Preface

Traditionally, Internet Protocol or IP networks have only offered a "best effort" delivery service for IP traffic; in these best-effort networks all traffic is treated equally. The service requirements – or more specifically service level agreement (SLA) requirements – of, voice, video, and mission critical data applications, for example, are not the same. Consequently, "best effort" IP networks have not been able to provide optimal support for multiservice applications with different SLA requirements.

Broadly speaking "quality of service" or QOS (either pronounced "Q-O-S" or "kwos") is the term used to describe the science of engineering a network to make it work well for applications by treating traffic from applications differently depending upon their SLA requirements. In the 5–10 years preceding this publication there have been significant developments in IP QOS to the point where the mechanisms, architectures, and deployment experience are now available to enable optimized support for multiservice applications on an integrated IP network. IP is becoming the convergence technology for multimedia services and consequently QOS is one of the hottest topics in IP networking, and yet currently it is still one of the least well understood from a practical perspective. Ten years ago, the design and implementation of large IP networks using routing protocols like OSPF and BGP was seen as a very specialist subject,

restricted to the gurus of the networking community. Today, however, with the proliferation of the Internet and large IP networks, much of the mysticism associated with these technologies has gone, and an understanding of them has moved into the mainstream. IP QOS today is seen as a specialist subject, much as OSPF and BGP were ten years ago.

In this book we hope to help to bring IP QOS more into the mainstream, through bridging the theory of QOS with the practice of deployment, from SLA definition to detailed design and configuration. We describe the key application SLA requirements, QOS functions, and architectures to help readers understand the concepts of IP QOS; case studies and examples are used to show how these concepts are applied in practice. In the process, we address some of the most common QOS questions:

- What's the difference between QOS, COS, and TOS? (Chapter 2, Section 2.1.1)
- Why use IP QOS rather than using layer 2 QOS capabilities? (Chapter 2, Section 2.1.4)
- What's the difference between a queue and a buffer? (Chapter 2, Section 2.2.4.2)
- What's the difference between a shaper and a policer? (Chapter 2, Section 2.2.4.3)
- Which QOS architecture should be used? (Chapter 2, Section 2.3)
- How do you assure an end-to-end SLA for Voice over IP (VoIP) traffic? (Chapter 3, Section 3.2.2.1.1)
- How many Diffserv classes should be used? (Chapter 3, Section 3.2.2.6)
- What packet marking scheme should be used? (Chapter 3, Section 3.2.2.7)
- Why isn't ECN widely deployed? (Chapter 2, Section 2.3.4.4)

- How do you convert between TOS, IP precedence, and DSCP? (Chapter 2, Appendix 2.A)
- How do you tune RED and WRED? (Chapter 3, Section 3.4)
- What options are there for admission control? (Chapter 4)
- How do you manage an IP QOS deployment? (Chapters 5 and 6)

In addition, we debunk some common IP QOS myths including:

- Jitter is more important than delay for VoIP. (Chapter 1, Section 1.3.1.2)
- The maximum VoIP load that can be supported on a link is 33%. (Chapter 3, Section 3.2.2.8)
- If a design does not use the recommended Diffserv marking scheme it is not Diffserv compliant. (Chapter 3, Section 3.2.2.7)
- IPv6 provides better QOS than IPv4. (Chapter 2, Section 2.3.5)
- MPLS provides better QOS than IPv4 or IPv6. (Chapter 2, Section 2.3.6)

Audience

We have tried to make this book accessible to a wide audience and to address beginner, intermediate, and advanced IP QOS topics. We hope it will be of value to anyone trying to design, support, or just understand IP QOS from a practical perspective. This includes, but is not limited to, network designers, engineers, administrators, and operators, both in service provider and enterprise environments, together with students looking to gain a more applied understanding of QOS. This book also serves as a general technical reference guide to IP QOS.

Previous knowledge or understanding of QOS is not a prerequisite to reading this book; however, we have assumed that readers have a

basic level of knowledge of data networking in general, and of the basic concepts of IP, IP routing, and MPLS, in particular. There are already many good books on these subjects.

Approach

The approach that we have taken in this book is one where we describe both theory and practice, linking them, wherever possible, through the use of case studies and examples.

From a theoretical perspective, we start by describing application SLA requirements, and then explain the range of QOS functions and features that can be used to support such SLAs, within the context of an overriding QOS architecture. Where we address theory and standards, we have aimed not to blindly reproduce information that is freely available in published standards, but rather we reference available standards and augment them with explanation, description, and context which are not available from these sources. The information here will make it easier to understand the technical detail provided in standards documents, such as Internet Engineering Task Force (IETF) drafts and Requests for Comments (RFCs), and the literature provided by network equipment vendors.

Where we use case studies, they are generally based upon real-life networking scenarios, and show how theoretical SLA requirements are translated into practical network designs, which are defined in terms of example configurations. To describe the example configurations we use a Diffserv meta-language. The meta-language provides abstraction from vendor specific configurations, thereby allowing it to be understood by readers that are not familiar with particular vendor QOS implementations and configuration. The meta-language can be easily translated into most vendors' specific configurations. The case studies presented are examples, and as such do not represent the only way of doing things; rather, they aim to describe possible methodologies and to bring out the key considerations.

In focussing on IP QOS, this book centers on the network layer, or layer 3, of the Open Systems Interconnection (OSI) 7 layer reference model. Hence, where we refer to network nodes or devices, in general

we are referring to IP routers and where we refer to lower layers, we mean with respect to layer 3. As no two networks are exactly the same, throughout this book we use a generalized network model to explain the application of QOS features and functions. This model is in line with the way in which many networks are designed, consisting of a hierarchy of core, distribution, and access routers; core routers (CRs) provide connectivity between distribution routers (DRs), which in turn aggregate connections to routers at remote sites, each of which have local access routers (ARs). When deploying QOS, there is often a difference between the functionality applied at the edge of the network compared to that applied in the core; in the context of the generalized network model, the edge of the network is represented by the connectivity between the ARs and DRs, while the core of the network provides the interconnectivity between DRs and CRs.

Throughout this book, where we use the terms service provider and customer, we use them generically. These terms are not intended to infer applicability only to network service provider environments, such as virtual private network (VPN) service providers; the networking department of an enterprise organization is also service provider to their enterprise. The terms are instead intended to distinguish between the provider of the service and the user of the service.

Content and Organization

The organization of the chapters is as follows:

- *Chapter 1: QOS Requirements and Service Level Agreements.* Service level agreements (SLAs) provide the context for IP quality of service. Application and service SLA requirements are the inputs and also the qualification criteria for measuring success in a QOS design. Chapter 1 considers the SLAs metrics that are important for IP service performance, reviewing the current industry status with respect to the standardization and support of these metrics, and then describes application SLA requirements and the impacts that these metrics can have on application performance.

- *Chapter 2: Introduction to QOS Mechanics and Architectures*. Chapter 2 provides an introduction and overview to the subject of QOS. In practical terms, QOS involves using a range of functions and features (e.g. classification, scheduling, policing, shaping), within the context of an overriding architecture (e.g. Integrated Service, Differentiated Services) in order to ensure that a network service delivers the SLA characteristics required by applications. This chapter describes and discusses the key QOS functions, features and architectures.

- *Chapter 3: Deploying Diffserv*. Diffserv is by far the most widely deployed IP QOS architecture; it is widely deployed in both private enterprise networks and in service provider networks providing VPN services to enterprises. Hence, in Chapter 3, we build on the foundations set by Chapters 1 and 2, to show how the Differentiated Services architecture (Diffserv) can be practically deployed at the network edge and in the network core in order to satisfy defined application SLA requirements. This is achieved through the use of end-to-end Diffserv design case studies, which are based upon experience gained from real-world deployments. These case studies show how SLA requirements are translated into practical network designs, which are defined in terms of example configurations using the Diffserv meta-language.

- *Chapter 4: Capacity Admission Control*. Capacity admission control is the process that is used to determine whether a new flow can be granted its requested QOS without affecting those flows already granted admission. There are a number of approaches to capacity admission control, and some technologies for admission control are still evolving. Hence, Chapter 4 describes the requirement for admission control and presents a taxonomy and review of the mechanisms available for capacity admission control in IP networks.

- *Chapter 5: SLA and Network Monitoring*. After a network design has been deployed, the ability to ensure that a network service continues to deliver the required SLAs is dependent upon SLA and

network monitoring. Chapter 5 discusses the technologies and techniques available for monitoring IP QOS enabled networks, considering both passive and active network monitoring.

- *Chapter 6: Core Capacity Planning and Traffic Engineering.* Capacity planning is the process of ensuring that sufficient bandwidth is provisioned to assure that the committed SLA targets can be met. IP traffic engineering is the process of manipulating traffic on an IP network to make better use of the network capacity, by making use of capacity that would otherwise be unused, for example. Hence, capacity planning and traffic engineering are related, where traffic engineering is a tool that can be used to ensure that the available network capacity is appropriately provisioned. This chapter describes a holistic methodology for capacity planning of the core network, and describes the theory behind traffic engineering in general, and analyses some of the options and deployment considerations for the possible approaches for traffic engineering in IP networks.