NORTEL NETWORKS

Carrier IP Network
Design for Performance and Dependability

10 October 2004.

Nortel Confidential: The information contained in this document is the property of Nortel. Except as specifically authorised in writing by Nortel, the holder of this document shall keep the information contained herein confidential and shall protect the same in whole or in part from disclosure and dissemination to third parties and use same for evaluation, operation, and maintenance purposes only.

Preface

Services need to satisfy quality requirements, at least when they use the ‘managed IP network’ that the service provider controls. The managed IP network can be characterised by the requirements at its external interfaces, regardless of its design. Then design deicisions can be taken to ensure that the network satisfies these requirements.

This paper considers how the managed IP network can be designed to satisfy performance and dependability requirements due to carrier telephony and multimedia. Specifically, it discusses implementation techniques that allow performance and dependability requirements to be satisfied, but it is not concerned with configuration details.

The information in the paper should help service providers to make design decisions, confident that Nortel carrier telephony and multimedia over IP can be implemented in many possible managed IP networks.

The chapters discuss:

❒ Network design considerations.

❒ Current performance levels.

❒ Current dependability levels.

❒ Specific performance techniques.

❒ Specific dependability techniques.

Table of Contents

Abbreviations

ABR	Available Bit Rate.
AF	Assured Forwarding.
AS	Autonomous System.
ATM	Asynchronous Transfer Mode.
b/s	Bits per second.
BGP	Border Gateway Protocol.
CBR	Constant Bit Rate.
CCS7	Common Channel Signalling system 7 (also called SS7).
CF	Control Forwarding.
CoS	Class of Service.
CPE	Customer Premises Equipment.
CS	Class Selector.
CS	Communication Server.
CS2000	Communication Server 2000.
CSPF	Constrained Shortest Path First.
DF	Default Forwarding.
DiffServ	Differentiated Services (for classifying and scheduling IP traffic).
DRR	Deficit Round Robin.
DSCP	DiffServ Code Point.
ECMP	Equal Cost Multi-Path.
EF	Expedited Forwarding.
ESR8600	Ethernet Routing Switch 8600.
EXP	EXPerimental use.
E-LSP	EXP-inferred-PSC LSP.
FEC	Forwarding Equivalence Class.
FIB	Forwarding Information Base.
Gb/s	Gigabits per second.
GFR	Guaranteed Frame Rate.
GoS	Grade of Service.
IAD	Integrated Access Device.
IEEE	Institute of Electrical and Electronic Engineers.
IETF	Internet Engineering Task Force.
IGP	Interior Gateway Protocol
IP	Internet Protocol.
ISDN	Integrated Services Digital Network.
IS-IS	Intermediate System to Intermediate System.
ISP	Internet Service Provider.
ITU-T	International Telecommunication Union -Telecommunication Standardisation Sector.

Kb/s	Kilobits per second.
LAN	Local Area Network.
LDP	Label Distribution Protocol.
LER	Label Edge Router.
L-LSP	Label-only-inferred-PSC LSP.
LSA	Link State Advertisement.
LSP	Label Switched Path.
LSR	Label Switching Router.
MAN	Metropolitan Area Network.
Mb/s	Megabits per second.
MCS5200	Multimedia Communication Server 5200.
MG	Media Gateway.
MG9000	Media Gateway 9000.
MGCP	Media Gateway Control Protocol.
MIB	Management Information Base.
MLT	Multi-Link Trunking.
MOS	Mean Opinion Score.
MPLS	Multi-Protocol Label Switching.
MPLS-TE	Multi-Protocol Label Switching Traffic Engineering.
MTA	Multimedia Terminal Adapter.
MTBF	Mean Time Between Failures.
MTTR	Mean Time To Recover.
MTU	Maximum Transmission Unit.
NOC	Network Operations Centre.
nrt-VBR	Non Real Time Variable Bit Rate.
OAM&P	Operations, Administration, Maintenance and Provisioning.
OSPF	Open Shortest Path First.
PDB	Per-Domain Behaviour.
PHB	Per-Hop Behaviour.
PoP	Point of Presence.
PPP	Point to Point Protocol.
PSC	PHB Scheduling Class.
PSTN	Public Switched Telephone Network.
PVG	Packet Voice Gateway.
QoS	Quality of Service.
RFC	Request For Comment (for defining Internet standards, describing best current practices or providing other information available through the IETF).
RIB	Routing Information Base.
RIPE	Réseaux IP Européens.
RMON	Remote MONitoring.
RSVP	Resource reSerVation Protocol.
RSVP-TE	Resource reSerVation Protocol Tunneling Extensions.
RTP	Real Time Protocol (for carrying media streams, including fax streams).
RTCP	Real Time Control Protocol (for monitoring delivery to complement RTP).
rt-VBR	Real Time Variable Bit Rate.

SCTP	Stream Control Transmission Protocol (for transporting multiple streams reliably).
SDH	Synchronous Digital Hierarchy.
SIP	Session Initiation Protocol.
SIP-T	Session Initiation Protocol for Telephony (for supporting SS7 encapsulation).
SLA	Service Level Agreement.
SMLT	Split Multi-Link Trunking.
SNMP	Simple Network Management Protocol.
SONET	Synchronous Optical NETwork.
SS7	Signalling System number 7 (also called CCS7).
STM	Synchronous Transfer Mode (SDH signal format).
STP	Signalling Transfer Point.
SSP	Signalling Service Point.
TCP	Transmission Control Protocol (for transporting single streams reliably).
TDM	Time Division Multiplexing.
TE	Traffic Engineering
ToS	Type of Service.
UBR	Unspecified Bit Rate.
UDP	User Datagram Protocol (for transporting streams without reliable delivery).
VBR	Variable Bit Rate.
VC	Virtual Circuit.
VC	Virtual Container.
VPN	Virtual Private Network.
WAN	Wide Area Network.
WFQ	Weighted Fair Queueing.
WRED	Weighted Random Early Detection.
WRR	Weighted Round Robin.

References

ITU-T

G.711 Pulse code modulation (PCM) of voice frequencies.
G.723.1 Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s.
G.729 Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP).

H.248 Gateway control protocol.
H.323 Packet-based multimedia communications systems.

T.38 Procedures for real-time Group 3 facsimile communication over IP networks.

X.146 Performance objectives and quality of service classes applicable to frame relay.

Y.1720 Protection switching for MPLS networks.

IETF

RFC 768 User Datagram Protocol.
RFC 791 Internet Protocol.
RFC 792 Internet Control Message Protocol.
RFC 793 Transmission Control Protocol DARPA Internet program Protocol specification.

RFC 1195 Use of OSI IS-IS for Routing in TCP/IP and Dual Environments.
RFC 1663 Integrated Services in the Internet Architecture: an Overview.
RFC 1771 A Border Gateway Protocol 4 (BGP-4).
RFC 1812 Requirements for IP Version 4 Routers.
RFC 1990 The PPP Multilink Protocol (MP).

RFC 2205 Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification.
RFC 2309 Recommendations on Queue Management and Congestion Avoidance in the Internet.
RFC 2328 OSPF Version 2.
RFC 2338 Virtual Router Redundancy Protocol.
RFC 2474 Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers.
RFC 2475 An Architecture for Differentiated Services.
RFC 2547 BGP/MPLS VPNs.
RFC 2597 Assured Forwarding PHB Group.
RFC 2676 QoS Routing Mechanisms and OSPF Extensions.
RFC 2686 The Multi-Class Extension to Multi-Link PPP.

RFC 2697 A Single Rate Three Color Marker.
RFC 2698 A Two Rate Three Color Marker.
RFC 2702 Requirements for Traffic Engineering over MPLS
RFC 2764 A Framework for IP Based Virtual Private Networks.
RFC 2833 RTP Payload for DTMF Digits, Telephony Tones and Telephony Signals.
RFC 2960 Stream Control Transmission Protocol.
RFC 2963 A Rate Adaptive Shaper for Differentiated Services.

RFC 3031 Multiprotocol Label Switching Architecture.
RFC 3034 Use of label switching on frame relay networks specification.
RFC 3035 MPLS using LDP and ATM VC switching.
RFC 3036 LDP Specification.
RFC 3086 Definition of Differentiated Services Per Domain Behaviors and Rules for their Specification.
RFC 3107 Carrying Label Information in BGP-4.
RFC 3168 The Addition of Explicit Congestion Notification (ECN) to IP.
RFC 3209 RSVP-TE: Extensions to RSVP for LSP Tunnels.
RFC 3246 An Expedited Forwarding PHB (Per-Hop Behavior).
RFC 3261 SIP: Session Initiation Protocol.
RFC 3270 Multi-Protocol Label Switching (MPLS) Support of Differentiated Services.
RFC 3272 Overview and Principles of Internet Traffic Engineering.
RFC 3550 RTP: A Transport Protocol for Real-Time Applications.
RFC 3564 Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering.
RFC 3623 Graceful OSPF Restart.
RFC 3630 Traffic Engineering (TE) Extensions to OSPF Version 2.
RFC 3662 A Lower Effort Per-Domain Behavior (PDB) for Differentiated Services.
RFC 3386 Network Hierarchy and Multilayer Survivability.
RFC 3689 General Requirements for Emergency Telecommunication Service (ETS).
RFC 3690 IP Telephony Requirements for Emergency Telecommunication Service (ETS).

Others

[1] Carrier IP Network Telephony Requirements, Nortel (April 2004).

[2] C. Alaettinoglu and S. Casner, Detailed Analysis of ISIS Routing Protocol on the Qwest Backbone: A recipe for subsecond ISIS convergence, *NANOG 24*, Miami (February 2002).

[3] C. Boutremans, G. Iannaccone and C. Diot (2002), Impact of link failures on VoIP performance, *12th International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV 2002)*, Miami (May 2002).

[4] C.J. Bovy, H.T. Mertodimedjo, G. Hooghiemstra, H.Uijterwaal and P. Van Mieghem, Analysis of End-to-end Delay Measurements in the Internet, *Passive and Active Measurements Workshop (PAM 2002*), Fort Collins (March 2002).

[5] B-Y. Choi, S.B. Moon, Z-L. Zhang, K. Papagiannaki and C. Diot, Analysis of Point-To-Point Packet Delay In an Operational Network, *IEEE INFOCOM*, Hong Kong (March 2004).

[6] B. Fortz, J. Rexford and M. Thorup, Traffic Engineering with Traditional IP Routing Protocols, *IEEE Communications Magazine* (October 2002).

[7] C.J. Fraleigh, S.B. Moon, J.B. Lyles, C.J. Cotton, M. Khan, D.A. Moll, R. Rockell, T. Seely, and C. Diot, Packet-Level Traffic Measurements from the Sprint IP Backbone, *IEEE Network* (December 2003).

[8] C. Fraleigh, F.A. Tobagi and C. Diot, Provisioning IP Backbone Networks to Support Latency Sensitive Traffic, *IEEE INFOCOM*, San Francisco (March 2003).

[9] G. Iannaccone, C-N. Chuah, S. Bhattacharyya and C. Diot, Feasibility of IP Restoration in a Tier-1 Backbone, *IEEE Networks Magazine* (March 2004).

[10] D. R. Kuhn, Sources of Failure in the Public Switched Telephone Network, *IEEE Computer* (April 1997).

[11] C. Labovitz, A. Ahuja, and F. Jahanian, Experimental study of Internet stability and wide-area network failures, *International Symposium on Fault-Tolerant Computing* (June 1999).

[12] C. Labovitz, R. Wattenhofer, S. Venkatachary and A. Ahuja, The Impact of Internet Policy and Topology on Delayed Routing Convergence, *IEEE INFOCOM*, Anchorage (April 2001).

[13] B. Mandeville, S. Sargood and D. Pullin, Optimising IP Network Performance Through Active Measurement, *The Journal of the Communications Network* (September 2002).

[14] A.P. Markopoulou, F.A. Tobagi and M.J. Karam, Assessing the Quality of Voice Communications over Internet Backbones, *IEEE Transactions on Networking* (October 2003).

[15] D. Newman, Core Competency: ISP backbones stand up in grueling 30-day performance test, *Network World* (December 2002).

[16] K. Papagiannaki, R Cruz and C.Diot, Network Performance Monitoring at Small Time Scales, *ACM SIGCOMM Internet Measurement Conference*, Miami (October 2003).

[17] M. Roughan, M. Thorup and Y. Zhang, Traffic Engineering with Estimated Traffic Matrices, *ACM SIGCOMM Internet Measurement Conference*, Miami (October 2003).

1 Network Design Considerations

To provide a context for the discussions in this paper, a managed IP network supporting real-time services is outlined schematically in section 1.1.

The support of real-time services imposes new demands on IP networks. Meeting these demands cost-effectively while at the same time supporting existing data services is especially challenging. The nature of the demands is summarised in section 1.2. Major mechanisms for meeting these demands are identified in section 1.3, with a particular emphasis on the cases in which service provider has to make fundamental choices about network infrastructures. These cases are described at length later in this paper.

1.1 Architectural Framework

On the managed IP network a service provider aims to provide a certain level of quality. Typical managed IP networks include aggregation networks as well as core networks, with a hierarchy such as the following:

- ❒ An aggregation network concentrates traffic from many customers. The links from an aggregation network router to an access network router have low capacity (perhaps 100 Mb/s or 155 Mb/s) relative to the links to a core network router.
- ❒ A core network uses high-capacity links which may themselves be arranged in a hierarchy of two or more levels as follows:
    - ■ A lower-level core network router provides entry points to the core network for aggregation network routers or for access network routers imposing enough demands. Its links to an aggregation network router, an access network router or another lower-level core network router have medium capacity (perhaps 622 Mb/s, 1 Gb/s or 2.5 Gb/s).
    - ■ A higher-level core network router provides the backbone for the core network. Its links to a lower-level core network router or another higher-level core network router have high capacity (perhaps 2.5 Gb/s or 10 Gb/s).

When access networks may connect directly to the core networks instead of passing through aggregation networks, edge router functions may have to be performed by core network routers as well as by aggregation network routers. In particular, DiffServ classification and marking may have to be performed by core network routers as well as by aggregation network routers.

Figure 1 depicts schematically a managed IP network based on a three-level hierarchy and intended to support carrier telephony over IP. In it the carrier-located trunk media gateways and some access networks are connected to lower-level core network routers, whilst the carrier-located line media gateways and other access networks are connected to aggregation network routers. This depiction is for illustrative purposes only; the discussions in this paper do not depend on the adoption of this hierarchy, and, in actuality, the connectivity and degrees of duplication of the nodes depend on the forwarding rates, interface rates and redundancy required.

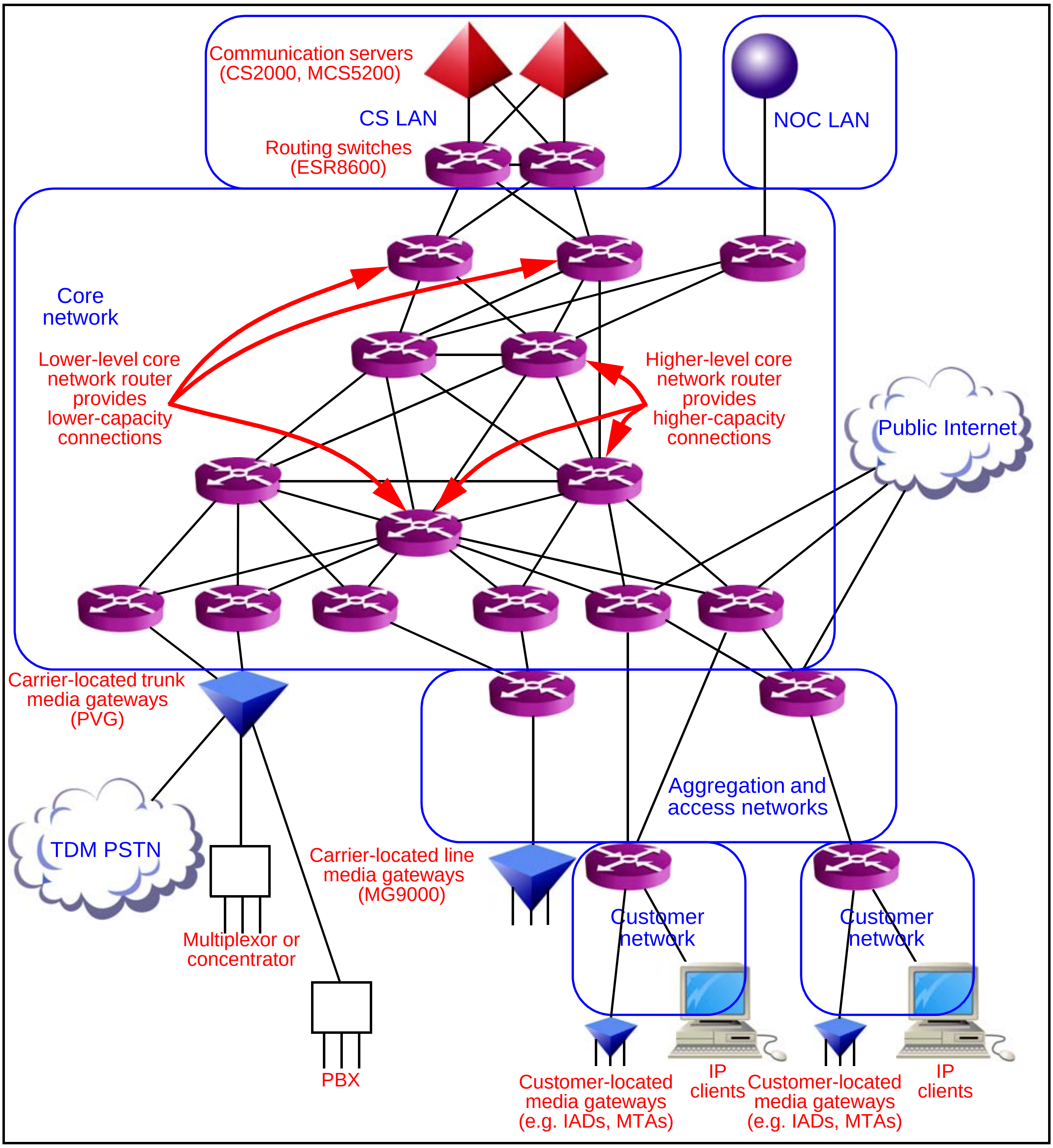


Figure 1 Roles of different types of router in one managed IP network

1.2 Specifying Requirements in Managed IP Networks

IP networks have typically been designed to satisfy the requirements of data services (such as Internet access, content hosting and IP VPNs). These services are principally ones that tolerate delay and packet loss: often they involve no fast user interaction and use TCP to adjust bandwidth consumption and to retransmit in the event of packet loss. The design of these networks has been based on the "best effort" service model and has required no significant extension to TCP or IP, as has been demonstrated by the immense growth in the Internet for data services. All of the traffic is carried across the IP network as rapidly as possible, but without assurances for timeliness or even reliability. Congestion due to inadequate capacity, route changes due to routing updates, and link or node failures all result in poorer response times which users accept to some extent.

Even within the intended range of data services, there are now applications that are not always handled well by the best effort service model. For instance, supply chain management applications may involve enormous numbers of client-server interactions in which users require immediate responses. They may have been deployed over a WAN but designed on the implicit assumption that the interactions occurred over a LAN having ample bandwidth. Although the applications themselves may appear to tolerate delay and packet loss, unless adequate capacity is assured the packet rate reduction or packet retransmission resulting from congestion can be unacceptable to users.

In addition, there are now potential services having requirements that are very different from those of data services. This is so not only for telephony and multimedia, but also for emulation of TDM switched or leased circuits over IP using a 'pseudo-wire'. In particular, for telephony over IP, there is likely to be an objective of achieving PSTN equivalence, in that quality should not be perceptibly worse when traffic is carried over the IP network than when traffic is carried over the TDM PSTN. This objective gives rise to targets (for delay and packet loss, for example) that are more demanding than those needed for data services. (However, the targets are not necessarily more demanding than those attained for data services, as indicated in section 2.1.) Moreover, these new services characteristically use UDP, which does not adjust its bandwidth consumption dynamically in the same way as TCP, so when deployed without proper planning they may not only have poor performance but also inflict poor performance on existing services.

The economic benefits of telephony over IP may be realised most effectively if the IP network supports services besides telephony, because using one network for several services allows costs to be reduced and new service combinations to be created. Moreover, even where core networks need to support only telephony, access networks (and therefore aggregation networks) need to support other services that customers require. Consequently, designing an IP network can entail making services co-exist cost-effectively and yet satisfy very different requirements. The IP network must then in particular satisfy performance requirements (in terms of targets for delay and packet loss) and dependability requirements (in terms of targets for availability) originating with telephony [1]. These requirements can then be incorporated in Service Level Agreements (SLAs), with the performance requirements, in particular, being used to define the Quality of Service (QoS) more rigorously than in many SLAs.

1.3 Satisfying Requirements in Managed IP Networks

An SLA covers only the portions of end-to-end paths that the service provider controls. These portions form the managed IP network. The following techniques, described in this section, should be used to ensure that SLAs can be satisfied by the managed IP network:

❐ Adequate bandwidth provision (to avoid general congestion affecting aggregate traffic flows).

❐ Routing control (to make traffic follow routes having known behaviours).

❐ Rapid traffic restoration (to keep disruption after a failure to acceptable levels).

❐ Admission control (to avoid local congestion affecting individual user sessions).

❐ Equitable bandwidth division (to keep packet delay variation to acceptable levels).

1.3.1 Adequate Bandwidth Provision

A service provider needs to provide enough capacity to satisfy the requirements but not too much capacity. Having excessive capacity results in under-utilisation, whilst having inadequate capacity can result in service degradation for some network users at peak times.

Congestion may occur because of bursts in demand, design problems or faults, even in core and aggregation networks. Traffic that does not tolerate delay or packet loss must be protected from it. Some degree of over-provisioning beyond the immediately apparent demand is therefore necessary for avoiding congestion. There are the following approaches to adequate bandwidth provision:

❐ Bandwidth over-provision irrespective of service differences

In this approach bandwidth is over-provisioned to accommodate the aggregate traffic demands, ignoring the differences in performance requirements between different services. Because real-time traffic requires low bounds on delay and packet loss, all of the traffic on the network must be given low bounds on delay and packet loss by ensuring that there is a low probability of encountering congestion.

When traffic demands are relatively low, this approach is often sufficient to ensure that performance remains satisfactory for all services. It can be justified by noting that the network should be simple to operate and manage, so that savings in operating costs cancel out capital costs incurred by providing excess capacity. However, as the number of customers and the traffic demand increase, over-provisioning irrespective of service differences may become an inefficient way of ensuring that the performance requirements can be satisfied for all services simultaneously. It may be economically viable only when the service provider is at the early stages in network deployment, owns the network infrastructure or has services that occupy the network at completely different times; even if these conditions are satisfied in core and aggregation networks they may not be satisfied in access networks, where multiple users of multiple services may share low bandwidth links.

In essence, over-provisioning irrespective of service differences ignores the savings that are made possible by multiplexing one physical network between several services simultaneously and by aiming to satisfy different performance requirements for different services.

❒ Bandwidth over-provision according to service differences

In this approach, traffic is classified, marked and treated on the basis of its performance requirements and on the basis of its importance for users or for the operation of the IP network. Different traffic classes use network capacity according to their relative priorities, so their different performance requirements are reflected in different bounds on delay and packet loss.

Doing this provides in particular a way to differentiate telephony and multimedia traffic from data traffic: real-time traffic can be given priority scheduling in routers, while traffic with a lower sensitivity to delay or packet loss can be delayed by degrees and discarded preferentially. Traffic that is sensitive to delay and packet loss but that has a predictable and bounded loading profile can thereby retain high performance, even as the network loading increases from other sources.

In fact IP networks have for some time used packet marking to differentiate network control traffic (such as routing protocol ‘keep alive’ messages and routing database update messages) from application traffic. This usage has been extended and adapted in the main standard for IP traffic differentiation, the IETF Differentiated Services (DiffServ) architecture, which is described in section 4.1. The DiffServ architecture allows different performance requirements to be satisfied in different ways, but it does not distinguish between services according to their dependability requirements: all of the types of traffic suffer equally from failures and are rerouted in the same ways after failures.

The choice between these approaches for a particular network will reflect the factors that are regarded as most important by the service provider. Essentially, the aim is to strike a balance between simplicity and under-utilisation of network bandwidth on the one hand, and increased complexity and improved efficiency on the other. For a service provider that owns the infrastructure, for example, additional bandwidth may be cheaper than it is for a service provider that leases the infrastructure, and the under-utilisation of some links may not be regarded as significant. Different service providers may also have different views of the operating costs and bandwidth savings due to differentiated bandwidth allocation, especially as they will have different views of the relative volumes and values of different services.

A quantitative comparison between the two approaches is provided in section 2.2.

1.3.2 Routing Control

When services are distinguished from one another, the network can be regarded as a set of virtual service networks supported on a common infrastructure. Doing this can achieve great flexibility in overall administration, service provisioning and configuration of the network, as it allows engineering tasks to be performed separately for the different services to which they refer. In particular, it allows telephony or multimedia traffic to be isolated from other traffic, and optionally it also separates different types of telephony or multimedia traffic with different performance and dependability requirements.

Packet marking is used to distinguish between services in order to ensure adequate bandwidth provision and maintain performance; explicit routing is used to distinguish between services in order to facilitate routing control and maintain dependability. Routing control is one aspect of ‘traffic engineering’; this term is sometimes used just for routing control and sometimes extended to cover, among other things, nodal traffic control such as is performed by DiffServ. (Also, the term ‘traffic management’ is sometimes used for nodal traffic control.)

The purpose of routing control is to make traffic follow routes having known behaviours. In particular, it is intended to ensure that routes do not over-utilise or under-utilise particular links and routers and that routes can be made available for traffic restoration after a failure. In the managed IP network different routes may sometimes be provided for different types of traffic but not for different user end points. (The routes are determined for aggregate traffic flows, not session-by-session, to avoid scaling problems.) There are the following approaches to routing control:

❒ Native IP routing

Link state routing protocols used within Autonomous Systems (ASs), like Open Shortest Path First (OSPF) and Interior System to Interior System (IS-IS), calculate the shortest paths for the traffic to end points based on weights assigned to each of the links. Load balancing can be used to distribute traffic equally between multiple routes having the same lowest weight; it thereby alleviates the link congestion that can be created by shortest path routing.

In this case routing control entails primarily adjusting the weights assigned to different links; when the adjustments are performed carefully with off-line tools the resulting traffic flows can be nearly optimal, as indicated in section 5.1. The routing can also take into account some degree of traffic differentiation.

Border Gateway protocol (BGP) is used to exchange routing information about peers between and within ASs.

❒ Connection-oriented link layer switching

Routes can be controlled explicitly by the administrator and moved away from shortest paths that would otherwise be selected by native IP routing using default weights. Doing this can allow network resources to be used more efficiently, by constraining the routing to take account of link attributes such as utilisation. Currently constraint-based routes are preconfigured, using off-line tools, because constraint-based routing can be complicated to use and slow to operate.

The use of a connection-oriented link layer can allow different routing policies to be enforced for different traffic types, so virtual service networks can use both packet marking and explicit routing, as indicated in section 4.3. The Virtual Circuits (VCs) of a connection oriented link layer, such as the Label Switched Paths (LSPs) of Multi-Protocol Label Switching (MPLS), can be dedicated to the use of specified traffic types and can be integrated with traffic differentiation schemes such as DiffServ. Because MPLS is currently expected to be the predominant connection-oriented link layer for IP, in this paper it is discussed more extensively than other possibilities. In particular its application to routing control and rapid traffic restoration is described in section 5.2.

Extensions of these mechanisms for routing control within ASs to routing control between ASs have not yet been fixed in specifications.

Different service providers make different decisions about whether to deploy a connection-oriented link layer for routing control. An approach to routing control using features extending MPLS can be valuable for large networks where telephony or multimedia traffic is present with other traffic. Recent deployments of MPLS for IP VPNs as described in RFC 2547 should provide valuable experience in design and operations. (Other IP VPNs, discussed in RFC 2764, do not require MPLS.) As a network grows, the benefits associated with MPLS may provide increasing justification for the additional operational investment required. In smaller networks, approaches that do not make use of MPLS may be appropriate.

1.3.3 Rapid Traffic Restoration

Although services using TCP, which provides retransmission, tolerate a lengthy loss of traffic, those using UDP do not. After a failure traffic is lost for a time that may be unsatisfactory for real-time services. Failures of links (involving interface cards, transmission layer network elements or fibres that may be cut) and nodes (involving control planes) are especially important. For this purpose, taking a node out of service for scheduled or unscheduled maintenance should be regarded as inducing a node failure if it has a similar disruptive effect on the network.

Many node failures can be confined within individual nodes by techniques like ‘non-stop routing’, which uses backup router components to maintain all pertinent state information and maintain adjacencies with surrounding routers. (However, techniques such as ‘non-stop forwarding’ that are based on routing protocol extensions do not confine failures in this way.) Other node failures, and all link failures, require that traffic be sent to and received from routes that avoid the points of failure.

The time during which traffic is lost as a result of a failure depends on the time taken to detect the failure and the time taken to recover after detection of the failure. The time taken to detect the failure can be reduced by detecting failures in the transmission layer instead of the IP layer when possible or by reducing the time interval between ‘keep alive’ messages. The time taken to recover after detection of the failure can be reduced by rapid rerouting of the traffic.

In practice rapid rerouting may involve two phases, in which first traffic near the point of failure is switched to predetermined alternative paths and then all traffic affected by the failure is switched to routes that are determined slowly but that make better use of network resources. Thus routing control can assist in making good choices of alternative paths.

As with routing control, there are the following approaches to rapid traffic restoration:

❐ Native IP routing

Shortest path routing results in traffic flows that compete for the bandwidth on a restricted set of paths and suffer the same fate when link and node failures occur. Load balancing can be applied at nodes where there is enough diversity in the network topology to broaden the set of paths and ensure that alternative paths are followed after a failure.

For link and node failures alike, the time taken to recover is affected by the time taken for routing protocol reconvergence. (Conventionally the rerouting of traffic onto an alternative path requires routing updates to be propagated throughout the network before shortest paths are calculated.) Careful network design can make this time fairly small (perhaps 1 second - 2 seconds) but not quite small enough to meet the most demanding dependability requirements of real-time services.

This use of native IP routing has worked very well for data services. It is discussed in section 5.1, along with an extension to ensure fast enough recovery for real-time services. This extension introduces the two phases mentioned above by supporting predetermined alternative paths that are used to effect local repairs while routing protocol reconvergence occurs.

❐ Connection-oriented link layer switching

Connection-oriented link layers can achieve fast rerouting after a link or node failure by moving traffic to predetermined alternative paths near the point of failure.

Once this local repair has been effected, routes that are closer to being optimal can be determined and established. These routes can be constrained to take account of link attributes such as utilisation.

The routes used in these repairs are typically preconfigured with the aid of off-line tools, but in principle they can be created dynamically by signalling. The mechanisms are described in section 5.2.

In general terms, routing control allows high value traffic to use routes with enough bandwidth and with high availability, while low value traffic competes for the remaining bandwidth. If a connection-oriented link layer is used, its use may be confined to the high value traffic by applying rules like those in section 4.3, so that the associated operating costs are not incurred for low value traffic. The decision about where and when to deploy a connection-oriented link layer for rapid traffic restoration will be closely aligned with the corresponding decision about routing control.

1.3.4 Admission Control

In the TDM PSTN there are few types of traffic and the demand for each type in normal operating conditions is known with a fairly small margin of error. The PSTN can therefore be designed cost-effectively, to achieve a given Grade of Service (GoS) (which is the likelihood of being unable to set up a call in normal operating conditions), by relying on call admission control to prevent unintended congestion. When the demands for capacity are excessive, the quality of calls in progress is maintained but new calls are not admitted.

When the demands are excessive in an IP network, higher layer protocols largely control the impact on services. However, real-time services do not usually respond to congestion: they do not reduce bandwidth consumption or retransmit lost packets, so performance deteriorates not only for them but also possibly for other services. In particular, if there is no call admission control, the quality is impaired both for new calls and for existing calls.

As real-time traffic flows become useless if too many packets are lost, a deployment of DiffServ may well try to ensure that packets in such flows are never discarded. (Even for non-real-time flows, for which packet discard leads to retransmission, a user may prefer to have no application session than to have one that is inadequately responsive.) Consequently a deployment of DiffServ may well arrange that capacity is used by services according to their performance requirements but may still not eliminate the potential for excessive demands for capacity. There may need to be some way of ensuring that entire traffic flows, rather than individual packets, are accepted or rejected. Moreover the acceptance or rejection of these traffic flows may need to be co-ordinated with operations on other flows in the same sessions; for instance, signalling messages may need to be sent when media traffic flows are blocked.

In the IP layer of a network there is no way of performing this co-ordination; the application layer (in the form of call processing for telephony or multimedia, for example) must be involved. As the application layer should admit traffic to the network only when there is capacity available, it must have some awareness of limits on capacity or some interactions with the IP layer or the link layer to ensure that limits are not exceeded.

A session (representing one occasion on which a service is used) can generate multiple media traffic flows. However, admission control techniques that are applied session-by-session to accept or reject sessions, rather than individual traffic flows within sessions, may be wanted by users; they can be viable if the numbers of sessions allowed are adjusted when there are changes in the proportions of media (or ‘user’ or ‘bearer’), signalling (or ‘control’) and management traffic for a service.

1.3.5 Equitable Bandwidth Division

In a core network, having high bandwidth links, a packet undergoes an insignificant delay when it enters a link. Accordingly a high priority packet is not significantly delayed while waiting for one low priority packet to enter the link. (It would be significantly delayed if many such packets entered the link before it, so it is scheduled ahead of them.)

However, when there are low bandwidth links (as in an access network using DSL, for example), the delay introduced when a packet enters a link from the head of the queue becomes important. (The delay is proportional to the packet size.) Giving certain packets higher priorities than others does not necessarily keep their delays low, as high priority packets may be delayed significantly while waiting for just one large low priority packet to enter a link in front of them. Either the transmission of the low priority packet must be interrupted and later resumed or the low priority packet must be small enough that the delay which it introduces can be tolerated by the waiting high priority packets.

If IP packets are kept small by keeping the end-to-end Maximum Transmission Unit (MTU) small, the load on the network, as determined by the packets transmitted, becomes large. If IP packets are fragmented at their sources, they are not always reconstituted correctly when fragments from multiple sources pass through a Network Address and Port translation (NAPT) device. (Such a device is very likely to be present at the edge of an access network.) In fact link layer mechanisms (such as those of Multi-class Multi-link PPP, Frame Relay and ATM) should be used over low bandwidth links to ensure that bandwidth is divided according to the bit rates required by the services, not just according to the packet rates.

2 Current Performance Levels

If PSTN equivalence is to be achieved by IP networks then certain performance requirements must be satisfied. The evidence presented in section 2.1 suggests that they can satisfy these requirements already if they are designed suitably.

Satisfying these requirements entails over-provisioning bandwidth. However, the ways of over-provisioning bandwidth adopted for current data services may not always be cost-effective when telephony and multimedia services are introduced. How they can be made more cost-effective is considered in section 2.2.

2.1 Internet Service Provider Core Network Performance

Achieving equivalent performance to that of the PSTN entails having an IP network that offers low delay (and implicitly low delay variation) and low packet loss. Various studies demonstrate that Internet Service Provider (ISP) core networks can be designed to achieve this. The studies summarised in this section considered the following:

- ❒ Routes within seven ISP networks (reporting on the delay and packet loss).
- ❒ Routes within a tier 1 ISP network (reporting on the delay and packet loss and on the suitability for carrying voice).
- ❒ Routes within seven ISP networks (reporting on the delay and packet loss and on the suitability for carrying voice).
- ❒ Routes between several ISP networks (reporting on the delay).

These studies use performance measurements that do not differentiate between different types of traffic. Nonetheless, they are sufficient to demonstrate that the performance requirements of telephony over IP can be satisfied in core networks provided that suitable designs are adopted. In particular, an important factor in achieving high performance is provisioning enough bandwidth to avoid congestion.

2.1.1 Routes within Seven ISP Networks

The study [15] assessed the performance of seven ISP networks in North America over 30 days. It used active performance measurements to sample delay and packet loss. The measurement devices continuously injected 1518 and 256 byte packets at four Points of Presence (PoPs) in each network; at each PoP the bit injection rate was kept below 512 Kb/s (but the packet injection rate was typically 66 p/s on average).

The data established that the fixed part of the delay was essentially due to propagation, at least on the one network for which absolute delay could be measured accurately. The variable part of the delay could rise to 200 ms for some networks. For all the networks packet loss was low.

Some caution must be exercised when using tests based on active performance measurements (of packets injected into the network), for the following reasons:

- ❐ The tests must use low proportions of the traffic (less than 1%) so that they do not influence the live traffic.
- ❐ Some tests may use low packet injection rates. However, transient performance degradation in IP networks arises primarily from congestion in which bursts of packets are lost; it occurs over intervals (10 ms - 100 ms) that can be too small to be observed consistently at low packet injection rates.
- ❐ The measurements resulting from the tests are then averaged over SNMP polling intervals (typically 5 minutes), and these average values are often averaged again over about 30 days to form the basis of SLAs. However passive performance measurements using subsecond sampling intervals can highlight transient fluctuations in the load, for which the link utilisation is much higher than the average over 5 minute intervals; packet loss may not always occur but packets can experience delays of some ms.

2.1.2 Routes within a Tier 1 ISP Network

One study [7] assessed the performance of a tier 1 ISP network in North America over three days. It used passive performance measurements of the departure and arrival times of packets within and between four PoPs along 30 links.

The data established that the fixed part of the delay was essentially due to propagation. The variable part of the delay was small (in comparison with the fixed part of the delay) in general, but some packets experienced delays of 100 ms due to router anomalies or routing changes. Packet loss was low.

Another study [3] assessed the performance of the same network over three days. It used active performance measurements to sample delay and packet loss; a packet was injected every 20 ms. The measurement devices continuously injected 200 byte packets at each of two PoPs. The study confirmed the results obtained by passive performance measurements (with a maximum delay of 33 ms for a packet loss ratio of 1×10^{-3} during periods when the route was regarded as available).

This study also reached conclusions about the capability of the network to support telephony. The voice quality to be expected was inferred by using the performance data as inputs to the E-model described in G.107 to derive a value for R. The inferred voice quality was found to have an average value for R greater than 90, with some transient dips below 80. (However, the effects of media gateways and IP or TDM access networks were not taken into account fully.) Packet loss was found to occur predominantly as single random events and could therefore be mitigated by introducing an algorithm for packet loss concealment.

2.1.3 Routes within Seven ISP Networks

The study [14] assessed the performance of seven ISP networks in North America over seventeen days. It used active performance measurements to sample delay and packet loss. The measurement devices continuously injected 50 byte packets into the networks, with measurements along 43 routes through the networks; for three days a packet was injected every 10 ms and for fourteen days a packet was injected every 100 ms.

The data established that the fixed part of the delay was essentially due to propagation, except on rather indirect routes. The variable part of the delay was small (in comparison with the fixed part of the delay) for some networks but could rise to ten times the fixed part of the delay for other networks. (Telephony over IP packets incurring such large delays would be lost, so the packet loss ratio of 2×10^{-3} that was observed on the worse routes would increase.) The spikes in delay had different characteristic patterns on different routes. Almost all of the routes suffered some packet loss, with intervals of packet loss ranging between 10 ms and 167 seconds; the longer intervals appeared to coincide with routing changes that could be ascribed to failures rather than to congestion.

The study also reached conclusions about the capability of the networks to support telephony. The voice quality to be expected was inferred by using the performance data as inputs to the E-model described in G.107 to derive a value for R and a representation on the Mean Opinion Score (MOS) scale. The implication was that some networks should already be able to support high quality voice but that others exhibited characteristics leading to periods of low quality, such as large spikes in delay, periodic patterns of delay, failures, and simultaneous packet loss on many routes.

2.1.4 Routes between Several ISP Networks

The study [4] assessed the performance of several ISP networks in Europe and elsewhere over 1 day. It used active performance measurements to sample delay. The measurement devices continuously injected 100 byte packets into the networks, with measurements along 963 routes through the networks; a packet was injected every 40 seconds.

The data established that the fixed part of the delay was essentially due to propagation, except on rather indirect routes. (There were some very indirect routes, having lengths that were many times the line of sight distance between the end points.) The minimum observed delay was often within a factor of 2 of the minimum calculated delay derived from the propagation delay and the number of routers.

The study found various distributions of delay, but most were gamma distributions and others were distributions with multiple peaks; multiple peaks are found in other networks, where they have been ascribed to different delays along different paths between the same end points [5]. The gamma distributions typically found have a high proportion of packets experiencing delays near the minimum observed delay and a low proportion of packets in an extended tail [13].

2.2 Bandwidth Engineering

2.2.1 Adequate Bandwidth Provision

As discussed in section 1.3, for ensuring that performance requirements are satisfied there are two approaches to bandwidth provision. Both approaches determine the transmission capacity required after over-provisioning by dividing the demand by an intended link utilisation. The demand could be estimated from network measurements, which might provide the mean bandwidth required at peak times, ideally measured using subsecond sampling intervals; in this situation the intended link utilisation would be selected to accommodate the forecast growth in demand and to ensure that the delay was low enough for a high enough proportion of the traffic (in other words, that the maximum delay and the packet loss ratio would not be greater than their targets). As noted in 2.1, measurements over coarse sampling intervals need to be treated with care; for instance, a

155 Mb/s link that appears 50% full over 5 minutes may actually be 70% full over intervals of 100 ms and even fuller over intervals of 1 ms and 10 ms [16].

Here the approaches to bandwidth provision are compared by means of an example thus:

❒ Bandwidth over-provision irrespective of service differences

The simplest method of providing enough bandwidth is to treat all of the traffic in the same way, making no distinction between the performance requirements of different types of traffic. In fact there is even no distinction between the dependability requirements of different types of traffic: all of the types of traffic suffer equally from failures and must be rerouted over the same alternative paths in the event of failures.

For the example, the aggregate demand is 1.9 Gb/s. The aggregate intended link utilisation, dictated by expectations about the most rapidly growing and most demanding services, is 40%. The transmission capacity required is then 4.7 Gb/s. In an SDH network using Virtual Containers (VCs), this transmission capacity requires two 2.5 Gb/s VCs, so the total capacity provided is actually 5.0 Gb/s and the actual link utilisation is 38%. If the VCs use SDH 1:1 automatic protection switching then four 2.5 Gb/s VCs are needed.

❒ Bandwidth over-provision according to service differences

When bandwidth over-provisioning exploits traffic differentiation, transmission capacity requirements are calculated using a different intended link utilisation for each traffic class (with more demanding classes having lower intended link utilisations). In the routers different traffic classes are associated with different outbound interface queues that can influence the relative proportions of the traffic classes on individual links.

For the example, the traffic is taken to fall into three classes, having the following demands and intended link utilisations:

- 0.4 Gb/s and 40% for high value real-time traffic.
- 0.3 Gb/s and 60% for high value non-real-time traffic.
- 1.2 Gb/s and 80% for low value traffic.

The aggregate demand remains 1.9 Gb/s but the mean intended link utilisation is 63% instead of 40% and the transmission capacity required is 3.0 Gb/s instead of 4.7 Gb/s. (Network control traffic is ignored in this example.)

In a refinement of the technique, the transmission capacity is reduced further by taking account of the differing dependability requirements of the traffic: the SLAs for the classes may imply that after a failure the high value traffic must be carried but the low value traffic may be discarded. If the traffic is distributed over two equal load-balanced routes in normal conditions, then each route must have a transmission capacity of at least (0.4 / 40% + 0.3 / 60%) Gb/s or 1.5 Gb/s, to accommodate all of the high value traffic after a failure. (The two routes together also accommodate all of the traffic in normal conditions.)

However, if the transmission capacity for each route is just 1.5 Gb/s then after a failure some of the high value traffic may be delayed or lost, because of transient fluctuations in the load. To avoid this, each route can be designed so that only 80% of the bandwidth is allocated to the high value traffic, in which case each route must have a transmission capacity of 1.9 Gb/s. This could be done by associating the traffic classes with the outbound interface queues thus:

- Strict priority queuing for high value real-time traffic.
- 57% weighting of the transmission capacity not allocated to real-time traffic for high value non-real-time traffic.
- 43% weighting of the transmission capacity not allocated to real-time traffic for low value traffic.

For these queue weights, if one route is to accommodate all the high value traffic after a failure it must have transmission capacity (0.4 / 40% + 0.3 / 60% / 57%) Gb/s or 1.9 Gb/s. (The division by the queue weights occurs because high value non-real-time traffic may get no more than 57% of the transmission capacity left over after allocating transmission capacity to high value real time traffic.)

Hence the traffic needs to occupy no more than two 2.5 Gb/s VCs, even when it is protected by diverse routing. (However, the diverse routing may not provide the same rapidity of response to failures as SDH automatic protection switching, if it requires routing protocol convergence before alternative paths become available.) In fact if 2.5 Gb/s VCs are used, perhaps 52% of the low value traffic will not be lost even after a failure.

Clearly bandwidth over-provision according to service differences requires less bandwidth than bandwidth over-provision irrespective of service differences (especially if high value traffic can displace low value traffic after a failure). It also allows any trade-off to be clearly identified. In the example, all of the high value traffic and 52% of the low value traffic could satisfy the performance requirements without provisioning a second SDH 2.5 Gb/s VC; the service provider could judge what would be an acceptable level of under-provisioning for low value traffic and what would justify the cost of a second SDH 2.5 Gb/s VC when high value traffic reached its maximum level or dependability requirements were established.

2.2.2 Link Utilisation

Modelling [8] suggests that fairly high link utilisations are possible on high bandwidth links, in which case the benefit of exploiting traffic differentiation is reduced, at least in core networks. The results of the modelling include approximately the following (for a packet loss ratio of 1×10^{-3}):

❒ If the maximum link delay is 1 ms and the link bandwidth is *b*, the allowed link utilisation rises from 25% to 60% proportionally to log(*b*) as *b* rises from 100 Mb/s to 1 Gb/s and flattens off when *b* > 1 Gb/s.

❒ If the maximum link delay is 10 ms and the link bandwidth is *b*, the allowed link utilisation rises from 25% to 80% proportionally to log(*b*) as *b* rises from 10 Mb/s to 155 Mb/s and flattens off when *b* > 155 Mb/s.

These results can be quite sensitive to the maximum delay (but less sensitive to the packet loss ratio); furthermore, they should not be extrapolated to lower bandwidths. They may provide ways of ensuring that the delay is low enough for a high enough proportion of the traffic. However, they do not accommodate any forecast growth in demand; in practice they need to be reduced to an extent dictated by the provisioning times of the service provider.

Figures for the link delays should be extended to the end-to-end delays by convolving the link delay distributions, not by adding the maximum link delays. (Doing this establishes that in some networks an end-to-end delay of 4 ms, discounting propagation delay, can be achieved with link utilisations of 80%.) A service provider without suitable forecasting and modelling capabilities would choose more conservative values than those indicated above. There is then still a case for some traffic differentiation, at least to the extent of allowing extra traffic, which does not have bandwidth assurances, whenever the bandwidth is otherwise unoccupied.

2.2.3 Design Maintenance

The relative proportions of applications and of protocols in an IP network change; a recent example of this is the striking growth in streaming (using UDP) and peer-to-peer content distribution (using TCP), which can now account for 80% of the traffic on some Internet links [7]. Different protocols have different effects on the network, especially in the event of congestion. Bandwidth provisioning must therefore be combined with network monitoring to ensure that performance is maintained for all applications. More generally, ensuring that the network continues to have adequate capacity as the services develop needs the following:

❒ Monitoring and measurement

The known current state of the network should include at least the network routing, the transmission capacity of each of the links and the operational status of each of the nodes and links. The actual link utilisations can be estimated from network measurements. The traffic matrices indicating the demands for routes across the network can be estimated from these same measurements and the network routing, for use in network planning and routing adjustment.

Network monitoring for performance assessment should extend to all services having SLAs. Performance metrics may be collected for each traffic class where traffic differentiation is use; they may also be collected for each route and, indeed, for each Virtual Circuit (VC), if a connection-oriented link layer supplements IP routing. This data may be correlated with other data from Management Information Bases (MIBs), such as actual node utilisations, and Remote MONitoring (RMON), such as actual link utilisations and packet forwarding rates.

Different performance metrics may be monitored for the different services or applications and require correlation with IP network performance measurements, particularly if degradation occurs. For example, from the user perspective, relevant telephony over IP measurements include call setup delay, call rejection ratio, voice quality, response times to train modems, and transmission times for standard fax pages; from the IP network perspective, one-way end-to-end performance measurements include packet delay, packet delay variation, packet loss and packet mis-sequencing.

❒ Forecasting and modelling

Traffic forecasting, with detailed trend analysis using historical data when possible, may be used in conjunction with the traffic matrices to establish likely future demands for routes and ensure that links have enough transmission capacity at the right time.

Traffic flows may be modelled using design tools specific to the routing protocol; to assist with this, routing tables can often be imported directly from the routers. The routes determined thereby can depend on link weights and, for some protocols, additional constraints such as link utilisation. Every constraint increases the complexity of the optimisation; its benefits and effect on the stability and scalability of the network design must be considered carefully.

- ❐ Configuration and control

  The network routing may need to be adjusted from time to time by modifying link weights and additional constraints. (This may be needed, for example, after a change in peering points, according to the hour of the day, or to accommodate a large enterprise customer.) The frequency of the adjustments should be minimised, subject to the need to ensure that performance requirements are always satisfied for all of the traffic types.

3 Current Dependability Levels

If PSTN equivalence is to be achieved by IP networks then certain dependability requirements must be satisfied. Several studies of commercial IP networks and the Internet are presented in section 3.1. Conclusions about the dependability of the PSTN are reviewed in section 3.2 to permit a comparison between the PSTN and IP networks.

There is in fact quite a substantial difference in dependability between the PSTN and many existing IP networks. However, recent developments in routers and routing control mechanisms can be used to remove the difference. These are outlined in section 3.3.

3.1 Internet Service Provider Core Network Dependability

Several studies implicitly indicate the extent to which IP networks based on current routers and routing control mechanisms satisfy dependability requirements equivalent to those of the PSTN. The studies summarised in this section considered the following:

❐ Routes within seven ISP networks (reporting on the availability of routes).

❐ Routes within a tier 1 ISP network (reporting on the availability of routes and on traffic restoration in the IP layer).

❐ Routes within a tier 2 ISP network (reporting on the availability of routes and on network outage root causes).

❐ Routes between three ISP networks (reporting on the availability of routes).

These studies illustrate that currently well designed networks in individual ASs can satisfy dependability targets which approach but do not reach those of the PSTN. However, the public Internet does not satisfy such demanding dependability targets.

3.1.1 Routes within Seven ISP Networks

The study [15] of seven ISP networks in North America, referred to in section 2.1, also considered dependability (in particular, the availability of routes between PoPs). For the dependability measurements, a supervision timer was started whenever measurement packet loss between PoPs was observed, and an outage duration timer was started if no further packets were received from the network within 10 seconds. The outage duration timer continued running until the network correctly delivered packets for a period of 10 seconds, at which point the duration of the outage was recorded and the timer was reset to zero. (All this corresponds with the definition of a reportable outage based on the notion of a Severely Errored Second in G.821 for 64 Kb/s TDM circuits.) At the end of the 30 day test period, the total number of outages and the total duration of outages were calculated for each network and used as an indication of availability. To prevent scheduled maintenance downtime from contributing to outage figures, outages recorded during maintenance windows were excluded from the totals.

The study concluded that four of the networks could achieve 99.999% availability, two could achieve 99.99% availability and one could achieve 99.96% availability. (In the last of these cases, certain transmission layer outages were corrected and discounted.)

Some caution must be exercised when interpreting the conclusions, for the following reasons:

- ❐ The measurements were taken only between four PoPs and for 30 days; they therefore provided a small sample relative to the network sizes and only a snapshot in time for assessing network element downtime.
- ❐ The recorded outages (the measured packet flow interruptions between PoPs exceeding 10 seconds duration) did not take into account the number of subscribers impacted, which is important to dependability.
- ❐ The measurements did not record outages of less than 10 seconds. Outages of less than 10 seconds are acceptable for some data services: user expectations are lower than for telephony, and standard TCP mechanisms ensure that packets lost during the outage can be retransmitted fast enough for minimal service interruption to be perceived. However, for telephony, some seconds of continuous outage on a single 2.5 Gb/s link could entail the loss of hundreds of voice packets from tens of thousands of voice channels, with the likelihood that many users would regard calls as terminated and redial.

3.1.2 Routes within a Tier 1 ISP Network

The study [9] of a tier 1 ISP network in North America provides insight into the ability of native IP routing control mechanisms to achieve dependability targets equivalent to those of the PSTN. In this network, routing control uses IS-IS with load balancing based on Equal Cost Multi-Path (ECMP) routing and relies on routing protocol reconvergence for rerouting traffic flows after a failure. Routing updates were monitored network wide by a dedicated passive listening router over about 5 months. The updates were timestamped for both duration and time of occurrence, and were classified as scheduled or unscheduled. The distributions of the durations and the times between events gave indications of the root causes of the failures. Both link failures and node failures were identified.

Failures were found to be widely scattered over weeks, days and hours and also over different links. Up to 50% of the failures also corresponded with maintenance windows. 80% of the failures were single isolated events lasting less than 10 minutes. Longer lived failures were often correlated over multiple links. The chief findings were the following:

- ❐ 46% of failures lasted less than 1 minute and were likely to be due to an overloaded router control plane or a faulty optical interface, both of which would cause a router to miss ‘keep alive’ messages and to declare that an interface had gone down.
- ❐ 40% of failures lasted 1 minute - 15 minutes and were likely to be software related (with router restarts or interface resets, for example).
- ❐ 4% of failures lasted 15 minutes - 45 minutes and were likely to have required some human intervention with substitution or maintenance of hardware equipment.
- ❐ 10% of failures lasted more than 45 minutes and were likely to be hardware failures (of optical fibres or interface cards, for example) that required human intervention.

The time taken to recover after detection of a failure was found to be dominated by the timers used during routing protocol reconvergence. Originally these resulted in a time of 8 seconds - 9 seconds, but fine tuning them gave subsecond times without loss of network stability. Accordingly routing protocol reconvergence alone may be rapid enough for rerouting traffic around link failures where transmission layer alarms are used to detect failures and timers are finely tuned.

3.1.3 Routes within a Tier 2 ISP Network

The study [11] of a tier 2 ISP network in North America examined causes of network failure. The study found that many failures had similar ranges of hardware and software problems to those for the PSTN (so they were not unique to the routing infrastructure of the Internet), at least for the trouble-ticket categories used by that particular network. Among these failures the main causes were as follows:

- 16% were due to scheduled and unscheduled maintenance.
- 16% were due to power failures in customer sites or PSTN facilities housing the ISP routers.
- 15% were due to fibre or carrier failures.
- 13% were due to unreachable or intermittent failures probably arising from, or experienced by, other service providers.

Most failures were associated with customer sites, not core network PoPs.

Some other conclusions were as follows:

- For interfaces between two core network routers, 5% of interfaces had a Mean Time Between Failures (MTBF) of at most 5 days, 40% of interfaces had an MTBF of at most 40 days and 75% of interfaces had an MTBF of at most 60 days. For interfaces between a core network router and a customer network router, the corresponding proportions were 2%, 70% and 80%. The difference between these figures reflects the fact that core network routers are more closely monitored, more completely protected, and housed in more specialised facilities.
- For interfaces between two core network routers, 20% of interfaces had a Mean Time To Recover (MTTR) of at most 20 minutes and 80% of interfaces had an MTTR of at most 120 minutes. For interfaces between a core network router and a customer network router, the corresponding proportions were 15% and 70%. The remainder typically requiring extended problem diagnosis or hardware replacement.
- Core network routers averaged greater than 99.8% availability, corresponding to about 15 hours downtime over a year.

3.1.4 Routes between Three ISP Networks

The study [11] of three ISP networks in North America examined Internet stability and the impact of router unavailability on the reachability of destinations involving routing between ASs of peering ISP networks, where BGP is used. In this case the impact of a router becoming unavailable is that all BGP peers of the failed router delete routes learned from the failed router. Up to 15 minutes are required for Internet routing to complete reconvergence after such a failure. The study was based on nine months of default-free BGP routing data.

The main conclusions were as follows:

- ❒ The Internet had significantly less availability than the PSTN. Route availability was less than 99.99% for 70% of Internet routes and less than 95% for 10% of Internet routes.
- ❒ 75% of Internet routes between ASs had an MTBF of at most 30 days (in that a path providing the route failed in that time). Failure of a single route does not necessarily means loss of connectivity between networks, as most ISP networks maintain multiple redundant connections with other ISP networks, and routers dynamically reroute around faults to bring about failover. Failover occurred within two days for almost all of the routes with redundant connectivity and within five days for more than 80% of the routes.
- ❒ 60% of Internet routes had an MTTR of at most 20 minutes (in that a path providing the route could be selected in that time after a failure). The remainder typically requiring extended problem diagnosis or hardware replacement. (As default-free routes announced by each ISP include routes passing through other ISP networks, the MTTR covers both the time for fault resolution and the time for routing information to propagate through the Internet.)
- ❒ BGP routing instabilities had patterns that recur daily and weekly and that correlate them with peak network usage. (IGP routing instabilities had no such patterns.) This suggests that they stemmed from congestion collapse, just as congestion collapse could cause outages in the PSTN.
- ❒ Some Internet routes contributed disproportionately to unavailability. Because of this 40% of routes exhibited multiple failures of paths lasting between one hour and several days.

3.2 PSTN Availability Analysis

Since 1992, telecommunication service providers in the United States have been required to notify the Federal Communications Commission about outages (or failures of service) lasting more than 30 minutes and affecting more than 30000 customers. A study [10] analysed the outages reported from 1992 to 1994. There were 303 such outages, which together lasted 16000 hours and lost 17 billion customer minutes. (For each outage the lost minutes were calculated by multiplying the duration of the outage by the number of customers affected.) Table 1 summarises the main findings.

Cause of outage	Proportion of outages	Average duration (minutes)	Number of customers affected	Proportion of downtime
Overload	6%	1100	280000	44%
Acts of nature	11%	830	160000	18%
Human error by operator staff (e.g. card removal)	25%	150	180000	14%
Human error by other people (e.g. cable cut)	24%	350	80000	14%
Hardware failure	19%	160	95000	7%
Software failure	14%	110	120000	2%
Vandalism	1%	450	85000	1%

Table 1 Causes and effects of reported PSTN outages (1992 - 1994)

Particular points of interest for comparison with ISP network outages are the following:

- Overloads contributed 6% of outages and 44% of downtime. In fact overloads can be regarded as expected outages, because service providers must balance technical and economic factors when determining the numbers of circuits provided.
- Hardware related errors and software related errors had roughly equal responsibility for failures due to human error by operator staff. Hardware related errors (including those arising in cable maintenance, power supply maintenance and power monitoring) accounted for about 15% of outages and 7% of downtime, and software related errors (including those due to version mismatches, incorrect data entries, and procedural mistakes during upgrades) accounted for 10% of outages and 7% of downtime.
- Hardware failures and software failures had effects that were kept short, probably by using extensive error detection and correction mechanisms, with human intervention where necessary.

A critical aspect of service dependability that cannot be quantified by downtime measurements alone is the impact on customers, as this can vary greatly depending on the locations, durations and types of failure. Figure 2 illustrates the overall impact of these types of outage on customers, in terms of durations and number of customer affected.

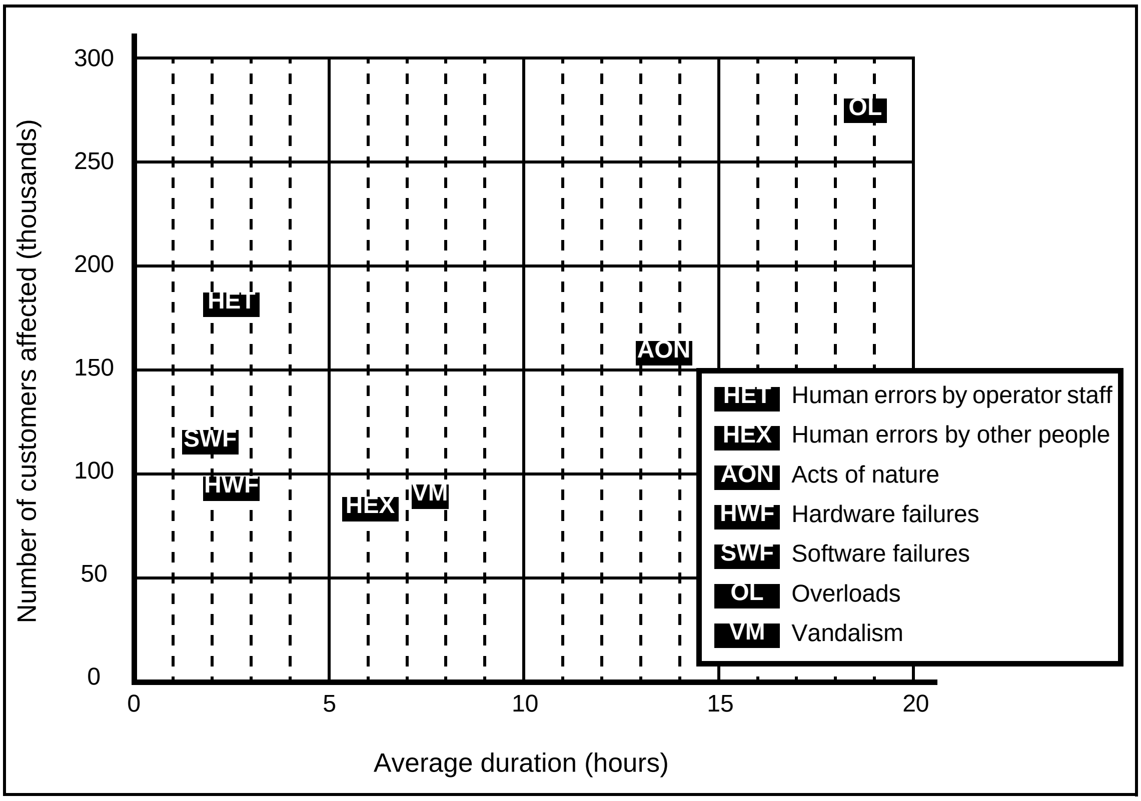


Figure 2 Impact of reported PSTN outages (1992 - 1994)

3.3 ISP Network Availability Improvements

An IP network may be affected by failures of several types, including the following:

- ❐ Link and port failures affecting packet forwarding.
- ❐ Hardware or software failures in control planes affecting forwarding and routing updates.
- ❐ Route flapping from changing link state advertisements.
- ❐ Configuration errors in the network itself (in, for instance, IGP weights or BGP policies) causing traffic to take poor paths.
- ❐ Configuration errors in other peering networks causing traffic to take poor paths or to be lost.
- ❐ Multiple successive failures with interactions leading to traffic congestion or outages.
- ❐ Malicious denial of service attacks consuming network resources.

Failures in the IP layer can be exacerbated by responses from users that perceive the failure. For instance, in telephony, failures could create congestion that impaired quality for calls in progress; if a failure persisted for some seconds many users would regard calls as terminated and redial, which in turn could lead to high transient signalling loads and further impairments in voice quality or high likelihoods of call blocking.

Improving the availability of a network involves considering the network elements as well as the network itself. Typically there should be no single points of failure: network element components should be duplicated to ensure reliability, with 1+1 or 1:*n* sparing being used to ensure that the failure of any single unit does not impair the operation of the network element. For instance, in the TDM PSTN, the signalling network architecture is based either on dual STPs or on highly meshed SSPs, and switches can immediately transfer traffic to predetermined backup routes after a failure.

For routers, improving availability entails reducing disruptions due to maintenance (by using hot equipment swapping and hitless software upgrades), interface failures (by aggregating multiple physical links into single logical links) and route processor failures. (All this can apply to customer network routers as well as to core network routers whenever customers require very high availability.)

The route processor is especially critical, as it maintains connectivity with peer routers and makes routing changes. There are the following approaches to enhancing its availability:

- ❐ Non-stop routing

  In this approach, the router does not need to interact with other routers. Hence the approach can operate with BGP and label distribution protocols as well as with IGPs. A backup route processor operates in conjunction with the main route processor. The backup route processor keeps enough information to continue routing after a failure of the main route processor; it may not keep all of the information, in order to simplify synchronisation between the processors.

- ❐ Non-stop forwarding

  In this approach (which is also referred to as ‘graceful restart’) the router needs to interact with other routers. The router uses extensions to IGPs, such as those in RFC

3623 for OSPF, to indicate that packets can still be forwarded, even though the route processor has failed. When the route processor resumes, it builds a new routing table by obtaining from the surrounding routers all the relevant routing information. If it does not complete its routing table update before the routing information changes then it could create routing loops; if it detects routing information changes it reverts to performing the normal shortest path first process for building a routing table. Extensions to BGP, currently being drafted, provide a similar technique for use between ASs.

In the network itself, achieving dependability targets equivalent to those for the PSTN involves having rapid traffic restoration after a failure of a link or a node. Fast failover mechanisms can be used to move traffic to alternative paths, without degradation in service perceptible to users. If necessary, these alternative paths may then be replaced by others that make more effective use of network capacity but that are established more slowly. Such mechanisms are discussed in section 5.1 and section 5.2.

These considerations apply to routing between ASs as well as routing within ASs. However, the routing protocols and scale of operations are different. Between ASs, physically separate alternative paths are commonly provided between peer routers but BGP sessions may take several minutes to be re-established once they have been terminated. Failovers between Internet ASs last 3 minutes on average, and some of them can cause routing table oscillations that last 15 minutes [12].

4 Specific Performance Techniques

IP networks have been designed primarily to support a range of data services (such as Internet access, content hosting and IP VPNs). However, telephony and multimedia services differ from data services in their degrees of sensitivity to delay and packet loss. In general different services have different requirements; to support them all cost-effectively, different classes of traffic can be given different treatments using the IETF Differentiated Services (DiffServ) architecture. This is described in section 4.1, along with mappings between applications, implementations and treatments.

The DiffServ architecture operates at the IP layer. Many IP networks include domains in which nodes control traffic using a link layer, instead of using IP. Traffic can be differentiated in such domains by using mechanisms that are related to those in the DiffServ architecture by the principles identified in section 4.2, whereby the IP packets tunnel through the link layer domains, as described in section 4.3.

4.1 Differentiated Services

4.1.1 Traffic Classification and Conditioning

The aim of traffic marking is to allow packet flows to be classified, segregated and either prioritised or assigned enough bandwidth to meet performance requirements. The IETF Differentiated Services (DiffServ) architecture defined in RFC 2475 provides a way of differentiating traffic types in order to make aggregate packet flows satisfy different performance requirements. An aggregate packet flow belongs to a traffic class for which the packet marking indicates the behaviour to be provided by routers, in terms of scheduling and, where appropriate, discard priority (or ‘drop precedence’ in RFC 2597). The packet marking is the Differentiated Services Code Point (DSCP) in the IP packet header, and the behaviour to be provided by routers is the Per-Hop Behaviour (PHB) of the traffic class (or ‘behaviour aggregate’ in RFC 2475).

Each packet in a traffic class has the same DSCP and has therefore the same PHB. Different DSCPs can correspond with the same or different PHBs. Various RFCs recommend making particular DSCPs correspond with particular PHBs.

Combining the PHBs provided by routers in paths across a domain results in a Per-Domain Behaviour (PDB). As discussed in RFC 3086, a PDB can underlie the performance targets for an SLA. A PHB is then merely the behaviour provided by a router to an aggregate packet flow that is intended to satisfy performance targets.

In the DiffServ architecture, packets are treated thus on ingress to a DiffServ edge router:

❒ The packets are classified, by examining their headers. The classification may depend on any existing DSCPs or on multiple fields in the headers (such as the source and destination addresses, source and destination port numbers, and protocol identifier as well as any existing DSCPs); in the latter case it is ‘multi-field classification’.

❐ The packets are conditioned by various functions. Among these is marking the packets by setting their DSCPs to indicate the PHBs that should be given by routers to the packets; the routers must implement defined policies on forwarding packets based on the DSCPs, including packets from untrusted sources.

A traffic conditioner in fact comprises one or more of the following functional elements:

❐ Meters measure the temporal properties (e.g. rate of an aggregate traffic flow selected by a classifier). The instantaneous state of these may be used to affect the operation of a marker, shaper or dropper based on policy in the traffic conditioner.

❐ Markers set DSCPs in IP packet headers. They do this with reference to the classification of the packets and any information from the meter that may change the classification (typically by changing the discard priority).

❐ Droppers discard packets where the traffic profile exceeds the configured rate. This is referred to as 'policing'.

❐ Shapers delay packets to conform with a defined traffic profile. Packets may be dropped if there is insufficient buffer space.

Figure 3 illustrates how these functional elements work together.

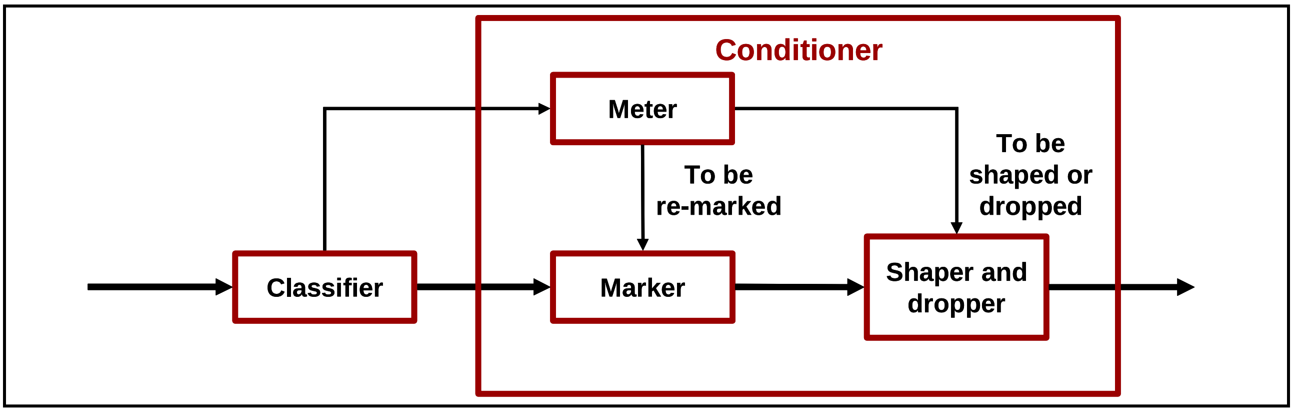


Figure 3 Classifier and conditioner functional elements

Packets destined beyond the DiffServ edge router are forwarded to the outbound interface, while network control packets destined for the router are forwarded to the route processor. The outbound interface acts thus:

❐ The packets are assigned to particular queues for scheduling, by examining their headers. Several scheduling methods are used; among them are strict priority and variants of Weighted Fair Queuing (WFQ), Weighted Round Robin (WRR) and Deficit Round Robin (DRR).

❐ Some packets may be discarded from some queues to manage congestion. Several discard methods are used; most are active queue management algorithms such as Weighted Random Early Detection (WRED).

The DiffServ edge routers perform all the processor-intensive functions, such as multi-field packet classification, packet marking and initial traffic conditioning. The DiffServ interior routers forward packets just by assigning them to the queues for the PHBs identified by the DSCPs, typically without performing traffic conditioning. Because the DiffServ edge routers perform all the processor-intensive functions, the DiffServ interior routers should be able to perform packet forwarding at wire rate.

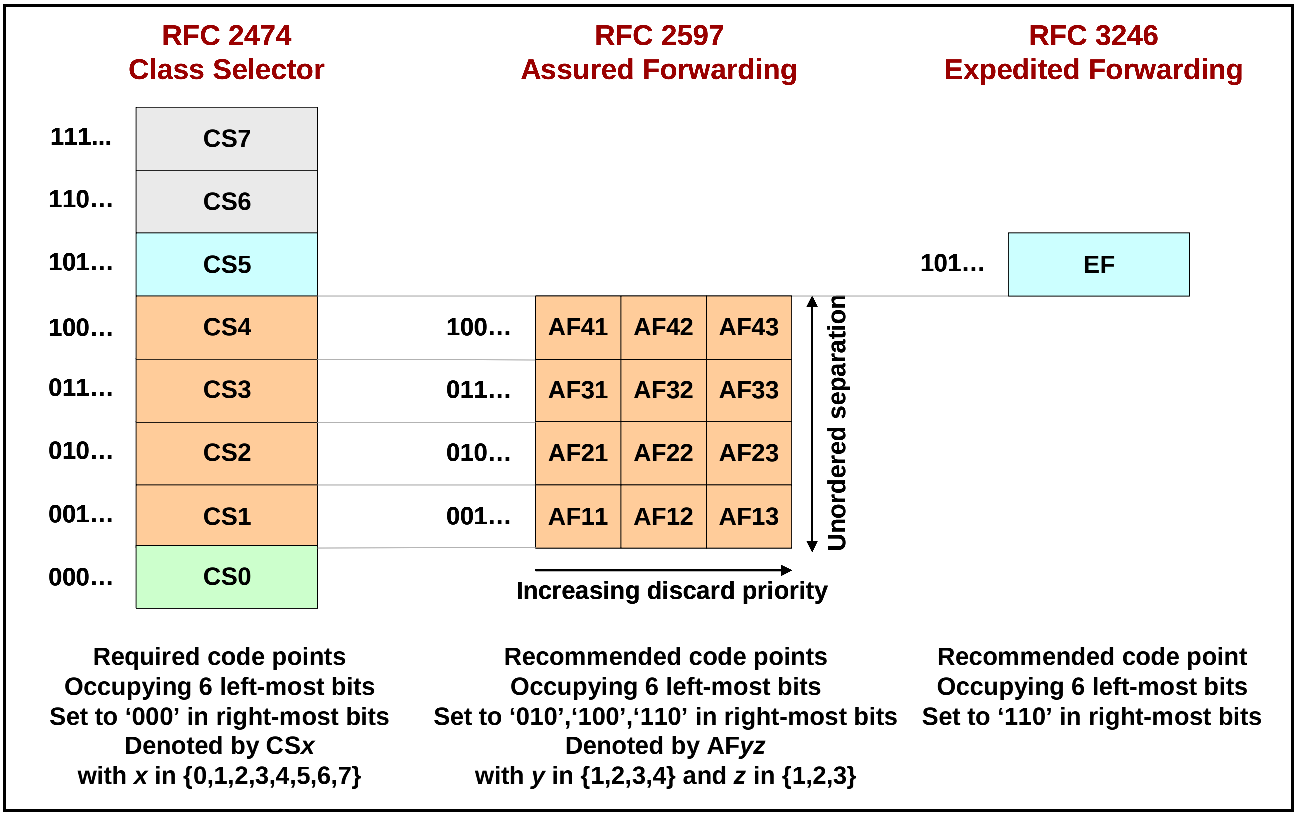
4.1.2 Standards for Behaviours and Markings

The main RFCs that describe particular DSCPs and PHBs are as follows:

- ❒ RFC 3246 specifies a PHB for Expedited Forwarding (EF) and recommends a DSCP for this PHB. This is intended for traffic that requires a low delay and a low packet loss; it is best suited to traffic that must be served at a constant rate. A router typically matches the output rate to the input rate by placing EF traffic in a strict priority queue. The ‘recommended’ DSCP may be denoted by ‘EF’, but traffic may have the EF PHB without having that DSCP.
- ❒ RFC 2597 specifies twelve PHBs for Assured Forwarding (AF) and recommends DSCPs for these PHBs. These are intended for traffic that needs a minimum bandwidth assurance; they are best suited to traffic that can be served at a variable rate and that may even burst above the rate committed by the network. The twelve PHBs are based on four independently forwarded PHB Scheduling Classes (PSCs), each with three packet discard priorities. For each PSC, a router typically handles congestion by queuing and selectively dropping packets. The ‘recommended’ DSCPs are denoted by ‘AF*yz*’ where *y* is in {1,2,3,4} and *z* is in {1,2,3}.
- ❒ RFC 2474 requires at least two PHBs for Class Selector (CS) and requires eight DSCPs for these PHBs. The eight DSCPs provide backwards compatibility with the 3-bit IP Precedence indicator in the former IP Type of Service (ToS) byte by extending these three bits uniquely to six bits. There must be at least two independently forwarded traffic classes, one of which provides the ‘best effort’ Default Forwarding (DF) treatment given by default to traffic that is not otherwise classified. The ‘required’ DSCPs are denoted by ‘CS*x*’ where *x* is in {0,1,2,3,4,5,6,7}.

Other PHBs are possible; for example, RFC 3662 proposes that ‘lower effort’ traffic have a minimum bandwidth assurance of 0% (unlike best effort traffic as it is sometimes configured, but not unlike best effort as it is often envisaged) but be permitted to scavenge for bandwidth not used by other traffic.

Some statements in the RFCs have been omitted from the descriptions above because they are not followed consistently in current practice. Figure 4 illustrates this minimal interpretation of the RFCs, in which the bits in the DSCPs are aligned and described in terms of the structure of the former ToS byte but there is no other ordering of CS*x* or relation between CS*x* and AF*yz*.



Required code points
Occupying 6 left-most bits
Set to ‘000’ in right-most bits
Denoted by CS*x*
with *x* in {0,1,2,3,4,5,6,7}

Recommended code points
Occupying 6 left-most bits
Set to ‘010’,‘100’,‘110’ in right-most bits
Denoted by AF*yz*
with *y* in {1,2,3,4} and *z* in {1,2,3}

Recommended code point
Occupying 6 left-most bits
Set to ‘110’ in right-most bits

Figure 4 Minimal interpretation of code points

4.1.3 Typical Traffic Classes

Some service providers may choose not to implement traffic differentiation in core networks, because it inevitably adds some management overhead. Other service providers may choose to implement merely a distinction between high value traffic (using an EF PHB), for which there are performance and dependability assurances, and low value traffic (using a DF PHB), for which there are no such assurances; the high value traffic is given strict priority in routing and the low value traffic may be discarded after a failure.

In certain environments, service providers may need at least four traffic classes in order to offer some evidently distinct services and use the available bandwidth most effectively while exploiting the capabilities of legacy routers (some of which have four queues). The following classes are representative of such schemes:

❐ Network control

This provides support to network operation, not a service to end users, having a minimum bandwidth assurance, a low or medium delay, and a low or medium packet loss. It is intended for management and control traffic for the managed IP network that must meet delay and packet loss constraints even when some or all of the network is severely congested. In terms of the DiffServ architecture, it has no named PHB and the CS6 DSCP or the CS7 DSCP. (Historically CS6 and CS7 were intended for time-critical network control traffic between administrative boundaries and within administrative boundaries respectively, but this distinction is not perpetuated in the DiffServ architecture.) In this paper a PHB for it is named ‘CF’ (for ‘Control Forwarding’).

- ❒ Real-time (or ‘premium’ or ‘virtual leased line’)

  This provides a TDM equivalent service with a low delay and a low packet loss. It can be regarded as a virtual leased line service if it satisfies all the delay and packet loss constraints imposed by the potential uses of leased lines. In terms of the DiffServ architecture, it has the EF PHB.

- ❒ Non-real-time (or ‘assured bandwidth’)

  This provides a Frame Relay equivalent service having a minimum bandwidth assurance and a low or medium packet loss. It lets traffic burst above the rate committed by the network to a customer, but when this happens its performance depends on the traffic load for all subscribers to all services. In terms of the DiffServ architecture, it has an AF PHB.

- ❒ Best effort (or ‘standard’ or ‘default’)

  This provides an Internet equivalent service; although in principle it offers no assurances, in fact it is often given a minimum bandwidth assurance of 5% - 10% but no limits on its delay or packet loss. In terms of the DiffServ architecture, it has the DF PHB and the CS0 DSCP.

The assured bandwidth class can evidently itself be subdivided by, for instance, designing variants offering different capacity and performance assurances, if the routers so permit. In fact modelling suggests that in access networks at least these classes are desirable for certain forms of multimedia traffic. For instance, the performance requirements for constant bit rate voice traffic, variable bit rate video traffic and file transfer data traffic may not be satisfied simultaneously on a 10 Mb/s link unless the link utilisation is low or different traffic classes are used. (However, video traffic may resemble voice traffic in its performance requirements and may be carried preferentially using a constant bit rate instead of a variable bit rate, especially if it is intended for multimedia conferencing rather than multimedia streaming.)

4.1.4 Applications using Behaviours and Markings

Service providers need to ascertain how traffic should be marked, so that it is given the appropriate treatment by the network to provide performance assurances. Different types of traffic receive different markings and different treatments. As far as possible the markings should be unchanged throughout the managed IP network but the treatments may vary, in that fewer distinctions between behaviours may be needed in core and aggregation networks than in access networks.

Several types of traffic are needed; they can be grouped into media (or ‘user’ or ‘bearer’), signalling (or ‘control’) and management traffic. (The types for telephony over IP are documented along with the requirements [1].) If DiffServ is used fully then these different types of traffic should be provided with behaviours that satisfy the following principles, expressed in terms of DSCPs and PHBs:

- ❒ Media
  - ■ Media traffic may use the EF PHB if it is not liable to create severe packet delay variation for other traffic using the EF PHB.
  - ■ Media traffic should use an AF PHB otherwise. This can be so for fax demodulated using T.38, creating large packets and carried over low bandwidth links.

- ❒ Signalling
  - ■ Signalling traffic may use the CF PHB if it is critical to restoring service in the event of severe network congestion.
  - ■ Signalling traffic may use the EF PHB if it is not liable to create severe packet delay variation for other traffic using the EF PHB.
  - ■ Signalling traffic should use an AF PHB otherwise. This can be so for protocols generating large messages and carried over low bandwidth links.
- ❒ Management
  - ■ Management traffic may use CF PHB if it is critical to restoring service in the event of severe network congestion.
  - ■ Management traffic should use an AF PHB otherwise.

Often stronger statements than these can be made. In fact Nortel has devised and implemented complete mappings (between traffic types, packet markings and behaviours) that satisfy the principles above, to clarify concepts and simplify management: network elements can have appropriate DSCPs and PHBs for different applications by default. These mappings are used as defaults throughout Nortel products. (They also form the basis of IETF work in progress to simplify the exchange of DSCPs between the networks of different service providers and customers.) Table 2 includes these mappings as far as they relate to DSCPs and PHBs and provides more general guidance on how traffic should be marked and treated by covering the following:

- ❒ Applications

  The example applications are categorised and exemplified according to the main types of traffic that they generate. The categories and examples are not necessarily appropriate or exhaustive for any service provider. They represent groupings of the traffic which might be sufficiently refined for access networks but unnecessarily refined for core and aggregation networks.

- ❒ Capacity and performance requirements

  The capacity and performance requirements are those for the main types of traffic generated by the applications. These are expressed in qualitative terms here; they can be read so that ‘high’ in the column on the packet loss ratio, for example, indicates that the application exhibits high tolerance to packet loss. (Quantitative forms of those relevant to telephony are documented elsewhere [1].) Although requirements for bandwidth assurance and packet delay variation occur in descriptions of many traffic classes, they are not mentioned in these capacity and performance requirements because they are implicit in the targets for the delay and the packet loss ratio.

- ❒ Typical DiffServ treatment

  The DSCPs map to the PHBs which routers should provide in order that the capacity and performance requirements are satisfied; the PHBs are implemented by interpreting the capacity and performance requirements in terms of traffic conditioners such as those in RFC 2697 and RFC 2698. They are used as defaults in Nortel products but do not preclude the use of other DSCPs where, for instance, service providers have deployed services already, the characteristics of applications are different, or the network implementation is constrained.

Application		Capacity requirement		Performance requirement		Typical DiffServ treatment	
Category	**Example**	**Upper bound on the bit rate**	**Upper bound on the packet burst size**	**Upper bound on the maximum delay**	**Upper bound on the packet loss ratio**	**DSCP**	**PHB**
Administration	Heartbeats, alarms	Low	Low	Low	Low	CS7	CF with suitable queue weights
Network control	Routing updates	Low	High	Medium	Medium	CS6	
Telephony	Voice, fax, circuit emulation, ISDN video as in H.320	High	Low	Low	Low	EF, CS5	EF
Multimedia conferencing	LAN video as in H.323, games	High	High	Low	Low	AF4*z*	AF with suitable queue weights and discard priorities
Multimedia streaming	Webcasts, closed-circuit television	High	High	High	Low	AF3*z*	
Low latency data	Client/server applications, transactions	High	High	Medium	Medium	AF2*z*	
High throughput data	Store/forward applications, email, billing, file transfer	High	High	High	Medium	AF1*z*	
Standard	Undifferentiated applications	High	High	High	High	CS0	DF

Table 2 Capacity and performance requirements for common applications

This choice of DSCPs for telephony over IP can be explained by applying the principles above to the different traffic types as follows:

- Media
  - Media traffic that requires low delay and low packet loss would have the EF PHB. This traffic would comprise not just voice but also fax upspeeded to G.711 and even fax demodulated using T.38 where the packets are small. In fact the predominant telephony over IP traffic is media traffic requiring low delay and low packet loss, so the IP network must be engineered to support a high proportion of such traffic.
  - Other media traffic would have an AF PHB. However, for telephony over IP the only such traffic comprises fax demodulated using T.38 where the packets are large.
- Signalling
  - Signalling traffic that is as critical to restoring service as signalling traffic for the routers could have the CF PHB and the CS6 DSCP or the CS7 DSCP. Signalling traffic such as SIP-T between two carrier-located communication servers might be treated thus.
  - Other signalling traffic would have the EF PHB or an AF PHB offering low delay where possible (as otherwise ring clipping might occur). Signalling traffic such as H.248 between a communication server and a carrier-located media gateway might be treated thus.
  - This signalling traffic could also need a distinction between DSCPs which correspond with the same PHB in core networks but with different PHBs in

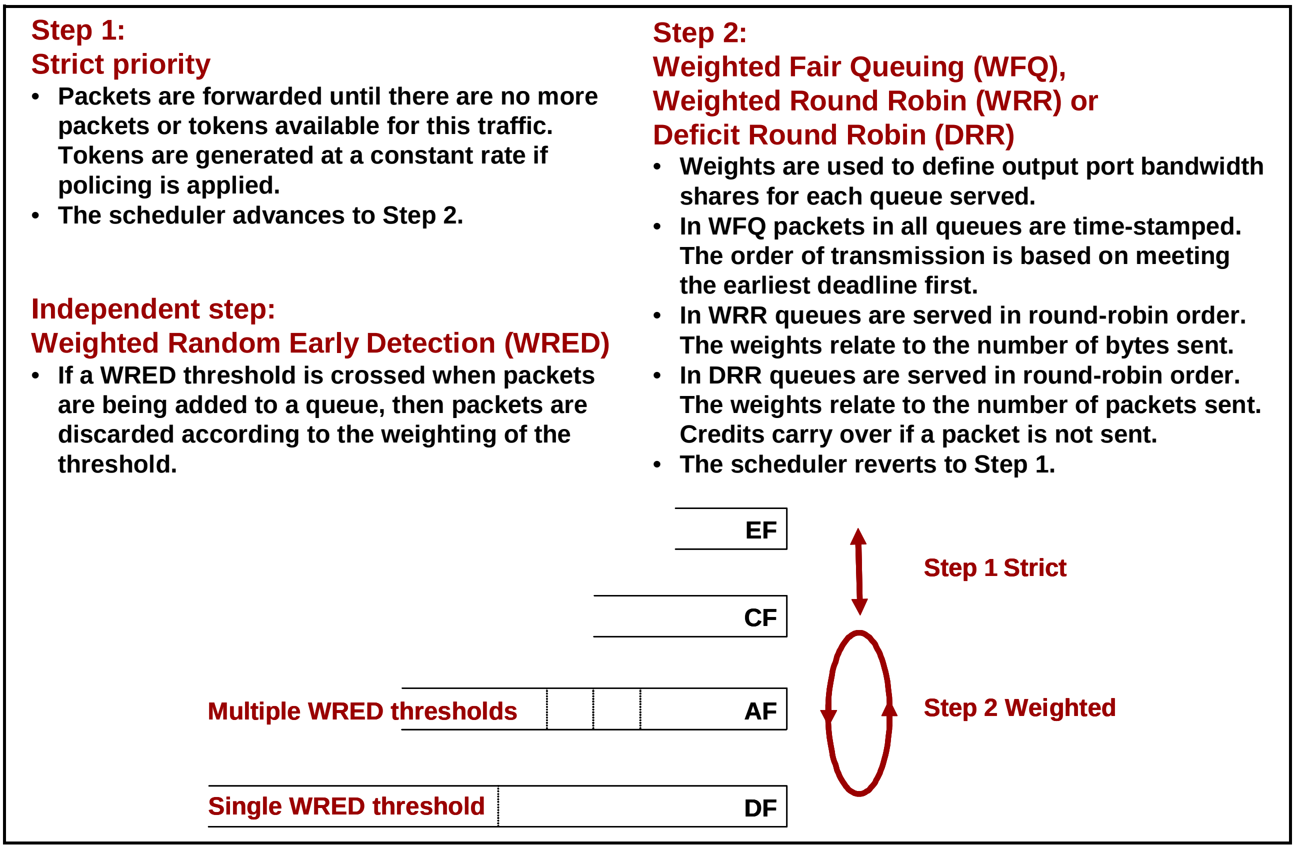
low bandwidth (less than 1 Mb/s) access networks. The distinction is expressed here in terms of EF and CS5. (The choice of the CS5 DSCP is due to the alignment of the bits in the DSCPs; an AF PHB would be adopted for the CS5 DSCP where necessary.) In low bandwidth (less than 1 Mb/s) access networks. media traffic may need to be separated from signalling traffic, with higher priority being given to media traffic that could otherwise suffer from high packet delay variation and consequential voice quality degradation. Signalling traffic such as SIP between a communication server and a customer-located media gateway or an IP client might be treated thus.

- ❐ Management
  - Management traffic that is as critical to restoring service as management traffic for the routers could have the CF PHB and the CS6 DSCP or the CS7 DSCP.
  - Other management traffic would have an AF PHB not offering low delay. It does not generally have the DF PHB and the CS0 DSCP, as it needs a minimum bandwidth assurance.

4.1.5 Implementations of Behaviours and Markings

Routers typically have four or eight outbound interface queues (a low delay queue using strict priority and other queues using a scheme such as WFQ, WRR or DRR) and manage congestion using WRED. Figure 5 illustrates schematically the interaction of these.

.



Step 1:
Strict priority

- Packets are forwarded until there are no more packets or tokens available for this traffic. Tokens are generated at a constant rate if policing is applied.
- The scheduler advances to Step 2.

Independent step:
Weighted Random Early Detection (WRED)

- If a WRED threshold is crossed when packets are being added to a queue, then packets are discarded according to the weighting of the threshold.

Step 2:
Weighted Fair Queuing (WFQ), Weighted Round Robin (WRR) or Deficit Round Robin (DRR)

- Weights are used to define output port bandwidth shares for each queue served.
- In WFQ packets in all queues are time-stamped. The order of transmission is based on meeting the earliest deadline first.
- In WRR queues are served in round-robin order. The weights relate to the number of bytes sent.
- In DRR queues are served in round-robin order. The weights relate to the number of packets sent. Credits carry over if a packet is not sent.
- The scheduler reverts to Step 1.

Figure 5 Application of WRED with strict priority and WFQ, WRR or DRR queues

PHBs are implemented in routers by adopting these scheduling and congestion management techniques. The following usage is fairly typical:

❒ CF traffic is forwarded from a queue using WFQ, WRR or DRR scheduling and taking perhaps 5% of the overall bandwidth. Congestion management is not applied to the queue.

❒ EF traffic is forwarded from a strict priority queue; delay is kept low because the scheduler will always output a packet when there is one present. Congestion management is not applied to the queue.

❒ AF traffic is forwarded from queues using WFQ, WRR or DRR scheduling. Congestion management is applied to the queue using discard priorities that correspond to the values of *z* in AF*yz*. One PSC (such as AF*yz* with different values of *z*) will generally use just one queue in a router, because otherwise packets in traffic flows having different PHBs in the same PSC might overtake one another and not just differ in their discard priorities.

❒ DF traffic is forwarded from a queue using WFQ, WRR or DRR scheduling and taking perhaps 5% - 10% of the overall bandwidth. Congestion management is applied to the queue.

Traffic from subscribers is policed on ingress to discard packets that are outside the contract and thereby maintain bandwidth assurances for other traffic. When the traffic does not require low delay it may be shaped.

The aggregate link utilisation for all of the traffic that has bandwidth assurances should not exceed 60% - 80% on any link even after a failure. (The link utilisation can depend on the link bandwidth and maximum delay, as outlined in section 2.2.)

4.2 Link Layer Mechanisms

IP traffic can be carried over different link layers as it traverses the network. These link layers may have their own Class of Service (CoS) mechanisms for nodal traffic control, handling such matters as traffic conditioning, queue management and scheduling. If the link layers are to be exploited fully instead of IP routing, the DSCPs must be mapped to link layer markings that use these mechanisms in ways that are consistent with the intentions in the IP layer.

Such mechanisms are available for the following link layer protocols:

❒ Multi-Protocol Label Switching (MPLS).

❒ Ethernet.

❒ Multi-class Multi-link Point-to-Point Protocol (PPP).

❒ Frame Relay.

❒ Asynchronous Transfer Mode (ATM).

Nortel has defined and implemented mappings between DSCPs and link layer markings, to simplify management by aligning the IP layer and link layer treatments of packets by default. This section provides an introduction to some of these mappings. The overall structure of a network using them is considered in further detail in section 4.3.

4.2.1 MPLS

MPLS has a marking scheme that uses a 3-bit field (commonly termed the 'EXP field') in the MPLS header. It can also use the label that identifies the Label Switched Path (LSP), with or without the EXP field. In fact, MPLS can operate without explicit MPLS headers, by using an underlying label switching link layer such as Frame Relay or ATM to provide VCs that act as MPLS LSPs; arguably, MPLS is not a 'link layer' at all.

The following approaches to relating DiffServ to MPLS are described in RFC 3270:

- ❒ EXP-inferred-PSC LSP ('E-LSP')

  An E-LSP relates DSCPs to EXP fields, allowing eight classes of service to be determined by EXP fields; for instance, the EXP fields might be identical with the three left-most bits of the DSCPs.

- ❒ Label-only-inferred-PSC LSP ('L-LSP')

  An L-LSP relates DSCPs to labels, using EXP fields for discard priorities. It is applicable even when the MPLS layer is stacked on top of another label switching link layer and there are no MPLS headers; in this case Virtual Circuit (VC) identifiers in the link layer underneath MPLS are used as labels and markings in the link layer underneath MPLS are used as discard priorities. This can happen when MPLS is used in conjunction with Frame Relay or ATM, where 1-bit discard eligibilities or cell loss priorities must be used for discard priorities.

4.2.2 Ethernet

IEEE 802.1Q has a marking scheme that uses a 3-bit field (commonly termed the 'IEEE 802.1p user priority') in the VLAN tag. IEEE 802.1Q therefore establishes eight classes of service that switches can use when forwarding traffic. Modern Ethernet switches can implement classes of service using both strict priority scheduling and weighted scheduling.

In an Ethernet portion of a network within an IP network using DiffServ there may be switches that ignore DSCPs or routing switches that can use both DSCPs and 802.1p user priorities; at the edge of the Ethernet portion there may be routers that ignore 802.1p user priorities or routing switches that can use both DSCPs and 802.1p user priorities. The routing switches need to map between DSCPs and 802.1p user priorities if traffic may subsequently encounter switches or routers that are not able to use both DSCPs and 802.1p user priorities. Hence:

- ❒ Frames should have their 802.1p user priorities marked (using the translation of DSCPs into 802.1p user priorities) when leaving routing switches if they may be going to switches that ignore DSCPs.
- ❒ Packets without trusted DSCPs should have their DSCPs marked (using DiffServ classification or the translation of 802.1p user priorities into DSCPs) when entering routing switches from switches if they may be going to routing switches or routers that ignore 802.1p user priorities.

4.2.3 Multi-class Multi-link PPP

Multi-class Multi-link PPP has a marking scheme that uses either a 2-bit field or a 4-bit field in the Multi-link PPP header. Both versions of the scheme are intended for use predominantly over low bandwidth links, where packets are fragmented into frames and frames having higher priorities are transmitted ahead of frames having lower priorities.

The choices of field size and header length depend on the bandwidth available and the number of queues supported by the devices at the ends of the links. (For instance, on a link of 128 Kb/s at most four service classes are likely to be used because otherwise lower priority classes receive too little bandwidth.) There are different ways of relating DSCPs to service classes, corresponding with these different choices.

4.2.4 Frame Relay

Frame Relay has extensive mature standards for performance metrics. These are used in X.146 to identify four service classes, each of which has delay and frame loss targets suited to particular applications. The service classes can be adopted for DiffServ thus:

- ❒ Service class 3 and discard eligibility 0 should be used for traffic having a CF PHB.
- ❒ Service class 3 and discard eligibility 0 should be used for traffic having the EF PHB.
- ❒ Service class 3 should be used for traffic having an AF PHB offering a low delay and offering a low packet loss.
- ❒ Service class 2 should be used for traffic having an AF PHB not offering a low delay and offering a low packet loss.
- ❒ Service class 1 should be used for traffic having an AF PHB not offering a low packet loss.
- ❒ Service class 0 and discard eligibility 1 should be used for traffic having the DF PHB.
- ❒ Discard eligibility 0 should be used for traffic having an AF*y*1 PHB.
- ❒ Discard eligibility 1 should be used for traffic having an AF*y*2 or AF*y*3 PHB.

The rules for interworking MPLS and Frame Relay use essentially the last two recommendations about discard eligibilities to express two discard priorities.

4.2.5 ATM

ATM has extensive mature standards for performance metrics. These are used to specify several service categories. The most important of these are Constant Bit Rate (CBR), real-time Variable Bit Rate (rt-VBR), non-real-time Variable Bit Rate (nrt-VBR) and Unspecified Bit Rate (UBR); Available Bit Rate (ABR) and Guaranteed Frame Rate (GFR) are rarely used. The service categories can be adopted for DiffServ thus:

- ❒ rt-VBR and cell loss priority 0 should be used for traffic having a CF PHB.
- ❒ CBR and cell loss priority 0 should be used for traffic having the EF PHB.
- ❒ rt-VBR should be used for traffic having an AF PHB offering a low delay.
- ❒ nrt-VBR should be used for traffic having an AF PHB not offering a low delay.
- ❒ UBR and cell loss priority 1 should be used for traffic having the DF PHB.
- ❒ Cell loss priority 0 should be used for traffic having an AF*y*1 PHB.

❒ Cell loss priority 1 should be used for traffic having an AF*y*2 or AF*y*3 PHB.

The rules for interworking MPLS and ATM use essentially the last two recommendations about cell loss priorities to express two discard priorities.

4.3 Link Layer Tunnels

Parts of an IP network may not use DiffServ or may not have DSCPs that can be trusted. Edge nodes, at the edge of the domain in which DiffServ is used, mark DSCPs. Interior nodes, interior to that domain, treat packets in accordance with those DSCPs.

The concepts of ‘domain’, ‘edge node’ and ‘interior node’ used here are applicable in other contexts besides that of DiffServ. There are several contexts in which a network may be partitioned into domains having different roles or statuses and formed from edge nodes and interior nodes. Edge nodes connect a domain to other domains (or other networks) and perform functions that are specific to the domain. (For instance, at the edge of a routing domain in which the nodes regard each other as trustworthy, the edge routers perform filtering on the traffic from adjoining domains.) Interior nodes interconnect to carry traffic across the domain. The differences between edge nodes and interior nodes therefore relate mainly to their functions (not to their forwarding rates and interface rates, for example).

In particular, in the DiffServ domain itself there may be various domains that use link layer nodal traffic control. The nodes at the edges of these link layer domains interwork DSCPs with appropriate link layer markings and Virtual Circuits (VCs) in the ways described in section 4.2 (with VCs being used just for MPLS, Frame Relay and ATM). The nodes interior to these domains, which are link layer switches, treat link layer frames in accordance with those link layer markings and VCs. In effect packets having DSCPs tunnel through a link layer domain: they enter the domain at a suitable edge node and leave at another suitable edge node, where their DSCPs may be used as they were or may be modified to match the link layer markings and VCs. (These link layer markings and VCs, however, may well use a coarser classification of behaviours than the DSCPs.)

Figure 6 illustrates how the interworking of DiffServ with a link layer mechanism can exploit the link layer mechanism inside the link layer domain but keep the use of the link layer invisible outside the link layer domain.

When the link layer is connection-oriented, separate VCs (in particular, separate LSPs for MPLS) may be defined for media traffic, signalling traffic and management traffic, as packets having different traffic types take different routes. In particular:

❒ Media traffic passing between media devices (such as media servers, media gateways, media proxies and IP clients) may sometimes use a mesh of VCs across the core network.

❒ Signalling traffic passing to media devices (such as media servers, media gateways, media proxies and IP clients) from communication servers (e.g. H.248, MGCP, H.323, SIP) will often use a star of VCs across the core network.

❒ Signalling traffic passing between communication servers (e.g. SIP, SIP-T) will need a mesh of VCs across the core network.

❒ Management traffic may be routed without using link layer switching. However, if VCs are preferred, the management traffic VCs might be combined with the signalling traffic VCs where the routes and desired PHBs are the same.

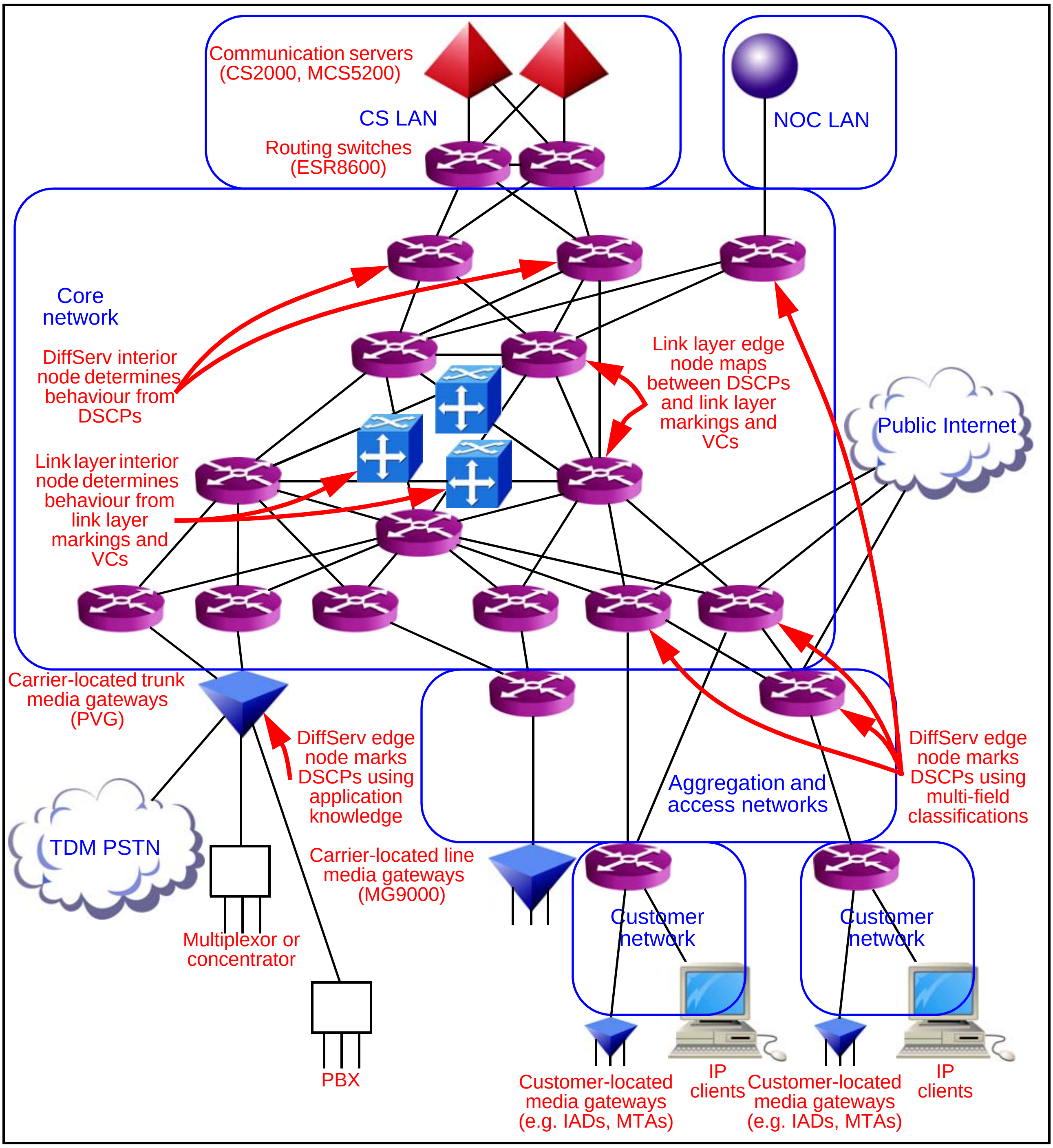


Figure 6 Tunnelling DiffServ DSCPs through link layer markings and VCs

5 Specific Dependability Techniques

Ensuring that IP networks have high availability requires careful design, to make traffic follow routes having known behaviours and to keep disruption after a failure to acceptable levels. Service providers that wish to use native IP routing, without a connection-oriented link layer, can adopt the techniques discussed in section 5.1 for these purposes, accompanying them with the IETF Differentiated Services (DiffServ) architecture where appropriate. These techniques are some of those often referred to as 'traffic engineering'.

MPLS with facilities for traffic engineering can also be used for these purposes in the way described in section 5.2. In this guise it may be extended to interwork with DiffServ for aggregate traffic flows at the present and for individual traffic types in the future.

5.1 Native IP Routing

5.1.1 Shortest Path Routing

Routing control allows a service provider to exploit the network cost-effectively by ensuring that the traffic flows do not vary in ways that lead to over-utilisation or under-utilisation of network segments. Using native IP routing control capabilities usually entails using OSPF or IS-IS as the Interior Gateway Protocol (IGP). These operate in essentially the same way as each other and can be used in a hierarchical manner to control the sizes, and the distribution between routers, of routing databases.

Each router maintains tables that:

❒ List each IP address prefix to which packets can be sent.

❒ Identify each router that can be used for the next hop towards each IP address prefix.

❒ Give each route a weight by summing the weights of the links along the route.

The route selected for a given protocol type and route type is the 'shortest path', which is the route with the lowest weight. If there are two or more routes with equal weights, load balancing can be used to distribute packets equally between the possible routes. (This should be done in such a way that all of the packets in a real-time traffic stream from a particular source to a particular destination use the same route, to prevent the delivery of packets out of sequence.) Routing control using OSPF or IS-IS involves determining the routing by adjusting the weights assigned to different links.

Careful assignments of link weights can lead to utilisations that are near to optimal in realistic networks [6]. Furthermore, these assignments can remain usable when there are errors in the traffic matrices estimated from the actual link utilisations and the network routing [17]. When link weights are the inverses of the link bandwidths (as they are, in widespread defaults) the utilisations can be far from optimal, by differing from the optimum by a factor of 2; however, such link weights still produce better results than link weights unrelated to the required bandwidth and the available bandwidth.

5.1.2 Routing Control

The underlying principles of routing control using OSPF or IS-IS are best understood by means of an example [6]. This uses a simple network in which all links have the same capacity, and each of four nodes (***q,r,s,w***) has an equal amount of 1 unit of traffic to send to node ***t***. The objective is to minimise the maximum load on all the links. Figure 7 shows three different loading outcomes determined by the weights assigned to each link in the network, with the level of traffic on each link indicated by the thickness of the line representing that link.

Since four units of traffic have to reach node ***t*** via its two incoming links, an optimal distribution of traffic is to have two units of traffic coming in via each of these links. This is not achieved by having the same weight of 1 on every link, which directs all of the traffic from nodes ***q***, ***r***, and ***s*** through node ***u***, and forces 3 units of load on link (***u,t***). It is also not achieved by simply increasing the weight of the overloaded link, which results in two shortest paths from nodes ***q***, ***r***, and ***s*** and therefore with load balancing an even splitting of traffic with 1.5 units over paths via ***u*** and ***v,*** and a load of 2.5 units on the link (***w,t***). Instead, assigning a high weight to link (***r,u***) has the effect of diverting traffic from ***r*** via ***v*** rather than ***u***, and results in an optimal distribution of traffic. No other routing scheme could produce a better solution with this objective.

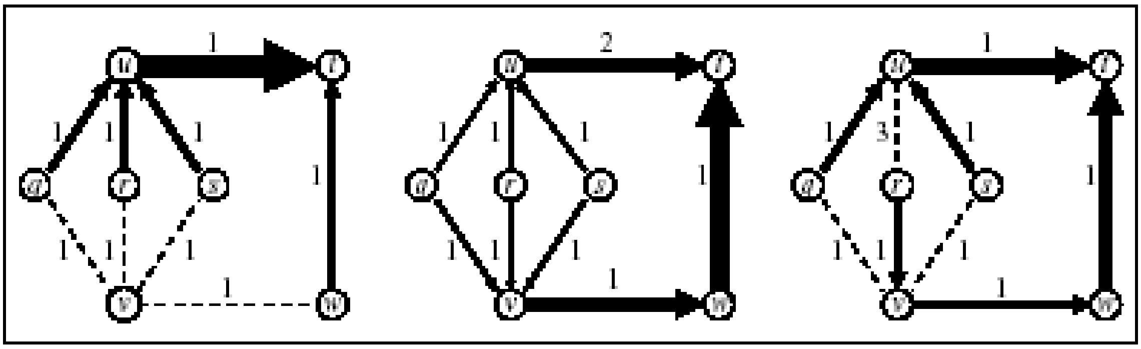


Figure 7 Example of routing control using weights

Changing link weights to alleviate congestion on the link (***u,t***) is an attractive alternative to buying and deploying additional bandwidth between routers ***u*** and ***t***. However, this example shows that diverse and complex routing patterns can be produced even with relatively small networks. Changing link weights can have a major impact on traffic flows across the network and may produce poor results in other parts of the network. Hence configurations should be evaluated through modelling before deployment, and updates should not be too frequent or performed without careful analysis of possible network instability.

To complement routing control by increasing network robustness are as follows:

- At least two physically diverse routes should be provided between each pair of end points under normal operating conditions.
- The routes surviving after a failure should be able to carry all of the traffic (including rerouted traffic) for which the SLAs impose suitable dependability requirements.

5.1.3 Rapid Traffic Restoration

Failures of links (involving interface cards, transmission layer network elements or fibres that may be cut) and nodes (involving control planes) can disrupt networks too much for real-time services. The disruption can be due to the time taken to detect failures or the time taken to restore traffic after failures. Node failures are the more problematic, because they may take longer to detect and may generate more messages during traffic restoration; they should be made very infrequent by ensuring that the nodes have suitably high redundancy, including hitless software upgrade and hot equipment protection, especially when they cannot be treated as link failures.

The options for detecting a link or node failure are as follows:

❒ Transmission layer failure detection

A transmission layer using SDH can pass alarms about detected failures to the IP layer in 10 ms - 20 ms. The detected failures can include failures of nodes having only SDH interfaces provided that when such a node fails it issues a ‘dying gasp’ by inducing suitable alarms on all its links. (Some link layers also provide failure detection mechanisms that could be used in this way.)

❒ Link layer failure detection

A link layer such as MPLS, Frame Relay or ATM can provide equivalents to the ‘keep alive’ messages of the IP layer; the failure of these to arrive can then be treated as an indication of failure of the link.

❒ IP layer failure detection

The detection of a failure requires routers to detect that some number of consecutive ‘keep alive’ (or “hello”) messages from a router is missing. The number of missing messages often defaults to three and the interval between the messages often defaults to 10 seconds. The interval can be reduced greatly, to make the time for detecting a failure fall to 1 second - 2 seconds, provided that the route processors do not become overloaded and routing stability is not jeopardised. Doing this may entail requiring that the ‘keep alive’ messages are used purely for indicating liveness and are not burdened with other functions.

Consequently a router having both SDH interfaces and Gigabit Ethernet interfaces might be configured to detect failures according to the presence of alarms on the SDH interfaces and the absence of enough consecutive ‘keep alive’ messages in the IP layer on the Gigabit Ethernet interfaces.

The options for restoring traffic after such a failure are as follows:

❒ Transmission layer traffic restoration

A transmission layer using SDH protects paths by switching to preconfigured backup paths having 1+1 or 1:*n* redundancy. The IP routers are normally unaware of the protection switch; they may miss one ‘keep alive’ message, but this will not in itself cause routing protocol reconvergence. The interruption to packet forwarding at the IP layer should then be the transmission layer rerouting time of 50 ms - 100 ms and apply only at the routers at the point of failure.

Although SDH can provide protection on end-to-end paths, typically the form of protection switching deployed in IP networks is point-by-point (using ‘multiplex section protection’ as opposed to ‘subnetwork connection protection’). This protects against link failures more effectively than it does against node failures.

- IP layer traffic restoration

  The IP layer can reroute traffic by inducing network-wide routing protocol reconvergence to restore stable packet forwarding on an alternative path around a failed link or node. Doing this requires the following operations:

  - On detecting a failure, a router waits for some time to confirm that the failure is not transient, before flooding Link State Advertisements (LSAs) into the network to inform all the other routers of the need to recompute routes. A fully meshed network having *n* nodes can generate O(n^2) LSAs after a link failure and O(n^3) LSAs after a node failure.
  - On receiving an LSA, a router waits for some time to aggregate LSAs on each router, before performing a shortest path calculation to create a routing information base and then updating the forwarding information base on its interface cards (during which packet forwarding may be interrupted). In the worst case, a network having *l* links and *n* nodes requires a shortest path calculation of complexity O($l \times \log(n)$), although different algorithms are suited to different network topologies. The calculation can be made more rapid by reducing it to an incremental shortest path calculation, which recomputes only the affected part of the routing tree.

  Thus routing protocol reconvergence is controlled by several timers and can take some time after the detection of a failure (depending on the network size and topology). The timers ensure that new consistent routing information is propagated network-wide, so that alternative paths become stable before packets are forwarded along them. In some networks careful reductions in the timer settings can bring the time for restoring traffic after a failure down to 1 second - 2 seconds while maintaining network stability.

Consequently a combination of transmission layer failure detection and IP layer traffic restoration can handle link and node failures quite fast in suitable networks [2]. However, this may not be quite fast enough to satisfy dependability requirements, especially if they are intended to be equivalent to those for the TDM PSTN [1]: failures might not only induce temporary impairments in the quality of calls in progress but also cause calls to be terminated prematurely, because of equipment time-out or user actions.

A further reduction in the time taken for traffic restoration can nonetheless be achieved within the framework of existing routing protocols. It entails the predetermination of alternative paths, so that successful routing can occur while new routing information bases and forwarding information bases are being produced. (Routes offering load balancing by ECMP can be regarded as such alternative paths.) In more detail this depends on the following:

- On detecting a failure, a router makes its interface cards switch to using alternative paths. It thereby performs a local repair that diverts traffic around the failure.
- On completing a shortest path calculation, a router updates its interface cards with the results of the shortest path calculation, which include definitions of replacement alternative paths that can be used if there is a subsequent failure.
- Extensions to link state IGPs, currently being drafted, define extra LSA fields; these allow a router to indicate that some of its links may be used by an upstream router in alternative paths to an arbitrary destination prefix or can break a single-hop forwarding loop to an upstream router. They thereby permit routing loops to be broken and increase greatly the coverage achieved by alternative paths.

For link failure an alternative to traffic restoration by rapid rerouting is the use of aggregate links; after a failure of one of the links, the remaining links carry at least the traffic for which the SLAs impose suitable dependability requirements. Such links can be available with SDH interfaces (in the form of ‘composite links’), as well as with link layers such as Gigabit Ethernet, Multi-class Multi-link PPP, Frame Relay and ATM.

5.1.4 DiffServ Interworking

DiffServ is independent from IP routing, in that different traffic classes have different behaviours but are subject to the same routing decisions. In particular, the weights assigned to different links in native IP routing do not usually distinguish between different types of traffic; instead, traffic flows between two end points take the same routes as each other even when packet marking is used to provide traffic differentiation. (However, traffic restoration after a failure may be able to exploit traffic differentiation if the SLAs permit, as demonstrated in section 2.2.)

Whether the independence of DiffServ from IP routing affects the network design depends on the network, for the following reasons:

❒ Across core networks there are usually several routes, along high bandwidth links. Routing control and rapid traffic restoration may provide more important design challenges than adequate bandwidth provision. Few traffic classes may be enough, because when networks are well designed and generously supplied with high bandwidth links all of the traffic on them can be given very similar performance.

However, even when the number of traffic classes is reduced to simplify operations, there can be a case for performing some traffic differentiation. At its simplest this could entail providing high performance and dependability for some of the traffic (by using, in particular, strict priority queuing and rapid rerouting) and explicitly offering no assurances for the rest of the traffic; in such a scheme, the overall link utilisation can be increased substantially, but when there is a heavy load or a high proportion of strict priority traffic, the rest of the traffic can experience high delay and high packet loss.

❒ Across aggregation and access networks there are typically few separate routes, along low or medium bandwidth links. Adequate bandwidth provision may provide more important design challenges than routing control and rapid traffic restoration. Several traffic classes may be wanted, for reasons mentioned in 4.1, but the routes may be unique.

Nonetheless, when the routes are not unique, routing control and rapid traffic restoration need to be considered.

When the independence of DiffServ from IP routing does affect the network design, there are several algorithms that can be used to determine routes. They perform constraint-based routing, or Quality of Service (QoS) routing, for the constraints appropriate to native IP routing. These algorithms have been examined mainly in the context of a split between traffic that has bandwidth assurances (such as can be provided in conjunction with strict priority queuing, for instance) and traffic that has no bandwidth assurances. The choice of algorithm can influence significantly the behaviour of traffic that has no bandwidth assurances. It can depend on the network size and topology: algorithms can differ in how consistently they produce better results than shortest path calculations that ignore bandwidth assurances and in whether they aim to achieve load balancing or to limit resource consumption (which is typically expressed in terms of the number of hops traversed by traffic that has bandwidth assurances).

5.2 Multi-Protocol Label Switching

5.2.1 Label Switching

Figure 8 provides a high-level functional view of how label switching in MPLS is used to convey traffic across a network. A Label Switched Path (LSP) is set up to convey packets along a sequence of Label Switching Routers (LSRs). For each IP packet incoming into an MPLS domain, an ingress edge LSR, or Label Edge Router (LER), assigns an MPLS label that is used to forward the IP packet across the network towards an egress LER. Forwarding across the network requires label matching and swapping in the LSRs; the full prefix matching at each hop required in IP routing is not needed in label switching.

A label distribution protocol can be used to inform LSRs about the labels assigned to the LSPs reaching particular destinations. Various such protocols have been proposed for different purposes, using either developments of new protocols or extensions to existing protocols.

MPLS can be used in conjunction with not only multiple transport and application level protocols (above IP) but also multiple link layer protocols (below IP), including PPP, Frame Relay and ATM. It can even adopt directly the labels provided by Frame Relay and ATM, which themselves use label switching. It can also be stacked above itself, so there can be nested switching domains, each using MPLS; thus MPLS permits a switching and routing hierarchy, with IP routing being required only at the edges of the outermost MPLS domains for routing between such domains.

MPLS was conceived initially to improve router performance, but this motivation has diminished considerably as advances in router design have lead to wire-rate forwarding of IP packets without link layer switching. However, MPLS has also been developed in other ways, to support VPNs and routing control. Most deployments to date have used MPLS for IP VPNs based on RFC 2547, but now standards for using MPLS for link layer VPNs are reaching maturity (especially for Ethernet). MPLS for routing control, in the form of MPLS Traffic Engineering (MPLS-TE), is the aspect of MPLS that is relevant to this paper, because it is now expected to become more widespread, especially with the introduction of services for which the performance and dependability are required to be equivalent to those in the PSTN.

In practice many service providers considering the use of MPLS-TE will already have used MPLS for IP VPNs. The use of MPLS for IP VPNs is neither inhibited nor required by MPLS-TE. For instance, telephony traffic going from a customer IP VPN to the TDM PSTN passes from the customer IP VPN into the managed IP network; the relevant partition of the managed IP network can be, but does not need to be, an IP VPN in its own right and can rely on native IP routing or on MPLS-TE.

In this context the term ‘traffic engineering’ is often interpreted rather narrowly, following the main content of RFC 2702; the term can also be used more broadly, as in RFC 3272, to cover nodal traffic control, handling such matters as traffic conditioning, queue management and scheduling, as well as routing control. Some of these matters are addressed by DiffServ, which can be deployed independently of MPLS-TE.

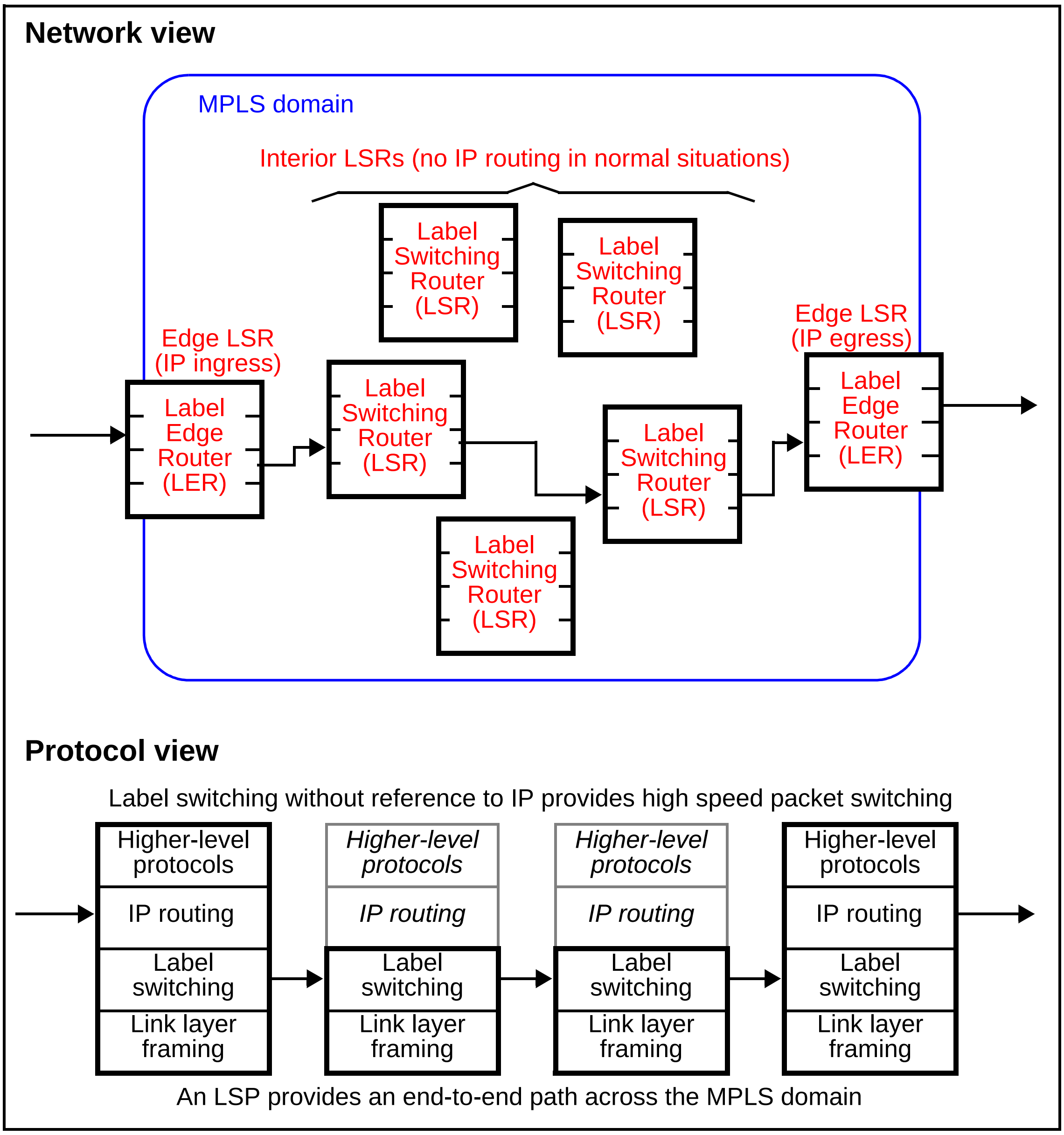


Figure 8 High-level view of label switching

5.2.2 Routing Control

MPLS-TE allows routes throughout an MPLS domain to be determined either hop-by-hop (in which case each LSR chooses the next hop, using for example the IGP routing information base) or explicitly (in which case a single LSR, which is usually the ingress LER or the egress LER, specifies some or all of the LSRs in the LSP). Explicitly routed LSPs that are laid down from ingress LERs are especially important for routing control.

MPLS-TE uses the following techniques:

❒ Extensions to link state IGPs, such as those in RFC 3630 for OSPF, define extra LSA fields; these allow a router to monitor, and exploit for routing purposes, the use of network resources. They could be used by routers to determine shortest paths according to link weights and additional constraints, without any use of MPLS. Thus far the constraints on links expressed in these fields (such as maximum bandwidth, maximum reservable bandwidth, unreserved bandwidth and administrative group) have been fixed only for certain protocol applications.

- ❐ Constraint-based routing can determine routes that match the network resources to the demands appropriately. It allows routes to be designed according to constraints on links rather than just weights derived from maximum bandwidths. It thereby lets the network capacity be used more cost-effectively.
- ❐ Extensions to Resource reSerVation Protocol (RSVP), documented in RFC 3209 as RSVP Tunnelling Extensions (RSVP-TE), define extra RSVP objects; these allow an LSR to establish LSPs that follow particular explicit routes (such as those determined by constraint-based routing). (Here 'TE' here means 'Tunnelling Extensions', but elsewhere it means 'Traffic Engineering'.) RSVP was originally developed to allow routers to reserve resources for individual traffic flows between hosts. However, doing this proved to be unscalable owing to the amount of state information that had to be maintained in each router. RSVP-TE follows RSVP in invoking admission control at every node on a route but differs from RSVP in being used for signalling aggregated host-to-host flows between LSRs (as opposed to single host-to-host flows between hosts). The admission control in MPLS-TE therefore limits admission for aggregate traffic flows, not for individual traffic flows or individual application sessions.

Constraint-based routing increases the complexity of network operation. As the number of points of ingress to the MPLS domain increases, the number of routes that MPLS LSPs may use increases steeply. This is further compounded where there are multiple traffic classes and multiple LSP attributes. Furthermore, design sensitivity to LSP placement order can occur with constraint-based routing, when the available link capacity left over after each LSP assignment is incorporated in the routing decision. Ideally designs should be insensitive to LSP placement order and to traffic flow changes within a reasonable scaling factor (perhaps 2 increase or decrease). In larger networks, the use of automated design tools becomes imperative. These tools may use heuristics to generate approximations to the optimal design, because optimisation problems in constraint-based routing for MPLS (and for the corresponding problems for native IP routing) can have high computational complexity.

5.2.3 Rapid Traffic Restoration

The ways of detecting a link or node failure do not change substantially when MPLS is adopted. However, RSVP 'keep alive' (or "hello") messages can be configured to be used for failure detection instead of their IGP counterparts.

The ways of restoring traffic after such a failure do however change somewhat when MPLS (or, strictly, MPLS-TE) is adopted. MPLS-TE fast rerouting (which is also known as 'fast reroute') makes use of predetermined backup LSPs so that, as with the predetermined alternative paths for native IP routing, restoration times comparable with those for SDH can be achieved. Only explicitly routed LSPs are protected in this way. In more detail the approach depends on the following:

- ❐ On detecting a failure, an LSR makes its interface cards switch to using backup LSPs. It thereby performs a local repair that diverts traffic around the failure.
- ❐ On completing a local repair (in perhaps 50 ms - 100 ms), an LSR may attempt to determine and establish new routes that are in some sense better than those used for

the local repair. This process may take some seconds. It is performed most effectively by the LSR at the ingress to the protected LSP (not at every point of local repair to the LSP), because that LSR alone has a global view that can provide a global repair. Optimising the routes for backup LSPs in general is extremely complex, because it involves considering not only constraints such as the maximum bandwidth on links but also the available routes in the transmission layer and other link layers and which kinds of LSPs are to use which kinds of backup LSP.

❒ Extensions to RSVP, currently being drafted, define extra RSVP objects; these allow an LSR to establish LSPs that are either 'detour' LSPs or 'bypass' LSPs. A detour LSP acts as a backup for one LSP; it is merged with the protected LSP upstream from the point of failure. A bypass LSP acts as a backup for several LSPs that have similar protection requirements; it encapsulates the protected LSPs using MPLS label stacking and removes the outermost label upstream from the point of failure. The choice between detour LSPs and bypass LSPs depends on the network size and topology; it can also depend on the choice of router, because some routers support only one kind of backup LSP, not two.

5.2.4 DiffServ Interworking

Basic MPLS does not ensure that the different traffic flows aggregated in an LSP receive different treatments; traffic flows are mapped into LSPs according to their Forwarding Equivalence Classes (FECs), which are essentially groups of packet flows destined for the same node at the edge of the MPLS domain. MPLS-TE can ensure that there is enough bandwidth for aggregate traffic flows in different LSPs; however, like basic MPLS, it does not provide traffic differentiation.

DiffServ provides an approach to satisfying performance requirements that is flexible and scalable (as it does not require flow-by-flow signalling and state). However, it must be supplemented by other mechanisms that influence the routes taken by packets if there are to be bandwidth assurances independent of how the network is affected by congestion or failures.

DiffServ and MPLS can be related in the following ways:

❒ Interworking DiffServ with MPLS allows the performance requirements satisfied by PHBs in a DiffServ domain for aggregate traffic flows based on DSCPs to be extended into an MPLS domain for aggregate traffic flows based on EXP fields and labels.

❒ Interworking DiffServ with MPLS and using MPLS-TE for routing control allows the performance requirements to be accompanied by bandwidth assurances for aggregate traffic flows based on DSCPs.

Interworking DiffServ with MPLS is possible with two kinds of LSPs, as indicated in section 4.2; traffic flows can be mapped into these LSPs according to their DSCPs as well as according to their FECs. An LSP of either kind may contain multiple aggregate traffic flows in different traffic classes, configured thus:

❒ EXP-inferred-PSC LSP ('E-LSP')

An E-LSP identifies the traffic class of a packet from its 3-bit EXP field (and uses the label just to indicate the FEC); for each traffic class there is a corresponding PHB (including the scheduling and discard priority to be applied). An E-LSP can distinguish between at most eight PHBs.

❒ Label-only-inferred-PSC LSP ('L-LSP')

An L-LSP requires that the traffic classes carried by it form a PHB Scheduling Class (PSC) so that the PHBs differ only in the discard priorities assigned to packets. The label therefore implicitly indicates the PSC as well as the FEC. If there are explicit MPLS headers for IP packets, the discard priority of a packet can be provided by the EXP field (in which case the L-LSP can distinguish between at most eight PHBs in one PSC; otherwise the discard priority of a packet can be provided by 1-bit discard eligibilities or cell loss priorities in the underlying Frame Relay or ATM (in which case the L-LSP can distinguish between at most two PHBs). (In practice there may be no need for an L-LSP to distinguish between more than three PHBs, as each of the PSCs currently defined for DiffServ has at most three PHBs.)

Extensions to RSVP, documented in RFC 3270, define extra RSVP objects; these allow an LSR to establish LSPs that are either E-LSPs (in which case it specifies the mappings between EXP fields and DSCPs) or L-LSPs (in which case it identifies the PSCs). In the absence of such specifications, LSPs can be taken to be E-LSPs using a preconfigured mapping between EXP fields and DSCPs; consequently E-LSPs are readily supported in routers.

Using MPLS-TE for routing control allows bandwidth to be reserved for LSPs. However, bandwidth reservation for LSPs is not directly related to bandwidth assignment by LSRs to their queues, because one traffic class may be present in several LSPs and several traffic classes may be present in one LSP. Bandwidth reservation for LSPs can be used to provide admission control for the LSPs, but it is a supplement to the PHBs given by DiffServ, not a replacement for them.

Interworking DiffServ with MPLS can be combined with using MPLS-TE for routing control by making each LSP contain only one traffic class (or actually one PSC). However, doing this can take considerable effort when many active and backup LSPs have to be established. (Of course, the effort can be reduced by using native IP routing for certain traffic classes, according to rules like those in section 4.3, or indeed by noting that MPLS is intended principally as a technology for core networks, in which traffic differentiation has questionable value, for reasons indicated in section 5.1.) Moreover making each traffic class be present in only one LSP at an LSR interface requires either merging LSPs or limiting severely the points of ingress to the MPLS domain.

The full integration of DiffServ with MPLS-TE calls for adapting the mechanisms of MPLS-TE to operate on individual traffic classes and thereby let bandwidth be assigned to queues. RFC 3564 identifies requirements for the enhancements to MPLS that are needed. Extensions to link state IGPs, currently being drafted, define extra LSA fields; these allow a router to apply shared bandwidth constraints to multiple aggregate traffic flows in different traffic classes. Extensions to RSVP, currently being drafted, define extra RSVP objects; these allow an LSR to establish LSPs that admit multiple aggregate traffic flows in different traffic classes subject to shared bandwidth constraints.