

# QoS: Solution Waiting For A Problem

Smriti Bhagat

Department of Computer Science

Rutgers University

Piscataway, New Jersey 08854

## Abstract

*Quality of Service (QoS) in data networks has been debated for atleast 35 years. Despite two decades of vigorous research and standards activity, QoS schemes have floundered. Research in QoS mechanisms assumes the existence of a large market for QoS-sensitive applications that will justify the expense of enabling QoS in the network. However, the lack of need for hard guarantees by both the customer and the internet service provider, is one of the fundamental reasons for the failure of QoS mechanisms. QoS research seems to provide a solution to a non-existent problem that may never arise.*

*In this position paper, we argue that it is needless to invest in building QoS solutions. The widely used overprovisioning technique is a sufficient and remarkable traffic engineering approach.*

## 1. Introduction

With the convergence of voice, video and data in enterprise networks, significant amount of research has focused on internet quality of service. In his statement, S. Keshav rightly articulates the fundamental goal of networking [8]:

*"The Holy Grail of computer networking is to design a network that has the flexibility and low cost of the Internet, yet offers the end-to-end quality-of-service guarantees of the telephone network."*

In an effort to work towards this goal, network researchers have proposed sophisticated QoS architectures intended to deliver performance guarantees. Nevertheless, none of these mechanisms are deployed widely across the public internet. In this position paper we argue that the time and effort spent

into QoS research has been wastefully expended, knowing the little commercial deployment of these techniques. The simple over-dimensioning techniques suffice to provide the desired network performance.

In the rest of the paper, we describe the mechanisms developed to cater to a speculated need for QoS and the various reasons for their failure. We provide facts to support for the idea that the current best-effort service can sustain the network traffic requirements of the future and present some thoughts on the future of QoS research.

## 2. What is QoS?

Quality of Service (QoS) is usually referred to as a measure of the ability of network to provide different levels of services to selected applications and associated network flows. The idea is to provide high priority to latency-sensitive applications, while low preference to all other traffic. More formally, there are two widely accepted definitions of QoS:

- "The collective effect of service performance which determines the degree of satisfaction of a user of the service." (ITU-T: International Telecommunication Union - Telecommunication Standardization Sector) [2]
- "A set of service requirements to be met by the network while transporting a flow." (IETF: Internet Engineering Task Force) [7]

Although, a user's perception of QoS is very subjective (usually indicating high quality), both these definitions describe QoS as some quantifiable performance. QoS can be measured in terms of different metrics like latency, jitter, bandwidth, packet loss etc.

## 2.1. QoS Mechanisms

Some classical models of traffic handling for QoS support are:

*IntServ*: The Integrated Services (IntServ) model integrates resource reservation and traffic control mechanisms to support special handling of individual traffic flows. A new flow with its QoS requirements is served only if the network side has enough available resources guaranteeing strict-QoS requirements.

*DiffServ*: The Differentiated Services (Diffserv) architecture [9] [3] uses traffic control enable service discrimination by aggregating packets of flows with similar QoS requirements into corresponding service classes (assured or premium). This ensures low priority for non-mission critical tasks and hence guaranteeing low latency for high premium traffic. There are other schemes that are gaining popularity like multiprotocol label switching (MPLS) etc. However, they provide traffic engineering solutions and not QoS guarantees.

The internet's default best-effort service also provides QoS by *overprovisioning* the network. This is achieved by providing more resources than actually required to accommodate demand fluctuations. This simple and highly effective technique provides high-quality services to all packets on an IP backbone. The simplicity of the approach and low management costs encourages service providers like Sprint [6] to employ overprovisioning in the core of the network.

## 2.2. QoS in-action or inaction?

Despite research on many technologically sound QoS mechanisms, none has had a significant implication on the public internet. The IntServ technique failed due to its implementation constraints. The underlying concept of the necessity of per-flow signaling elaboration in routers is not scalable. Maintaining and processing per-flow control information in each router is not realistic and makes the router's performance very poor, since the number of flows the router accommodates in a core network is extremely large. Although DiffServ does not have a scalability problem like IntServ, it provides less flexibility. In practice, the deployment of both these techniques is rare. The failure of QoS mechanisms and the reasons behind it in greater detail in Section 3. We highlight the fact that most of the issues QoS is attempting to solve are not real problems in today's well-provisioned internet.

## 3. Need for QoS?

It is believed that customers working on mission-critical and real time applications have an economic incentive to in-

vest in QoS capabilities so that acceptable response times are guaranteed within certain tolerances. Advocates of QoS defend the necessity of QoS for high performance distributed applications and mission-critical applications. This would enable a new class of real-time applications like video conferencing, internet telephony etc. We do acknowledge the necessity of low-latency guarantees for critical applications like medical-surgery, military applications etc. Nevertheless, we argue that customers with such mission-critical applications (with disastrous failures), can purchase extra bandwidth to ensure availability of resources.

QoS researchers argue that overprovisioning is expensive and inadequate for ensuring low latency for real time applications. If the problem is to provide a good service to a certain class of traffic, then, in a modern world of inexpensive fiber, the solution is clear: by supplying sufficient capacity, provide the same good service to all traffic. A survey by University of Twente at Netherlands [10] reports that fiber has become so inexpensive that for the costs of a few miles of highway, it is now possible to create a nation-wide backbone in the Tbps range. Thus, network link capacity is no longer a scarce resource.

Low costs also enable very good best-effort service with abundant capacity, which wins over rationed capacity with only select traffic getting better than average treatment. In the past five years the growth of available backbone capacity has exceeded the growth of Internet usage, which remained stable at approximately 100% per year. Current and the predicted future demands can be easily handled with simple over-engineering schemes. Another argument against overprovisioned networks is that they are wasteful as they work at a very low utilization levels, typically below 15% of their capacity. We argue that even if QoS techniques are deployed, some level of overprovisioning is still mandatory to provide headroom for usage fluctuations. When the huge overheads linked with QoS mechanisms can be easily avoided by adding cheap resources, it is clear that there is no strong motivation behind these techniques.

*Sprint Example*: We present the statistics [1] from a major ISP, Sprint, which believes in a simple overprovisioning architecture, with no complex QoS techniques. The average one-way delay is 18.77ms and the minimum delay 8.71ms. Although the maximum delay observed over a period of 6 months was 924.11ms, it was an outlier, as the standard deviation is 7.74ms. The mean delay values are too small to affect the performance of any real-time application.

Note that most QoS measurements are based on bandwidth allocation, while the primary concern for real-time applications is latency. In the example above, we use the network delay as a measure of performance. In case of IntServ, it can be argued that presence of bandwidth may

not always guarantee low latency.

The applications that require strict-QoS are not well identified. One of the most cited application by QoS advocates, voice over IP (VoIP) is already widely used with good performance. Most VoIP customers are satisfied with the quality of the service they get. It is intuitive to believe that these customers would not be willing to pay extra for the same service, with some extra guarantees that protect them from the rare worst-case of long network latencies. It seems that QoS research is building QoS-sensitive solutions for QoS-insensitive customers.

As exemplified above, the current best-effort network performance is sufficiently high for real-time voice and video applications. This raises the question of whether QoS research is creating solutions for an illusory problem?

## 4. QoS Failure

A statement issued by the Internet2 QoS Working Group in May 2002, (and updated recently in Jan 2006) is an affidavit of the failure of QoS solutions [11]. This group, whose mission is to support the development of IP traffic differentiation services, claims that Premium IP Services based on DiffServ principles have failed: *“Despite considerable effort and success with proof-of-concept demonstrations, this effort yielded no operational deployments and has been suspended indefinitely.”* [11]. There is no future of QoS research when the promoters of the technology acknowledge its failure.

There has been a lot of research on engineered QoS designs that not necessarily address the real need of customers and providers. On the other hand, there has been insufficient research into new internet pricing and business models. The downfall of QoS mechanisms can be attributed to a multitude of issues ranging from technological to business-related [12]. In this section we list some of these issues that impacted the failure of QoS mechanisms. Although we have tried to categorize these issues, they are not independent and have some degree of overlap.

### 4.1. Deployment

The limited (if any) deployment of QoS architectures, evinces their glaring deficiencies. There has been no smooth transition from best-effort to QoS. The complexity of QoS systems discourages large ISPs from widely utilizing them. Upgrading to QoS routers and hardware significantly decreases the manageability of the network. Current QoS-enabled routers have an insufficient number of queues to support large scale service differentiation. These routers are (in real time) designed to test for an additional parameter,

the forwarding priority of each packet. This may lead to some performance impact on the routers particularly when performing classification functions at the edges. This is a reasonable speculation even though there aren't any large scale deployment of QoS architectures to verify it.

### 4.2. Costs

A representative majority of service providers believe that, in today's fiber-based networks, it is cheaper to provision generously than to deploy any form of QoS that gives better treatment to a privileged traffic class. Besides the one time deployment costs, QoS techniques add a substantial amount of manageability costs. Trained network operators are required to configure and maintain the QoS architectures. Agreements have to be made between service providers to ensure inter-operable QoS implementations. A QoS functional model needs to be developed to support stringent service level agreements (SLA). Developing and negotiating on a service level specification (SLS) is another overhead for the service provider, which adds the costs of SLA auditing. The mechanisms to monitor SLA are expensive. Which makes it significantly harder for a customer to claim loss if the SLA was not met. With a limited scope for return on investment from QoS architectures, the cost of their deployment is prohibitively high.

### 4.3. Pricing Model

Success of a new protocol or standard largely depends on its economic benefits. The greatest criticism of QoS is the absence of a business model. There is a paradoxical situation here: there is no model to charge for high quality of service, and hence no economical benefit in providing the services, which in turn leads to the absence of a pricing model. This deadlock situation further adds to the lack of motivation in investing in QoS mechanisms. In particular, price discrimination is economically efficient when there exist distinct market segments with different willingness to pay for basically the same service. QoS guarantees through traffic contracts (for individual flows or traffic aggregates) are only advantageous in situations of overload. This, of course, may be a useful distinction if overloads are frequent or have very serious consequences when they do occur. However, there is no means to ensure that a premium service is consistently better than best effort on a regular basis [5].

### 4.4. End to End Requirements

QoS is typically considered to be an end-to-end phenomenon. Even if one link in the end-to-end path fails, the application does not receive the QoS it needs. Thus enterprise

QoS deployment is important when viewing QoS from the application perspective. A critical issue is the assurance of end-to-end QoS coherence in the face of multiple intervening parties with heterogeneous characteristics. The need for negotiations across service providers is inevitable and undesirable. ISPs do not have strong incentives to reach agreements with other providers to enable end-to-end QoS services. Since there are no well defined services categorized as premium, it makes the issue of negotiating the QoS-related aspects of an SLS harder, than if the customer was offered the same set of services by multiple providers.

A fundamental issue in the quest for end-to-end QoS is that by definition, communication is bidirectional. If more than one consumers are involved, like in a VoIP telephonic conversation, there is no mechanism to ensure quality of the overall experience in case of conflicting quality levels. Furthermore, the nature of IP data transport is unidirectional and the network does not establish any relationship between both directions of the same flow. Although filtering and/or signaling in the network could help tackle this issue, it is computationally very intensive to be realized on a per packet basis.

#### 4.5. Security

Additional mechanisms need to be implemented to handle critical security issues related to QoS. The questions of accountability and non-repudiation need to be effectively addressed. A network operator must have means to authenticate the user who has a traffic contract with the ISP. Also, the control data of QoS mechanisms needs to be secured. An attacker must not be able to change the control data (the priority class information) to flood a provider with prioritized traffic. This could have serious implications and make network behavior even worse than without QoS mechanisms. Like mentioned earlier, all these overheads come with monetary and performance costs.

#### 4.6. Quality Experience

QoS service is about the performance that would be experienced by a customer in the event of a network denial of service (DoS) attack. That is, QoS is about the assurance. What networks cannot offer today is a guarantee, though such guarantees are sought only by customers who perceive that they are necessary and are willing to pay extra for them. Customers do not have any apparent incentives in investing in a service that is indistinguishable from already excellent best-effort in the average case and offers no additional assurance of delivering the quality they need in the worst case.

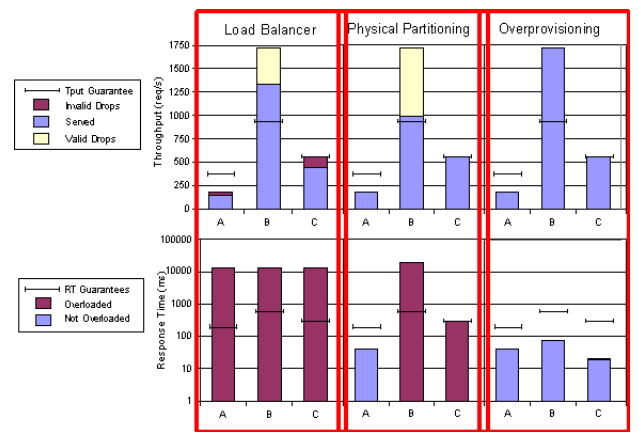


Figure 1. QoS vs Overprovisioning: Results from the Teoma benchmark. Source: [4]

## 5. Overprovisioning

If the problem is to allow use of the full capacity of the network with some traffic still getting good service, then differentiated services are not the answer. Overprovisioning is a simpler solution that is much easier to deploy and does not require dramatic changes to operational practices or pricing models. The idea is to throw bandwidth at the problem of providing QoS. Its a low cost solution because of the huge amounts of bandwidth in backbones, which has increased per-fiber bandwidths hugely. Studies have shown that in most big ISPs, if they do have a congestion problem, adding bandwidth to that part of the network is a lot a less complex than putting in specific network-wide mechanisms to provide QoS. Overprovisioning alleviates some of the fundamental issues related to QoS deployment as mention in the previous sections.

## 6. Experimental Study

We now present a case study performed by a group in the University of California, Santa Barbara [4]. They propose a new QoS solution, a *Quorum* architecture, for traffic shaping and admission control. They compare this with existing QoS Solutions, they call *Physical Partitioning*, no QoS or *Load Balancing* and with *Overprovisioning*. We use the results of this study to highlight the performance of overprovisioning and compare it with QoS using differentiated services. (We do not show the complete result of the Quorum system.) Three services are emulated e-Commerce, Stocks and Search, with labels A,B,C resp. in figure 1. The details of the experimental setup and further explanation is available in [4]. Also, refer to figure 2.

Service Class	QoS Guarantees		
	90th percentile Resp. Time	Average Throughput	Avg Compute Requirement
e-Commerce	150ms	350 req/s	10 ms/req
Stocks	500ms	175 req/s	30 ms/req
Search	600ms	18 req/s	100 ms/req

(a)

Experimental Workload	
Average Input Traffic	Service Status
170 req/s	Not Overloaded
331 req/s	Overloaded
12.4 req/s	Not Overloaded

(b)

**Figure 2. QoS Guarantee and Traffic Workload. Source: [4]**

As observed in Figure 1, the amount of traffic served with no QoS is directly proportional to the input demands, and it fails when the demand exceeds the available resources. The throughput guarantees are maintained by the QoS enabled system. It drops packets, when the traffic is more than the agreed amount. However, overprovisioning by 4 times, or link utilization of average 25% can be seen as an extremely efficient technique. It is able to serve all the traffic.

The response time results in figure 1, show that overprovisioning is the only mechanism that can provide response time guarantees for any class of traffic. Even though QoS enables service provides the promised guarantees for the premium services, its performance on the low-priority services is 10 times the maximum allowed.

## 7. Conclusion

QoS researchers have spent lots of time and effort in inventing novel QoS mechanisms on the assumption that the market for these exists. However, these services are neither deployable nor marketable across the public network. The significant cost and complexity of QoS services makes them far from a solution. Best-effort service already provides near-zero loss and low delays disregarding any traffic classification. Other services that provide utilization improvement over best-effort and do not have the complexity of QoS are yet to be engineered to a level of large-scale deployment. As of now, well-provisioned networks are likely to remain the norm. With the available low-cost fiber, and simpler overprovisioning techniques, it is surely time to give up on the idea of QoS mechanisms.

## References

- [1] <http://www.networkworld.com/research/2002/1216isptestside1.html>.
- [2] Recommendation E.800: Terms and definitions related to quality of service and network performance including dependability. *ITU-T*, August 1994.
- [3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. *RFC 2475*, 1998.
- [4] J. M. Blanquer, A. Batchelli, K. Schauer, and R. Wolski. QoS for internet services done right. *11th ACM SIGOPS European Workshop*, Sept 2004.
- [5] L. Burgstahler, K. Dolzer, C. Hauser, J. Jahnert, S. Junghans, C. Macian, and W. Payer. Beyond technology: the missing pieces for QoS success. In *RIPQoS '03: Proceedings of the ACM SIGCOMM workshop on Revisiting IP QoS*, pages 121–130, New York, NY, USA, 2003. ACM Press.
- [6] C. Diot, D. Meyer, and P. Whiting. MPLS and the Sprint E—Solutions IP backbone network. June 2001.
- [7] B. R. E. Crowley, R. Nair and H. Sandick. A framework for QoS based routing in the internet. *RFC 2386, IETF*, August 1998.
- [8] S. Keshav. *An Engineering Approach to Computer Networking*. Addison-Wesley, 1997.
- [9] V. P. Kumar, T. V. Lakshman, and D. Stiliadis. Beyond best effort: Router architectures for the differentiated services of tomorrow's internet. *IEEE Communications*, 36(5):152–163, 1998.
- [10] A. Pras, R. van de Meent, and M. Mandjes. QoS in hybrid networks - An operator's perspective. *IWQoS, LNCS*, 2005.
- [11] B. Teitelbaum and S. Shalunov. Why Premium IP service has not been deployed (and probably never will). *Internet2 QoS Working Group Informational Document*, 2002.
- [12] B. Teitelbaum and S. Shalunov. What QoS research hasn't understood about risk. In *RIPQoS '03: Proceedings of the ACM SIGCOMM workshop on Revisiting IP QoS*, pages 148–150, New York, NY, USA, 2003. ACM Press.