# 8 PHAROS: An architecture for next-generation core optical networks

Ilia Baldine[†], Alden W. Jackson[‡], John Jacob[⊗], Will E. Leland[‡], John H. Lowry[‡], Walker C. Milliken[‡], Partha P. Pal[‡], Subramanian Ramanathan[‡], Kristin Rauschenbach[‡], Cesar A. Santivanez[‡], and Daniel M. Wood[⊕]

[†]Renaissance Computing Institute, [‡]BBN Technologies, [⊗]BAE Systems, [⊕]Verizon Network Systems

## 8.1 Introduction

The last decade has seen some dramatic changes in the demands placed on core networks. Data has permanently replaced voice as the dominant traffic unit. The growth of applications like file sharing and storage area networking took many by surprise. Video distribution, a relatively old application, is now being delivered via packet technology, changing traffic profiles even for traditional services.

The shift in dominance from voice to data traffic has many consequences. In the data world, applications, hardware, and software change rapidly. We are seeing an unprecedented unpredictability and variability in traffic patterns. This means network operators must maintain an infrastructure that quickly adapts to changing subscriber demands, and contain infrastructure costs by efficiently applying network resources to meet those demands.

Current core network transport equipment supports high-capacity global-scale core networks by relying on higher speed interfaces such as 40 and 100 Gb/s. This is necessary but in and of itself not sufficient. Today, it takes considerable time and human involvement to provision a core network to accommodate new service demands or exploit new resources. Agile, autonomous, resource management is imperative for the next-generation network.

Today's core network architectures are based on static point-to-point transport infrastructure. Higher-layer services are isolated within their place in the traditional Open Systems Interconnection (OSI) network stack. While the stack has clear benefits in collecting conceptually similar functions into layers and invoking a service model between them, stovepiped management has resulted in multiple parallel networks within a single network operator's infrastructure.

*Next-Generation Internet Architectures and Protocols*, ed. Byrav Ramamurthy, George Rouskas, and Krishna M. Sivalingam. Published by Cambridge University Press. 

Such an architecture is expensive to build and operate, and is not well-suited to reacting quickly to variable traffic and service types. This has caused the network operators to call for "network convergence" to save operational and capital costs.

In the area of traffic engineering and provisioning, IP services now dominate core-network traffic, but IP networks utilize stateless per-node forwarding – costly at high data rates, prone to jitter and packet loss, and ill-suited to global optimization. Layer 2 switching mechanisms are better by some measures but lack fast signaling. Generalized multiprotocol label switching (GMPLS) attempts layer 2 and 3 coordination but is not yet mature enough for optical layer 1 and shared protection over wide areas. Today's SONET 1+1 method of provisioning protected routes for critical services consumes excessive resources, driving down utilization, increasing cost, and limiting the use of route protection.

Thus, network operators are looking to integrate multiple L1–L2 functions to reduce cost and minimize space and power requirements. They also aim to minimize the costly equipment (router ports, transponders, etc.) in the network by maximizing bypass at the lowest layer possible. To allow maximum flexibility for unpredictable services, they require a control plane that supports dynamic resource provisioning across the layers to support scalable service rates and multiple services, e.g., Time Division Multiplexing (TDM), Storage Area Networking (SAN), and IP services. Such a control plane also enables automated service activation and dynamic bandwidth adjustments, reducing both operational and capital costs.

Surmounting these challenges requires a re-think of core network architectures to overcome the limitations of existing approaches, and leverage emerging technologies. In response to these challenges, the US Defense Advanced Research Projects Agency (DARPA) created the Dynamic Multi-Terabit Core Optical Networks: Architecture, Protocols, Control and Management (CORONET) program with the objective of revolutionizing the operation, performance, survivability, and security of the United States' global IP-based inter-networking infrastructure through improved architecture, protocols, and control and management software. CORONET envisions an IP (with Multi-Protocol Label Switching (MPLS)) over Wavelength Division Multiplexing (WDM) architecture on global scale. The target network includes 100 nodes, has aggregate network demands of between 20 and 100 Tb/s using up to 100 40 or 100 Gb/s wavelengths per fiber (higher demand uses higher capacity waves), and supports a mix of full wavelength and IP services.

The network is highly dynamic with very fast service set up and tear down. A key CORONET metric in this regard is very fast service setup (VFSS) in less than 50 ms + roundtrip time. There are also fast services (FSS) with 2 second setup requirements, scheduled services and semi-permanent services. The IP traffic includes both best effort and guaranteed IP services with a variety of granularities as low as 10 Mb/s per flow. The network must be resilient to multiple concurrent network failures, with double- and triple-protected traffic classes in addition to singly protected and unprotected services. Restoration of services is enacted within 50 ms + round trip time. To ensure efficiency in

handling protected traffic, CORONET specifies a metric, $B/W$, where $B$ is the amount of network capacity reserved for protected services and measured in wavelength-km, and $W$ is the total working network capacity, also in wavelength-km. B/W must be less than 0.75 for the CONUS-based traffic in the CORONET target network.

In this chapter, we present PHAROS (Petabit/s Highly-Agile Robust Optical System) – an architectural framework for next-generation core networks that meets the aggressive CORONET objectives and metrics. Through its framework, optimization algorithms, and control plane protocols, the PHAROS architecture:

- substantially improves upon today's 30-day provisioning cycle with its automated systems to provide less than 50 ms + round trip time in the fastest case;
- replaces opaque, stovepiped layer 1, 2, and 3 management systems with accessible administration;
- qualitatively improves the tradeoff between fast service setup and network efficiency; and
- assures network survivability with minimal pools of reserved (and therefore normally unused) capacity.

The CORONET program is the first program to explore control and management solutions that support services across global core network dimension with 50-ms-class setup time, and also to respond to multiple network failures in this time frame. The PHAROS architecture is designed in response to that challenge. The PHAROS architecture has been designed with awareness of current commercial core network practice and the practical constraints on core network evolution. The design of PHAROS also includes support for asymmetric demands, multicast communications, and cross-domain services, where a domain is a network or set of networks under common administrative control.

PHAROS aims to build upon recent research that highlights intelligent grooming to maximize optical bypass to reduce core network costs [1, 2, 3, 4, 5, 6, 7, 8]. The program also exploits the use of optical reconfiguration to provide bandwidth-efficient network equipment that responds gracefully to traffic changes and unexpected network outages [9, 10, 11].

While a large body of work exists on several exciting research problems in next-generation networks, our focus in this chapter is on the *system architecture* – how we can leverage individual solutions and clever breakthroughs in transport and switching from a signaling, control and management perspective in order to hasten deployment. We therefore see PHAROS as a bridge between the state of art in research and the next-generation deployed system.

Architecting any system requires selecting choices within a tradeoff space. In this chapter, we not only describe the choices we made, but in many cases, we also discuss the alternatives, their pros and cons and the reasons for our choice. We hope this gives the reader a flavor of the typical strategies in this space, and an appreciation of how requirements drive the choice.

The remainder of this chapter is organized as follows. After surveying background work, we begin with an overview of the PHAROS architecture. Following that we describe three key components of PHAROS, namely, the cross-layer resource allocation algorithm (Section 8.4), the signaling system (Section 8.5), and the core node implementation (Section 8.6). Finally, we give some preliminary results on performance estimation.

## 8.2 Background

We briefly survey prior work on some of the topics discussed in this chapter, namely, path computation, protection, and node architectures. Unlike IP networks, path computation in optical networks involves computation of working and protection bi-paths. Approaches can be classified by the nature of the required paths (e.g., node-disjoint, link-disjoint, k-shortest), the order for computing them (e.g., primary-then-protection vs. joint-selection), and the cost associated with each path. Some works include [12, 13]. Our approach is a hybrid one and uses the concept of joint or shared protection.

The various levels of protection defined for different traffic demands in a core optical network, along with the low-backup-capacity targets, motivate the use of shared-protection schemes for this application. Such techniques fall into broad categories of the various computational and graph-theoretic approaches: constrained shortest paths [14], cycle covers [15], and ILP formulations like the p-cycles [16]. As these techniques can guarantee only single-protection for all the flows, they would have to be augmented to guarantee double or triple protection for the set of flows that require it. In this chapter, we have outlined the preliminary formulation of a shared-mesh-protection algorithm based on virtual links and jointly protected sets that deliver double- and triple-protection services.

The sophistication of optical-network-node architectures has risen as the state of the art for the optical components within these nodes has advanced. Recent advances in optical-switch reliability and functionality, along with the size of the available switch fabrics, have motivated node architectures that allow such multiple functionalities as reconfigurable add/drop, regeneration, and wavelength conversion [17]. The cost, power, size, and reliability calculations for these different implementations are highly technology-dependent and are changing rapidly as new technologies are transitioned into the commercial market. As a result of this rapidly changing trade-space, we have chosen to remain agnostic to the exact switch architecture in our nodes, a feature we discuss further in the next section.

## 8.3 PHAROS architecture: an overview

In designing the PHAROS system, we were guided by some high-level principles and tenets, such as technology-agnosticism, fault-tolerance, global optimizations,

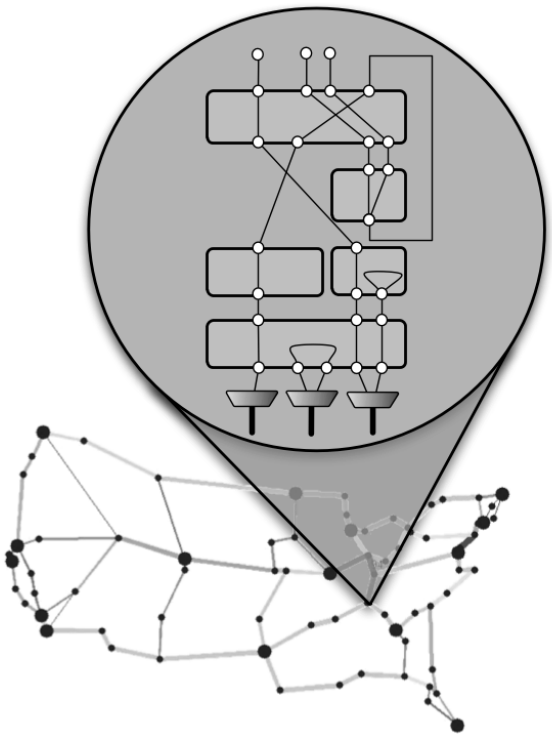


**Figure 8.1** Multi-level topology abstractions make PHAROS technology agnostic.

etc. These motivated some innovations such as topology abstractions, triangulation, etc. In this section, we first discuss these principles guiding our architecture, along with features associated with them. We then give a brief overview of the logical functional blocks and their roles.

A basic tenet of the PHAROS architecture is a *technology-agnostic* design that maximizes bypass to achieve lower cost-per-bit core network services and accommodates future generations of switch technology for long-term graceful capacity scaling. Current systems employ some degree of abstraction in managing network resources, using interface adapters that expose a suite of high-level parameters describing the functionality of a node. Such adapters, however, run the twin risks of obscuring key blocking and contention constraints for a specific node implementation, and/or tying their interfaces (and the systems resource management algorithms) too tightly to a given technology.

The PHAROS system avoids both of these problems by using *topology abstractions* – abstract topological representations for all levels of the network. The representations extend down to an abstract network model of the essential contention structure of a node, as illustrated in Figure 8.1, and extend upward to address successive (virtual) levels of functionality across the entire network.

With a uniform approach, common to all levels of resource representation and allocation, PHAROS accurately exploits the capabilities of all network elements, while remaining independent of the switching technology. At the signaling and control level, the PHAROS architecture also provides a set of common mechanisms for its own internal management functions (such as verification and failover); these mechanisms provide significant architectural immunity to changes in the technologies used in implementing specific PHAROS functional components.

The PHAROS architecture uses multilevel topological abstractions to achieve *global multi-dimensional optimization*, that is, efficient integrated resource optimization over the fundamental dimensions of network management: network extent, technology levels, route protection, and timescales. Abstraction allows a given request to be optimized across the network, simultaneously trading off costs of resources within individual network levels as well as the costs of transit between levels (such as the optical–electrical boundary). Resources of all levels can be considered, including wavelengths, timeslots, grooming ports, and IP capacity.

PHAROS optimization unites analysis of the resources needed to deliver the service with any resources required for protection against network element failures. Protection resources (at all levels) are allocated in conjunction with the resources required by other demands and their protection, achieving dramatic reductions in the total resources required for protection (the CORONET B/W metric). Our optimization design allows PHAROS to unify the handling of demand timescales, exploiting current, historical, and predicted future resource availability and consumption. Timescales are also addressed by the overall PHAROS resource management strategy, which selects mechanisms to support available time constraints: for example, PHAROS employs pre-calculation and tailored signaling strategies for very fast service setup; selects topology abstractions to perform more-extensive on-demand optimization where feasible; and evaluates long-term performance out of the critical path to enable rebalancing and improve the efficiency of the on-demand optimizations.

Finally, the PHAROS architecture achieves a high degree of *fault-tolerance* by using a design construct that combines redundancy and cross-checking in a flexible way to mitigate single point of failure and corrupt behavior in a Cross-layer Resource Allocator (CRA), a critical component of the PHAROS architecture described in Section 8.4. This design construct, which we refer to as *triangulation*, pairs up the consumer of the CRA function (typically a network element controller) with a "primary" and a "verify CRA." The verify CRA checks that the primary CRA is performing correctly, and corrupt behavior can be detected by using appropriate protocols amongst the consumer and the primary and verify CRAs.

PHAROS is a system that dynamically applies network resources to satisfy subscriber requests in an efficient and timely manner. It can be applied to a broad range of service models, topologies, and network technologies. Such broad

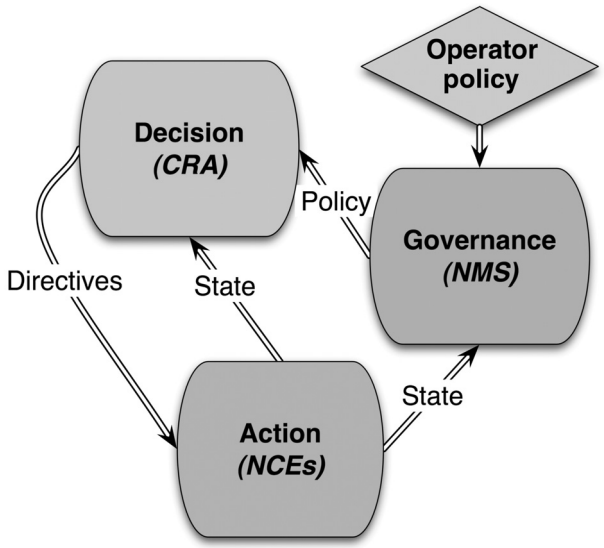



**Figure 8.2** Functional components comprising the PHAROS system.

applicability is possible because our high-level architecture remains technology-agnostic; the benefit is that PHAROS can provide new capabilities and services whether they entail working with legacy infrastructures or with technologies not yet envisioned.

The PHAROS functional architecture separates governance, decision making, and action, streamlining the insertion of new services and technologies. The relationships among these roles are summarized in Figure 8.2.

The governance role is critical for correct operation but is not time-critical. Governance establishes policy and reaction on a human timescale. It is not on the critical path for service instantiations. The network management system (NMS), described further below, is the primary repository of nonvolatile governance information and the primary interface between human operators and the network. For the human operator, PHAROS maintains the functionality of a single, coherent networkwide NMS; this functionality is robustly realized by an underlying multiple-agent implementation.

The decision role is the application of policy to meeting subscriber service requests, and is therefore highly time-critical. That is, the role lies in the critical path for realizing each service request on demand: the decision process is applied to each subscriber service request to create directives for control of network resources. The cross-layer-resource allocator (CRA) function, described further in the next section, is the primary owner of the decision process. Because of the critical contribution of the decision role to the network's speed, efficiency, and resilience, the CRA function is implemented by a distributed hierarchy of CRA instances. In conjunction with our mechanisms for scoping, verification, and failover, the instance hierarchy autonomously sustains our unitary strategy

for resource management: each service request and each network resource within a domain is managed by exactly one CRA instance at any point in time, with a dynamic choice of the particular CRA instance. The result is globally consistent resource allocation, with consistently fast service setup.

The action role implements the decisions made in the decision role. It is a time-critical function. The responsibility of the action role is limited to implementing directives. The network element controllers (NECs), described in a later section, are the primary architectural components responsible for this role.

Our approach allows network operators to take advantage of technological improvements and emerging service models to meet the increasing requirements of their subscribers' applications. The governance function controls the behavior of the PHAROS system, establishing those actions and parameters that will be performed automatically and those that require human intervention. The decision function applies these policies to effectively allocate resources to meet real-time subscriber service requests. The action function implements the decisions quickly, reporting any changes in the state of the system.

## 8.4 Resource allocation

We begin with a discussion of possible resource allocation strategies and describe our approach. In Section 8.4.2 we discuss means of protecting allocated resources from failure. To be agile, we use the concept of "playbooks" described in Section 8.4.3. We conclude in Section 8.4.4 with a short description of our "grooming" approach for increasing resource utilization.

### 8.4.1 Resource management strategies

A key architectural decision in any communications network is the organization of the control of resources. Two of the most important aspects are whether global state or just local state is tracked, and how many nodes participate. Based on these and other choices, approaches range from "fully distributed" where each node participates using local information, to "fully centralized" where resource control is in the hands of a single node utilizing global information. We first discuss the pros and cons of several points in this spectrum and then motivate our choice.

The fully centralized or *single master* strategy entails a single processing node that receives all setup requests and makes all resource allocation decisions for the network. This approach allows, in principle, global optimization of resource allocation across all network resources. It has the further virtue of allowing highly deterministic setup times: it performs its resource calculation with full knowledge of current assignments and service demands, and has untrammeled authority to directly configure all network resources as it decides. However, it requires a single processing node with sufficient capacity for communications, processing, and

memory to encompass the entire network's resources and demands. This node becomes a single point of failure, a risk typically ameliorated by having one or more additional, equally capacious standby nodes. Moreover, each service request must interact directly with the master allocator, which not only adds transit time to service requests (which may need to traverse the entire global network) but also can create traffic congestion on the signaling channel, potentially introducing unpredictable delays and so undercutting the consistency of its response time.

At the other end of the spectrum is the fully distributed or *path threading* strategy. Each node controls and allocates its local resources, and a setup request traces a route between source and destination(s). When a request reaches a given node, it reserves resources to meet the request, based on its local knowledge, and determines the next node on the request path. If a node has insufficient resources to satisfy a request, the request backtracks, undoing the resource reservations, until it fails or reaches a node willing to try sending it along a new candidate path. This strategy can yield very fast service setup, provided enough resources are available and adequately distributed in the network. There is no single point of failure; indeed, any node failure will at most render its local resource unavailable. Similarly, there is no single focus to the control traffic, reducing the potential for congestion in the signaling network. However, the strategy has significant disadvantages. Setup times can be highly variable and difficult to predict; during times of high request rates, there is an exceptionally high risk of long setup times and potential thrashing, as requests independently reserve, compete for, and release partially completed paths. Because backtracking is more likely precisely during times when there are already many requests being set up, the signaling network is at increased risk of congestive overload due to the nonlinear increase in signaling traffic with increasing request rate. The path-threading strategy is ill-suited to global optimization, as each node makes its resource allocations and next-hop decisions in isolation.

One middle-of-the-road strategy is *pre-owned resources*. In this strategy, each node "owns" some resources throughout the network. When a node receives a setup request, it allocates resources that it controls and, if they are insufficient, requests other nodes for the resources they own. This strategy has many of the strengths and weaknesses of path-threading. Setup times can be very quick, if sufficient appropriate resources are available, and there is no single point of failure nor a focus for signaling traffic. Under high network utilization or high rates of service requests, setup times are long and highly unpredictable; thrashing is also a risk. Most critically, resource use can be quite suboptimal. Not only is there the issue of local knowledge limiting global optimization, there is also an inherent inefficiency in that a node will pick routes that use resources it owns rather than ones best suited to global efficiency. In effect, every node is reserving resources for its own use that might be better employed by other nodes setting up other requests.

*Which of these strategies, if any, are appropriate for the next-generation core optical network?* Future core networks present some unique factors influencing

our choice of PHAROS control organization. First, there is adequate signaling bandwidth and processing resources available, which allow for global tracking of resource use if necessary. Second, nodes are neither mobile nor disruption prone, again making it feasible to concentrate control functionality. Third, under high loads, efficient (preferably optimal) allocation is required. Fourth, the stringent service requirements and expectations make the user of the core optical system highly intolerant of stability issues.

We believe that these factors shift the optimum point significantly toward a centralized control for PHAROS although not completely. In essence, our approach is to move away from a single point of failure but retain the ability to use global information for resource allocation decisions resulting in a strategy that we term *unitary resource management*. The unitary strategy relies upon the previously described CRA function to determine the optimal joint resource use for a service and its protection, integrating optimization across multiple layers of technology (e.g. wavelengths, sub-wavelength grooming, and IP).

In the unitary strategy, system mechanisms autonomously sustain the following three invariants across time and across network changes: (1) the integrated CRA algorithm is sustained by a resilient hierarchy of CRA instances; (2) for each request for a given combination of service class, source, and destination(s), there is exactly one CRA instance responsible at any time; and (3) for each network resource there is exactly one CRA instance controlling its allocation at any time. Each CRA instance has an assigned scope that does not overlap with that of any other CRA instance; its scope consists of a service context and a suite of assigned resources. The service context defines the service requests for which this CRA instance will perform setup: a service context is a set of tuples, each consisting of a service class, a source node, and one or more destination nodes.

The unitary strategy allows a high degree of optimization and highly consistent setup times, as a CRA instance can execute a global optimization algorithm that takes into account all resources and all service demands within its scope. There is no backtracking or thrashing, and no risk of nonlinear increases in signaling traffic in times of high utilization or of high rates of setup requests. There is some risk of suboptimal resource decisions, but the PHAROS architecture allows for background offline measurement of the efficacy of allocation decisions and the reassignment of resources or service contexts by the NMS function. The unitary strategy uses multiple CRA instances to avoid many of the problems of the single master strategy: under the PHAROS mechanisms for scoping, failover, and mutual validation, a hierarchy of CRA instances provides load distribution, fast local decisions, and resilience against failure, partition, or attack. Moreover, by concentrating routing and resource-assignment decisions in a few computationally powerful nodes, the strategy allows for complex optimizations based on a global picture, while reducing switch cost and complexity. The CRA instances are maintained on only a small subset of the network nodes, which can be accorded higher security and a hardened physical environment if network policy so chooses.

In the case of the CORONET target network, three CRA instances are used, one CONUS based, one in Europe, and one in Asia.

### 8.4.2 Protection

Protection can be *link-based*, *segment-based* or *path-based*. We summarize the pros and cons of these approaches below.

In *link-based protection*, for each interior element along the primary path, a protection route is found by omitting that one element from the network topology and recalculating the end-to-end path. Thus for each protected path there is a set of $n$ protection paths where $n$ is the number of interior elements on the primary path. These paths need not be (and usually are not) interior-disjoint from the primary path or from one another. For a single failure, link-based protection may give an efficient alternate route; however, the approach faces combinatorial explosion when protecting against multiple simultaneous failures.

In *segment-based protection*, like in link-based protection, a protected primary path is provided with a set of $n$ protection paths, one for each interior link or node. A given protection path is associated with one of these interior elements; it is not based on the end-to-end service requested but simply defines a route around that element. A classic example of segment-based restoration can be found in SONET Bi-directional Line Switched Rings (BLSR), where any one element can fail and the path is rerouted the other way round the ring. Because segment-based restoration paths are independent of any particular primary path, they may be defined per failed element instead of per path. However, they can also be highly non-optimal from the perspective of a specific service request, and are ill-suited to protecting against multiple simultaneous failures.

*Path-based protection* defines one or more protection paths for each protected primary path. A primary with one protection path is said to be singly protected; a primary with two protection paths is doubly protected; and similarly for higher numbers. Each protection path for a primary is interior-disjoint with the primary path and interior-disjoint with each of the primary path's other protection paths (if any). Practical algorithms exist for jointly optimizing a primary path and its protection path(s).

In the current PHAROS implementation, we use the path-based protection strategy. Relative to other protection approaches, path-based protection maximizes bandwidth efficiency, provides fast reaction to partial failures, and is readily extended to protection against multiple simultaneous failures. Further, fault-localization is typically not required to trigger the restoration process. In the past, one of the drawbacks was the number of cross-connections that might need to be made to create a new end-to-end path; however, with schemes based on pre-connected subconnections, invoked in the PHAROS implementation, this is less of an issue. The main drawback is higher signaling load for protection.

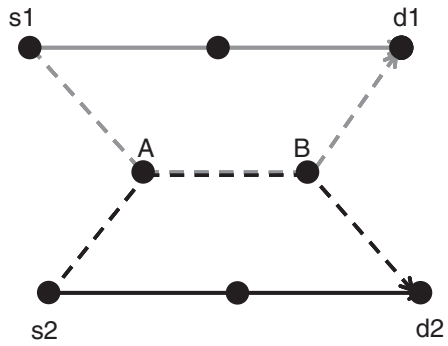



**Figure 8.3** Illustration of shared protection.

#### 8.4.2.1 Shared protection

Having selected path-based restoration for the CRA function, there is still a choice of approaches for allocating the network resources required to ensure that each protection path is supported if a failure (or combination of failures) requires its use. Broadly speaking, there are two general protection strategies for path-based restoration: dedicated and shared [18].

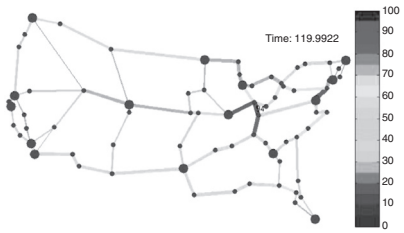
In dedicated protection, each protection path has reserved its own network resources for its exclusive use. In shared protection, a subtler strategy significantly reduces the total amount of network resources reserved for protection while providing equal assurance of path restoration after failures. It is based on the observation that for a typical given failure or set of failures, only some primary paths are affected, and only some of their protection paths (in the case of multiple failures) are affected. Thus, a protection resource can be reserved for use by an entire set of protection paths if none of the failures under consideration can simultaneously require use of that resource by more than one path in the set.

PHAROS uses shared (or "joint") protection, which significantly reduces the total amount of network resources reserved for protection while providing equal assurance of path restoration after failures. Shared protection is illustrated by an example in Figure 8.3, where there are two primary paths (the solid gray and black lines), each with its own protection path (the dotted gray and black lines).

No single network failure can interrupt both primary paths, because they are entirely disjoint. So to protect against any single failure, it is sufficient to reserve nodes A and B and the link between them for use by either the gray or the black protection path: a failure that forces use of the dotted gray path will not force use of the dotted black path, and vice versa. For an in-depth treatment of shared protection in practice, the reader is referred to [18].

Shared protection provides substantial savings in bandwidth. Figure 8.4 shows an example network and the capacity used by dedicated and shared protection


(a) Dedicated Protection




(b) Shared Protection

**Figure 8.4** Capacity (in lambdas) used by dedicated (top) and shared (bottom) protection strategies.

strategies respectively. Shared protection uses about 34 percent less overall capacity in this example.

### 8.4.3 Playbooks

One significant contribution to agility in the PHAROS architecture is a strategy we term playbooks. A *playbook* is a set of pre-calculated alternatives for an action (such as selecting a protection path) that has a tight time budget. The playbook is calculated from the critical path for that action using the CRA function's global knowledge and optimization algorithms. The playbook is stored on each instance that must perform the action; on demand, each instance then makes a fast dynamic selection from among the playbook's alternatives. Playbooks are used to ensure fast, efficient resource use when the time constraints on an action do not allow the computation of paths on demand. In the PHAROS architecture, we use playbooks in two situations: for *Very Fast Service Setup (VFSS)*, and for *Restoration*. We describe the approach used in PHAROS for both of these below.

#### 8.4.3.1 Very Fast Service Setup (VFSS) playbooks

Our current approach is that for each (src, dest, demand rate) combination, we precompute and store two bi-paths:

- First bi-path: The bi-path with the minimum sum of optical edge distances in the working and protection paths. It is computed a priori based only on the topology and as such ignores the network load and protection sharing.
- Second bi-path: Saved copy of the last computed optimal bi-path. The bi-path is optimal in the sense that it minimizes a combined load- and shared-protection-aware cost metric. Since some time has elapsed since the bi-path was initially computed, it may no longer be optimal (or even valid).

The first bi-path is dependent only on the network topology, and needs to be computed only during initialization or after topology changes (a rare event). Note that a link failure is not interpreted as a topology change. For the second bi-path the CRA is constantly running a background process that iterates through the list of valid (src, dest, rate) triplets and computes these optimal bi-paths based on the instantaneous network conditions. Once all (src, dest, rate) combinations have been considered, the process starts once again from the top of the list. Thus, when a new demand arrives, the last saved copy of the corresponding bi-path is just a few seconds old.

In addition, a third bi-path is computed when a new demand arrives. The primary path is computed using Dijkstra's shortest path first (SPF) algorithm where the optical edge costs are related to the current network load. Once the primary path is computed, its links and nodes are removed from the topology, the costs of protection links conditional on the primary path are determined, and then the protection path is computed by running the Dijkstra algorithm again. Since the removal of the primary path may partition the network, there is no guarantee that this bi-path computation will succeed.

These three bi-paths are incorporated into a playbook for that (src, dest, demand rate) combination and cached in the primary CRA (pCRA) instance for the source node. Because VFSS playbooks reside uniquely in the source node's pCRA, there is no possibility of inconsistency. Finally, when an instance receives a demand for that (src, dest, demand rate) combination, it computes the costs of these three bi-paths, taking into account the current network resource availability, and selects the cheapest valid bi-path.

#### 8.4.3.2 Restoration playbooks

A particular failure, such as a fiber cut, may affect thousands of individual demands. Computing alternative paths for all of these demands (for path-based restoration) within the restoration budget is not feasible. Furthermore, unless the resources in the protection path are preallocated, there is no guarantee that a particular demand will successfully find an alternate path after a failure. Thus, path-based protection requires the protection path to be computed along with the primary path, and the resources in the protection path to be reserved.

For each existing demand, there is a playbook entry specifying the path (or paths, for doubly and triply protection demands) to use in case the primary path fails. Each entry specifies the path and regeneration and grooming strategies,

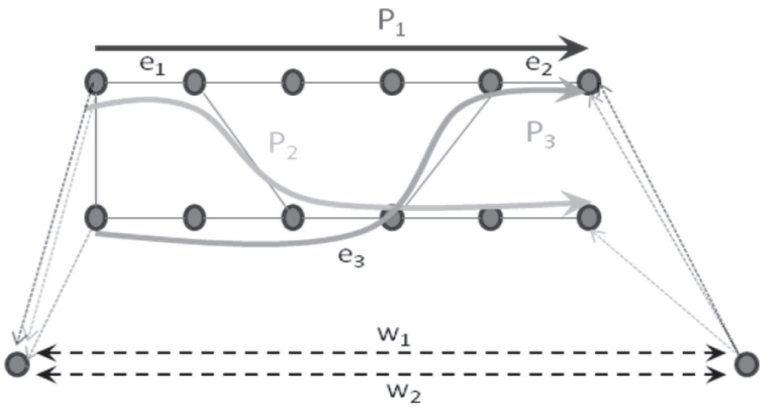



**Figure 8.5** PHAROS restoration playbooks allow efficient on-demand selection of restoration paths.

and identifies the pool of resources (such as wavelengths) to choose from upon failure. The playbook does not specify the resource to use, as such assignment can be made (efficiently under shared protection) only after a failure occurs, as illustrated in Figure 8.5.

Paths $W_1$ and $W_2$ are enough to protect $P_1$, $P_2$, and $P_3$ against any single failure. However, it is not possible to uniquely assign a wavelength to each demand before the failure occurs. For example, suppose we were to assign $W_1$ to $P_1$. Since $P_1$ and $P_2$ are both affected by link e1 failure, then $P_2$ should be assigned $W_2$. Similarly, since $P_1$ and $P_3$ are both affected by link $e_2$ failure, $P_3$ should also be assigned $W_2$. However, $P_2$ and $P_3$ should not be assigned the same wavelength, since this will result in blocking if link $e_3$ fails.

### 8.4.4 Sub-lambda grooming

Finally, many demands do not fill a full wavelength. If one such demand is uniquely assigned to a full wavelength, without sharing it with other demands, it will result in wasted bandwidth and long-reach transponders. To alleviate this problem, demands can be aggregated into larger flows at the source node. They can also be combined with other nodes' demands at intermediate nodes (a process we refer to as sub-lambda grooming, or SLG) so that wavelength utilization at the core is close to 100%. Once demands are sub-lambda-groomed, they can be optically bypassed.

Deciding where and when to sub-wavelength-groom demands is a difficult optimization problem. It must take into account different tradeoffs among capacity available, the cost (both capital and operational) of the SLG ports and transponders, and the fact that constantly adding or removing demands will unavoidably result in fragmentation inside a wavelength. What may appear to be a good grooming decision now may hurt performance in the future. Grooming decisions,

then, must balance medium- to long-term resource tradeoffs and be based on medium-term traffic patterns.

Within our topology-abstraction-based architecture, grooming is a generalized operation where each level packs its smaller bins into larger bins at the level immediately below. Currently, we have a three-level system where we aggregate and groom sub-lambda demands into full wavelengths, and wavelengths onto fibers. However, aggregation and grooming of smaller bins into larger bins constitutes a fundamental operation that repeats itself at multiple layers.

## 8.5 Signaling system

The PHAROS signaling architecture is designed to support operations in the control as well as management planes. Its function is the delivery of data between the elements of the architecture in a timely, resilient, and secure fashion. The main requirements for the signaling architecture are:

- *Performance*: the architecture must support the stringent timing requirements for connection setup and failure restoration.
- *Resiliency*: the architecture must be resilient to simultaneous failures of several elements and still be able to perform the most critical functions.
- *Security*: the architecture must support flexible security arrangements among architectural elements to allow for proper authentication, non-repudiation and encryption of messages between them.
- *Extensibility*: the architecture must be extensible to be able to accommodate new features and support the evolution of the PHAROS architecture.

The PHAROS signaling and control network (SCN) is the implementation of the PHAROS signaling architecture. It allows NECs to communicate to potential CRA/NMS, with signaling links segregated from the data plane to minimize the risk of resource exhaustion and interference attacks. The PHAROS architecture supports an SCN topology that is divergent from the fiber-span topology, and does not require that the network element controllers and network elements be co-located. For next-generation core optical networks providing deterministic and minimal delay in signaling for service setup and fault recovery, it is recommended that the SCN be mesh-isomorphic to the fiber-span topology, and the network element controllers be collocated with the network elements as shown in Figure 8.6. This configuration minimizes the signaling delay for service setup and fault recovery.

Based on bandwidth sizing estimates that take into account messaging requirements for connection setup, failure signaling and resource assignment, a 1 Gb/s channel is sufficient to maintain stringent timing for setup and restoration under heavy load and/or recovery from multiple fault scenarios. Two performance goals drive the channel size requirements for the PHAROS SCN: very fast service

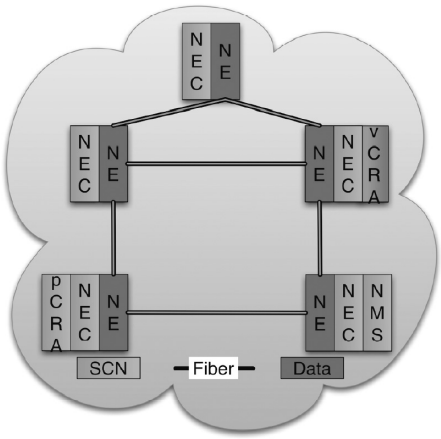



**Figure 8.6** The signaling and control network (SCN) connects network elements (NE) and their associated network element controllers (NEC), cross-layer resource allocator (CRA) and network management system (NMS).

setup and 50-ms-class restoration from simultaneous failures. The sizing estimates assume worst case signaling load for a 50-Tb/s-capacity 100-node global fiber network with service granularity ranging from 10 Mb/s to 800 Gb/s. Fibers connecting nodes were presumed to carry 100 100-Gb/s wavelengths.

The majority of the signaling traffic (with some exceptions) travels through the links that constitute the SCN topology. Thus the signaling architecture accommodates both the control and the management planes. Each link in the SCN has sufficient bandwidth to support the peak requirements of individual constituent components. This is done to reduce queueing in the signaling plane, thus expediting the transmission of time-critical messages. Additionally, to ensure that the time-critical messages encounter little-to-no queueing delay, each link is logically separated into a *Critical Message Channel (CMC)*, and a *Routine Message Channel (RMC)*. All time-critical traffic, such as connection setup messages and failure messages, travels on the CMC, while the rest (including the management traffic) use the RMC.

As in traditional implementations, the SCN is assumed to be packet-based (IP) and to possess a routing mechanism independent of the data plane that allows architectural elements to reach one another outside of the data plane mechanisms.

### 8.5.1 Control plane operation

In our approach to connection setup, the two competing objectives are:

- The need to perform connection setup very rapidly (for Fast Service Setup (FSS) and Very Fast Service Setup (VFSS) service classes).
- The need for global optimization of protection, which requires the entity responsible for path computation, the primary CRA instance (pCRA), to have a global view of the network.

There are two basic approaches to connection setup: *NEC-controlled* and *CRA-controlled*. They vary in complexity of implementation, the tradeoffs being between the connection setup speed and the need for a global view of resource allocation.

In the NEC-controlled approach, the NEC instance at the source node communicates with its assigned pCRA to receive the information about the routes for working and protection paths and then, in a manner similar to RSVP-TE with explicit route option, sends signaling messages along these paths to the affected NEC instances to service this request. (RSVP-TE reserves resources on the forward path and configures the network elements on the reverse path from destination to source.). This approach has the advantage of fitting into the traditional view of network traffic engineering. One issue in the approach is connection-setup latency: the NEs on each path are configured serially, after the response from the pCRA is received, with processing at each NEC incurring additive delays. Adding to the initial delay of requesting path information from the pCRA makes this approach too slow to be applied in the case of very fast and fast connection classes.

In the CRA-controlled approach, the NEC instance in the source node communicates the service setup request and parameters to its assigned pCRA and leaves it up to this CRA instance to compute the optimal path and to instruct individual NEC instances on the computed path to configure their NEs for the new connection. This approach has several advantages over the NEC-controlled approach. First, NEC configuration occurs in parallel, which serves to speed up connection setup. Second, only CRA instances are allowed to issue NE configuration requests to NECs, which is a desirable property from the viewpoint of network security, as it allows PHAROS to leverage strong authentication mechanisms in NEC-to-CRA communications to prevent un-authorized node configurations. The disadvantage of this approach is its scalability, as a real network may contain a large number of NEC instances, and having a single pCRA presents a scalability limit.

Given our requirement of supporting very fast connection setups, the serialization delay incurred by the NEC-controlled approach is prohibitive. We therefore use the CRA-controlled approach, but address the disadvantage by exploiting the unitary resource management strategy (see Section 8.4). In other words, by dividing the space of all possible service requests into disjoint scopes, a hierarchy

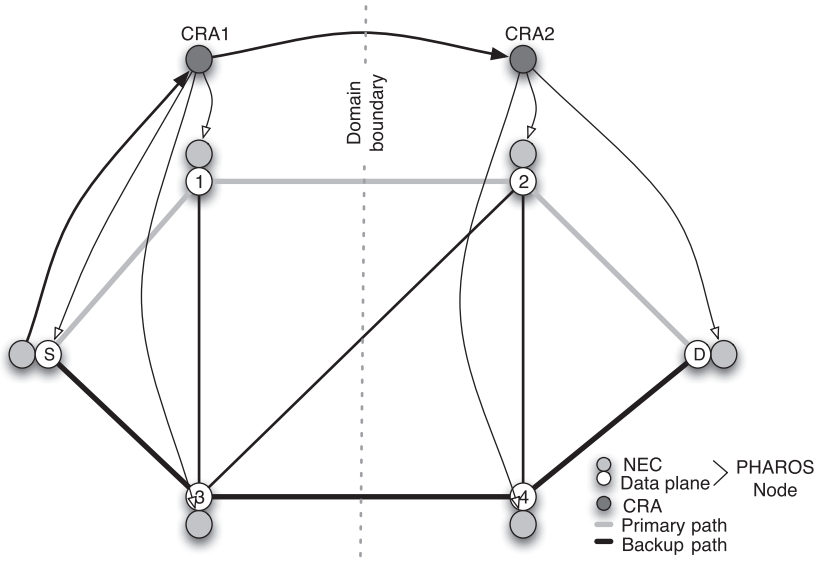



**Figure 8.7** The PHAROS *federated* approach for signaling across domains.

of pCRA instances can divide the load while providing more local pCRA access for localized requests.

In an approach that we term the *federated* approach, the initial NEC contacts the pCRA in its scope with a service-setup request as illustrated in Figure 8.7. The pCRA maps the service's path onto domains based on its network state and provisions the service path across its own domain, while at the same time forwarding the request to the appropriate pCRAs in the neighboring domains. This approach deals with the interdomain state consistency problem by leveraging the fact that a pCRA is not likely to have up to date information about its own domain and somewhat stale information about other domains. This approach also accommodates security requirements by restricting the number of CRA instances that may configure a given NEC to only those within its own domain. It also retains the parallelizing properties, thus speeding up connection setup.

### 8.5.2 Failure notification

Traditionally, in MPLS and GMPLS networks, failure notifications are sent in point-to-point fashion to the node responsible for enabling the protection mechanism. This approach works when the number of connections traversing a single fiber is perhaps in tens or hundreds. In PHAROS, assuming the worst-case combination of fine-granularity connections (20 Mbs) and large capacity of a single fiber (10 Tbs), the total number of connections traversing a single fiber may number in tens or hundreds of thousands (500 k connections in this case). It is infeasible from the viewpoint of the signaling-plane bandwidth to be able to sig-

nal individually to all restoration points in case of a failure, because the number of points and connections whose failure needs to be signaled may be very large. Additionally, point-to-point signaling is not resilient to failures, meaning that, due to other failures in the network, point-to-point failure messages rely on the SCN routing convergence to reach intended recipients and to trigger protection, which may be a lengthy process.

The solution we adopted in PHAROS relies on two simultaneous approaches:

- Signaling on aggregates that can indicate failure of a large number of connections at once.
- Using intelligent flooding as a mechanism to disseminate failure information.

The first approach significantly cuts down on the amount of bandwidth needed to signal a failure of many connections resulting from a fiber cut, but it requires that the nodes receiving failure notifications are able to map the failed aggregates to specific connections requiring protection actions.

The second approach, in addition to reducing bandwidth requirements, also has the desirable property that a signaling message always finds the shortest path to any node in the network even in the presence of other failures, without requiring the signaling-plane routing to converge after a failure.

Combined, these two approaches create a PHAROS solution to the failure-handling problem that is resilient and scalable and addresses the stringent restoration-timing requirements.

## 8.6 Core node implementation

In this section we discuss a core node implementation that is designed to optimize the capabilities of the PHAROS architecture. We note that the PHAROS architecture does not depend upon the core node being implemented this particular way – as mentioned earlier, it is technology agnostic.

The PHAROS core node design focuses on maximizing flexibility and minimizing the complexity of intra-node ports required to provide the complete range of PHAROS services and reducing the capital and operational costs per unit of bits. The primary objectives identified to satisfy this vision include: (1) arrange subscriber traffic onto wavelength and sub-wavelength paths to enable switching at the most economic layer, (2) enable shared protection, and (3) enable transponders to be repurposed to service both IP and wavelength services and also service transit optical–electrical–optical (OEO) regeneration functions. When combined with a control plane designed for optimum resource allocation, the PHAROS optical node is highly adaptable to incoming service requests. The PHAROS node architecture defines the principal hardware systems extending from the

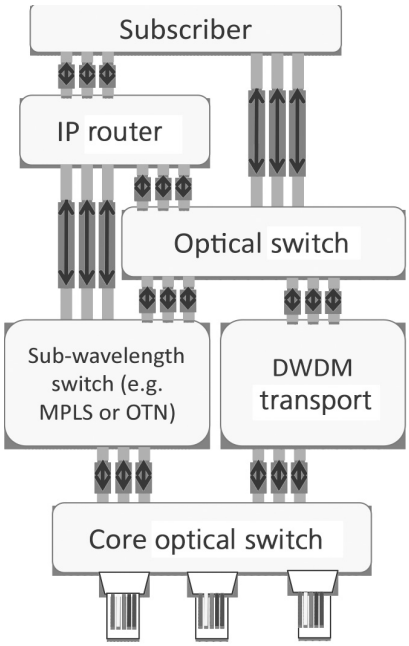



**Figure 8.8** PHAROS core node implementation showing various optical network elements.

fiber connections with the subscriber facility to the fiber connections in the physical plant of the core network, as illustrated in Figure 8.8.

The PHAROS node is composed of the following elements:

- Subscriber service layer connections to bring client services into the core node.
- Edge router (packet switch) to support best effort IP services.
- Fast optical switch to allow sharing of sub-wavelength switch and transport ports.
- Sub-lambda grooming switch and DWDM transport platform to support full and sub-wavelength switched (via MPLS, OTN or SONET) and packet services with fast setup, tightly bounded jitter specifications. This equipment also provides OEO regeneration.
- Core optical switch to manage optical bypass, optical add/drop, and routing between optical fibers.

Note that these elements may or may not be instantiated in the same hardware platform. The PHAROS architecture emphasizes configuration, and can be applied to a variety of different network element configurations.

The core node implementation results in reduced excess network capacity reserved for protection via protection sharing between IP, TDM, and wavelength services that arise at the Subscriber ports. The desire to support guaranteed QoS

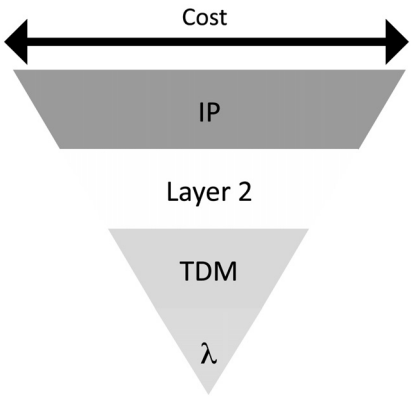



**Figure 8.9** Qualitative representation of "hop cost" for services on the network. IP hops are the highest cost, optical (lambda) the lowest. Cross-layer resource management chooses the network path with the minimum total "hop cost."

and high transport efficiency is supported either via TDM switching to realize the hard guarantees for latency or MPLS, for higher transport efficiency, depending on the needs of the particular carrier. Across the network, reduced equipment and port costs are realized by minimizing average hops at high-cost layers via dynamic cross-layer resource allocation. Layer cost is represented pictorially in Figure 8.9. Most high-performance packet-based router/switches include hidden convergence layers in the switch, which adds more buffering and switch PHY costs. TDM switches (SONET, OTN) operate directly at their convergence layer, which is the main reason they are much simpler/less costly. The minimum cost node hop is at the optical layer. The use of colorless and directionless all-optical switching, though not required since OEO processes can be used, can reduce the number of OEO ports by as much as 30 percent in the global network configuration. In colorless switching, any wavelength may be assigned to any fiber port, removing the restriction common in today's reconfigurable optical add/drop multiplexers where a given wavelength is connected to a particular fiber port. Directionless switching means the ability to cross-connect any incoming port to any outgoing port in a multi-degree configuration.

## 8.7 Performance analysis

We have created a high fidelity OPNET simulation of the PHAROS system. Figure 8.10 compares the performance of three protection approaches: (1) dedicated protection in which each primary path receives its own protection path; (2) shared protection, where a set of protection paths may share a resource as explained in Section 8.4; (3) opportunistic shared protection, a sophisticated

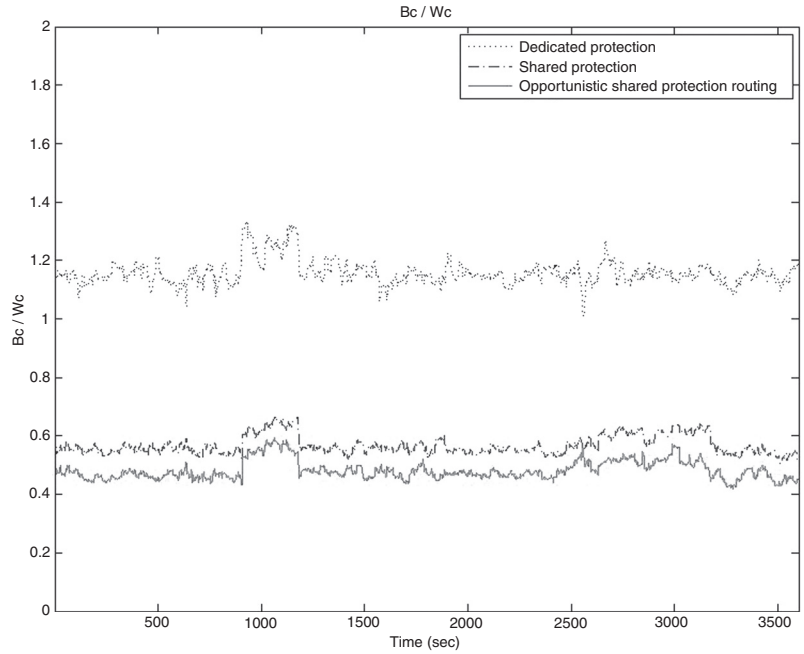



**Figure 8.10** B/W comparison of different protection strategies

version of (2) where the protection paths are chosen to maximize shared protection opportunities.

Requests for bandwidth are generated over time. For each approach, we plot the $B/W$ metric as a function of time; $B/W$ is defined as the Backup (Protection) over Working capacity ($B$ and $W$ are in units of km-wavelength), which is a rough measure of the relative cost incurred in protection. Thus, the lower the $B/W$, the better.

Results shown here are for a 100-node optical network model with 75 nodes in the Continental US (CONUS), 15 in Europe and 10 in Asia. The line rate is 40 Gb/s, and the aggregate traffic is 20 Tb/s of which 35% is IP traffic and 65% is wavelength services; 90% of the source-destination pairs are within the US. The bit-averaged distance for the intra-CONUS traffic is about 1808 km. The B/W numbers shown in Figure 8.10 are for CONUS-contained resources only.

We see that the PHAROS shared protection strategies significantly outperform dedicated protection. Specifically, shared protection has about 50% lower B/W than dedicated, and opportunistic improves this further by about 10%.

## 8.8 Concluding remarks

The emergence of data as the dominant traffic and the resultant unpredictabilty and variability in traffic patterns has imposed several challenges to the design

and implementation of the core network – from agile, autonomous resource management, to convergence and L1–L2 integration to the signaling system.

In this chapter we have described the architecture of a future core network control and management system along with a node implementation that enables future scalable and agile optical networks developed as part of the DARPA/STO CORONET program. This work provides control and management solutions that support services across core network dimension with 50-ms-class setup time, and also to respond to multiple network failures in this time frame. It provides a method of cross-layer resource allocation that delivers efficient allocation of bandwidth, both working and protection, to services at all layers in the network, including IP and wavelength services. Preliminary evaluations show significant advantages in using PHAROS.

The architecture described in this chapter enables core network scale beyond 10X of today's networks by optimizing path selection to maximize optical bypass, and minimize the number of router hops in the network. As a result, a higher capacity of network services can be supported with less network equipment.

## References

[1] Simmons, J. On determining the optimal optical reach for a long-haul network *Journal of Lightwave Technology* 23(3), March 2005.

[2] Simmons, J. Cost vs. capacity tradeoff with shared mesh protection in optical-bypass-enabled backbone networks *OFC/NFOEC07, Anaheim, CA, NThC2* March 2007.

[3] Dutta, R. and Rouskas, G. N., Traffic grooming in WDM networks: past and future *Network, IEEE*, 16 (6) 46–56, Nov/Dec 2002

[4] Iyer, P., Dutta, R., and Savage, C. D. On the complexity of path traffic grooming *Broadband Networks, 2005 2nd International Conference*, pp. 1231–1237 vol. 2, 3–7 Oct. 2005.

[5] Zhou, L., Agrawal, P., Saradhi, C., Fook, V. F. S. Effect of routing convergence time on lightpath establishment in GMPLS-controlled WDM optical networks *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 3 pp. 1692–1696 16–20 May 2005.

[6] Saleh A., and J. Simmons Architectural principles of optical regional and metropolitan access networks, *Journal of Lightwave Technology* 17(12), December 1999.

[7] Simmons J. and A. Saleh The value of optical bypass in reducing router size in gigabit networks *Proc. IEEE ICC 99*, Vancouver, 1999.

[8] Saleh A. and Simmons, J. Evolution toward the next-generation core optical network *Journal of Lightwave Technology* 24(9), September 2006, 3303.

[9] Bragg A., Baldine, I., and Stevenson, D. Cost modeling for dynamically provisioned, optically switched networks *Proceedings SCS Spring Simulation Multiconference*, San Diego, April 2005.

[10] Brzezinski, A and Modiano, E., Dynamic reconfiguration and routing algorithms for IP-over-WDM networks with stochastic traffic *Journal of Lightwave Technology* 23(10), 3188–3205, Oct. 2005.

[11] Strand, J., and Chiu, A. Realizing the advantages of optical reconfigurability and restoration with integrated optical cross-connects *Journal of Lightwave Technology*, 21(11), November 2003. 2871.

[12] Xin, C., Ye, Y., Dixit, D. and Qiao, C. A joint working and protection path selection approach in WDM optical networks *Global Telecommunications Conference, 2001. GLOBECOM '01. IEEE*, pp 2165–2168 vol.4, 2001

[13] Kodialam, M. and Lakshman, T. V., Dynamic routing of bandwidth guaranteed tunnels with restoration *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings*. pp. 902–911 vol. 2, 2000

[14] Ou, C., Zhang, J., Zang, H., Sahasrabuddhe L. H., and Mukherjee, B. New and improved approaches for shared-path protection in WDM mesh networks *Journal of Lightwave Technology*, pp. 1223–1232, May 2004.

[15] Ellinas, G., Hailemariam, A.G., and Stern, T.E. Protection cycles in mesh WDM networks *IEEE Journal on Selected Areas in Communications*, 18(10) pp. 1924–1937, Oct 2000

[16] Kodian, A., Sack, A., and Grover, W.D., p-cycle network design with hop limits and circumference limits *Broadband Networks*, 2004. Proceedings of the First International Conference on Broadband Networks, 244–253, 25–29 Oct. 2004

[17] Gripp, J., Duelk, M., Simsarian, M. *et al.* Optical switch fabrics for ultra-high-capacity IP routers *Journal of Lightwave Technology*, 21(11), 2839, (2003).

[18] Simmons, J. M. Optical network design and planning in *Optical Networks*, B. Mukherjee Series editor, Springer 2008