Enhancing Transport Networks with Internet Protocols

Alan Chapman, Nortel
H. T. Kung, Harvard University

March 1998

Abstract

Transport networks are facing new challenges and opportunities because of the explosive growth of data traffic. Besides having to meet the ever increasing bandwidth demand, transport networks need to provide new functionalities for the support of data applications. For example, they should provide elastic data pipes which guarantee minimum bandwidths but also allow expansion when extra bandwidth is available. They should support instant and automatic discovery and configuration of layer-3 nodes which, traditionally, is possible only for LANs. Enhanced transport systems with inherent packet multiplexing can meet these challenges. It will be natural to use Internet protocols and technologies to implement these packet transport systems, and thereby re-use much of the existing Internet infrastructure already widely deployed.

Introduction

Transport systems such as the North American SONET networks are arrangements of multiplexers, switches (often called cross-connects) and transmission links that provide virtual “pipes” between service points. The pipes are administered and relatively static and the complexity of providing them is hidden from the higher layer services.

Most of the traffic carried on the transport network will soon be data. The increasing proportion of data traffic will require the transport systems change to handle data packet flows more efficiently, and also to work well with features of data traffic such as high-level end-system protocols like TCP.

Existing transport systems are circuit-based and have developed as a way of efficiently managing networks based on the time division multiplexing (TDM) voice bandwidth hierarchy. They provide the same fixed bandwidth TDM pipes for both voice and data. This causes inefficiency for data applications which are inherently bursty. The TDM hierarchy has a coarse granularity at high bandwidths resulting in under-utilization and yet data flows are unable to exploit the leftover bandwidth of other data pipes. As data traffic becomes the dominant load for transport systems, high efficiency in carrying data is critical. A few percent of efficiency increase can translate into a large cost saving by avoiding new investment or by capturing new revenue from the existing installed base.

On the other hand, many data services now demand increased performance guarantees, such as the guaranteed bandwidth that is inherent in the TDM system, while not wanting to depart from the current open usage style of the Internet. This is a problem beyond just Internet QoS protocols; transport systems, which after all carry all the traffic, can play an essential role in the total solution.

This paper suggests that we use Internet protocols (IP) and technologies to enhance transport networks for efficient handling of data packets, exploiting the rich functionalities of IP protocol suites and their widespread use. ATM, X.25, and Frame Relay have also attempted to address the need of packet transport systems. But these other approaches lack the ubiquity of IP, and require development of much of the infrastructure that IP already has.

Current TDM-Based Transport Systems

Although we talk about networks for voice and data in terms of point-to-point links and mesh topologies, it is important to remember that these are logical views. These logical topologies are overlaid on physical transmission systems using multiplexing technologies such as SONET and the logical topology is implemented by configuring cross-connect points in the physical network. A central office in a voice network will see a direct link to its neighbor but that link will actually traverse many multiplexing and switching points in the transport network. Two Internet routers may see themselves as immediate neighbors without having to understand that there is another transport level network, complete with its own management and recovery systems, which provides this logical proximity.

As depicted in Figure 1, the logical data network of (a) can be implemented by a transport fibre ring with add/drop points (b), or by a mesh with multiplexers and cross-connect switches(c).

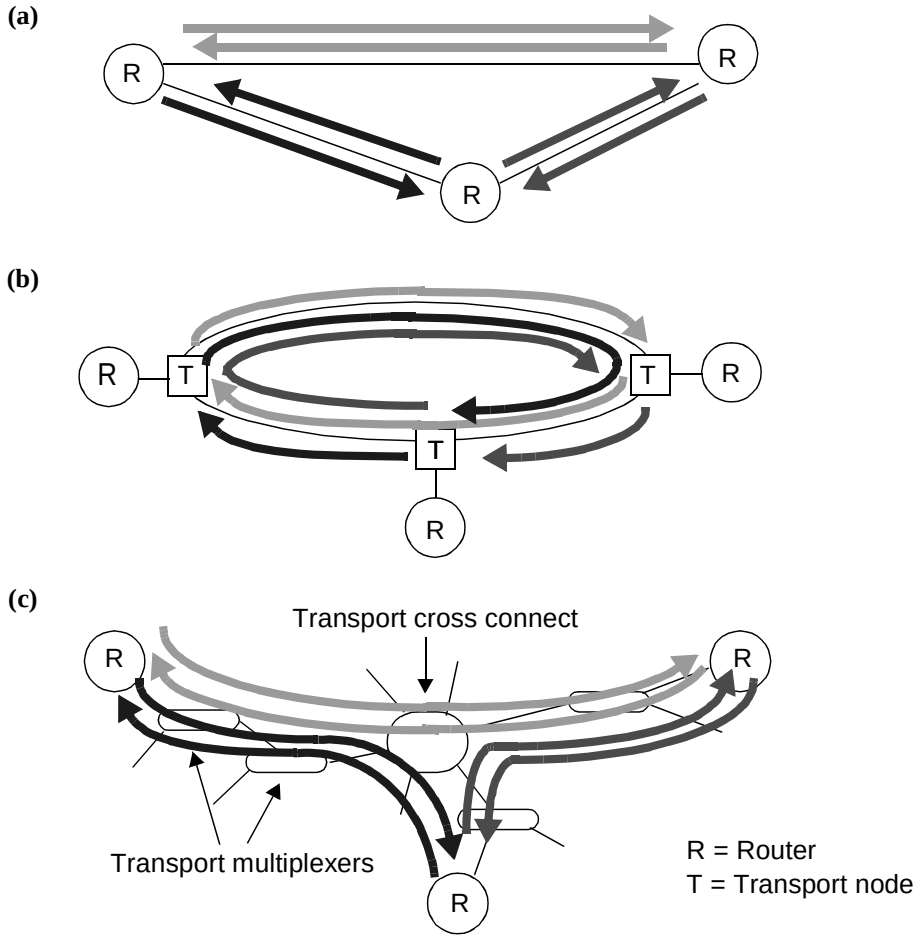

Figure 1(a) Logical data network, and (b) ring-based and (c) mesh-based transport system

Note that, in both the example implementations of Figure 1 (b) and (c), paths pass through multiple nodes but if the bandwidth in a path is not being used it can not be made available to other paths passing through the same nodes. This means that links are usually either under utilized or under provisioned. The waste of bandwidth in a circuit cross-connected network is potentially much higher than the ATM cell tax that is more usually discussed.

The current TDM systems are based on fixed bandwidth pipes whose rates are in a hierarchy derived from aggregating 64Kb/s circuits. This granularity becomes coarser as bandwidth increases, typical rates are 1.5Mb/s, 45Mb/s, 155Mb/s, 622Mb/s, 2.4Gb/s. Using a 2.4Gb/s pipe to deliver 1Gb/s of traffic is inefficient in using network bandwidth.

Goals of Packet Transport Systems

In a transport network which was designed for packet traffic rather than voice traffic, the switches and multiplexers would be packet-based. In TDM systems the bitstreams from one node to another are steered through the network by pre-configured circuit switches at each multiplexing point. In a packet-based system this steering would be done by inspecting the header of each packet to determine its destination.

Customer packets should be encapsulated in transport packets. The networking scheme for the transport packets must be separate and not visible at the customer level. The encapsulation model must be recursive such that traffic can be aggregated at multiple levels in a similar way as it is in the TDM hierarchy.

Instead of pre-configured, fixed-bandwidth circuits there would be a virtual path through the network from any node to any other node based on the destination transport address. No bandwidth resource would be used unless traffic was present. More flexible pricing, based on actual traffic carried, would be facilitated.

When providing services with bandwidth guarantees, a packet transport should be able to emulate the circuit-based mesh in that a defined minimum bandwidth can be allocated between any pair of nodes. However, unused bandwidth should be made available to other flows in a dynamically shared fashion so that a flow can opportunistically exceed its minimum. The guaranteed bandwidth should be allocated at a much more granular level, than in the TDM networks.

In addition, the transport must support many customer networks and must provide protection of one customer from the traffic of another. In a circuit-based transport system there is an established belief in security and isolation between users. The security attributes of the TDM world must be equally demonstrable in the packet-based system.

A powerful feature of local area networks is multicast and broadcast. A good model for this is Ethernet where broadcast between members of the same community is used to bootstrap up communications to new or previously unknown nodes. Within the transport network customers should be given multicast and broadcast capability for their community.

To accommodate the expected need for differentiated services in data networks, the packet transport must also provide for guaranteed levels of delay and packet loss. The Resource Reservation Protocol (RSVP) [1] defined for IP networks can be re-used in the packet transport network to allow the set-up of paths with a particular performance over and above best effort. Advances in routing to accommodate quality of service can also be re-used. It should be noted that routes through a transport network will be much less subject to change than the public internet making it simpler to ensure paths with a defined quality of service.

Architecture of Packet Transport Networks

Packet transport networks will support more flexible and dynamic allocation of network resources than their circuit-based counterparts. In TDM transport, a circuit is set up between each pair of nodes, as depicted in Figure 2. Each remote node appears as a separate port and traffic can be sent at any time to any remote node. However, the bandwidth of these circuits is then committed whether it is used or not. The bandwidth of the physical access link is partitioned to reflect the committed bandwidths and there is no flexibility.

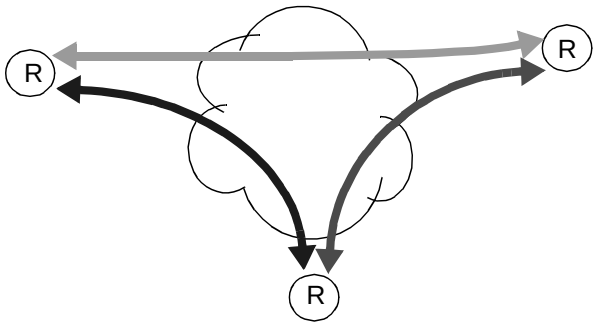

Figure 2. Three circuits in a transport system

In contrast, packet transport allows the access bandwidth to be dynamically allocated. Remote nodes can be represented as logical ports but there is no commitment of bandwidth when this is not needed. The physical access link is fully available for traffic to any destination.

Another attribute of packet transport is automatic configuration. As an example we can look at Ethernet, an existing LAN transport technology. In an Ethernet network, new nodes can announce themselves on joining the network. They each have a unique transport (Ethernet) address and no manual configuration is needed. Any node can search for a resource, such as an IP address, using a multicast protocol and get an answer from the node which owns the resource. In this way it is very simple to build tables of associations.

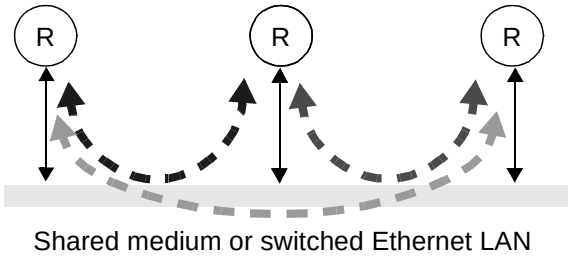

Figure 3. Ethernet as a transport system in implementing the three circuits of Figure 2

We could envisage building a packet transport network out of Ethernet switches which emulate the Ethernet shared medium but allow geographical and capacity scalability. Figure 3 illustrates an Ethernet implementation of the three circuits of Figure 2.

However, there are some drawbacks to using Ethernet equipment as it is in a large, wide area network:

- Ethernet switches learn the paths to end points by observing passing traffic. There are no standard routing protocols equal to those available in the IP domain.
- IP routers are currently farther ahead in implementation for interfacing to high-speed SONET pipes.
- A packet transport should support the encapsulation of multiple customer frames in one transport frame to increase efficiency of network use. Ethernet does not support this.
- Ethernet does not support fragmentation of packets to handle changes of path characteristics.
- Ethernet is limited to a small maximum packet length.
- There is no standard for multiple levels of encapsulation of Ethernet frames.

Thus, we propose that IP equipment and protocols be used for the transport network.

A general architecture for an IP-based packet transport network is depicted in Figure 4. The system has the following subsystems

- *Transport Routers (TR)*. These are standard IP routing equipment but are not visible to the customer network. They route transport IP (TIP) packets, which encapsulate customer's packets [9], from one access point to another.
- *Transport Access Point (TAP)*. This is the interface between the transport system and the customer. At this point the customer's frames are encapsulated in a TIP packet. The TIP destination address corresponds to the location with the matching destination of the customer's frame. The TAP also implements functions such as policing and control of bandwidth, accounting and TCP trunking which is described later.
- *Dynamic Host Configuration Protocol (DHCP) Server.* The transport system uses dynamic configuration as described in [2, 5] to provide transport IP addresses for new access points. The TIP address of the access point will be in the routing table of the local transport router and will thus be propagated through the transport network by standard routing protocols.

The next section looks at getting the benefits of Ethernet at the access level but using IP protocols and technology to implement the underlying network.

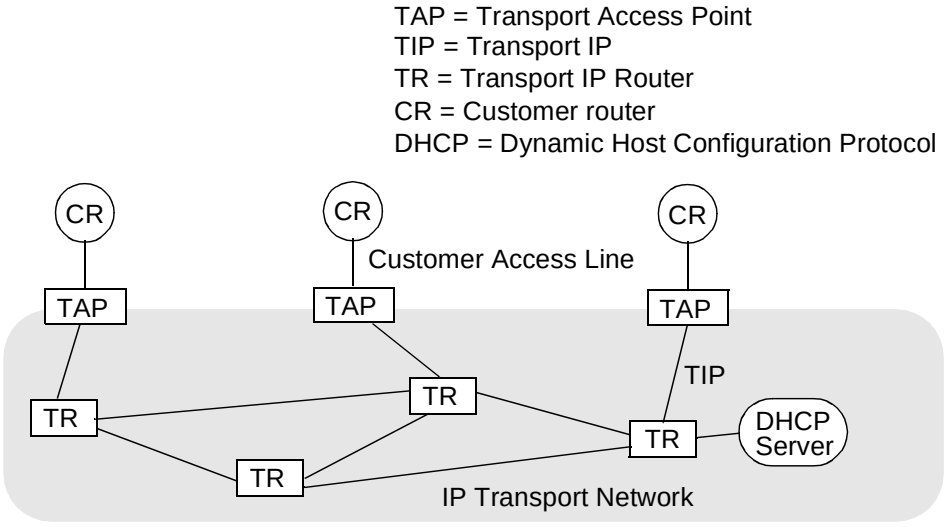

Figure 4. General architecture for an IP-based packet transport network

IP-Based Packet Transport Systems Using the Ethernet Model

In this arrangement, as shown in Figure 5, the IP-based packet transport system emulates an Ethernet network. To the customer's protocols, the transport network is transparent in the sense that it behaves like an Ethernet rather than a routed network. For example, the customer's routing protocols do not involve TRs or TAPs. The customer's IP packets do not increment their hop counts when journeying over the transport network.

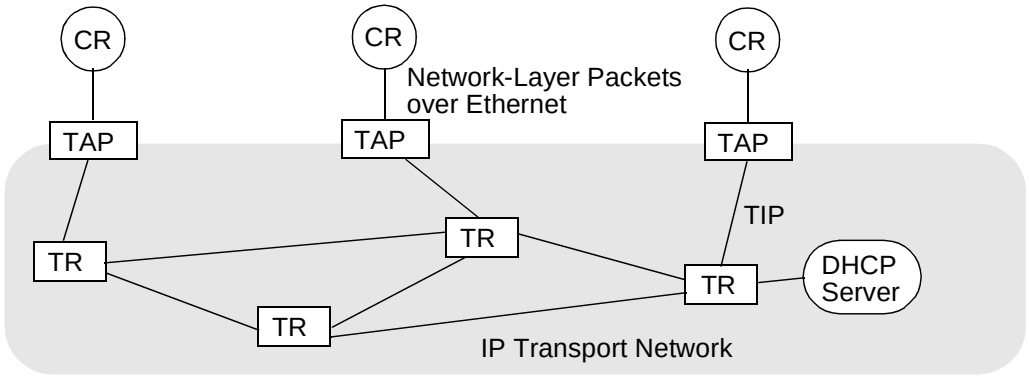

Figure 5. IP-based packet transport system using the Ethernet model

This Ethernet model allows the customer to use the transport system with conventional and easy network management methods. When a new customer router (CR) is attached to the transport network via configuration, the customer can peer it with other CRs already on the system, and run routing protocols such as BGP4 [11] between them to set up their routing tables. The transport network will automatically discover the new CR. That is, the TAP connected to the CR will receive from the CR's interface its Ethernet address, and will obtain from the DHCP server a TIP address for the CR to be used by the transport network. Thus, for every customer Ethernet address there is a corresponding TIP address. Via the broadcast feature, the CR will be able to receive ARP-request messages and respond to them over the transport network, as if it were contacted to an Ethernet network.

When a CR sends a packet to the transport network, the TAP connected to the CR will remove the packet from its Ethernet frame and encapsulates it in a TIP packet. The TIP destination address corresponds to the location with the matching destination Ethernet address of the Ethernet frame. Packets arriving from the transport network are put back into Ethernet frames with the correct source and destination Ethernet addresses. Ethernet broadcast or multicast frames will be translated into IP multicast packets for the transport system and copied to all of that customer's interfaces.

Suppose that a CR has a TIP interface instead of an Ethernet interface. Then the cost of translation between Ethernet addresses and the corresponding TIP addresses at the TAP will be alleviated. Moreover, the TIP interface can be implemented over any layer-2 protocol such as Ethernet and SONET, to allow any of these links to be used. Replacing Ethernet interfaces with TIP interfaces is desirable from the view point of IP-based packet transport systems, but would represent a new set of standards for the networking infrastructure

IP-Based Packet Transport Systems with TAP-ROUTERs

In the absence of a TIP interface standard, it is possible to handle customers IP packets over any layer-2 protocol by having the TAP also function as a customer router. Figure 6 depicts this method of using a TAP-ROUTER for connecting a CR to the transport system. To the CR, the TAP-ROUTER is a neighboring router capable of participating in the customer's routing protocols. At the TAP-ROUTER, IP packets arriving from the CR are encapsulated into TIP packets, and conversely, TIP packets arriving from the transport system are decapsulated into IP packets. Each TAP-ROUTER has an IP address in the customer domain and a TIP address in the transport domain. TAP-ROUTERs will learn the mapping between IP and TIP addresses, in a way similar to how CRs learns the mapping between IP and Ethernet addresses under the Ethernet model.

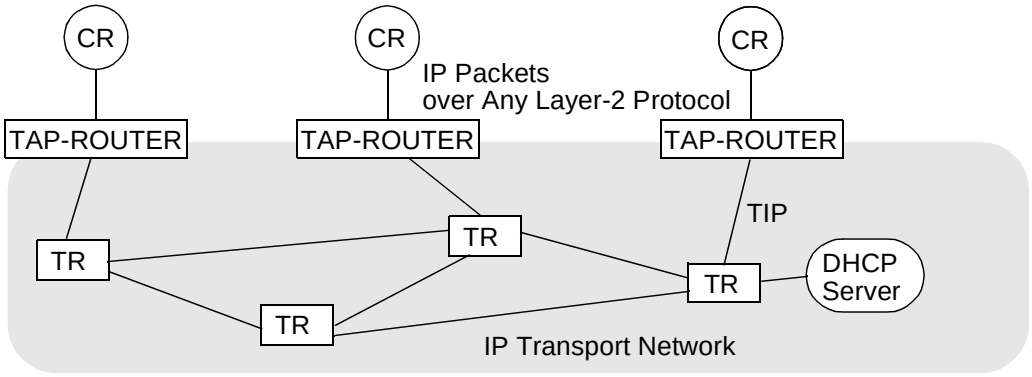

Figure 6. IP-based packet transport system with TAP-ROUTERs, which are TAPs capable of participating in the customers's routing protocols

An advantage of IP is that it has been implemented over a large number of layer-2 links. Use of TAP-ROUTERs allows the transport system to support all these links. Like the Ethernet model of the preceding section, this approach does not require any modification to CRs in order for them to be connected to the transport system.

TCP Trunking

One effect of encapsulating customer traffic is that all the traffic from a source to a destination looks like one flow or trunk (also, in Internet terminology, a tunnel). That is, all the packets over the trunk appear in the transport network with the same source and destination addresses. This reduces the size of tables in the transport routers and the number of contending connections. However, the objective of sharing unused bandwidth means that there will be congestion and therefore packet discard at internal transport routers. The loss of many packets can cause the TCP connection to reduce its bandwidth too much or even to enter the timed out state [7]. Furthermore, we expect that multiple customer packets might be encapsulated in a single transport packet to reduce header overheads. The discard of a transport packet would result in the discard of multiple customer packets with potentially undesirable effects on the customer's TCP connections.

We propose that the transport access points (TAPs) use the transport control protocol (TCP) for the transport trunks which need to compete for bandwidth. The customer packets would be encapsulated in a transport packet with both an IP header and a TCP header. In TCP the sender constantly tests the network to see if more bandwidth is available, and uses the loss of a packet as an indication to decrease its rate. Any lost packets are sent again so that there is a reliable flow of traffic. Thus, these TCP trunks would dynamically adjust their rates when competing for bandwidth, and, moreover, retransmit lost packets when experiencing packet loss. They can be viewed as "virtual paths" capable of providing reliable and flow controlled connections.

An admission process can limit the number of TCP connections over a TCP trunk and also the number of other TCP trunks which may compete for the shared network resources. This will assure some guaranteed minimum bandwidth and maximum delay for the TCP trunk and, thus, each of the customer's TCP connections within the trunk. TCP control will allow the TCP trunk and the TCP connections it carries to use more than their guaranteed bandwidths when extra bandwidth is available. Our simulation has confirmed this elastic nature of TCP trunks.

There are additional advantages of TCP trunking. For example, the TCP receiver will automatically put packets in order in situations where packets may arrive out of order due to route changes or packet striping over multiple links by a router. By limiting the number of TCP trunks, they can compete fairly in bandwidth [6, 8] even when they share the same FIFO in a router.

Quality of Service

Quality of service (QoS) is generally used to describe some controlled or predictable level of delay, packet loss and information transfer rate. Some applications are sensitive to absolute delay or to the variation in delay. These applications tend to have a fairly predictable rate requirement and to use the unacknowledged datagram protocol (UDP). Other applications are "greedy" being less concerned about delay and more with achieving as fast a transfer rate as possible. Greedy traffic flows typically use TCP. Some amount of packet loss is natural in a network carrying TCP but excessive loss (more than a few percent) can result from inadequate or poorly managed buffers as well as from unconstrained numbers of UDP flows which do not adapt to network loading.

To minimize delay for packets it is necessary to give them priority at each point where congestion is occurring. High priority traffic could totally dominate if it were not limited in volume so an overall admission control must be provided to limit the aggregate rate at which this type of traffic can flow. Since the transport network is providing "pipes" between pairs of nodes and is changed by administration rather than dynamic user requests, it is not difficult to assure arithmetically that on any link the sum of all the rates for high priority traffic does not exceed the link rate and indeed leaves enough headroom for other types of traffic. The transport access points would be configured to know the maximum rate for high priority traffic to any destination point and would prevent any excess. By adding a class mark to the encapsulating packets they would be given priority at each transport router without that router having to do any accounting itself. The type of service (TOS) field in the IP header provides for an indication to "minimize delay".

It is also required that the transport system provide some minimum level of bandwidth for the total traffic between any pair of access points. This minimum would typically include both the allocated allowance for low delay traffic and some allocation for greedy traffic. The access point would again do the accounting and ensure that, when there is traffic to send, it is able to achieve the promised minimum. It is also required that the access point facilitate opportunistic use of bandwidth over and above the allocation when it is available in the network.

Accounting at the access point would use a moving average over some window of time. At any time when the average rate achieved is less than the allocated minimum, the access point would mark the packets with a higher delivery priority. For traffic sent opportunistically after the minimum is met the packets would be marked as lower priority for delivery and therefore discardable. If the allocations of bandwidth over the network are done conservatively then the higher priority packets should rarely be discarded. The TOS field in the IP header provides for an indication to "maximize reliability".

Since the access point has control over which packets are marked as discardable it can bias loss toward less fragile connections [8] and maintain a more fair sharing between different flows. It can also use this control to allocate guaranteed bandwidth to preferred classes within the total traffic. The concept of dynamic QoS [4] describes how traffic classification and admission control can be implemented in the absence of application-based reservation by observing the traffic characteristics of flows in order to classify them for treatment.

Keeping the complexity of QoS management at the edges of the network and having simple mechanisms within the network is consistent with the scalable networks and with the trends in the internet world [10].

Discussions

Why not ATM?

ATM has been developed to provide good bandwidth granularity while preserving the performance for delay and delay variation for traditional voice services. The relatively small fixed length packet reflects the need to minimize the time to assemble packets and the latency of converging streams in the network. To provide dynamic sharing of bandwidth for data requires either a lossless flow control method or a packet-aware (and perhaps flow-aware) discard policy. There is some inefficiency in encapsulating packets in cell streams since many data packets are small and the packing factor can be poor. Where the granularity of ATM is not needed because the links are high speed or there is little voice traffic, the use of IP can be more efficient in the use of bandwidth. All the developments of IP networking such as multicast sessions and resource reservation schemes can be re-used.

What about MPLS?

MPLS [3] is being defined with the aim of simplifying packet forwarding in IP networks. It forwards using simple labels instead of the IP header and allows for aggregation of many streams under one label. The labels are stackable to allow the equivalent of multiple level IP encapsulation. It is designed to handle multiple protocols and so could be the basis for a packet transport even though it is being defined to work at a peer level within the Internet. The simpler lookup allows faster header processing and the aggregation results in smaller tables. The price of these advantages is a new set of protocols. Encapsulation in IP provides for less flows and therefore smaller tables; having a separate, and well structured, addressing scheme would produce many of the enhancements to header processing speeds. However, if MPLS demonstrates an advantage, it can be re-used inside the transport network.

Interworking with customer networks

We are proposing IP-based packet transport as a direct replacement for the circuit-based bandwidth management where point-to-point links are provided on an administered basis. Certainly, the body of work to date on ATM, Multiprotocol over ATM (MPOA) and MPLS could be tapped to provide simple discovery mechanisms and on-demand resource reservation. In general, however, we feel that the transport system must remain transparent to the networks it supports.

Co-existence

The recursive use of one networking model has obvious advantages. However, we would expect that the IP transport would run on top of partitioned TDM transport and over ATM networks. Many hybrid mixes can be accommodated in today’s heterogeneous networks.

Conclusions

There is always a need for an underlying transport network that supports multiple customer networks. Traditionally that has been provided by the TDM infrastructure. More recently frame relay and ATM have been introduced to improve flexibility of provisioning and dynamic sharing of bandwidth. ATM has some efficiency issues for packet data as well as requiring a new set of protocols. Frame relay also brings its own protocols and does not provide for the performance guarantees that are expected for the future.

There is an opportunity to re-use existing IP protocols and technology to provide a packet transport. Such a transport could support multiple data networks with good bandwidth efficiency and introduce quality of service guarantees that will be needed for new services. In addition, the transport could provide easy and fast configuration for customers. The recursive use of one protocol suite would minimize investment and interworking problems.

It is unfortunate that the development of routing equipment and protocols has, by and large, been focussed on the public Internet rather than on providing general-purpose transport for the large volumes of data that currently are carried on TDM transport systems.

References

[1] Braden, R. (Editor), “Resource ReSerVation Protocol (RSVP),” RFC 2205, September 1997.

[2] Bradner, S. O. and Mankin A. (Editors), “IPng: Internet Protocol Next Generation,” Addison-Wesley Publishing Co., Reading, MA, 1996.

[3] Callon, R. et al., “A Framework for Multiprotocol Label Switching,” draft-ietf-mpls-framework-02.txt, November 1997.

[4] Chapman, A. and Kung, H. T., “Automatic Quality of Service in IP Networks,” Proceedings of the Canadian Conference on Broadband Research, Ottawa, Canada, April 1997, pp. 184-189.

[5] Droms, R., “Dynamic Host Configuration Protocol,” RFC 2131, March 1997.

[6] Floyd, S. and Jacobson, V., “Random Early Detection Gateways for Congestion Avoidance,” Transactions on Networking, August 1993.

[7] Lin, D. and Kung, H. T., “TCP Fast Recovery Strategies: Analysis and Improvements,” Proceedings of IEEE INFOCOM’98 (The Conference on Computer Communications), San Francisco, California, April 1998.

[8] Lin, D. and Morris, R., “Dynamics of Random Early Detection,” SIGCOMM’97.

[9] Perkins, C., “IP Encapsulation within IP,” RFC 2003, October 1996.

[10] Nichols, K., Jacobson V. and Zhang, L.“A Two-bit Differentiated Services Architecture for the Internet,” Internet Draft draft-nichols-diff-svc-arch-00.txt. (Also available at http://ftp.ee.lbl.gov/papers/dsarch.pdf)

[11] Rekhter, Y. and Li, T., “A Border Gateway Protocol 4 (BGP-4),” RFC1771, March 1995.