Next Generation Core Network Architectures

A White Paper

IronBridge Networks
55 Hayden Avenue
Lexington, MA 02421
http://www.ironbridgenetworks.com

781-372-8000

Table of Contents

Exhibits

Introduction

Internet Service Providers (ISPs) can choose from among a number of existing and emerging technologies to create their backbone IP networks. In many cases core networks have been comprised of datagram IP routers interconnected via SONET links. In other cases ATM has been used to interconnect core routers. MPLS is emerging as another option for network backbones. Each technology presents trade-offs in the service provider's quest for building the optimal core network.

The choice of one core networking technology over others is dependent upon the ISP's current service offering, installed network facilities and future objectives. Each technology presents opportunities for ISPs to enhance the operation of their business. They can enable ISPs to:

- Simplify network engineering
- Improve bandwidth utilization
- Increase network reliability
- Create a multi-service, IP-based network.

Exhibit 1 - ISPs are faced with backbone technology choices

As with all networking technologies, the issues of interoperability and investment protection must be considered when evaluating IP backbone alternatives. ISPs desire multi-vendor solutions that foster a competitive marketplace. They also want to ensure that equipment will not be rendered obsolete by the introduction of newer technologies.

This paper describes each of the core networking technologies and compares and contrasts them in accordance with the following criteria:

- **Technical Stability and Interoperability:** Stable, open and interoperable industry standards protect ISP investments in equipment.
- **Traffic Engineering:** Facilitates analysis of traffic patterns and balancing loads across multiple links for optimum utilization.
- **Network Reliability:** Enhances over-all network resiliency, including recovery from link or equipment failures.
- **Bandwidth Efficiency:** The amount of overhead associated with delivery of IP datagrams.

- **Support for Multi-Service Networking:** Ability to support delivery of multiple classes of service (CoS) and service level agreements (SLA) .

Background

ISP networks are typically constructed in a hierarchical manner. The hierarchy is usually two or three layers and is typically based on link and equipment speed and aggregation of traffic (Exhibit 2).

At the carrier edge, traffic is accepted at relatively low speeds from the ISP's customers and concentrated onto a small number of higher speed links. These edge devices are referred to as "edge routers" which are capable of making hundreds of low-speed user connections and provisioning the attributes of the ISP's service for the user.

The edge router links are connected either to hub routers or to very high-speed core routers. Hub routers provide another level of aggregation before traffic reaches the core network. Core routers provide a very high level of aggregation for transport across high-

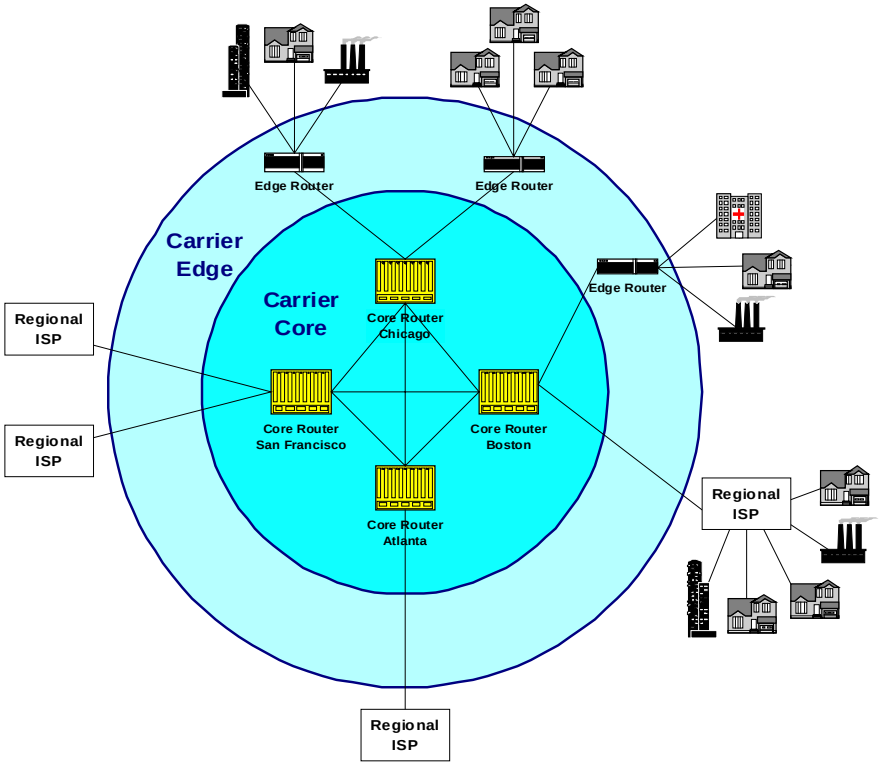

Exhibit 2 – ISP networks aggregate traffic as it travels across the core

speed fiber-optic links. The core router also performs statistical multiplexing of traffic to increase bandwidth efficiency.

Some of the key requirements for a core router are that it:

- Scale in capacity to meet the increasing demands for high bandwidth and port count
- Provide a high availability backbone network service to the lower tier routers
- Provide tools to simplify the operation and engineering of the network
- Support IETF Differentiated Services for multi-service networking

Differentiated Services

Differentiated Services, IETF RFC2474 and RFC2475, are specifications that define a framework for delivery of multiple service levels through an IP router and implementation of that framework in IPv4 and IPv6. Differentiated Services provides a common way for routers to discriminate between different classes of IP traffic and apply different forwarding treatments to each class. It doesn't specify the service or performance characteristics for each class —this detail is left to the router vendor.

With Differentiated Services the type of service (ToS) byte in the IP datagram header is marked to indicate the assigned class of service for each packet. For example, a carrier might choose to divide IP traffic into that generated by network control (routing protocols and network management), circuit emulation traffic, voice traffic, high-priority Internet traffic, and regular "best-effort" Internet traffic. The carrier would configure the edge routers to mark each packet with a ToS value corresponding to each of these traffic classes.

Once the classes are identified, the router can isolate each class in separate queues and apply priority forwarding in accordance with the service objectives for each class. This classification technique allows flows to be aggregated across the inner tiers of the network and relieves the core routers of the need to perform fine-grained flow classification. It also makes multi-service networks much easier to build on a large scale.

The range of service options that are based upon this approach can be greatly widened by introducing policing, traffic engineering and provisioning techniques. For example, each class of service can be

assigned a bandwidth limit and the router can police the received traffic on the ingress link to determine whether actual usage is within contract specifications. The router can apply different forwarding and congestion management techniques based on whether packets are within contract or out of contract specifications. In addition, the links used to handle high priority traffic can be engineered to exceed the aggregate contracted bandwidth with very high probability. This traffic engineering technique allows the service provider to ensure that this traffic will receive access to its contracted bandwidth.

When policing, traffic engineering and provisioning tools are employed together, a range of services can be supported and service level agreements (SLAs) can be assured.

IP over SONET

IP datagram routers are widely used in core networks because they are simple, offer good performance and are compatible with the edge routers used in the aggregation network. As core networks evolve to support larger traffic loads and multiple services, datagram routers need to provide support for classes of service, increased availability and traffic engineering. A new generation of core routers with features to meet these needs is now becoming available.

Differentiated Services support enables datagram routers to provide many of the class of service and traffic engineering benefits of IP over ATM or MPLS without the additional overhead and complexity. In addition, datagram routers can integrate SONET functions into the Layer 3 forwarding decision to improve network availability.

Short Term Congestion Control in Datagram Networks

Early implementations of IP were designed to transmit data whenever the host had data to send, without concern for network congestion. This introduced the possibility of congestion collapse: When packets were lost due to congestion, hosts would respond by retransmitting data, resulting in more network traffic and more congestion. This problem has been solved by having hosts control the amount of traffic that they offer to the network through a process called TCP slow start. Using slow start, the maximum number of unacknowledged TCP frames (known as the window size) starts out small and increases over time. If data is lost, hosts reduce their window size. This ensures that the total amount of traffic in the network is controlled when congestion occurs.

However, TCP slow start introduces another potential problem: When congestion occurs, data from multiple TCP sessions will be lost simultaneously. This can result in a number of hosts reducing their window size and then subsequently increasing their window at the same time. Gradually, the traffic oscillations from a very large number of hosts can become synchronized, resulting in large relatively short-term oscillations of the aggregate traffic levels across

links in the core. Obviously, the bandwidth utilization of the high-speed core links under these conditions is very poor.

In order to prevent synchronization of host window oscillations the core routers must implement Random Early Detect (RED) and also provide adequate buffering. RED is a technique by which routers detect that a link has the potential to become congested, and react by discarding a small percentage of traffic early, well before running out of buffer space. The hosts whose traffic is discarded will then reduce their window size early. This ensures that some hosts will be forced to be out of synchronization with overall traffic oscillations. Over time, RED will defeat the synchronization of traffic oscillations, causing a smooth aggregate flow on large links in the core of the Internet.

It is also essential that routers offer sufficiently large buffers for user data. This is critical for two reasons: first, effective use of the RED technique requires a large variation between the times when different flows experience data loss, and secondly routers can buffer (rather than discard) data during the time it takes RED to become effective. Some router vendors have been tempted to improve port density and cost by skimping severely on the amount of buffers offered. This is a poor tradeoff because it ensures that the core of the network will be limited to very low average link utilization.

Traffic Engineering

There are two factors that influence bandwidth efficiency in datagram IP routed networks: congestion management and distribution of traffic across multiple network paths. Many datagram-routed IP networks achieve only 60 percent average bandwidth utilization. The increasing volume of backbone Internet traffic and the expense of high-speed, long distance links are driving carriers to improve bandwidth utilization. Congestion can be properly managed by using routers that implement RED. The task of improving network efficiency by balancing the traffic on links throughout the network is known as traffic engineering.

Effective traffic engineering can be applied to the pure datagram network architecture using traditional mathematical techniques and Optimized MultiPath (OMP) extensions to the IGP. Under reasonable conditions, these techniques enable optimal traffic engineering.

Mathematical techniques are implemented by centrally computing metric values and distributing them via network management

commands to each core router. The routers advertise the metrics in the routing protocol (OSPF or IS-IS). Each router can then run a normal Dijkstra route computation and forward packets on minimum cost paths.

In many cases this technique results in equal cost paths. This in turn requires the use of a method to split traffic appropriately across them. An emerging method to accomplish this is through use of Optimized MultiPath (OMP). OMP allows link usage information to be advertised into OSPF (draft-ietf-ospf-omp-02.txt) or IS-IS (draft-ietf-isis-omp-01.txt). The link usage information enables routers to distribute packets over multiple paths with the aim of equalizing network traffic and improving link utilization.

When splitting traffic over multiple paths, it is important to ensure that packet order is maintained for each individual flow. Typically, this is accomplished by applying a hash algorithm to the source and destination address pair. Flows are then distributed amongst the available paths based on the result of the hash function. This ensures that packet ordering is maintained for each flow.

These traffic engineering techniques for datagram routers provide a fully automated way of achieving very high network utilization. They are relatively straightforward and achieve provably optimal traffic engineering. The quality of traffic engineering achievable with datagrams in the steady state is identical to that achievable with MPLS in typical networks. The difference between the two techniques shows up in the rate of response to sudden network changes (such as major links failures) and in one detail of support of circuit emulation which is discussed below.

Multiservice Networking

Datagram routers can provide support for multiservice networking by implementing Differentiated Services and other features that deliver predictable behavior for specific traffic types. In addition, hard service guarantees can be made for some services if traffic levels can be characterized and the network is carefully provisioned end-to-end to accommodate them.

Supporting specific service guarantees with Differentiated Services requires:

1. The ability to separate traffic into multiple services;
2. Accurate prediction of the level of traffic associated with any particular service class;
3. Knowledge of service characteristics (delay and loss) that will be achieved in routers and links for each particular service class for each particular service level.

The first item is the entire point of Differentiated Services. The other two points will vary according to the type of service supported and the characteristics of network equipment, respectively.

Consistent delivery of Differentiated Services requires special queuing mechanisms to separate each class of traffic. Once separated, each queue is managed independently, according to a scheduling algorithm that ensures classes receive access to bandwidth in proportion to the desired level of service.

Traffic must also be policed to ensure conformance with service level agreements. Non-compliant user data traffic is detected and discarded, or can be marked as non-conformant and forwarded on a 'bandwidth available' basis. The router handles the non-conformant traffic in a manner that ensures that it will not have an impact on compliant traffic.

The combination of policing and RED provides effective tools to manage short-term network congestion.

Prediction of traffic levels can be done very accurately for some types of traffic but is more difficult for other types of traffic. For example, with virtual leased line service (circuit emulation) the traffic characteristics for each circuit (network ingress, egress and bandwidth) are fixed.

Similarly, voice traffic characteristics are easy to predict because each voice call is comparatively small in size and the dynamics of voice are well known. In addition, voice over IP services currently represent only a small percentage of total Internet traffic and are predicted to remain so, even as a large portion of voice traffic gradually shifts to the Internet. If all voice traffic is put into another single class of service, the network can be provisioned to allocate very high quality service to voice traffic.

Traffic prediction for general Internet service is more difficult. In principle, the next web search from a particular host could go anywhere in the world. However, carriers can collect statistics regarding Internet traffic levels and offer relative guarantees for general Internet traffic.

Proper support for Differentiated Services requires routers that exhibit predictable performance for all traffic levels. For example, the packet loss rate and delay for each service class must be quantified, per service class, for all traffic levels. Determining these relationships may be difficult for router architectures with limited interconnect bandwidth. Ideally, a router architecture with full bandwidth any-to-any connectivity is required.

Stability and Interoperability

The IETF has agreed upon a differentiated services architecture (RFC2474) and packet marking (RFC2475). Router treatments of packets in the available service classes have been specified, but service offerings based upon such service classes and per-hop behaviors have not been specified yet. However, current standards are adequate to allow equipment vendors to implement interoperable equipment on a per-hop treatment basis. There is a good deal of flexibility in mapping service classes to carrier service offerings.

OMP is currently being defined as an option to the OSPF and IS-IS protocols.

Bandwidth Efficiency and Reliability

A traditional datagram routed network does not incur any of the additional overhead associated with ATM and MPLS encapsulation techniques. In terms of raw encapsulation overhead, it is therefore the most bandwidth efficient of the three alternative core network architectures.

IP over SONET routers can take advantage of SONET signaling to enhance the ability of the router to re-route around a failure. SONET provides identification of link failures and recovery within 50 msec. Failure signals are used by SONET terminals operating at Layer 2 to automatically switch traffic between the affected link and the protection link. However, a router with integrated support for SONET can use the failure indication to intelligently re-route traffic at Layer 3. This allows the router to decide what traffic among all the links in the protection group should continue to flow over the remaining links. It also allows the router to select particular classes of service to drop as a result of the failure.

Terabit routers have an advantage over smaller routers when it comes to SONET recovery. The sheer size of terabit routers imply that the interconnection between two core terabit routers is likely to consist of multiple OC-48 or OC-192 links. Assuming that terabit routers implement SONET APS, it is therefore straightforward to provision a single backup link for every N physical links between any two core routers. In contrast, the interconnection between smaller routers may consist of a single SONET link, which makes using SONET APS in the router very costly. This is one of many reasons that a large number of small routers is not a sufficient alternative to a moderate number of terabit routers in high bandwidth core networks.

Statistical gain can be increased considerably by providing both premium and bulk services over the same infrastructure. Premium services, such as VPNs require a relatively high assurance of service (low delays, low loss rates, very high probability of service availability). If these services were delivered on a separate network, the overall load would have to be limited to ensure that short-term traffic fluctuations do not overload the network. By adding bulk Internet service over the same network, the excess network capacity (needed to ensure the premium service SLA) can be used by the bulk service. In addition, there are likely to be time of day differences between the peak period for VPN service and the peak period for bulk Internet service. By combining lower priority bulk Internet traffic and premium service traffic on the same infrastructure, it is possible to combine the high revenue potential of VPN service with the high utilization that is acceptable to the cheaper bulk traffic.

IP over ATM

ATM is frequently used as an alternative or a complement to the core routing technology described above. Previously, ATM was the only way to trunk IP networks at high-speeds. Routers have since equaled the link speeds offered by ATM switches, although ATM still features the ability to provide virtual connections among routers. The virtual connections reduce the number of physical connections between routers and improve the network's traffic engineering capabilities.

Under this mature technology, an ATM core is used to interconnect multiple IP routers. In many cases, the same ATM core may be used to support other services, such as ATM PVCs or Frame Relay. This is often referred to as an overlay network configuration, where the Internet service is laid over the ATM transport network (Exhibit 3).

A long-term consideration for ATM is the pace of development for

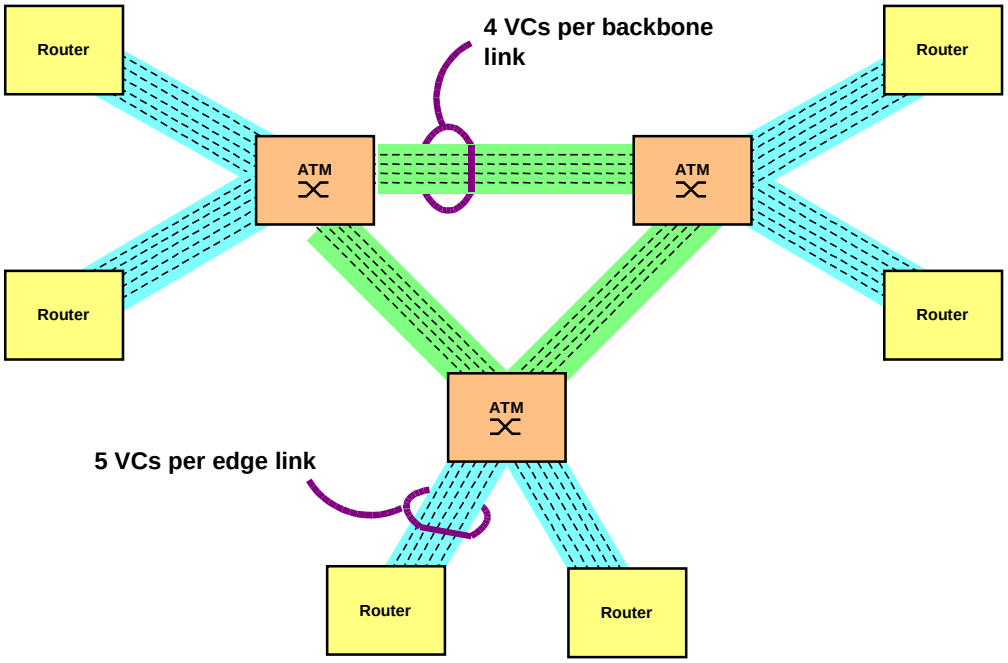

Exhibit 3 – ATM cores feature a full mesh overlay architecture to connect multiple IP routers

high performance SAR chips. Generally speaking, the fastest SAR chips available are not as fast as the fastest optics, implying a constraint on the speed of ATM interfaces on high performance core routers and ATM switches.

Support for Multiservice Networking

ATM has been designed to support multiple types of service in a single network infrastructure. A single ATM network can support IP Differentiated Services traffic, circuit emulation and VPN traffic, as well as voice and video traffic.

Support for multiservice IP networking over an ATM core network is complicated by the fact that ATM's QoS capabilities were designed for a fundamentally different purpose than support for IP CoS. There are two issues:

- Mapping IP classes of service into ATM service categories that deliver an equivalent service adds administrative complexity;
- Three of ATM's service categories [constant bit rate (CBR), variable bit rate (VBR) and available bit rate (ABR)] were not designed to support the TCP slow start flow control protocol. This reduces the number of service categories that can be used to prioritize IP traffic to the GFR and UBR service categories with Early Packet Discard (EPD) and Partial Packet Discard (PPD) enabled.

Each of these factors is discussed below.

Mapping IP CoS to ATM QoS

Support for IP over ATM requires mapping IP Differentiated Services to ATM service categories. This adds an extra level of administration to the network since the lower tier routers will use Differentiated Services to discriminate between services. Mapping is also somewhat complicated by a lack of standards.

Most ATM overlay networks do not currently use ATM's multi-service capabilities. They use only the UBR service category, which does not require configuration of a traffic descriptor. Differentiated Services support requires mapping each IP class of service (CoS) to a different ATM virtual circuit, each with a specific ATM service category and traffic descriptor.

Service Category		Traffic Descriptors
CBR	Constant Bit Rate	The network reserves a constant amount of bandwidth, the Peak Cell Rate (PCR, for this VC and does not exceed tolerances for Cell Transit Delay (CTD, Delay Variation (CDVT and loss (CLR
VBR	Variable Bit Rate	Characterized by an average or Sustained Cell Rate (SCR, Peak Rate (PCR) and Maximum Burst Size (MBS. Data that exceeds the SCR is eligible for discard, but short term bursts up to the PCR are always delivered
ABR	Available Bit Rate	Traffic may be submitted up to a Peak Cell Rate (PCR), the network reserves a Minimum Cell Rate (MCR, the remainder is delivered only if bandwidth is available, and all transmissions are subject to ATM flow control
GFR	Guaranteed Frame Rate	Similar to VBR, except the burst size is equal to the packet MTU, allowing all cells in a frame to be admitted at line rate
UBR	Unspecified Bit Rate	Traffic may be submitted up to a Peak Cell Rate (PCR) but no bandwidth guarantees are offered by the network

Exhibit 4 - ATM service categories provide a range of alternatives for handling multiservice IP traffic

The ATM switch forwards traffic on each VC according to ATM standards. Traffic that is within the descriptor limits receives priority treatment. Bursts above each of these descriptors, and traffic on the UBR connections, are accommodated if bandwidth is available.

Each connection that offers traffic within its minimum traffic descriptor will experience predictable delay, delay variation and cell loss, as specified by the ATM standards. Thus the ATM core network can provide hard service level assurances for IP Differentiated Services flows that are fully characterized.

Congestion Control

ATM switches are designed to manage congestion by discarding data that does not adhere to the traffic contract. The service categories that provide bandwidth guarantees (CBR, VBR and ABR) police received data according to traffic descriptors. They assume the attached device delivers a well-behaved flow.

This assumption can interfere with the operation of TCP slow start and limit average link utilization. A burst of traffic from a single

host can be discarded by the switch, causing TCP slow start to significantly drop its transmission rate. Over time, a network of hosts can oscillate in synchronization between high and slow transmission rates.

ATM switches offer Early Packet Discard (EPD) and Partial Packet Discard (PPD) mechanisms on the UBR and GFR service categories. These mechanisms ensure that when congestion occurs and the switch must discard cells, the entire packet is discarded. It is reasonable to expect that these mechanisms will be enhanced with Random Early Detection (RED) techniques, similar to routers. Thus, the GFR and UBR service categories will soon provide priority forwarding that is compatible with Differentiated Services and TCP slow start.

Bandwidth Efficiency

ATM imposes additional bandwidth overhead on the IP data, resulting from two factors: The extra cell headers and AAL5 trailer; and the "round-off error" due to partially filled cells. This cell tax is typically 20 percent.

An alternative scheme to reduce the ATM overhead was recently studied in a research network. The study found that if AAL5 were used with the null encapsulation (restricting a particular VC to carry only IP, but eliminating the SAP/SNAP header) then the ATM overhead is approximately 15-16 percent. This is because nearly all IP packets are either large (500 bytes or larger), or exactly 40 bytes. Large packets are more efficient because the AAL5 header is amortized over a larger packet, and only the last cell is potentially incomplete. 40-byte packets are relatively efficient because with an 8-byte AAL5 trailer they fit exactly into the 48-byte ATM cell.

This scheme is not interoperable with FRF.8, which may be a limitation for carriers with frame relay gateways. The overhead with this scheme is nonetheless greater than that of the other approaches covered in this paper.

In the future, a new Frame Network-Network Interface (FNNI) may become available that is designed to eliminate the cell tax. FNNI standardization is progressing, although FNNI implementations are not yet available.

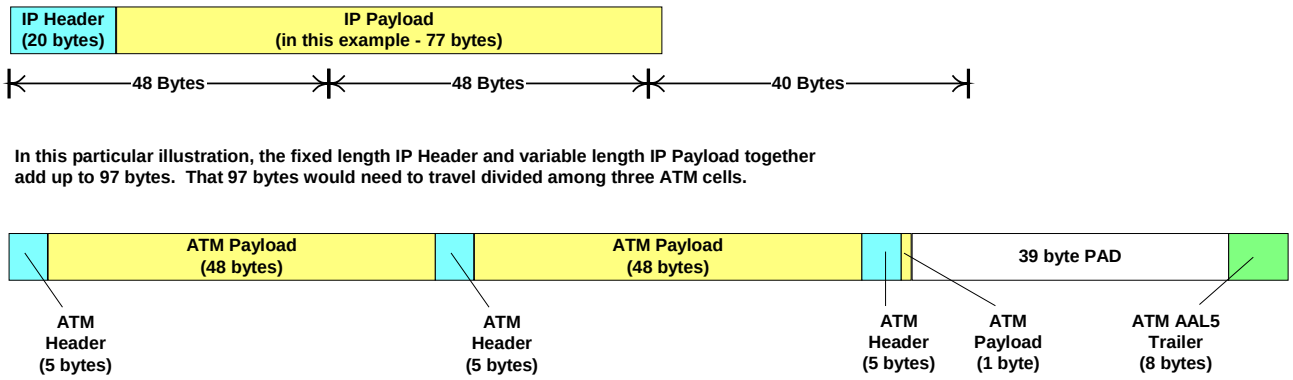

Exhibit 5 - ATM data structure can result in a "cell tax"

Network Reliability

ATM can use many of the same reliability mechanisms as IP over SONET. In particular, ATM links typically make use of Layer 2 SONET APS. In addition, the Private Network to Network Interface (PNNI) routing and signaling standard allows SVCs and soft PVCs to be rerouted upon link failure.

The PNNI link failure recovery time can be substantially longer than native SONET APS, or a datagram router that has integrated SONET APS. This is due to PNNI's call set-up function, which must be used to reroute the failed PVC. Call set-up is dependent upon the number of PVCs affected by the failure and may take anywhere from one to tens of seconds. The 50 millisecond recovery time available with SONET APS makes it a better alternative than PNNI.

If SONET APS is not desirable, then an alternative using redundant ATM VCs can be used. In this scenario, redundant VCs are set up a priori between each source and destination. While this doubles the number of VCs that must be maintained, it also ensures that an alternate VC is immediately available for reroute, in case of failure.

Traffic Engineering

ATM technology provides rich sets of tools that can be used to monitor network traffic patterns and actively manage network performance. Perhaps the most powerful ATM tool is the ability to explicitly route traffic between routers connected by the ATM network. The trade-off with ATM is the administrative overhead associated with engineering the large number of VCs needed in a core network. ATM does not scale well in large core networks.

Typically, each router in an ATM overlay network is connected to all other routers in a full mesh VC topology. ATM switches provide traffic counters to measure the actual bandwidth utilization on each VC. This allows traffic patterns between routers to be isolated and the computation of a future traffic matrix.

Given a known traffic demand matrix, network topology and capacity, it is relatively straightforward to run a centralized traffic engineering application. Explicit routing is then used to provision VCs along the desired paths to implement the desired solution. ATM traffic descriptors ensure that each VC on a shared link receives the bandwidth needed to support the traffic between routers. This allows efficient traffic engineering through centralized route computation.

Alternatively, connection-oriented technologies such as ATM and MPLS also allow for fully distributed traffic engineering. Generally speaking, it is not possible to ensure that distributed traffic engineering results in optimal traffic allocation, but very good allocation can be achieved without the disadvantages of centralization. Distributed methods can also be used as short-term approaches following link failure, to allow reasonable network operation while waiting for a centralized approach to optimize the overall network operation.

ATM PNNI routing can be used to support distributed automatic traffic engineering. With one approach the first ATM switch for each VC selects the path for the VC based on network capacity and utilization information advertised in PNNI routing, as well as the requirements for the VC as advertised in UNI/PNNI signaling. PNNI uses the ATM ingress and egress addresses and traffic descriptors for each VC, among other parameters, to provision VCs automatically. This method has the advantage of providing an extra layer of fault tolerance. If a link fails, PNNI automatically recalculates the route along the remaining links without the need for a centralized computation.

It is also possible to move the PNNI computation to the ATM-attached routers through use of PNNI Augmented Routing (PAR). Where routers are multiply-connected to the ATM cloud, this has the advantage that a router can pick paths via any of its directly connected ATM switches based on the status of the ATM network. This also facilitates the computation of multiple non-overlapping VCs between any pair of core routers for redundancy. Finally, this method optionally allows each core router to optimize the entire set of VCs from itself to every other core router in the network.

Finally, some ATM switch vendors have implemented a derivative of the distributed automatic traffic engineering approach. These derivatives, sometimes referred to as constraint based routing, allow policy-based (configured) constraints on the paths taken by some of the traffic. This allows the user to exert policy controls for portions of the network while retaining automatic routing.

The use of connection oriented techniques (including ATM and MPLS) has an advantage over datagrams when supporting circuit emulation. As discussed above, datagram traffic engineering makes use of multipath, and multipath forwarding is typically based on a hash function on the IP source and destination address. This implies that while a relatively accurate split of traffic is possible, it is generally not possible to achieve perfect splitting. Where there are large unsplittable flows (such as a large emulated circuit), it becomes increasingly difficult to obtain fine grain traffic distribution over multiple paths. Connection oriented techniques allow precise allocation of emulated circuits to specific paths, with accurate allocation of individual flows to specific links.

A substantial drawback to using ATM overlay networks is the significant number of VCs that must be administered in a large core network configuration. To establish a fully meshed network, the number of VCs is:

$$\frac{N*(N-1)}{2}$$

where N is the number of routers. If Differentiated Services are implemented, then the number of VCs is multiplied by the number of service classes that must be supported by the core network. Many large ATM networks are already having difficulty administering the number of VCs required to deliver best effort service.

Stability and Interoperability

ATM is a very stable, open and interoperable networking technology. The ITU-T and ATM Forum have issued a comprehensive set of standards for ATM sufficient to build high speed networks with strong QoS control.

The ATM standards ensure multi-vendor interoperability and satisfy a wide range of networking applications. Successful wide-scale field deployments of ATM multi-vendor networks provide ample proof of its maturity.

IP Over MPLS

Multi-Protocol Label Switching (MPLS) provides a connection-oriented infrastructure, similar to virtual circuits, designed to support IP traffic. MPLS integrates label-based (i.e., connection-oriented) forwarding with IP routing and provides the traffic engineering advantages of ATM overlay networks without the bandwidth efficiency and performance limitations of ATM.

MPLS uses IP routing protocols with a label-switched forwarding paradigm. A new Label Distribution Protocol (LDP) and/or a version of RSVP provides the signaling mechanism that MPLS uses to set-up Label Switched Paths (LSPs) between specified ingress and egress routers.

Technical Stability and Interoperability

The IETF has completed much of the standardization work for MPLS. However, several issues must still be addressed and early MPLS implementations are likely to solve these issues in an independent and non-interoperable manner. As standards are completed, vendors are likely to converge on a single, interoperable implementation.

Interoperability is easier to achieve among core routers or between core and edge routers, and is relatively harder to achieve from edge router to edge router. For example, core routers will primarily be transit points for MPLS Label Switched Paths (LSPs) carried between edge routers. Edge routers need to be concerned about when LSPs are set up and the manner in which IP packets and VPNs are mapped onto LSPs. This requires cooperation between the ingress and egress edge routers. For this reason, it is likely that large core IP routers will be able to support a wide range of services between edge routers from multiple vendors, even in the short term. However, interoperability between edge routers using MPLS may be limited in functionality in the short term.

Bandwidth Efficiency

MPLS adds a relatively small, four-byte header to every IP packet. For carrying Internet traffic, MPLS is substantially more efficient than ATM and nearly as efficient as traditional IP. MPLS is more efficient than ATM or IP encapsulation for implementing VPNs. Also, MPLS allows for a wide range of traffic engineering approaches, which again increases the ability to make efficient use of network resources.

ATM creates overhead due to the AAL5 trailer, the SNAP/SAP header (if used), the 5 byte cell header attached to every cell, and the requirement to round up the payload to a multiple of 40 bytes. MPLS is up to 20 percent more efficient than ATM.

MPLS can make use of the same traffic engineering advantages as ATM (but with IP routing rather than PNNI routing). MPLS therefore combines powerful traffic engineering capabilities with an efficient packet encoding. For these reasons, MPLS allows very efficient use of network resources.

MPLS may be used to create virtual private networks (VPNs) by encapsulating IP datagrams. MPLS encapsulation is substantially more efficient than IP-over-IP encapsulation, because MPLS headers are smaller than IP headers. In addition, MPLS supports two levels of encapsulation, allowing networks with large numbers of VPNs to also leverage MPLS for traffic engineering. In this application, one level of MPLS encapsulation supports the VPNs while the second level performs traffic engineering. This requires 8 bytes of header. In contrast, IP-in-IP encapsulation for VPNs uses a single, 20-byte IP header, which is considerably larger than two MPLS headers.

Also, MPLS can provide an APS-like fail-over capability without the requirement to maintain unused backup links. MPLS provides better link efficiency than SONET 1:N or 1+1 reliability mechanisms.

Network Reliability

MPLS can be used to establish diversely routed redundant paths between ingress and egress nodes. ECMP can protect against a link failure by re-distributing traffic among the redundant paths. This provides a protection capability with detection and fail-over speed similar to SONET APS without the need to maintain unused links for

protection. This can be used in two ways: A backup path can be used to backup a single LSP or a physical link.

Exhibit 6 illustrates the use of an LSP segment to backup a single LSP. In this case there is an existing LSP (illustrated as a solid blue line) from router A, to B, D, and terminating at ingress E. A backup LSP segment is set up from B, via C, to D. This backup LSP is then merged into the existing LSP at node D. If the link from B to D fails, then node B immediately forwards all traffic received on the LSP onto the backup LSP segment.

A similar approach, combined with the MPLS multi-level encapsulation feature, can be used to enable an LSP segment to backup an entire link. Upon detection of a link failure, a Label Switch-Router (LSR) can switch all the LSPs (and datagram traffic) that had been using that link to the backup LSP. For example, in figure X, rather than using the LSP segment from B to D via C to backup a single LSP, the LSR could switch all traffic travelling over the link from B to D. Then, if the link from B to D were to fail, all traffic that would have gone over that link is instead sent over the backup LSP. The multi-level encapsulation feature keeps track of the multiple individual LSPs that are mapped into the backup LSP and returns them to their previous state at node D.

In each case, the fail-over time is similar to SONET because the LSR makes use of paths that are created a priori.

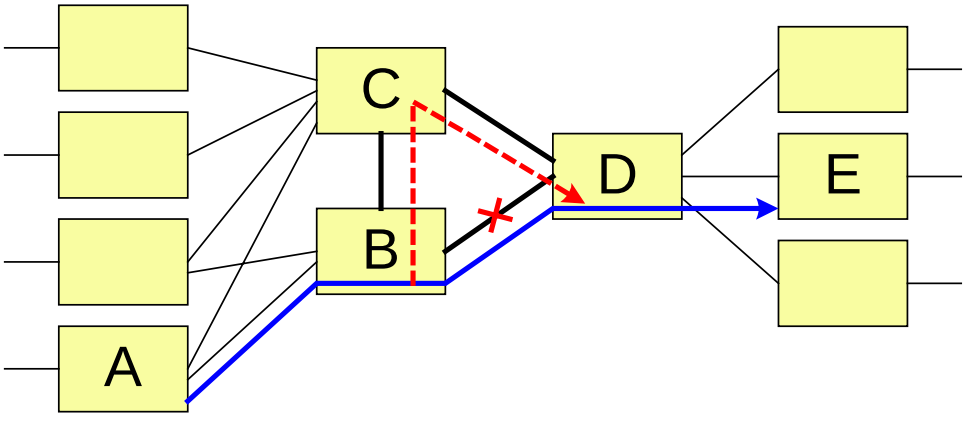

Exhibit 6 - Backup LSPs can be used to provide route diversity in an IP network

Traffic Engineering

MPLS provides a connection-oriented overlay network for transmission of IP datagrams. Similar to ATM, a full mesh of LSPs can be set-up between routers, allowing traffic to be explicitly routed across links between each router. Traffic patterns between routers can be isolated for analysis and optimization.

The traffic engineering capabilities of MPLS are very similar to those of ATM. The one difference is that routing in the connection-oriented MPLS infrastructure is based on IP routing protocols (BGP, OSPF, and/or IS-IS). Therefore, MPLS allows any of the traffic engineering methods described for datagram routers (distributed computation) or for ATM (central computation) to be used. With MPLS, ingress routers can compute paths across the core without having to participate in a "non-IP" routing protocol, such as PNNI.

MPLS also supports constraint-based routing, which allows policy constraints to be applied to the path set-up for some of the traffic.

MPLS allows bandwidth guarantees, similar to ATM, to be established on LSPs. These guarantees can be set up along the LSP using either RSVP or LDP signaling.

Support for Multiservice Networking

MPLS can be used in conjunction with Differentiated Services to support multiservice networking. MPLS offers straightforward mapping between Differentiated Services datagrams and LSPs. This is an advantage over the cumbersome mapping required when using ATM. MPLS allows each Differentiated Services class of service to be isolated and separate switching treatments applied. An MPLS label switched router can provide independent queuing and policing either on a per-LSP basis or using multiple queues per LSP.

When multiple queues are supported per LSP (e.g., for different CoS classes), MPLS has an advantage over ATM. MPLS minimizes the number of LSPs needed to support multiservice networking.

MPLS may be used to establish LSPs with bandwidth guarantees that support circuit emulation and VPNs, for example. MPLS may also be combined with datagram IP in a number of ways. MPLS LSPs could be treated as simple links between datagram routers, giving a specific traffic type an explicit route and queue to traverse.

Summary

As we have seen in the above analysis, each architecture for supporting IP service has advantages and disadvantages.

IP datagram routing is a mature technology with proven performance and interoperability. When used with Differentiated Services, provisioning and network modeling techniques, IP datagram routing provides support for a wide range of service classes, including voice over IP, circuit emulation, and VPNs. Mathematical techniques plus OMP provide robust traffic engineering strategies for a very efficient datagram network.

IP over ATM is a mature connection-oriented technology that can be immediately deployed to improve traffic engineering. The major long-term issue with ATM is the availability of higher speed commercial SAR chips. ATM also suffers from a "cell tax" that reduces bandwidth efficiency by about 20 percent compared to the other choices.

In many ways MPLS combines the best features of the datagram IP routed and IP over ATM approaches. MPLS provides a wide range of traffic engineering approaches and allows efficient use of network bandwidth. MPLS avoids the disadvantages of ATM (cell tax, the need for SAR functions in routers, and the need to run another routing protocol for ATM), while providing a very efficient encapsulation feature. MPLS standards are advancing quickly and will soon provide a basis for completely interoperable core and edge devices.

The three architectures provide great flexibility to develop innovative new IP-based services for customers, while also improving network efficiency and reliability.

IronBridge Networks

IronBridge supports all three principal architectural options (IP over ATM, IP over MPLS, and IP datagram routing) discussed in this paper. The IronBridge networking platform scales to multiple-terabits and features full support for IP Differentiated Services. IronBridge provides powerful tools that will help in the network

provisioning and modeling necessary to simultaneously support premium and bulk Internet services.

IronBridge plays a leading role in the definition of IETF standards. The company is committed to delivering open and interoperable solutions for core Internet backbone networks.

For more information, refer to our web site: http://www.ironbridgenetworks.com.

Bibliography

"ATM User-Network Interface Specification V3.1," af-uni-0010.002, 1994, The ATM Forum.

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z. and Weiss, W. "An Architecture for Differentiated Services," IETF RFC 2475, December 1998.

Comer, D., *Internetworking with TCP/IP. Volume I.* New Jersey: Prentice-Hall, 1991.

"A Framework for Multiprotocol Label Switching," <draft-ietf-mpls-framework-05.txt> September 1999, IETF, R. Callon, G. Swallow, N. Feldman, A. Viswanathan, P. Doolan, A. Fredette.

Halabi, S., *Internet Routing Architectures.* Indianapolis, IN: Cisco Press, 1997.

Huitema, C., *Routing in the Internet*. New Jersey: Prentice-Hall, 1995.

Moy, J., *OSPF, Anatomy of an Internet Routing Protocol.* Reading, MA: Addison-Wesley, 1998.

"Multiprotocol Label Switching Architecture," <draft-ietf-mpls-arch-o6.txt> August, 1999 IETF, R. Callon, A. Viswanathan, E. Rosen.

Nichols, K., Blake, S., Baker, F, and Black, D. "Definition of the Differentiated Services Field (DS Field) in the Ipv4 and IPv6 Headers," IETF RFC 2474. December 1998.

"Private Network-Network Interface Specification Version 1.0," af-pnni-0055.000, March 1996, The ATM Forum Technical Committee.

"Traffic Management 4.0," af-tm-0056.00, April 1996, The ATM Forum Technical Committee. The

Glossary

ABR. available bit rate

ATM. Asynchronous Transfer Mode

CBR. constant bit rate

CDVT. Delay Variation

CLR. Cell Loss Rate

CoS. class of service

CTD. Cell Transit Delay

Diff Serve. Differentiated Services

ECMP. Equal Cost Multipath

EPD. Early Packet Discard

FNNI. Frame Network to Network Interface

GFR. guaranteed frame rate

IETF. Internet Engineering Task Force

IP. Internet Protocol

IS-IS. intermediate system to intermediate system

ISP. Internet Service Provider

LDP. Label Distribution Protocol

LSP. Label Switched Path

LSR. Label Switch Router

MBS. Maximum Burst Size

MCR. Minimum Cell Rate

MPLS. Multiprotocol Label Switching

OMP. Optimized MultiPath

OSPF. Open Shortest Path First

PAR. PNNI Augmented Routing

PCR. Peak Cell Rate

PNNI. Private Network to Network Interface

PPD. Partial Packet Discard

PVC. permanent virtual circuit

RED. Random Early Detection

RSVP. Resource Reservation Protocol

SAP/SNAP. Service Access Point/Subnetwork Access Protocol

SCR. Sustained Cell Rate

SLA. service level agreement

SONET. Synchronous Optical Network

TCP. Transmission Control Protocol

ToS. type of service

UBR. unspecified bit rate

VBR. variable bit rate

VPN. virtual private network