Guaranteed QoS Synthesis for admission control with shared capacity

D J Songhurst, P L Eardley, B Briscoe, C Di Cairano Gilfedder, J Tay

BT Research

{david.songhurst, philip.eardley, bob.briscoe, carla.dicairano-gilfedder, june.tay}@bt.com

Abstract

Guaranteed QoS Synthesis (GQS) is a distributed measurement-based admission control scheme. It is designed as a simple and scalable approach to providing strong service guarantees using bulk packet congestion marking across a core network region. We describe the operation and performance of GQS, with particular reference to its use for fair resource-sharing between guaranteed traffic and a rate-responsive non-guaranteed class. This analysis includes a detailed simulation study which fully represents the interactions between events at packet and session timescales. Results confirm that GQS provides strong guarantees under normal conditions, is robust to different traffic configurations, and readily recovers from network failure events.

1. Introduction

The Internet carries traffic of many different kinds, but we can make a broad distinction between flows that are responsive and flows that are not. Responsive flows are able to vary their rates in response to some indication of congestion such as dropped packets or explicit congestion notification (ECN) marking. Such flows might include web browsing and file transfers, and would typically use TCP's rate control protocol. Other types of traffic, such as interactive or streaming speech or video, may be unresponsive or have very limited ability to change their rate. Users may well desire a service that offers strong quality-of-service guarantees (limits on packet delay and packet drop probability) for these types of traffic. In order to provide such a service, differentiated from other traffic types, ideally we would like the network to apply admission control at the session level, but without losing the simplicity of a packet network.

The traditional approach to admission control is to use explicit capacity reservation. Each router (or a bandwidth manager on behalf of a set of routers) keeps track of all flows in progress and hence knows whether there is sufficient spare capacity to accept a new flow request. This approach is unscalable because of the need to maintain full flow state throughout the network. More recently there has been interest in the possibility of distributed measurement-based admission control, whereby admission decisions are made on the basis of a current measurement of load or congestion [1], but across the required network path [2, 3, 4].

Setting aside for now the technical problem of how rate control and admission control are achieved, there is an economic problem of how, in principle, network resources should be allocated between flows of all different kinds. This is an issue of fairness, which economists would interpret as maximising social welfare. Kelly et al [5] have shown that a TCP-like rate control responding to congestion indication achieves a rate allocation ('proportional fairness') that is socially optimal for responsive flows.

By interpreting congestion marks as prices this approach can be applied more generally to incentivise appropriate rate-sharing between flows of all kinds. Applications that are not responsive can send probe

packets to determine whether the current congestion price is lower than their 'willingness-to-pay', thus pushing the problem of admission control out to the edge of the network [2].

The European collaborative project M3I [6] examined various approaches to market-based pricing and resource control, including congestion pricing. One scheme to emerge from the project was the concept of a network region surrounded by gateway routers which perform admission control to that region on the basis of measurements of congestion across the region [7, 8]. This has subsequently been developed within BT as Guaranteed QoS Synthesis (GQS) [9]. The GQS scheme differs from previous measurement-based admission control schemes through its combination of the following key factors:

It operates within a defined region of the network that is fully protected by admission control at all border routers, and it is designed to interwork with other admission control schemes so as to provide end-to-end service guarantees.

It uses explicit congestion notification (ECN) as a means to signal congestion to the boundaries of the region.

It ensures 'pre-congestion notification' through the use of a virtual queue for early congestion marking.

Uniquely, it enables fair resource sharing between guaranteed and responsive traffic classes in a way that is adaptive to their relative demands.

Ming, Hoang and Simmonds [10] have also proposed a scheme for fair admission control by edge routers. However this scheme uses a more complex internal signalling mechanism (using a 'Resource Discovery Protocol'), and the fairness is defined by fixed shares of the capacity rather than being adaptive to relative demand.

In this paper we examine the performance of the GQS admission control scheme with particular reference to the way in which network resource is shared between guaranteed and non-guaranteed traffic. In Section 2 we discuss the concept of 'fair' resource-sharing between guaranteed and responsive traffic. Section 3 gives an overview of GQS, and Section 4 presents some of the simulation work. In Section 5 we describe a bandwidth protection mechanism which can prevent starvation of either traffic class. Section 6 describes how GQS would be configured to cope with variable bit-rate reservations. Sections 7 and 8 present further development of the GQS approach, and conclusions.

Note that we refer variously to guaranteed and non-guaranteed traffic, non-responsive and responsive traffic, and reservation and non-reservation traffic. Within the scenario considered in this paper these are all references to the same two traffic classes.

2. Resource-sharing for guaranteed and responsive traffic

Responsive and non-responsive applications can be characterised by notional utility functions representing the benefit gained as a function of data rate (Figure 1). This utility can be expressed as a 'willingness-to-pay'. This suggests that if congestion marks are interpreted as congestion charges then network resources can be shared fairly between these two kinds of traffic by subjecting non-responsive demands to admission control on the basis of a fixed threshold applied to measured congestion marking rate. This concept underlies the GQS system of measurement-based admission control, described in Section 3.

Now in the current Internet environment, responsive applications do not actually have an incentive to reduce their rates in response to congestion marking. This might be achieved through congestion pricing, or through policing (see [11] for a possible policing mechanism). In any case we assume that there is a class of responsive traffic (which we call 'non-guaranteed') which may be distinct from the lowest-priority 'best-effort' traffic. This class should share resources in a fair way with rate-guaranteed connections – 'fair' in the sense that packets in non-guaranteed flows have a small (but non-zero) value relative to the per-packet value of guaranteed flows.

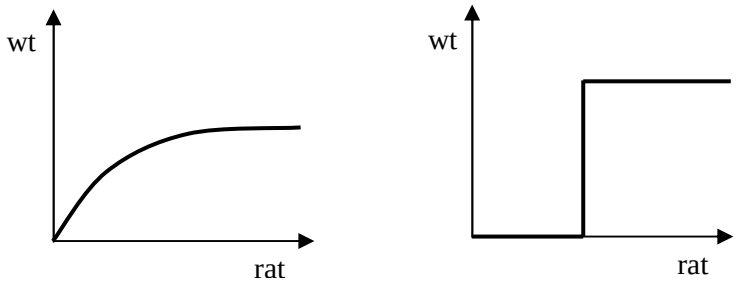


Figure 1 Utility (willingness-to-pay) of responsive and non-responsive applications

This approach has several advantages. Compared to the simple approach of segregated capacity it is more fair, efficient and robust. It provides a fairer level of service to non-guaranteed (responsive) traffic, because admission of guaranteed flows is constrained by strong non-guaranteed demand – even though that demand is responsive. It is more efficient under varying traffic mix because it does not require capacity to be explicitly allocated for each class – there is a flexible resource allocation boundary between the classes which adapts to relative demand. It should also be more robust to network failures that necessitate routing changes – the resource allocation boundary will automatically adapt to new load conditions.

We recognise that this approach to fair resource allocation may not be acceptable to network operators in all circumstances. In particular, if non-guaranteed traffic is not subjected to an adequate level of usage-charging then there is the possibility that excessive non-guaranteed traffic load can cause denial-of-service to guaranteed traffic (through admission control). This can be countered by ensuring the availability of minimum guaranteed levels of bandwidth to each traffic class, as discussed in Section 5.

3. Guaranteed QoS Synthesis

The key concept of Guaranteed QoS Synthesis (GQS) is that an admission-controlled guaranteed service is 'synthesised' from congestion measurement across a core region of the network, when the routers within this region are not flow-aware and simply do bulk packet congestion marking. This GQS region may cross several operator domains – the important requirement is that all routers on the border of the region act as GQS gateways performing admission control to the region. Figure 2 shows a GQS region.

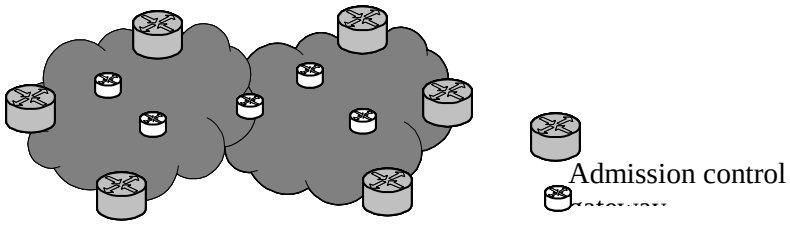


Figure 2 A GQS region Core router

The GQS gateway and core routers have the following functions:

Core routers use priority queueing to give precedence to guaranteed packets over non-guaranteed packets. They also do explicit congestion notification (ECN) marking in order to give early warning of approaching congestion. Queueing and marking algorithms are described below.

- Gateway routers have both ingress and egress functions with respect to the GQS region. The egress function is to monitor the proportion of guaranteed traffic that is congestion-marked, separately for each ingress-egress path, and to signal this information to the appropriate ingress router. The ingress function is to perform admission control for new guaranteed connection requests by comparing the measured

congestion with a fixed threshold. The ingress function also includes queueing and marking as for core routers.

Note that guaranteed packets only carry ECN marks within the GQS region – they are removed by the egress gateway. GQS is designed to interwork with end-to-end protocols for reservation, such as RSVP and SIP. Only the gateway routers need to communicate with these protocols. Note also that a GQS region can encompass more than one operator domain. See [9] for a fuller discussion of GQS.

Core router queueing and marking functions

By using priority queueing within core routers, the GQS region can be configured so that guaranteed packets have negligible queueing delay, and there is very low packet drop probability (for either class) except under failure conditions. Where guaranteed and non-guaranteed packets share the same physical queue in a router it is also possible to use a packet pre-emption mechanism whereby guaranteed packets can always have access to the full queueing space by pre-empting waiting non-guaranteed packets – this further reduces the likelihood of guaranteed packet drops.

Various possible ECN marking algorithms can be used in core routers, and two of these are illustrated in Figure 3. Both of these use a virtual queue, which is a counter (equivalent to a leaky token bucket) whose size is used to determine the probability that an arriving packet is marked. The virtual queue output rate is slightly less than the configured route capacity (by an amount *delta* in Figure 3) in order to ensure that ECN marking precedes any possibility of dropped packets.

In the simple algorithm of Figure 3(a), all packets cause the virtual queue to be incremented, and all packets are marked probabilistically according to a RED-type algorithm depending on the size of the virtual queue (graph on right of Figure 3(a)). Hence flows from both traffic classes are marked with equal probability.

Figure 3(b) illustrates a more sophisticated algorithm which provides for differentiated marking of reservation and non-reservation packets. This algorithm has the following differences from Figure 3(a):

Non-reservation packets are not fed into the virtual queue, but are marked according to the size of the non-reservation packet queue. This is because latency is considered less critical for this class so it is not necessary to keep the real queue short by using a virtual queue.

The marking of non-reservation packets does not depend on the size of the virtual queue. This is because non-reservation packets do not affect reservation packet performance (because of the priority queueing), so their marking probability need not reflect the number of reservation packets in the virtual queue.

Reservation packets are marked according to the sum of the virtual queue and the non-reservation packet queue.

GQS performance issues

We have used a combination of analysis and simulation studies to evaluate GQS performance. The following factors are of particular interest:

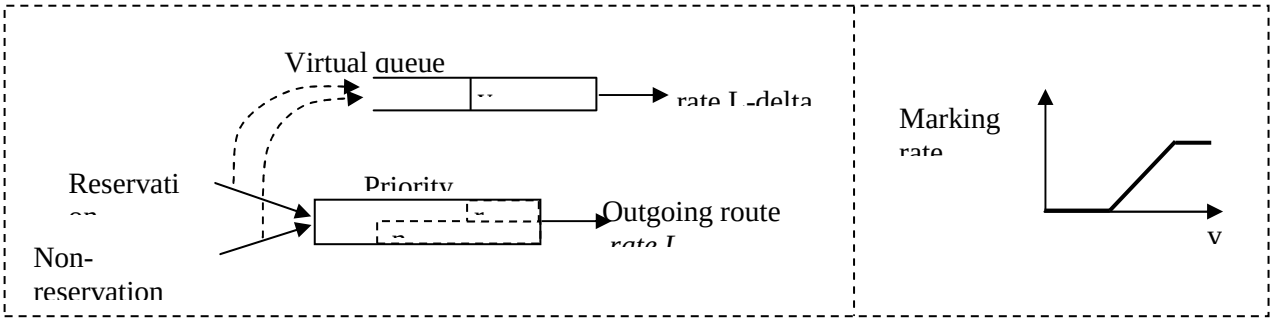
Does GQS provide, under normal conditions, a solid guarantee of low delay and no packet drop for reservation traffic?

What utilisation can be achieved consistent with this guarantee?

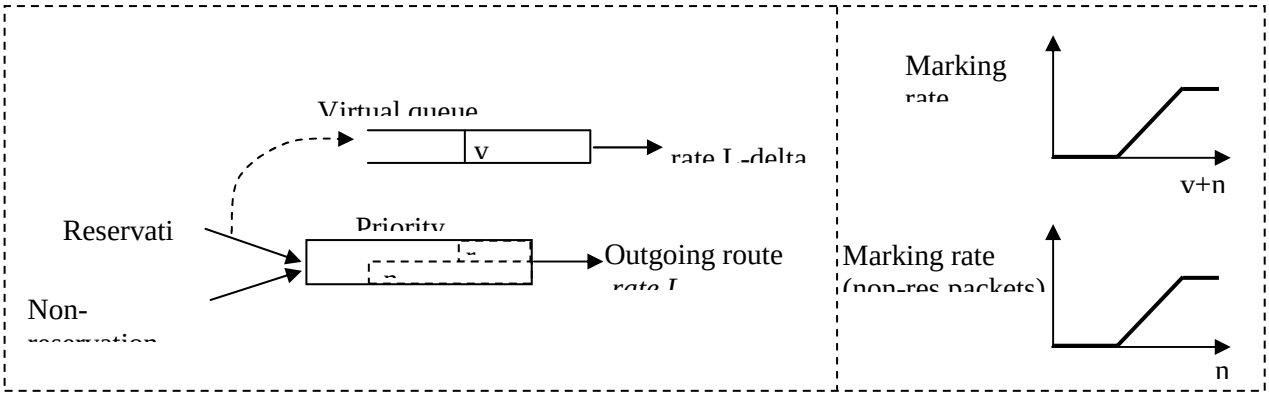
- Under what conditions (high load, failures) might the guaranteed performance fail?

How is bandwidth shared between reservation and non-reservation traffic under different patterns of relative demand?

How robust are GQS parameter settings – do they need to vary with different network and traffic configurations?



(a) Algorithm for identical marking



(b) Algorithm for differentiated marking

Figure 3 GQS core router congestion marking algorithms

4. GQS simulation study

The ns-2 Network Simulator [12] was used to develop a simulation model of the topology illustrated in Figure 4. This comprises a single ingress-egress path through a GQS region, with up to two core routers on the path. The core and gateway routers operate the GQS system as described above, including the ECN marking algorithm shown in Figure 3(a). Traffic on the ingress-egress path comprises responsive flows and guaranteed flows. The core network links have capacity either 100Mbit/s or 1Gbit/s. A fixed number of responsive (ECN-responsive TCP) flows are modelled as FTP sessions that last throughout the simulation (constantly bandwidth greedy). Short duration web-like responsive flows and long-range dependence arising from a superposition of such flows were not modelled. Guaranteed flows arrive in a Poisson process, having a mix of bandwidth demands ranging from 64kbit/s to 512kbit/s, at constant bit-rate (see, for example, [13] which found that the Poisson process is a good model for user-generated session arrivals). The 'cross traffic' on each link is modelled as background packet load, not as separate flows.

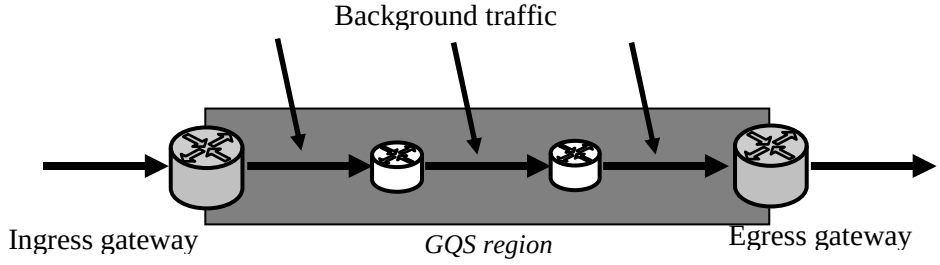


Figure 4 Simulated network topology

A significant problem for the simulation design was the wide range of event timescales. On a 1Gbit/s link the time interval between packet arrival and departure events is of the order of a microsecond. The timescale for TCP rate reaction is of the order of 30 – 200 milliseconds, dependent on round-trip time. The timescale for arrivals and durations of guaranteed flows is multiple seconds. It was considered necessary to encompass all of these event timescales in a single simulation model in order to capture their interaction. This meant that runs would be very long – a typical simulation run of 20 to 30 minutes of simulated time would require the generation of many millions of packets.

Typical results traces from high-load run

Figure 5 shows results from a typical high-load simulation run. In this example the ingress-egress path traverses two core network links (and one core router), each link having capacity 100Mbit/s. The total demand at each link (guaranteed and non-guaranteed traffic) is substantially more than 100Mbit/s. The admission control threshold is set at a marking probability of 0.15. In this scenario the non-guaranteed traffic throughput is reduced (through rate-adaptation) from a normal 23Mbit/s to an average 0.5Mbit/s. The remaining traffic load on each link (background traffic plus end-to-end guaranteed traffic) is constrained (through admission control) to 98.5Mbit/s. There are no packet drops.

Throughput under overload

Table 1 shows some average statistics for a configuration using 1Gbit/s links, at three different load points. At normal load (96.8% utilisation) there is negligible ECN marking. At high load, responsive traffic throughput is reduced in response to ECN marking but non-responsive traffic is unaffected. At very high load, responsive traffic throughput is severely reduced and non-responsive throughput is reduced by admission control. In each case there were no packet drops, and core network queueing delay was negligible (less than 0.2ms) for both traffic classes.

Robustness

A series of experiments was made to test the robustness of GQS to varying traffic scenarios. Throughout this series the GQS parameters were fixed as follows:

admission control threshold was 0.15,
virtual queue output rate was 0.99 of the link capacity,
the minimum and maximum thresholds on the virtual queue size for RED-style marking were 0.2 and 0.8 of the actual outgoing buffer size.

The following variations were made in the scenario:

Link capacity: 100Mbit/s, 1Gbit/s
Guaranteed traffic (% of total traffic demand on each link): 30%, 50%, 70%
Ingress-egress traffic (% of total traffic on each link): 20%, 50%
Core links having equal or differing loads
Number of core links between ingress and egress: 2, 3

In all cases results were essentially similar to Table 1.

Summary of conclusions from simulations

The following overall conclusions were found from the simulation results:

- Under all ‘normal’ (non-failure) conditions GQS ensured excellent QoS for admitted guaranteed connections, with no packet drop and negligible queueing delay.

- It is possible to run at very high throughput (99% of line rate) with no packet loss. Non-guaranteed traffic is protected from packet loss by rate-adaptation in response to ECN marking.

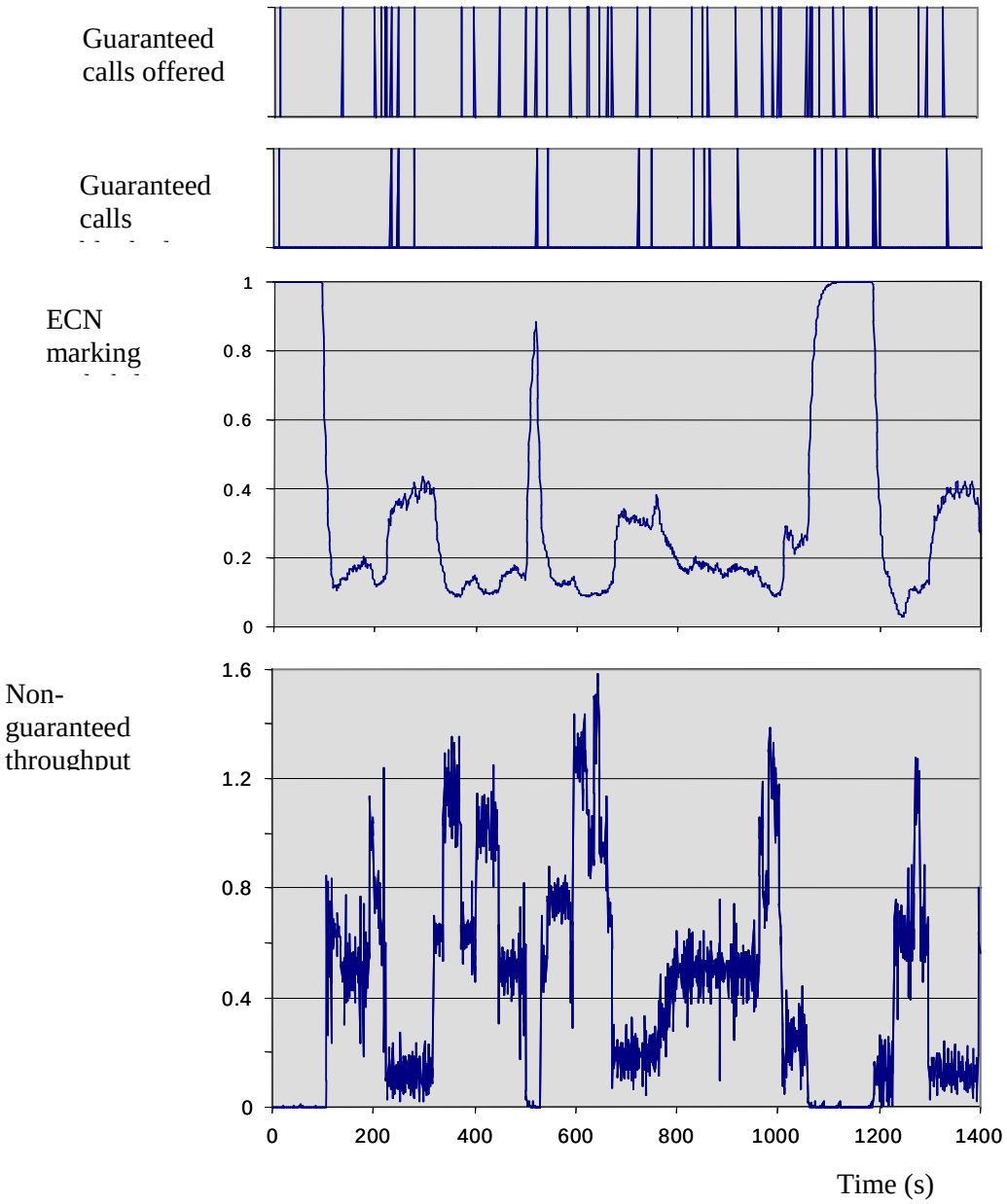


Figure 5 Traces from high-load simulation run

Load point	Non-responsive traffic (Mbit/s)		Responsive throughput (Mbit/s)	Mean ECN rate	Utilisation
	Offered	Throughput			
Normal	736.9	736.9	231.5	0	0.968
High	835.7	835.7	136.0	0.004	0.972
Very high	1033.3	984.5	4.9	0.176	0.989

Table 1 Throughput per traffic class at different load points

As guaranteed traffic load increases it first causes non-guaranteed traffic to back-off and is ultimately constrained by admission control. It is straightforward to ensure no guaranteed packet drops, by setting the virtual queue output rate and the admission control threshold so as to ensure that admission control is triggered when there is still some spare capacity (more than the maximum rate of a new connection).

The important configurable GQS parameters are the virtual queue output rate reduction factor *delta* and the admission control threshold. Experiments with various different traffic configurations showed that these parameters are robust and can be fixed regardless of relative levels of guaranteed and non-guaranteed traffic.

Some simulation runs were started in an artificially overloaded state (excessive number of guaranteed connections). In practice this situation could arise due to network failures. In this scenario packet drops were experienced for a limited period until sufficient guaranteed connections had terminated. The system then found a new equilibrium with zero packet loss.

- The TCP rate reaction of non-guaranteed traffic is on a faster timescale than arrivals and departures of guaranteed connections. This is, in part, because relatively few distinct guaranteed flows were simulated – most guaranteed traffic was simulated as constant rate background traffic. Figure 5 shows how, in a high-load scenario, the non-guaranteed throughput switches between different equilibrium levels whenever a guaranteed connection starts or terminates. This might be the case in practice if there are not many guaranteed flows on a link, or if they have long durations. By contrast, if there were many guaranteed connections (such as voice calls) then the total guaranteed throughput could fluctuate by small amounts on a similar timescale. However the total responsive traffic load can show large fluctuations on a small timescale because all flows vary their rate simultaneously.

5. Bandwidth protection

On an overloaded link, what determines the relative throughputs of guaranteed and non-guaranteed traffic? Average guaranteed throughput is limited (by admission control) to a maximum level no greater than the virtual queue output rate (*L-delta* in Figures 3(a) and 3(b)). However strong non-guaranteed demand can constrain guaranteed throughput to a lower level, and this is dependent on the value of the threshold at which the measured marking rate triggers admission control. Even with a rather small value (e.g. 10%), responsive flows will back off very considerably before admission control is triggered. However if there are many responsive flows then their total throughput – even when their individual rates have adapted to a 10% marking rate – may be large. So it is possible for guaranteed demand to be starved of capacity. If non-guaranteed flows are not suitably charged or policed, this might be considered a denial-of-service issue.

In fact, it is possible to enhance the ECN marking algorithm used by core routers so as to ensure that each class of traffic (guaranteed and non-guaranteed) receives a given minimum level of bandwidth regardless of the demand of the other class. This is illustrated in Figure 6, which is an enhancement of the algorithm shown in Figure 3(b).

The algorithm of Figure 6 has two enhancements:

The virtual queue output rate is reduced by an amount M_N. This ensures that guaranteed flows will receive heavy marking, so triggering admission control, when the guaranteed load approaches $L-M_N$ where L is the line rate (or configured capacity). Hence non-guaranteed traffic always has access to at least bandwidth M_N

An additional token bucket is introduced to filter out guaranteed packets, up to rate M_G, which are not added to the virtual queue and are not subject to congestion marking. The virtual queue output rate is reduced by a corresponding rate M_G. This ensures that guaranteed traffic always has access to at least bandwidth M_G.

We assume that $M_N + M_G << L$, so that there is an adequately large region where traffic is subject to congestion marking.

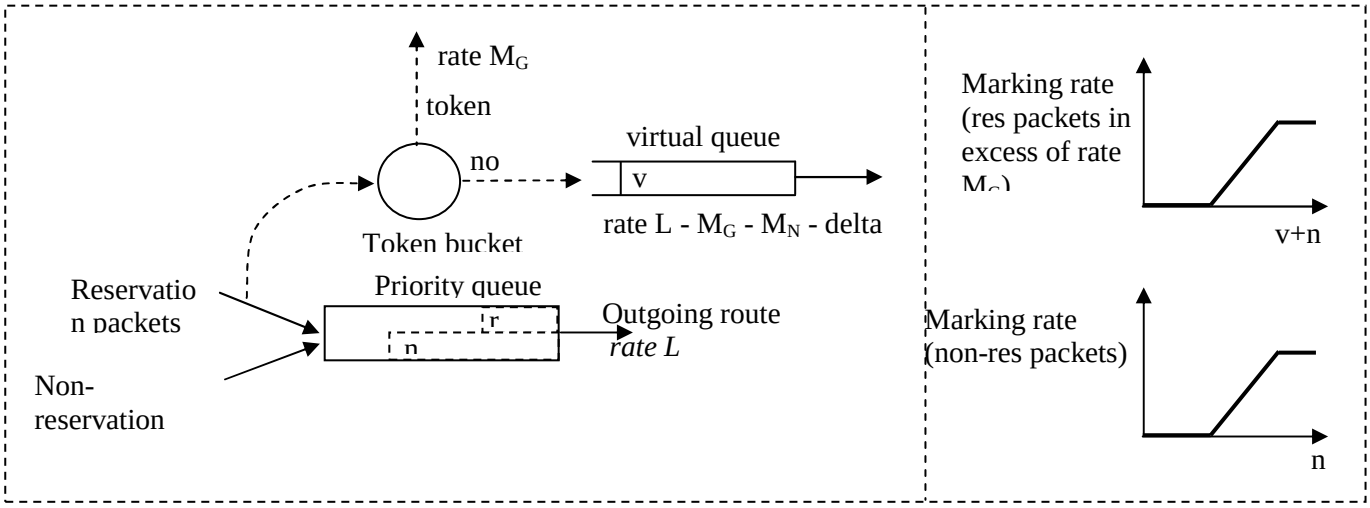


Figure 6 Marking algorithm for bandwidth protection

Figure 7 illustrates how this bandwidth protection algorithm shares capacity differently within three different load regions defined by the relative guaranteed and non-guaranteed traffic demands.

In the left-most region the guaranteed demand is less than M_G. In this case no guaranteed packets receive congestion marking and so new guaranteed connection requests are always accepted, even protecting against DoS attacks from non-guaranteed traffic. Non-guaranteed traffic can use any spare capacity within this region.

In the central region, if guaranteed demand is greater than M_G and non-guaranteed demand is greater than M_N then resource is shared according to relative demand as normal.

At the start of the right-most region, if the guaranteed traffic level reaches $L - M_N$ then new guaranteed connection requests are blocked by admission control. Hence this region is effectively reserved for non-guaranteed traffic.

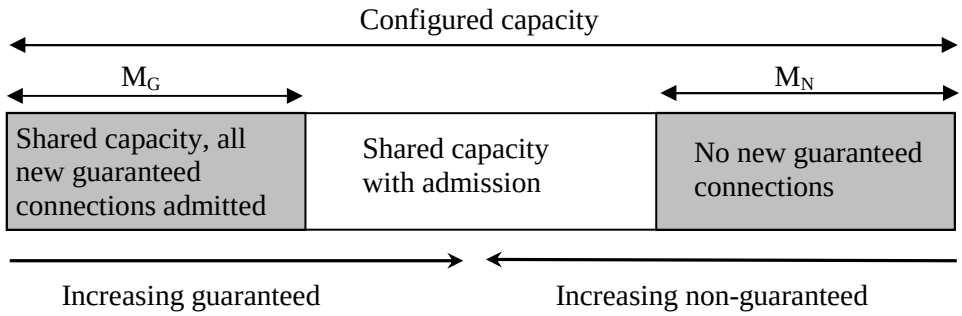


Figure 7 Bandwidth protection regions

6. Variable bit-rate reservations

Up to now we have assumed that guaranteed connections have constant bit-rate. Suppose now that they may have variable bit-rate (VBR) up to a known peak rate. This poses a problem for any kind of admission control system. If an explicit capacity reservation system (such as Intserv) is used, then how much capacity should be reserved per connection? Peak rate allocation could be very wasteful of capacity if mean rates are substantially less than peak rates. This problem has previously been addressed for an ATM environment, using effective bandwidth models [14, 15]. 'Effective bandwidth' is a measure of the bandwidth required per

flow, in order to achieve a given QoS target, when a number of variable bit-rate flows are multiplexed togther. Hence effective bandwidth depends not only on traffic characteristics but also on network characteristics such as link capacity and buffer size. Effective bandwidth lies between the mean rate and peak rate of a flow, and when many flows are multiplexed on a large link it tends to be closer to the mean rate.

Using a measurement-based admission control system such as GQS, the difficulty is that under-used capacity is not reserved. The congestion measured by egress gateways only reflects the current load, not the under-used capacity. Therefore it is possible for too many guaranteed flows to be admitted with the risk of packet drops if some VBR flows increase their rates.

This might be a problem if GQS is used at the edges of the network where the rates of individual flows may be a significant proportion of link capacity. A capacity-based system of admission control is more appropriate here. However in the core of the network where links have large capacity it is possible for GQS to accommodate VBR connections efficiently. The key mechanism is to reduce virtual queue output rates appropriately (M_N in Figures 6 and 7). Then admission control is applied when the capacity unused by guaranteed traffic falls to M_N – the right-most region in Figure 7. This capacity is available for non-guaranteed traffic, and can also be used by existing VBR reservations that burst above their average rate. If the safety factor M_N is substantially larger than the largest individual reservation then this mechanism can provide an ‘almost-certain’ assurance of no packet drops for guaranteed connections (except, of course, in network failure situations). This approach is illustrated in Figure 8.

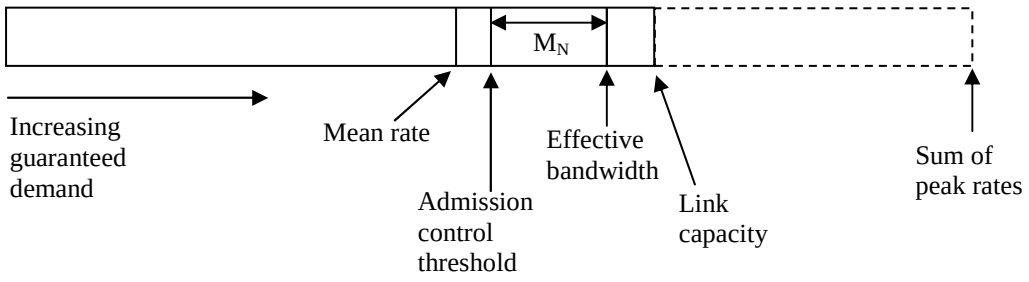


Figure 8 Admission control for variable bit-rate reservations

Figure 8 illustrates how we might use an estimate of the total effective bandwidth of guaranteed load, as an excess over the mean rate, in order to set the safety factor M_N. For large core network links with many flows this relationship could be expected to be fairly consistent.

How large would the factor M_N typically need to be? We can start from analyses of the characteristics of VBR traffic, including silence-suppressed voice [16] and video [17]. Calculations indicate that effective bandwidth is typically no more than 10% over the mean rate for voice traffic multiplexed on links of at least 100Mbit/s and for video traffic multiplexed on links of at least 1Gbit/s.

Note finally that peak rates of VBR flows are likely to be much larger than mean rates. For silence-suppressed voice the peak rate is 2-3 times larger than the mean rate, and for many video formats the factor can be much larger. Hence naïve admission control based on capacity reservation, with peak rate allocation, would be very wasteful of capacity, as indicated in the right-hand part of Figure 8.

7. Further development

GQS can also be used in a more traditional scenario where admission control of guaranteed traffic is not adaptive to non-guaranteed demand. This is achieved by just one small change in the congestion marking algorithm – in Figure 3(b) (and Figure 6) reservation packets are marked according to the virtual queue size

v, taking no account of the non-reservation queue size. Then non-reservation load has no effect on admission control, so that guaranteed traffic always has access to the full configured capacity, with priority over other traffic classes. Hence it is not necessary to use the token bucket mechanism of Figure 6 to ensure minimum bandwidth M_G – but it is still appropriate to reduce the virtual queue rate by some amount M_N in order to ensure a minimum bandwidth available to non-guaranteed traffic and to provide a safety margin for admission-controlled calls.

Recently we have been pursuing standardisation within the IETF Transport Area Working Group [18]. We are collaborating with the authors of RTECN [19] and RMD [20] and with Cisco. The most significant change has been the addition of a flow pre-emption mechanism. It is possible to use similar ECN marking to that described earlier so that ECN measurements trigger flow pre-emption of existing reservations. This is a valuable mechanism in extreme failure situations – by pre-empting sufficient reservations the total load is brought below the current capacity, rapidly restoring QoS for the remaining reservations. Flow pre-emption is selective so it is possible to ensure service for high-priority (emergency services) traffic.

8. Conclusions

Guaranteed QoS Synthesis is a very effective technology for providing an admission-controlled service for guaranteed QoS. It combines simplicity and scalability with fair and efficient use of resources.

The analysis of GQS operation has been supported by detailed simulation studies which confirm the excellent QoS provided by GQS and its robustness to different traffic patterns.

GQS can be configured to provide absolute priority to guaranteed services, or to operate fair sharing of resources between guaranteed and non-guaranteed services. The resource-sharing mode has advantages of economic efficiency and high robustness to traffic configuration and failures. It is straightforward to configure minimum bandwidth guarantees that ensure neither traffic class can be starved of resources.

GQS can be configured to cope with variable bit-rate reservations in a way that is very efficient within large core network links.

We are pursuing IETF standardisation for the essential components of distributed measurement-based admission control using ECN marking, of which GQS is one possible realisation.

References

[1] *A measurement-based admission control algorithm for integrated service packet networks*, S Jamin, P Danzig, S Shenker, L Zhang, IEEE/ACM Transactions on Networking, Vol 5 No 1, February 1997

[2] *Distributed admission control*, F P Kelly, P B Key, S Zachary, IEEE JSAC 18 (2000) 2617 – 2628

[3] *Egress admission control*, C Cetinkaya and E Knightly, Proc IEEE Conference on Computer Communications (Infocom'00), http://citeseer.nj.nec.com/cetinkaya00egres.html, March 2000

[4] *Endpoint admission control: Architectural issues and performance*, Breslau, Knightly, Shenker, Stoica, and Zhang, Proc ACM SIGCOMM'00, Computer Communication Review, 30(4), October 2000

[5] *Rate control in communication networks: shadow prices, proportional fairness and stability*, F P Kelly, A Maulloo, D Tan, Journal of the Operational Research Society, 49, No 3, pp 237 – 252, March 1998

[6] EU Fifth Framework Project M3I – Market-Managed Multiservice Internet, 2000-2002, http://www.m3i.org

[7] *Admission control based on packet marking and feedback signalling - mechanisms, implementation and experiments*, M Karsten and J Schmitt, Technical Report TR-KOM-2002-03, TU-Darmstadt, http://www.kom.e-technik.tu-darmstadt.de/publications/abstracts/KS02-5.html, May 2002

[8] *Packet Marking for Integrated Load Control*, M Karsten and J Schmitt, in IFIP/IEEE Proc 9th Int'l Symposium on Integrated Network Management (IM'05), http://www.cs.uwaterloo.ca/~mkarsten/papers/im2005.html, May 2005

[9] Guaranteed QoS Synthesis – an example of a scalable core IP quality of service solution, P Hovell, R Briscoe, G Corliano, BT Technology Journal, Vol 23 No 2, April 2005

[10] *Fair intelligent admission control over DiffServ network*, Ming, Hoang and Simmonds, Proc International Conference on Networks (ICON'03), http://www-staff.it.uts.edu.au/~simmonds/ICON03ming.pdf, September 2003

[11] *Policing Congestion Response in an Internetwork using Re-feedback*, B Briscoe, A Jacquet, C Di Cairano-Gilfedder, A Salvatori, A Soppera, M Koyabe, Sigcomm 2005

[12] *The Network Simulator – ns-2*, http://nsnam.isi.edu/nsnam/index.php/User_Information

[13] *Wide-area traffic: the failure of Poisson modelling*, V Paxson and S Floyd, IEEE/ACM Transactions on Networking, Vol 3, pp 226 – 244, 1995

[14] Charging communication networks: from theory to practice, D. J. Songhurst (editor), Elsevier, 1999

[15] Prof F P Kelly page on effective bandwidths, http://www.statslab.cam.ac.uk/~frank/eb/

[16] *Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation*, Jiang and Schulzrinne, 9th IEEE International Conference on Computer Communication Networks, 2000

[17] Professor V Siris pages on large deviation techniques, http://www.ics.forth.gr/netgroup/msa/

[18] *A Framework for Admission Control over DiffServ using Pre-Congestion Notification*, B Briscoe, P Eardley, D Songhurst, F Le Faucheur, A Charny, J Babiarz, K Chan, Internet Draft draft-briscoe-tsvwg-cl-architecture-01, October 2005

[19] *Congestion Notification Process for Real-Time Traffic*, J Babiarz, K Chan, and V Firoiu, IETF draft-babiarz-tsvwg-rtecn-05 Work in Progress, October 2005

[20] *RMD-QOSM - The Resource Management in DiffServ QoS model*, A Bader, L Westberg, G Karagiannis, C Kappler, T Phelan, IETF draft-ietf-nsis-rmd-06 Work in Progress, February 2006