

Sprint Global Quality of Service: Guarantor of Application Delivery

White Paper

April 2006



Together with NEXTEL

1.0 Introduction

1.1 Application Management Focus

Enterprises, faced with employee, affiliate, and user populations that are increasingly outside the confines of the organization campus, expand their ability to serve through the implementation of application software that is delivered to users across networks. Data show that headquarters office application access is a much smaller percentage of application use than access initiated from major sites located far removed from the servers on which the applications run.

A result of this phenomenon is that “the majority of business managers see the value of IT as coming primarily from applications, rather than the supporting infrastructure,” according to Jim Metzler, (Network World, February 16, 2004).

According to Johna Till Johnson (Network World, February 23, 2004), most companies claim to run between 50 and 300 applications over WANs. Many companies have more than 1200. With this level of application activity, the issues of prioritization, availability, security, visibility, and reporting of application and data flow become very important.

Quality of Service (QoS) is all about application delivery. Sprint’s QoS provides a comprehensive solution tailored to the multiple network environments over which applications operate.

1.2 Why Is Network Management Through QoS Important?

A communications network forms the backbone of any successful organization. Networks complement, add value to, and enhance every business process.

- Network functionality has enabled an additional, and totally separate, channel of distribution, giving a completely new meaning to direct selling.
- Networks provide a competitive advantage, enabling collaboration of geographically dispersed teams, shortening product time to market.
- Networks provide a real-time connection with customers and target markets, delivering the capability to push messages to targeted markets electronically. They also provide the ability, through search engine operations and electronic use of traditional marketing materials, to pull prospects and customers into contact through Websites.

- Networks span distance in the development of partner collaboration programs.
- Networks reduce operating costs, making contributors more productive.

Networks transport a multitude of applications and data. These include high-quality video and delay-sensitive data such as real-time voice. Bandwidth-intensive applications, such as these, stretch network capabilities and resources.

QoS provides a means of focusing on the performance of applications across the network, rather than on the network itself.

As organizational success becomes ever more intertwined with the effectiveness of its network capabilities and applications that depend on those capabilities, QoS will grow in importance.

1.3 What Is QoS?

Quality of Service (QoS) refers to a set of techniques used to manage and measure the effectiveness of critical network resources. QoS enables the ability to provide better service to certain flows (for example, specific business applications).

To be effective, networks must provide secure, predictable, measurable, and sometimes guaranteed services. Achieving the required QoS targets becomes the secret to a successful end-to-end business solution. QoS techniques enable you to control the delay, delay variation (jitter), and packet loss for different types of traffic on a network. Principal QoS controls include traffic classification, relative prioritization, bandwidth allocation, plus ways to mitigate and manage congestion.

Historically, different traffic types (for example, voice and video) were transmitted over network mechanisms specifically suited for these applications. The development of the packet-based network, on which voice, video and data are converged, provides opportunities to expand business scale, optimize costs, and enhance organizational returns. However, converging to a packet-based network, with variable-length packets, requires considerable effort to ensure the predictability necessary to support all service types.

1.4 QoS Delivery

QoS is the result of a set of techniques employed to ensure proper end-to-end network treatment of various traffic types. Proper network treatment is a subjective description of the performance that an organization expects applications and identified traffic types to exhibit across its network.

Although it is delivered through technologies, QoS has more to do with defining the performance, reliability, and availability metrics for the applications a business operates across a network. Specification of these performance metrics is defined in agreements made with service providers called service-level agreements (SLAs).

The SLA specifies commitments for the ability of a network and/or protocol to deliver guaranteed performance/throughput/latency bounds based on mutually agreed measures, usually by prioritizing traffic.

Within the framework of SLAs, parameters such as jitter, delay, and throughput can be addressed for a variety of network types. For example:

- Streaming multimedia might require guaranteed throughput.
- IP telephony might require strict limits on jitter and delay.
- Dedicated link emulation requires guaranteed throughput and imposes limits on maximum delay.
- A safety-critical application, such as remote surgery, might require a guaranteed level of availability.

QoS, as defined in SLAs, ensures that all applications can coexist and function at acceptable levels of performance.

1.5 QoS Boundaries

Complicating the delivery of QoS is the variability in bandwidth availability across the wide area network (WAN) over which applications operate.

Applications function well within a local area network (LAN) environment where high capacity (10 Gigabit Ethernet) exists, yielding virtually infinite bandwidth.

As application traffic moves outside the LAN, available bandwidth is variable. Applications, many of which are not tailored to WAN characteristics and distances, often provide suboptimal responsiveness as a result.

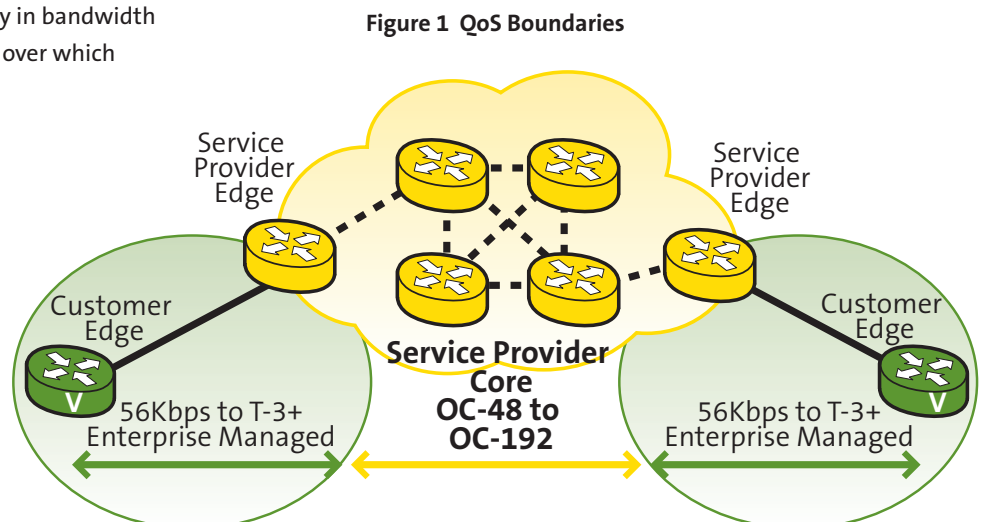
As is illustrated in Figure 1, the WAN can be divided into two environments. The area where bandwidth restrictions are most pervasive is the WAN connection between the end user organization location (customer edge [CE]) and the service provider network edge (PE). This WAN access connection is provided by and managed by the end user organization (Enterprise in the figure). It typically ranges from a 56 kbps capacity to a T1 (1.5 megabits per second) or T3 (45 megabits per second) capacity.

The major constraint to bandwidth availability on this customer edge to provider edge (CE-to-PE) link is cost. Because local access bandwidth is expensive, that relatively narrow pipe must be shared among multiple applications with diverse characteristics. These range from videoconferencing to ERP database lookups. Some traffic is very delay sensitive, but not as sensitive to packet loss (for example, voice, video). Other traffic uses a lot of bandwidth, is not very delay sensitive, but is very sensitive to packet loss (for example, ERP, database).

On this WAN connection bandwidth constraints and bandwidth cost are major drivers of QoS demand.

The other segment of the WAN is the service provider core. This is the beige area of Figure 1. The service provider is responsible for the configuration, management, deployment, and cost of this segment. The service provider customer is concerned about this WAN segment because the manner in which it is provisioned and managed provides evidence of the capability of the service provider to deliver the end-to-end application performance and QoS specified in the SLA.

The philosophy of service providers in the management of their cores varies and should be understood prior to entering any SLA. This philosophy will impact their ability to deliver QoS.



2.0 Sprint Global QoS

2.1 QoS Delivery Overview

There are essentially three ways through which to deliver the application effectiveness delineated in QoS guarantees.

2.1.1 Over-provisioning Approach to Application Assurance

One method is to provide network resources sufficient to meet the peak demand with a substantial safety margin. Resources include the bandwidth, routing devices, and other infrastructure necessary to provision the network such that bandwidth is virtually infinite. With this level of provisioning, application performance is unhindered by network constraints or congestion.

A service provider committed to this approach can make overprovisioning a viable, cost-effective alternative in the service provider core. Since the 1990s, backbone capacity costs have decreased drastically, creating an opportunity for service providers to scale core bandwidth in a cost-effective manner.

However, the approach can be very expensive when applied at the network edges (CE to PE). There is greater risk of peak demand increasing faster than predicted, so it is not generally considered an appropriate standalone solution at the network edges (CE to PE). At the CE-to-PE network segment, response to unexpected application or network expansion may easily be negatively impacted by delays in the deployment of additional resources.

2.1.2 Differentiated Services

Differentiated Services (DiffServ) is a method of providing QoS on large networks. DiffServ deals with aggregate flows of data rather than single flows and single reservations. A single negotiation will be made for all of the packets to or from, for example, a single ISP, or packets that are of a specific traffic type, such as voice. Classes of traffic are defined. The guarantees needed for each class and how much data will be allowed are defined for each class.

It is generally most effective to implement DiffServ at the network edge and to avoid clogging up the core. DiffServ is defined on a per-hop basis, with the QoS policies configured separately on each individual router involved in the process.

All the policing and classifying is done at the boundaries between DiffServ clouds. This means that the core routers can get on with doing the job of routing and not care about the complexities of collecting payment or enforcing agreements.

For most applications, DiffServ QoS is more than adequate. It is beneficial to provide more priority of all-voice calls than for all-data calls. It is appropriate to provide a higher level of service for a class of “business-critical applications” than for a class of “peer-to-peer” applications.

2.1.3 Integrated Services

Integrated Services (IntServ) is a fine-grained QoS system that specifies the elements to guarantee QoS on networks. IntServ can, for example, be used to allow video and sound to reach the receiver without interruption.

When using IntServ, each individual phone call or flow is identified and managed separately. The Resource ReSerVation Protocol (RSVP) is used to define source IP address, source port, destination IP address, destination port, and protocol.

Every application that requires a guarantee has to make an individual reservation.

The process is best implemented in the edge network, leaving the core network resources available for aggregate flows only.

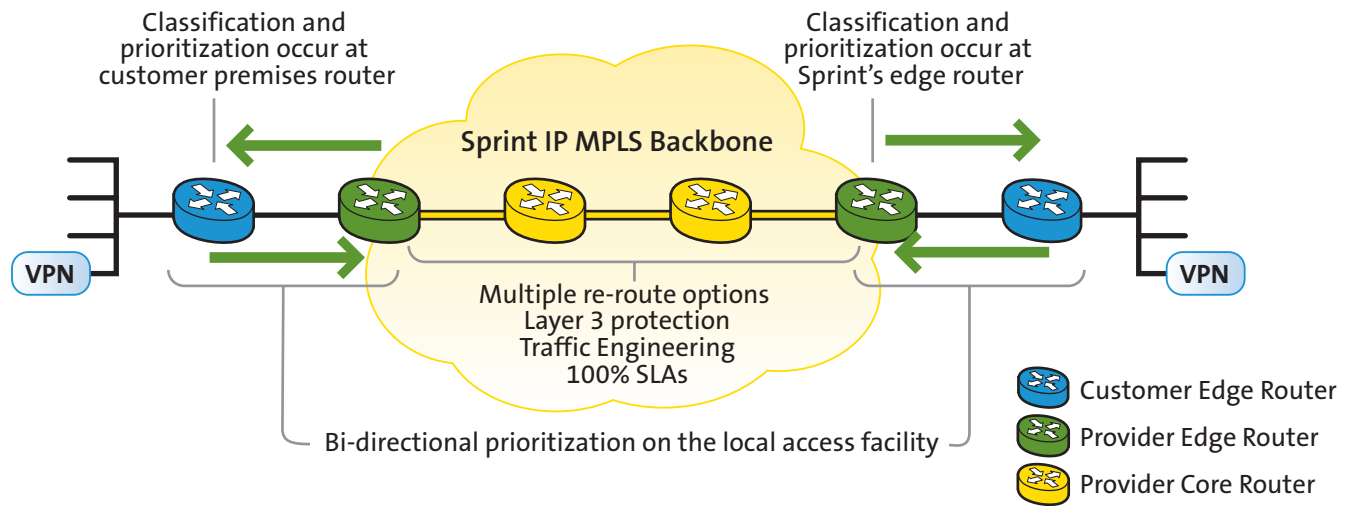
This method of providing QoS is the most operationally intricate. It is also the most costly from the perspective of memory and CPU usage.

IntServ is, however, used to provide the required QoS for a voice-on-demand (VoD) application or for other applications that a content provider who is providing content/applications that have stringent QoS characteristics.

2.2 Sprint Application Delivery Overview

The Sprint approach to network application performance is to deliver an end-to-end solution that ensures that application performance requirements are defined, delivered, and perceptible at all user interfaces. The approach is two-fold:

Figure 2 Sprint QoS Solution



1. Sprint delivers a complete QoS solution supporting DiffServ and IP Precedence on the WAN link between the customer edge (CE) and the Sprint provider edge (PE).
2. Sprint provides an absolutely congestion free and totally unconstrained core through which customer applications operate at maximum speed across an environment that is similar in bandwidth to the LAN environment for which the applications were designed.

Figure 2 portrays the manner in which the Sprint solution is implemented. This solution takes advantage of the capabilities of DiffServ and IP Precedence at the point where it is most effective, at the customer network edge.

The solution then ensures that no further traffic flow manipulations negatively impact application, data, and traffic type performance by making sure that the core network is open to move these flows at the maximum rate possible. Figure 2 also identifies some of the additional features of its core environment that, while they are not technically QoS features, add reliability and other protections to the total solution.

2.3 Sprint's Core Provides Robust Application Operations

At the core of every Sprint VPN solution is a robust OC-192 (10 Gbps) IP/Multiprotocol Label Switching (MPLS) enabled backbone. This infrastructure ensures that QoS concerns are eliminated for all traffic passing through its core. The commitment that Sprint has to

provision its core network with bandwidth (capacity) that is more than double the expected network traffic load is unique among service providers. It is, however, the most direct approach to ensuring a virtually congestion-free network environment.

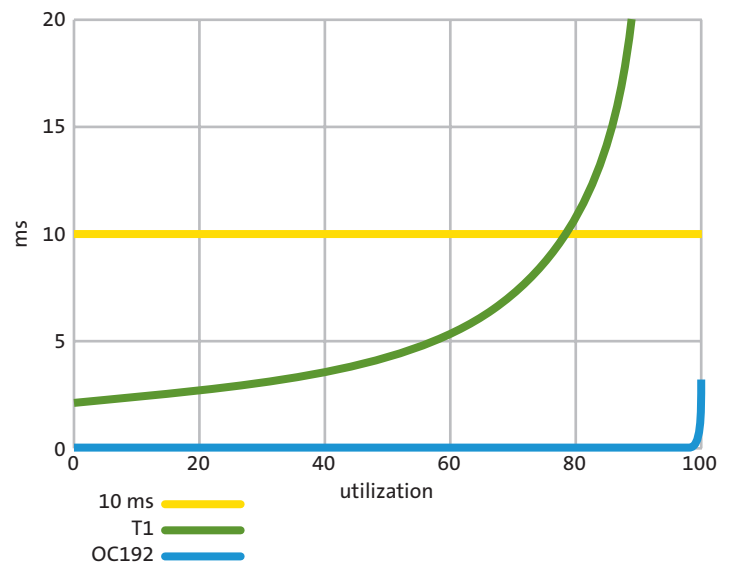
Sprint's approach to traffic engineering on its IP backbone is to provision ahead of the demand curve. The Sprint philosophy is one of total congestion avoidance. Therefore backbone links are maintained at traffic loads not to exceed 40 percent utilization, providing ample network capacity to effectively ensure high-quality service to all packets on the IP backbone. Sprint's congestion avoidance philosophy makes sure that data travels through the network routers on a first-in first-out basis, minimizing delay, jitter, and packet loss.

To achieve this, Sprint provides MPLS VPN services on an OC-192 native IP platform composed of Cisco System® routers. The Sprint backbone is managed through Cisco® 12000 and 7500 Series Routers. By delivering Sprint Global MPLS VPN over a native IP network, Sprint decreases networking complexity and eliminates unnecessary overhead, leading to a more cost-effective solution for customers.

Sprint's premise that the backbone can be provisioned such that it is congestion free is validated by the results of tests Sprint has conducted. These are reflected in Figure 3. The graph demonstrates that, at network backbone speeds (OC-48 [2.4 Gbps] and OC-192 [10 Gbps]), queuing delay is not significant until utilization levels approach 100 percent. Sprint's policy of provisioning the backbone such that it is no more than 40 percent utilized ensures that its IP network implementation delivers optimal QoS levels across its core.

Sprint test process:

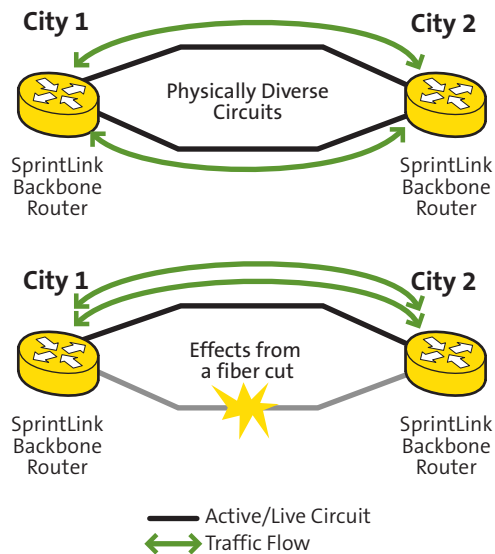
- Testing was coordinated by Sprint.
- Implemented prioritization scheme on T1 and OCn circuits to address potential queuing delays. Used traffic generator to load circuits.
- Results showed that optical circuits are not affected by queuing delay until nearly 100 percent utilization.
- Testing consisted of Internet-Mix (I-Mix) traffic.
- Enterprise Mix (E-Mix) traffic showed similar results at points tested.
- Tests were conducted using fixed utilizations and measuring end-to-end delay through a router interface of different types for different speeds.
- All SONET/SDH (OCn/STM-n) interfaces were executed using fiber interfaces.
- All TDM (DS3/T1) interfaces were copper.
- Graph represents best-fit line over data.

Figure 3 T1 Access Versus OC-192 Backbone Circuit

The use of Intermediate System-to-Intermediate System (IS-IS) and its fast reroute capabilities adds assurance that traffic flowing across the Sprint core moves through a network that truly delivers congestion-free bandwidth.

QoS levels are further guaranteed through the use of the IS-IS dynamic routing protocol to compute metrics supporting traffic distribution over the IP backbone links. These links are provisioned in pairs using per-flow load balancing between the diverse paths. In the event of a fiber cut, traffic is automatically rerouted to the other link in the pair using the fast reroute capabilities associated with IS-IS. All links in the network are live and can accept traffic if other links go down. Figure 4 illustrates the fast reroute of traffic that results from a cut in a link.

Because the network is fully interconnected, Sprint is able to provide multiple redundant paths through the network. Provisioning MPLS VPN over a native IP core provides customers with a greater level of reliability and redundancy for their mission-critical, enterprise WAN traffic. This is a critical differentiator in Sprint's QoS solution.

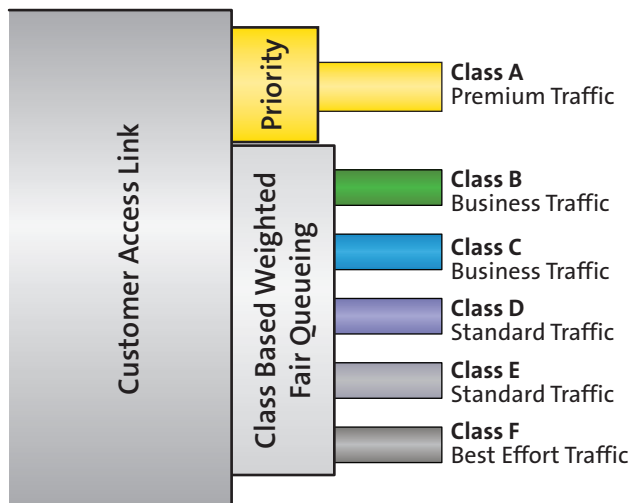
Figure 4 Traffic Fast Reroute Capability of Sprint MPLS VPN Architecture

2.4 QoS Between Sprint and the Customer Edge

At the edge of the network, between the customer routers and Sprint's edge routers, bandwidth is defined by the customer and is typically more constrained. To bring QoS capabilities to this network segment Sprint enables traffic prioritization (class of service [CoS]) from the customer site to the network edge.

At the edge, there is a growing need for CoS and application prioritization as customers move to converge multiple applications over a single data network. Real-time, end-to-end CoS is required to support such time-sensitive applications as voice and video over IP. Using Cisco innovations such as Class-Based Weighted Fair Queuing and Low-Latency Queuing, Sprint provides customers the ability to prioritize their traffic on the access links, which are the most likely places for congestion to occur because these links are frequently low bandwidth, overutilized, and tend to experience “bursty” traffic patterns. Customers have the flexibility to determine the number and size of the queues as needed to support their applications. Figure 5 demonstrates the manner in which traffic is segregated into classes in order to prioritize traffic types.

Figure 5 MPLS-Enabled CoS Traffic Segregation Module



2.5 Management of CoS Policies

CoS policies are determined by the customer interactively with the Sprint Account Team Solutions Engineer.

The Sprint Account Team and the Sprint customer collaborate with the Sprint IP Implementation Team to ensure that the CE and the PE routers are properly configured to support the desired customer policy.

Once implemented, changes to the policy are submitted to Sprint by the customer via an online change request form. The Sprint IP Implementation Team will enable the policy change on the PE routers and coordinate with you, the customer, to make sure that the CE is properly configured as well.

2.6 Sprint QoS Certification

Sprint has industry recognition for its unique and fully developed QoS solution.

A third-party assessment validated that Sprint’s MPLS virtual private network (VPN) service meets Cisco best practices and standards for delivering QoS. This earned Sprint recognition as a Cisco certified IP VPN Multiservice QoS provider.

The new Cisco Powered Network designation, IP VPN-Multiservice QoS Certification, verifies that the managed IP VPN service operator has committed to meeting best-practice criteria, defined by Cisco, for delivery of real-time voice and video services end-to-end.

To obtain and maintain this certification the service provider (in this case Sprint) is required to undergo an annual third-party, on-site assessment, to validate that the service provider follows best practices for delivering recommended levels of network performance (including latency, jitter, and packet loss) and customer support. The certification also specifies that the SLA must span the network from customer edge to customer edge in order to provide the end-to-end transparency and seamless management essential in a converged voice and data environment.

2.7 Sprint Global QoS Benefits

- Sprint’s backbone QoS philosophy is based on simplicity and structured to transmit all customer data as quickly as possible across the core.
 - Characteristics include a high-bandwidth core with a high level of diversity.
 - Eliminating contention for resources ensures all customer traffic is treated with highest QoS level.
- CoS provides standards at the edge to complement Sprint’s core QoS capabilities and allow customers to effectively manage their bandwidth.
- Sprint delivers a cost-effective solution.
 - As customer CoS needs change, pricing does not.
 - Customers can change policies on the edge without increasing costs and with the knowledge that core QoS levels will remain constant.

3.0 Managed Services Option Offering

Sprint offers a Value-Added Management Option which places the management of the customers' routers (CEs) under Sprint responsibility. In this scenario, the CoS policy setting is determined by the customer, in collaboration with the Sprint Account Team, but the Sprint Managed Network Operations Group owns the process of configuration of the customer router in addition to coordination with the Sprint IP/MPLS implementation organization for the management and configuration of the PE router.

With a fully managed solution, the customer receives maximum Sprint support:

- Router configuration management
- Proactive monitoring
- Notification for the router
- Break/fix support
- Customized reporting

4.0 Sprint Solutions Differentiators

A combination of Sprint's engineering philosophy and standard product features provide customers the opportunity to experience enhanced QoS with Sprint MPLS VPN:

- Service is provided across a converged IP backbone
- Sprint enablement of IP/MPLS VPNs facilitates network management and maintenance and provides predictable performance.
- Backbone links are provisioned in pairs, delivering fast reroute provided by IS-IS.

- The solution is engineered for congestion avoidance as links are maintained at less than 40 percent of capacity.
- 100 percent Cisco Powered Network.
- Sprint is the first "IP VPN-Multiservice QoS" Cisco certified global provider.
- End-to-end SLAs, end-to-end reporting and Class of Service included standard to enhance the customer experience.

5.0 Conclusion

Sprint's unique QoS engineering philosophy is based on maintaining a simple, flat core architecture designed to be virtually congestion-free, with SLAs that provide for better than 99.9 percent data delivery rate and up to 100 percent availability. It also provides network transmission latency of less than 55 milliseconds. This QoS philosophy is validated by its extensive use in Sprint customer accounts and by industry-leading experts that have certified the Sprint QoS solution.

With Sprint's proven, industry-leading native IP architecture the customer receives:

- A high degree of bandwidth scalability
- Intelligent routing
- IP convergence

The Sprint implementation of MPLS VPN adds secure, global any-to-any connectivity backed by end-to-end SLAs and delivers a comprehensive, flexible, and cost effective QoS solution:

- QoS: Sprint offers industry-leading SLAs and delivers availability, latency, packet loss, jitter, and installation services that meet agreed-upon customer requirements.
- CoS: Sprint maximizes the use of existing bandwidth and allows users to realize the best possible performance of delivery of mission-critical traffic.



Together with NEXTEL