CONVERGED PACKET TRANSPORT

Evolution of Core Networks: from Circuit to Packet

Table of Contents

Table of Figures

Executive Summary

This white paper summarizes Juniper Networks' vision for the continuous telecommunications shift from circuit to packet.

The challenge of building a cost-efficient and survivable packet-optimized transport layer is studied from different angles, including technology progress, industry experience, and economic incentives. This paper explores the long-term effects of data convergence and presents the case for migration toward transport based on the next-generation packet platforms, such as Juniper Networks® PTX5000.

Introduction

Modern society is increasingly dependent on connectivity and mobility.

In addition to the steady growth of business and consumer data traffic in fixed networks, smart phones are pushing the limits for data transmission over air, and fixed-mobile convergence is moving voice traffic directly onto IP backbones. The use of traditional landline handsets is quickly going the way of VHS tapes, both at home and in the office. Watching a movie hosted in the cloud data center, checking corporate email on the run, and placing a call on a fourth-generation (4G) network are very different media experiences, but they employ the same network-layer IP protocol. This "packet revolution" in the area of communications was not heralded with any press releases but simply issued forth to all but terminate circuit switching at the network's application level, which had powered all means of connectivity for several decades. Whatever we download, see, or hear today is (in most cases) delivered to us in chunks of information of variable length—that is, in the form of network packets.

Standing in the middle of the "new network" storm, it is easy to imagine the worldwide communications industry to be converging around packet transport. If the source and the receiver of virtually any human (or robotic) transaction—be it a voice call, HDTV program, traffic light change, or a file download—invariably process and encapsulate data in variable-length units, it is only logical to expect unified core networks to take advantage of the statistical nature of packet flows to utilize available bandwidth effectively.

Amazingly as it sounds, this is not always the case. Transport layers of some of the largest networks in the world still run over legacy circuit-based SONET/SDH equipment, which was originally developed in the early 1990s and has gradually been upgraded over time to cope with increasing voice and data demand. This conservative approach to transport evolution has ensured persistent stability of network services for a large array of applications, and it has kept the technology pains of early Internet growth away from the majority of business and residential customers. It has also resulted in high overhead on top of packet services and has grown to represent one of the largest opportunities for cost reduction in service provider networks. This white paper intends to bridge this gap, discussing the challenges and opportunities for end-to-end packet transport.

Legacy Core Infrastructure and Associated Costs

Robustness, performance, and cost constraints have historically dictated a unique infrastructure for every digital data application. Voice calls were carried on switched circuits; television broadcasts were delivered using a combination of satellite and terrestrial distribution networks; and pre-Internet data communications were limited in volume. Not surprisingly, packet data was often added as an "afterthought" to existing transmission units. At the time that SONET/SDH were introduced as standardized multiplexing protocols, they became an important step forward and a chance to provide the same underlying network for a large variety of digital data.

Their transport-oriented features and protocol neutrality quickly made SONET/SDH the de facto standard for voice, data, and streaming media applications. Since SONET/SDH essentially are a time-based multiplexing/demultiplexing technology, application-specific equipment has had to adapt to the hierarchy and groom digital data streams into fixed bandwidth frames available in a limited array of sizes (STS1 to STS192).

As a result, by the end of the 1990s, a core network infrastructure typically included a series of SONET/SDH rings carrying digital data from various sources (see Figure 1).

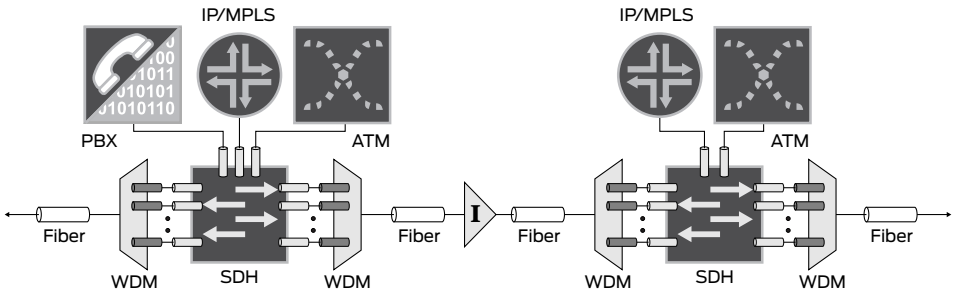


Figure 1: Legacy core network (IP over SONET over WDM)

The protocol neutrality of SONET/SDH allowed ATM, Frame Relay, packet data, and voice calls to be multiplexed over the common core. Failure recovery was usually done with linear Automatic Protection Switching (APS) or ring-based protection [1]. A wavelength-division multiplexing (WDM) layer was later added to provide multiple SONET/SDH signals with the ability to share physical media.

However, the payload filling the SONET/SDH containers has changed dramatically over time.

Voice and video delivery have migrated to an all-IP infrastructure, as did the formerly multiprotocol data networks. In many cases, different digital applications have still maintained independent distribution paths, but internally they nearly always represent data in packet form.

When packetized traffic traverses the SONET/SDH backbone, bandwidth loss from adaption to fixed-size time-division multiplexing (TDM) containers and envelopes can be relatively high. For instance, terrestrial video transport in the U.S. typically uses 270 Mbps video circuits, which does not fit cleanly into the TDM bit rates [2]. For this reason, only two 270 Mbps signals can fit into an OC-12 circuit, instantly wasting about 10 percent of available bandwidth. However, each 270 Mbps video signal is itself a container that can carry compressed SD or HD streams wrapped into an MPEG-2 transport envelope (and possibly multiplexed to form a DVB/ASI signal). The end result is that only 60 percent or less of line bandwidth is available to the actual MPEG video stream compared to what could be delivered over a native packet-oriented transport (for example, Ethernet over WDM). This property of TDM transport is called “grooming inefficiency.”

Elastic data sources (such as IP flows or Layer 2 Ethernet trunks), on the other hand, can adapt themselves to fixed-size containers and fit into a SONET/SDH hierarchy fairly well. When running IP/MPLS traffic over SONET within High-Level Data Link Control (HDLC) framing, effective “goodput” ranges from 94 to 97 percent [3]. Nevertheless, since any given link of a fixed bandwidth is rarely utilized to full capacity, this uncovers yet another downside of TDM transport. All TDM equipment classes cannot detect and multiplex individual packets, so a virtual circuit may run full even if there is plenty of bandwidth available in the same path—an effect called “lack of statistical multiplexing.”

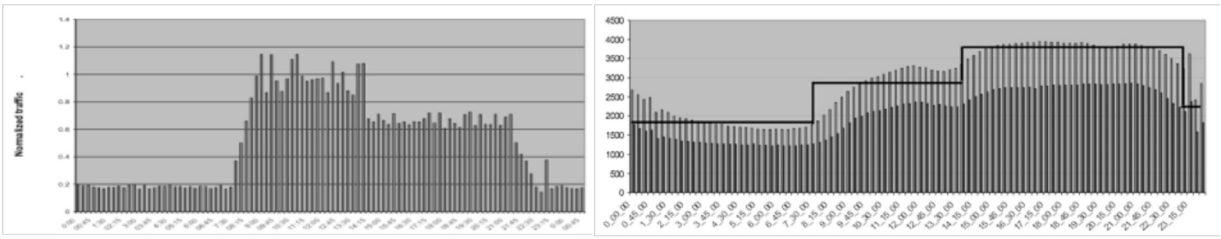


Figure 2: Traffic distribution (24-hour period) at mobile and fixed nodes (source: IST NOBEL)

When running variable-rate traffic over the network backbone, efficiency can be vastly improved with multiplexing at the packet level. Looking at the example in Figure 2, normalized traffic dynamics are shown for mobile (variety of sources) and fixed network nodes [4]. Tidal-chart time scales of dissimilar traffic types make statistical multiplexing a very effective instrument in optimizing the core capacity, as an unused network path can be filled with traffic sharing the same direction. This includes (but is not limited to) multiplexing of wireline and wireless data, traffic originating in regions within different time zones, packet flows with distributed (peer-to-peer) or centralized (data center-bound) affinity, and so on.

Therefore, in the network where traffic running over SONET/SDH transport is predominantly packet-based, TDM transport itself represents an overhead that can be quantified in terms of bandwidth loss. Although the exact figure is highly dependent on network topology and time-bound traffic distribution, it would be safe to conclude that migration from circuit to packet-based transport alone can reclaim a significant percent of the bandwidth otherwise wasted due to grooming and multiplexing deficiencies. Since packet networks typically employ dynamic redundancy (such as MPLS fast reroute or Ethernet link aggregation group), wavelength utilization can be further improved with fast path restoration superseding legacy circuit-level (1:1) path protection [5].

To get an idea of what the savings might be, we can think about a typical optical transport network with 80 10 Gbps channels in its transmission layer. Considering the conservative case, migrating this network from TDM to a native packet-based architecture might recover over half of the total bandwidth and provide an immediate opportunity to resell the spare transit capacity. With wholesale IP prices ($ per Mbps per month) ranging from below $10 in New York to as high as $80 in Mumbai [6], filling the unused capacity goes way further than ordinary "savings." It means that an upgrade toward an all-packet transport could pay for itself many times over.

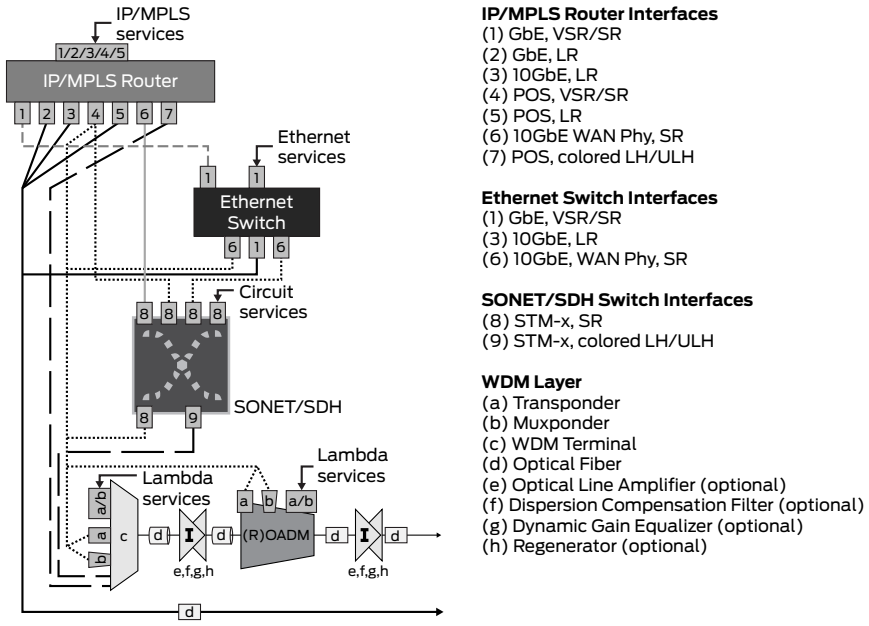


Figure 3: Interface interconnect map in the legacy transport architecture

While the benefits of removing the legacy SONET/SDH layer are fairly straightforward, one can legitimately ask whether a dedicated transport layer is required.

The traditional packet service architecture (Figure 3) has long included an option for interconnecting edge devices (IP/MPLS service routers or Ethernet switches) directly over dark fiber via long-range transceivers. This architecture was later enhanced with a Packet over DWDM option, where colored lasers are provisioned either directly on the services router or via a WDM transponder, allowing wavelength-level signal separation. Such an approach allows multiple packet networks to share the same fiber capacity, with spare lambdas relegated for legacy TDM transport.

To make a decision about whether the dedicated transport layer is needed, it is useful to check Table 1, where running packet services over dark fiber represents the simplest core architecture. In such an instance, a single IP/MPLS network can span over a vast geographical region and deliver L2 and L3 traffic to Internet and VPN customers. An attempt to extend this model into multiple packet networks sharing the same topology would require wavelength-level separation in the WDM layer. With a core based on Packet over DWDM, every overlay network has to be represented with at least one physical port at each node, and the number of wavelengths limits the total number of colocated networks. For example, a metro network might use one wavelength, a long-haul core network might use another, and a legacy TDM network might use a third. Ultimately, the entire wave plan would have to be tailored according to the network's colocation needs, while also factoring resiliency and traffic growth needs—indeed a very complex task.

Table 1: Packet Delivery Options With and Without a Dedicated Transport Layer

Transport Technology	Separation Units	Separation Granularity	Bandwidth Utilization	Number of Overlay Networks (Independent Service Planes)
None (packet over dark fiber)	N/A	N/A	Good	Single (packet) network
Packet over wave over DWDM	Wavelength-based	2.5 Gbps, 10 Gbps, 40 Gbps, 100 Gbps	Good	Low
Packet over SDH over DWDM	TDM channels	SDH hierarchy	Very Poor	High
Packet over OTN over DWDM	TDM subchannels	NxODU0 hierarchy	Moderate	Very high
Packet over packet over DWDM	Label-switched path	Any rate (stat mux)	Very good	Very high

The fact that wavelength-based transport can easily run out of available wavelengths spurs the need for the electrical transport layer, where an additional level of multiplexing allows for growth of services beyond the limits of a wavelength plan. At this high end of the backbone design, bandwidth utilization also matters the most, strongly calling for departure from legacy SONET/SDH topologies.

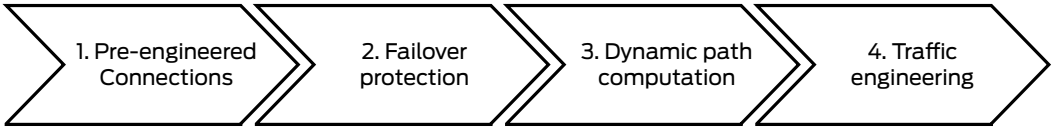
Therefore, a new, packet-oriented architecture is desirable for the transport role. Such a dedicated transport layer is mostly needed in the backbones of Internet service providers (ISPs) and large enterprises, and it represents the focal topic of this document.

Transport Layer Components

Control Plane

The term "control plane" in the transport layer describes the collective intelligence that drives route computation, failure recovery, and traffic management decisions. The oldest transport layer in use (SONET/SDH) is also the simplest one. Being widely available commercially, SONET/SDH transport provides only circuit services with available 1:1 protection and rudimentary control plane capabilities.

The growing mismatch between the needs of packet services and static circuit provisioning has long sparked interest in alternative technologies such as Frame Relay, ATM, and Dynamic Packet Transport (DPT). More recent "transport runner-up" efforts have included Provider Backbone Bridge-Traffic Engineering (PBB-TE) and Transport MPLS (T-MPLS). Without going into the details of why those technologies eventually failed to displace SONET/SDH (or even get into widespread adoption), it is interesting to note that their control plane development took the following general steps:



Although many transport technologies stopped short of going the full way, the time and effort that went into development of their control plans was typically very significant. Moreover, efforts to replace or modify a well-established framework have almost invariably resulted in duplication of effort and compatibility issues, which makes the control plane's de jure (standardization) and de facto (interoperability) statuses extremely important for network design (Table 2).

Table 2: Common Transport Technologies with Control Planes

Transport Technology	Pre-Engineered Path	Path Restoration	Control Plane	Standardization/ Adoption
SONET/SDH	Yes, TDM-based	No	Static	Yes/Widely used
ATM	Yes, cell-based PVC	No	PNNI	Yes/Gradually phased out
OTN	Yes, TDM-based	Based on MPLS	GMPLS	Early stage/Early stage
MPLS/MPLS-TP	Yes, packet LSP	Fast reroute	MPLS	Yes/Widely used

On the requirement side, the modern transport control plane must support the following features:

- Path protection for unicast and multicast traffic (ability to fail over to prearranged path)
- Path restoration for unicast and multicast traffic (ability to restore reachability after multiple failures)
- Traffic engineering (dynamic bandwidth allocation, link affinity)
- Prioritization and preemptive connection treatment

While this list looks deceptively simple, the effort required to fulfill all of the requirements on it should not be underestimated. Moreover, the basic principles of sharing link states, path computation, and traffic engineering are essentially the same, and they have been extensively refined in the industry over the last decade. The rules of convergent evolution dictate similar solutions for similar problems, which means that many transport technologies entered life with "simple" pre-engineered services, only to discover later that the more complex capabilities (such as path computation and restoration) are calling for functions already implemented elsewhere.

Decoupling the control and data planes offers an elegant solution to this problem. Connection-oriented, abstract path establishment, and manipulation appear to be universally useful for all transport equipment types, regardless of the technology underpinning the data plane. Once developed and accepted by the industry, they represent an achievement too large to displace.

This is why, at present, MPLS appears to be the control plane of choice. There is little indication of transport networks widely adopting any other protocol stack in the near to midterm future.

Data Plane

As discussed previously, the transport data plane can consist of virtual circuit (VC)-oriented or packet-oriented hardware. Regardless of the underlying technology, the transport data plane must have the hardware suitable for label stack manipulation and QoS support. Multicast replication is an optional, but highly desired feature as well. On the long-haul interface side, strong optical (LH/ULH) integration is required to reduce dependency on external transceivers and provide preemptive path restoration based on optical signal quality prior to error correction.

The choice between circuit and a packet-based data plane is pivotal to transport network efficiency and cost, so it helps to mind the difference between alternatives from various perspectives (Table 3).

Table 3: Data Plane Options

Technology	MPLS stack support	QoS	Multicast	Relative cost per port
OCS	GMPLS	One class	Rudimentary	Lower (50-90%)
OTN	GMPLS	One class	Rudimentary	Medium (80-110%)
Packet	MPLS/G-MPLS	Eight classes	Full support	Higher (100-130%)

Fault Tolerance and Manageability

When evaluating vendor claims about transport robustness, it is useful to keep the main reasons for network failures in check. Aside from human errors, fault descriptions can be broadly sorted into hardware, software issues, and link failures.

Considering legacy SONET/SDH switches operating at Layers 1 and 2 of an OSI network model [OSI] to be the reference systems, it is interesting to gauge them against more complex devices. As expected, according to data reported by IST NOBEL, the total FIT rate (number of failures for 109 hours) does grow between L1, L2, and L3 platforms (Figure 4).

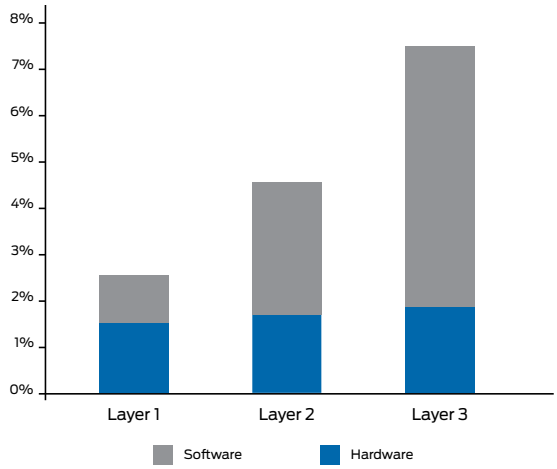


Figure 4: The number of equipment components in failure at least one time per year (source: IST NOBEL)

Yet, the ratio of hardware-specific failures does not change between L1/L2 (for example, SONET/SDH) and L3 (for example, IP/MPLS) equipment significantly, so the observed difference is mostly driven by software. This makes sense because L3 equipment traditionally provides more complex services (Internet routing, VPN, multicast, etc.) compared to the static switching typical of SONET/SDH transport. The extra functionality is driven by software code in the control plane, which, together with an increased variety of use cases, can negatively affect reliability.

However, the software reliability gap reasonably decreases once L1/L2 equipment is retrofitted with a dynamic control plane that employs comparable or the same software complexity as the regular MPLS stack.

In particular, any TDM switches (such as OTN platforms) operating under GMPLS control are not likely to offer noticeable software-driven reliability advantages over their packet-based transport counterparts. Any such differences between equipment from various vendors (if observed) should be attributed to software quality (for example, nonstop routing [7]), coding discipline, and regression methodology [8] rather than genuine equipment class differences.

In addition, it should also be noted that MPLS equipment in a transport profile has a limited function set compared to "service-oriented" IP/MPLS networks and should experience lower failure rates. Moreover, when integration of service and transport IP/MPLS domains is not required, a packet transport platform can operate in its own plane with a dedicated management stratum and remain unaffected by service-level IP or MPLS domains. In this case, a separate Network Layer operates its own label stack and does not maintain routing adjacencies with client IP and MPLS networks (Figure 5). From a client standpoint, transport network remains transparent and looks like a tunnel to upper layers.

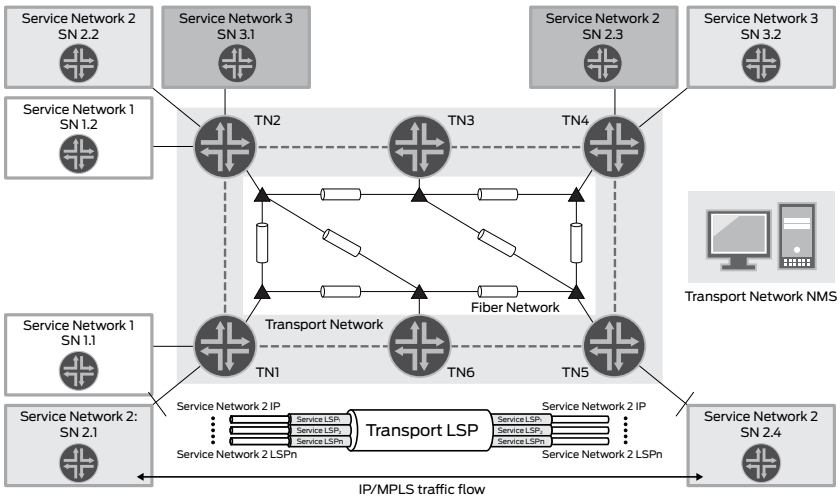


Figure 5: Sample packet transport network with independent control plane

Finally, some aspects of reliability of modern L3 equipment might actually improve over legacy SONET/SDH parameters. One such aspect is detection and processing of link failures.

In "classic" SONET/SDH architecture, the long-haul transponders are external to the electrical exchange switch (EXC), and details of the optical plane operation remain hidden from the SONET layer. Optical signal degradation (if detected) ultimately manifests itself in the SONET/SDH overhead sections and a faltering link is declared down. This declaration triggers reconvergence at TDM (if APS enabled) and packet layers (if TDM restoration failed). Thus, failure recovery function exists on two layers and requires a spare path for backup. Adding agile photonic layer into equation make the picture even worse – now every path needs a spare circuit on both optical and TDM layers and traffic is triple-protected at the cost of three times the overprovision. Yet, path restoration remains reactive and nothing happens until traffic is physically lost.

On the other hand, if packet transport equipment operates directly over colored transceivers (or virtual transponders in a separate optical shelf such as OTS1000), its control plane might detect optical signal degradation straight from forward error correction (FEC) logic and provide restoration before the failing link is declared nonoperational (Figure 6). Such a feature allows link failures to be resolved much faster simply by monitoring the planned margin of detectable and correctable errors (order-dependent on cyclic redundancy code performance) in optical-electrical converters to inform the control plane of the pending need to restore the path. This can improve switchover times and (in some cases) achieve hitless restoration—a result unachievable without synergy between transport and optical layers.

Further, packet transport needs no bandwidth overprovisioning – spare capacity can be used by best-effort traffic with no restrictions, while high-priority LSPs may preempt low-priority paths at any time.

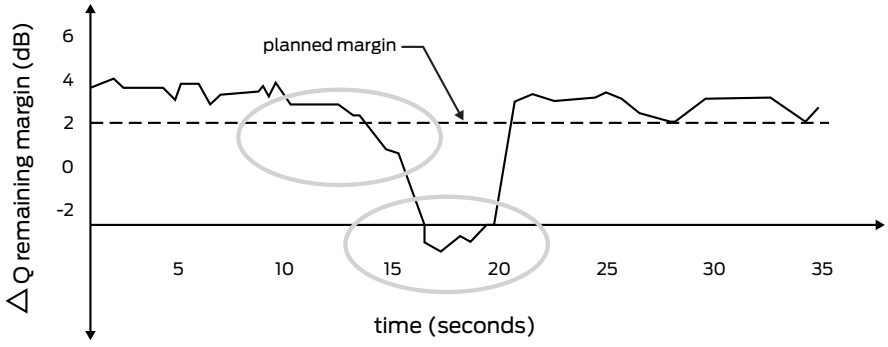


Figure 6: Pre-FEC bit error rate (BER) monitoring in the optical domain integrated into a transport layer

With this in mind, we can conclude that a next-generation transport layer can be built to match and exceed reliability levels established by SONET/SDH equipment. The combination of carrier-grade hardware, thoughtful control plane development, and sound network architecture might actually result in the transport design exceeding the 50 ms convergence target historically reserved for voice traffic. Newer payload types—such as video, trade floor transactions, and gaming—can be even less tolerant to packet loss and delay, thereby being successfully delivered over next-generation transport networks.

Transport Technology Drivers

To create the packet-aware, highly robust, and economical core infrastructure is challenging, but the task is made much easier by leveraging the intellectual property the industry has built over the last two decades. The four most relevant factors in the network "game-changing" arena are:

- Ubiquitous presence of IP as the main network-layer protocol (v4/v6)
- Advanced state of IP/MPLS control plane development with tight synergy between IP and MPLS
- Fast packet processing in the silicon-based data plane
- De facto acceptance of Ethernet as the framing protocol of choice

The combined power of these trends is especially vivid when contrasted against earlier generation technologies. For example, ATM switches demonstrated robust line-rate performance, QoS, and dynamic control plane (private network-to-network interface or PNNI) as early as the mid-1990s. However, the failure of ATM to become the endpoint protocol of choice has created the packet-to-cell adaption layer (segmentation and reassembly or SAR), which quickly became a performance and cost bottleneck at the IP/MPLS network boundary. The new packet network can now provide the same degree of performance and service assurance without cell overhead—packet-oriented properties of an IP control plane are nicely complemented with circuit-oriented services of MPLS.

Another example can be drawn from SONET/SDH development. Long touted for its diagnostics, management, and protection capabilities, the advantage of SONET framing gradually became almost identical to the Ethernet WAN PHY frame (10GBASE-W) defined in IEEE 802.3ae. Combined, the native Ethernet error detection (link, encoding, framing) and alarm/performance monitoring available in the 10GbE WAN PHY today are equivalent to or better than the OC192 frame.

Augmented by Ethernet economy of scale, the new network adhering to 40GbE and 100GbE standards developed by the IEEE P802.3ba Ethernet Task Force is not only capable of robust monitoring and link diagnostics, but it also features port pricing significantly lower than Packet over SONET ports of equivalent capacity. This is attested by the fact that Juniper shipped over 300 long-haul 100GbE interfaces in the first year this product was introduced on T-series routers.

Transport Economy

The first thing to notice when migrating away from legacy architecture (Figure 3) is the fact that the modern handoff between transport and services layers is now almost invariably an Ethernet port with payload expressed in frames. The data frame entering the transport network through such a port can be routed in a variety of layers (provided they have enough intelligence). Multilevel topologies capable of routing decisions are often called "multilayer transport networks."

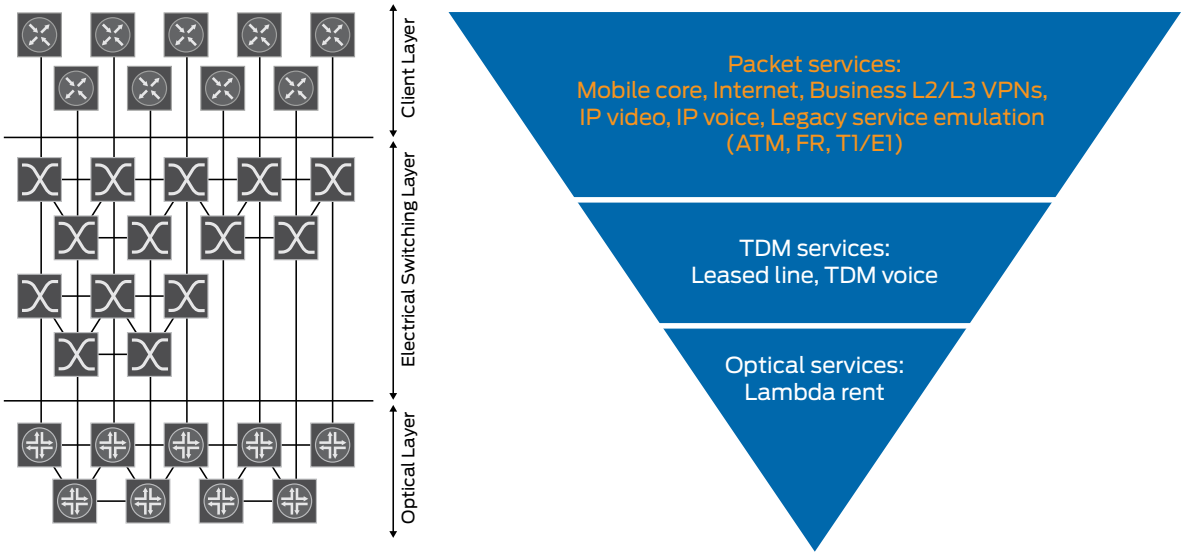


Figure 7: Common multilayer transport structure versus client services

Once the traffic profile, node topology, and cost boundaries are well defined, the task of the transport network design can be reduced to a multilayer optimization (MLO) problem and as such is a popular subject for academic and applied studies. Such studies consider transport that consists of up to three layers—optical circuit switching layer, electrical circuit switching layer, and packet transport layer (sometimes also referenced to as "L2 / Ethernet /MPLS")—each potentially capable of its own routing decisions (Figure 7, left).

While service providers can provision services at any level of hierarchy, a typical mix of ISP revenues favors packet-based services (Figure 7, right), so non-revenue equipment should be minimized or gradually phased out.

With that in mind, the task of multilevel network planning is reduced to finding the most cost-efficient transport topology to serve the capacity requirements for clients. Such an MLO task is always case-specific and uses equipment cost, node topology, and traffic requirements as inputs. But before delving into details of the optimization process, we need to look deeper into constituent transport layers and related architectures.

Optical Switching: Transparent Savings?

At a first glance, the solution for an MLO problem should be trivial. Since optical circuit switching (OCS) offers the lowest cost per bit, the optical layer can be made the sole transport layer, accepting packet and TDM payloads directly from client (edge) devices. Such "transparent" transport can be managed statically (via NMS) or dynamically (with a GMPLS control plane), adapting optical topology to traffic and resiliency needs.

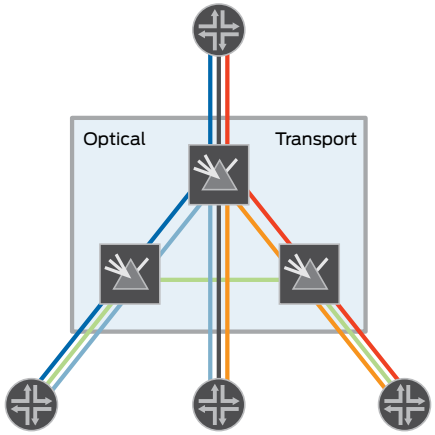


Figure 8: Transparent transport design for N=4

However, even a simple modeling exercise proves that the "fully transparent" network is not economically viable.

This is mostly related to the fact that optical switching data plane can typically switch one lambda speed and multiplex one more (for example, provide 100 Gbps switching plus 10 Gbps multiplexing). With the lowest connection unit being a wavelength, a purely optical transport would require at least [N*(N-1)] high-speed ports for full connectivity between N service routers (Figure 8). For example, an OCS network of only 17 nodes would need a minimum of 272 edge interfaces to provide one-hop optical connectivity.

Fast proliferation of wavelengths alongside network growth is yet another issue of photonic switching—building a full-mesh optical transport quickly proves to be impractical due to exhausting economically acceptable wave plans. While optical packet switching (OPS) and sub-wavelength optical burst switching (OBS) have not progressed beyond feasibility studies, the OCS wavelength usage remains proportional to the square of nodes—O(N2).

In a practical example of cost comparison, one optimization study based on a reference German network of 50 optical nodes and 89 links [14] have demonstrated that deployment of client services directly over 10Gbps wavelength quickly grew impractical upon expansion of demand matrix or an upgrade of underlying optical speeds – all of that would be avoided with packet transport.

A third significant challenge of "transparent networks" is agility—or dynamic characteristics of optical switching technology—where even relatively simple optical topologies can face operational difficulties (see insert).

OCS technology limits

" ...Path computation at the optical/photonic layer is very complex, involving multiple optical parameters that must be reconciled before a light path is assured to work. Algorithms tend to be very complex, require network-wide inputs for successful completion, and <are> not easily implemented <in> realtime.

The expectation is that the restoration times in OCS will be in the range of 1-2 minutes."

IST NOBEL Phase 2, D2.3 p37

In short, we can say that photonic switching alone is not sufficient for transport and has to be augmented with electrical sub-wavelength processing.

Such "hybrid" network design can include optical elements in the form of optical cross-connects combined with electrical switching, thus forming a compromise between "transparent" and "opaque" design options. The exact degree of network opacity should necessarily depend on topology, traffic matrix, and relative equipment costs—in other words, it should be a valid solution for the MLO problem.

OTN: Down the Sub-Slot Road

Recognizing the need for electrical switching layer, some vendors of legacy SONET/SDH gear are now offering new TDM device classes based on revised recommendation G.709 for an Optical Transport Network (OTN), published by ITU-T Study Group 15. Starting life based on original SONET/SDH specs in 2001, by 2009 the OTN standards gained improvements such as a new 1.25 Gbit/s transport unit (ODU0) suitable for Gigabit Ethernet CBR streams; a 111.8 Gbit/s transport unit (ODU4) for 100GbE; and a generic, variable-sized transport unit (ODUflex) with corresponding payload mapping procedure [9].

The main idea behind an OTN was to build the transport layer around the old SONET/SDH principles, but with less grooming overhead. Instead of being forced onto the rigid hierarchy of SONET/SDH rates, payloads could now be mapped into fairly efficient GbE-sized (ODU0) or variable-sized ODUflex (n x ODUk, where k = 0,1,2,3,4) time slots. In turn, service devices can be connected to OTN switches via channelized OTN (cOTN) interfaces—or with regular packet ports such as Ethernet with IEEE 802.1q encapsulation—eliminating the need for a separate port per every destination.

Furthermore, with an agile control plane (such as GMPLS), the OTN transport network can potentially provide dynamic path establishment and protection, alongside some auto-bandwidth capabilities with discrete steps based on ODUflex.

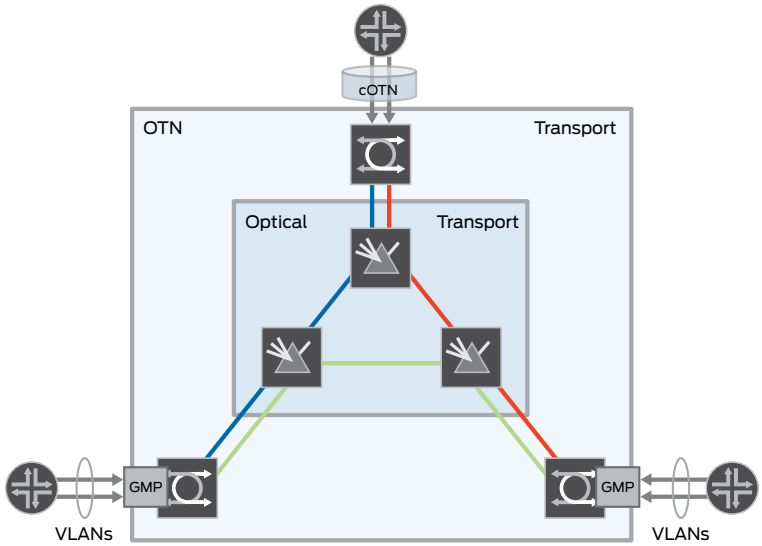


Figure 9: OTN in the transport layer

However, all OTN equipment is still principally based on time slot manipulation and remains incapable of packet recognition. This means that packet payloads need to be adapted to TDM transport—either with special "GMP" blades on OTN devices grooming Ethernet VLANs into time slots—or with channelized OTN (cOTN) interfaces on packet routers and switches (Figure 9). In the networks where most of the payload is packet-based, such adaptation can be a burden.

Furthermore, the cost of packet-to-OTN adaptation can spill over into the control plane as well. Not only do OTN switches require explicit bandwidth planning for O(N2) virtual paths, but they also have to interact with service devices (edge routers), communicating time slot parameters and requesting VLAN policer or cOTN configuration changes via UNI or static configurations. Scaling a transport networks that consists of exponentially growing "soft" circuits thus becomes a management and provisioning challenge.

Finally, an OTN hierarchy remains incapable of statistical multiplexing and multicast. Practically speaking, this means that an OTN switch manages bandwidth utilization much like SONET/SDH equipment. For instance, a GbE port transported in ODU0 always occupies 1.25Gbit/s in transmission path—even if the tributary packet interface is essentially idle.

Such inconveniences can be economically justified only if, for a given network topology and traffic composure, an OTN transport option is convincingly less expensive. And high-capacity OTN systems complete with port-level multiplexers, circuit fabric and packet adaptation cards are not necessarily less complex than an MPLS packet switches.

Packet Transport: Ethernet and MPLS

The phenomenal growth of the Internet made it very clear that the prevailing payloads were expressed in packets, not circuits. This shift marked the beginning of all-packet network transport. Although early generations of transport routers were mostly oriented toward Internet protocol (IP) processing, they later embraced MPLS as a method for conveying arbitrary L2 and L3 frame payloads. The resulting packet transport devices offered no-nonsense benefits in accommodating variable-rate demands and statistical multiplexing, thus using bandwidth efficiently and simplifying the overall topology and traffic planning in the backbone (Figure 10).

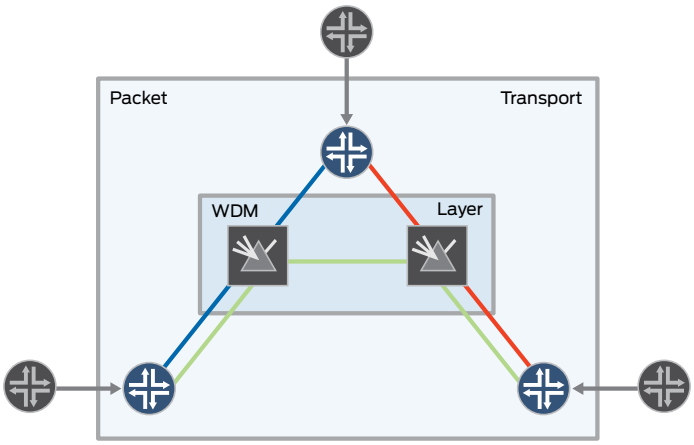


Figure 10: Packet transport network (with OXC shortcut shown in WDM layer)

On the other hand, the rise of fixed and mobile Internet has led to price wars between operators with consecutive erosion of revenues. This creates significant pressure on packet transport vendors to increase capacity of their devices in parallel with lowering the cost of data processing. This lower cost, however, is not expected to compromise equipment quality. This latter reason is exactly why, despite numerous attempts, low-grade Ethernet L2 platforms have failed to gain traction in core and metro networks.

In addition to being cost-effective, packet transport architectures are also frequently confronted with the need to support legacy TDM services. For operators collecting a significant share of revenue from TDM and leased-line services, an all-packet transport was (until recently) not a viable design option.

A third set of requirements for modern transport revolves around resiliency and manageability. While migrating from a (relatively trivial) SONET/SDH infrastructure, many operators wanted to retain the in-service upgrade, diagnostics, and service/transport separation capabilities similar to what they had on the circuit switches.

This unique combination of challenges has led to a situation, where major service providers cannot practically get “god box” systems capable of every packet and circuit operation at attractive cost points and thus have to choose.

And the choice is increasingly tilted towards packet gear as the native adaptation platform for customer traffic.

In the meanwhile, packet platforms are improving driven by a combination of continuous manufacturing process improvements in the semiconductor industry, where the cost and speed of integrated circuits keep improving at a higher pace than those of optical or circuit switching technologies. As a result, today’s packet transport platforms are approximately 100 times faster and roughly 50 times cheaper per bit than they used to be 12 years ago (see Figure 11)[1].

[1] By comparison, photonic interfaces only improved in speed by the factor of 40 for the same period of time while also growing in size. As a result, the real estate and power budget of optical transponders became major problems for transport system designers.

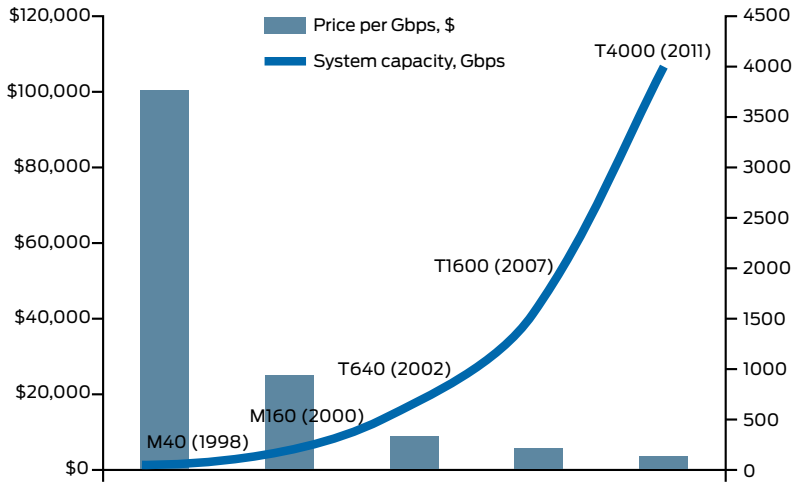


Figure 11: Packet transport evolution, 1998-2010

Today's packet systems also massively benefit from "Ethernet economy," where high-speed PHY chips and pluggable optics are produced in volume once a new Ethernet step becomes a commodity.

Finally, the ongoing software and protocol development brings packet systems features like in-service software upgrade (ISSU) or extensive OAM capabilities —something never available before on this class of devices.

Multilevel CapEx Optimization: Hybrid-Optical approach

From the previous sections, it might be obvious that virtually any transport network can be constructed and fine-tuned using a variety of technologies available in optical and electrical switching layers. The colored optical signal can originate in the electrical switching layer—or be sourced via transponders as part of DWDM shelves. OXCs can be installed or removed in the transit nodes, packet services can be offered over packet or OTN infrastructure, and so on.

Therefore, finding the acceptable solution for the MLO problem is a non-trivial task. In the most complete case, it requires considering cuts through three candidate layers (OCS, OTN, and packet) being active at the same time—a daunting exercise, which quickly grows in computational complexity with the number of nodes. Moreover, any mismatches between models and the real world might result in significant errors, which suggests certain caution at interpreting MLO results.

Nevertheless, running various MLO algorithms (such as described in [10][11]) against modeled connectivity and resiliency requirements can offer some very valuable insights. Such algorithms typically take four sets of input variables:

(1) Relative costs of candidate equipment (IP/MPLS, OTN, and ROADM/WDM)

(2) Node count and fiber topology (including distances)

(3) Target traffic parameters in the form of node interconnection matrix

(4) Resiliency requirements

Many models are also designed to take additional constraints into account—for example, distance ranges for built-in LR optics in packet and OTN devices, the number of available wavelengths, device capacities, and so on. The end result of a typical MLO computation is a solution (or a set of solutions) that demonstrates the relative cost of transport design options within model boundaries—whether packet or OTN switches should be used or where optical shortcuts should be installed.

For example, in a NOBEL IST study of a 68-node European network topology supplied by France Telecom, a network based on traditional core routers (IP over WDM, Architecture 1) was shown to have a definite potential for cost savings when select nodes were fitted with OXCs. At the same time, using a dedicated, cost-optimized packet transport option (IP over Ethernet, Architecture 4) proved to be the most cost-effective solution that did not need additional equipment [5].

Another set of authors from NSN and Swisscom [13] studied a 16-node two-layer IP/WDM reference network to research the minimum link load levels, where introduction of optical switching even makes sense. According to this work, the break-even case happens at 40 percent of link utilization—lower values resulted in elimination of the photonic layer.

Yet a different group of researchers [14] from Atesio and Juniper Networks working on optimization scenario for 17 IP locations out of 50 WDM nodes found that MLO results are sensitive to shifts in media speed and traffic patterns. In their study, the combination of growing wavelength speeds and demands resulted in significant growth of optimal network opacity. In another interesting discovery, the authors concluded that their cost-driven optimization algorithm tends to minimize not IP ports, but the total number of transponders in the network, while opportunistically using MPLS or WDM switching. In resulting 100Gbps network, the optimal degree of network opacity varied between 9 and 17 percent.

So, seemingly, the case for the right transport layer critically depends on the relative cost of equipment, fiber topology, and planned link speeds/utilization. The good news is that at the time of network planning, equipment pricing and fiber topology can be seen as well-formed variables with small error margins. But what exactly do the traffic parameters mean?

Traffic Composure: The Invisible Hand of TCP/IP

Historically, MLO optimizations have focused on traffic models coherent with SONET/SDH networks—that is, all payloads considered to be stationary CBR streams with well-known affinity. This is an assumption that does not hold true not only during long-term traffic shifts (such as shown in Figure 2), but, more importantly, over short-term intervals. For example, let us consider the screenshot of a typical network segment utilization (Figure 12).

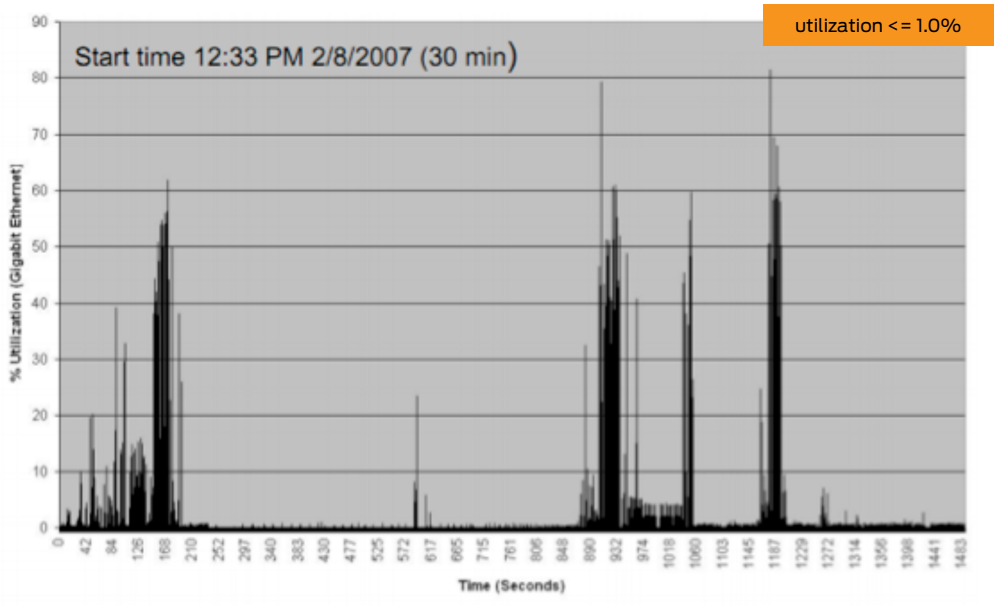


Figure 12: File server link utilization on GbE segment (source: LBNL)

What we can see in the previous picture is that the average link utilization (as measured over "long" intervals ranging in minutes) tends to be fairly light. On the opposite, the "peak" utilization level strongly depends on the interval of measurement—the smaller the interval, the higher the maximum peak recorded. For instance, the period between seconds 900 and 932 in the previous screenshot seems to have at least 40 percent utilization if calculated with a 10-second measurement interval and grows up to 100 percent if using measurement intervals comparable to packet serialization time.

Therefore, even in such a trivial example, it is not straightforward to estimate the traffic requirements for MLO—is this source really sending traffic at 10 Gbps, 10 Mbps, or something in between?

This difference becomes crucial when we consider what happens to this traffic in the transport network. An OTN-based solution, for example, allocates a full ODU0 container to accommodate a 1 GbE link, with two links of this type equating the OC48 circuit. Considering the two tributary GbE interfaces are lightly loaded, delivery of empty frames from New York to Los Angeles can cost as much as $30,000 per month [15]. On the opposite, this spare bandwidth could be reused when using all-packet transport—an effect called statistical multiplexing gain (SMG).

But before we discuss SMG, let's further define the terms of traffic characterization.

As we already noted, the terms “average” and “peak” make sense only together with an interval over which the rate measurement was done. Since the most dire consequence of bursts is potential of traffic loss due to overrun of transit buffers, it makes sense to measure “peak” rate over periods comparable to delay buffer length—which ranges from 50 to 100 ms.

On the other hand, an “average” rate can have many functions depending on intervals— for example, average Internet access rates can be compared between days, months, or periods of the day. For traffic planning purposes, the definition of an “average” interval typically correlates to uninterrupted human attention span—for example, it is defined on the order of seconds.

Now, let’s assume we know that a set of traffic sources (1..r) operates with an array of peak and average rates defined as previously described. Is there any way to predict what would be the new peak and average rates when the r sources are multiplexed together into one link? The answer is definitely “no.”

This is related to the fact that packet traffic (as seen from an intermediate network node) is a product of complex hardware and software interactions on various levels of computer systems, which makes it a non-stationary stochastic process with long-range dependencies [16]. So, at best, we can only describe traffic with probability distribution parameters—but we cannot predict link utilization at any moment with certainty. This leads to fairly interesting conclusions.

First off, when multiplexing several bursty sources into a link smaller than the cumulative sum of respective source links, we can never guarantee that the result of statistical multiplexing would not trigger packet loss.

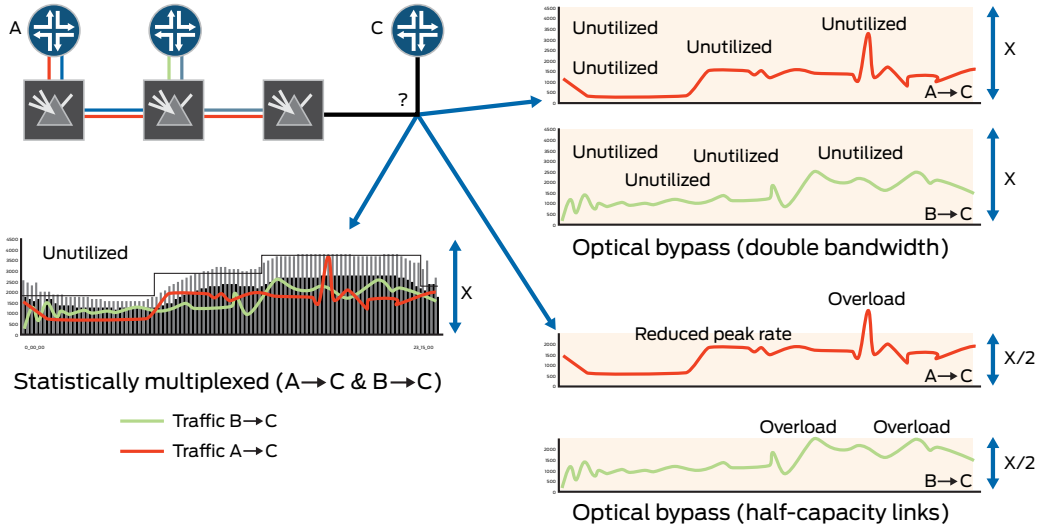


Figure 13: The effects of optical bypass introduced for an intermediate node

The opposite is also true—when demultiplexing traffic onto TDM circuits (for example, when installing an optical bypass), lossless operation can only be guaranteed when the original bandwidth is doubled (Figure 13). Any smaller pipe is not enough to handle traffic bursts at least at some time.

Whether traffic loss at a given peak rate is acceptable or not depends on the service-level agreement (SLA) between the user and the operator. In a typical N-class QoS model, there are “priority” and “non-priority” application classes, each with its own range of requirements and tolerances. Quite often, an operator aims to satisfy priority traffic requirements at (or close to) zero-loss rate—for instance, when delivering IP video traffic to subscribers. Conversely, flow sizing for such payload should be done based on peak parameters of respective traffic streams (for example, VBR encoded HDTV channels). If only priority streams drive the entire network plan, the SMG factor should have no impact on the MLO solution because multiplexed streams have peak rates exactly equal to the sum of tributary peak rates.

However, the same conclusion is not valid in the presence of best-effort applications, such as bulk information transfers.

According to the ATLAS Internet Observatory report [12], over 52 percent of the Internet traffic in 2009 was delivered using the HTTP protocol, which operates on top of closed-loop Transmission Control Protocol (TCP). Additionally, 37 percent of Internet traffic was not classified to the application level but likely included peer-to-peer traffic running primarily over TCP.

Unlike open-loop streaming protocols, TCP is designed to adapt to available bandwidth using a variety of congestion avoidance algorithms. As a result, many best-effort applications such as delivery of emails, files, and Web pages can continuously adapt to available bandwidth, filling the link capacity unused by priority traffic. This opportunistic behavior remains constrained without statistical multiplexing. When multiple sources sharing the same direction are allocated with TDM slots, elastic best-effort traffic at one source has an upper transmission limit equal to its slot rate (Figure 14).

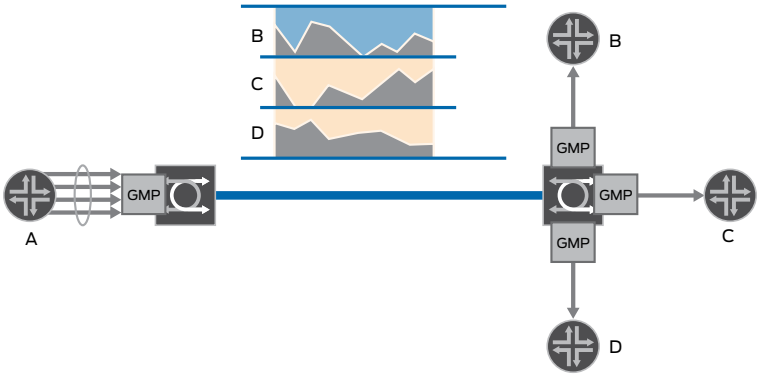


Figure 14: TDM slot allocation for traffic flows in the same direction

On the opposite, the same best-effort traffic is able to gain significantly more bandwidth if the same sources are statistically multiplexed together in one concatenated link of equal cumulative capacity (Figure 15).

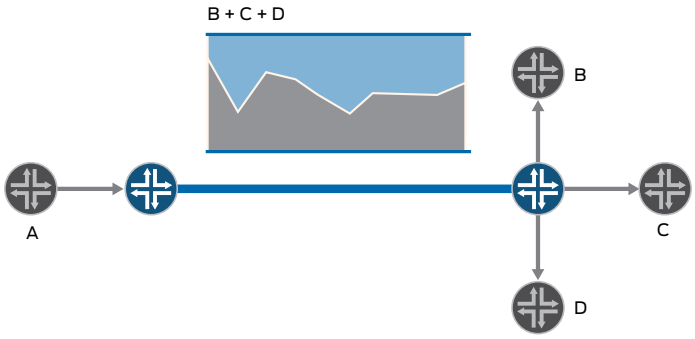


Figure 15: Result of statistical flow multiplexing in packet transport network

Therefore, effective bandwidth gain due to multiplexing is dependent on the number of multiplexed sources and their relative activity. The previous illustration refers to a situation when multiple sources of best-effort traffic do not overlap in time (such as in the situation featured in Figure 2). More complex interaction can result in different bandwidth allocation scenarios, but (on average) it should always remain more efficient than the TDM case shown in Figure 14.

Formalizing the same statement for network planning purposes, nevertheless, is not a straightforward task.

Constructing an efficient analytic traffic model that fits the empirical data for arbitrary network type is a challenging endeavor that remains unresolved in the context of multilayer optimization. Such statistical characterization should necessarily include “slow-changing” parameters such as time-of-the-day dependencies as well as degrees of self-similarity as observed in existing networks [17].

However, we can still make significant improvements to multilayer network modeling algorithms under consideration that certain probability characteristics of traffic behavior can be reduced to static approximations. As an example, we can assume that when aggregating r flows, only K of them can reach peak rate at the same time. If every traffic flow q is described by a mean rate and an additional rate , such as the peak rate is described as , then for K=1 the SMG approximation becomes the following:

$$SMG = \frac{\sum_{q=1}^{r} f'_q + \sum_{q=1}^{r} f''_q}{\sum_{q=1}^{r} f'_q + \max_{q=1...N} f''_q}$$

Higher values for K increase the number of additional rates in the denominator and thus reduce SMG. In a sense, K can be considered a proxy for traffic loss probability—best-effort traffic can be aggregated with K=1, while priority traffic requires K to exceed r for lossless operations. Note that controlled packet loss in elastic traffic groups is normal and helps the closed-loop transport protocols like TCP to gauge the available bandwidth and adapt. This "invisible hand" of TCP cooperates very well with statistical multiplexing to make sure that links are efficiently filled.

In the case of an N-class QoS model, where each flow q is characterized by 2*N rates $\{f'_{q1}, f''_{q1} \ldots f'_{qN}, f''_{qN}\}$, parameter K can be different for priority and non-priority traffic. If the mean rate for best-effort traffic exceeds the peak rate of priority, SMG is simply equal to link capacity divided by the mean traffic rate.

Understandably, any MLO algorithm accounting for statistical multiplexing of VBR sources yields solutions with higher opacity compared to the algorithms dealing with CBR demand matrices. For example, a study by Professor Pietro Belotti [18] demonstrated that the effects of statistical multiplexing (modeled with MPLS transport devices) on network capital cost could be quite remarkable, with improvements ranging from 6.3 to 26.7 percent for a 39-node reference network relative to network design operating merely over TDM transport.

Choosing the Right Solution

control and data planes defines DNA of the transport layer, with combinatorial variety supplied by a diversity of available network equipment. As demonstrated in this document, transport traffic can be switched at the all-optical circuit level, electrical circuit level, packet level—or any combination in between.

However, the "survival of the fittest" rules of economy mean that not all combinations are viable, effectively limiting the set of choices. This said, all network services can be effectively mapped onto a common transport foundation, where the control plane is invariably based on MPLS and the data plane can take packet-only (for example, MPLS), VC-only (for example, OTN), or hybrid form. With this picture in place, any perceived differences between vendor approaches ("have optics, will add packets" or "have packets, will add optics") are reduced to competitive price, density, and the ability to fill optical media efficiently.

To illustrate the previous points, Juniper Networks teamed with researchers from the Department of Mathematical Sciences at Clemson University to assess the CapEx impact of technology choices using a sample network of 20 core and 107 edge nodes [19]. We have focused primarily on capital costs because operational and legacy adaptation expenses are essentially "soft dollars," which are much harder to quantify with analytical methods.

The overall traffic demand in this sample network (8 Tbps) was divided according to a "top-heavy" model, with the top 10 percent of demands holding 58 percent of the total traffic and the bottom 40 percent of demands holding 8 percent of the traffic.

We have also accounted for dynamic traffic characteristics using two variables—K and α designating the number of simultaneous peaks and peak-to-average ratio (burstiness), respectively. Such characterizations should allow a network architect to obtain a range of MLO solutions to gauge model dependency on traffic parameters:

$K \in \{1,2,4,8,16,32\}$, $\in \{1.5,3,4.5\}$.

We also assumed that all photonic optimization was already done and technology choices were limited to an electrical switching layer. This left us with four possibilities: (a) OTN layer added to IP routers layer (bypass model), (b) optimized hybrid MPLS + OTN topology, (c) MPLS-only packet transport network, and (d) OTN-only circuit transport network (see Figure 16). For optimization purposes, we have assigned equipment in different classes with relative price tags expressed in abstract units, with MPLS equipment assumed to carry 30 percent premium over OTN gear of the same capacity.

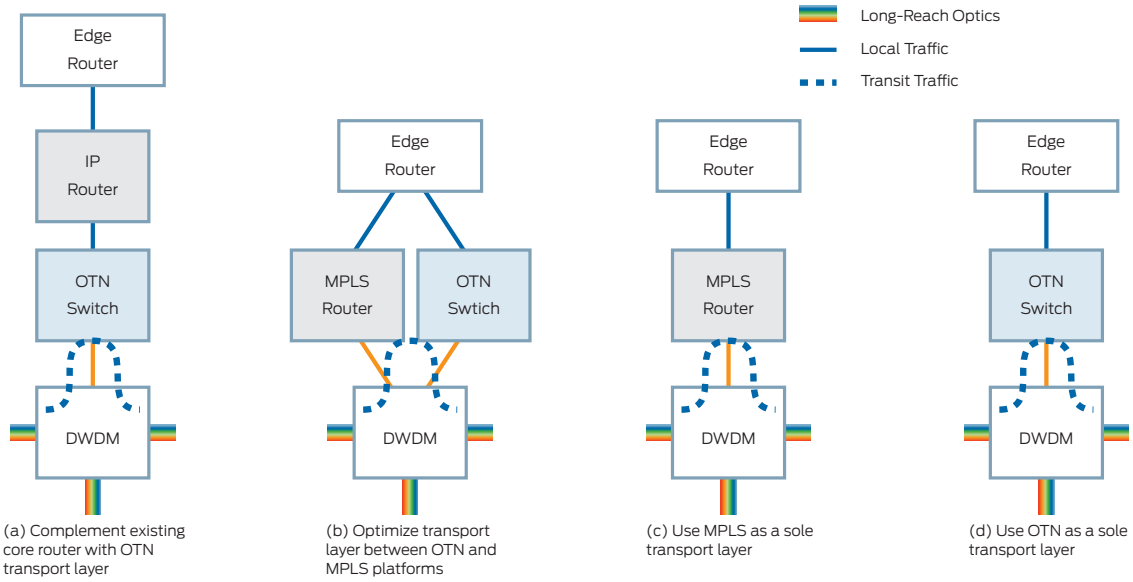


Figure 16: Reference architectures for MLO optimization study by Clemson University and Juniper Networks

We have implemented optimization models using the AMPL modeling language [20] and solved them with Gurobi Optimizer, which implements a parallel branch-and-bound algorithm. In cases where finding an optimal solution was beyond the boundaries of commercial-grade SMP computers, we have accepted feasible solutions that had an optimality gap within 1 percent.

Table 4: Comparison of Solution Costs (in Abstract Units) for Several Traffic Parameter Sets

Traffic Parameters		OTN / MPLS bypass			OTN + MPLS Optimized			MPLS Transport		OTN Transport	
α	K	Links	Links	Cost	Links	Links	Cost	Links	Cost	Links	Cost
1.5	1	1750	767	35141	86	1662	22466	1742	22646	3846	38460
	2	1820	787	36301	458	1430	23170	1810	23530		
	4	1898	816	37748	828	1206	23958	1904	24752		
	8	1986	854	39502	994	1112	24396	1974	25662		
	16	2098	906	41818	1038	1154	25382	2106	27378		
	32	2218	958	44214	1116	1140	25980	2206	28678		
3	1	1996	846	39418	106	1836	24928	1990	25870	6934	69340
	2	2236	927	43681	86	2148	28784	2226	28938		
	4	2556	1048	49664	240	2344	32872	2556	33228		
	8	2930	1201	56923	868	2184	37072	2934	38142		
	16	3370	1398	65854	1482	2078	41834	3370	43810		
	32	3836	1614	75482	2482	1628	45984	3830	49790		
4.5	1	2246	927	43781	94	2160	29020	2238	29094	9898	98980
	2	2672	1074	51422	134	2468	33424	2672	34736		
	4	3232	1282	61806	344	2936	41608	3232	42016		
	8	3890	1549	74527	1190	2880	49340	3894	50622		
	16	4646	1894	90022	2010	2918	58034	4646	60398		
	32	5482	2269	107007	3470	2396	65848	5464	71032		

The outcome of MLO computation (Table 4) offers us a new and valuable insight. In addition to pure photonic switching that we already discounted before, the "OTN bypass" (a) and "OTN only transport" (d) options exhibit prohibitively high capital cost and are not practically viable. This effect is evident across all traffic variations.

The two remaining topologies (MPLS and MPLS+OTN) appear to have very similar cost at the low values of K and remain within 10 percent at K=32 for all burstiness values. To understand what that means, we should recall that K has the physical sense of the number of overlapping peak values and is inversely proportionate to probability of traffic loss due to collision of bursts. If the network is sized merely based on priority traffic demands, values for K should remain conservatively high, while sizing for mixed-priority traffic yields optimal results at low to medium K values[2].

Notwithstanding the cost of operating two equipment classes (OTN and MPLS) in parallel, the capital gap between pure MPLS and the MPLS+OTN solution is negligible at low K values and never exceeds 10 percent up to K=32. This means there is some value in combining MPLS and OTN processing for transit traffic but it is far less than the initial difference in port costs (remember we assumed MPLS to carry a 30 percent premium). In addition, MPLS+OTN topology uses more fiber capacity than a pure MPLS solution (three to seven percent growth in link consumption at K=32).

The previous results are specific to the model design and the reference network chosen, but they highlight the relationships between transport technology choices, traffic demands, and the network expenses[3]. A full business case should incorporate the cost of bandwidth, wavelength planning, and operational expenses based on the state of technology and the needs to overlay or interoperate between devices with dissimilar control and data planes (if any).

In conclusion, we can say that a pragmatic, informed attitude toward technology selection invariably means selecting the best design among many alternatives. While doing so, it is useful to remember the following:

- Optimal transport architecture is principally a multidimensional problem.
- Bandwidth is not free—traffic dynamics are critical to network planning.
- Multilevel optimization (MLO) is a useful tool to understand relations between input variables and design outcome.
- The inverted pyramid of ISP revenues is not isomorphic to CapEx—the MLO solution might not favor the low-price ports.

While working with service providers worldwide on a daily basis, Juniper Networks fully understands the challenges of today's transport networks. We employ state-of-the art research, analysis, and one of the world's most respected R&D organizations to solve them. In the next sections we discuss how the theory and practice of transport come together in the Juniper product lineup.

Juniper Networks PTX Series Packet Transport Router: Introduction

In a move coherent with its history of innovation and industry trends, in March 2011 Juniper became the first company to introduce the first-ever packet transport router, Juniper Networks PTX5000. What's fundamentally new about this device is the sole focus on transport operations, which spans all the way from hardware; software feature sets to mechanical form factors and energy efficiency.

Not surprisingly, Juniper Networks PTX Series delivers the highest density and capacity per rack space, which has been a signature of Juniper Networks since the inception of the company. The PTX Series represents the company's strategy in the transport environment. With this new platform, Juniper makes a statement—modern transport revolves around packets, and the packet forwarding is the most logical foundation for transport intelligence.

Historically, the discussions on packet optical transport were a bit convoluted.

[2] We also tested the same model for higher values of K but found the results beyond K=32 to stabilize within 4 percent.

[3] Juniper Networks Professional Services consultants can provide customer-specific network modeling and planning.

There was almost no "optical" content to them, but quite a bit of non-packet (for example, TDM) matter. The truth is that photons are great signal carriers but are very hard to manipulate in logical and storage operations, making pure optical operations very limited in intelligence. Electrons are not so great at delivering information at long distance, but are very usable to memorize and process information. The PTX Series aims to realign network transport with its key role—processing packets electrically and transporting them over optical media.

In doing that, the PTX Series is forming a new product category. First off, the new device is squarely oriented toward MPLS processing and does not support other network protocols (such as IPv4 or IPv6) at scale in roles other than client services. Second, the PTX Series removes the need for a hierarchy of electrical and photonic equipment—a multi-service transport node based on the PTX Series at the bare minimum needs an optical shelf with a passive multiplexor. This stopgaps the "technology-bloating" phenomenon the carriers experienced for years, when vendors insisted on developing "intelligence and savings" within each and every layer of equipment. Single transport layer means no duplicating functions, no uncoordinated protection, and no unneeded bandwidth planning. It also means that the entire transport network is smaller, more resilient, and has exactly one management plane.

At the same time, the PTX Series fully supports the MPLS-TP server layer function and can run the label stack and control plane separate from any of the network clients (such as IP traffic or MPLS services). This gives the operators an option to keep the transport group logically and/or operationally separate from services, thus maintaining high predictability and minimal client-server coordination.

This new level of functionality can be best augmented by the fact that the PTX Series is not based on any existing Juniper chipset or mechanical chassis design.

It also does not employ merchant silicon in the critical packet path. Built from scratch with the purpose of being the new network's transport foundation, the PTX Series is best compared to all-electric sports cars in its no-compromise of speed, performance, and operational efficiency. In the following sections, we show exactly how the PTX Series aligns to the high-level transport vision that we have outlined so far.

Converged Packet Transport

In the discussion of transport architectures, we have mentioned that traditional packet platforms did not offer the means to support legacy circuit-switched services. This caused some operators to combine packet and circuit switching equipment in the overlay or ship-in-the-night fashion, both modes being suboptimal from a cost and bandwidth utilization perspective.

Juniper's answer to this challenge is very simple. Since a packet-switched platform operates over interfaces with OTN framing, customers must have a roadmap for circuit-to-packet migration with an OTN emulation service[4]. Such a service allows the core-facing links to have TDM "sub-slots" that can be mapped to MPLS LSPs. This way, the minimally required bandwidth to support OUT clients can be dynamically allocated, with the rest of the bandwidth of the same link automatically available for packet services. This unique ability makes the PTX5000 a drop-in replacement for any legacy transport system, with immediate recovery of spare bandwidth.

This ability to absorb and accommodate packet and circuit traffic over a high-speed core defines the new class of transport layer—the "Converged Packet Transport." This principal difference between converged transport and OTN devices with built-in packet aggregation blades is the lack of the packet-to-circuit adaptation. A packet switch processes any incoming IP, MPLS, or Ethernet traffic natively. Converged packet transport solves the problem of split personality for core networks with long-term circuit and packet service coexistence: regardless of traffic demand, any service can be deployed at any point across the entire network. When coupled with extensive use of built-in transponders, this offers the path to a streamlined, simplified equipment interconnect map with zero non-revenue interfaces (see Figure 17).

[4] Roadmap item.

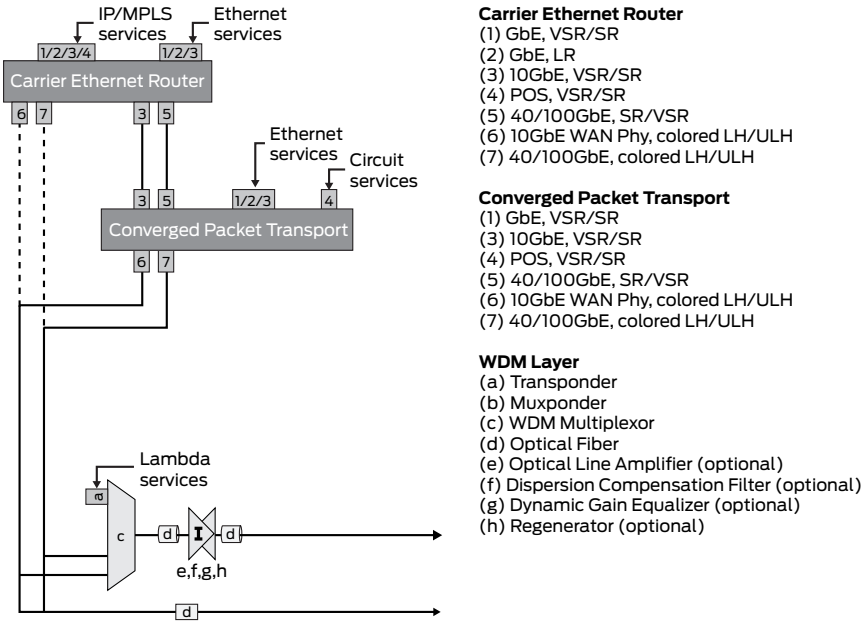


Figure 17: Simplified interface interconnect map with "Converged Packet Transport"

Cost-Bit Benchmark

Historically, packet core platforms claimed capital efficiency improvements based on the gains in density coherent with technology curve. Dennard's scaling law and the economy of scale ensure that the cost per bit is dropping with every product generation. Riding the scaling wave allows vendors to claim that every new product is more cost-efficient than a previous one, but it does not allow operators to compete in cost-bit environments, where IP transit prices are falling faster than the price per transistor of an integrated circuit.

The problem of price can also have a profound effect on network architecture. As we have shown in the previous chapter, the relative cost of MPLS, OTN, and WDM equipment affects the optimal transport topology as a function of an MLO solution. Knowing this fact, Juniper engineers aimed to create the next-generation label-switching router (LSR) with a packet forwarding path streamlined to a degree when it undercuts the traditional IP routers in price and overshadows any circuit-based platform in utility. Such precision in silicon planning is possible due to Juniper's extensive in-house ASIC and NPU engineering expertise—something simply not possible with off-the-shelf components available today.

The PTX Series is driven by Juniper Networks purpose-built Junos® Express chipset, providing a full MPLS feature set but limited IP scale. The resulting reduction in lookup budgets and memory structures allows the Junos Express chipset to achieve the remarkable 120 Gbps of full-duplex forwarding capacity in just two chips, with up to four PFE complexes per linecard (Figure 18).

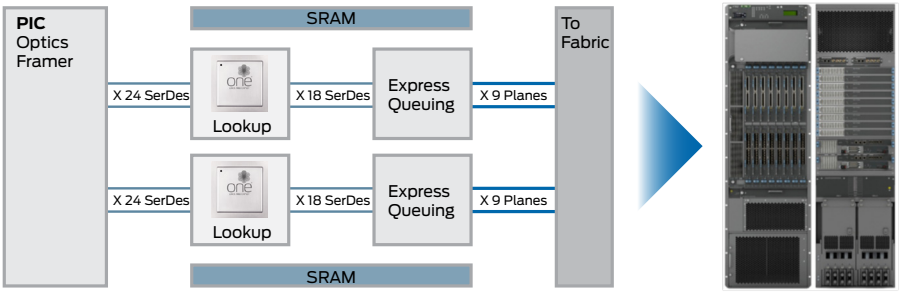


Figure 18: PTX Series high-level packet forwarding architecture

With packaging as dense as 480 Gbps of revenue-generating bandwidth per slot, Juniper Networks PTX Series Packet Transport Router clearly provides the new cost-bit benchmark in transport applications. This is the answer to the mismatch between the growth of best-effort traffic due to the ever-increasing content consumption and downward trends of the IP wholesale markets. With the PTX Series, service providers and operators can profit not only from traditionally high-margin services such as leased lines and CBR video delivery, but also from the "white noise," "fractal," and "tidal chart" dynamics of Web page transfers, VBR video streams, and file downloads.

The PTX Series is also designed with CO planning in mind. To gain maximum capacity without moving into multichassis domain, the platform is available in singlewide and doublewide rack formats, the latter occupying exactly two bays to keep energy density, cooling and floor requirements within conventional limits. The platform also offers more than twice the energy efficiency of nearest competitors, scaling the network while reducing its environmental footprint.

Deployment Architectures

One important conclusion from discussion of multi-level transport architectures is that all of them try to solve the same three problems: path computation, path resiliency and network management. We have shown that independent solution of these problems across several levels (for example, packet-TDM-optical planes) is not only unnecessary but also fails from CapEx perspective. This makes a strong case for collapsing the layers of transport. The required and sufficient transport functionality should be fully covered within packet-aware label switch router, such as Juniper Networks PTX Series and statically configured optical network in photonic layer.

In greenfield deployments, PTX Series simplifies the deployment architecture to packet transport routers serving client-layer Ethernet services routers. The latter can request Ethernet, IP, or MPLS transport from the network core as needed. However, LSRs do not provide IP VPN, and VPLS functions for CE devices (Figure 19).

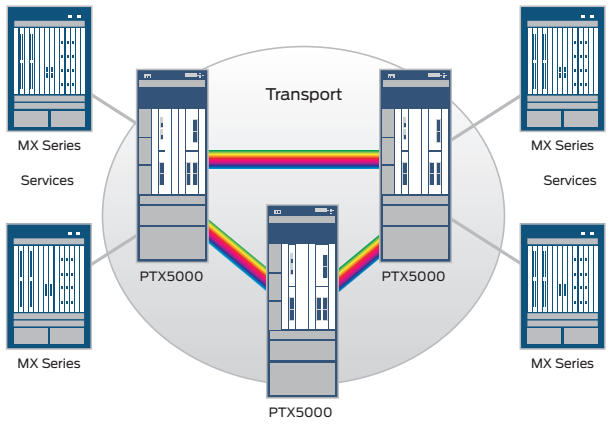


Figure 19: Baseline packet transport architecture

In an architecture where the existing network core is served by IP routers, the PTX Series can form the supercore, scaling the network at the new efficiency point. Existing "outer core" routers can be refocused to providing feature-rich IPv4 and IPv6 services such as IP VPN, large-scale filtering and DDoS mitigation, carrier-grade NAT, and so on (Figure 20).

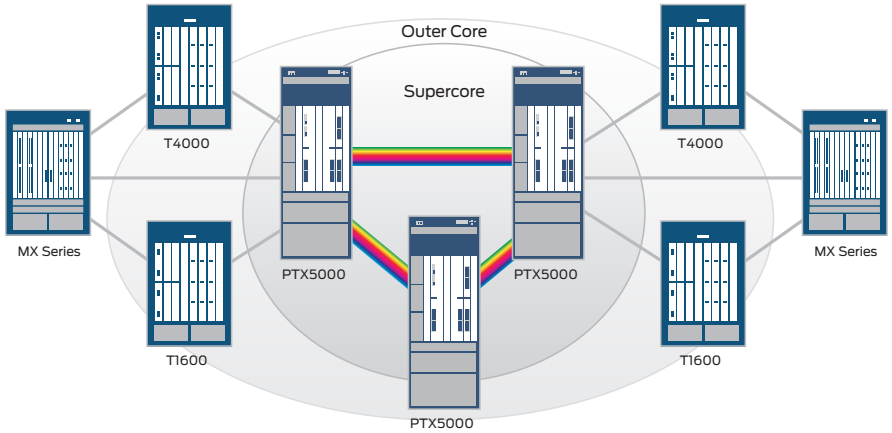


Figure 20: Packet supercore architecture

Finally, when the existing core network is based on the TDM, the PTX Series can provide the drop-in replacement for legacy circuit transport while retaining circuit service interfaces5. This application effectively recovers the core bandwidth otherwise lost due to grooming inefficiency and the lack of statistical multiplexing in TDM platforms (Figure 21).

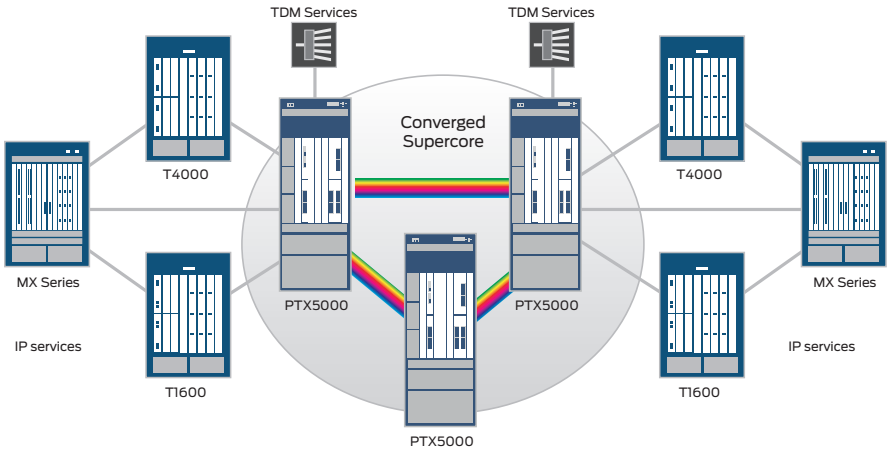


Figure 21: Converged supercore architecture

In all cases, the PTX Series can operate in the same or independent MPLS stack as connected devices and provide statically and dynamically provisioned services with a high degree of independence—the signature property of the modern, packet-oriented transport design.

Conclusion

The complex, packet-centric core networks of today clearly require an effective and economically viable transport layer. Realistically speaking, such a layer should incorporate the benefits of dynamic control plane combined with the reliability and management characteristics previously delivered by SONET/SDH technology.

While the past of the network evolution is not necessarily a good predictor for the future, the development of network technology clearly favors the most economically reasonable and simple way to do the job. More complex, baroque, and elaborate technologies tend to remain in the “limited deployment” mode, often stranding adopters behind the mainstream trends of technology.

This is why the combination of equipment capital cost, bandwidth utilization, and operational simplicity requirements is strongly driving the network design away from circuit and into the all-packet domain. Recognizing this trend, Juniper leads the industry with PTX5000 as a cut-through path to a new, packet-oriented provider network. This vision enables Juniper customers worldwide to take advantage of IP and Ethernet economy and fuels the transformation of mobile and fixed networks alike.

Standing in the “packet storm” is confusing only until one reaches the point of realization that surging packet traffic represents an opportunity, not a threat. Just like the real ocean waves can be harnessed to produce electricity, millions of packet flows aggregated in the transport layer offer an opportunity to utilize every bit of available bandwidth while maintaining full control over the cost and quality of service.

This is the point where the “packet revolution” is finally reaching the core.

[5] Roadmap.

Bibliography

[1] "Synchronous Optical Networking" **wikipedia.org**

[2] Wes Simpson. "Wasting Bandwidth" **11 Mar. 2008 TV Technology**

[3] James Manchester, Jon Anderson, Bharat Doshi, and Subra Dravida. *"IP over SONET"* IEEE Communications Magazine, May 1998

[4] IST IP NOBEL Phase 2. ""Migration Guidelines with Economic Assessment" Deliverable D2.4

[5] IST IP NOBEL Phase 2. "Report on Resilience Mechanism for NOBEL Solutions" Deliverable D2.3

[6] Telegeography CommsUpdate. "Global network access drives local IP transit prices", **Oct 1 2009**

[7] "Continuous Systems, Nonstop Operations with Junos OS Software" white paper, © 2009 Juniper Networks, Inc.

[8] "Network Operating System Evolution" white paper, © 2010 Juniper Networks, Inc.

[9] Martin Carroll, Josef Roese, Takuya Ohara. *"The Operator's View of OTN Evolution,"* IEEE Communications Magazine, September 2010

[10] Iyad Katib, Deep Medhi. "A Network Optimization Model for Multi-Layer IP/MPLS over OTN/DWDM Networks," IP Operations and Management, 9th IEEE International Workshop

[11] Ralf Huelsermann, Matthias Gunkel, Clara Meusburger, Dominic A. Schupke. "Cost modeling and evaluation of capital expenditures in optical multilayer networks," Journal of Optical Networking, Vol. 7, Issue 9, September 2008

[12] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, F. Jahanian, M. Karir. ATLAS Internet Observatory 2009 Annual Report

[13] Thomas Engel, Achim Autenrieth, Jean-Claude Bischoff. "Packet Layer Topologies of Cost Optimized Transport Networks." Proceedings of 13th International Conference on Optical Networking Design and Modeling, 2009

[14] U. Menne, R. Wessaly, C. Raack and D. Kharitonov "Optimal degree of optical circuit switching in IP-over-WDM networks." Proceedings of 16th Conference on Optical Networks Design and Modeling, 2012

[15] Sample BANDWIDTH PRICING REPORT **OCTOBER, 2009**, Telegeography.com

[16] A. Adas, A. Mukherjee. "On Resource Management and QoS Guarantee for Long Range Dependent Traffic." IEEE Infocom, pages 779-787, April 1995

[17] Matthew T. Lucas, Dallas E. Wrege, Bert J. Dempsey, Alfred C. Weaver. "*Statistical Characterization of Wide-Area IP Traffic*." IC3N '97 Proceedings of the 6th International Conference on Computer Communications and Networks

[18] Pietro Belotti, Antonio Capone, Giuliana Carello, Federico Malucelli. "*Multi-layer MPLS network design: The impact of statistical multiplexing.*" Computer Networks: The International Journal of Computer and Telecommunications Networking, Volume 52, Issue 6, pages 1291-1307

[19] P. Belotti, K. Kompella, L. Noronha. "Robust Optimization Models for Networks with Statistical Multiplexing", September 2010

[20] R. Fourer, D.M. Gay, B. Kernighan. "AMPL: A mathematical programming language." In algorithms and model formulations in mathematical programming, pages 150-151. Springer-Verlag, New York, NY, USA 1989

About Juniper Networks

Juniper Networks is in the business of network innovation. From devices to data centers, from consumers to cloud providers, Juniper Networks delivers the software, silicon and systems that transform the experience and economics of networking. The company serves customers and partners worldwide. Additional information can be found at **www.juniper.net**.

Corporate and Sales Headquarters
Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or 408.745.2000
Fax: 408.745.2100
www.juniper.net

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: 31.0.207.125.700
Fax: 31.0.207.125.701

To purchase Juniper Networks solutions, please contact your Juniper Networks representative at 1-866-298-6428 or authorized reseller.

2000402-004-EN Mar 2013 Printed on recycled paper