

Convergence through packet-aware transport [Invited]

John Wei, K. K. Ramakrishnan, Robert Doverspike, and Jorge Pastor

AT&T Labs Research, Florham Park, New Jersey

Luis Aguirre-Torres

Corrigent Systems, San Jose, California

Charles Kalmanek

AT&T Labs Research, Florham Park, New Jersey

Thomas Afferton

Northrop-Grumman, Chantilly, Virginia

RECEIVED 2 MAY 2005; REVISED 3 NOVEMBER 2005;
ACCEPTED 3 NOVEMBER 2005; PUBLISHED 21 MARCH 2006

We provide a comprehensive description of the Packet-Aware Transport Network (PATN) architecture first introduced in [IEEE Commun. Mag. (3) 120 (2004)]. The PATN architecture provides a more efficient transport of data services such as Ethernet and IP than traditional time-division multiplexing- (TDM) centric access network architectures. The PATN architecture builds upon several emerging technologies and improvements in silicon processing to seamlessly converge multiple network functions onto a single platform, optimize network utilization, and reduce operating costs. The PATN architecture is also designed to support all existing TDM voice, private line, and data services. To validate this architecture we implemented a comprehensive test bed that provides a microcosm of local, metro, and long-distance transport networks. We describe this evaluation methodology and present experimental results of key aspects of the PATN. © 2006 Optical Society of America

OCIS codes: 060.4250, 350.0350.

1. Introduction and Motivation

Today's access and metro area networks have evolved from the need to support traditional voice and time-division multiplexing (TDM) private line services. As a result, carriers have historically relied on a hierarchical TDM structure [i.e., Plesiochronous Digital Hierarchy (PDH) and Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH)] as the underlying technology for all services, including conventional voice, private line TDM and new data services such as Ethernet. However, current projections for the near future indicate traffic growth will be dominated by access to packet switches that provide Frame Relay (FR), Asynchronous Transfer Mode (ATM), Ethernet and Internet Protocol (IP) services. This is in contrast with a decline in traditional voice and TDM private line traffic. Currently these packet-switching functions are performed by Layer 2 technology-specific network edge devices that typically also provide a gateway function between the access/metro and the core network of large network service providers. In an effort to minimize access costs and optimize network utilization, service providers are driving to the convergence of these functions to a common platform called the Multiservice Edge (MSE).

Therefore, an increasingly important requirement of access and metro networks is to direct packet-based traffic from customer premises and hand-offs from third party access providers to the MSE.

Additionally, Ethernet has become the dominant link layer technology used in small-to-medium-size campus and enterprise networks. The proliferation of Ethernet in the enterprise market can be mainly attributed to ease of operation and the fact that Ethernet interfaces and network equipment have historically been significantly less expensive than TDM interfaces of similar bandwidth. Together they give enterprise customers the incentive to deploy Ethernet interfaces to their network service provider. Consequently, major carriers are exploring means to provide Ethernet interfaces and services in addition to conventional TDM services. However, the underlying transport technology for the access and metro networks used to backhaul Ethernet traffic to Ethernet switches still relies on traditional TDM transport. This prevents the network service provider from fully realizing the benefits of statistical multiplexing of *bursty* Ethernet packet traffic.

This paper provides a comprehensive description of the Packet-Aware Transport Network (PATN) architecture first introduced in Ref. [1] along with results from experimental validation of this architecture. The PATN architecture provides a more efficient transport of data services such as Ethernet and IP than traditional TDM-centric access network architectures. The PATN architecture builds upon several emerging technologies such as Resilient Packet Rings (RPR), Multi-Protocol Label Switching (MPLS) and Next Generation SONET, as well as continual improvements in silicon processing. The PATN architecture combines these technologies seamlessly to allow a network operator to consolidate multiple network functions onto a single platform, optimize network utilization, and reduce operating costs. The PATN architecture is also designed to support all existing TDM voice, private line and data services. In addition, it presents network service providers with an architecture that allows the introduction of data services such as point-to-point and point-to-multipoint Ethernet switching. Some of the salient features of the PATN architecture are

- Deep channelization and distributed multiplex bundling of traffic (called *grooming*)
- High-density TDM data termination and Idle Packet Suppression (IPS)
- Local Ethernet switching in the access and metro area networks
- Port-mapped MPLS encapsulation of customer data circuits to and from the MSE
- Aggregation and reliable transport of MPLS virtual circuits, enabling statistical multiplexing of MSE-bound data traffic
- Bandwidth fairness and efficiency for the convergence of TDM and data services
- Introduction of a gateway switch between the metro and core MPLS network
- A high-speed packet interface between the gateway switch and the MSE that provides per-customer and per-service class of service differentiation
- Inter-working with a MPLS core (long distance) network for end-to-end differentiated transport service support.

In the next section, we provide an overview of the existing TDM-based metro transport infrastructure, and provide a perspective on its limitations in supporting emerging packet services. In Section 3 we present the PATN architecture and provide details on some of the key features of the PATN listed above. Section 3 also provides an overview of how the emerging packet-oriented services may be supported efficiently in the PATN. Section 4

describes a PATN test bed that we have used in the laboratory to evaluate implementations of equipment that form the key components of the PATN. Also presented are experimental results of important aspects of the PATN.

2. Current Metro Transport Networks

Today's terrestrial telecommunications transport networks in the United States may be loosely described as comprising three segments: access, metro, and core (long distance). The access network connects customers to transport equipment that is owned and operated by a telecommunications carrier. The metro network generally includes transport equipment and facilities in metropolitan areas, but may vary greatly in size, extent and complexity. The core network consists of long-haul switching and transport equipment providing transport services between major cities. Packet switches and routers are typically deployed in large metro areas. However, since packet switches and routers are operationally more complex than transport network equipment, packet-switching equipment tends to be consolidated in a smaller number of locations than transport equipment. Packet traffic is "back-hauled" through the metro transport network to a packet switch or router where it is either switched locally to another customer location within the same metro area, or switched to another core location.

Figure 1 depicts an example of today's metro network. Transmission and multiplexing equipment resides in central offices (buildings) and is connected between offices by fiber optics. Most carriers use SONET/SDH today. Furthermore, in metro networks, transport equipment is normally configured into SONET self-healing rings to provide extra reliability. Broadband Digital Cross-Connect Systems (B-DCS) are interconnected by the core optical transport network and interface between the metro and core networks. The network elements that serve as the gateway between metro and core networks are shown in colored background in Fig. 1.

When one considers the end-to-end service provided to a customer, this picture becomes more complex, since portions of a customer's transport circuit or packet service may be provided by a combination of carriers, Internet Service Providers, etc. For simplicity, Fig. 1 depicts one carrier that provides both metro and core transport services and packet services, as well as another metro transport provider in the same metro area. This simplified diagram illustrates how integration of packet-aware transport might occur within one carrier's metro and core networks and between two different carriers.

2.A. Network Services

Most customers connect to a carrier's network services via TDM interfaces varying from DS0 to DS3 PDH signal rates, and from STS-3 to STS-192 SONET signal rates in North America (STM-1 to STM-64 SDH signal rates outside of North America.) For example, a Frame Relay customer may connect to a service provider's Frame Relay switch at 256 kbit/s, the equivalent to a 4xDS0 signal rate. For the purposes of this paper, we broadly categorize the services offered by network service providers as follows:

- Circuit-Switched Voice and Private Line TDM services
- Packet transport/connectivity services
- Packet-based voice, Virtual Private Network (VPN) and Internet access services
- Ethernet services

Conventional Voice and Private Line services have for many years dominated the traffic distribution in the access, metro, and core transport networks, and as a consequence

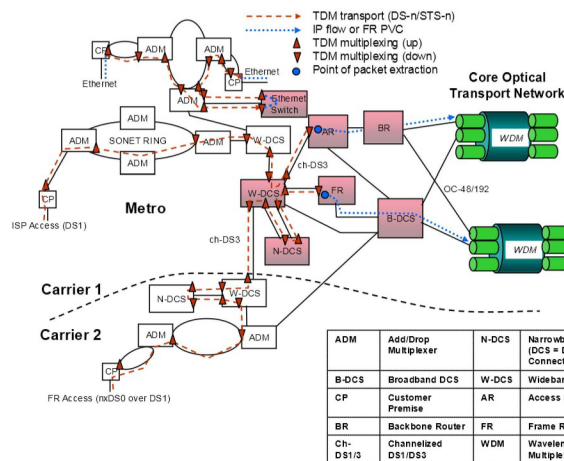


Fig. 1. Metro Transport Network.

have traditionally been the main business focus for many network service providers. As a result, network operators have tried to optimize their transport infrastructure to improve performance and meet increasing customer demands to carry this type of traffic. In contrast, VPNs, Internet access, switched Ethernet services and packet-based voice services have in recent years become an increasingly prominent business opportunity for network service providers; this has catalyzed the transformation of the carrier networks. In this transformation, multiple service terminating network elements (i.e., network edge routers, packet switches and voice switches that form the demarcation point for Service Level Agreements (SLA) and performance measurements) are being converged to a common service edge platform. These dominant trends to packet-based and converged service edges will reduce the importance of designing networks around TDM voice and private line services.

2.B. Limitations in Current Metro Transport Networks

Figure 1 illustrates two packet network access circuits that exemplify some of the limitations of today's network transport for packet services. The circuit in the middle of the figure represents a DS1 interface from a customer router connected to an access router in an ISP network. In the example shown, the metro transport and ISP services are provided by the same carrier. The access router de-encapsulates packets from the DS1 signal and performs Layer-3 routing and switching. The second circuit, shown at the bottom of the diagram, represents an access circuit to a Frame Relay network. This circuit also uses a DS1 interface from a customer router, but the metro carrier provides an nxDS0 service, which terminates at the Frame Relay switch. The Frame Relay switch de-encapsulates the packets from the nxDS0 channel and maps them to one or multiple Permanent Virtual Circuits (PVCs). This access circuit may be provided by a different metro carrier other than the Frame Relay service provider. There are some obvious limitations in the present network structure:

- The metro network has many points where circuits are subject to TDM multiplexing (upward triangle in Fig. 1) and demultiplexing (downward triangle). The TDM multiplexing hierarchy creates circuit bundles at different rates; these must be multiplexed

in order to optimize link utilization by aggregating circuits going to the same destination. Three types of Digital Cross-Connect Systems (DCS) have been developed to handle this task: Narrowband (N-DCS), Wideband (W-DCS), and Broadband DCSs cross connect signals at the DS0, DS1 (or SONET VT-1.5), and DS3 (or SONET STS-1) signal rate granularity, respectively. This is shown in Fig. 1, in which the Frame Relay access circuit traverses an N-DCS and W-DCS to create DS1s (which may then be packed into DS3s) which are bound for the Frame Relay switch. The grooming function at the DCS enable the DS1s (and DS3s) to have a reasonably high density (or fill) of DS0s packed into it.

- As a result of the hierarchical TDM structure, network service providers are unable to take advantage of statistical multiplexing across multiple packet access circuits. The TDM multiplexing hierarchy limits the size of customer access circuits to fractional DS1, DS3, and STS granularities. This means the customer needs to purchase an access circuit large enough to accommodate peak demand, normally resulting in higher cost and lower link utilization.
- The transport of Ethernet traffic over existing transport infrastructure is also inefficient. Currently, in order to provide these services, customer Ethernet frames are encapsulated into TDM circuits and homed to a dedicated Ethernet switch at the gateway office in a “hub and spoke” architecture. Since most of the demand for Ethernet services is intrametro, this architecture is inefficient. This is illustrated in the top circuit flow of Fig. 1, where a point-to-point Ethernet Virtual Circuit is provided between two customer Ethernet interfaces at different customer locations on the same ring. The Ethernet signal is mapped at each Add-Drop Multiplexer (ADM) into a TDM circuit (e.g., a 10/100 Mbit/s Ethernet signal is mapped into an STS-1/STS-3 SONET signal), and routed to a separate Ethernet switch. The Ethernet switch must then de-encapsulate the packets from the two interfaces and switch the Ethernet frames between them.
- Multiple low-rate customer access circuits are aggregated into a single higher-rate signal and homed to a common packet switching platform. This requires most access routers and packet switches to support channelized TDM interfaces, where a higher-rate TDM signal is demultiplexed to its lowest constituent signal level, before packet switching and routing functions may be applied to each customer access circuit. Given the lower customer density, greater power consumption, and space requirements, channelized interfaces on access routers and packet switches have traditionally been more expensive than packet interfaces, radically affecting the current cost structure of the access and metro area networks.

3. Packet-Aware Transport Networks

The PATN architecture is proposed to address the following challenges:

- Reduce capital and operational costs through more cost-effective interfaces, more efficient use of bandwidth and the deployment of a network architecture with fewer network elements.
- Create new service opportunities through packet-aware capabilities, including more granular or variable access connections to packet services.
- Enable transition from a network designed to efficiently support TDM-dominated traffic to one designed to efficiently support packet-dominated traffic.

Figure 2 shows two carrier networks. Carrier 1 has both a PATN and a core network, whereas Carrier 2 has a traditional TDM metro network. Note that the architecture described here focuses upon packet-aware transport confined to a single carrier. The concepts of the proposed architecture can be extended to a multi-carrier environment; however, the interfaces between carriers in such an architecture is beyond the scope of this paper. The Carrier 1 PATN provides aggregation and transport of TDM and packet traffic from multiple central offices (CO) in the metro area into a large gateway CO containing a *Gateway Multi-Service Switch* (G-MSS). The G-MSS separates TDM and packet traffic to be handed off to a B-DCS for intercity TDM transport or an MSE platform performing service terminating functions for Frame Relay, ATM, Ethernet, and IP.

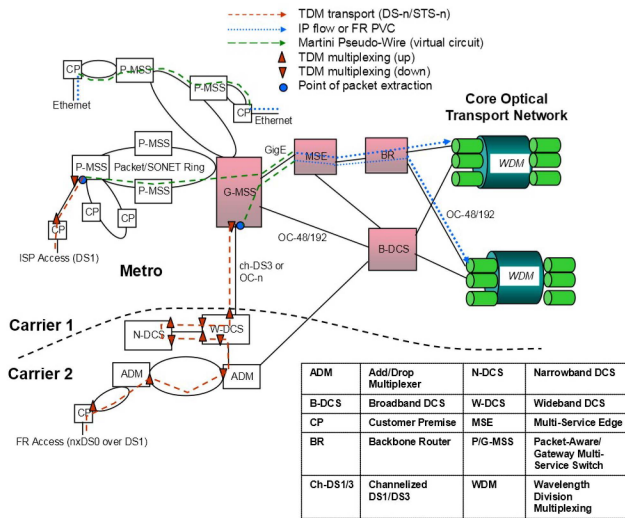


Fig. 2. Packet-Aware Transport Network.

The proposed PATN contains one or more *Packet-aware Multi-Service Switches* (P-MSS), which interface with customers' equipment as well as with other carriers (the G-MSS is a specialized version of a P-MSS). TDM and packet traffic from P-MSS to P-MSS or from P-MSS to G-MSS is carried over metro access rings (typically SONET rings) that terminate on the G-MSS. The P-MSS and G-MSS network elements support both electrical interfaces (e.g., T3, Fast Ethernet) and optical interfaces (e.g., OC-n, GigE). In particular, customers may interface with a Customer Premise (CP) device that connects to the P-MSS via a packet interface or a dedicated TDM circuit. Many physical layer ("last-mile") technologies can be used to transport the CP equipment, such as 2-wire or 4-wire copper, fiber, Broadband Wireless or WiMax. All are consistent with the PATN architecture (e.g., see Ref. [2]).

The PATN eliminates the need to multiplex and demultiplex TDM channels multiple times, as occurs in today's network, thus reducing or eliminating the need for DCSs. Upon their first entry into the PATN at the P-MSS and G-MSS, TDM channels are demultiplexed down to the lowest granularity needed to extract and switch (i.e., groom) the embedded TDM channels. Since this may require demultiplexing a high rate TDM signal down to the lowest constituent signal level, we refer to this function as *Deep Channelization*. This

is followed by TDM data termination and Idle Packet Suppression functions at both the P-MSS and G-MSS.

As mentioned in Section 1, some network service providers are driving to converge various service-terminating network functions into a single MSE platform. The proposed PATN architecture is consistent with this plan, as it aggregates multiple customer access circuits in a single large-capacity link between the G-MSS and the MSE. This reduces the need for channelized TDM interfaces on the MSE, thus optimizing link utilization. The elimination of TDM access circuits requires the creation of Virtual Circuits (VCs) for the transport of MSE-bound customer traffic. The PATN makes use of IETF-defined Pseudo-Wire Encapsulation (PWE) procedures [3] to create customer VCs between P-MSS and the MSE. The PATN multiplexes VCs over P-MSS rings to get the benefit of statistical multiplexing for packet traffic destined to the MSE. This is shown in the middle flow from the CP to the MSE in the Carrier 1 metro network in Fig. 2.

For Ethernet services, the inefficient *Hub and Spoke* architecture and the dedicated Ethernet switch shown in Fig. 1 are replaced with simpler, more efficient Ethernet switching architecture. For example, intra-ring Ethernet services can be transported between P-MSSs over a single ring as Ethernet VCs (top flow of Fig. 2). In contrast, inter-ring Ethernet services (same metro) are switched via the G-MSS. To enable this, the P-MSS and G-MSS include the capability to provide Layer 2 switching (transparent bridging) using the capabilities defined in IEEE 802.1 [4].

The PATN architecture described above relies on various technologies such as RPR and MPLS plus the capabilities of Next Generation SONET, such as Virtual Concatenation (VCAT), Generic Framing Procedure (GFP), and Link Capacity Adjustment Scheme (LCAS) [5], as well as progress in programmable silicon solutions. The following paragraphs describe the use of these technologies in the PATN architecture.

3.A. Deep Channelization and TDM Data Termination

The PATN reduces the need to transport and groom low-rate TDM circuits (e.g., Frame Relay nXDS0 customer circuits) across the access and metro networks. This is accomplished via TDM data termination and is performed at variable-rate customer TDM interfaces at both the P-MSS and G-MSS. At times, low-rate customer traffic is multiplexed into a higher-rate optical signal, requiring wideband and narrowband grooming prior to being forwarded to a TDM data termination module. For example, a network service provider may offer both TDM private line and VPN access multiplexed over a single TDM interface. This requires the P-MSS (or G-MSS) to perform wideband and narrowband grooming in order to separate data traffic embedded in TDM channels from TDM private line traffic. This is typically performed at a cross-connect, such as an STS/VT Time Slot Interchange (TSI) shown in Fig. 3. The lower-rate DS0s would be demultiplexed and then remultiplexed into DS1s, as shown on the right in Fig. 3.

Data traffic embedded in TDM channels is forwarded (normally through a back-plane interface) to a *Framer* device, where the TDM customer channels are extracted from the optical signal. These TDM channels are terminated in a TDM data termination module, to extract the packets.

The Layer 2 data traffic is extracted and separated into logical channels, where idle frames and cells are suppressed. The packets are switched to the appropriate output port through a packet fabric, as shown in Fig. 3.

3.B. Idle Packet Suppression and Statistical Multiplexing

Customers today use different Layer 2 framing technologies such as Frame Relay, ATM, and Packet-over-SONET (PoS), to connect to VPN and Managed Internet Access services.

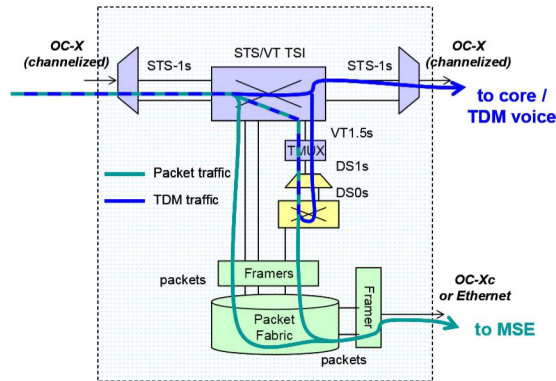


Fig. 3. Deep Channelization, TDM Data Termination and Packet Extraction in a P-MSS and G-MSS.

While these Layer 2 technologies are all amenable to IPS, to illustrate the concept, we focus on PoS and its encapsulation format, as shown in Fig. 4.

PoS packets are encapsulated in a High-Level Data Link Control- (HDLC) like frame delineated by a flag byte. Empty frames (consisting of only flag bytes) are inserted to match the bit rate of the underlying TDM signal. The Point-to-Point Protocol (PPP) [6] is used for link management between the customer interface and service terminating network element. In the PATN, when IPS is performed, the P-MSS (or G-MSS) transmits the PPP packets without the empty frames, thus providing significant efficiency improvements while still maintaining complete service transparency. An alternative to IPS could be to terminate the PPP at the P-MSS. However, if we wish to use the PATN to *transparently* migrate today's MSE access links (i.e., circuits that terminate on TDM interfaces at the MSE), then we must replicate the functionality that now exists at the customer end. In other words, although we replace much of the TDM circuit with a VC, this should behave as a TDM circuit to the customer interfacing equipment (typically DS-n line cards in routers and switches). For example, today's access circuits use the Link Control Protocol (LCP), a subset of the protocol suite of PPP. Some of the functions of LCP are to send messages to establish and tear down the link, as well as to ascertain its functionality through echo requests. With today's TDM access circuit, if a failure occurs somewhere in the transport network between the customer and MSE interfaces, the LCP protocol would detect this failure, and both interfaces would react, e.g., by declaring the link *out of service*. If the PATN terminates the PPP protocol at the P-MSS, then the LCP would operate only between the customer interface and the P-MSS interface. If a similar failure occurs, then the terminating equipment would not be able to use LCP to determine the state of the link. We note that although higher-layer protocols may be used, to be compliant with existing customer TDM interfaces, it is most consistent to transmit the PPP across the VC.

Idle Packet Suppression and early termination of the packet stream carried by TDM channels enables the PATN to derive the benefits of statistical multiplexing. To quantify the potential gain obtainable from statistical multiplexing with the PATN in comparison to carrying packet services using TDM access, we analyzed Simple Network Management Protocol (SNMP) data collected from a large number of T1 links carrying packet traffic from business customers terminating on access routers of a large ISP. We showed in Ref. [1] that the gains measured over an entire day, across several days of a week for different

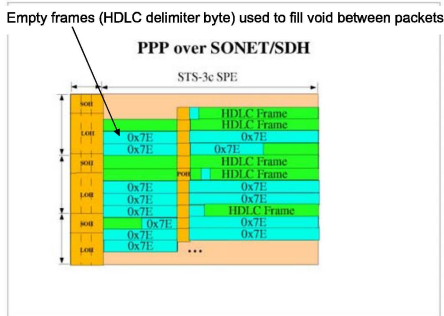
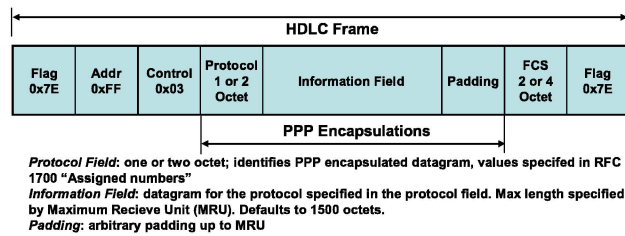


Fig. 4. Packet over SONET encapsulation.

levels of aggregation, ranging from 28 T1 links (that would comprise a fully populated T3) to 336 T1 links (in an OC12 link). The gain ranges from approximately 3.7 to 5.3, depending on the day of the week.

3.C. Pseudo-Wire Encapsulation and MPLS Transport

Deep channelization, TDM data termination, and Idle Packet Suppression are considered the first stages in the PATN architecture. The payload frames extracted after IPS are then forwarded to a next generation network processor, where it is encapsulated and ready to be transported across the PATN. To do so, these packets (e.g., Frame Relay frames or PPP packets) are port-mapped into an MPLS Pseudo-Wire (PW) (also called a *Martini Pseudo-Wire*). Pseudo-wires are introduced in the encapsulation and service emulation framework in the PWE3 Working Group for Layer 2 Transport [3]. The IETF PWE3 Working Group defines two ways of mapping data traffic into pseudo-wires: Port mode and Service mode. The latter requires service-specific as well as service interworking functions to be implemented on the P-MSS and G-MSS, thus increasing the complexity and cost of these network elements. The PATN architecture assumes these functions are only implemented at the MSE.

A pseudo-wire emulates a Layer 2 service (e.g., Frame Relay, ATM, Ethernet, TDM) over a packet-switched (Layer-3) network (PSN) that uses MPLS or IP. Different pseudo-wire encapsulations are defined by the IETF for different Layer 2 data encapsulation technologies. Each pseudo-wire represents a customer Virtual Circuit, and is identified by a VC ID (i.e., an MPLS label).

The application of pseudo-wire encapsulation to the PATN is shown in Fig. 5. We adapt the PWE framework in the PATN for access, and use the term *access pseudo-wire* to describe the pseudo-wire containing customer Layer 2 payloads. An access pseudo-wire in the PATN begins at the P-MSS, traverses the PATN, and terminates at the MSE. A pseudo-

wire is identified by a unique label encapsulating the Layer 2 payload at the ingress of the pseudo-wire. The pseudo-wire label is used to identify the specific outbound interface (whether it is a physical, logical, or virtual interface) to which the native Layer 2 payload should be forwarded upon exiting the pseudo-wire. An access pseudo-wire terminates at the MSE where a “back-to-back” pseudo-wire may be implemented to carry the traffic across the core PSN. The service functionality associated with the pseudo-wire label resides entirely at the MSE interface.

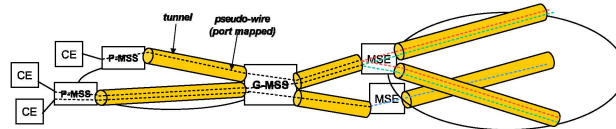


Fig. 5. Access pseudo-wire encapsulation in the PATN.

Figure 5 illustrates how access pseudo-wires are multiplexed over a tunnel LSP, transported across the PATN, and are demultiplexed and delivered to the MSE. An access pseudo-wire may contain multiple Layer 2 connections from a CE (e.g., frame relay DLCIs.)

At the MSE, a virtual interface defines a “port” that terminates the access pseudo-wire (see Fig. 6.) The virtual interface is attached to a service-specific (e.g., Ethernet, Frame Relay, ATM) processing instance providing the desired service functionality at the MSE. Thus, an access pseudo-wire can transport multiple customer Layer 2 connections to the MSE where they are de-encapsulated and independently attached to Layer-3- or Layer-2-based services based on their respective pseudo-wire label (VC ID).

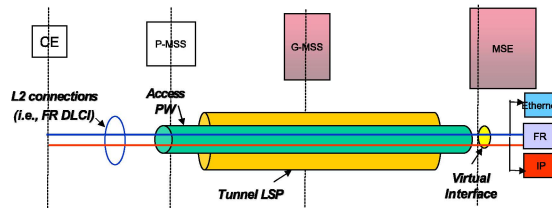


Fig. 6. Port-mode pseudo-wire encapsulation.

This protocol framework and the creation of virtual ports at the MSE-PATN interface enable a service provider to substantially simplify the interface on the MSE platform. Typically, an MSE port to a traditional TDM access network uses channelized interfaces whose implementations are both expensive and resulting in low port density.

Figure 7 shows the functionality of a MSE port card to interface to the access network that is depicted in typical vendor schematic diagrams. Components such as the TDM de-multiplexing and packet termination functions for each of the major rate hierarchies in the TDM domain consume considerable space and power on the line card. This translates to a much higher real cost for the card. By eliminating such functionality (as identified in Fig. 7) in a pure packet-oriented implementation, MSE port cards that interface with the PATN can be made substantially cheaper and much more densely populated.

Having a label associated with each customer VC allows for VC differentiation according to SLAs (e.g., contractually committed information rates) and reliability requirements.

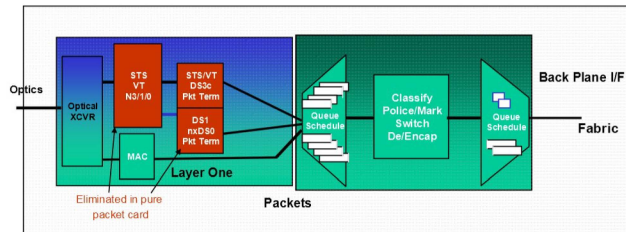


Fig. 7. Multiservice edge line card: interface components.

Between the G-MSS and the MSE, to enable a more efficient implementation of the MSE line card, VCs are further classified and aggregated into MPLS LSPs (i.e., LSP tunnels) according to the Class of Service (CoS) requirements. This aggregation is enabled by incorporating a second level of encapsulation for CoS tunnels between the G-MSS and the MSE. Application of pseudo-wires and their support for QoS in PATN is elaborated further in the following subsection.

3.D. Applicability of Resilient Packet Rings and Next Generation SONET

The bursty nature of data traffic (as opposed to TDM) demands a Medium Access Control (MAC) mechanism to provide fair ring access and efficient bandwidth management for the transport of MPLS customer VCs over the SONET/SDH infrastructure. RPR provides these functions as well as added resiliency, spatial reuse, and an efficient method for CoS differentiation [7, 8].

The RPR MAC enables efficient network utilization by data services. However, the PATN architecture is intended to support both TDM and data services. For SONET/SDH-based metro rings, VCAT may be used to carve out a portion of the channel capacity to carry data traffic exclusively (e.g., VCAT may be used to virtually concatenate two or more synchronous transport signals, independent of the conventional signal hierarchy, in order to match the bandwidth granularity required by (aggregate) customer data traffic.) GFP may then be used to map all data services universally into SONET/SDH (either as RPR MAC frames or MPLS packets.) This enables the multiplexing of multiple data services into a single SONET/SDH STS while still preserving the entire Layer 2 MAC information of an RPR or Ethernet frame. A change in traffic distribution or the allocation of bandwidth to data and TDM may require dynamic resizing of the respective virtually concatenated group (VCG). Link Capacity Adjustment Scheme may be used to perform this resizing and to provide hitless bandwidth modification of any VCG following the failure of any element of the group. The components of a virtually concatenated signal are not constrained to the same path. Rather, common processing and delay equalization occur only at the end points and are therefore completely transparent to the intermediate nodes.

3.E. QoS Support in the PATN

Besides the ability to scale to support a large number of customer ports, support for pseudo-wire class-of-service (CoS) is also an important consideration. In the PATN, the access pseudo-wire CoS would be identified using pseudo-wire encapsulation EXP bits. The P-MSS would map the access pseudo-wire traffic onto Layer 2 CoS (e.g., RPR CoS classes) locally, and mark the EXP bits. Support for existing Frame Relay services, with their current quality of service (QoS) definitions, and ATM QoS for CBR, VBR and ABR services

all require isolation of flows. Traffic in flows going between the P-MSS and the MSE receives CoS treatment subject to the appropriate policing and shaping. Since there may be limits on the number of policers and shapers that can be cost effectively implemented at the P-MSS, G-MSS, or MSE nodes, the PATN allows aggregation of the virtual circuits into larger aggregates with respect to CoS treatment. Thus, traffic coming from multiple P-MSSs are aggregated at the G-MSS onto a CoS-based tunnel toward the MSE.

It should be noted that current MSE interface implementations are already capable of supporting a large number of virtual ports, so that a line card can support the required high density. However, the primary constraint appears to be in the number of distinct policers and shapers that may be implemented on the line card, which will directly affect its ability to support CoS for each pseudo-wire on an individual basis. By allowing for the aggregation of multiple pseudo-wires at the G-MSS into a single MPLS tunnel based on the required CoS for the pseudo-wire, the number of MPLS tunnels that have to be distinctly policed and shaped at the MSE interface is substantially reduced. With this aggregation, the primary function that increases with the number of pseudo-wires at the MSE interface is the lookup to deliver the pseudo-wire to the appropriate service function on the MSE, which is technically much more scalable.

3.F. G-MSS and Interface to the MSE

The G-MSS is a fundamental part of the PATN architecture. It terminates multiple SONET access rings as well as RPR logical overlay rings, and provides an interface between the access/metro networks and the MSE. Moreover, just like the P-MSS, the G-MSS performs TDM data termination, IPS and MPLS pseudo-wire encapsulation on channelized interfaces directly connecting to customer equipment or other network service providers. This is illustrated in Fig. 2, where the G-MSS interfaces to another carrier over a TDM interface.

The G-MSS carries the MPLS LSP tunnels to the MSE. Customer MPLS VCs and LSP tunnels originating from the P-MSS rings or from G-MSS constitute the primary mechanism for connecting customer interfaces to the services implemented at an MSE interface. These are transported to the MSE over a high-speed packet interface such as PoS, Gigabit Ethernet or Next Generation SONET, all of which are considered to be more efficient and less expensive than channelized TDM interfaces. The MSE performs a *look-up* based on the MPLS VC label to perform the appropriate service function on customer circuit. It should be noted that no service function is performed at any point in the PATN architecture, except at the MSE, with the exception of Layer 2 Ethernet transparent bridging.

3.G. Services Supported by the PATN

3.G.1. TDM Private Line Services

The PATN architecture allows a range of implementation alternatives for P-MSS and G-MSS vendors and multiple options for transporting TDM traffic across metro access rings. Since P-MSS nodes implement deep channelization, the demultiplexing and switching of TDM channels is distributed throughout the PATN. One option for the transport of TDM traffic uses circuit emulation to carry TDM traffic through a packet-only access network (such as described in Ref. [9]). In this case, (following one direction of transmission) the P-MSS extracts embedded TDM channels, encapsulates the TDM payload in a format supporting circuit emulation, such as that developed in the IETF PWE3 Working Group, and switches encapsulated packets through the PATN with the appropriate QoS. The G-MSS converts the packets back to TDM and multiplexes the TDM channels onto a channelized TDM interface to the B-DCS (as shown in Fig. 2).

A second option subdivides the capacity on the metro access rings using an RPR *cutout* for packet traffic. Virtual concatenation may be used to allow packet traffic to be carried

in a subset of the STS-1s, with TDM traffic carried in the remaining STS-1s. Additional implementation options relate to the switch fabric in the P-MSS and G-MSS nodes. One example is where the P-MSS and G-MSS contain both packet and TDM fabrics, with the TDM fabric supporting TDM grooming down to at least VT1.5 or DS0 granularity (as shown in Fig. 3.)

3.G.2. Ethernet Services

The PATN is designed to support Ethernet services such as Ethernet virtual circuits (EVC) and Transparent LAN services more efficiently than today's access methods. Ethernet virtual circuit service offers a range of service options such as the port-mapped service discussed earlier. It may also allow VLAN-tagged traffic from customers to be mapped onto different virtual circuits for switching to different locations. Customers typically specify the traffic profile for a port in terms of a committed information rate (CIR), and a peak information rate (PIR) (or alternately as excess information rate (EIR) where $EIR = PIR - CIR$). Customers may also specify the CoS treatment associated with all of the traffic associated with a port. Ethernet virtual circuit services are directly supported by P-MSS nodes, using pseudo-wire encapsulation for transport of virtual circuits between P-MSS nodes and across the core network.

Transparent LAN service is also readily supported by integrating transparent bridging functionality into P-MSS and G-MSS nodes. IEEE 802.17 defines *flooding* procedures, which support broadcast on the ring while avoiding packet duplication even in the presence of ring topology changes. Customer VLAN tag information is carried transparently through the provider-based bridged LAN. The PATN encapsulates customer traffic using a provider-based VLAN tag that is provisioned at the P-MSS and G-MSS nodes. Thus, traffic from all ports associated with a specific customer is bridged over the provider-based bridged LAN within the metro area. Transparent LANs may be extended across the core network by use of the emerging Virtual Private LAN Service standards [10].

4. Packet-Aware Transport Network Test Bed and Evaluation

As part of our effort to refine and validate the PATN architecture presented in Section 3 of this paper, we built a research test bed consisting of equipment from multiple vendors, emulating a microcosm of the access/metro/core environment within which it would operate. This experimental test bed was used to evaluate the different enabling technologies described above and their ability to support the PATN architectural goals. In this paper, we describe experiments and results to demonstrate the following features:

- Differentiated classes of service (CoS) for access pseudo-wires from a P-MSS to the MSE. Multiple classes of transport services with varying SLA requirements can thus be supported over the converged packet-aware metro network.
- Fair ring access for data services, and efficient bandwidth management for the support of data traffic, including applications with elastic bandwidth requirements that are capable of opportunistically using unreserved available bandwidth.
- Reliable transport of both intrametro and MSE-bound customer VCs via pseudo-wire encapsulation and MPLS transport.
- Local Ethernet switching in the access and metro area networks, including point-to-point virtual circuits and transparent LAN service.
- Emulation of TDM circuit services over a pure packet-based network, both across a single metro and across the MPLS core via the MSE enabling effective convergence of data and TDM traffic.

- Packetization and IPS of data traffic embedded in a TDM signal.
- Distributed grooming (early packet extraction and idle packet suppression for packet traffic) and statistical multiplexing of intrametro and MSE-bound customer virtual circuits.
- G-MSS functionality between the metro and core MPLS network using a high-speed packet interface, particularly for aggregated gigabit Ethernet trunks between the G-MSS and the MSE.
- MPLS *labeled* interface between the G-MSS and the MSE, providing aggregation of port-mapped pseudo-wires while maintaining CoS differentiation between customer VCs.

4.A. Test Bed Overview

To set up this test bed, we collaborated with leading-edge transport equipment vendors, mostly in the metro space, whose products rely on different technologies such as RPR, Next Generation SONET, and MPLS. The test bed represents multiple metro rings, interconnected together through an emulated MPLS core network as depicted in Fig. 8.

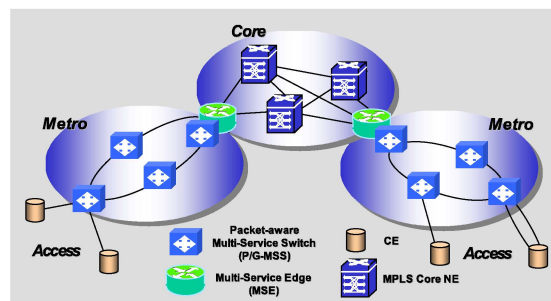


Fig. 8. Test-bed conceptual model.

This test bed supports multiple types of transport service, including those that are common in today's access and metro networks as well as emerging metro network services such as Ethernet switching. The test bed supports provisioning of native TDM circuits at DS1, DS3, and STS-n signal rates that are commonly found in existing metro networks developed for enterprise private line applications and Internet access to network services, using next generation SONET rings. For metro rings based on pure packet technologies such as RPR, support of TDM private line is provided via circuit emulation. Both of these implementations also allow for the possibility of performing IPS on TDM access applications that carry HDLC-like packet payloads. The test bed also supports basic point-to-point Ethernet VCs as well as point-to-multipoint Transparent LAN service both within a single metro ring, and across multiple rings.

4.B. Test Bed Experiments

Our experimentation focused on evaluating the basic enabling technologies, followed by evaluation of all the features and capabilities within a single metro, and then followed by end-to-end intermetro experiments across the emulated MPLS core.

4.B.1. CoS Support and Bandwidth Management Experiments

In our test bed, both RPR ring implementations and Next Generation SONET equipment support CoS differentiation for different types of traffic. While CoS implementations vary among vendors, they all share the following characteristics:

- Support of a high priority (HP) class with minimal jitter that can be used to transport constant-bit-rate traffic such as TDM circuit emulation. This class is defined by a CIR specification, which is also the allowable peak rate.
- Support of one or more medium priority (MP) classes that can be used for bursty traffic requiring a guaranteed minimum CIR, such as variable-bit-rate video and enterprise networking applications. This class is also characterized by a EIR/PIR in addition to the CIR requirement.
- Support of a low priority (LP) class that is used for best-effort traffic, typically used in ISP access where customers share the available bandwidth opportunistically. This class is simply characterized by an EIR specification.

To test and illustrate the different CoS differentiation capabilities of the participating vendors, we conducted the following Ethernet packet transport experiments in our test bed. We set up four point-to-point Ethernet flows, also called Ethernet Virtual Circuits (EVCs), with different priority classes and bandwidth assignments across a 2.5 Gbit/s metro ring, as depicted in Fig. 9. These intraring EVCs are implemented using pseudo-wire encapsulation. The EVCs were connected to packet traffic test equipment, which allowed us to generate traffic and measure the resultant throughput of the injected flows. In this experiment, the LP and MP EVCs were sending a constant load of 1 Gbit/s. In contrast, the traffic from the two HP EVCs were varied from zero to 100% load on the ring. The sum of the CIRs for the HP and MP point-to-point VCs was equal to the capacity of the ring. We recorded the corresponding throughputs for all the flows; the results are shown in Fig. 10.

The measured throughput results showed that the HP VCs could fully satisfy their applied load as long as they stayed within the flows' allocated CIR bandwidth. As the corresponding input loads on the HP VCs increased, bandwidth was initially taken out only from the LP VC, while the MP VC continued to achieve its peak rate (PIR = 800Mbit/s). When the HP VC load was increased beyond about 700 Mbit/s, the MP VC's throughput also started to decrease. However, the MP VC throughput never dropped below its guaranteed CIR of 500 Mbit/s. In this implementation, the LP traffic competed with the excess portion of the MP traffic allocation (EIR) for the remaining uncommitted bandwidth on the ring. Thus, we are able to see that the MP traffic does not completely starve the LP flow. Together, these results demonstrated the CIR and PIR mechanisms across the different classes of service on the metro access ring.

4.B.2. Congestion Control and Fairness Experiments

In addition to the CoS differentiation, different vendors and technologies also implement different policies for the allocation of excess available bandwidth. In RPR, a distributed congestion control scheme is defined and implemented to ensure fairness among individual flows (e.g., see Ref. [8]). The Next Generation SONET ring implementation evaluated employs a centralized bandwidth broker to arbitrate competing traffic demand and allocate bandwidth appropriately across the different classes.

To examine the congestion control and fairness mechanisms, we configured our test bed as shown in Fig. 11. In this setup, the link between nodes C and D was shared by all three LP VCs that terminated at node D. It became congested as we increased the load on each of the VCs. All other links had sufficient capacity to meet the demand. In this setup we also

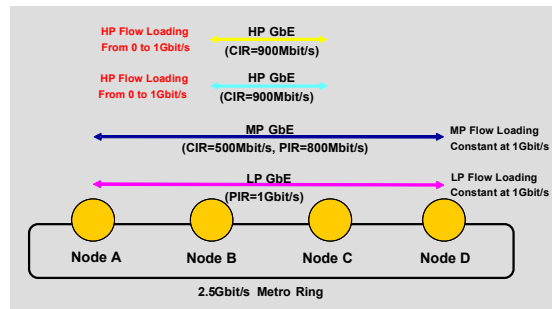


Fig. 9. Ethernet CoS traffic experiment setup.

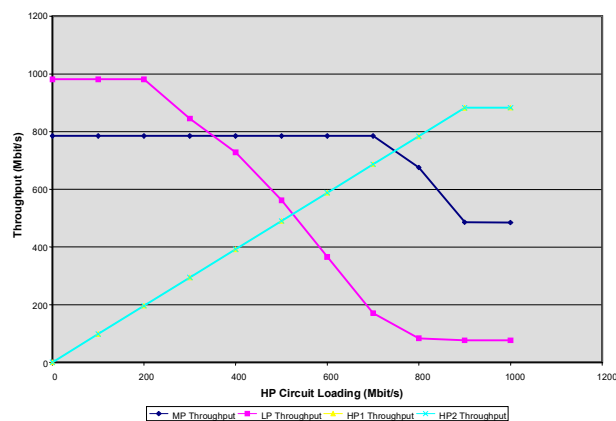


Fig. 10. Ethernet CoS traffic test throughput results.

have a fourth LP circuit between nodes D and B. This flow should not be affected by the congestion control mechanism, as it does not compete for bandwidth over the congested link (C to D). Figure 12 shows the corresponding throughput measurement results.

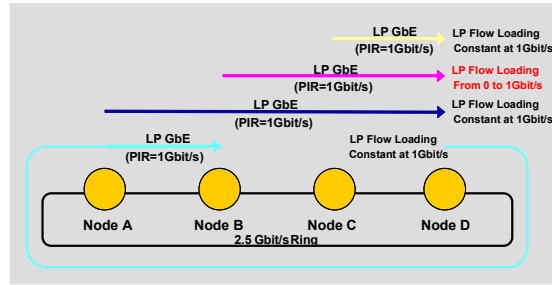


Fig. 11. Ethernet fairness test experiment setup.

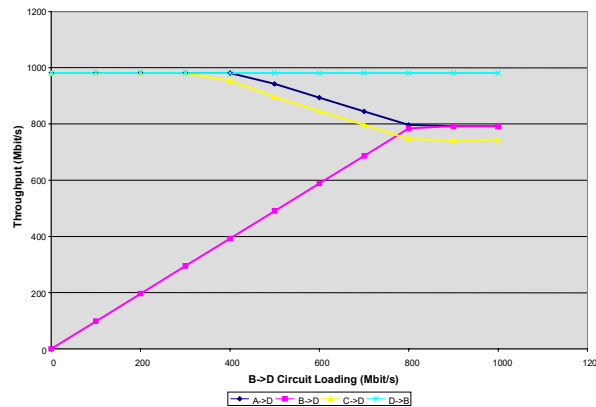


Fig. 12. Ethernet fairness test throughput results.

Observe that as we increased the load of the flow from B to D, all other flows initially achieved their full peak rate, as expected. However, as the load increased to approximately 400 Mbit/s, the throughput of the two competing flows decreased as the link between nodes C and D began to experience congestion. At approximately 800 Mbit/s, load for the flow from B to D, and for higher demand, the three competing flows converged to a roughly equal (to within 6%) share of the uncommitted bandwidth.

The results also showed that when the ring is congested, the competing flows received their fair share of the bandwidth regardless of their specific source location while also avoiding downstream node starvation. Furthermore, the fourth LP flow that did not route over the congested link was not affected. It maintained its full 1 Gbit/s bandwidth throughout the test. This illustrates noninterference and spatial reuse. While this result is encouraging, in our experiments we found that fairness and congestion control in existing vendor implementations are still relatively immature. This is an area that is being actively addressed in the IEEE 802.17 standard working group [7].

4.B.3. TDM Private Line Circuit Emulation

In a packet-aware transport network (PATN), TDM private lines may be supported differently depending on the underlying platform. In an implementation based on Next Generation SONET, TDM traffic will be supported natively over a SONET VCAT group set aside for TDM services. On a pure packet platform such as RPR, TDM support will require circuit emulation. Currently these are mostly proprietary implementations. Standardization of TDM circuit emulation is being addressed in the IETF PWE3 working group [9]. Figure 13 depicts our TDM Private Line circuit emulation experiment set up over a four-node RPR ring of 2.5 Gbit/s capacity. In this experiment, we set up a SONET OC-12c TDM circuit over the ring from node A to node D. The OC-12c signal over this emulated circuit was generated by an OmniBER SONET test set, which allowed us to perform the bit error rate (BER) and jitter measurements.

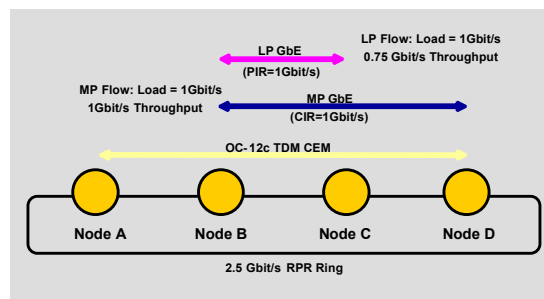


Fig. 13. TDM circuit emulation experiment setup.

We also set up two Ethernet virtual circuits across the intermediate nodes spanned by the OC-12c circuit, in order to simulate an overloaded network condition. One EVC was provisioned as a medium priority (MP) circuit with a CIR of 1 Gbit/s, and the other EVC as a low-priority (LP) circuit with an EIR of 1 Gbit/s. We measured the throughputs for the two background Ethernet flows. With both Ethernet flows (MP and LP) transmitting at the full 1 Gbit/s rate, we observed that the MP flow managed to get the full 1 Gbit/s throughput, whereas the throughput for the LP flow was about 0.75 Gbit/s. This is equal to the uncommitted bandwidth remaining after the HP OC-12c virtual circuit and the MP Ethernet flow demands are satisfied.

In this experiment, with the background traffic to overload the network, we took the BER measurements on the test set for more than eleven days, and we encountered zero bit errors. Additional standard jitter generation, tolerance, and transfer tests were also performed on the interfaces in the presence of the aforementioned background network traffic. In a TDM system, jitter can potentially cause signal bit errors and will reduce the noise margin of the system, making it more prone to errors. To accommodate for this, receivers are designed to withstand and tolerate some degree of jitter in the incoming signal. Jitter may also accumulate and be magnified across multiple timing/retiming imperfections. Both ITU-T and Telcordia have produced standards on jitter test specifications for operational TDM network elements [11, 12]. In our experiments, all the jitter tests passed well within the standard Telcordia GR-253 requirements. The jitter generation and tolerance results were taken for the OC-12 interface used to support an emulated circuit, and they were typical of such interfaces. We also performed jitter transfer measurements, although such measurements are mostly irrelevant in packet-based circuit emulation. In circuit emulation,

the bytes within the TDM signal are encapsulated into packets, transmitted, and buffered at the egress, where the signal byte-stream is recreated and re-timed for further transmission. This process effectively eliminates transfer of jitter from ingress to egress.

Besides jitter measurements, we also measured the end-to-end delay for DS1/3 and OC3/12 emulated circuits across a metro ring. In circuit emulation, delay is incurred in several places. First, there is the insertion delay at the ingress where the incoming TDM signal is packetized. Second, there is the standard packet processing and transmission, and propagation delay in transmitting the packet across the network. Finally at the egress, there is additional delay due to buffering in order to absorb packet delay variation. Both the insertion delay and the play-out buffer delay are configurable. A higher insertion delay will allow larger packets to be formed, thereby reducing the packet overhead. Likewise, a larger play-out buffer can accommodate larger packet delay variation, and is less sensitive to network congestion. However, both of these are achieved at the cost of increasing the overall end-to-end delay. The exact parameters are therefore selected based on trade-offs between bandwidth efficiency, delay variation tolerance, and overall delay. For example, to carry a DS1 signal at 1.5 mbps, one implementation incurs roughly a 10% overhead when packing 8 DS1 frames (equal to a 1 ms insertion delay) into a single packet for transmission. Play-out buffers of 4 to 10 packets are observed in different vendor implementations. In our test bed, we observed one-way delays ranging from 1.5 ms to 8 ms, depending on specific vendor implementation and signal rates. It should be noted that most carriers deem acceptable a total delay of up to 120 ms for Frame Relay services. Furthermore, regardless of vendor implementation, deployment of CEM in a PATN should avoid multiple packet-to-TDM conversions as this would invariably induce multiple insertion and play-out buffer delays. Instead, packetization and depacketization should be performed just once, and be as close to the customer ends as possible.

4.B.4. Idle Packet Suppression Experiments

To gauge the viability and demonstrate the effectiveness of IPS, we performed the following IPS experiment over the TDM test circuit configuration depicted in Fig. 13. For this experiment, we enabled IPS on the emulated OC-12c circuit, and connected it to a tester with a PoS interface to transmit IP packet traffic in PPP frame format over our emulated OC-12c SONET circuit.

For this experiment, we kept the loadings of the EVCs constant at the full 1 Gbit/s rate and varied the PoS virtual circuit loading rate from 0 to 600 Mbit/s. Figure 14 depicts the corresponding throughput results on all three virtual circuits. First, it showed that the IPS functionality was transparent to the client PoS interfaces. Second, the packet flow inside the IPS enabled TDM VC consumed only the bandwidth needed to sustain the injected packet load. This is in contrast to a dedicated TDM circuit, which consumes the full OC-12c bandwidth at all times. Under IPS, unused PoS bandwidth can be claimed by other packet flows in the network. Finally, notice that as the PoS circuit input load increased, bandwidth was taken only from the LP EVC in accordance with the CoS guarantees. These results illustrate the potential statistical gain and bandwidth savings afforded by idle packet suppression.

4.B.5. Ring Protection Switching Experiments

Protection switching for customer VCs in the event of a failure typically follows the priority ordering stipulated in the CoS definitions. In general, the HP class guarantees sub-50 ms switching time, as required by the more stringent TDM application requirements. Beyond this, different vendors' equipment implements protection switching mechanisms defined in the standards corresponding to the technology. For example, the RPR vendor equipment

supports both wrap and steer protection in accordance to the IEEE 802.17 RPR draft. However the implementations differ in their flexibility, which may be provisioned on a per-ring, per-class, or per-virtual circuit basis.

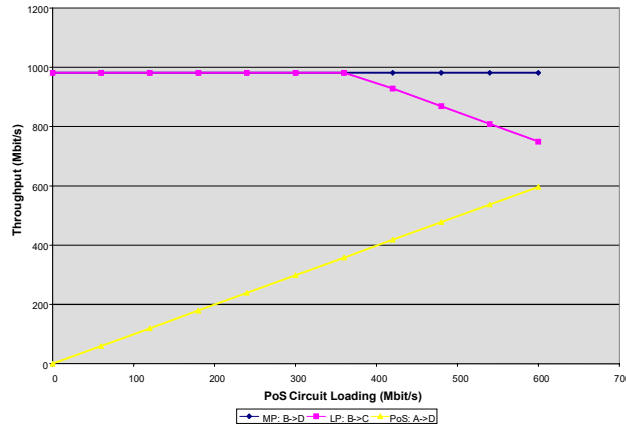


Fig. 14. Packet throughput result under varying PoS loadings.

To test the ring protection implementations in our test bed, we set up the test VCs across the ring as depicted in Fig. 15. We injected two types of failures into the ring: a link failure and a node failure. In both cases, we measured the ring protection switching time under RPR steering as experienced by both a TDM circuit and an Ethernet virtual circuit. We used the OmniBER's Service Disruption Time tests for TDM, and the packet-tester packet loss for the EVC.

We collected data for both an unloaded and a loaded network in order to examine the effects of background loading on the protection switching time on the different classes of services. For the loaded network test, both EVCs were transmitting at their full 1 Gbit/s rates. Upon failure, the TDM and the MP EVC will both be switched over to the opposite direction of the ring, thereby preempting the LP EVC traffic due to insufficient total capacity on the ring. The measurements for the protection switching times under RPR steering for the different services are summarized in Table 1.

Table 1. RPR Ring-Steering Protection-Switching Time Measurements

	Emulated TDM Circuit	MP EVC
Link Failure (unloaded)	≈ 17 ms	≈ 2.1 ms
Link Failure (loaded)	≈ 16 ms	≈ 210 ms–315 ms
Node Failure (unloaded)	≈ 15 ms	≈ 4.5 ms
Node Failure (loaded)	≈ 16 ms	≈ 200 ms–300 ms

In our experiment, we repeated each test case several times to ascertain the stability of the measurements. The switching time for the emulated TDM circuit (via the HP CoS) obtained was relatively stable (± 5 ms or so) and were unaffected by the background (MP and LP) Ethernet traffic. On the other hand, the switching times for the MP Ethernet virtual circuit exhibited a considerable increase in delay under a loaded network condition.

However, they remained in the sub-second range, which we believe should be acceptable for common Ethernet applications. We also measured the resultant throughput of the EVCs both before and after the protection switching events. The corresponding measurements indicated that the MP EVC retained its full 1 Gbit/s bandwidth after the switching event. In contrast, the throughput of the LP EVC was decreased to about 0.73 Gbit/s. This observation is consistent with the expected bandwidth allocation policy for each priority class.

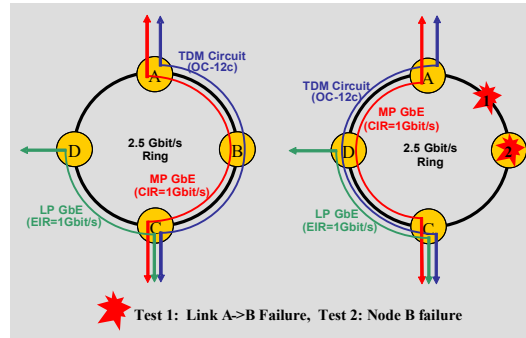


Fig. 15. RPR ring-steering protection-switching time measurement test setup.

4.B.6. Gigabit Ethernet Aggregated Link Protection Experiment

Another protection switching technology that the PATN could take advantage of is Ethernet link aggregation (LAG) scheme. In particular, we experimented with the IEEE 802.3ad-based [13] LAG approach for protecting Gigabit Ethernet links. These Ethernet links may be access trunks between enterprise customer premise equipment (CPE) to the metro network drops, or uplinks from the metro gateways to the Multiservice Edge at the core. The original 802.3ad specification calls for a link failover time on the order of seconds. However, in our proposed architecture, we envision the use of this technology to support various classes of services including TDM circuit emulation applications, which demand a stringent protection requirement. We therefore worked with the vendors to reduce their LAG failover time to the sub-50 ms range, consistent with current practice.

In this experiment, we set up an end-to-end emulated T1 circuit where part of the circuit rides over a two-link aggregated Ethernet trunk, as depicted in Fig. 16. We utilized CPE equipment from one vendor and P-MSS equipment from a different vendor across the two ends of the aggregated link in the setup. Therefore this experiment also demonstrated interoperability of the 802.3ad implementations.

We simulated a link failure on one of the component links, which forced the T1 emulation traffic to switch over to the other link. Using the OmniBER service disruption test, we recorded LAG failover times that range from 5 ms to 32 ms across different trials, and cover both unidirectional and bidirectional cuts.

4.B.7. Intermetro Transport Experiments

Sections 4.B.1 to 4.B.6 described test bed experiments for intrametro transport over the packet aware network. In this section, we describe an extended setup that brings together our proposed service and architectural features in an overall end-to-end setting.

In an end-to-end application, intermetro traffic originating on one metro area and destined for another metro area requires handoff to the network core. Depending on the type

of service, it may be handed off to either a packet transport core for packet traffic or a TDM transport core for TDM traffic. Alternatively, in a purely packet-based architecture, emulated TDM circuits could also be handed off to the packet core for intermetro transport. The handoff uplinks between the G-MSS in the metro and the MSE at the edge of the core network are typically SONET based. However it may also be IEEE 802.3ad link aggregated Gigabit Ethernet, which is capable of delivering an equivalent level of protection at a substantially reduced cost, as we have demonstrated earlier. Intermetro traffic flows are multiplexed onto the uplink trunks via MPLS, or via VLAN mapping if the trunk is Ethernet based. Multiplexing via VLAN mapping could be used for near-term deployment. To address the VLAN scalability issues (from the relatively small VLAN ID space of 2^{12} (4096), various schemes have been proposed. For example, in the “Q-in-Q” approach, stacking of VLAN IDs is used to provide a hierarchy of VLANs, separating the carrier and customer VLAN zones.

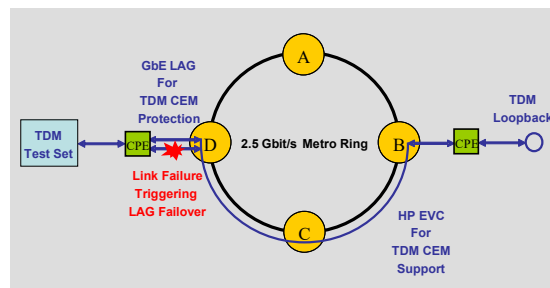


Fig. 16. GbE LAG-based protection experiment setup.

Test Bed Configuration

Figure 17 depicts the test bed setup for our intermetro trunking experiments. Two MSEs were connected back-to-back to emulate an MPLS core network that provides intermetro transport connectivity between two metro rings. The metro rings themselves comprised equipment from the same vendor. We configured an Expedited Forwarding (EF) class and a Best Effort (BE) class LSP tunnel between the two MSEs to aggregate and transport traffic across the metro rings. In this test bed, nodes B and C in both rings serve the role of a G-MSS node. Gigabit Ethernet uplink trunks were used to interface between the G-MSS nodes and the MSEs.

To experiment with the different possibilities, we configured separate MPLS-based and VLAN-based uplinks/handoffs in this setup. Three emulated OC-3c circuits and a best effort EVC were routed over the MPLS-based uplink. Likewise, three emulated DS3 circuits and a best effort EVC were routed over the VLAN-based uplink. Both the TDM and EVC circuits were routed end-to-end through the MSEs using the established LSP tunnels, following conventions outlined in the pseudo-wire encapsulation and transport section. To preserve end-to-end QoS, the circuit-emulated TDM VCs were aggregated and mapped into the EF tunnel, while the background EVCs were aggregated and mapped into the BE tunnel. An additional best-effort EVC was also established between the two MSEs and routed over the BE tunnel, which generated further overloading in the MPLS core. All best-effort EVCs were transmitting at the maximum 1 Gbit/s rate. The uplink MPLS label and VLAN ID mappings were manually assigned via the vendor-supplied EMS. In an actual deployment, an automatic assignment scheme would be more desirable.

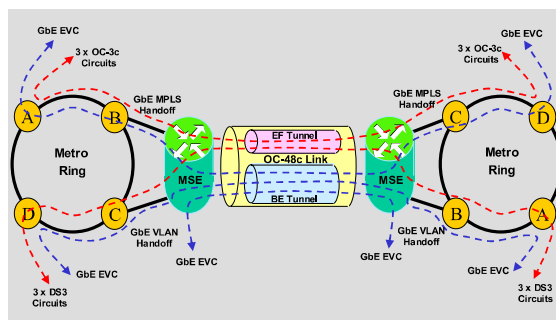


Fig. 17. Intermetro trunking experiment.

Experimental Results

We measured the signal performance for the emulated TDM circuits across the test-bed network under the overloaded condition. We ran the test and took the BER measurements for approximately seven days without encountering any bit errors. We also performed the standard jitter generation, tolerance, and transfer tests under the same overloaded condition and all the jitter tests passed well within the standard Telcordia GR-253 specifications. We also measured end-to-end one-way delay for both the emulated OC-3c and the DS3 circuits, and obtained figures ranging from 1.9 ms to 8.4 ms, depending on vendor implementations and signal rates. As explained before, these latency figures were primarily caused by the ingress insertion delay, and at the egress, play-out buffer delay used for circuit emulation. The additional switching and propagation delays across the MPLS core in our test bed was negligible (approximately 0.1 ms), and the measured TDM circuit delay variations are all insignificant and remain within a range of 0–3 ms. We also tested this setup under various levels of bursty background traffic, and despite the different conditions, the emulated TDM VCs remained stable and operational without any error. This illustrates that with a suitable traffic flow aggregation and bandwidth management strategy, it is possible to provide robust end-to-end TDM service over a pure packet-based infrastructure. Table 2 summarizes the throughput measurements for the best-effort EVCs in our setup.

To understand the figures, observe that the best-effort EVC traffic in this experiment had several bottlenecks along their paths. The first bottleneck was the Gigabit Ethernet uplink. The MPLS-based uplink had an uncommitted bandwidth of about 530 Mbit/s ($1\text{ G} - 3 \times \text{OC}3$) and the VLAN-based uplink had an uncommitted bandwidth of about 860 Mbit/s ($1\text{ G} - 3 \times \text{DS}3$). The tightest bottleneck was at the OC-48c link joining the two MSEs. It had an uncommitted bandwidth of about 1.75 Gbit/s ($2.5\text{ G} - 3 \times \text{OC}3 - 3 \times \text{DS}3$). Since the MSEs implement weighted fair queuing (WFQ) in their schedulers, each of the three competing background BE EVC flows received roughly its proportional share of the remaining bandwidth of 382 Mbit/s, 626 Mbit/s, and 740 Mbit/s, respectively. These figures clearly demonstrated the CoS priority and bandwidth management mechanisms across both the metro and the MPLS core. Such mechanisms enable the convergence of differentiated services over a common packet infrastructure while providing the benefits of bandwidth sharing. These experiments demonstrated that it is possible to offer end-to-end transport of both packet and TDM circuits over a converged network based on our PATN architecture, and an MPLS core that supports differentiated CoS.

Table 2. Intermetro EVC Throughput Measurements

	Applied Load	Available Bandwidth After the 1st Bottleneck	End-to-End Throughput (After 2nd Bottleneck)
EVC over MPLS uplink	1Gbit/s	530Mbit/s	382Mbit/s
EVC over VLAN uplink	1Gbit/s	860Mbit/s	626Mbit/s
EVC across MPLS core	1Gbit/s	N/A	740Mbit/s

4.B.8. Transparent LAN Service (TLS)

In addition to the above point-to-point services, the equipment in our test bed also supported multipoint services in the form of basic TLS. TLS is particularly attractive from a client's perspective, because it is easy to use and simple to deploy and maintain. Direct support of TLS in metro P-MSS equipment is also attractive to a carrier, since it helps to reduce the amount of backhauling and eliminates the need for separate Ethernet switching equipment. At the time of writing, most of the vendor TLS implementations are rudimentary and can only support service within a single metro ring. However, industry efforts like those under the IETF VPLS working group are in progress to develop scalable solutions that can extend support across multiple metro regions over an IP or MPLS backbone.

In our test bed, we experimented with these basic vendor TLS implementations and verified their MAC learning capabilities that enable the efficient support of distributed Ethernet switching without resorting to a centralized hub-and-spoke architecture. Yet, in these experiments we found that the vendor TLS implementations will require more refinement before they become mature enough for actual deployment.

5. Conclusions

The PATN architecture leverages emerging technologies such as RPR, Next Generation SONET, and the continual improvements in silicon processing, and combines them with new architectural innovations such as deep channelization, idle packet suppression, pseudo-wire encapsulation over MPLS transport. It thereby enables a network operator to consolidate multiple network functions onto a single platform to support a seamless transition from TDM-based to packet-based services. In addition, it presents network service providers with an architecture that allows the introduction of new data services such as Transparent LAN Service and Ethernet switching in the wide area.

To provide a "proof of concept" of the PATN architecture, we constructed a PATN test bed, where we evaluated the key technologies and their required interworking. This required us to "incubate" advanced prototype versions of multiple vendors' equipment that would provide the new functionality (some of which are now commercially available). We performed experiments that covered a wide range of features such as the creation of intrametro and intermetro pseudo-wires, differentiated CoS support, TDM circuit emulation, Idle Packet Suppression, packet-based protection switching mechanisms including RPR and Ethernet Link Aggregation, and MPLS LSP Tunnel aggregation and handoff to the Multiservice Edge Switch. We also demonstrated end-to-end service capabilities such as the reliable transport of Ethernet virtual circuits, and emulated TDM circuits. Corresponding results of these laboratory experiments are provided as an existence proof of the PATN architecture.

The PATN architecture addresses simultaneously a network service provider's need to support traditional circuit-oriented traffic, and the desire to grow the emerging Ethernet and packet-oriented services, while reducing its underlying capital and operational investments. As the metro transport services become increasingly packet-centric, the PATN architecture presented in this paper will become an eventuality.

L. Aguirre-Torres' and T. Afferton's contributions to this work were made while they were employees of AT&T.

References and Links

- [1] T. Afferton, R. Doverspike, C. Kalmanek, and K. K. Ramakrishnan, "Packet aware transport for metro networks," *IEEE Commun. Mag.* **42**(3), 120–127 (2004).
- [2] R. Doverspike, K. K. Ramakrishnan, J. Wei, J. Pastor, and C. Kalmanek, "Service-driven networks for packet-aware transport," presented at the Optical Fiber Communications Conference (OFC 2005), Anaheim, Calif., March 6–11, 2005.
- [3] IETF, Pseudo Wire Emulation Edge to Edge (PWE3) Working Group, <http://www.ietf.org/html.charters/pwe3-charter.html>.
- [4] IEEE 802.1 Working Group, <http://grouper.ieee.org/groups/802/1>.
- [5] D. Cavendish, K. Murakami, S. Yun, O. Matsuda, and M. Nishihara, "New transport services for next-generation SONET/SDH Systems," *IEEE Commun. Mag.*, **40**(5), 80–87 (2002).
- [6] IETF RFC 1548, "The Point-to-Point Protocol (PPP)," <http://www.ietf.org/rfc/rfc1548.txt>, Dec 2003.
- [7] IEEE Standard 802.17–2004: "Resilient Packet Ring," <http://www.ieee802.org/17>.
- [8] D. Wang, K. K. Ramakrishnan, C. Kalmanek, R. Doverspike, and A. Smiljanic, "Congestion control in Resilient Packet Rings," in *Proceedings of the 12th IEEE International Conference on Network Protocols (ICNP 2004)* (Berlin, 2004), pp. 108–117.
- [9] M. Riegel, IETF Internet Draft, "Requirements for edge-to-edge emulation of TDM circuits over packet switching networks," <http://www.ietf.org/internet-drafts/draft-ietf-pwe3-tdm-requirements-08.txt>, April 2005.
- [10] L. Andersson and E. C. Rosen, eds., IETF Internet Draft, "Framework for layer 2 virtual private networks (L2VPNs)," <http://www.ietf.org/internet-drafts/draft-ietf-l2vpn-l2-framework-05.txt>, June 2004.
- [11] ITU-T Recommendation G.824, The control of jitter and wander within digital networks which are based on the 1544 kbit/s hierarchy.
- [12] Bellcore Generic Requirements GR-253-CORE, Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria.
- [13] IEEE Standard 802.3 Part 3, "Carrier sense multiple access with collision detection (CSMA/CD) access methods and physical layer specifications," Chap. 43, Link Aggregation.