

# VoIP Network Design for Service Providers

---

A look at the challenges involved in  
designing a converged network to  
support VoIP and other applications

---

How service providers can manage migrating existing  
voice services or building new managed IP networks

This white paper addresses:

- The complexities of VoIP network design
- Designing a framework to support VoIP migrations
- The unique performance, restoration and reliability requirements of voice services



# Contents

Introduction .....	3
Framework .....	3
IMS Architecture .....	4
Service Provider VoIP Architecture .....	5
A Multi-Service MPLS Network .....	6
Multiple Classes of Service .....	6
Multiple Domains .....	7
Bearer Network Design .....	9
Generic MPLS Network Design Process .....	9
Supporting VoIP in Generic MPLS Network Design Process .....	12
Design Steps .....	14
Signaling Network Design .....	15
QoS Design .....	17
The Integrated Services (IntServ) model .....	17
The Differentiated Services (DiffServe) QoS model .....	18
Supporting Intra-Enterprise VoIP Traffic .....	20
QoS for Signaling Traffic .....	20
Conclusions .....	21
Acronyms .....	22
About the Authors .....	23
Jayant G. Deshpande .....	23
David J. Houck .....	23

## Introduction

Driven by competitive pressure and the desire for new service generating opportunities, major Service Providers (SP) have begun rolling out VoIP (Voice over IP) services to business and consumer markets. The challenge is to design the network to support VoIP service requirements for strict connectivity, latency, jitter, packet loss, and reliability objectives that are normally expected from the (circuit-switched) PSTN services. In addition the network design must support new voice applications – that are made possible by the new converged voice/data network.

SPs are either migrating voice services from the circuit switched networks to their existing data networks, or building new managed IP networks to support voice and other applications. In all cases, these data networks need to be carefully designed to support their most important and challenging application – VoIP.

Some of the key components of network design for supporting VoIP are:

- Topological design: Core network topology, softswitch and router placement, and backhaul. Design for reliability and scalability.
- Capacity design: Softswitches, routers, facilities, servers, and other network elements to support voice traffic and signaling.
- Signaling network design: Interconnecting voice end points independent of their access arrangements and corresponding signaling protocols (*e.g.*, SIP, SS7, H.323, H.248/Megaco, MGCP).
- QoS design: For end-to-end quality of service objectives for latency, jitter, and packet loss. Traffic policing, queuing, and shaping.

In a few cases, the addition of VoIP traffic may not require significant modifications to the existing data network topology or router capacities, but signaling networking design and QoS designs represent new elements in traditional data network design.

Network management and network security are integral parts of offering any new service and must be considered; however, they are not specifically addressed in this document.

## Framework

Designing a SP converged network is a complex endeavor since the SP must carry voice and data traffic from many customers with varying QoS requirements for multiple classes of service including VoIP and support access from a variety of access networks.

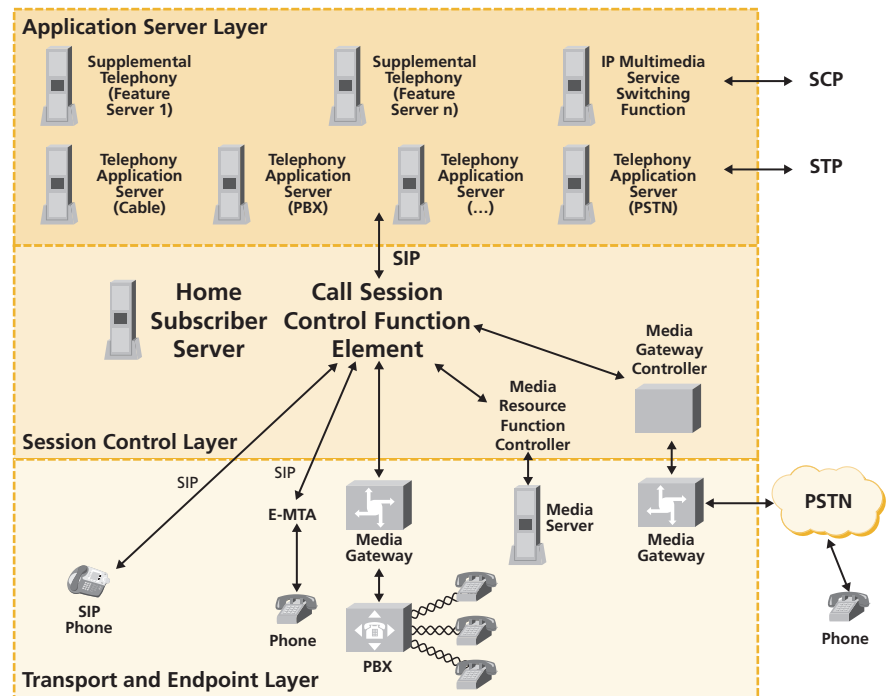
We begin with a description of the SP architecture which is the most important input to a network design. This includes not only the connectivity of network elements but also the functional relationships between the network and signaling elements and the corresponding protocols.

First, we set the context with a generalized architecture known as the *IP Multimedia Subsystem (IMS)*<sup>1</sup>.

<sup>1</sup> For an overview of the IMS architecture, see the White Paper, *IP Multimedia Subsystem (IMS) Service Architecture*, by Lucent Technologies, January 2004. ([http://www.lucent.com/livelihood/090094038005df2f\\_White\\_paper.pdf](http://www.lucent.com/livelihood/090094038005df2f_White_paper.pdf))

## IMS Architecture

The IMS architecture was created jointly by the 3<sup>rd</sup> Generation Partnership Project (3GPP), European Telecommunications Standards Institute (ETSI), and Parlay Forum. As shown in the simplified Figure 1, IMS is a three layer architecture developed around SIP-based communication between the VoIP and other telephony and non-telephony component systems with connectivity to existing voice network technologies and legacy systems supporting these technologies. Additionally, the architecture supports convergence of wireline and wireless services.



**Figure 1 Simplified IMS Architecture**

Functionally, the voice end-points and the media gateways in the *Transport and Endpoint Layer* are controlled from the endpoint logic in their (respective) Telephony Application servers.

The middle *Session Control Layer* includes the *Call Session Control function* of registration of endpoints and routing of the SIP signaling messages. The user profiles are stored at the *Home Subscriber Server*. The signaling layer also includes the control functions for specific media gateways (such as the trunk gateways) and media servers (such as the announcement server).

The Specialized Telephony servers, sometimes called Feature servers, are also at the Application server layer. Similarly, connectivity to the SS7 network to the STPs and the SCPs is provided through the respective (signaling) gateways in the Application Server layer.

It will be some time before the abstraction of Figure 1 is realized through vendor products. Based on the currently available vendor technologies and near-term expectation, we use the following terminology that is closer to vendor realization of products:

The term “softswitch” is generalized in this paper to include at least one Telephony application server responsible for the domain of end users that it controls. Depending on the domain of its end users, the softswitch may also contain the media gateway controller and/or signaling gateway.

The Supplemental Telephony servers will be denoted by their specific functions (such as unified messaging server) or simply as feature servers.

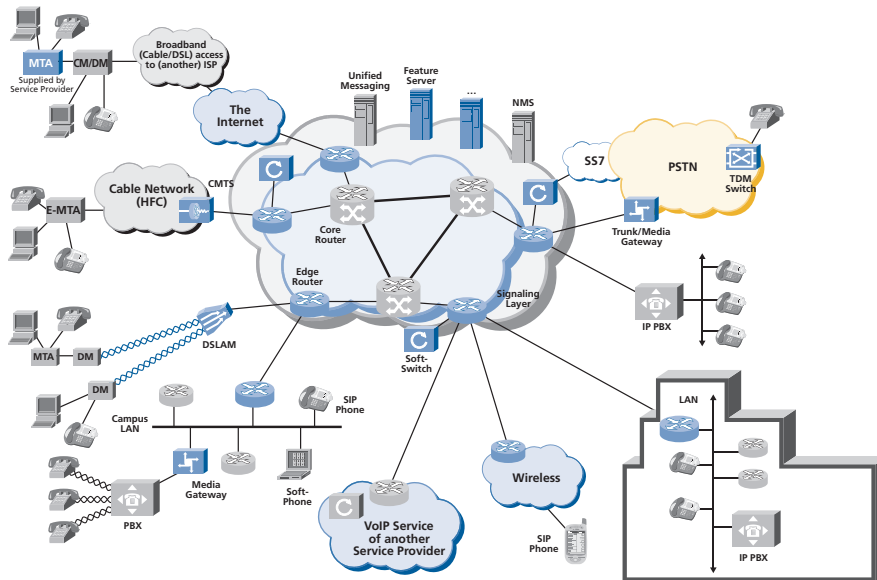
Depending on the signaling entities or endpoints connected to it, a softswitch must be able to communicate over one or more of the signaling protocols: SIP, ISUP, H.323, H.248/MEGACO, MGCP, etc. Some of the examples of softswitches defined here are: a combined system of connection and signaling gateway for connecting to PSTN, the PacketCable™ CMS, or a SIP based Session border controller (SBC) or the “so called” border element.

Over time, much of the communication between the signaling elements is expected to be based on SIP, but in the interim many legacy protocols will need to be supported.

The softswitch may be distributed over several physical systems.

### Service Provider VoIP Architecture

A Service Provider architecture for VoIP is shown in Figure 2.



**Figure 2: VoIP Architecture**

VoIP is one of the many services that the SP offers over its IP infrastructure (such as Internet access, MPLS VPN<sup>2</sup>, L2 VPN, and Video). However, much of Figure 2 emphasizes the networking connectivity for VoIP services.

<sup>2</sup> As defined in RFC 2547bis.

## **A Multi-Service MPLS Network**

Access to VoIP services and other data services is provided at the edge routers, strategically located by the SP in its area of operations. Often, an edge router supports access to multiple services. But, that is not necessary, either because a particular router technology is not able to support access to all SP services or the SP may provide access to a particular service from dedicated edge routers for security, stability, and/or ease of manageability.

<sup>3</sup> For small Service Provider networks, it is not necessary that there be a network of separate core routers as shown in Figure 2. The edge routers can be directly connected with each other. The edge routers will perform the functions of both a Label Edge router (LER) and a Label Switching Router (LSR).

MPLS is becoming the core technology<sup>3</sup> of choice for SPs since MPLS has many advantages over pure IP connectivity for edge router interconnection supporting multiple services. Therefore, even though VoIP does not specifically require it, it is recommended that MPLS technology be used as the core of the IP infrastructure connecting the edge routers.

- With MPLS, each edge router is only one (layer-3) hop away from another edge router, thus there is no layer 3 lookup at the intermediate routers.
- Separation of control plane and data forwarding plane is inherent to MPLS. Thus, vendors implementing these tasks from separate processors result in router stability: failure of control plane does not affect data forwarding for the existing flows through the router.
- MPLS supports fast reroute options that can provide improved reliability for VoIP.
- Many L3 and L2 VPN services are easy to deploy on a common MPLS architecture.
- If the network uses an edge-core architecture as in Figure 2, additional network stability is achieved by not requiring storing of the customer routes or the Internet routes in the core routers.
- MPLS can support enhanced traffic engineering by establishing multiple paths between a pair of edge routers for reliability and finer QoS treatment by class of service.
- Additionally, MPLS supports improved network security and management by enabling the use of separate LSPs for management, signaling and bearer traffic, if necessary.

The contents of this paper is still applicable for SPs who do not deploy MPLS core technology for VoIP service. The impact of VoIP traffic on the generic design (Section 3), all of the signaling design (Section 4), and the need for differentiated service for QoS (Section 5) are relevant even if MPLS is not chosen as the core technology (eg, end-to-end QoS can be supported using only the DSCP markings).

## **Multiple Classes of Service**

Providing multiple services from the same infrastructure necessarily implies classifying the traffic according to each individual service need and delivering to customer expectations (through SLAs) for that service.

For the VoIP service, maintaining voice quality requires low latency, low jitter (small delays variation between successive VoIP packets at the destination), and low packet loss.

Queuing delays at all network nodes must be minimized for the voice packets. As far as possible, the end-to-end latency of the voice packets should approach the unavoidable propagation time of the path taken by the VoIP packets.

To account for the network jitter, a *jitter buffer* is used at the destination. However, queuing delays in the network could result in unacceptable values for jitter even if the average end-to-end latency is acceptable. Thus, a late-arriving packet is equivalent to a lost packet, if it is not available at the *play-out* time.

There also may be additional packet loss in the network due to traffic congestion resulting in large queues at the routers.

In summary, VoIP traffic must be given preferential QoS treatment in the network to maintain the required voice quality.

The service provided may maintain several additional classes for its non-VoIP traffic to render each class its required QoS treatment (*e.g.*, network control traffic, video, critical business data, best effort data).

Section 5 will deal with QoS design.

### **Multiple Domains**

In addition to providing the VoIP service to its own customers, the SP must provide connectivity to voice services (both VoIP and circuit switched) of other SPs including connecting to the Public Switched Telephone Network (PSTN). In fact, for any voice call carried over the SP network, one or more end points of the call may not be the customers of that SP.

A domain may be loosely defined as a collection of voice endpoints and one or more signaling entities. The signaling entities within a domain may communicate with each other for *on-net* calls whereas signaling entities in different domains must communicate with each other for inter-domain calls –the *off-net* calls.

The SP IP network carries the “bearer” VoIP traffic (that is made up of the IP packets carrying encoded voice using the RTP protocol) and the signaling traffic between the signaling entities based on the signaling protocol. Both these traffic streams of IP packets enter at the edge routers of the network.

In Figure 2, the SP has deployed several softswitches (that includes Telephony servers for the respective domains, feature servers and media gateway controllers as necessary) in its network.

### Service Provider Domain:

The SP domain interconnects many *access (sub-) domains* over the IP infrastructure.

- Voice connection for telephones connected to IP PBXs are controlled by a softswitch over SIP or H.323 connectivity to the PBX.
- Traditional PBXs connect to the enterprise media gateways. Voice encoding is performed at the gateway and H.323 (or lately, SIP) is generally used for signaling between the media gateways and an SP softswitch.
- IP phones and soft phones are directly controlled by a softswitch using SIP protocol over IP; we will call these SIP phones
- The SP may offer voice connectivity to subscribers from different enterprise customers for intra-enterprise voice communication. The intra-enterprise communication may be over L3 or L2 VPNS, but that is not necessary.
- If the SP offers (or collaborates with another provider to offer) broadband services to end users over cable, small business, remote users of large enterprises, and residential customers can be provided VoIP service over the cable connection.
- VoIP service over cable has been defined in the PacketCable™ standards from Cable Labs®.
  - The customer connects telephone(s) into the telephone jack(s) of PacketCable-compliant Embedded Multimedia Terminal Adapter (E-MTA) for VoIP connectivity. The PC/router connects into the Ethernet port of the E-MTA. The E-MTA has an integrated cable modem that connects to the CMTS over a DOCSIS-compliant IP connection carrying both the voice and data traffic from the E-MTA<sup>4</sup>.
  - The PacketCable-compliant softswitch, called Call Management Server (CMS), controls the E-MTA for Voice calls over Network-based Call Signaling (NCS) protocol that is based on MGCP.
  - The CMTS provides differentiated treatment to VoIP using the PacketCable Dynamic QoS (DQoS) standard.
- Recently, some SPs have been offering VoIP service to broadband service subscribers (Cable or DSL ISPs). Note that these ISPs need not have any relationship with the VoIP SPs. The VoIP connectivity over Cable does not follow the PacketCable standards. Further, the VoIP traffic may be carried over the Internet (including the cable/DSL access) before being forwarded to the SP. The SP supplies a (non-PacketCable-compliant) MTA – often called a phone adapter or telephone adapter – to the VoIP subscriber. The customer connects the telephone(s) and PC/router into the MTA telephone and Ethernet ports. The MTA and the phone together act as a SIP phone and are under the control of the SP Softswitch that is a SIP control element
  - This MTA does not include a cable modem or DSL modem embedded in it. This MTA connects into the (existing) cable of the DSL modem.

<sup>4</sup> The standard allows for an MTA to be a stand-alone element (without cable modem function) connecting into a regular DOCSIS modem. However, stand-alone PacketCable-compliant MTAs are not common.

- Many new mobile phone vendors' products are supporting SIP technology. It will be possible for the SP to offer VoIP services in collaboration with the mobile telephone services.
- The SP may also offer many other value added services to its customers such as IP Centrex or voice VPN (not shown in Figure 2).

#### **Connecting to PSTN:**

The SP VoIP service must be connected to the PSTN, since PSTN is the most prevalent voice service.

- PSTN voice is converted to IP at the media gateway (also called trunk gateway) that may be located at a PSTN central office or the SP location. These gateways connect to the edge routers.
- The SP softswitch also connects into the SS7 network for ISUP (signaling) connectivity to the PSTN (TDM) switches and other PSTN signaling entities.
- It is assumed here that the softswitch includes both the signaling gateway and media gateway control functions. It is possible that these two functions are delivered from different network elements.

#### **Connecting to VoIP Services of other Service Providers:**

The SP may connect to external VoIP domains.

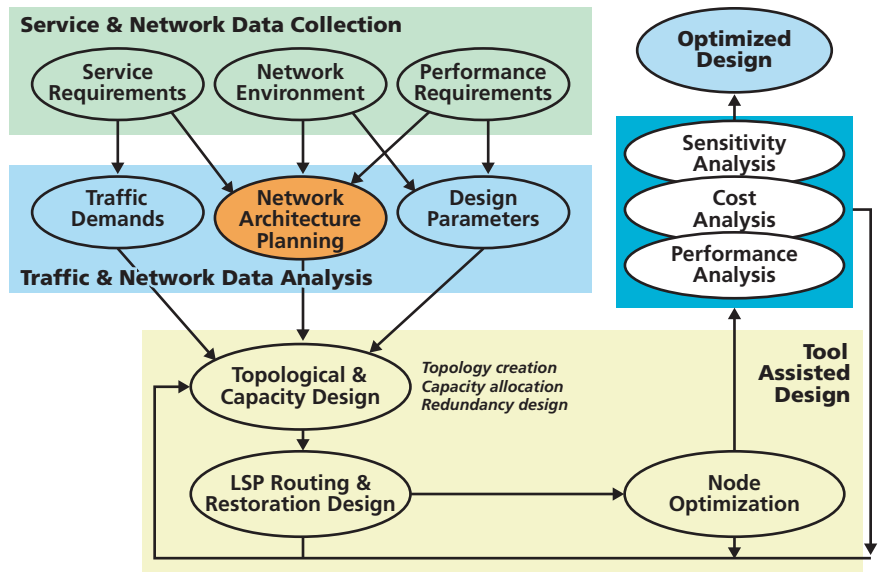
- The VoIP services from different SPs can be connected allowing end users of these VoIP services to communicate with each other.
- The SP's IP networks connect with one another over an inter-AS (Autonomous System) connection.
- There are no standards for interconnecting VoIP services of two SPs. Such interconnection is possible only through mutual understanding and agreements. In particular, the two services need to agree upon the signaling protocol between their respective softswitches.

## Bearer Network Design

The VoIP bearer traffic is only a part of the overall IP traffic on the SP MPLS/IP network. Therefore, the network design for capacity and topology of the network has to be based on all traffic and not just the VoIP bearer traffic. When the SP is only looking to migrate voice services from the existing network, the overall network capacity and topology design may be an incremental activity for support of the VoIP traffic.

#### **Generic MPLS Network Design Process**

A generic design process for a multi-services MPLS network is illustrated in Figure 3.



**Figure 3 Generic IP/MPLS Design Process**

A brief description of the modules in Figure 3 will be presented here. Note that implementing VoIP has varying degree of effect on individual modules.

#### Service and Network Data Collection

- **Service Requirements:** The service requirements are usually specified in terms of SLAs including QoS parameters. These requirements help determine the number of traffic classes that must be supported. For VoIP service, the types of codecs used by the endpoints may also be specified. Choice of codecs has an impact on estimating the bearer traffic.
- **Network Environment:** Evaluate the existing network environment including the network elements already deployed. It is important to understand how much of legacy equipment must be retained as well as any issues related to multivendor inter-operability.
- **Performance Requirements:** For each supported service, there may be performance requirements on the network including end-to-end latency and/or packet loss in the network. Services performance requirements may also include the network restorability and availability objectives. Additionally, voice services will need to meet the goals for components of call set up time.

#### Traffic and Network Data Analysis

- **Traffic Demands:** Traffic demands must be calculated from the service requirements. Depending on the tools used, the traffic demands need to be specified between every pair of endpoints (pipe model) or traffic to/from each endpoint (hose model).

- **Network Architecture Planning:** The service requirements, network environment, and performance requirements help determine the detailed network architecture that is at the heart of the network design. These requirements include network hierarchy, interconnection of network components and the necessary function of each component for end-to-end support of the required services.
- **Design Parameters:** Some of the design parameters required to control the behavior of the network design procedure are network level performance and reliability requirements, facility types and their ownership, possible vendor equipment model types, restoration bandwidth requirement, and hop count. The network environment and performance requirements help determine some of these design parameters.

#### Tool Assisted Design

- **Topological and Capacity Design:** Based on these inputs, the topological and capacity design can be carried out (with the help of tools) to determine the number, size, and locations of routers, switches, and links and the optimal access arrangement including backhaul, as necessary. The output of the design process can also help choose among many alternatives; *eg*, lease or own the facilities, oversubscription values. A few details of the topological and capacity design are presented later.
- **LSP Routing and Restoration Design:** Network design must allow for efficient traffic engineering. This is particularly important for support of the voice traffic with its stringent performance and reliability requirements. Traffic engineering of the design will help determine traffic routing for various services such as VoIP and MPLS VPN services. Further, by mapping each individual demand to (primary) LSPs, the bandwidth of each link can be utilized to maximize the overall network bandwidth efficiency for the given traffic pattern in the design. Restoration bandwidth can be added through implementation of the secondary LSPs as appropriate.
- **Node optimization:** Often, the network design will result in many smaller similar devices (routers/switches) being collocated. Optimizing the nodal equipment (into larger devices) may result in removal of unnecessary intra-node traffic and associated complexity such as unnecessary adjacencies.

#### Optimizing the design

Reaching optimized design requires the following three analyses and iterating over the design to reach optimality (See Figure 3).

- **Performance and Reliability analysis:** Evaluate the design against performance and reliability objectives at the network level as well as at the packet level. For VoIP, it is also necessary to meet the goals at the call level.
- **Cost analysis:** The topological and capacity design helps estimate the capital and operations cost estimates for the SP.
- **Sensitivity analysis:** Sensitivity analysis is about evaluating the *what-if* scenarios for changes in the values of some of the network parameters, such as fluctuations in the traffic demands. A good design requires that a small fluctuation should not have disproportionate effect on the rest of the network.

## Supporting VoIP in Generic MPLS Network Design Process

The need to support VoIP has the most impact on the following aspects of the generic design process for designing a multi-services MPLS network

- Voice traffic estimation
- Performance
- Restoration and reliability

### Estimating VoIP traffic

Although G.711 (64 kbps) is a common codec choice, various voice compression options, notably G.726, G.729 and G.723.1, allow for codec bit rate ranging from 32 kbps to 5.3 kbps per voice channel. Thus, instead of dedicating 64 kbps for each voice channel in circuit-switched network, carrying voice over an MPLS network with compression and silence suppression – resulting in voice bandwidth lower than 64 Kbps – enables statistical multiplexing and allows for better utilization of bandwidth.

Traffic demand calculations for VoIP over MPLS often start with the traditional PSTN descriptors characterized by point-to-point traffic demands and engineered link blocking probabilities. The SP can estimate the average (busy hour) VoIP traffic demand at each edge router based on the knowledge of the number of subscribers connecting through that router, codecs used, busy hour call attempts (BHCA), average call length, etc. Then tools such as the Erlang B formula can be used to calculate the number of simultaneous calls.

Irrespective of the codec, there are layers 1, 2, and 3 overheads. For each VoIP packet, there will be 40 bytes of the IP/UDP/RTP header, 4 bytes of the MPLS header and the layer 2 overhead (*e.g.*, 6 bytes for PPP).

(Note that if there are multiple MPLS headers required, such as in the case of MPLS VPNs, hierarchical MPLS topology, etc., each MPLS header will require an additional 4 bytes).

Depending on the codec used, the VoIP traffic payload in the VoIP packet will be of different sizes. Frequently, the voice payload in each VoIP (RTP) packet is worth 20 ms of voice.

As an example, for the G.711 PCM codec, the required bandwidth for each call in each direction will be 64 Kbps for the voice packets plus about 31% of layer 2 and layer 3 overhead resulting in about 84 Kbps (160 bytes payload + 50 bytes header with PPP); whereas for G.729a CS-ACELP codec, the required bandwidth for each call in each direction will be 8 Kbps for the voice packets with 230% of “IP tax” resulting in about 28 Kbps (20 bytes of payload +50 bytes of overhead).

The voice bandwidth calculations become further complicated depending on the layer 1 connectivity in the network including allowing for the necessary separation between the successive packets. For example, for SONET links, the SONET framing overhead must be added. Additionally, Packet over SONET (POS) links may require increased bandwidth to

account for byte stuffing. (A case in point is the use of G.711 codec without silence suppression. The byte pattern for the “quiet time” is often the same as that of the POS flag and control bytes. Thus, these quiet bytes must be *escaped* with byte stuffing, thus increasing the number of bytes that must be transmitted).

Tools are available that will help calculate the bandwidth requirements for different codec types and protocols used and a packet loss requirement with silence suppression enabled.

If the SP offers *Unified Messaging* service, including voice mail, the demand for such traffic should be included in the overall voice demand. This traffic can be modeled either as just additional voice calls to the server or it can be separately estimated and added to the other VoIP bearer traffic. Unified messaging may also require non-voice data communication between the server and user PCs. Estimates of that data traffic, if considered significant, must be added to the overall data traffic demand on the SP network

#### **Network Performance for Voice traffic**

Sections 4 and 5, that deal with Signaling and QoS design respectively, will address the signaling performance and voice quality performance in finer detail. But it is also possible to incorporate some of the VoIP performance goals in designing the MPLS network, so that the signaling and QoS performance can be efficiently implemented.

For example,

- The “voice” LSPs may be constrained to be carried over a limited number of hops to help improve the delay performance.
- Later we will see that a certain amount of bandwidth is dedicated for voice traffic at each link. It will help if the network design itself is able to allocate higher bandwidth links wherever possible. Consider the case of an SP where the same central office connects the SP to the IP network as well as to the PSTN. SP network links connecting to the edge router in the central office (CO) should be of a higher bandwidth than dictated by design tools, so that it is possible to allocate higher voice bandwidth on these links to improve the performance.

#### **Network Restoration and Reliability for Voice traffic**

Typically, the reliability and restoration requirements for voice services are higher than for data services. Customers still expect a “five 9s” reliability (*i.e.* 99.999% service availability) as provided by the traditional circuit-switched PSTN. Mere addition of routers, softswitches, and links may not be enough for voice service reliability. Restoration methods such as MPLS “fast reroute” must be implemented. It is important that sufficient bandwidth is allocated at the design time for supporting fast reroute for voice traffic. (Note that for the MPLS core, many SPs have opted for the rerouting capabilities of the IGP rather than investing in SONET restoration).

## Design Steps

As indicated earlier, the design is based on the consolidated VoIP and other data requirements.

- **Geographical Locations:** The SP has a good idea as to the COs where the edge routers can be placed based on the customer coverage areas. In some cases, the SP may want to backhaul traffic from its remote offices to these COs. If the number of edge router locations is small, there may not be a need for the core network. Otherwise, the hierarchical edge-core design must be considered. Typically, the core router locations will be a subset of the edge router locations.

Since a significant amount of bearer traffic flows through the media gateways, their placement must be carefully chosen. The media gateways for connecting to the PSTN must be closer to PSTN connection point(s). (The media gateways do not have to be collocated with the corresponding MGCs or SGs). If the media gateways connecting the business customers' PBXs over private lines/PRIs are to be placed in SP locations, such locations should also be identified.

Iteratively, the design tools will help converge to the optimum geographic locations for the edge and core routers (and backhaul if needed) based on costs, performance, reliability, and routing design.

- **Design parameters and constraints:** The design must satisfy the performance (*e.g.*, QoS treatment for VoIP, number of hops), reliability (*e.g.*, PSTN-like reliability and survivability), and scalability (*e.g.*, say, little topology or capacity change with 10% additional demand, multi-year design, support for future protocols) goals specified by the SP including VoIP-related requirements. There may be constraints on the network facility choices (SONET/SDH, DWDM) or their ownership (leased or constructed).

Choice of the internal gateway protocol and implementing its hierarchical implementation has considerable influence on the network topology. MPLS rerouting to support VoIP goals also affects the network design. Note that the choice of hierarchy of the Interior Gateway Protocol may dictate the edge-core design.

- **Data Analysis:** In most cases, these traffic estimates will be in "hose mode" where, the aggregate traffic into an edge router bound for any destination is estimated. The design process requires that the traffic demand be computed for every pair of edge routers ("pipe mode"). A gravity matrix model can be used to compute the required point-to-point traffic matrix.

Generally, the aggregate VoIP traffic is the same in each direction on a link.

In spite of the possibility of including significant amount of text in the signaling messages in new signaling protocols like SIP, it can be assumed that the total signaling traffic carries by the IP/MPLS network is small in comparison to the total VoIP traffic and certainly miniscule compared with the overall IP traffic over the network. Thus, the signaling traffic volume can be ignored for the capacity and topology design.

- **Topology generation:** A network design tool is essential for larger networks to provide the network topology and capacity that supports the performance and restoration requirements for each class of service and minimizes the cost of the network. For smaller networks, some heuristics will work reasonably well.

The first step is to generate the basic topology without redundancy.

Then modify the topology to support the performance, QoS and routing objectives. Then generate a modified topology that will support the required redundancy, survivability, and restoration.

The design can be further “tweaked”; *e.g.*, the links can be resized to available bandwidths, backhaul links removed/added, if there are multiple edge routers in a CO, their interconnectivity within the CO can be optimized.

## Signaling Network Design

A SP may deploy many softswitches in the network either because it needs to support a large number of end-points or because any one softswitch product may not support all the needed signaling protocols, or both.

Every pair of softswitches needs to exchange signaling information with each other for connecting calls of end points they control. Additionally, the softswitches must communicate with the Feature servers for call features other than those required for a simple voice connection.

As far as possible, a single protocol must be used for communication between the softswitches and feature servers deployed by the SP. Session Initiation Protocol (SIP) does support the features for this communication and is currently favored by most SPs.

Note that there needs to be a full mesh of connectivity among all softswitches. Generally, this cannot be avoided even for a large SP with many softswitches.

The main objectives of the Signaling design are:

- Support the required BHCA objective for all VoIP connections carried over the SP network.
- Support the call set up time and call disconnect time objectives.
- Support the service availability requirement by providing redundancy in signaling connectivity.

As indicated earlier, there is seldom any need for considering the actual signaling traffic volume between the softswitches and other signaling entities since the signaling traffic volume is a very small percentage of the total IP traffic carried in the network. This is the case even when signaling traffic may carry some user data such as that used in the popular SMS (Short Messaging Service).

A high level procedure for designing the Signaling network follows:

- Determine the signaling protocols that must be supported by the SP.
  - For PSTN, the SP will need to support connectivity to the SS7 network over ISUP as well as MGC functionality.
  - All softswitches must support SIP, which is a protocol of choice for communication between the SP's softswitches, Feature Servers, and other signaling points. This will also help the evolution towards the IMS architecture.
  - Additional protocols such as NCS, H.323, MGCP, and Megco/H.248 will also need to be supported depending on the needs of the access domains.
- Decide on the vendor products that may be deployed in the network to support these protocols.
  - Include all existing softswitch products already in use in the network (such as the CMS in the Cable network) that must be continued to be used in the signaling design.
  - Subject to unit cost constraints, select additional softswitch products that support multiple signaling protocols.
  - Carefully evaluate the support for redundant configuration of softswitch deployment: primary/secondary, active/standby, state replication, *etc.*
  - It may be prudent to distribute certain VoIP functions based on flexibility, security, and cost considerations. For example, it may be necessary to use a vendor product that is deployed only as a SIP proxy, for connecting the SIP endpoints.
- Determine the number of softswitch vendor units required to support the estimated BHCA from all domains.
  - The first rough calculation will add the BHCA for all domains with signaling protocols supported by each selected vendor model and divide that aggregate estimated BHCA by the BHCA capacity<sup>5</sup> of that vendor model resulting in the number of softswitches of that type. Repeat for each vendor model. (Do not double count the demand if a particular signaling protocol is supported by multiple vendor models).
  - Refine calculations by removing the demand that can be supported by the existing softswitches.
- Provide for additional softswitch units for redundancy based on reliability calculations. At least n:1 redundancy must be provided for each softswitch managing the same (access) domain, since most SPs require diversity. Load sharing may be configured over two or more softswitches; however, the entire load must be supportable when one softswitch in the redundant configuration is out of order.
- Since all softswitches are connected with each other over a high speed IP network, they can be placed anywhere in the network with the following considerations
- Each softswitch must be connected to two edge routers. Often, the softswitches will be located in a CO where there is an edge router. The other connection should be to an edge router in another CO if possible.

<sup>5</sup> Vendor-supplied BHCA value may need to be modified based on experience, lab testing, and confidence in the vendor-supplied numbers.

- Further refine and finish the design with following considerations:
  - Generally, there should not be any concerns about latency in the signaling network, provided that signaling packets are marked properly and given priority over other data packets. Any softswitch should support domains that may be connected into a different part of the network. But if the call setup time or call disconnect time objectives are not satisfied because of high latency of the design, it may be necessary to deploy additional softswitches and rearrange the softswitch placement in different geographical areas.
- Evaluate the design against the network management connectivity requirements. Modify if necessary
- Evaluate the reliability of the design against objectives. Modify as necessary.
- Evaluate the design against regulatory policy compliance and SP's constraints about connectivity to the PSTN at various points in the network. For example, the SP may want to carry the off-net traffic to the nearest PSTN gateway.

## QoS Design

As indicated earlier, the SP must provide preferential QoS treatment<sup>6</sup> to the VoIP traffic to support voice quality; requiring low latency, low jitter, and low packet loss.

The VoIP traffic shares networking resources – both processing and transmission – with other data traffic on most access links and all of the IP infrastructure routers and links. The overall QoS design for the network must support end-to-end voice quality objectives for the bearer traffic as well as the design objectives for the signaling traffic.

There are two possible QoS models that the SP can use: integrated services (IntServ) model and differentiated services (DiffServe) model.

After briefly describing the IntServ model, QoS design using the recommended DiffServe model is presented.

### **The Integrated Services (IntServ) model**

End-to-end network resources are reserved for each VoIP flow using the RSVP protocol for preferential treatment to the voice traffic resulting in bounded values for latency, jitter and packet loss.

However, IntServ requires a high processing load in the routers and is not very scalable with the current technology for per call signaling. Use of RSVP-TE may still be used to create LSPs and for maintaining LSP rerouting. But, in that case, resource reservation requests are few and far between. Using IntServ for VoIP QoS will require frequent creations and destruction of “voice” LSPs.

PacketCable and some wireless standards have specified Integrated Services model for QoS

<sup>6</sup> Note that most vendor implementations give the highest priority for the network control traffic such as routing and vendor-proprietary protocols over all other traffic including voice. However, such network control traffic, being very small, should have little effect on the QoS performance for VoIP and other traffic.

## The Differentiated Services (DiffServe) QoS model

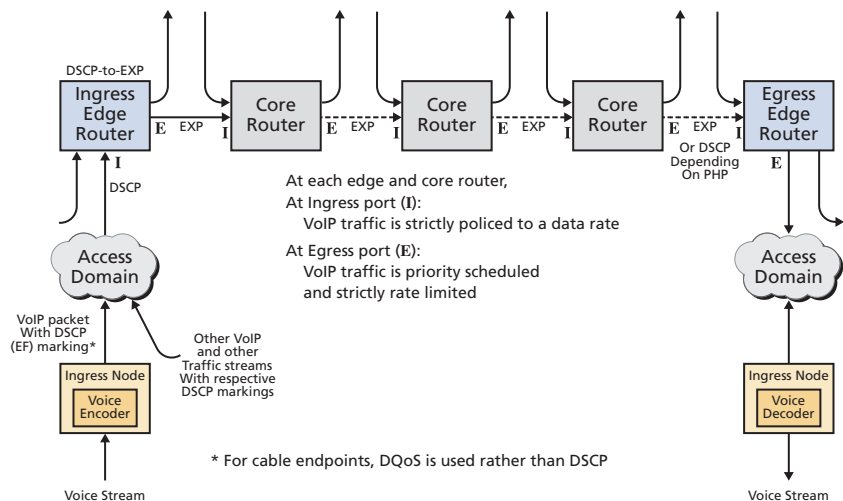
The differentiated services model renders each packet a QoS treatment based on the class of service of that packet at each hop through the network – the specific per-hop-behavior (PHP) is specified and configured at each router instead of the end-to-end resource reservation of the IntServ model.

The SP may support many classes of service including a *real-time* class for the VoIP bearer traffic, in addition to several *data* classes and the *best effort* traffic class. Packet classification markings / mechanisms for the corresponding classes varies depending on the protocol used at a particular network “hop”. Some of these classification markings/mechanisms are:

1. Differentiated Services Code point (DSCP) that is a 6-bit value in the 8-bit TOS byte in the IP header.
2. Three bit EXP value in the MPLS header of the packet
3. Dynamic QoS (DQoS) mechanism defined by PacketCable standards that differentiates voice packets from data packet at a Cable Modem Termination System (CMTS).

Generally, the SP should require that the IP packets entering the edge routers be marked with DSCP values corresponding to the classes of service (CoS) that the SP offers<sup>7</sup>. VoIP bearer traffic packets (real time packets) are usually marked with the DSCP value for the EF (Expedited Forwarding) treatment. (See Figure 4)

<sup>7</sup> In some instances, (eg cable access), the edge router may use other mechanism (eg, DQoS) to directly mark the EXP value based on the class of the incoming packets.



**Figure 4: DiffServe for QoS for VoIP traffic**

Data packets – of classes other than the best effort traffic – are marked with the DSCP value of the corresponding AF (Assured Forwarding) class. Best effort traffic is marked with the DSCP value of the BE (Best Effort) class.

Figure 4 illustrates the DiffServe QoS for the IP infrastructure. Note that at each ingress port of an edge router, VoIP and data streams from multiple sources enter the router. Similarly, VoIP and data streams from multiple sources exit the router for multiple destinations. Similarly for all the core routers.

The entering VoIP traffic and traffic of other classes at each port is the aggregate of the traffic from many sources. Similarly, the exiting VoIP traffic and traffic of other data classes at each port is the aggregate of voice traffic bound for many destinations. The QoS treatment at each of these routers is based on packet classification rather than on the source or destination.

The ingress router inserts the MPLS header in each IP packet received from the access domain. The router must map the packet DSCP value to the 3-bit EXP marking within the MPLS header. The MPLS header will be removed by the last core router before delivering the packet to the egress edge router<sup>8</sup>.

<sup>8</sup> Assuming penultimate hop popping (PHP). Otherwise, the MPLS header is removed by the egress edge router.

There are only eight possible EXP values out of which only up to six may be available for use by the SP for QoS. (Other values are often designated by the vendor for network control traffic including routing protocol packets and vendor-proprietary protocol packets).

We propose that all VoIP traffic be mapped to the same EXP value by the SP. At a high level, the following processing elements at the edge and core routers for VoIP packets will help support the VoIP latency, jitter, and packet loss objectives under most circumstances.

- **Incoming traffic:** The incoming aggregate VoIP traffic is strictly policed: The aggregate incoming rate of the VoIP packets at each port of the edge or the core router is limited to a configured bandwidth value based on the estimate of the VoIP packet rate at that port. Procedures for estimating traffic demand were covered earlier. The other (data) classes of traffic may also be policed to their individual rates, possibly with bursting allowed.
- **Outgoing traffic:** At each egress port of each edge or core router the VoIP traffic is placed in a single queue that is served with strict priority. Non-VoIP packets are not scheduled for transmission if there is any packet in this strict priority queue of VoIP packets. Further, the voice packets are shaped to a strict configured rate. The scheduling of traffic for other (data classes) may follow a variety of schemes, *eg*, weighted round robin with random dropping of the packets with drop probabilities based on the relative priority of those classes.

With the recommended differentiated services QoS architecture described above, the VoIP design parameters that need to be determined are the VoIP bandwidth that must be supported on every link of the infrastructure, which can be computed from the VoIP demand matrix that was computed in the topological design and the LSP routing through the network.

Supporting efficient QoS is a complex task. In addition to many nuances of polling and shaping policies, vendor options vary in the implementation of QoS policies. It is not uncommon to arrive at the desired result after trying out several networkwide configurations.

### **Supporting Intra-Enterprise VoIP Traffic**

The SP network carries VoIP traffic for calls managed by its softswitches as well as the intra-enterprise VoIP traffic that is completely controlled by the individual enterprise gateways. Traffic engineering for supporting these two classes of VoIP traffic requires careful planning.

The QoS design outlined earlier will treat the aggregate of these two traffic streams as one single class of VoIP traffic – use the aggregate policing/shaping rate and place the traffic in one single priority queue. For some SPs, this aggregate treatment may be appropriate.

However, a SP may want to separate the VoIP traffic attributed to the calls controlled by its softswitches and subject to its CAC policies from the intra-enterprise traffic that is not subject to the SP CAC. The SP may want the CAC based entirely on the availability of the bandwidth that is earmarked for the VoIP traffic for the calls controlled by its softswitches. On the other hand, the SP must still maintain the SLAs for voice quality for its enterprise customers even for their internal voice traffic. Therefore, these two VoIP traffic types need to be differentiated from each other at each router and yet support the SP CAC and customer SLAs.

Typically, these two classes of traffic will be strictly policed to their respective estimated rates at the ingress ports of the routers, whereas at egress the SP may allocate higher priority to the VoIP traffic controlled by its softswitches over the intra-enterprise VoIP traffic, but shaping both streams to their strict respective rates. The complexity of such packet scheduling schemes is not discussed here.

Since the VoIP traffic from the VPN customers belong to the respective VPNs, security issues must also be considered if there is to be voice communication between endpoints from two different VPNs.

### **QoS for Signaling Traffic**

There does not seem to be any agreement among network practitioners about the right DiffServe class and the associated treatment for the signaling traffic. All agree that the signaling traffic must be given preferential treatment with excellent delay and packet loss performance. That would suggest classifying the signaling traffic as real-time traffic (the same class as the VoIP bearer traffic). However, signaling packets vary in length, and with the advent of SIP and other networking protocols these signaling packets are increasingly becoming longer. Real-time classification for the signaling traffic does not bode well for the VoIP bearer traffic, since this QoS strategy will increase the end-to-end jitter for voice.

Assigning signaling traffic to the class of network control traffic, with precedence higher than the VoIP bearer traffic, has its merit in that the signaling traffic is extremely important for managing the voice service as the routing and other network control traffic is to running of the network.

But this classification for the signaling traffic will significantly increase traffic volume of higher priority than the VoIP bearer traffic. This will have adverse effect on the jitter for voice communication, if only slightly.

The third option is to assign the signaling traffic the highest “data” class just below the real-time class. Provided that VoIP bandwidth allocation has been diligently made larger than estimated, the signaling traffic will get the required latency and packet loss treatment in most cases, without affecting the jitter for voice communication. Rather than allocating a separate class for the signaling traffic, the signaling traffic is included in the same class as the highest non-VoIP data traffic class.

## Conclusions

SPs have begun rolling out VoIP services for business and consumer markets. That has posed network design challenges to meet the strict performance and reliability requirements for PSTN-like voice quality and service availability while supporting VoIP and other data services over a single IP infrastructure.

This white paper identifies essential aspects involved in including VoIP service in a multi-services environment and provides guidelines to meeting the design challenges for supporting the SP’s voice customers over a variety of access arrangements and connecting them to PSTNs as well as voice services of other SPs. The complexities of network design have also been brought in this paper.

We identify MPLS as the most appropriate core network technology for supporting multiple services of the SP and present the necessary modifications required to support VoIP bearer traffic in a generic network design for an IP/MPLS network. In addition, we describe procedures for estimating VoIP traffic and supporting the unique performance, restoration, and reliability requirements of voice services.

Signaling is an integral part of voice service. Elements of signaling design for capacity and placement of the softswitches are presented in this paper for the needed performance and reliability of signaling, so that the PSTN-like goals for call set-up times and service availability can be met.

Without QoS support, it will be impossible to sustain VoIP service in a multi-services environment over an IP infrastructure. The need for priority treatment of the VoIP traffic has been emphasized to meet the latency, jitter, and packet loss objectives for voice quality. Design guidelines for QoS treatment using the differentiated services model has been presented. Finally, QoS treatment for the signaling traffic as well as for supporting intra-enterprise VoIP bearer traffic have also been included in this paper.

## Acronyms

Acronym	Definition
AF	Assured Forwarding
BE	Best Effort
BHCA	Busy Hour Call Attempt
CAC	Call admission Control
CMS	Call Management Server
CMTS	Cable Modem Termination System
CO	Central Office
CoS	Class of Service
CS-ACELP	Conjugate Structure – Algebraic Code-excited Linear Prediction
DOCSIS	Data Over Cable Service Interface Specification
DQoS	Dynamic QoS
DSCP	Differentiated Services Code Point
EF	Expedited Forwarding
E-MTA	Embedded MTA
EXP	Experimental (bits)
IMS	IP Multimedia Subsystem
IP	Internet Protocol
ISUP	ISDN User part (ISDN: Integrated Services Digital Network)
LAN	Local Area Network
Megaco	MGCP according to H.248
MGCP	Media Gateway Control Protocol
MGC	Media Gateway Controller
MPLS	Multi-Protocol Label Switching
MTA	Multi-media Terminal Adapter
NCS	Network-based Call Signaling
PHB	Per Hop Behavior
PHP	Penultimate Hop Popping
POS	Packet over SONET
PRI	Primary Rate Interface
PSTN	Public Switched Telephone Service
QoS	Quality of Service
RSVP	Reservation Protocol
RSVP-TE	RSVP – Traffic Engineering
RTP	Real Time Protocol
SBC	Session Border Controller
SCP	Signaling Control Point
SIP	Session Initiation Protocol
SP	Service Provider
SS7	Signaling System 7
STP	Signaling Transfer Point
TDM	Time Division Multiplexing
VoIP	Voice over IP
VPN	Virtual Private Network

## About the Authors

**Jayant G. Deshpande**  
**Lucent Technologies– Bell Laboratories**

Jayant is a Network Design Engineer in the Network Technologies and Performance Department at Lucent Technologies Bell Laboratories in Holmdel, New Jersey. Jayant's work focuses on data and voice networking services development, network architecture, design, planning, performance analysis, and QoS. Jayant received his Ph.D. in Electrical Engineering from University of Texas at Austin in 1973.

**David J. Houck**  
**Lucent Technologies– Bell Laboratories**

David is a technical manager in the QoS Management and Assessment Group in Holmdel, New Jersey. David leads a team that focuses on performance modeling and traffic management of converged packet networks with QoS requirements. Dave received a B.A. in mathematics in 1970 and a Ph.D. in operations research in 1974, both from The Johns Hopkins University.

To learn more about our comprehensive portfolio, please contact your Lucent Technologies Sales Representative.

Visit our web site at [www.lucent.com](http://www.lucent.com).

This document is for planning purposes only, and is not intended to modify or supplement any Lucent Technologies specifications or warranties relating to these products or services. The publication of information in this document does not imply freedom from patent or other protective rights of Lucent Technologies or others.

Copyright © 2004  
Lucent Technologies Inc.  
All rights reserved

LWS VOIP 10/04

**Lucent Technologies**  
Bell Labs Innovations

