.e. one byte per $1/r$ second) and a maximum storage b. It can be viewed that the average rate of the traffic flow is also r. The parameter b is equivalent to the maximum allowed burst length of traffic. Through the token bucket, the traffic is reshaped to fit the *Tspec* and the amount of data sent during a time period t cannot exceed $r * t + b$.

However, to process the policing module every $1/r$ second causes a heavy load. A virtual clock approach, suitable for both hardware and software implementations, presented here may be used to avoid the overhead of implementing a "real" token bucket. For each flow, two additional variables are needed. One variable, denoted as t_p, is used to record the time of last packet departing the policing module. Another one, denoted as b_p, is used to record the token bucket depth at time t_p. When a new packet arrives, the number of available tokens, b_n, can be calculated as

$$b_n = min((\lfloor t_n * r \rfloor - \lfloor t_p * r \rfloor + b_p), b),$$

where t_n represents the time the new packet arrives. Then, a comparison between the size of the arriving packet, denoted as s, and the number of tokens is made. b_p and t_p for the flow are modified as

$$b_p = \begin{cases} b_n - s & \text{if s} < b_n, \\ 0 & \text{if s} \geq b_n, \end{cases}$$

and

$$t_p = \begin{cases} t_n & \text{if s} < b_n, \\ \lfloor t_n * r \rfloor / r + (s - b_n)/r & \text{if s} \geq b_n, \end{cases}$$

where $\lfloor t_n * r \rfloor / r$, instead of t_n, is used in order to obtain an integral number of $\frac{1}{r}$.

C. Scheduling Model

Figure 4 shows the bandwidth sharing between different QoS classes in a round robin fashion. The link bandwidth is shared between GQ, CL, and Best-Effort flows. The actual bandwidth used by a class may vary with time. The shaded vertical bar in Figure 4 represents bandwidth usage within a round robin cycle time. Figure 5 shows the detailed bandwidth allocation for flows of various classes within a cycle. During a cycle time, each GQ and CL flow is allocated a time period to transmit. The length of each period is determined by the reserved bandwidth for each flow. The remaining time in a cycle is used to serve the Best-Effort flows in a first-come-first-serve order.

The structure of a per-flow scheduler is shown in Figure 6 where the GQ and CL flows have per-flow queues and the Best-Effort flows are multiplexed into one queue. The scheduler is responsible for selecting

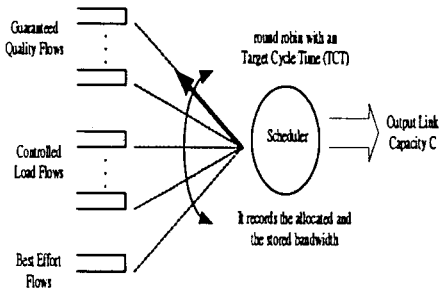


Fig. 6. Scheduler: a packetized cyclic implementation of weighted fair queueing

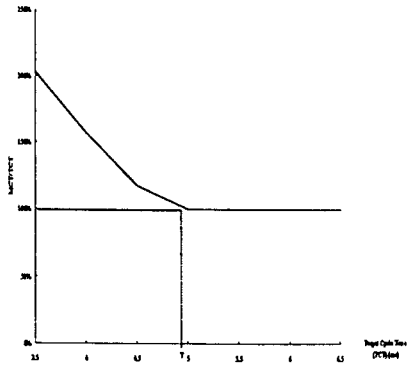


Fig. 7. The applicable minimum TCT to approximate weighted fair queueing while avoiding instability

the queues to serve, according to the bandwidth allocated to each flow. For the round robin scheduling, the first parameter is the cycle time, called Target Cycle Time (TCT). The TCT determines how often a flow can be served. The time periods are then allocated as

$$\frac{R}{C} * TCT \quad for\ each\ GQ\ flow\ and$$
$$\frac{r}{C} * TCT \quad for\ each\ CL\ flow,$$

where C is the output link capacity, R is the request bandwidth in *Rspec*, and r is the token rate in *Tspec*. However, the reserved time for a flow may not be used because of no more traffic, and the scheduler should "remember" the remaining time and add it to the allocated period of the next cycle. A flow, in a sense, stores some reserved bandwidth to handle its future bursts. If too many flows are in bursty period in the same time, the actual cycle time (CT) may exceed TCT. It may not violate the QoS requirement of each flow when this condition happens, but it puts the system in an unstable and unpredictable situation. Thus, the determination of the TCT is important.

Our scheduling algorithm, based on a round robin mechanism, can be viewed as a packetized cyclic implementation of weighted fair queueing (WFQ). WFQ, proposed in [5] and analyzed in [14], is based on a weighted bit-by-bit round robin mechanism. In WFQ, flows sharing an output link are transmitted as if they were serviced in a bit-by-bit, round robin order. The "weighted" means that some flows can send more bits than the others in a cycle. However, in a packetized network, a packet should be transmitted as a contiguous one. The scheduler cannot send a bit for the packet and come back later to send the next bit. Thus, to approximate WFQ as close as possible, in addition to remember the allocated bandwidth of each flow, an approach is to reduce the TCT. However, cutting the TCT too short results in the frequent occurrences of having the CT exceeding the TCT. A maximum cycle time (MCT) is defined as a measure during a simulation period with a given TCT. The ratio MCT/TCT can be used to measure the stability of the scheduler. Figure 7 shows the relationship, from our simulations, between the MCT/TCT and TCT when the output link utilization is 70% . T, the minimum applicable TCT for having a stable scheduler, is given as

$$T = \frac{\max_{i \in F}(M_i) * n}{C},$$

where C is the link bandwidth, M_i is the maximum packet size for flow i, F is the set of QoS flows (i.e. GQ and CL flows), and n is the size of F. This formula means that approximately every QoS flows can send one packet in a cycle. However, in order to prevent the impact of TCT on the performance, the simulations in this study adopt a fixed TCT, $5ms$.

IV. Cost and Performance of Various Traffic-Service Scenarios

Now we are ready to report our simulation results which may help service subscribers to make their decision in choosing GQ or CL service for their applications, and provide information to service providers about effective provisioning of the GQ and CL services. It may also help to decide a reasonable billing policy and estimate the revenue. The performance is measured by the requirements of different services, i.e. maximum delay bound for the GQ service and average delay for the CL service, the output link utilization, and the number of admissible flows.

A conservative computation-based admission control computes the number of admissible flows by allocating a fixed bandwidth to each flow, but a well-performed measurement-based admission control [4] may accept as many flows as possible and drive the link utilization to a higher level without violating the QoS requirements. The cost of a flow is defined as the total cost of the link divided by the maximum number of admissible flows. Assume that the cost of a link with a capacity of 1 Mbps is 1 dollar. If a 45 Mbps link can admit up to 29 flows, the cost per flow is thus 1.55 dollars. The maximum number of admissible flows for GQ and CL are different. We compare the cost in providing the GQ and CL services. The influence on cost of GQ and CL services, when the admission control is changed from the conservative computation-based to the measurement-based, is examined. We also compare the performance

degradation of different traffic-service pairs when the admission control has admitted too many flows.

A. Homogeneous Traffic-Service Scenarios

The first study is to measure the cost and performance when the system is filled with the same kind of traffic flows requesting the same service type. A T3 transmission media with a bandwidth of 45 Mbps is adopted in this simulation for video and data traffic. However, considering the simulation scalability, a T1 with 1.5 Mbps is adopted for the low-bit rate voice traffic.

In Figure 8–10, we present the delay behavior of video, voice and data traffic under the GQ and CL services driven by a well-performed measurement-based admission control. Figure 8 shows the behavior of video traffic, which is defined to have a constant bit rate, receiving the GQ and CL services. A delay bound of $10ms$ is set for the GQ service. It is observed that the maximum number of allowable flows for both GQ and CL is 29. The maximum utilization approximates to 96.5% . According to the cost formula, the cost is 1.55 for both services. For service providers, the costs to provide these two services for the constant bit rate traffic are similar and should be reflected to the price.

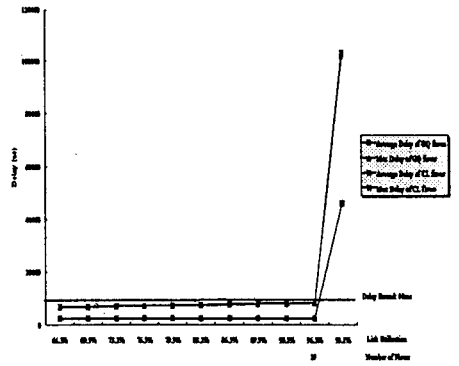


Fig. 8. Behavior of video traffic receiving GQ or CL service

As Figure 9 shows, the behavior of voice traffic is not as good due to the bursty nature of the On-Off model. For the GQ service, the number of allowable flows is 23 with link utilization of 32.8%, and the cost is 0.065. For the CL service, the number is 32 with utilization of 44.7%, and the cost is thus 0.04.

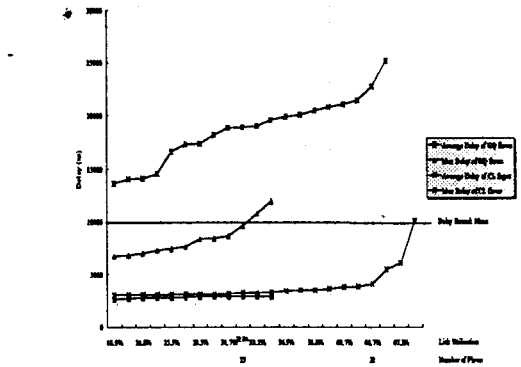


Fig. 9. Behavior of voice traffic receiving GQ or CL service

An interesting comparison can be made between different admission control mechanisms. If the service provider wants to make the same revenue for providing different service types, by running a conservative computation-based admission control, the following condition must hold:

$$\frac{C}{512kbps} * X_1 = \frac{C}{25.6kbps} * X_2,$$

where 512 kbps and 25.6 kbps are given according to Table 1 and the allocation method in section 3.C, X_1 is the cost for each GQ flow, X_2 is the cost for each CL flow, and C is the link capacity. It is derived that the cost of GQ is 20 times larger than the cost of CL. However, if a well-performed measurement-based admission control is used, the formula becomes

$$23 * X_1 = 32 * X_2,$$

where 23 and 32 are taken from Figure 9. The cost ratio drops dramatically to 1.41. The cost of the GQ service is not necessarily so expensive when compared with the CL service.

Figure 10 shows the behavior of data traffic. The number of allowable flows to meet the maximum delay bound for the GQ service is 28 and the CL service can afford 32. The corresponding link utilizations are 60.2% and 69.5%, and the corresponding costs are 1.6 and 1.7.

Again, for the conservative computation-based admission control, we want

$$\frac{C}{8Mbps} * X_1 = \frac{C}{1Mbps} * X_2,$$

where 8 Mbps and 1 Mbps are given according to Table 1 and the allocation method in section 3.C. Thus, the cost ratio is 8 times. However, with a well-performed measurement-based admission control, the ratio is reduced to 1.14 according to the equation

$$28 * X_1 = 33 * X_2.$$

Table 2 and Table 3 summarize the cost and the number of admissible flows with different traffic-service pairs and admission control algorithms. For service providers, the billing policy should not be made by considering only the bandwidth request, i.e. R or r. A mechanism to identify the characteristics of traffic flows is important because a bursty flow costs more than a

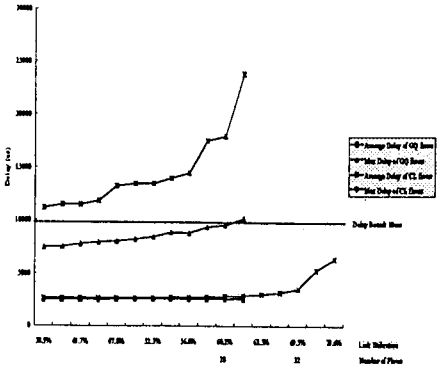


Fig. 10. Behavior of data traffic receiving GQ or CL service

	Conservative		Well-Performed	
	GQ	CL	GQ	CL
Video	1.6	1.5	1.55	1.55
Voice	0.5	0.026	0.065	0.046
Data	9	1	1.6	1.4

Table 2. Cost of various traffic-service pairs and admission control algorithms

constant bit rate flow. The effect of a well-performed admission control is more significant for bursty traffic. Besides, it can be observed that the cost to provide the GQ service is not so high as expected. Hence, by providing the GQ service with a well-performed admission control, service providers can convince their customers to subscribe this service and obtain more revenue. For service subscribers, to identify the characteristics of applications helps to select which service to subscribe. For non-bursty traffic, the GQ service is a good choice while enjoying a maximum delay bound, with just a little higher cost. However, if the traffic is bursty, the cost of the GQ service is much higher than the CL service. Nevertheless, being charged, on the GQ service, more than twice as the CL service seems unreasonable. The service provider might be over-charging or running a poor algorithm for admission control.

In summary, for service subscribers, we have the following suggestions:

1. Understanding the traffic characteristics of your applications is important.
2. For traffic with no or low burstiness, you may choose the GQ service to enjoy a transmission delay bound with a little higher payment. However, the cost and performance for these two services should not differ much.
3. With a larger traffic burstiness, the cost difference between these two services is larger. It is necessary

	Conservative		Well-Performed	
	GQ	CL	GQ	CL
Video	28	30	29	29
Voice	3	58	23	32
Data	5	45	28	32

Table 3. Number of admissible flows for various traffic-service pairs and admission control algorithms

to consider whether a maximum delay bound is actually required. If not, the CL service is a much cheaper choice.

For service providers, we provide the following suggestions:

1. It is unreasonable to make your billing policy by considering only the bandwidth request. Differentiating the traffic characteristics is necessary.
2. The cost difference between the GQ and CL services arises when the traffic burstiness increases. However, the GQ service is not so expensive as expected when a well-performed admission control mechanism is used.
3. Be careful to the ill-behaving admission control. It has a great influence on the cost of service provisioning.

B. Heterogeneous Traffic-Service Scenarios

The impact of ill-behaving admission control should be studied. There are totally 6 pairs of traffic-service types. By increasing the number of flows of a specific traffic-service pair, the impact on the existing flows are observed. The simulation results show that the most affected pair of existing flows is the voice-CL pair. For example, Figure 11 and Figure 12 show the impact of increasing the numbers video-GQ and data-GQ flows on the existing flows, respectively. It is noticed that the voice-CL flows suffer more from the overbooked flows. The data-CL flows are influenced with a smaller degree. All other heterogeneous scenarios lead to the similar results.

In this study, it is obvious that the flows receiving the CL service are more sensitive to overload than those receiving the GQ service when the admission control is too aggressive or ill-behaving. Another observation is that bursty traffic suffers more when the system is overloaded.

V. Conclusion and Future Work

In this paper, we compared two services, Guaranteed Quality (GQ) and Controlled Load (CL), defined in IETF Integrated Service Working Group for video, voice and data traffic. To conduct this study, we first characterized the video, voice, and data traffic by defining their traffic arrival processes and service specifications, i.e. *Tspec* and *Rspec*. Control mechanisms were then specified. They include a leaky bucket traffic policer, a packetized cyclic scheduler of weighted

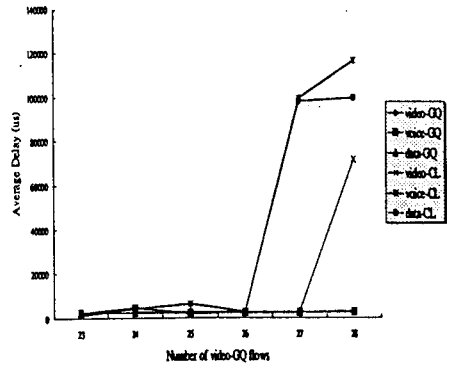


Fig. 11. Impact of increasing the number of video-GQ flows on the existing flows

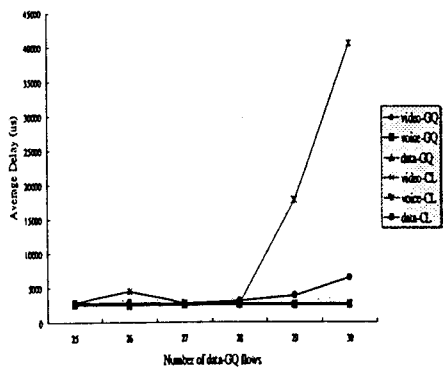


Fig. 12. Impact of increasing the number of data-GQ flows on the existing flows

fair queueing, and two admission control algorithms, i.e. a conservative computation-based algorithm and a well-performed measurement-based algorithm. Our simulation obtained results of link utilization, cost of service provisioning, number of admissible flows, average delay, and maximum delay.

Implications for service subscribers and providers were drawn from the numerical results. Traffic burstiness plays an important role in determining how different the GQ and CL services can be. Higher burstiness leads to larger difference between them in terms of both cost and performance. However, it was observed that this difference can be much reduced if a well-performed measurement-based admission control algorithm is in place. Nevertheless, subscribers are suggested to use the CL service when the traffic burstiness is high and the maximum delay bound is not critical. Subscribers should however keep in mind that traffic of high burstiness running over the CL service is more vulnerable to the overload situation perhaps caused by the ill-behaving admission control mechanism.

There are places where future works may be pursued. More realistic traffic models [15] and their corresponding *Tspec* and *Rspec* can be used for a similar study to verify the implications reported here. Control parameters in the scheduler and the admission control module should be further studied. Especially, how to design a well-performed measurement-based admission control module that can achieve the utilization reported here is worth pursuing. Cost estimation is another area that should be further explored. In our study, we estimate the cost of a flow as the total cost of the link divided by the maximum number of admissible flows. Other sophisticated methods may be used.

REFERENCES

[1] Braden R. Ed. et al. Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification. *RFC 2205*, September 1997.

[2] Shenker S. et al. Specification of Guaranteed Quality of Service. *RFC 2212*, September 1997.

[3] Wroclawski J. Specification of the Controlled-Load Network Element Service. *RFC 2211*, September 1997.

[4] Sugih Jamin, Peter B. Danzig, Scott J. Shenker, and Lixia Zhang. A Measurement-Based Admission Control Algorithm for Integrated Service Packet Networks. *IEEE/ACM Transactions on Networking*, 5(1), February 1997.

[5] A. Demers, S. Keshav, and S. Shenker. Analysis and Simulation of a Fair Queueing Algorithm. *Proc. ACM SIGCOMM '89*, pages 1–12, September 1989.

[6] Seaman M. et al. Integrated Services over IEEE 802.1D/802.1p Networks. *Internet Draft*, June 1997.

[7] ISO/IEC 13818-x. Information Technology – Generic Coding of Moving Pictures and Associated Audio Information.

[8] O. Rose. Simple and Efficient Models for Variable Bit Rate MPEG video traffic. *Performance Evaluation*, 30:69–85, 1997.

[9] P. Pancha and M. E. Zarki. Bandwidth Requirements of Variable Bit Rate MPEG Sources in ATM Networks. In *Infocom '93*, pages 902–909, 1993.

[10] Simon S. Lam, Simon Chow, and David K. Y. Yau. An Algorithm for Lossless Smoothing of MPEG Video. In *SIGCOMM 94*, pages 281–293, 1994.

[11] ITU Rec. G.711. Pulse Code Modulation (PCM) of Voice Freuencies.

[12] Richard V. Cox. Three New Speech Coders from the ITU Cover a Range of Applications. *IEEE Communications Magazine*, September 1997.

[13] G. Pal and S. Agrawal. Window-based Congestion Control in a Packet Switched Network with Voice and Data Transmission. *Computer Communications*, (19):612–618, 1996.

[14] Parekh A. K. and Gallager. R. G. A Generalizaed Processor Networks – the Multiple Node Case. In *Infocom '93*, pages 521–530, 1993.

[15] Paxson V. and Floyd S. Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, June 1995.