Jean-François Labourdette
& Zhensheng Zhang
Editors

Layered Architecture for Scalability in Core Mesh Optical Networks

Jean-François Labourdette
Tellium

In a previous column[1], I talked about the interconnect penalty incurred when building a switching node complex with small size switches. In this column, we examine the scalability of core optical mesh networks, and specifically how the interconnect penalty, which results from traffic at a node exceeding the maximum switch size, impacts the core network architecture.

1 Introduction

As traffic demand grows and evolves, core network nodes need to switch ever-larger amounts of traffic at different rates (DS1, DS3, OC-3, OC-12, OC-48, OC-192). Furthermore, the traffic mix shifts towards higher rates over time. This evolution, if it outpaces the evolution of switch size, requires that multiple switches be deployed at certain core network nodes. As described previously, interconnecting multiple switches within an office wastes ports, and this inefficiency has implications regarding the overall network architecture. For example, our analysis showed that if traffic forecast is only 70% accurate when deploying and connecting transmission and switching equipment in an office, as much as 30% of the switch ports may end up being used for connecting the switches together. While many other aspects impact the core mesh network architecture, we focus here on the effect of switch size.

The core optical network consists of backbone nodes interconnected by point-to-point WDM fiber links in a mesh interconnection pattern. Each WDM fiber link carries multiple wavelength channels (e.g., 160 OC-192 channels). Transmission rates of wavelength channels on long-haul WDM systems are currently evolving from OC-48 to OC-192 and are expected to evolve to OC-768 in the future. Multiple conduits (each containing multiple fibers) are usually incident at the backbone nodes from adjacent nodes. Figure 1 illustrates a core optical network. Diverse edge equipment, such as IP routers, FR/ATM switches, and Multi Service Provisioning Platforms (MSPP), is connected to an optical switch.

An OEO-based core optical switch converts optical signals into the electrical domain at the ingress port, switches the electrical signals through an electrical switch matrix, and then converts signals into the optical domain at the egress port. The switch fabric of the OEO switch is typically strictly non-blocking allowing complete interconnection flexibility between the ports of the switch. The granularity of the switch fabric drives the grooming granularity, i.e., the lowest rate at which the equipment can switch, multiplex, and demultiplex signals. The grooming granularity of the optical switch determines the granularity at which network bandwidth is managed. The optimal grooming granularity of a network depends on the mix of traffic rates and on the traffic volume supported by the network.

Traffic carried in the core optical network consists of data traffic that is packet-groomed by IP routers and FR/ATM-switches into OC-N (N = 12, 48, 192) trunks. TDM switches aggregate traffic at lower rates such as STS-1 (52 Mb/s) and VT1.5 (1.7 Mb/s) and feed into the core at OC-N rates as well. Figure 2 illustrates a traffic mix for a carrier backbone network. In this instance OC-48 (2.5 Gb/s) and OC-192 (10 Gb/s) services and trunks are a dominant (57%) and growing component of the core traffic mix.

Two core network architectures of interest are a flat STS-1 network architecture and a two-tier or layered STS-1 and STS-48 network architecture. In the flat network architecture, each network node contains an optical switch that can switch at STS-1 and higher rates in the SONET hierarchy[2]. As traffic grows, the switching capacity of a node may be exceeded but can be scaled by interconnecting multiple such switches. However, scaling the network in this manner incurs a possibly severe penalty in terms of interconnect capacity between the switches. We have previously quantified the interconnect capacity required in such a switching system to maintain a given level of blocking, and we make use of these results later on. In the layered network architecture, the network is scaled by organizing it in layers, with each layer optimized to switch and groom at a different rate, typically an STS-1 switching layer, and an STS-48, or wavelength, switching layer. Our key observation is that there is a crossover

[1] *J-F. Labourdette, "The Interconnect Penalty of Small Size Switches," Optical Network Magazine Tutorial Corner, Sept/Oct 2002.*

[2] *Similarly in the corresponding SDH hierarchy.*

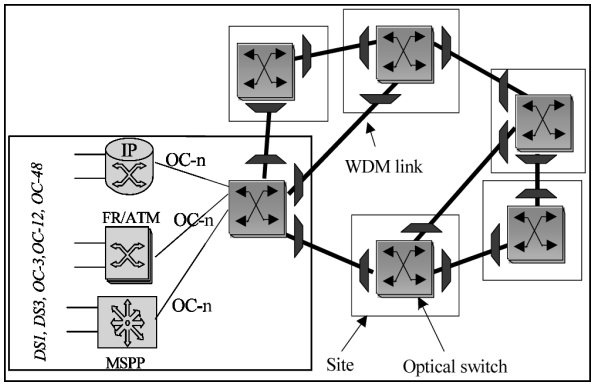

Figure 1: Core Mesh Optical Network.

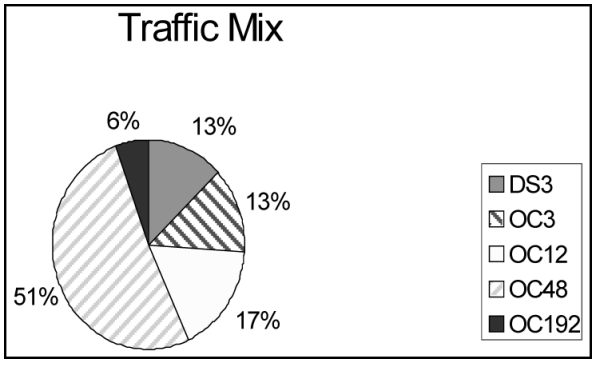

Figure 2: Example of traffic mix in the core optical network, (as % of total bandwidth) shifting toward OC-48 and above.

point beyond which a layered architecture becomes more cost-effective, as the total traffic grows and as the traffic mix evolves toward higher rates. Beyond this crossover point, a scalable network is more efficiently and cost effectively realized by a layered network architecture, with each layer optimized to groom at a different rate.

Indeed, large-scale networks have always been organized in multiple layers. It may be a service provider's dream to have a single switch that is scalable, manageable, low-cost, and that can switch all rates and protocol formats. But practical considerations such as hardware and software scalability, manageability, and reliability have always led to layered architectures, with each layer optimized independently. In the layered architecture, scalability and manageability are achieved by multiplexing traffic flows into larger streams as they traverse from the edge to the core, and demultiplexing them as they traverse from the core to the edge. Effectively, traffic flows are groomed and switched at a coarser granularity at the network core and at a finer granularity at the edge.

2 Architecture: Flat vs. Layered

Let us consider in more detail the two architectures for the core backbone network: (1) the flat architecture, and (2) the layered, or two-tier, architecture.

In the flat architecture (as shown in Figure 3), the core optical switch operates at STS-1 granularity. The STS-1 switch terminates all OC-N services from client equipment (such as IP routers, ATM/FR switches, Multi-Service Provisioning Platforms (MSPPs)) on optical ports[3]. The STS-1 switch also terminates wavelengths (OC-48/OC-192) from the WDM systems connecting offices together. The flat architecture allows network bandwidth to be managed in STS-1 increments. OC-N streams are switched by first demultiplexing into component STS-1 streams at the input port and multiplexing the STS-1 streams at the output port. As the traffic grows beyond the capacity of the STS-1 switch, multiple STS-1 switches are interconnected to yield a larger STS-1 switching complex. This wastes ports for interconnecting STS-1 switches together, the key factor in our analysis.

The flat network architecture faces challenges to achieve fast and capacity efficient restoration of OC-N circuits. Shared mesh restoration of all individual OC-N circuits is difficult due to the complexity of handling so many circuits. For example, if a fiber carrying 160 OC-192 channels breaks, potentially, $160 \times 48 = 7680$ STS-1 circuits could be affected by the failure. This makes it very difficult to achieve restoration time of the order of few hundreds of ms. Fast restoration in the flat architecture is thus likely to require dedicated mesh $(1 + 1)$ protection, thereby incurring the capacity penalty of dedicated mesh $(1 + 1)$ protection compared to shared mesh restoration.

In the layered architecture (as shown in Figure 4), the core optical switch operates at STS-48 granularity. Connected to the core optical switch are switches that groom at STS-1 granularity. The STS-48 switch directly terminates OC-48 and OC-192 services. To groom traffic at a lower rate, edge STS-1 switches terminate OC-N ($N < 48$) services. We term this a layered, or two-tier, architecture because there are STS-48 switches performing "core-grooming" at STS-48 rates, and STS-1 switches performing "edge-grooming" at STS-1 rates. In this layered architecture, wavelengths are managed in increments of OC-48.

The STS-48 switch also terminates wavelength from WDM systems. Connected to the STS-48 switch are one or more STS-1 switches. The STS-1 switch terminates services below STS-48 rates (e.g., DS3, OC-3, OC-12), and aggregates them into OC-48 pipes. OC-48 or OC-192 services between backbone nodes are set up as lightpaths by finding a route in the core network and configuring the STS-48 switches along the route. Services below OC-48 rates (called subrate services) between backbone nodes are set up as follows: A set of OC-48 or OC-192 lightpaths between the core switches serve as trunks (or an overlay topology) for purposes of routing

[3]Some OC-48 or OC-192 trunks may be directly connected to the WDM systems since sub-rate grooming is not required. This would be done at the expense of provisioning flexibility.

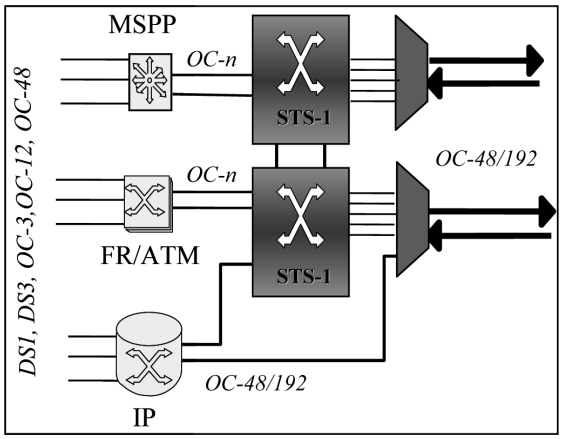

Figure 3: Node architecture in a flat core network.

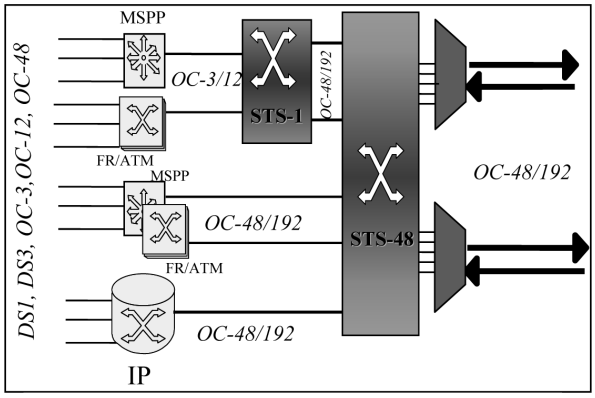

Figure 4: Node architecture in a two-tier core network.

subrate services. For example, the overlay topology may be identical to the physical topology, with a direct lightpath between all neighbors. If there is a direct lightpath between a pair of backbone nodes, and there is enough capacity on that lightpath, a subrate service between the node-pair can use that lightpath. If there is no direct lightpath with enough capacity, then the subrate service has to traverse multiple lightpath hops, and at each intermediate node hairpins into the STS-1 switch and gets switched onto the lightpath on the next hop. Hairpinning is a natural inefficiency of routing on an overlay topology. Because of the possibly less-efficient packing of sub-OC-48 connections into OC-48 trunks, the resulting OC-48 wavelength utilization would in general be lower than in the flat network architecture. However, the amount of hairpinned traffic between a node-pair is bounded and less than a fraction of an OC-48, because, as soon as hairpinned traffic between a pair of nodes exceeds a threshold, a more economical solution is to set up a direct lightpath between that pair of nodes.

The layered network has a fast, efficient, and scalable restoration architecture. OC-48 and OC-192 lightpaths between STS-48 switches are restored using shared mesh restoration, allowing maximum sharing efficiency. Sub-rate circuits that ride on lightpaths are automatically restored thanks to the restoration of the lightpath. Connections between the STS-1 switch and STS-48 switch may be protected by 1:N protection. In this restoration architecture, sub-rate circuits are effectively restored as if using shared-link restoration on the overlay topology.

3 Analysis & Results

Using results on interconnect penalty, the analysis of the flat and the layered network architectures shows that the layered network architecture eventually becomes more cost-efficient as the traffic scales and as the traffic mix evolves towards higher rates.

We provide now some representative qualitative graphs that capture this fundamental behavior. For the sake of simplicity we assume that the traffic mix consists of two service rates: OC-48 and OC-3. We also assume that all wavelengths are at OC-48 rates. The salient assumption is that an STS-48 switch can be built to larger size (in number of OC-48 equivalent) than an STS-1 switch[4]. We also make the following cost assumption: (a) the cost of an OC-48 port is the same on the STS-1 and the STS-48 switches, and (b) the Gb/s cost of an OC-48 port is less than the Gb/s cost of an OC-3 port on the STS-1 switch.

With the flat network architecture, we consider two cases, first where the flat network supports dedicated mesh (1+1) protection, and secondly where the flat network supports shared mesh restoration, despite the expected lower restoration performance. With the two-tier layered architecture, the STS-48 switches support shared mesh restoration.

Figure 5 illustrates the cost comparison between the two architectures when the traffic mix shifts from 0% OC-48 to 100% OC-48 while keeping the total amount of traffic constant. The two single-tier curves represent the flat STS-1 network: the higher curve uses dedicated mesh (1 + 1) protection, the lower curve uses shared mesh restoration (with much slower expected restoration performance). The two-tier curve represents the layered network capable of shared mesh restoration in the STS-48 layer for OC-48 connections and OC-48 trunks between the STS-1 switches. For the flat architecture, the bandwidth overhead of dedicated mesh (1 + 1) protection over shared mesh restoration results in more network bandwidth and more network-side interface capacity. The cost curve of the flat network when shared mesh restoration is used is below the cost curve of the flat network when dedicated mesh protection is used, as expected. The cost curves also decrease as the traffic mix evolves towards OC-48 since the cost of an OC-48 interface is less per Gb/s than the cost of an OC-3 interface. The layered

[4] *In currently deployed products, this ratio is 2.*

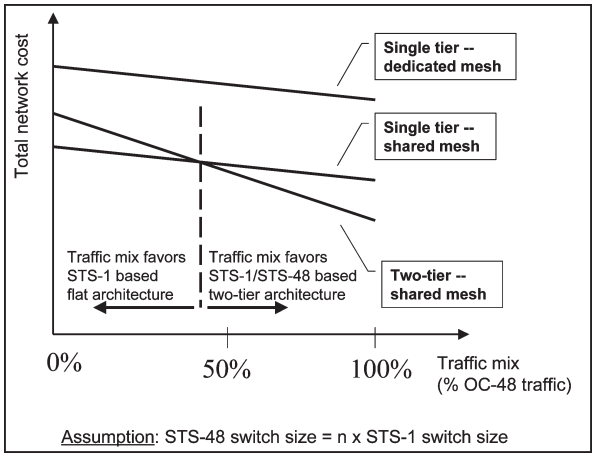

Figure 5: Cost comparison when (a) flat network is capable of dedicated (1 + 1) mesh protection, (b) flat network is capable of shared mesh restoration, and (c) the layered network is capable of shared mesh restoration.

network is cheaper than the flat network using dedicated mesh protection for the entire traffic mix. When the proportion of OC-48 in the traffic mix exceeds a certain value, and if the total traffic through a node exceeds the capacity of the STS-1 switch (so that multiple STS-1 switches have to be interconnected) then the layered architecture becomes cheaper than the flat architecture with shared mesh restoration. The cost of the layered network declines more sharply than the cost of the flat network as the traffic mix shifts towards OC-48. This occurs because less STS-1 switching capacity, and therefore less interconnect penalty, is required, and because fewer interconnects between the STS-1 and STS-48 switches are required. In addition, and similarly to the flat network architecture, the cost of the OC-48 is less per Gb/s than the cost of the OC-3 interface, further decreasing the network cost as the traffic shifts to higher rates.

4 Conclusion

Large-scale networks have historically been organized into multiple layers for scalability and manageability. In a previous column, we analyzed the penalty that results from interconnecting several switches within an office when the total traffic in that office exceeds the capacity of a single switch. In this column, we use these results and show that a layered (two-tier) STS-1/STS-48 network becomes cheaper than a flat STS-1 network when two conditions are met. First, the total traffic per node scales beyond the capacity of the STS-1 switch, and secondly the proportion of OC-48 and above connections in the traffic mix exceeds a certain value. While we carried out this qualitative comparison based mostly on switch sizes, there are many other aspects that come into play and would impact the cost of single-tier and two-tier core optical network architectures and change the comparison [1]. On the other hand, the argumentation presented here reflects a fundamental behavior and applies to networks in general.

5 Reference

[1] C. Olszewski *et al.,* "Two-Tier Network Economics," NFOEC, Dallas, TX, Sept. 2002.