Commercial Models for IP Quality of Service Interconnect

Bob Briscoe
<bob.briscoe@bt.com>

Steve Rudkin
<steve.rudkin@bt.com>

02 Jun 2005

Abstract

Interconnection of IP QoS capabilities between networks releases considerable value. In this paper we show where this value will be realised. We give technical and economic arguments for why QoS will be provided in core and backbone networks as a bulk QoS facility incapable of distinguishing or charging differentially between sessions. While between edge networks a vibrant mix of retail QoS solutions will be possible, including Internet-wide per-flow guarantees.

We outline cutting edge research on how to coordinate QoS between networks, using a session-based overlay between the edges that will extract most surplus value, underpinned by a bulk QoS layer coordinating the whole. We survey today's interconnect tariffs and the current disconnected state of IP QoS. Then we describe a commercial 'model of models' that allows incremental evolution towards an interconnected future.

The paper covers intertwined engineering and economic/commercial issues in some depth, but considerable effort has been made to allow both communities to understand the whole paper.

1 Introduction

Interconnection of networks allows every individual and every device on one side to connect with every individual and every device on the other, immediately releasing huge reserves of value for very little cost. Interconnecting the quality of service (QoS) capabilities of networks releases similar reserves of value, again for relatively little cost.

Without QoS interconnection, customers have to choose between service providers by trading-off features and price against inaccessibility of some of their favourite contacts. After interconnection, the market share of each provider is no longer a differentiator for potential customers. So, although market share is still important to each provider, they have to focus more on features and price.

This paper aims both to describe the present landscape and to predict its evolution. We describe the cutting edge of new technical and economic research that unifies the whole area of IP QoS, both technically and commercially. This allows us to predict how economic forces will drive the IP QoS industry over the next decade or so.

In engineering terms, the main issues in IP QoS interconnection are the scalability of trust and security mechanisms at boundaries between operators. In business terms, the main issue is to define a flexible commercial model that allows evolution in how both revenue and costs can be shared between players who are all trying to maximise profits.

We present a model that simultaneously solves all these engineering and commercial problems, allowing a range of value-based charging options for QoS to co-exist around the edge, bounded at the lower limit by a floor of cost-based signals that extend beneath the whole internetwork. We explain why it is inevitable that QoS margins will be most squeezed in backbone networks — a squeeze that will grow outwards. This justifies our model, which can coordinate the engineering and commercial aspects of QoS between the edges, even if the middle is only interested in covering its costs.

We start (§2) with an analysis of where QoS management will and won't be needed, why QoS interconnect is valuable and how much value it will release. §3 describes the very disconnected state of the art and why it is that way, ending with an attempt to show that there is some similarity between all the diverse approaches we see around us in the industry.

The main body of the paper is then divided into two sections. §5 takes an engineering perspective, explaining the issues in coordinating service between operators. While §6 takes an economic/commercial perspective to coordination of interconnected networking businesses. The service coordination section describes issues with the state of the art, and ends with the results of recent research that unifies IP QoS technology, based on an understanding of the economic issues. The business coordination section explains the issues of cost and surplus value in QoS provision and where competition will erode margins fastest. We survey the state of the art in interconnect tariffs commenting on their robustness relative to more ideal tariffs.

Before drawing conclusions, a final section (§7) recommends an industry structure that will allow natural selection to evolve the tariffs and interworking necessary to provide interconnected IP QoS.

2 Market need?

The need for IP QoS, and the applications it might enable, is well rehearsed elsewhere [25, 1]. To briefly summarise, QoS is primarily required for interactive streaming between humans, which may be a small proportion of long-term future demand, but in the medium term it could make up about half of the traffic on a converged IP network. QoS is also desired for the IP networks of enterprise customers to assure lower utilisation than is typical on the public Internet. The focus of this section is purely whether there is a need for IP QoS capabilities to be interconnected.

2.1 Access, core or both?

Providing differentiated QoS primarily concerns managing the risk of congestion.[1] So there will only be a market for discriminating QoS provision if there is some tangible risk of congestion.

Although unattended computing applications such as peer-to-peer file-sharing don't themselves place stringent QoS demands on a network, they do search out all available capacity. As we move to providing all communications over a multi-service IP network, this profligate demand creates a market for differentiated QoS for other, interactive uses of the same network. Economies of scale for provision of capacity will always lead to access bandwidth being more costly than core (Fig 1).[2] So whenever demand exceeds capacity, the bottlenecks it encounters will invariably arise in access networks.[3]

Initially network operators avoided the need for QoS across interconnected networks, solely creating products to differentiate IP QoS in their own access networks. Such 'balkanising' behaviour is a classic initial approach of network businesses. It attacks the problems that have least external constraints in order to capture some early revenues. IP QoS interconnect was avoided in this initial phase by targeting applications that didn't require real-time interaction across networks.

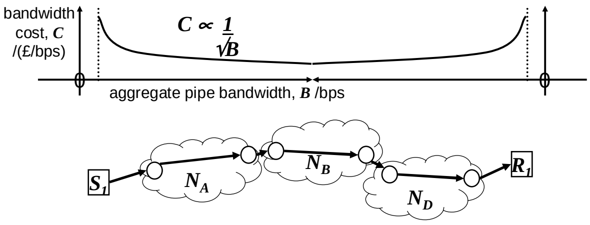

Figure 1: Scaling of bandwidth cost with aggregate pipe bandwidth. The mirrored plot is shown merely to highlight where the costs lie along a typical end to end path shown below.[5]

So, the initial focus in the home user market has been on QoS for content delivery, where best efforts interconnect is merely used to fill local caches in advance, thus avoiding any need for interconnected IP QoS.

In the business market, the most pressing initial problem was to reduce costs by connecting together multi-site businesses over a converged IP network rather than using costly private wires. But, the general levels of congestion on the public Internet were too great for business-critical applications.[6] So virtual private networks were built, engineered for lower utilisation than the public Internet. Again, QoS interconnect at the IP level was avoided by purchasing underlying link capacity from wholesalers around the world, so that a single operator could run IP virtual private networks (VPNs)[7] over dedicated logical circuits.

So, to summarise so far, by a general cost economics argument, access networks will always need QoS control more than core networks, but QoS control is also needed in core networks for more demanding (enterprise) customers willing to pay more per unit bandwidth than the average public Internet user, in order to avoid congestion effects. We have also shown how early IP QoS products have targeted applications that avoid interconnection.

[1] Other concerns are minimisation of propagation delay by shortest path routing and maximising availability. But discriminating QoS along these dimensions is a more niche business for some high-value customers.

[2] Geographical dispersion is inherent to access networks, leading to inherently higher operations & maintenance and capital costs. For instance, switch/router interface costs dominate capital costs and, at least for a future all-optical network, the cost of a unit of capacity depends on the capacity of the interface from which the capacity is partitioned by an inverse square root law [28] (i.e. a unit of capacity from an interface one quarter the size will cost twice as much).

[3] But core networks are not immune [25, 22].

[5] No implication that the path has linearly increasing then decreasing pipe bandwidth is intended.

[6] And confidentiality was required, but that is outside the scope of this paper.

[7] Typically also using multi-protocol label switching (MPLS)

2.2 The value of interconnected QoS

Metcalfe's Law [32] predicts that as the number of end-points N grows, a network's value rises with the square of N, that is $O(N^2)$. So interconnecting previously isolated QoS capabilities would seem to release huge amounts of value. But Lyons [30] points out that this 'law' is based on the flawed assumption that every single individual values connectivity equally with each other individual in the world. Having reworked the analysis [6], we have arrived at the far less optimistic result: the value of an internetwork grows more like $O(N \ln N)$. For instance, connecting the QoS of a small network of three million users to an existing QoS internetwork of seven million users would only increase its value from about 28% to 30% of the market's value.

But, and it is a big but, this more pessimistic result only applies where the two networks address distinct markets with no particularly strong community of interest across them (networks in Eire and Romania, say). In the special case where two networks compete for exactly the same market (e.g. a cable and a DSL network covering the same city), if each network were isolated with market share λ, its value would be of the order of the square of its share, $O(\lambda^2)$ (the same result as Metcalfe's Law but for different reasons). In the above example with a $3 : 7$ split, but dividing a single market rather than ostensibly separate ones, the smaller network would be worth about $30\%^2 = 9\%$ of the whole when isolated, but 30% when interconnected.

In summary, where a market is ostensibly a single community of interest, but it is carved up between competing network operators, the value to each operator of one interconnected IP QoS capability across the whole market is far greater than each could realise with isolated QoS products. And since IP QoS is most pressing for access networks, QoS interconnection will require coordination of QoS management between access networks. Paradoxically this implies that an access-dominated company like BT should invest most effort in QoS interconnection with its fiercest access competitors (e.g. the UK cable operators).

This raises the interesting question of whether the business of QoS coordination will be conducted through the intervening backbone networks, or whether an overlay market will evolve (transparent to intervening networks) that will coordinate access network QoS 'edge-to-edge'. We will return to this question later in §7 on market structure.

3 Disconnected diversity: the state of the art

In 1993 the Integrated Services effort commenced, aiming to specify a single approach to quality of service for the whole Internet. The first commercial implementations appeared in 1996 and their performance was dreadful. Since, properly crafted implementations have been built with decent performance. But the damage was done. Further, the Integrated Services Architecture (Intserv) included no ability to aggregate in core networks [2]. So unfortunately every aspect of the effort became associated with the word 'unscalable' (Mustill & Willis [33] expand on this story). In particular the reservation protocol (RSVP [43, 4]) became tarred with the same brush. So more recent architectures with aggregation capabilities still suffer accusations of poor scaling, simply because they use RSVP, which some confuse with the unscalable Intserv architecture[8].

The above is not just a cautionary tale against the use of buzz-word engineering in system design. It is an explanation of why IP QoS is in such a fragmented state. The success of the Internet was because it offered a single overlay internetworking technology. It didn't glue together lots of different packet networking technologies side by side using gateways — it replaced the few that did exist with a more generally useful abstraction. But in the last few years, probably due in large part to the failure of the Intserv effort, different operators and different sectors of the industry have all chosen different IP QoS solutions for their access networks:

- Many access technologies that can carry IP are built over ATM technology (GPRS, UMTS, DSL, Satellite DVB). Some operators still choose to use the dedicated virtual circuits that ATM provides to assure the QoS of IP data. However, many are moving to QoS solutions at the IP layer, both in order to take advantage of the economies of packet multiplexing and to avoid being tied to particular access technologies (e.g. some are considering moving from ATM to Ethernet).
- In 1997, a new tactical approach was proposed by Clark [11], which became standard-

[8] There is also confusion over the term 'scalable'. Some engineers use it to mean that a box can be built with current technology that meets current demand at a price that customers are willing to pay, and as demand increases there is a modular way to add more boxes. In computer science, scalable means that computational complexity grows less quickly than growth in demand on the system (so if demand doubles, less than twice as many boxes will be needed). Intserv is termed unscalable because in core networks it exhibits linear growth of complexity with demand.

ised as the Differentiated Services architecture or 'Diffserv' [3]. This was aimed primarily at solving the immediate and high-value problems of IP QoS in enterprise networks. In access networks, Diffserv is often chosen as a substrate to differentiate aggregate classes of service, over which other QoS solutions are used to manage per-session QoS (e.g. the bandwidth broker below).

- In the late 1990s the cable industry chose the Intserv approach in its access network specifications [9, 8] using Diffserv in backbone networks.

- There is no common approach to IP QoS across the DSL access operators. While some have used the standard IETF QoS capabilities still available in routing and switching equipment, others have chosen various proprietary centralised bandwidth broker approaches (see next item).

- A bandwidth broker is a central server (per domain) that arbitrates access to a statically provisioned logical partition of the network's resources. Typically Diffserv is used to create a logical partition of the network's resources as the substrate over which the bandwidth broker works. All session requests are directed to the bandwidth broker, which holds a map of the network and keeps track of utilisation of each resource.[9] The bandwidth broker idea is reminiscent of how ATM switched virtual circuits were planned to be set up between providers. The idea was first proposed for IP in 1995 [34] and re-invented for the IETF in 1997 [36]. No bandwidth broker standardisation emerged from the IETF[10]. Proprietary bandwidth brokers became commercially available in 2003, but as Cuevas [15] confirms, still no clear inter-bandwidth broker standards have emerged.

- Some IP service providers have gone ahead with provision of telephony services using the best effort Internet, even though most operators have preferred to deploy QoS controls so that such services work more reliably.

- Similarly, some operators have invested heavily in raw access capacity (e.g. in Korea) to effectively avoid the need for access QoS.[11]

- And, of course, as different applications converge onto IP networks, we will still have to interconnect with the QoS of legacy access networks, such as the PSTN.

Another reason Intserv was deprecated was its lack of support for policy control over admission of sessions, which was required in both commercial and public sectors. In the late 1990s, the policy-based admission control architecture [42] was defined to allow interception of a QoS request to be redirected to a policy decision point (using the COPS protocol [16]) in order to apply admission policy. Commercial bandwidth broker solutions usually include (or promise to include) integrated policy control facilities.

Nowadays, operators rarely give their customers the freedom to make direct QoS requests to the network anyway. Instead, they are expected to make session requests to a session server, which makes the QoS request to the network on their behalf.[12] Intercepting the request at the application layer reveals more information about the user's intent, so policy control can be much richer. However, giving customers no other choice than this approach raises public policy concerns over both privacy and bundling.

All these approaches can rightly claim to be standards-based, but they all pick different standards from the wide variety available (Mustill & Willis provide a useful overview [33]).[13] So, given QoS is mostly needed in access networks, and interconnected QoS appears to release so much value, we have to find some way to connect all these different IP QoS approaches together.

Worse, on top of all these approaches, different operators have different ideas on what commercial model they will offer their customers to sell QoS. For a start, there are different markets to address: public Internet, managed networks, QoS bundled with high level services, and so on. Then operators will want different tariff models: subscription, quotas, usage.

So, in order to bring out the commercial issues in IP QoS interconnection, we will draw a line under

[9] Because bandwidth brokers deal with all session set up requests, they exhibit exactly the same lack of scaling (in the complexity growth sense) as Intserv. However, in the engineering sense, bandwidth brokers exhibit worse scaling than Intserv, because there is no natural way to distribute them across multiple machines.

[10] The Simple Interdomain Bandwidth Broker Signaling (SIBBS) proposals from the Internet 2 Qbone project were rejected as too immature

[11] On 29 Mar 2005, KT announced they would be moving to usage-based charging over the next two years, because even their over-provisioned access network is being congested by file-sharing traffic (there were also regulatory reasons).

[12] Usually ignoring any network QoS set-up required in the customer's network.

[13] Indeed, the choice is still widening. For instance, the next steps in signaling (NSIS) working group of the IETF is working on a more generic signaling architecture than RSVP for the Internet — whether it is ever widely adopted is another matter though.

this mêlée and reduce it to abstractions that allow us to understand the main technical factors that will have a bearing on commercial issues.

Given QoS concerns managing the risk of congestion, we are primarily concerned with the nature of the commercial relationship that the network provider has with end-customers who are in control of network load. The main distinction we need to draw is whether there is i) a direct or ii) an indirect relationship (through an intermediate ISP). In the latter case, the network provider cannot directly influence QoS in real time, so relies on static provisioning:

i) Load control (direct). If capacity that a user needs is congested, the network provides feedback intended to reduce demand. In networking control terms, this is called **closed-loop control**, because there is a direct feedback loop to the source of the load. There are two qualitative types:

Rate adaptation involves the application adapting its sending rate to the currently experienced congestion level. Such applications are termed elastic. Because this is how TCP works, few people consider it as a QoS mechanism as it is associated with the single QoS best-effort Internet. However, mechanisms now exist in IP for the network to give early notification of congestion before it affects QoS, so it has become feasible to provide better QoS by allowing the data rate of certain customers to respond less strongly than others to incipient congestion.[14]

Admission control involves the user requesting a capacity reservation for a session and the network returning feedback either accepting or denying the request. Reservations are necessary for inelastic applications.

ii) Static conditioning (indirect). This involves assessing likely demand on each of the separate links of a network and engineering the capacity assigned to a class of service on each so that congestion is rare. Customers can only be offered an assurance of high QoS if their traffic is conditioned to a certain profile (otherwise capacity cannot be engineered). Diffserv works on this model. Traffic entering each logical traffic class ('colour') is policed to this profile, with excess 'recoloured' to a lower class. This is called **open-loop control**, because there is no immediate feedback loop to control the loads applied.

Note that this distinction is categorised by what the customer is expected to do (response vs. contracted profile), not what level of QoS the customer gets. Whether the customer gets generally improved QoS, or guaranteed QoS would be another dimension of classification that would cut across this one.

For instance, to some degree guaranteed bandwidth can be provided through any of the three approaches, not just admission control. Allowing zero rate adaptation in response to approaching congestion[15] provides guaranteed bandwidth. Or guarantees can be offered without any end-user load control mechanism, just with static traffic conditioning.[16] Some ISPs propose to offer voice over IP (VoIP) this way. So static traffic policing could be categorised as a third form of load control for direct end-customers. Of course, with no immediate load control, there is a small chance that any network element might occasionally overload, causing random losses spread across all flows, potentially making them all fail [25]. That is why a statically provisioned network is often used as a substrate over which QoS based on load-control (rate adaptive or admission control) is added.

Our aim was to provide sufficient technical abstractions to be able to discuss the commercial issues of interconnection between IP QoS approaches without getting buried in the myriad of detailed differences across the industry, as given earlier. The above three-way categorisation will serve our purpose.

4 The problem

Throughout the rest of the paper we will use the scenario shown in Fig 2 to illustrate the issues.

[14] An explicit congestion notification (ECN [38]) field was defined within the IP header (v4 & v6) in 2001, whereas previously the only way to signal congestion was to drop packets. Also a so-called virtual queue [14, 27] can detect incipient congestion before it even causes queuing delay. A virtual queue is just a bulk token bucket being filled at the data arrival rate but emptied just below the line rate (e.g. at 99%).

[15] Up to a threshold that triggers admission control [22].

[16] The strength of such a guarantee depends on the topology required:

Pipe model If the customer is willing to be constrained to a point to point path, strong guarantees can then be given.

Hose model However, often customers wish to spread traffic to multiple destination (hose model), in which case, any assurances depend on the spread of traffic at any instant and are therefore prone to QoS failures. The more predictable is the spread of the customer's traffic, the lower the probability of guarantee violation for the same engineered utilisation [39, 25].

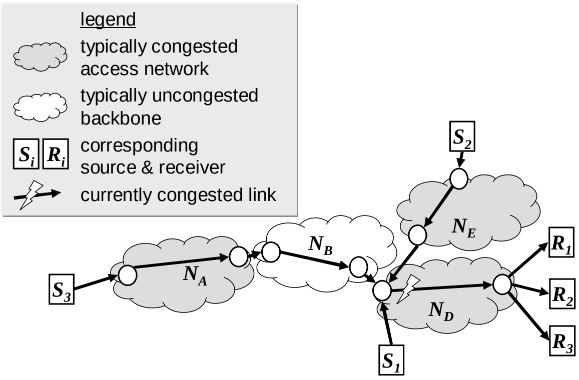
legend
typically congested access network
typically uncongested backbone
S_i R_i corresponding source & receiver
currently congested link
S2
NE
NB
NA
S3
R1
R2
R3
ND
S1

Figure 2: Interconnect scenario (topology only).

Three representative potential flows, S_1R_1, S_2R_2 & S_3R_3 are shown converging on a congested link through egress access network N_D. The potential flows traverse one, two and three networks respectively. This allows us to contrast the coordination of QoS between networks that focus strongly on managing QoS and other scenarios which include networks on the path that are not particularly concerned about QoS management.

By potential flows we mean there is demand for the flows but the networks have to decide whether to meet the demand, either by reducing their data rate or controlling admission of one or more flows. One network, N_B, is a rarely congested backbone. The others, N_A, N_D & N_E, are frequently congested access networks. As well as managing the potential congestion on the shared link, each flow shares capacity with other flows (not shown) in the other access networks. So, to determine its rate, or whether it is admitted at all, all these flows must be coordinated at once.

The next two sections use the abstractions from the previous section to develop an understanding of the two main issues that network businesses need to address in order to interconnect:

Service coordination involves coordinating all domains to collectively supply the required QoS — or to ascertain that they can't. With interconnect, flows from customers of other networks intermingle with those of a network's own customers. Networks and end-customers have to collectively determine the share of the congested link that each flow should receive, or whether it should be admitted at all - an apparently complicated capacity allocation problem.

Business coordination involves determining what to charge end-customers and how to split revenue — or refund for failure — between domains. This in turn comprises determining which networks earn what from each flow (but there is no implication that there has to be per-flow charging — even with bulk usage charging, decisions to adjust flow rates affect everyone's revenues).

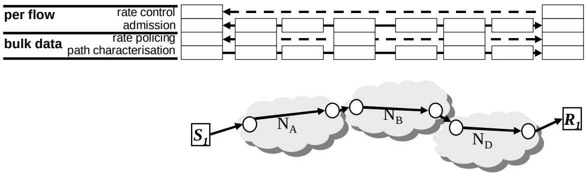

Figure 3: Layering of QoS control

It is clear that service and business coordination are intrinsically linked: if the networks agree to reduce the data rate[17] of one flow, different operators gain or lose revenue; if they fail to agree, service to all flows suffers random losses. And, of course, the agreement process must be simple and cheap, given in practice many millions of flows must be handled daily.

The next two sections approach this intertwined subject from two angles. First, the service coordination section takes a technical approach introducing economic aspects in this technical context. Then, the business coordination section takes an economic approach, introducing the technical aspects in this economic context.

5 Service coordination

The service coordination problem divides into two control layers for the two granularities of service: bulk data and flows.[18,19] Each of these is further divided into two sub-layers as shown in Fig 3. We will start from our lowest control layer and work upwards. We then describe a simplified target architecture that has recently emerged from the research community.

[17] Rate control is used as the general case, because denying admission reduces the rate to zero.

[18] This is a control plane layering that reflects the layering of the data plane. So there is no implication that lower layers provide service to higher layers — it is not a service stack. Indeed, admission control and rate control are typically mutually exclusive, rarely being applied to the same flow.

[19] Note that in traditional telecoms, capacity for a flow is considered to be layered beneath data transfer, with a partition of capacity set aside for each flow. Whereas in packet networks, packet multiplexing is the base service and allocation of capacity to a flow is typically achieved by modifying packet scheduling or by controlling packet load.

5.1 Bulk path characterisation

QoS is determined by the dynamic characteristics of different paths through an internetwork. A network that has agreed to deliver packets for a sending customer with a certain QoS may have to forward through a string of downstream networks but still meet its obligation to its customer. The simplest contractual model for this is the recursive one where network N_A contracts with N_B to provide downstream service, leaving N_B to subcontract with N_D and so on (see §7 on market structure for details). Metrics for impairments to QoS like delay, congestion or loss-rate accumulate along the path. So the sending customer's impairment budget must be shared across the string of networks. At each contractual boundary, the upstream network asks the next downstream network to keep within the impairment budget that is what remains of the overall budget after it has subtracted its own share.

Operators considering deploying inter-provider IP QoS are starting to discuss deployment of echo-responders at strategic interconnect points between networks around the Internet. Then the quality of paths across different networks could be actively measured using probes, both for service management and to verify these contractual obligations with neighbours. Whether such a measurement fabric could ever be proofed against cheating is yet to be determined — whether it were independently operated or shared by those being measured.

A less costly and less complex alternative seems possible. At the same time as data is forwarded, routers on the path traversed by each flow could characterise the path by modifying the packets. Indeed they already do. For instance, the rate that routers on the path drop packets due to transient or persistent overload conditions is an implicit way for the receiver to characterise the path. Or the recently standardised explicit congestion notification (ECN) capability can be used. The more congested a router interface is, the more it randomly sets (or 'marks') the ECN field in the IP header to warn of approaching congestion.[20] By piggy-backing path characterisation on data packets, highly valuable dynamic metrics can be used to dynamically manage QoS. All impairment metrics can already be measured this way: path delay (approximately using TTL), loss rate and incipient congestion [7].

Not only can end-points use this path characterisation. By passive metering at interconnect boundaries, it should be possible to establish how much of a particular path characteristic is due to which network. However, at any interconnect point in the current Internet, it is only possible to measure the impairment that has been introduced so far along the upstream path. So two neighbouring networks can establish whether the upstream network has kept to its obligations, but not whether the downstream network has. This doesn't support the recursive contractual model. A network needs to prove to its upstream, not its downstream, neighbour that it has kept to its agreement.

In recent architectural work to fix the Internet's capacity allocation and accountability problems [7], we have proposed a trivially simple mechanism to re-align path characterisation metrics to a common reference at the destination (rather than at the source as is traditional). We call this re-feedback, short for receiver-aligned feedback. The above referenced paper also proposes a simple technique to introduce re-feedback into the Internet without needing to change the Internet protocol or IP routers.

Further, re-feedback has been carefully designed to prevent cheating. A framework of simple mechanisms has been proposed so that strategising players, whether network operators or end-users, will report path characterisation honestly, even when it is used as the basis for interconnect charging or preferential allocation of capacity to different customers.

For example, in the Internet as it stands (Fig 4a), a meter (the eyeball) at the interconnect between N_A & N_B might measure an average of 0.2% congestion in passing traffic destined for subnet R_1. This implies 0.2% incipient congestion has already been experienced upstream, but says nothing about the remaining congestion on the path. With re-feedback (Fig 4b), routers on the path increment the metric exactly as they did before. But the sender ensures that it initialises the metric to whatever value is necessary to reach a standardised number (zero in the example shown) at the destination. The sender does this based on previous experience of the path using feedback from the receiver. So, if re-feedback were used over exactly the same path as above, the meter would measure -0.3%, implying 0.3% congestion remains on the path downstream (it has to invert the sign).

To be clear, if the same meter were to look at the path at another time t_2 on the right of the figure, it might measure 0.1% upstream congestion if classic feedback were being used by the sender. But if re-feedback were being used by the sender, the meter would read -0.6% in this case, implying 0.6% congestion downstream.

It is much more useful for a network to know how much congestion traffic is going to cause, rather than how much it has already caused. Indeed, we

[20] Although ECN is standardised and implemented, its deployment is still patchy.

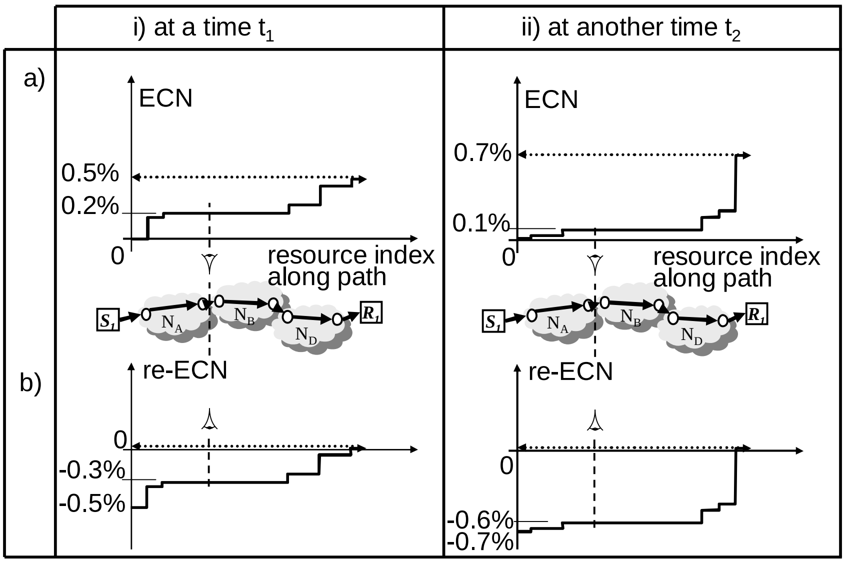

Figure 4: Alignment of explicit path congestion (ECN) metric at a) sender and b) receiver

will see later that N_B should charge N_A for the downstream congestion caused by traffic that N_A forwards into N_B.

Of course, an operator does not want to reveal any more about the internal status of its network than absolutely necessary. Its asymmetric access to information about its own network is a source of market power. Or, seen from everyone else's perspective, it is a barrier to the effective working of the market [12]. However, an operator cannot help but reveal impairments such as congestion or delay. So a system that makes them cheaper to measure should not be seen as a threat. Delay is a physical fact of transmission across a network. While, if an operator attempts to deny the existence of the early signs of congestion, it will not be able to push back against rising load, resulting in degraded service for all its customers.

Path characterisation is a major part of the Internet's mechanisms for determining how much capacity each competing flow would get in our interconnect scenario (§4). However, to control load, path characterisation must be fed back to the source. This feedback is a pre-requisite of the rate control layer (Fig 3), which we will return to later, when we discuss per-flow control. At this point, all that is necessary to note is that path characterisation in the forward direction can be achieved as a side-effect of forwarding, without regard to flows. So it scales extremely well as demand grows.

5.2 Bulk rate policing

Traditionally, the load that could be applied to a network was limited by the physical capacity of the customer's link. The load one network could inject into its neighbour was similarly constrained by their interconnect link. It is however possible to logically constrain a link to a lower capacity, for instance using a bulk token bucket regulator. Since the addition of the Diffserv field to IP in 1998, it has been possible to define different classes of service, then separately logically constrain the capacity available to each class of packets, defined by which Diffserv code point (DSCP [35]) they carry. In this way, a single set of networked IP resources can be made to appear as multiple logical networks, one for each class. And with suitable scheduling and policing policies, idle capacity of higher classes can be borrowed by lower classes, thus also preserving the advantages of packet multiplexing, rather than creating wasteful hard partitions.

With bulk rate policing no service coordination is necessary each time a flow arrives or departs. All coordination must be done in advance, making QoS highly sensitive to the accuracy of traffic predictions (or equivalently, extra investment in over-capacity is required to minimise the impact of poor predictions). Ideally a customer should be able to predict levels of traffic from each source to each destination. A network model can then determine the likely load on each link in the internetwork, taking account of average traffic from each source, what proportion of it is likely to be directed over the link in question, what variation is likely, and how correlated the variations are likely to be, to determine how often the peaks will all focus on the same link at once. In practice only average traffic is known with any degree of certainty, and allowances for peak load are made using rules of thumb to predict variations.

In our interconnect scenario, let us imagine that all three flows are entitled to use a premium 'assured forwarding' class of service (for instance, they could all be flows originating from customers who have paid to use an interconnected VPN service[21]). N_D might calculate that this VPN will consume 5% of each of its links' capacity at peak. It might then allocate say 8% of the capacity of every resource[22] in N_D to this premium class. The non-VPN traffic that is congesting the link in N_D will still be able to use it, but whenever a VPN packet requires serving, the congested interior router will give the packet priority over 8% of its capacity.

The above procedure protects the VPN traffic from lower priority traffic, but not from excessive traffic of the same class. N_D must also set up policers at its interfaces with its neighbours (N_B, N_E & S_1) to limit traffic into this VPN class to 8%[23] (shown as bulk data rate policing at every network ingress in Fig 3). All traffic above this limit would be recoloured to a lower class. In this way, if our three flows and others like them were all of a higher priority, their risk of congestion on the link shown would be much reduced, instead being largely confined to the lower classes.

However, occasionally, traffic in the VPN class might still congest the interior link's allocated VPN capacity. Even though all incoming traffic is policed so as not to exceed 8% of ingress capacity, traffic from multiple ingress points might all occasionally happen to converge on the same resource within N_D.

Similarly, all traffic from other networks might happen to focus on one of the ingress points into N_D. Then the policer would re-colour the excess, implying a randomly selected proportion of the traffic would no longer carry the VPN class of service. If this traffic encountered congestion on an interior link within N_D, its QoS would degrade [39].

Depending on the application mix, such risks might be acceptable. If they are not, below we describe how admission control can be used to protect the flows trying to use the congested link.

5.3 Flow admission control

Let us assume that the applications in our scenario have a minimum usable rate (inelastic — e.g. interactive voice or video). Then, during congestion, rather than reduce all their rates so they all become unusable, it is best to sacrifice whole flows — admission control. Ideally we would want to discard the lowest value flows, but as long as capacity is engineered so that refusal of admission is rare, the extra complexity is not worth the bother. The most straightforward approach is to simply admit flows until a new request would cause congestion.[24]

Traditionally, service coordination involves signaling a request for each new flow to the network elements along the path, to check if any are unable to accept the new request. The outcome then needs to be signaled to the applications that made the request and to all the network elements so that every element only reserves capacity for successful requests (or they might be expected to time out if a positive response isn't received). The signaling system must ensure that all networks have a consistent view of whether the request succeeded. Otherwise one network might start charging, while another declines service.

Also, in a packet network, inelastic traffic given a service guarantee must be treated separately to rate-adaptive traffic. If the two were mixed together with the same priority, admission control of inelastic flows alone would give no guarantee of service. Even if more inelastic flows were refused admission to the system, an increasing load from elastic flows could increase congestion. The simplest way to separate the two modes is to schedule packets of the inelastic class in priority over elastic traffic. So, when we say 'admission control', we mean control of admission into an inelastic class.

5.3.1 Admitted flow rate policing

Once an application's request has been accepted, it could maliciously send more data than it requested. So each network will want to check that it is only giving priority service to traffic within the requested rate envelope (rate policing). In our interconnect context, if we assume N_E pays N_D an interconnect charge proportionate to the size of each reservation, we have a more insidious problem. N_E can fraudulently make a profit by reducing the size of the reservation request before passing it to

[21] As mentioned in the Introduction, interconnected VPNs are currently rare. Each VPN is currently built by one operator over dedicated capacity, leased where necessary to acquire global coverage. But the necessary standards are starting to be put in place so that VPNs can be created by interconnecting VPNs built over the capacity of different operators. As noted earlier, the motivation is to use IP to multiplex use of the underlying capacity, which is cheaper than buying dedicated capacity for the perhaps small numbers of VPNs that any one operator serves at any specific global location.

[22] This example is simplified. In practice, different VPNs are likely to selectively use some links more than others, so the traffic matrix of each VPN would be predicted and capacity allocated proportionately to predicted demand on each link separately.

[23] As above, in practice, policed levels would be tailored to the predicted traffic matrix. Also more complex policing policies are typical that allow a certain level of burst, etc.

[24] Assuming no pre-booking is supported.

N_D but still accepting the original reservation from its customer S_2. Therefore, N_D has to rate police at its boundary with N_E. And by extension, rate policing must occur at every trust boundary.

This rate policing problem lies at the heart of the scalability problem found with Intserv in 1997. The problem applies to all similar approaches for reserving capacity over packet networks. Rate policing requires fairly complex per-packet operations to associate each packet with its flow and then test the rate of each flow over the sequence of packets. Clearly this requires state about each flow to be maintained at every trust boundary. It is hard (i.e. expensive) to build boxes for interconnect that can rate police at the very high data rates required at interconnect boundaries. Certainly, this problem will prevent us moving to all-optical interconnect any time soon, given all-optical state storage is some way off.

If bundles of flows could be policed on an aggregate basis at inter-domain boundaries, the problem would be reduced. For instance, MPLS gives the same label to all flows forwarded the same way across a domain (a forwarding equivalence class or FEC). So one would think that this aggregate could be rate policed rather than policing each internal micro-flow.

But MPLS was never designed or intended for interconnection. The number of FECs required for a network grows with the square of the number of endpoints. So if MPLS were to be spread across interconnection boundaries, the FEC aggregates used today for one domain would all have to be broken down into smaller and smaller aggregates. So at inter-domain boundaries, there would be many more, smaller aggregates to police, exacerbating, rather than solving the flow rate policing problem. Instead, the solution to this problem lies in a more subtle form of aggregation based on path characterisation, which is presented in §5.5.2.

5.3.2 Signaling coordination

Of course, for all the networks in our scenario to coordinate reservation signaling, they must all understand each other's signaling. The session initiation protocol (SIP) is becoming widely adopted, but it was never designed for signaling QoS requests to network elements. It is ideal for the end-applications to find each other (via proxies) and to coordinate their requirements with each other. But SIP is an application layer protocol, not streamlined for talking directly to routers. QoS fields are being added to SIP for requirement coordination. But a protocol designed to coordinate QoS with routers, such as RSVP, is also required. However, RSVP takes three passes because its design was generalised for both unicast and multicast, which adds unnecessary delay when setting up unicast flows. Therefore, it will be necessary for RSVP to interwork with other approaches being developed (and with legacy PSTN signaling of course).

Thus, in order to realise the huge predicted value of QoS interconnect, we need a signaling coordination mechanism that allows each domain to choose its own signaling approach. This will cope with the diversity that has already arisen in the industry and will also allow evolution of further diversity if competitively advantageous. Or consolidation can evolve just as easily, if the industry finds it can converge on sufficient solutions.

Elsewhere, we have proposed a signaling coordination solution that uses the same protocol elements as RSVP, but can be switched into different modes (sender initiated, receiver initiated, etc.) dependent on the IP protocol ID. It is designed to overlay existing signaling approaches, allowing diversity of approaches along the path. The alternative would be for each domain to have to deploy different gateways between itself and each other type of neighbouring network. For n approaches to QoS, that would require each approach to implement $n-1$ types of gateway. With the overlay approach, each QoS approach only has to interface with one common approach laid over all the others. Not only does this make interconnecting the current approaches cheaper and less prone to 'Chinese whispers' errors, but it also puts up less of a barrier to innovative approaches.

5.4 Elastic flow rate control

For applications that can tolerate a varying data rate (elastic), the question of what rate each of our three flows should be given seems difficult, given whatever rate one is given might affect other flows sharing other congested links with it in other networks. Indeed, at first glance, given flows arrive and depart with anything down to sub-second hold-times, expensive, continuous coordination between networks would seem to be required.

In fact, in 1989, after some catastrophic congestion collapses on the early Internet, an ingenious solution to this problem [24] was added to TCP/IP. It required no explicit coordination messages, solely using path characterisation carried almost for free in the data flows, as described in §5.1. Those flows passing through our congested link all experience losses (or ECN marking in the near-future) caused by that link. The TCP receiver feeds this characterisation back to the source at most every other packet. The source runs TCP's rate control algorithm, which backs off sharply when congestion is notified or slowly ramps up otherwise. New flows are allowed to push in with an exponential ramp up. The sharp back-off of existing flows quickly makes way for them. The whole system continually tries to converge to a position where all flows cause the same congestion per round trip as each other.[25]

Also, as everyone's TCP algorithms continuously do their work, networks implicitly exchange signaling (in the path characterisations carried in packet headers). In the next section (§5.5) we explain why this implicit signaling is an ideal candidate for inter-domain QoS coordination and in §6.2 we explain the role of incipient congestion as a metric for interconnect settlements.

Currently about 98% of traffic on the public Internet identifies itself as TCP[26]. The balance either does not adapt to congestion or adapts sluggishly. Adding all the traffic that BT plan to serve over IP (most significantly PSTN and private wires), it is predicted the proportion of adaptive (elastic) traffic on BT's networks might drop to 50-70% on initial deployment of the 21st Century Network. After initial deployment, this proportion is expected to rise again as demand growth for elastic data is expected to continue to outgrow inelastic.

Discussions of QoS invariably dismiss the value of this huge majority of elastic traffic. Further, the extremely low cost of TCP's elegant capacity allocation mechanism is never recognised for the extremely valuable business asset that it is. It works so well, few people know it is there. The next section introduces recent research that promises a service coordination mechanism as elegant and cheap as TCP's, but which also naturally solves our capacity allocation problem for different qualities of service, including admission controlled flows.

5.5 Simplified target architecture

Although TCP works, and appears to have kept the Internet stable, it was developed by intuition rather than science. There was no proof of what the best algorithm for allocating capacity would have been, so we could not know how close TCP's capacity allocations were to optimality. In 1997, Kelly created a model to solve the capacity allocation problem posed in our three flows scenario, but for the whole Internet at once [26]. It was an economic optimisation that globally maximised the value derived by every customer across the whole Internet, while minimising incipient congestion. Kelly also proved it would be stable. It is therefore a very important result, as any other allocation would be sub-optimal. So networks that cooperated in reaching the optimal allocation would better satisfy their customers, gaining competitive advantage over those that didn't.

The resulting mechanisms were nearly identical to TCP/IP's[27] — just the algorithms were different. The resulting optimal allocations had superficial similarity to TCP's but with significant differences.[28] Kelly's main extension beyond TCP was to allow for differences in willingness to pay (i.e. weight) in order to allow for different qualities of service. The resulting optimal Internet would continually aim to converge to a position where the incipient congestion caused by each flow was proportional to its weight.

Another way to say this is that, if each user had to pay for the incipient congestion her applications caused then, as conditions continually changed, everyone would choose software that adapted the rate for each of their flows so that allocations would remain optimal across the whole Internet. That is, total value would be maximised and total congestion would be minimised.

The model is recursive for interconnected networks [7]. Using our scenario as an example, recursion means S_1 should pay N_A for all the incipient congestion it causes on the path to R_1, N_A should pay N_B for the remaining congestion on the path downstream of N_A and so on at each interconnect point down the path.[29] Passive congestion pricing emulates active policing at inter-domain boundaries by recursively giving each network a financial incentive not to allow its upstream customers to cause congestion in downstream networks (see [7] for details). So the considerable complexity in correctly setting up bulk inter-provider traffic conditioning agreements (§5.2) and the risk of wrongly predicting demand all disappear.

Fortunately, the metering for such interconnect charging is very cheap to deploy. If congestion notification is piggy-backed on data and aligned to zero at the receiver as in Fig 4b), all that is necessary to determine the charge for incipient congestion downstream of the interconnect boundary is

[25] That is, if the round trip time is T and the packet loss rate (or ECN packet marking rate) on the path is p, every flow's TCP algorithm converges towards a bit rate x such that $xT\sqrt{p}$ is constant.

[26] This invariably implies it adapts to congestion as TCP should, the exception being a small proportion of malicious traffic.

[27] Assuming deployment of the ECN extensions to TCP/IP standardised in 2001.

[28] With Kelly's optimal algorithm, every flow would converge towards a bit rate x such that $xp = w$, where w is no longer a constant, but the user's willingness to pay (i.e. weight) for congestion per unit time for each flow. Note that, unlike TCP, there is no direct dependence on round trip time, T, although there is indirectly, because congestion p itself depends partially on T.

[29] Note that our choice of the 'sender-pays' model is deliberate. See the direction of payment discussion in §7.2.

to passively count the congestion metrics passing the interconnect point — in bulk without regard to flows.[30] Thus, we have an extremely low cost mechanism for coordinating quality of service for the whole Internet that ensures optimal use of network investment.

5.5.1 Congestion pricing underlay

However, in surveys across a wide range of economic sectors [37] there is considerable evidence that people are highly averse to unpredictable charges. So, since publication of Kelly's seminal work, our research has focused on how to exploit this new understanding of the fundamental role of congestion in the economics of networks, but without forcing end-customers to accept congestion pricing.

We have developed ways to use congestion charging just for bulk wholesale and interconnect pricing, but then allow much more flexibility in the design of retail tariffs and retail services layered on top (§6 gives examples of various possible retail service plans). The idea is to replace congestion charging at the first link in the above recursive chain of charges between S_1 & N_A. Instead we want to allow a rich variety of more human-friendly tariff models such as volume caps, volume charging, per-session charging, flat pricing, etc. The choice would depend on the particular sector of the retail market. But congestion charging is a sufficient wholesale and interconnect tariff layered beneath the retail market.

Note that the terms retail and wholesale are used generically. We are not implying any recommended separation of business between actual retailers and wholesalers, such as BT Retail and BT Wholesale. Indeed congestion pricing might only ever be used as an internal price within a networking wholesaler to manage internal traffic policers and provisioning, whilst selling service based on a more traditional tariff plan, but treating congestion pricing as a lower bound on the price in use.

To be clear, the underlying congestion pricing metrics would stay the same as we described in the recursive model above, but only for the wholesale and interconnect parts. N_A paying N_B paying N_D and so on. But S_1 would not be expected to pay the congestion price to N_A. So at network edges the congestion 'price' would only be used internally, in order to correctly set the chosen retail parameters layered above (the volume price, the volume cap or whatever). For example, S_1 might be on a volume charged tariff with two time-of-day prices. The retailer would determine these time of day prices by averaging the internal congestion price it had to pay to the wholesaler. Alternatively, the congestion price might only be metered internally by the wholesaler before transforming into a more stable price to the retailer.

Note that the different retail models can pick and choose between using pricing and throttling to manage congestion and therefore QoS. Dynamic throttling is an exact complement to dynamic pricing[31], and the evidence shows that dynamic throttling is more acceptable to customers for many applications. We use the term throttling to include policing traffic to ensure the customer's computers are throttling themselves correctly [7]. The term throttling also includes caps or quotas.

To summarise so far, recent research has delivered the potential for the industry to adopt an elegant, extremely low cost mechanism that could unobtrusively coordinate inter-domain capacity allocation for all qualities of service, not just the single quality level of TCP. It would also significantly simplify and improve inter-domain traffic policing. It involves two inversions to traditional QoS thinking:

- Instead of network equipment providing different qualities of service (i.e. enhanced priority during congestion), the sending customer device is allowed a laxer response to incipient congestion. Those customers that are allowed a lax response (or zero response) to congestion, get more of the capacity of the congested resources, which is equivalent to the network equipment giving priority access, but with none of the trust-boundary complexity of the traditional alternative[32].
- Rather than a packet picking up congestion characterisation as it traverses the network, the source pre-loads each packet with a sufficiently negative level of congestion charac-

[30] Having fixed a price for incipient congestion, over an accounting period (e.g. a month) the meter would simply need to count the bulk volume of traffic marked with the ECT(0) and CE code-points in the ECN field of the IP header and subtract the two. At the end of the month, multiplying the result by the price of congestion advertised earlier would determine the charge a network should pay its downstream neighbour.

[31] Congestion pricing from the underlying wholesale market could be used to control throttling at the retail level.

[32] This approach has the important advantage that high speed network elements (e.g. at interconnect boundaries) do not have to verify the identity of traffic to verify its authorisation to use QoS. Network elements merely piggyback their congestion status on packets, so it is impossible to masquerade as someone else's identity for QoS purposes, as there *is* no separate identity for QoS purposes. Our paper on policing congestion response [7] explains how the approach moves this hard security problem to the very first ingress edge of the network, where customer identity can be verified much more scalably.

terisation appropriate to the path (using feedback). Then as congestion characterisation is picked up as the packet traverses the path, the metric tends to zero on reaching the destination (Fig 4b)).

The first inversion of thinking can emulate any of the diverse QoS mechanisms given in §3. The second aligns metrics correctly at interconnect points. It allows network operators to prevent customers abusing their freedom to respond to congestion however they wish [7].[33] Each downstream network can incentivise the next upstream network and so on to eventually prevent the ultimate sending customers from abusing the freedom to use a lax response to congestion. We have therefore managed to remove both bulk and per-flow rate policing from interconnect trust boundaries and still supply the same services.

5.5.2 Simplified & scalable guaranteed QoS

We have explained that ideally we want to superpose guaranteed sessions over a packet network in order to exploit gains from packet multiplexing. But this seems to lead to scalability problems, particularly at interconnect boundaries, as explained in §5.3.1 on rate policing for admitted flows. We have claimed that it is possible to build a variety of retail service plans on top of our simplified target architecture, using congestion characterisation piggy-backed on each packet as a generic QoS coordination mechanism. We now back up that claim, by showing how to provide guaranteed reservations over this architecture, at the same time solving the scalability problems of previous approaches. Full details are given in Hovell *et al* [22], but we provide a brief outline here in order to be able to bring out the interconnection aspects.

Rather than each flow request being put to every network element on a path (or at least one in every domain), the technique determines whether the path across a very large hop is congested, all in one go. The hop can encompass many domains, so as more neighbouring operators adopt the approach, the system becomes increasingly simple and scalable. Such a large hop must be surrounded by a ring of gateways capable of reservation signaling, but signaling is ignored by all elements within the ring. From outside the ring the gateways effectively appear to have synthesised guaranteed QoS even though they only use a non-reserved forwarding service within the ring, hence the name: guaranteed QoS synthesis (GQS).

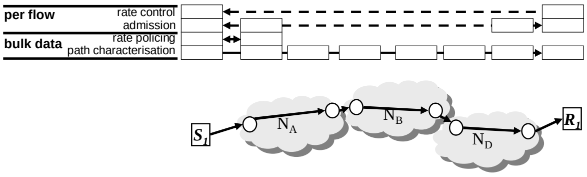

Figure 5: QoS layering simplified by generalising the use of path characterisation

Each egress GQS gateway monitors ECN in passing data to characterise the aggregates from each active ingress gateway across the ring. When a new reservation request arrives, it is only accepted if incipient congestion, on the path that it will take across the ring, is below a defined threshold.

To determine what share of our congested link should be assigned to each of our three flows, in §5.1 we explained how the Internet piggy-backs characterisation of each path's congestion on the data. GQS gateways use identical underlying path characterisation. But instead of varying the rate of each flow continuously, the gateways ensure admitted flows get their full requested rate by blocking non-admitted flows completely — admission control.

If one compares the resulting QoS layering of Fig 5 with the original without GQS in Fig 3, one can see session admission control functions are no longer necessary anywhere but at the network edge. Routers within the ring (including at interconnect boundaries) are not configured to recognise QoS signaling, which passes across the ring transparently as if it were ordinary data.

At interconnect boundaries within the ring, there is no flow or session awareness. There is only bulk data transfer (carrying path characterisation). When we introduced this simplified target architecture (§5.5) we explained how recursive congestion pricing between networks is very cheap to implement and precisely emulates active rate policing. It creates the correct back pressure on upstream networks so that they have an incentive to block only the flows that would cause congestion, even though nothing within the ring is aware of flows. So the GQS solves the scalability problem of rate policing at inter-domain boundaries, as promised in §5.3.1 on admitted flow rate policing.

Note that a different product (sessions) is sold to end-customers by retailers than is traded at interconnect boundaries (incipient congestion). The commercial implications of this technical development on the interconnect market are explored in §6.2 on value-based vs. cost-based charging.

[33] Internet users have always had this freedom, but it has never been promoted as a QoS mechanism because there was no way to police it.

One class congesting another. GQS gateways control admission to a prioritised class of service, set aside only for admission controlled traffic. However, we can make the whole capacity of each resource available to either class. The partition between the two can be determined by instantaneous relative demand for each type of traffic at each resource, rather than error-prone static provisioning, which requires the traffic matrix to be estimated in advance. Nevertheless, once a flow is admitted into the higher priority class its guaranteed service is preserved; it cannot be affected by increasing demand for lower priority traffic.

The trick is in the bulk congestion marking algorithm on interior routers. Instead of marking traffic based solely on congestion of its own class, it should be marked based on the incipient congestion it causes to all classes [19].

Whenever any incipient congestion is present, higher priority traffic will therefore carry a higher rate of congestion marking than other traffic (given it causes more congestion of lower priority traffic). The price per congestion mark should be the *same* in each class. But higher priority traffic will generally still cost more because it gets marked more often.

6 Business coordination

We have shown how QoS is all about managing the risk of congestion, and that predominantly this risk arises in access networks. QoS would also be beneficial in core and backbone networks to provide end-to-end consistency of service, but scalability and cost-efficiency at interconnect boundaries is critical. We have predicted that a tiered solution will be adopted that allows for a diversity of charging schemes at the edge; including per-session, per-flow and bulk charging. We have shown that such a tiered QoS solution is feasible based on bulk congestion marking to coordinate end-to-end service.

We have shown that it would be trivially simple to implement passive bulk congestion charging between networks. This gives each network the incentive to ensure its customers behave correctly, thus avoiding the need for active per-flow rate policing, which is the critical scalability issue at interconnect boundaries.

In this section, we consider the desirability of congestion charging from economic and commercial points of view. We have implied that congestion pricing alone would be a sufficient generic lowest tier of the solution, over which higher level charging for session QoS can be added around the edges. So in §6.1 we examine why this lowest layer can be sensitive to everyone's value preferences, even where human users are mixed with extremely efficient, strategising machines maximising the value they extract from the internetwork on behalf of their users.

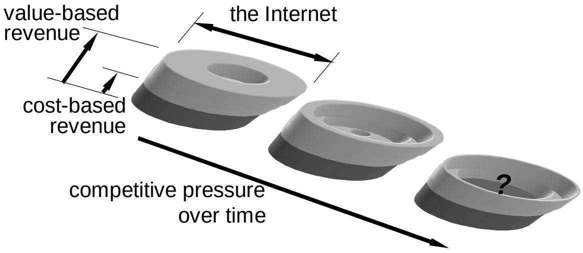

Figure 6: Value-based charges tiered over cost; and erosion of margins growing from the middle of the Internet outwards.

As competing providers undercut each other their prices fall, but no further than the marginal cost of supplying capacity by the cheapest possible means. Over time, the theory is that providers will upgrade the capacity of each resource (link or router) to balance demand for each resource — mediated by congestion pricing for each resource. So not only can congestion pricing keep hoards of extremely hostile customers under control, it is also the end result of fierce (perfect) competition.

In reality, competition for each route will be far from perfect. So above the base cost of the lower tier exposed by congestion pricing, there is space in the higher tier for service providers to add value-based charges. In §6.2 we explain why a vibrant market in this upper space should be able to extract surplus value, particularly around the access to the internetwork, fortuitously just where investment costs are greatest (Fig 6). But we also explain how the process of competition will erode these margins, albeit from the middle of the network outwards. Note that we are imagining these competition-driven changes to gradually bite over the next decade or so, although they are of course in motion as we speak.

Even though offered prices will usually exceed the congestion price, to coordinate pricing (or service constraints) it will still be necessary to expose the congestion price, as a time-varying lower bound for each route through each network. So congestion pricing is not just some far distant theoretical scenario for a highly competitive world that we may never reach. It is a lower bound we should take account of in pricing networks today, in order to manage such problems as peer-to-peer file-sharing traffic.

At the end of this section (§6.3) we discuss the merits of the various interconnect tariffs in use or under consideration today, particularly because of their ultimate influence on retail tariffs. Invariably these tariffs are designed to extract value (the upper tier) without taking full account of how underlying costs vary. So whenever customer behaviour causes costs to exceed value, the tariff doesn't push back sufficiently.

In §7 we explain the market structures necessary for a mix of tariffs to co-exist so that natural selection through competition can cause piecemeal growth of more robust tariffs, which is how congestion pricing will take hold and spread.

Investment in new network infrastructure, and hence growth of higher level services, stagnates if the risk that it will not produce returns is too high. Having provided mechanisms to push back against demand whenever cost exceeds value, we have greatly reduced this risk (at the same time simplifying the technology). By improving understanding of the competitive process, our aim is to further reduce the risk associated with fear of the unknown.

6.1 Congestion pricing

A data sender can cause congestion in its own access network or the access network serving its destination. Congestion is a classic example of a negative economic externality, that is, a detrimental side-effect on others. So, in economic terms, QoS coordination is all about ensuring a sender internalises (i.e. directly experiences) the cost of the externalities it causes others to suffer.[29] The problem is complicated by having to pass on the cost of this externality, not just between the two access networks but across the intervening networks, which are deliberately designed to rarely experience congestion themselves. The bulk path characterisation mechanisms, in the simplified target architecture just described (§5.5), carry the cost of the externality to its source with absolute precision, and have already been implemented very cheaply in today's routers (but not deployed).

But the sender need not pay money to internalise the varying cost of its actions. For the majority of customers who are averse to a constantly changing price, it can be flattened out and the peaks replaced by a constraint of equivalent value. As long as the lost value caused by the constraint is equivalent to the lost value others suffer, the externality can be considered to have been fully settled. But still, those customers willing to pay a higher flat charge, suffer the constraints much less often. Examples of constraints are admission control and rate policing.

For instance, the retail customer might pay predictable per-session charges for reserving session QoS based on bandwidth and duration similar to traditional telephony charging. Internally, the session retailer can make its admission control decision by comparing the revenue it will receive for the session with the likely congestion charge from the underlying network (§5.5.2). Another example is where the customer pays a flat charge for Internet access, which is treated as a credit limit or quota. An internal account of congestion charges is maintained against each customer. The closer the account approaches the quota, the more traffic destined for congested paths is throttled, but traffic into uncongested paths proceeds unhindered [7].

Importantly, the price signals for QoS throughout the internetwork can be the same, whether the customer at the edge is exchanging them for constraints or actually paying the variable price. That is, the most universally useful metric for QoS co-ordination is money, being a well-understood unit of exchange anywhere in the world[34].

Once QoS co-ordination is viewed as a flow of money, one can see the dual nature of microeconomics at work. Every flow of money has a source and a sink, with a chain of demand and supply interfaces along its length each mediated by price, as it traverses the value chain down through network layers and across interconnected networks. So the role of interconnect agreements is to carry money from where demand enters the network industry to the operator(s) that supply the bottleneck resources needed to satisfy the demand.

Demand side. Money flows from end-customers into the network against the back-pressure of the price of QoS. Sufficient demand will lead to sufficient willingness to pay to overcome the back pressure, thus ranking demand and rationing service to those willing to pay most.[35] This service rationing at the edge avoids congestion on the network elements in most demand, thus preserving QoS for those willing to pay. And where necessary a constraint of equivalent value can serve in place of money at the edges.

Supply (provisioning) side. Money flows to those parts of the network that are most under-

[34] The International Accounting Rate System (IARS) [23] used in public telephony is already based on a virtual currency dependent on a basket of major global currencies.

[35] Some regions of an internetwork also have to rank demand based on policies other than willingness to pay (e.g. a government policy of universal telephony service, a corporate policy of dual vendor supply, or a military policy of pre-emption priority for senior officers). It is best to convert such policies into their equivalent financial value, which all the other networks in the world can understand, rather than expecting every network to be sensitive to every regional policy.

provisioned, driving investment in network resources where they are most needed. Network resources that are highly used but already oversupplied don't attract any revenue flow (we would expect congestion charges to be complemented by capacity charges in these cases). So, a sufficient business coordination system allows free flow of money to share out revenues across the whole internetwork, dropping them only where investment is most required. This is what congestion pricing achieves [26].

6.2 Value or cost?

In §2.2 we described how the full value of potential connectivity to others with QoS capabilities can be released by interconnection. Of course, operators can charge each other and end-customers for access to this value, which is discussed in a separate paper [6]. In this section, rather than the value of potential usage of capacity, we focus on the value of actual usage — and its relation to cost.

Each session should only take place if its value exceeds its marginal cost[36]. But if one network tries to capture most of the surplus value, another competing network can undercut it and still cover its costs. So, in a competitive (or well-regulated) retail market, usage charges will tend towards cost.

So the key questions that determine whether a network can sustain a healthy surplus from value-based charging are:

1. How feasible is it for networks to infer customer value anyway?
2. How strong is competition likely to be?

By examining each question in turn below, we show that access networks are in a much stronger position to sustain value-based charging than backbones.

We end by explaining what it means to talk about the cost of transmitting information anyway. Understanding the cost of each session is the primary contribution of this paper. The price of any session shouldn't fall below its cost, so that only those sessions where value exceeds cost will proceed. Further, it is also important to understand how to determine cost, given it is the limit that competition can drive a network to.

6.2.1 Inferring value

From the customer's point of view the aim of the game is to reveal as little about their value as possible. For instance, many people pay 10p for delivery of a 100B SMS message. But no-one would pay the equivalent (£1000) for delivery of a 1MB audio track, let alone £1 million for a 1GB video! But if a network offered transfer of 1GB of data for just £1, customers would disguise SMS messages as general data — if they could.

If a network delves into packets to determine customer intent (deep packet inspection or DPI), it risks falling foul of various national regulations (anti-competitive behaviour, anti-trust or common carrier to name a few). And as routine encryption becomes more common (for VPNs, e-commerce etc) the applicability of this technique for price discrimination will diminish. However, the majority of customers are not expert enough (or bothered enough) to thwart such price discrimination by encrypting their traffic deliberately. So price discrimination will still be possible (though not necessarily ethical) against more naïve customers.

A more defensible strategy (both ethically and competitively) is to at least bundle some difference in quality of service for the difference in price, no matter how trivial the extra cost to the operator. By requiring the customer to ask for QoS on a per-session basis, the edge network is better placed to infer intent and price by value.

But a backbone network does not have the luxury of naïve customers. If it tried to delve into packets and charge by value, edge networks would simply pass the traffic between themselves through encrypted tunnels. In any case, DPI at the high speeds typical in backbone networks would be prohibitively expensive. If it tried to over-charge for insurance against the very small risk of congestion in its backbone, edge networks would encrypt their QoS signaling or simply route through another backbone.

6.2.2 Competition

To be competitive, an operator doesn't necessarily have to offer a competitive price for every path. Customers buy access to a basket of routes when they choose a provider. But competition could cause subscription (capacity) charges to reduce and usage charges to increase, making it more economic for a customer to subscribe to multiple providers (termed multi-homing). Then, fast route selection based on price [21] could become more common.

Few end-customers currently multi-home — the exceptions being larger businesses that place a pre-

[36] For brevity we will take the term 'cost' to include a 'normal' profit margin sufficient to reward the risk of the initial investments.

mium on never losing connectivity. However, network providers generally interconnect with many other networks. So route selection is highly competitive in the middle of an internetwork, but less so around the edges. As regulatory measures, such as local loop unbundling, intensify competition between access networks, we might expect to see fast, price-based selection of routes spreading from the middle of the internetwork to the edges.

6.2.3 Cost

The costs of network resources are sunk. Whether they are used or not, they cost the same. So transmitting information would appear to add nothing to the cost.[37] But for any congestible resource, as total usage approaches capacity, each customer's usage causes the others to experience a 'social' cost; where everyone affected by congestion is also to blame for it. To manage congestion, someone should charge each user i in proportion to how much they are to blame for approaching congestion. So they should be charged both in proportion to their own rate of usage x_i and in proportion to the probability p that the resource is going to become congested. So each user should at least be charged $c_i = \Lambda p x_i$, where Λ is the agreed price of incipient congestion, which should remain relatively constant over time. The trick is to charge users just before congestion is experienced in order to avoid anyone experiencing degradation in service. An extremely cheap mechanism to do this in bulk, without regard to each user's flows, is already available in every vendor's routers.

We said 'someone' should levy the congestion charge. The owner of each resource doesn't actually experience any direct cost. But if resource owners do levy the charge themselves, they can offset it against the cost of upgrading capacity. In fact, the relative levels of congestion revenue from each resource indicate which resources most require upgrade. In theory, congestion charges should fully cover the marginal cost of capacity.

But provisioning additional network resources often involves long lead times. So, if it is possible to configure a resource to appear to have a lower capacity than it actually has, it can be made to generate congestion marking sufficiently early to upgrade it in time. Alternatively, traditional fixed capacity charges can complement congestion charges in order to cover these fixed upgrade costs.

6.2.4 Erosion of margins

To summarise, access network retailers can expect to be able to raise 'excess' profits from value-based charging by selling QoS (both connectivity and usage) to end-customers, but backbone operators whose only customers are other networks should not expect to take any more than 'normal' profits over cost in the longer term — edge retailers will tend to keep the value-based profit to themselves. So, as the market matures we expect value-based per-session QoS to be confined to edge networks (Fig 6), while a hole in the value-based market grows outwards from the middle backbones of the Internet.

Access wholesalers will be in a half-way position — they will probably be able to share in these excess profits, for instance where they provide retailers with the facilities to support value-based charging by identifying customer intent. So, access networking should be sufficiently profitable to be able to risk the large investments required for access infrastructure expansion.

Congestion marking will serve to coordinate the edges across the 'hole' by exposing the underlying cost for each network. By ensuring the price never drops below that required for congestion pricing, demand can be managed correctly and the costs of necessary capacity expansion will be covered. Any excess above this cost will be the icing on the cake necessary to de-risk infrastructure expansion.

Also, we can now confirm that disenfranchising the interconnect market from per-flow QoS is not only technically feasible as outlined in §5.5.2 (Fig 5) but also economically inevitable.[38]

6.3 Interconnection tariffs

6.3.1 State of the art

As we said earlier, the IP QoS market is currently balkanised (i.e. intra-provider only) with little experience of which inter-provider tariffs might work and which won't. The purpose of this section is to survey tariffs[39] being proposed or in use. Interconnect charges represent a major cost to access networks, so the choice of interconnect tariffs tends to

[37] Of course, operational costs in planning, fault handling and managing a network will remain, but the simplicity of congestion notification as a management tool should also help to reduce even these.

[38] This is not surprising, given the original design of the GQS was motivated by an understanding of the economics.

[39] A tariff is a formula used to derive a charge. It is some function of metrics and prices, usually a simple addition of the products of prices and metrics. For instance the formula for charge $C = aV + bt + c$ is a three-part tariff where a is the price per data volume V, b is the price per time t and c is a one-off constant charge. Formally, the price of a metric is defined as the partial derivative of the charge with respect to that metric. For instance the price of data volume, $\frac{\partial C}{\partial V} = a$.

strongly influence the structure of charges passed onwards to end-customers. We will comment on whether various proposals under discussion create perverse incentives, or whether they are robust to strategising.

Peering Tariffs used on the best efforts Internet started with no-fee peering, where neighbours consider they derive approximately equal value from interconnecting with each other and each causes the other similar costs.

Connectivity-capacity tariffs As the Internet has commercialised, more highly connected networks have started to charge smaller networks for the privilege of access to their richer connectivity, usually with a monthly charge priced relative to the interconnect link capacity, but also dependent on relative connectivity. The connectivity element is value-based, while the capacity element is an attempt to cover costs.

The value-based tariff element of the tariff sufficed while the market was booming — weak competition during growth allowed operators to extract a significant portion of customer value, rather than being driven to cost. But at the turn of the millennium peer-to-peer file-sharing rose dramatically in popularity at about the same time as the technology investment market crashed. The risk of investing in even more capacity wasn't justified by sufficient customer value behind the traffic. Many operators reacted by moving to two-part interconnect tariffs, adding a usage element, but still deriving a proportion[40] of revenues from capacity pricing. These pricing changes at the interconnect level caused the beginning of attempts to control IP QoS on the retail market.

Usage-charging Different agreements favour different ways of determining the usage element of the charge. The two most commonly used are:

- Volume charging, which simply involves charging by bytes transferred (usually in both directions with the larger network charging the smaller) over the accounting period;
- The 95th percentile peak demand method is occasionally used, because it is supported by the dominant router vendor. Traffic volume in each direction is counted over every 15 minute period throughout each day and the set of readings is ranked. The day's charge is then based solely on the 95th percentile reading;

A good example of the current state of the art in interconnect charging is the London Internet Exchange (LINX), which charges a three-part tariff for i) the maximum capacity of the port purchased; ii) the volume of traffic transmitted in either direction; and iii) a port congestion charge[41] [29]. A recent interim report from the European CoCombine project gives a full survey of published peering agreements around Europe [20].

Volume tariffs Volume charging is an exceedingly blunt instrument. It can best be thought of as a first stab at demand management in an immature market. Controlling congestion costs (and hence capacity investment costs) seems to be its motivation, rather than value extraction, but it is insensitive to the time and the place where congestion occurs. It does dampen demand from low value traffic, leading to capacity investment being used more by higher value traffic.

An improved variant of volume charging is time of day volume pricing. Rather than the price unpredictably rising and falling with congestion, two or three discrete steps are advertised in advance, set based on typical experience. Fulp *et al* [18] provide an analysis of the trade-offs between using lots of small steps or a few big steps.

Peak demand tariffs The 95th percentile peak demand method is intended to incentivise a network to incentivise its customers to smooth out daily peaks in demand in the hope that aggregate demand from all networks will smooth as a result, leading to improved daily utilisation.

Both peak-demand and volume charging are link-based rather than network-based. That is, they take no account of whether traffic spreads out evenly across a network or is concentrated down certain routes. They simply look at the traffic on the interconnect link, not where it came from, or where it is going, or the status of the path it traverses. So neither can push back against congestion patterns that arise from demand that happens to be spread differently from 'normal' experience.

Session-based tariffs. The tariffs that are most well-understood by humans are session-based — the natural unit in which humans conceptualise communications, rather than the individual flows of data packets that the network understands. We envisage all sorts of tariff models for media sessions will be used in the retail market.

[40] Currently half to two-thirds is typical.

[41] A stringent penalty that doubles the capacity charge if per-port volume is more than 80% of the possible volume that could have been transferred in a month.

We have already explained why we believe session-based charging will evolve to an edge-to-edge overlay model, without session-based charging at every intervening interconnect. We also explained earlier why charging for setting up session QoS will be a more defensible strategy than merely applying charges to session signaling, whether or not QoS is required.

Session-based tariffs have the advantage that they can be based on session value, but set against that is the extra cost of bill itemisation and consequent per-item queries and auditability. The ETSI open settlement protocol (OSP) [17] is well-established for authorising and accounting for inter-provider VoIP call charging (see the white papers of Transnexus [41]).

VPN interconnect tariffs Currently IP VPNs are invariably built over a single operator's capacity, leased if necessary from other operators at the logical link layer. So the only interconnect is at the link layer, which is no different from today's traditional leased circuit market and outside the scope of a paper on IP QoS interconnect. Some operators are preparing the ground for interconnection of VPNs at the IP or MPLS level using Diffserv QoS (see earlier). Tariffs in such scenarios would be no different from those within existing Diffserv service level agreements, essentially charging a capacity-based premium for priority whilst constraining the customer to a traffic conditioning agreement.

Comparison against congestion pricing The above tariffs can be thought of as human-friendly approximations to the ideal congestion price. The congestion price varies packet by packet. It varies over time as other demand comes and goes. It varies over space, depending on how congested different paths are through the network. And it varies by class, depending on how much congestion higher priority traffic causes to lower priority traffic.

The problem with trying to be friendly to humans is that they share networks with computers — computers that can be programmed to extract maximum value from the network on a much more fine-grained basis than humans. Congestion pricing correctly incentivises even computers with absolute precision in time, space and class — at the minimum granularity possible: the packet. So the closer a service plan tracks the underlying congestion price, the less likely it can be abused. This is what makes congestion pricing ideal as an interconnect tariff, where there is no need to be friendly to humans. Interconnect tariffs only need to deal with the worst case customers: computers.

Retail tariffs that address human friendliness such as those above can then be layered over congestion pricing at the retail edges of the network. Alternatively, service constraints that cost the customer an equivalent amount to the congestion price can be applied (as per the examples in the Demand Side discussion in §6.1).

Operators tend to resist new usage-based tariffs because they make it hard to predict revenues accurately. In this respect the introduction of congestion charging would be no different to that of volume or peak demand charging. The uncertainty could be softened by introducing congestion charging at a low price, in conjunction with a predictable metric (such as capacity). Then its relative contribution could be gradually increased in subsequent accounting periods, just as was done with volume and peak-demand tariffs when they were new. But throughout this evolution, underlying congestion marking could be used to coordinate QoS technically.

6.3.2 Revenue sharing

Revenue enters an internetwork at the edges. Interconnect tariffs determine how much of it is moved to the networks in the middle. We have argued that operators will try to charge for QoS by value, but that charges should always exceed a cost-based lower bound. So operators will need to at least share revenues based on cost. And those in the middle will want a share in the value.

The example below shows it is relatively easy to revenue share based on costs (congestion). But having covered everyone's costs, we think it unlikely that there will ever be a systematic way to apportion the value-based surplus among all the networks on each path. Interior networks are likely to charge whatever excess over cost their market power allows them to demand, rather than depending on an institutionalised system such as the IARS used in telephony.

Cost apportionment example. Congestion accumulate along the path. Let us imagine that charges are based on congestion for the path across the internetwork shown in Fig 4b) at time t_1. For the moment, let us assume we have a global congestion price of £0.10/MB. Path congestion is 0.5%, so for 1Gbyte of video, say, the sender will be charged for $0.5\% \times 1\text{GB} = 5\text{MB} \quad @10p/\text{MB} = £0.50$. At the boundary between networks N_A and N_B they measure 0.3% congestion, so for each Gbyte N_A pays N_B $0.3\% \times 1\text{GB} = 3\text{MB} \quad @10p/\text{MB} = £0.30$. Given there is no congestion in N_B, congestion measured at the boundary between N_B and N_D

is still 0.3%. So N_B pays N_D 30p as well. The remainder of the congestion arises in N_D so it pays the receiver nothing. So in this case, only 50p enters the internetwork and is shared in proportions $20p : 0 : 30p$.[42,43]

Value apportionment example. Imagine the sender S_1 in Fig 2 is a video server and that N_A charges the video server £1 — the value it believes the server places on video QoS. N_B might charge N_A at a price of £0.55. And N_D might charge N_B at a price of £0.40. So the £1 paid by the sender is shared $45p : 15p : 40p$ between $N_A : N_B : N_D$. None of the networks believe that the value changes along the path. They just have to fiddle the prices until the apportionment is correct.

Clearly it would become extremely complicated to set all these fiddle-factors throughout the Internet to correctly share revenues between all the operators on each path, given one might find five operators along typical paths, with nine not uncommon [31][44]. Further, given IP QoS is intended to support multiple applications, different fiddle-factors would have to be set for each.

Thus to share revenues by cost, the metric itself reveals how costs have accumulated, so inter-provider prices can be roughly the same along the path. Whereas to share revenues by value, which doesn't change along the path, interconnect prices have to be a confusion of a market price and a fiddle-factor to get the revenues to share out as desired.

We showed earlier (§6.2.1) that edge networks are better placed to infer customer value and hide it from the middle. We have also shown that any systematic attempt at value-based apportionment would be extremely complicated. So we believe each edge network will tend to 'bill and keep', or at most share value only with the remote edge network involved in the session.

7 Commercial model: diversity and uniformity

We have already proposed a simplifying architecture (§5.5) that overlaid value-based session QoS above cost-based bulk QoS. Our task in the present section is to build on this, to develop an industry structure capable of fostering competitive innovation for the value-based sector around the edges, whilst ensuring interoperability and efficiency within and across the middle.

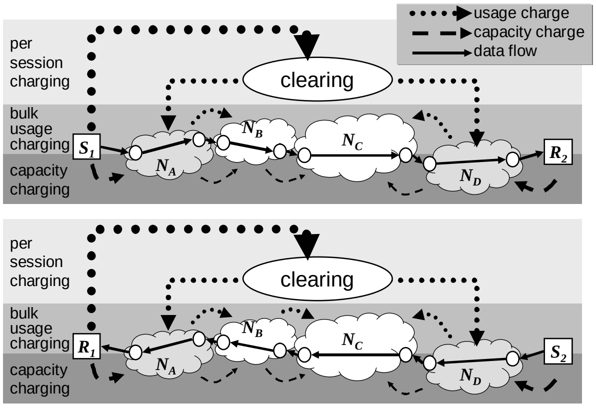

Figure 7: End-to-edge clearing for value-based charging. Two duplex flows for the same session are shown, one each in the upper and lower halves of the figure.[46]

7.1 End-to-edge clearing model

The missing piece is support for the session-based overlay that can share value-based revenues across edge networks, whether or not the middle is involved. A session will be initiated by whoever derives value from it, whether the data sender, receiver, both or even some third party (not shown). We therefore introduce a clearing role that can receive payment for a session and forge customer relationships with arbitrary service and network providers in order to distribute appropriate proportions of the revenue (assuming appropriate accounting records).

Based on [5], Fig 7 shows this clearing function in action for a session consisting of two duplex flows. Three layers of charging are shown: i) capacity charging as a foundation that continues irrespective of usage; ii) bulk usage charging at the interconnect interfaces and iii) per-session charging at the end-customer interfaces and between the clearing function and the edge networks. Of course, at the

[42] If a network is tempted to fake a higher level of congestion to attract extra revenue, its upstream networks will find a cheaper route around it (see [7]).

[43] It is of no concern that N_B receives no income from congestion charging. If its operators take a conscious decision to dimension it generously, they can set capacity charges to whatever level is required to top up expected congestion charges.

[44] These statistics are from a 1998 study. We suspect the numbers may have fallen a little since due to consolidation, but we have not found more recent statistics.

[46] Although the two halves of the figure look almost identical (except for one usage charge) the quantitive levels of session and usage charges may differ for each flow direction, dependent on relative flow rates and prices at each service interface (for instance N_C charges N_B in one direction but the price is zero in the other).

[48] Capacity charging is identical between the two halves, but session and bulk usage charging differ.

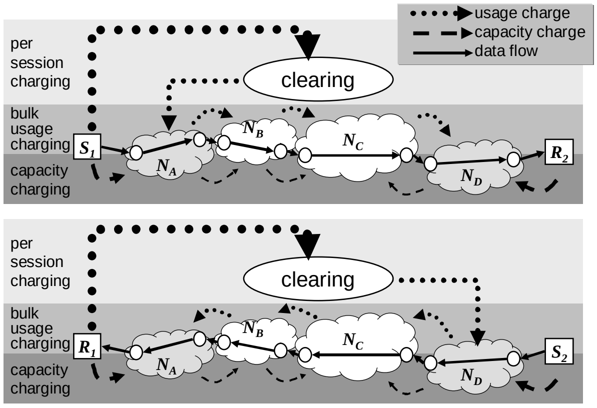

Figure 8: End-to-edge clearing for sender-pays charging. Two duplex flows for the same session are shown, one each in the upper and lower halves of the figure.[48]

capacity and bulk usage levels there are no transactions specific to the flows. They just contribute to an overall increase in usage charging (and theoretically to an overall increase in the need for capacity). In this case, the middle networks take a share of some of the surplus value by usage charging and capacity charging their smaller customer networks (nearer the edge) irrespective of the direction of transmission.

The clearing function provides the overlay needed to allow the two edge networks to coordinate apportionment of the value of QoS, over the heads of the intervening backbones. Note that the clearing function (not either edge network) interfaces directly with the end-customers.[49]

We have deliberately called it a function, because it is not necessarily a separate business. It could be combined with a traditional networking business (or not) — typically both the edge networks in the figure might offer this service, and the paying end-customer would use the service of their local edge network. The above paper envisaged multiple competitive clearing functions, and proposed a simple Web-based or DNS-based directory service that corresponding end-applications could use to lookup an appropriate clearing function that traded with both their edge networks. Essentially the function is very similar to a SIP-proxy and likely to be associated with one.

Fig 8 shows how the same function could be reused to apportion revenues where bulk usage charges track the direction of the data, following the sender-pays model. The congestion charging model that precisely recovers costs would work this way. Again, a scenario with a session consisting of two duplex flows is used. So the two parts of the per-session charge must be cleared across to the respective sending edge networks.[50]

The clearing function enables a separation between the end-edge-edge-end overlay market in QoS sessions and the market in bulk QoS that religiously follows the same hop-by-hop path as the data. That is, it enables the market separation between the per-flow layer and the bulk data layer shown in Fig 5.

7.2 Interior revenue apportionment

Although interior networks might be disenfranchised from session-based charging they can continue to share in value-based charging on a bulk basis (e.g. volume charging as now). Each network can agree an interconnect tariff with each of its neighbours independently. So in Fig 7 network N_B has agreed a two-part tariff with N_A where it receives both a capacity and a usage (e.g. volume) charge from N_A for data irrespective of direction (however, the agreed prices for each direction need not be the same). From this income it must subtract the capacity charge it pays N_C, and the usage charge (in one direction but not the other in this case — for some reason they have agreed a zero price in one direction). The task of each network is solely to ensure it agrees prices with each of its neighbours that will result in a profit overall. The same paper [5] gives the formula that each network would use to calculate its profit, dependent on all its agreed interconnect prices.

So, whether bulk usage charges were flowing towards the middle of the network (Fig 7) or with the direction of transmission (Fig 8), each network's task would be the same: to set prices with its neighbours to ensure that over all the expected incoming and outgoing flows of money (both fixed and variable), it makes a profit. We should clarify that the whole industry doesn't have to choose between the two models in the figures. The proposed model encompasses both figures and other permutations, because each network can apply bulk usage pricing to data flowing in either direction in order to make a profit. Therefore, even if a model is prevalent where money flows towards the middle, a sender-pays model like congestion pricing can be introduced piecemeal, by the independent choices of networks anywhere in the system.

[49] An alternative model can be envisaged where an end-customer always pays its local edge ISP, which then pays the other end through clearing, but this model leads to more transactions so it will be less competitive.

[50] Of course, a clearing function will have an account relationship with each edge network so actual currency does not have to move for each session. Only the account is incremented or decremented, then settled say monthly.

This model of agreeing tariffs independently with each neighbour is called split edge pricing, because at an interconnect boundary (e.g. that between N_A&N_B in Fig 7), neighbour N_A levies a conceptual 'split-price' on N_B for the QoS of bringing sent data from S_1 to the interface. While, N_B mirrors this, levying a split-price on N_A for the QoS of delivering received data to R_2. The actual price for data in this direction is the difference between the two split prices (with the sign determining who pays whom). For data in the other direction (lower half of the figure), they each levy a second split-price on the other, the difference between them again determining who pays whom and how much.

At each interconnect boundary there will be four split usage prices per class of service. A similar approach can be applied to capacity pricing or the pricing of an SLA.

Responsibilities for fetching both directions of traffic from their ultimate source and delivering to their ultimate destinations are split either side of their mutual boundary, even though N_B subcontracts to N_C who subcontracts to N_D. So network neighbours are each providers for the other and customers of the other. But for any class of service and direction of traffic, one may weigh heavier as a provider and lighter as a customer relative to the other.

The basic network subcontracting model of split-edge pricing has been implicit in the industry for a long time. As far as we know, it was first explicitly articulated in an Internet context by Shenker *et al* who called it edge pricing [40] (they acknowledged Van Jacobsen as their private source). More recently, it has also been called the virtual pairwise model [44]. Alternative models are discussed in §7.3.

Tariff diversity. Edge pricing is a powerful model because it allows unfettered tariff innovation.[51] Each network can choose to agree an interconnect tariff with each of its neighbours separately without the metrics chosen in one agreement constraining the metrics used in the others. So if a network comes up with a new tariff idea, it can introduce it without having to take it to standards for agreement.

The metric (e.g. bit volume, session duration or congestion marks) that one network agrees with one neighbour can be multiplied by the agreed price to produce a result in the common units of money. So if N_B is planning to transit a large new flow of traffic from N_A to N_C, even if it has agreed to use completely different metrics for the networks either side, it can work out the implications of one agreed price on how it should set the other, by normalising everything to money units.

Direction of payment. If usage-charging is going to be used, maximum value can be realised by enabling either end or both to pay, to combine the value available from both ends. However, moving money across networks incurs transaction costs. So, the above paper [5] carefully considered what the default bulk usage-charging model should be for the Internet. Any model other than the default would require clearing, but the default would require no clearing, and hence no extra transaction costs. On the basis that the large majority of communications proceed with the consent of both ends, in pure economic terms the default should be sender and receiver both pay.

However, back in 1999 the paper also predicted that unsolicited traffic would become a problem if customers were usage-charged for traffic they received — termed denial-of-funds attacks. The paper worked through an imaginary game where some networks offered 'sender-pays', some 'receiver-pays' and some 'both-pay'. At that time, both-pay won that conceptual game — the estimated cost of denial-of-funds attacks was considered less than the value released by allowing receivers to share usage costs with senders. That was 1999.

Now we all experience huge volumes of spam, background traffic from virus-infected zombie hosts, flooding attacks and gratuitous advertising attached to Web pages & e-mail. With the current Internet, you have some hope of controlling what you send, but no control over what you receive. So if we played that game in 2005, sender-pays would win, therefore we would recommend a sender-pays default model for bulk cost-based usage charging (Fig 8).

The current best-efforts interconnect model is 'both-pay' so we expect the industry to use this model for QoS interconnect initially. But we expect piecemeal change to the sender-pays model as QoS charging is exploited more frequently to launch 'denial of funds' attacks.

7.3 Avoiding brittle structures

Edge pricing has been such a ubiquitous model, we are in danger of becoming complacent about its worth. Sometimes, the way a new capability is planned to be introduced would violate the model. Few people understand how much would be at stake

[51] Probably the most well-known telephony tariff innovation of recent times is 'friends and family', originally introduced by MCI. But the Internet market has seen numerous innovations in tariffs in recent years.

if we lost the edge pricing model, because nothing continually reminds us how thankful we should be for the flexibility it brings.

Carrier selection is a clear example of a violation of edge pricing. Carrier pre-selection has been imposed by telephony regulators in many countries as an attempt to improve the competitiveness of the carrier market. However, it involves the end-customer contracting with two networks at once for the same service. It would cause end-customers too much confusion if both tariffs had different structures. So neither network can change its tariff without the other agreeing. Every end-customer doing carrier selection has a pair of relationships with different pairs of networks. So, overall, every network must support the same tariffs. So, ironically, imposing carrier selection prevents healthy competition. It prevents any network deploying an innovative tariff, without taking it through standards first.

Therefore, it would be counter-productive to regulate for carrier selection in an immature market such as that for IP QoS, where encouraging tariff innovation should be paramount.

Roaming is another example of a violation of edge pricing that forces a standard set of industry tariffs. But an intermediary (a virtual mobile network operator) can alleviate this problem [13] by converting charging under one tariff to another. The end-to-edge clearing function described above has to play a similar intermediary role.

8 Conclusions

We have brought together all the disconnected approaches to Internet quality of service under an integrated model — a 'model of models' that interconnects commercial and technical diversity, relying only on the uniformity of the Internet protocol. The whole commercial and technical approach is ultimately based on the economics of the two most recently standardised bits in the IP header: the ECN field. It requires the final piece of the jigsaw: our own proposal to realign the meaning of the ECN field, but without changing IP.

The central plank of the approach is for networks to charge their neighbours a dynamically rising price as their customers cause congestion to rise. Edge networks can convert these price signals into their local approach to quality of service, which then still interworks with other local approaches. For instance, edge networks can provide bandwidth guarantees across the breadth of the Internet, without requiring any special flow guarantee mechanisms in the intervening networks. Or one edge network can offer some of its customers faster transmission to any destination in the world without requiring any special arrangements with the intervening networks. Other networks receive compensation for the preferential use of their capacity solely through congestion charging at the interconnect boundaries.

The interconnect accounting is extremely cheap and simple, but it preserves a precise association between congestion anywhere in the internetwork and the customers that cause it. This means that, for instance, if two customers paying different flat monthly subscriptions both make heavy use of peer-to-peer file-sharing, the amount of congestion that each is allowed to cause anyone else can be limited in proportion to their subscription fees.

Thus at the edges of the internetwork we encourage a vibrant mix of commercial (and technical) diversity, but in the middle and in wholesale markets, we expect much more uniformity. We have argued that i) in edge networks per-flow QoS guarantees will be sold under a range of retail models, while ii) in core and backbone networks a bulk QoS facility incapable of distinguishing different sessions will inevitably emerge as the sufficient charging model. Thus, margins in the middle will erode as a 'QoS value hole' grows outwards (Fig 6). An end-to-edge clearing intermediary will become a critical part of the infrastructure to move surplus value between the edge networks across this hole.

Having surveyed the interconnect tariffs and technologies used in today's balkanised mêlée, we propose our model to support evolution towards an interconnected future, allowing inter-domain tariffs to evolve piecemeal from those in use today to the more robust ones we have outlined. The architecture is a classic example of the recently articulated new design principle for the Internet: "Design for Tussle" [10].

Acknowledgements

Michael Lyons, Steve Cassidy, Gabriele Corliano, Andy Reid (BT), Siddhartha Sengupta (LBS).

9 Glossary

AS Autonomous system

ATM Asynchronous transfer mode

CDN Content distribution network

CIDR IETF classless inter-domain routing

CE IETF congestion experienced code-point of the ECN field

Diffserv IETF Differentiated Services Architecture

DPI Deep packet inspection

DSCP IETF Diffserv code-point

DNS Domain Name Service

DSL Digital subscriber line

DVB Digital video broadcasting

ECN IETF explicit congestion notification

ECT IETF ECN-capable transport code-point of the ECN field

FEC MPLS forwarding equivalence class

GPRS General Packet Radio Service

GQS Guaranteed QoS synthesis(er)

IARS ITU International Accounting Rate System

ICT Information and communications technology

IETF Internet Engineering Task Force

Intserv IETF Integrated Services Architecture (ISA)

IP Internet Protocol

IPsec IETF IP security

ISP Internet service provider

ITU International Telecommunication Union

MPLS Multi-protocol label switching

NSIS IETF Next Steps in Signaling working group

OSP ETSI Open Settlement Protocol

QoS Quality of service

PSTN Public Switched Telephone Network

RFC IETF Request for comments

RSVP IETF Resource Reservation Protocol

SIP IETF Session Initiation Protocol

SLA Service level agreement

SMS Short Message Service

TCP Transmission Control Protocol

TTL IP time to live

UMTS Universal Mobile Telecommunications System

VoIP Voice over IP

VPN Virtual private network

References

[1] John Adams, Lawrence G. Roberts, and Avril IJsselmuiden. Changing the Internet to support real-time content supply from a large fraction of broadband residential users. *BTTJ*, 23(2), April 2005.

[2] F. Baker, B. Braden, S. Bradner, A. Mankin, M. O'Dell, A. Romanow, A. Weinrib, and L. Zhang. Resource ReSerVation protocol (RSVP) — version 1 applicability statement; Some guidelines on deployment. Request for comments 2208, Internet Engineering Task Force, URL: rfc2208.txt, January 1997.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. Request for comments 2475, Internet Engineering Task Force, URL: rfc2475.txt, December 1998.

[4] R. Braden (Ed.), L. Zhang, S. Berson, S. Herzog, and S. Jamin. Resource ReSerVation Protocol (RSVP) — version 1 functional specification. Request for comments 2205, Internet Engineering Task Force, URL: rfc2205.txt, September 1997.

[5] Bob Briscoe. The direction of value flow in multi-service connectionless networks. In *Proc. International Conference on Telecommunicatons and E-Commerce (ICTEC'99)*, URL: http://www.m3i.org/papers/valflow_ictec99.pdf, October 1999.

[6] Bob Briscoe. The value of network interconnection. Technical Report TR-CXR9-2005-004, BT, URL: http://www.jungle.bt.co.uk/projects/2020comms/net/ixval/ixval_tr.pdf, February 2005.

[7] Bob Briscoe, Arnaud Jacquet, Carla Di Cairano-Gilfedder, Andrea Soppera, and Martin Koyabe. Policing congestion response in an inter-network using re-feedback. *Proc. ACM SIGCOMM'05, Computer Communication Review*, 35(4), September 2005. (to appear).

[8] CableLabs. PacketCable 1.5 dynamic quality-of-service. Specification PKT-SP-DQOS-I10-040721, Cable Television Laboratories Inc, URL: http://www.packetcable.com/downloads/specs/PKT-SP-DQOS-I10-040721.pdf, July 2004.

[9] CableLabs. PacketCable 1.5 architecture framework. Technical Report PKT-TR-ARCH1.5-V01-050128, Cable Television Laboratories Inc, URL: http://www.packetcable.com/downloads/specs/PKT-TR-ARCH1.5-V01-050128.pdf, January 2005.

[10] David Clark, Karen Sollins, John Wroclawski, and Robert Braden. Tussle in cyberspace: Defining tomorrow's Internet. *Proc. ACM SIGCOMM'02, Computer Communication Review*, 32(4), August 2002.

[11] David D. Clark. A model for cost allocation and pricing in the Internet. In *Proc. MIT Workshop on Internet Economics*, URL: http://www.press.umich.edu/jep/works/ClarkModel.html, March 1995.

[12] Ioanna D. Constantiou and Costas A. Courcoubetis. Information asymmetry models in the Internet connectivity market. In *Proc. 4th Internet Economics Workshop*, URL: http://www.m3i.org/papers/ie.pdf, May 2001.

[13] Gabriele Corlianò and Kashaf Khan. Economic tussles in the public mobile access market. *BTTJ*, 21(3), July 2003.

[14] Costas Courcoubetis and Richard Weber. Buffer overflow asymptotics for a switch handling many traffic sources. *Journal Applied Probability*, 33:886–903, 1996.

[15] Maria Cuevas. Admission control and resource reservation for session-based applications in next generation networks. *BTTJ*, 23(2), April 2005.

[16] David Durham, Jim Boyle, Ron Cohen, Shai Herzog, Raju Rajan, and Arun Sastry. The COPS (common open policy service) protocol. Request for comments 2748, Internet Engineering Task Force, URL: rfc2748.txt, January 2000.

[17] Open settlement protocol (OSP) version v2.1.1. Technical specification 101321, ETSI TIPHON, URL: http://www.etsi.org/, August 2000.

[18] Errin W. Fulp and Douglas S. Reeves. The economic impact of network pricing intervals. In *Proc. Internet Charging and QoS Technology (ICQT'02)*, number 2511, pages 315–324, URL: http://www.springer.de/cgi-bin/search_book.pl?isbn=3-540-44356-8 or http://arqos.csc.ncsu.edu/papers/2002-07-icqt02-pricing-intervals.pdf; Slides URL: http://www.tik.ee.ethz.ch/~qofis02/icqt_s11_1.pdf, October 2002. Springer LNCS.

[19] Richard J. Gibbens and Frank P. Kelly. On packet marking at priority queues. *IEEE Transactions on Automatic Control*, 47(6):1016–1020, June 2002.

[20] Emanuele Giovannetti, Alessio D'Ignazio, Joerge Lepler, Cristiano Ristuccia, and Stefanie Brilon. Initial data-set on transit prices and quality. Deliverable D5-WP1, CoCombine IST project IST-2004-2012, URL: http://www.cocombine.org/pdf/D5_final.pdf, November 2004.

[21] David K. Goldenberg, Lili Qiu, Haiyong Xie, Yang Richard Yang, and Yin Zhang. Optimizing cost and performance for multihoming. *Proc. ACM SIGCOMM'04, Computer Communication Review*, 34(4):79–92, October 2004.

[22] Peter Hovell, Bob Briscoe, and Gabriele Corlianò. Guaranteed QoS synthesis (GQS): An example of a scalable core IP quality of service solution. *BTTJ*, 23(2), April 2005.

[23] Reforming the international accounting rate system. Web site, ITU, URL: http://www.itu.int/osg/spu/intset/, 2001.

[24] Van Jacobsen. Congestion avoidance and control. *Proc. ACM SIGCOMM'88, Computer Communication Review*, 18(4):314–329, 1988.

[25] Darren Johnson. QoS control versus generous dimensioning. *BTTJ*, 23(2), April 2005.

[26] Frank P. Kelly, Aman K. Maulloo, and David K. H. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49(3):237–252, 1998.

[27] S. Kunniyur and R. Srikant. Analysis and design of an adaptive virtual queue (AVQ) algorithm for active queue management. *Proc. ACM SIGCOMM'01, Computer Communication Review*, 31(4), October 2001.

[28] Robert Lechner. Competitive network evolution towards broadband IP. In Bruce Wiltshire, editor, *Proc. BT Alliance Engineering Symposium (AES'99)*, page Session 1A Paper 2. BT Alliance, June 1999.

[29] LINX — Fees schedule. Web page URL: http://www.linx.net/joining/fee-sched.thtml, July 2004. Version 5.0.

[30] Michael Lyons. Information, networks and economics. *The Journal (IBTEJ)*, Jan–Mar 2000:40–44, 2000.

[31] Sean McCreary and kc claffy. How far does the average packet travel on the Internet? Technical report, CAIDA, URL: http://www.caida.org/Learn/ASPL/, May 1998.

[32] Robert M. Metcalfe. Metcalfe's Law: A network becomes more valuable as it reaches more users. InfoWorld opinion column, "From the Ether" Web URL: http://www.infoworld.com/cgi-bin/displayNew.pl?/metcalfe/bmlist.htm, October 1995.

[33] Dave Mustill and Peter Willis. Delivering QoS in the next generation network — A standards perspective. *BTTJ*, 23(2), April 2005.

[34] Klara Nahrstedt and Jonathan M. Smith. The QoS broker. *IEEE Multimedia*, 2(1):53–67, 1995.

[35] K. Nichols, S. Blake, F.Baker, and D. Black. Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers. Request for comments 2474, Internet Engineering Task Force, URL: rfc2474.txt, December 1998.

[36] Kathleen Nichols, Van Jacobson, and Lixia Zhang. A two-bit differentiated services architecture for the Internet. Request for comments, Internet Engineering Task Force, URL: rfc2638.txt or URL: http://diffserv.lcs.mit.edu/Drafts/draft-nichols-diff-svc-arch-00.pdf, July 1999. Based on draft-nichols-diff-svc-arch-02 (April, 1999, originally Nov 1997).

[37] Andrew Odlyzko. A modest proposal for preventing Internet congestion. Technical report TR 97.35.1, AT&T Research, Florham Park, New Jersey, URL: http://www.research.att.com/~trmaster/TRs/97/97.35/97.35.1.body.ps or URL: http://www.research.att.com/~amo/doc/modest.proposal.pdf, September 1997.

[38] K. K. Ramakrishnan, Sally Floyd, and David Black. The addition of explicit congestion notification (ECN) to IP. Request for comments 3168, Internet Engineering Task Force, URL: rfc3168.txt, September 2001.

[39] Andy Reid. Economics and scalability of QoS solutions. *BTTJ*, 23(2), April 2005.

[40] Scott Shenker, David Clark, Deborah Estrin, and Shai Herzog. Pricing in computer networks: Reshaping the research agenda. *ACM SIGCOMM Computer Communication Review*, 26(2), April 1996.

[41] The value of IP clearing & settlement; new revenue opportunities for IP carriers. White paper, Transnexus, URL: http://www.transnexus.com/White%20Papers/Value%20of%20IP%20Clearing%20and%20Settlement.pdf, 1997?

[42] R. Yavatkar, D. Pendarakis, and R. Guerin. A framework for policy-based admission control. Request for comments 2753, Internet Engineering Task Force, URL: rfc2753.txt, January 2000.

[43] Lixia Zhang, Stephen Deering, Deborah Estrin, Scott Shenker, and Daniel Zappala. RSVP: A new resource ReSerVation protocol. *IEEE Network*, September 1993.

[44] Raymond Zhang. QoS-enabled inter-provider MPLS VPN services. Web URL: http://cfp.mit.edu/meetings/oct04/agenda_oct-20-21_04.html, October 2004. (Presentation to CII-CFP inter-provider QoS working group).