Guaranteed QoS synthesis — an example of a scalable core IP quality of service solution

P Hovell, R Briscoe and G Corliano

With the transition of services like telephony to be carried over IP networks there is the potential for catastrophic numbers of calls to fail whenever sufficient demand is focused on unpredictable points in the core IP network. This is well known; service differentiation helps but does not alleviate the problem — call admission control is required but seems expensive for the few occasions it is required. This paper describes a BT-developed experimental mechanism called guaranteed QoS synthesis (GQS) that performs call admission for core IP networks for constant bit rate streams (voice and video). The mechanism is primarily aimed at Internet services but it may be possible to extend it for VPN applications. The GQS mechanisms is economic to deploy and operate, and scales without any increase in complexity. It achieves these properties by keeping no flow state in the network and basing call admission decisions on the measured congestion across the network. The paper describes the high-level GQS architecture as well as some of the deployment issues and potential savings in the operational support area. How GQS enables the separation of the interconnect QoS and retail business models is also explained.

1. Introduction

Within the communications industry there is a trend towards converged networks where all types of traffic are carried over a single core IP network. In particular services like telephony over IP (VoIP), video on demand (VoD) and video telephony will be carried on the network alongside best effort Internet traffic. The former services, in general, generate near constant bit rate traffic that does not react (i.e. reduce their sending rate) to network congestion. Congestion, in the core of the IP network, can occur when sufficient demand happens to focus on a particular point in the network. When this occurs there is a risk of catastrophic numbers of call failures. To remove this risk, mechanisms to guarantee quality of service (QoS) for calls once accepted are required.

In core networks there are proposals to give priority to certain services, this greatly reduces the risk of catastrophic failure, but does not remove it. Other proposals divide up the available bandwidth and add call admission control (e.g. MPLS) to alleviate the catastrophic failure condition. Unfortunately, engineering guaranteed QoS into core networks will add significantly to operational and capital costs. This additional cost is difficult to justify for the few occasions it will be necessary, although catastrophic failures will be hard to hide from the public eye (particularly when TV phone-ins will be both the most likely cause and the most likely casualty). A mechanism is needed that gives QoS guarantees but at minimal extra cost. During 2000/02 BT initiated and led a collaborative project[1] investigating a new approach to the provision of QoS services based on the measured congestion of the network.

During 2003/05 additional work (design, comprehensive simulation and a demonstrator) has been undertaken to prove that a generic idea generated within that project could be a solution to providing guaranteed QoS, economically, over either a single core or interconnected core networks — the resulting solution is called guaranteed QoS synthesis (GQS).

The GQS solution provides guaranteed QoS to constant bit-rate traffic like voice or video via flow (call) admission control. The challenge since the early 1990s has been to achieve such guarantees without losing the characteristic benefits of the Internet approach. It is well known that the robustness and low cost of the Internet depends on not remembering anything about passing packets. Offering flow guarantees seemed to be

[1] Market Managed Multiservice Internet — http://www.m3i.org/

impossible without remembering something about the flow from one packet to the next. The GQS solution was the first, and remains the best, solution to this problem, particularly because it is built on the recently developed understanding of the fundamental economics of networks. To a large extent, the carrier-grade properties that we are looking for emerge naturally because we have paid attention to these technical and economic features from the outset. Its stateless nature gives GQS its resilience and simplicity of configuration. Its foundations in both supply and demand side economics links it naturally to typical capacity planning practices and ensures it is proof against attempts to cheat by any of the players in the market-place.

2. GQS overview

2.1 QoS guarantee requirements

The most fundamental requirement that GQS was designed to support was the provision of economic but strong (IntServ-like [1]) QoS guarantees for a reservation class while allowing the overall balance between reservation and non-reservation classes to be configured according to economic and service requirements. The various types of reservation and non-reservation traffic and how they are handled by GQS are outlined below.

- Constant bit-rate reservations

  GQS is designed to offer a guaranteed service (low packet-queuing delay, no packet drops) for reservations sending at constant bit rate. This guarantee is maintained by applying acceptance control to constrain the total reservation traffic load, and by applying strict priority (order of service and packet drop) to reservation packets over non-reservation packets. The priority queuing algorithm ensures that non-reservation packets cannot cause queuing delay or drops in reservation packets. The acceptance control algorithm uses congestion measurements as its metric to make call acceptance decisions.

- Variable bit-rate reservations

  Not all reservations are constant bit rate; some applications may generate variable bit-rate traffic requiring strong QoS guarantees. While these are not a problem for QoS mechanisms using explicit capacity reservation (e.g. IntServ [1]/RSVP [2]), they do present a problem for GQS because of its reliance on measurements. The current GQS design does not support variable bit-rate reservation although various ideas are being worked on that may allow GQS to be extended to handle this type of traffic reservation.

- Non-reservation traffic

  Essentially the same 'best-effort' service models can be applied as would be used in the absence of GQS, but two sources of interaction between reserved and non-reserved traffic should be noted:

  — the priority given to reservation calls (relative to non-reservation packets) is configurable, e.g. the effect of setting a high call acceptance threshold would be high packet-level congestion for non-reservation traffic (i.e. relatively high levels of packet drop and packet delay) being allowed before reservation calls are blocked — however, the GQS system is such that non-reservation traffic is not completely starved of resources;

  — the call blocking probability for reservation traffic is affected by the total load (including non-reservation demand), so that an economic balance of traffic carried between reserved (guaranteed) and non-reserved (best-effort) traffic is achieved.

2.2 GQS mechanisms

From a high-level point of view, the GQS architecture consists of a set of mechanisms, which extends the functionality of existing network elements. GQS mechanisms are structured in three broad classes — data path, internal control path, and end-to-end control path mechanisms. These three classes of mechanisms are shown in Fig 1 and outlined below.

- Data path mechanisms

  They include all mechanisms enabling gateways and core routers to treat packets differently (depending on whether they receive guarantees or not), to mark/drop packets (depending on the local congestion state), and forward them to the next hop.

- Internal control path mechanisms

  They include all mechanisms that enable gateways, upon external request, to configure and query data path mechanisms. They allow gateways to gain views of the congestion state from other peer gateways, decide whether to admit a flow or not, and provide a response to the external requester.

- End-to-end control path mechanisms

  They include all mechanisms that enable gateways to interface to the external world; in particular, they enable gateways to receive requests for guaranteed traffic and to give responses to the external requester.

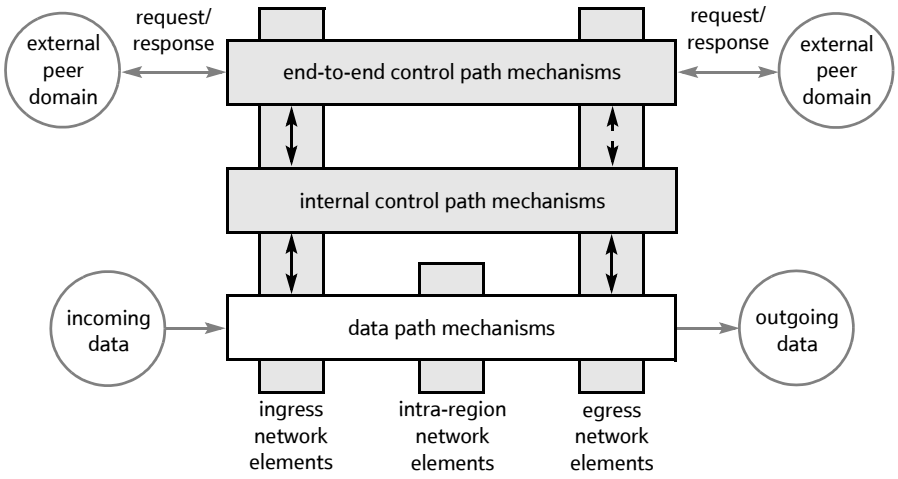


Fig 1 GQS mechanisms.

2.3 GQS network elements

Figure 1 also shows how the three sets of GQS mechanisms are mapped on to the three classes of network elements — ingress, intra-region, and egress elements. Traffic traversing a GQS region experiences all three types of element. A device can, however, act as both an ingress and an egress element. Note that the overall operation of each element depends on the signalling protocol used for the end-to-end control path; hence, any operation described here is an alternative among many.

- Ingress network elements

  These represent the entry points to a GQS region. Each flow has exactly one entry point which is determined by the inter-domain routing between network operators. The ingress of the region requires that all three types of the above-mentioned mechanisms be implemented. What we refer to as a GQS ingress gateway implements at least the data path mechanisms — internal and end-to-end control path mechanisms can be implemented either on the same element or by some other element at ingress of the region — the GQS design is agnostic to such a choice, in that the design copes with both. In terms of operation, ingress elements do the following:

  — end-to-end control path mechanisms intercept requests for guaranteed flows, pass them to the internal control path mechanisms and, depending on the end-to-end transactional model, they either forward the request to the following (QoS) domain, or wait for an internal response;

  — internal control path mechanisms process requests appropriately and return a response to the end-to-end control path mechanisms, and, in particular, as part of the processing, they:

  query egress elements to get a load report stating the congestion state of the link between ingress and egress elements of the GQS region,

  if a load report does not exist (i.e. there is currently no traffic flowing between the ingress and the requesting egress gateways), initiate a probe mechanism (between ingress and egress), the outcome of the probing being the creation of a load report for the pair of gateways — a query to the egress elements will then get the probe-initiated load report,

  once a load report has been received, perform admission control which will generate a success/ failure response,

  if successful, configure data path mechanisms;

  — for all admitted traffic, data path mechanisms classify incoming packets into either guaranteed or best-effort classes, policing traffic to make sure it is conformant to the service agreement.

- Egress elements

  These represent the exit points of the GQS region. Each flow has exactly one exit point which is determined by a combination of inter- and intra-domain routing. The egress of the region requires that all three types of the above-mentioned mechanisms be implemented. What we refer to as a GQS egress gateway implements at least the data path mechanisms — internal and end-to-end control path mechanisms can be implemented either on the same element or by some other element at egress of the region. In terms of operation, egress elements do the following:

— for all admitted traffic, data path mechanisms count the number of explicit congestion modification (ECN) marks, producing a load report per ingress gateway (for further information on ECN, see the Appendix),

— internal control path mechanisms query the data path mechanisms to get the appropriate load report and send it to the requesting ingress.

- Intra-region elements

  These carry out the standard tasks performed by ordinary core routers; hence they only implement data path mechanisms. In particular, these mechanisms do the following:

  — classify incoming packets into guaranteed or non-guaranteed classes, depending on the differentiated services code-point (DSCP) header field,

  — depending on the service class, mark the ECN bits of incoming packets (or drop them if appropriate) according to a certain marking algorithm,

  — schedule packets with a certain queuing management mechanism which gives guaranteed traffic strict higher priority over non-guaranteed traffic.

2.4 GQS operation

GQS can be deployed within one or more core network domains. A set of domains that collectively implement GQS is a GQS region. As far as the end-to-end (QoS-enabled) application service is concerned, a GQS region is perceived as a single network resource. The routers at the edge of a GQS region, which intrinsically form a ring topology, play a special role, in that they are the network elements responsible for synthesising the guaranteed services. We refer to these routers as GQS gateways.

Figure 2 shows a typical GQS arrangement; a few data flows are shown entering or leaving each gateway, representing its attached access network. For clarity these flows are not shown crossing the core, except for one, which is highlighted along its length. On the outer, access network side of each gateway, any traditional QoS solution (e.g. bandwidth brokers or IntServ) can be used. With BT's current network design, the GQS domain would stop at the metro node, but, if IP-aware DSLAMs where introduced, the edge of the GQS domain could become the DSLAM.

However, in order to describe the overall operation of a GQS system, we need to first decide the end-to-end control path transactional model as this affects the internal control path mechanism. How end-to-end QoS reservations are handled in the access network is unimportant, as GQS can cope with any model; but to make the description concrete, we assume SIP ('a' in Fig 2) is used to co-ordinate the end applications and agree the QoS parameters and RSVP is used as the end-to-end QoS control path mechanism. In particular, RSVP is used so that the ring-fence of GQS gateways are enabled to intercept and process RSVP QoS messages,

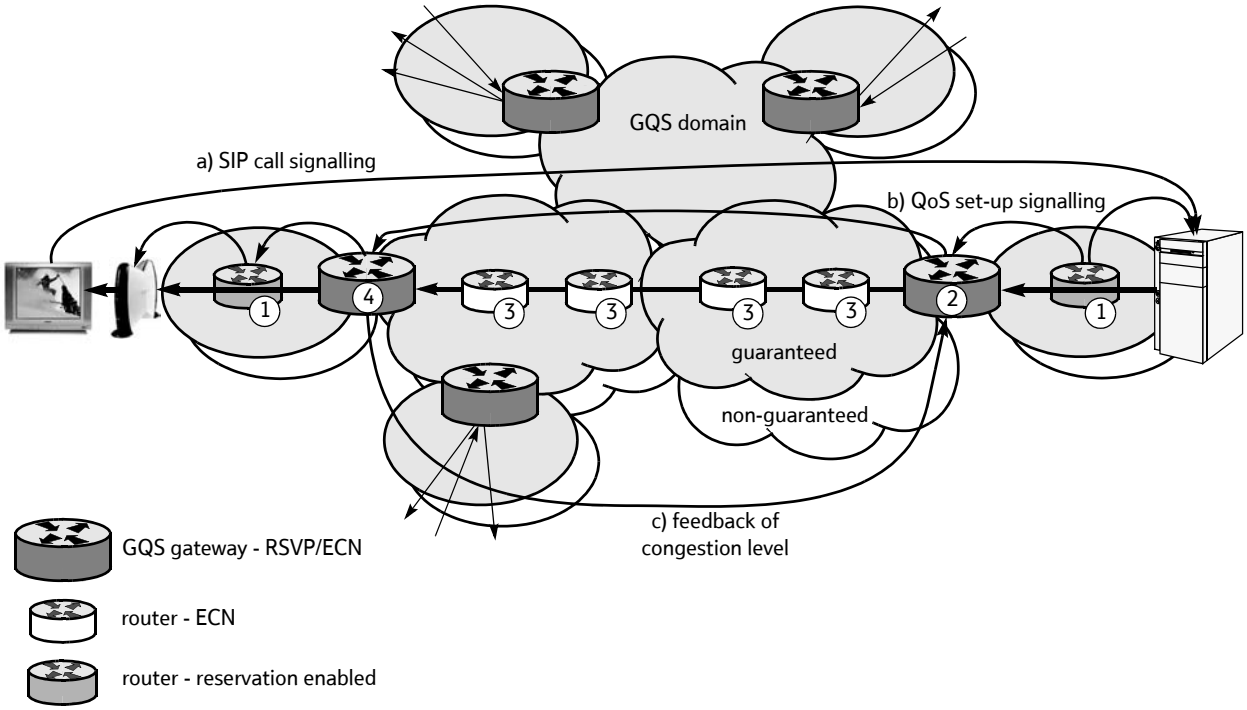


Fig 2 Example of GQS system.

whereas intra-region elements are not (i.e. RSVP messages are opaque to these elements and hence silently treated as data packets in the interior).

The resultant end-to-end QoS transactional model is the traditional one of RSVP, with one exception. The data sender prepares routers along the data path by announcing the flow specification it intends to send, with each hop passing its address to the next hop ('b' in Fig 2), the exception being that, within a GQS region, only gateways can intercept RSVP messages — hence RSVP treats the whole region as a single hop. After reaching the data destination, a response returns back along the same set of routers (not shown in Fig 2). Again, because all intra-region routers cannot see RSVP, the whole region appears as a single reservation hop, with the egress gateway sending its response straight to the address given earlier by the ingress gateway. If the end-to-end RSVP signalling exchange completes successfully, reservation state is added to each gateway so that data path processing can commence. It must, however, be noted that GQS does not depend on RSVP and any signalling system can be used to the GQS gateway to request a guaranteed connection.

The various data path processing steps applied to this flow are represented by circled numbers. In access network equipment, step 1 represents traditional policing of the data to keep it within the reservation. The GQS gateways keep guarantees by only allowing data matching an accepted reservation to be tagged with a DSCP chosen to represent 'guaranteed'. Any traffic not under a reservation, including traffic with a bit rate in excess of that reserved for it, is re-classified (i.e. downgraded) to the non- guaranteed class of service before being allowed into the region by the admission control mechanism (step 2). This is just standard traffic policing and re-classification — no different from that used in DiffServ except that all guaranteed traffic is also marked as ECN-capable (otherwise it would be dropped rather than marked by interior routers in the event of congestion onset).

In the data path of all intra-region elements, guaranteed traffic is given strict priority over other classes and allowed to pre-empt the place of other traffic in shared buffers if they are too full[2]. If any intra-region router experiences congestion, it will mark a proportion of all the guaranteed packets it forwards with ECN (step 3). A proportion of best-effort packets will similarly be marked if ECN capable, or dropped if not. Note that ECN marking has nothing to do with flows, of which intra-region routers are unaware. The proportion of packets ECN marked is determined via a virtual queue mechanism that has the effect of predicting congestion, i.e. the virtual queue becomes congested earlier than the real queue because it is emptied slightly slower than the actual line rate. We are currently investigating whether these techniques can be supported by today's core routers — indications are that with a firmware update they will potentially be able to.

On reaching the egress GQS gateway, the fraction of ECN marks in arriving guaranteed traffic is metered and stored (step 4). A load report is produced and stored for the aggregate of traffic from each upstream GQS gateway as long as at least one reservation is active.

Upstream admission control (step 2) is determined by this congestion metric. This arrangement is called measurement-based admission control (MBAC), but previous MBAC schemes have been confined to a single node. With GQS, the congestion measurement is accumulated along the path across the region, and fed back as a load report to the ingress, where admission can be controlled. The internal control path mechanisms are responsible for feeding back load reports ('c' in Fig 2); but with our example of using RSVP as the end-to-end control signalling, the load report can be piggy-backed on the RSVP response message.

If the ECN fraction of traffic on the path from the relevant upstream gateway exceeds a fixed threshold, a new reservation request will be denied. If required, this threshold could be different for different services so that some call-acceptance differentiation during congested periods is achieved. If a new request arrives between a pair of gateways where no other active reservations are in place, sufficient probe packets are sent across the ring to establish the ECN fraction for that path before admission control continues. Probe packets are those data packets using the guaranteed service that are injected at the ingress destined for the egress gateway. Their only function is to enable an estimation of the congestion between an ingress and egress gateway to be generated if no guaranteed traffic is flowing between these particular gateways.

The result of the above is that GQS pushes all per-flow complexity out of the set of core networks that form the core of the Internet, including interconnect points. The complexity of the ingress GQS gateways grows with flow volumes at the same rate as a DiffServ node's complexity grows (but DiffServ requires an edge router, to perform policing, etc, at every network boundary, whereas this is not required in the GQS architecture). Flow classification and metering at egress GQS gateways is an additional function not present in comparable solutions. But the prize is that neither core nor interconnect routers have any awareness of flows

[2] There is no need to pre-empt *ongoing transmission* of non-guaranteed traffic when a guaranteed packet arrives, given the vanishingly small per-packet transmission delay at core link speeds.

only requiring queue management functions working on bulk data.

The egress gateway records the congestion between a particular ingress and itself; further consideration is required where there can be more than one route between a particular ingress and egress gateway, as would be the case when the intelligent gateway protocol (IGP) equal-cost multi-path routing is employed within a particular network, or border gateway protocol (BGP) load sharing between networks within a GQS domain.

It can also be seen that GQS is based on three standard Internet protocols, but all used in a different arrangement to that for which they were originally designed:

- a reservation signalling protocol such as RSVP [2] is used, but in a scalable arrangement unlike the original integrated services architecture [1],
- DSCPs [3] are used, but not the complexity of service level agreement handling in the DiffServ architecture [4],
- ECN [5] is used, but not in its original end-to-end congestion control architecture.

In all cases, we have not contravened the standards, because the architectures that we avoid using are merely informational — it is the protocols that are standardised. An informational IETF draft is, however, being considered to document the GQS technique.

3. GQS deployment issues

It is unrealistic to expect the complete Internet to deploy GQS overnight (or even ever) and hence the following three issues associated with the incremental deployment of GQS need to be understood.

3.1 Interconnect between GQS networks

Figure 3 shows a scenario where some interconnected networks have adopted GQS, while others have not. In practice deployment will start with GQS gateways around just one operator's core (which has been upgraded with the techniques described). As other operators take up the solution, they will interconnect at the bulk packet level rather than the flow level. No connection-oriented (CO) gateways will be needed at the interconnect points, as bulk congestion charging will be sufficient. The GQS gateways form a ring surrounding a set of interconnected connectionless networks (i.e. no gateways are needed between networks), whereas all other core admission control solutions need a connection-oriented gateway even between networks that use the same technology.

3.2 Interconnect with non-GQS networks

To connection-oriented networks outside the ring, the whole ring appears to be one single reservation hop. Therefore more than one ring can exist on a path, as shown in Fig 3. At the interconnect between a GQS ring and any other connection-oriented approach a connection-oriented gateway will be required. In all cases, gateways would implement the GQS gateway functions on its GQS side and the appropriate functions of the alternative approach on its other side (reusing any functions common to both).

3.3 Compatibility of end-to-end signalling

End-to-end signalling further breaks down into signals initiating application sessions, and the resulting signals to reserve QoS. One might imagine that the two could be achieved together, but the first involves the session initiator finding the correct destination, which is a prerequisite to finding the resources on the resultant path to the destination in order to reserve them. Therefore

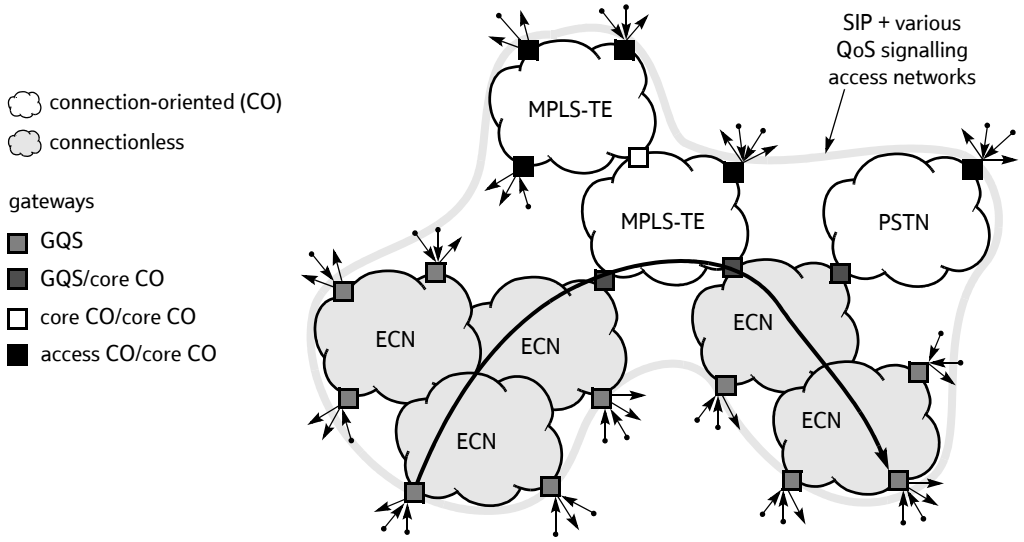


Fig 3 Interworking during incremental deployment

session signalling cannot be assumed to be along the data path, while reservation signalling must be.

- Session signalling

  Session signalling finds the correct destination for a session by resolving an abstract correspondent address into a process address on a particular device (allowing for user mobility, call forwarding, etc). Also, the required QoS is negotiated between the correspondents at this stage. Resource reservation and therefore the GQS cannot be involved until these two tasks have completed. The session initiation protocol (SIP) has become the predominant standard that fulfils these purposes across the Internet, although its QoS negotiation is still in the process of standardisation. For end-to-end sessions across non-Internet technologies (e.g. to the PSTN), gateways are required to convert the signalling formats and functions.

- Reservation

  Once the QoS specifications and the addresses of the ends of the flow(s) have been established, finding, and then reserving, the resources for each flow can commence. One mechanism is that RSVP PATH signals can trace the path the data will take by following normal data routing, but alternate routing can be forced. RSVP signalling is designed to cater for various resource reservation models in each domain it encounters. In some domains the signal may touch every resource along the data path as in the integrated services architecture [1]. In others (e.g. in the policy-based admission control framework [6]), when the QoS request arrives at the ingress router, it may trigger a request to a centralised policy decision point that acts on behalf of all resources in the domain. In parallel, the request will continue to the next domain. On its return, the RSVP reservation response co-ordinates the responses from the resources it touched on its way out (including the results from any requests delegated to centralised servers), and communicates the overall success or failure of admission control to the source and to each resource.

  To RSVP a GQS domain appears much as if it were operating along the lines of the policy-based admission control model. The ingress GQS gateway acts on behalf of the whole set of resources on the path across the ring of gateways — RSVP does not know and does not need to know that the gateway may be responding on behalf of multiple domains. The end-to-end reservation also allocates resources in domains before the ingress GQS gateway and continues to reserve resources in domains beyond. Even if there are other GQS rings on the path, RSVP just treats each as one more reservation hop.

  If each network used different QoS signalling, the gateways would have to recognise and convert the message formats. Of course, this may not be straightforward if one signalling system only implements a subset of the functions of another.

We must add that no examples of interworking gateways have been designed in detail. The above points are merely statements about what is likely to be possible. It must also be stated that GQS does not rely on RSVP but could be interfaced to any signalling mechanisms.

4. GQS operational support features

GQS has been designed to minimise operational support costs; the following indicates how it achieves this in three areas.

4.1 Capacity allocation and configuration

GQS was deliberately designed to remove any need to configure and manage capacity allocations between guaranteed and non-guaranteed responsive traffic on any core routers. Guaranteed and non-guaranteed traffic automatically balance their shares of all resources throughout the network. Any capacity not being used for reservations can be borrowed by elastic traffic. As soon as reserved traffic needs the capacity, it is strictly prioritised over elastic traffic. Just one parameter, set equally on every ingress GQS gateway — the threshold congestion level — determines the balance between guaranteed and non-guaranteed traffic throughout the network. If desired, different thresholds can be set for different types of reservation, perhaps depending on the revenue they attract. But the set of thresholds would be the same on each GQS gateway. Eventually it will be necessary to design modifications to network management tools to set common thresholds for different types of reservation across a number of GQS gateways.

In practice, however, to avoid best-effort starvation, which potentially could occur under very extreme traffic conditions, a small proportion of the link capacity is exclusively allocated to best effort, this is achieved by router buffer configuration. Similarly a small proportion of the link capacity is taken out of the ECN feedback mechanism which has the effect that guaranteed traffic will be admitted even when a particular path through the network is saturated by best-effort traffic.

In an over-provisioned core network, complicated traffic matrix predictions have to be produced to enable the network to be designed and deployed with a degree of certainty that ‘normal’ traffic can be carried without

suffering congestion. In a GQS network these traffic matrix predictions are not so important because congestion notification through all the paths automatically adapts to the prevailing pattern of demand.

4.2 Capacity provisioning and planning

Congestion signalling statistics accumulated from each network interface are a highly valuable input to procedures for planning capacity growth. Congestion signalling can be triggered long before true congestion arises, so that capacity provisioning can always be put into effect well ahead of need. This is achieved through the use of early marking — using virtual queues as described earlier. This does not cause sessions to be blocked unnecessarily because admission control thresholds can be set to trigger blocking at a higher notification rate which takes account of this early congestion notification. Through the interpretation of congestion signals as shadow prices [7], these signals correctly reflect the relative value of demand from each resource and from differing services, and therefore provide a direct and unambiguous indication of exactly where upgrades are desired. However, we are not saying that congestion signalling can replace longer term planning to take account of demographic trends, etc, before any market effect is noticed. Therefore capacity planning tools would need to be developed, for which congestion notification statistics would be the main but not the only input.

4.3 Resilience to failure

GQS was also deliberately designed to ensure reservations survive failures within the core. Reservations are decoupled from core routing rather than pinned to it, so that once routing re-converges to bypass a failure, the reservation continues, strictly prioritised over any unreserved traffic on the new route in order to make way for itself. However, this only works if the new route has sufficient capacity along its length for the sum of previous reservations and the re-routed ones. As long as non-guaranteed traffic forms the majority share of demand this will be the case. However, a catastrophic failure can cause reserved traffic from multiple routes to crowd into a route with insufficient capacity for all the guarantees. In this case, as soon as sufficient customers with reservations give up their degraded sessions, the rest of the sessions will immediately recover. Over time, as sessions depart naturally, new sessions will be blocked until the correct balance for the new network topology has been reached. Once the original route is repaired and traffic returns to it, more new sessions will be allowed in until the original balance is reached again. Planned downtime of core equipment can be scheduled when demand is low, so that re-routed reservations will survive, in a similar way to above. If the equipment cannot be returned to service in time for the next peak period, guaranteed and non-guaranteed traffic will automatically find their shares of the alternative capacity as above.

We have not designed for failure of a GQS gateway (or planned downtime). It would seem possible to use a replicated pair of machines for each gateway if such resilience were required. Solutions that replicate dynamic state and automatically switch traffic to the hot standby are available, but we have done no design in this area.

5. Charging models for GQS

Figure 4 shows an example value chain built around GQS gateways. The important feature to note is that a different product is traded outside the GQS gateways than inside. GQS gateways are designed to be placed in a ring around the core networks. Therefore the single unidirectional data flow from customer A to B that we focus on in Fig 4 passes through one gateway to enter the ring, and another to leave it. Within the ring an example string of interconnected tier 1 and tier 2 providers are shown. An access network is shown at each end of the flow, outside the ring.

Outside the ring of gateways, guaranteed bandwidth flows can be sold by network wholesalers to retailers and by retailers to customers. Thus, looked at from the outside, familiar business models like those we find for telephony, for ATM or for Frame Relay connections, can be preserved. For instance, in Fig 4, the charges from each access network are met by a clearing broker, which collects revenue from the sending customer to cover both access network charges (and its own costs and profit)[3]. To emulate classic telephony charging, flows in each direction would be reserved together, and the broker would charge the originating end for both.

Of course, other charge reapportionment regimes different to that shown in the figure can be created across edge networks. The essential point is that the combination of both a ring of GQS gateways and a suitable broker cleanly isolates the business model of the QoS interconnect market from the complications of a lively retail market. For VoIP there is currently no interconnect QoS charging model (interconnection being done via conversion to PSTN), but when a QoS IP interconnect regime is introduced it is likely that the QoS revenue from each session would have to be shared

[3] In Fig 4 we have deliberately separated each role in the value chain. In practice, a player in the market-place might well take on more than one of the adjacent roles. For instance, an access wholesaler may also act as an end-to-end clearing broker. Indeed, the same access wholesaler may also operate a GQS gateway.

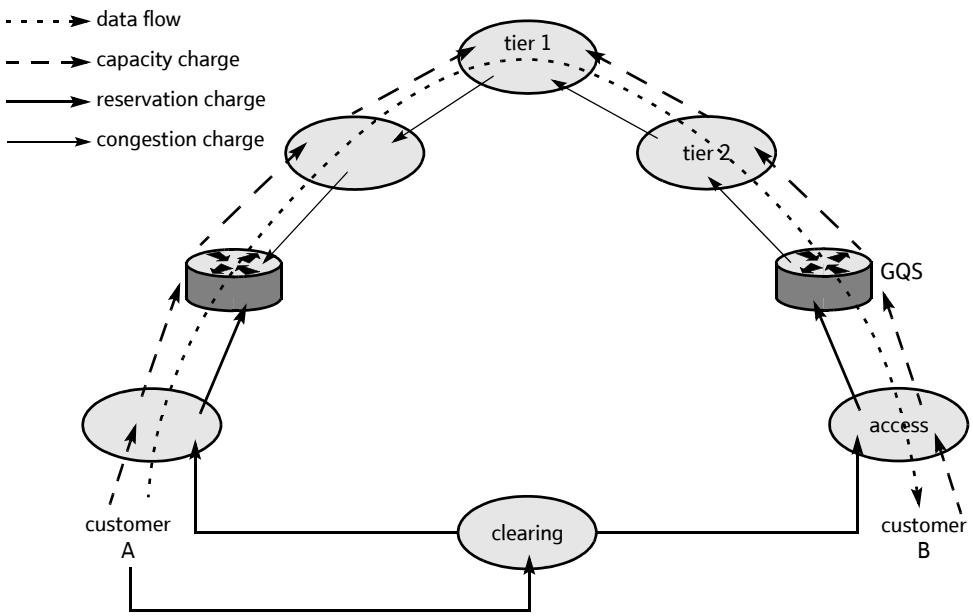


Fig 4 Example value chain around guaranteed QoS synthesis.

across intermediate networks. The settlement clearing-house would therefore have to understand routing to find which networks to credit. However, GQS would enable the simpler model to be preserved.

Inside the ring, however, no trace of any connection or flow is apparent. The data packets from each flow are all carried together in bulk, with no need to classify them into flows. Most significantly, no flows are apparent at the points interconnecting networks. Instead, at the interconnect points within the ring bulk congestion charging could be applied. Otherwise GQS gateway operators would have no incentive to block revenue-earning flows before congestion risked damaging quality.

By congestion charging, we mean that, as IP packets cross the networks within the ring, each router flags them with standard 'congestion experienced' marks with a small probability related to congestion experienced at that router. Consequently, the proportion of marked packets at any point represents how much upstream congestion has been experienced, i.e. ECN marks accumulating as data traverses congested resources. Congestion charging involves charging for the bulk volume crossing each interconnect point, but only counting marked packets. Such bulk accounting is as cheap to operate as bulk charging by volume, which is already becoming common because of its low cost. Congestion charging serves the dual role of providing all the correct incentives for each party to respond correctly to congestion **before** it degrades service, and over time it focuses revenues on network resources equal to their required upgrade costs.

The above value chain of products — bandwidth guarantees created from bulk congestion pricing — is entirely concerned with **usage** charging. The outer arc of the money flows in Fig 4 also shows how capacity charging would be included in the value chain. We make this point to emphasise that usage charging is complementary to, not a substitute for, capacity charging. Incidentally, capacity charges are shown being paid in the direction towards the networks giving the greatest extra connectivity, as is usual.

We should also clarify that non-guaranteed service can continue to be sold Internet-wide 'underneath' the gateways — as a distinct service — whether sold flat rate, by volume, or whatever. In other words, the proposed guaranteed service complements the best-effort Internet, although the two are deliberately not assigned hard partitions of capacity. The GQS is designed to automatically find the optimum share of capacity that each service (guaranteed and non-guaranteed) should use. The share will change from one minute to the next and from one path through the network to another. Note that non-guaranteed traffic is not necessarily lower value than guaranteed; it is just a different mode of service where, once a guarantee has been assigned, it must be kept.

6. Summary

This paper has discussed the requirements for an economic QoS mechanism in core IP networks that performs call admission control so that services like VoIP will not suffer degradation in periods of unpredicted congestion. A BT-developed mechanism called guaranteed QoS synthesis has been described that uses

measured congestions in the network as the basis of call admission decisions. Although GQS is still at the embryonic stage, it shows enormous potential as a way to reduce complexity and costs in the network while providing bandwidth guarantees.

GQS will, however, never be deployed everywhere on the Internet and hence an incremental deployment mechanism has been devised where GQS domains can interwork with other core IP core QoS mechanisms.

Finally the business models that GQS makes possible are introduced. These can be a complete separation of retail and core IP business model with inter-provider interconnect settlements based on a mix of capacity and bulk congestion metrics.

Acknowledgements

The authors wish to thank the whole GQS team of BT's Networks Research Centre at Adastral Park for their work on GQS, feedback and suggestions, which have contributed to the research described here. Also thanks to the M3I team, Dr Martine Karsten et al, for their original work in this area.

Appendix

Explicit congestion notification (ECN)

Before ECN, the only way a router could signal its congestion was by dropping packets. ECN was designed to allow a router to signal that it was approaching congestion by marking packets, thus allowing early avoidance of both congestion and retransmission delays. ECN involved redefinition of the IP packet header itself (specifically the last two bits of the differentiated services byte in both IPv4 and IPv6).

Each ECN-capable router probabilistically marks packets in proportion to the severity of the prevailing congestion as they enter its egress queue. The random early detection (RED) algorithm is used to determine the likelihood of marking each packet, dependent on the moving average of the recent (exponentially weighted) queue length. The simple RED algorithm is applied equally to all the packets arriving at an egress router interface, with no regard to flows.

In summer 2001, ECN was accepted by the Internet Engineering Task Force as a Proposed Standard, although it had already been implemented for some time by the major router manufacturers. Standardisation of ECN was a significant event in the history of the Internet, given that it is a far more robust way to achieve closed-loop control at the packet level than loss-detection, and given that designs using closed-loop control tend to be far simpler than open-loop.

References

1 Braden R, Clark D and Shenker S: 'Integrated Services in the Internet Architecture: An Overview', IETF RFC1633 (June 1994) — http://www.ietf.org/rfc/

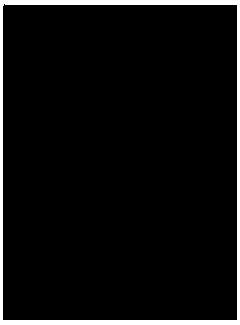
2 Braden R, Zhang L, Berson S, Herzog S and Jamin S (Eds): 'Resource ReSerVation Protocol (RSVP) — Version 1 Functional Specification', IETF RFC2205 (September 1997) — http://www.ietf.org/rfc/

3 Nichols K, Blake S, Baker F and Black D: 'Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers', IETF RFC2474 (December 1998) — http://www.ietf.org/rfc/

4 Blake S, Black D, Carlson M, Davies E, Wang Z and Weiss W: 'An Architecture for Differentiated Services', IETF RFC2475 (December 1998) — http://www.ietf.org/rfc/

5 Ramakrishnan K K, Floyd S and Black D: 'The Addition of Explicit Congestion Notification (ECN) to IP', IETF RFC3168 (September 2001) — http://www.ietf.org/rfc/

6 Yavatkar R, Pendarakis D and Guerin R: 'A Framework for Policy-based Admission Control', IETF RFC2753 (January 2000) — http://www.ietf.org/rfc/



P Hovell

Bob Briscoe joined BT in 1980 and now directs the research programme of BT's Networks Research Centre. In the late-1980s he managed the transition to IP of many of BT's R&D networks and systems. In the mid-1990s he represented BT on the HTTP working group of the IETF and in the ANSA distributed systems research consortium, which led to the creation of the OMG and CORBA. In 2000 he initiated and was technical director of the Market Managed Multi-service Internet (M3I) consortium, a successful European collaborative project investigating the feasibility and user acceptability of controlling Internet quality on fast time-scales through pricing.

His published research, standards contributions and patent filings are in the fields of loosely coupled distributed systems, scalable network charging and security solutions (especially multicast), managing fixed and wireless network loading using pricing and on the structure of communications markets. He is also studying part-time for a PhD at University College London.

Gabriele Corliano holds an MSc in Computer Engineering from the Polytechnic of Turin, awarded in October 2000.

In July 2000, he also gained a diploma in Telecommunication Engineering from the Eurecom Institute (Sophia Antipolis, France). He first worked for Motorola in the domain of Personal Communications.

As a software designer, he participated in the design of end-to-end IP over EDGE communication systems, and then in the design of GSM-UMTS dual mode cell reselection.

He joined BT's Research unit in June 2001. Working in the Edge Laboratory as a senior research scientist, he is currently investigating the domain of technical and contractual mobility in next generation communication networks.