Architectures and Protocols for Capacity Efficient, Highly Dynamic and Highly Resilient Core Networks [Invited]

Angela L. Chiu, Gagan Choudhury, George Clapp, Robert Doverspike, Mark Feuer, Joel W. Gannett, Janet Jackel, Gi Tae Kim, John G. Klincewicz, Taek Jin Kwon, Guangzhi Li, Peter Magill, Jane M. Simmons, Ronald A. Skoog, John Strand, Ann Von Lehmen, Brian J. Wilson, Sheryl L. Woodward, and Dahai Xu

***Abstract*—The Core Optical Networks (CORONET) program addresses the development of architectures, protocols, and network control and management to support the future advanced requirements of both commercial and government networks, with a focus on highly dynamic and highly resilient multi-terabit core networks. CORONET encompasses a global network supporting a combination of IP and wavelength services. Satisfying the aggressive requirements of the program required a comprehensive approach addressing connection setup, restoration, quality of service, network design, and nodal architecture. This paper addresses the major innovations developed in Phase 1 of the program by the team led by Telcordia and AT&T. The ultimate goal is to transfer the technology to commercial and government networks for deployment in the next few years.**

***Index Terms*—Core networks; Dynamic networks; Network architecture; Network evolution; Network restoration; Optical networks.**

I. Introduction

The Core Optical Networks (CORONET) program addresses the development of architectures, protocols, and network control and management to support the future advanced requirements of both commercial and government networks, with a focus on highly dynamic and highly resilient multi-terabit core networks [1]. The program includes dynamic services with connection setup requirements that are two to three orders of magnitude faster than what is currently possible. It also encompasses catastrophic network failure in that recovery from multiple concurrent failures is required for a subset of the network traffic. The aggregate traffic demand, which is composed of both IP and wavelength services, represents, for the largest case considered, more than a twenty-fold increase over today's traffic levels for the largest of any individual carrier. It is the desired goal of the program to achieve transition of these advances to commercial and government networks in the next few years. Thus, the aggressive requirements must be met with solutions that are scalable, cost effective, and power efficient, while providing the desired quality of service (QoS).

CORONET is a two-phase program, with Phase 1 focused on the development of architectures, protocols, and algorithms, and Phase 2 focused on developing prototype network control and management systems to implement these advances. Phase 1 was completed in 2010 and included extensive simulations to validate the proposed solutions. Phase 2 will extend into 2012.

This paper addresses the major innovations developed in Phase 1 by the team led by Telcordia and AT&T. With respect to wavelength services, the innovations include 1) a novel distributed connection setup protocol with numerous advantages as compared to GMPLS (Generalized Multi-Protocol Label Switching) [2], 2) a powerful routing engine that provides multiple diverse routing paths, 3) a restoration scheme that combines both pre-set (for speed) and dynamic (for cost and capacity efficiency) aspects, and 4) a methodology for determining the number of transponders to deploy at a location. With respect to IP services, standards-based protocols were combined with an "elastic" capacity paradigm for connection setup, and an innovative two-step restoration mechanism was developed to achieve both speed and efficiency.

Section II presents an overview of the CORONET assumptions, and Section III summarizes the CORONET vision, to motivate the work that was performed. The architecture and protocols developed for IP services are covered in Section IV; wavelength-service innovations are covered in Section V. Strategies for handling scheduled services are briefly covered in Section VI, and traffic generation is discussed in Section VII. The nodal architecture is described at a high level in Section VIII, followed by a discussion of a methodology for sizing the pre-deployed nodal transponder pools in Section IX. Security issues are discussed in Section X. Finally, some of the extensions planned for Phase 2 of the program are highlighted in Section XI. Note that an overview and preliminary results of the program were presented in [3]. Additional material on the program may be found in [4–10].

Manuscript received May 4, 2011; revised September 29, 2011; accepted November 2, 2011; published December 7, 2011 (Doc. ID 146874).

Angela L. Chiu, Gagan Choudhury, Robert Doverspike, Mark Feuer, John G. Klincewicz, Guangzhi Li, Peter Magill, John Strand, Sheryl L. Woodward, and Dahai Xu are with AT&T Laboratories, USA.

George Clapp (e-mail: clapp@research.telcordia.com), Joel W. Gannett, Janet Jackel, Gi Tae Kim, Taek Jin Kwon, Ronald A. Skoog, Ann Von Lehmen, and Brian J. Wilson are with Applied Research, Telcordia Technologies, USA.

Jane M. Simmons is with Monarch Network Architects, USA.

Digital Object Identifier 10.1364/JOCN.4.000001

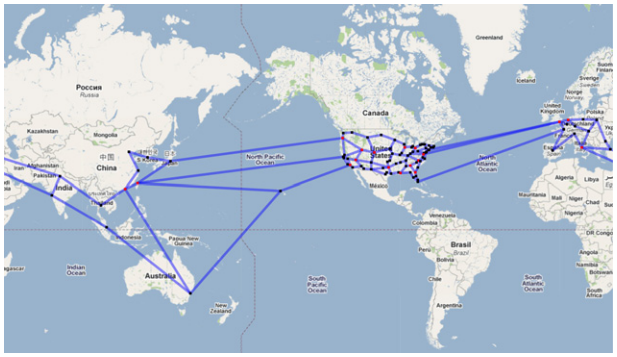

Fig. 1. (Color online) Global CORONET topology.

II. CORONET Overview

The representative global core topology used in CORONET is shown in Fig. 1. The location and connectivity of the nodes are loosely modeled after existing commercial networks, but they are not intended to characterize an actual network in use today. The network is composed of 100 nodes, with 80 designated as *small* nodes and 20 as *large*, and with 75 nodes located inside the continental United States (CONUS). The total number of network links, including the transoceanic links, is 136, yielding an average nodal degree of 2.7. The maximum degree of any node is five. Full details of the topology can be found in [11].

CORONET encompasses both IP services and wavelength (WL) services, with 75% of the traffic bandwidth being IP. Consequently, there are both IP routers and optical switches at each node. The IP traffic is composed of guaranteed bandwidth (GB) services with strict QoS requirements, as well as best-effort traffic. The wavelength services range from single-wavelength to octal-wavelength (i.e., eight wavelengths bundled into a single connection). Table I provides an overview of the characteristics and requirements of both the IP and wavelength services. Realistic traffic models are used, as opposed to uniform all-to-all traffic, with the 20 large nodes participating in the majority of the total IP and wavelength-service bandwidth. To make the traffic model commercially realistic, the wavelength services are restricted to the large nodes and to 20 of the 80 small nodes. This was done because the wavelength services are very demanding and a carrier must deploy significant network resources to support them. The average routed distance for intra-CONUS traffic was roughly 1600 km for IP traffic and 2600 km for wavelength services (only 30 CONUS nodes act as the source or destination of wavelength services, thus resulting in longer paths).

The provisioning time for the *very fast setup* (VFS) traffic class, which applies only to wavelength services, is shown in Table I as 100 ms for CONUS and 250 ms for global connections. This is an approximation for a calculation of the provisioning time requirement as the *round-trip delay* (RTD) + 50 ms, where RTD represents the round-trip fiber transmission delay of the shortest end-to-end working path. The provisioning time for *fast setup* is 2 s without consideration of the RTD. The requirements of the various traffic classes are discussed in detail in Sections IV to VI.

The baseline CORONET architecture is IP-over-optical layer, as shown in Fig. 2. The IP layer is populated with core routers and the optical layer is populated with all-optical switches (AOSs). The term AOS is used here to designate a fully flexible reconfigurable optical add/drop multiplexer (ROADM), as is discussed further in Section VIII.

Four scenarios were investigated, with the aggregate traffic demand ranging from 20 Tb/s to 100 Tb/s, representing a significant increase over the traffic supported by any of today's largest carriers. There is a maximum of 100 wavelengths per fiber, where the wavelength data rate is either 40 Gb/s or 100 Gb/s, depending on the scenario. While hybrid 40 Gb/s and 100 Gb/s scenarios are likely in the future, our studies were based on the original CORONET requirement that there would be only one wavelength rate per fiber. This was done to simplify the analysis. However, even if this condition is relaxed, we do not expect the majority of our results to change in any significant way. In the baseline physical topology, each link is populated with either one or two fiber pairs, depending on the traffic demand scenario, with the CORONET metrics allowing up to 10% of the links to have one additional fiber pair. This tight limit on overall network capacity ensures that the CORONET requirements are addressed by solutions that are capacity efficient.

Optical bypass is assumed in all scenarios, where traffic transiting a node can be kept in the optical domain rather than undergoing electronic regeneration. An optical reach of 2000 km is assumed for 40 Gb/s line rate and 1500 km for 100 Gb/s line rate.

III. CORONET Vision

A. Dynamic Services

Transport networks today are mostly quasi-static, with connection setup typically requiring on-site manual involvement and connections (also called *circuits*) often remaining established for months or years. As an initial transition from this relatively rigid environment, transport networks are becoming configurable, with connections established remotely through software control, assuming that the necessary equipment is already deployed in the network. Configurable networks take advantage of flexible network equipment such as advanced ROADMs and tunable transponders.

The next step in this evolution is dynamic networking, where connections can be rapidly established and torn down without the involvement of operations personnel. Dynamic offerings are currently limited to subwavelength rates with setup times of the order of a minute [12]. CORONET greatly pushes the envelope for dynamic services, extending dynamic networking to wavelength services and providing setup times as fast as 100 ms.

While clearly not all services require a setup time of 100 ms, some applications do depend on rapid network response time. For example, with highly interactive visualization and data fusion, a user may pull together large chunks of data from numerous global locations to form a comprehensive situational awareness. One approach is to set up permanent connections

TABLE I
CORONET TRAFFIC CHARACTERISTICS AND REQUIREMENTS

	Very fast setup (wavelength services only)	Fast setup	Scheduled setup	Semi-permanent
Setup time	CONUS: ≤100 ms Global: ≤250 ms	≤2 s	≤10 s	N/A
Holding time	1 s to 1 min	10 s to 10 h	1 min to 1 month	Months
Restoration	Single failure only	Up to 3 failures	Up to 3 failures	Up to 3 failures
Blocking	10^{-3}	10^{-3}	10^{-4}	N/A

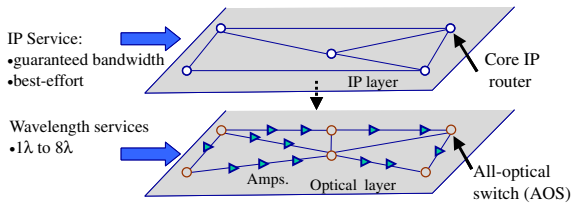

Fig. 2. (Color online) IP-over-optical layer architecture.

TABLE II
BREAKDOWN OF IP AND WAVELENGTH TRAFFIC CLASSES

	Very fast setup	Fast setup	Scheduled setup	Semi-permanent	Best effort
IP (75%)		15%	15%	15%	30%
Wavelength services (25%)	10%	5%	5%	5%	N/A

between all locations that may participate in the process. However, given the numerous locations that may be involved and the relatively small proportion of the time that any one connection is needed, establishing permanent connections can be prohibitively expensive. Dynamic networking is an attractive alternative, where the connection setup time must be of the order of 100 ms to meet the human tolerance for delay with interactive applications. Similarly, global-scale distributed computing can benefit from dynamic networking, where the connections must be established and torn down very rapidly in order to realize significant savings in required capacity. Another application that becomes feasible with very fast service setup is "path hopping" for purposes of added security, where the connection is moved rapidly to a new path to avoid eavesdropping. Additionally, setup times of the order of 100 ms allow for the possibility of restoring a failed connection by issuing another setup request.

We envision that the need for dynamic services will burgeon over the next five to ten years. This growth will likely be a "push–pull" evolution, where the need for on-demand services by technical developments such as cloud computing drives the implementation of dynamic networks, and the availability of a dynamic network infrastructure fuels the development of more applications that can take advantage of the dynamic services.

In CORONET, 50% of the aggregate demand (in terms of bandwidth) is treated as dynamic, either immediately on demand or scheduled, as shown by the shaded cells in Table II.

B. Resilient Networks

Two important characteristics of any restoration scheme are speed of restoration and cost. While it is relatively straightforward today to provide rapid restoration with simple schemes, such as 1 + 1 protection, it proves very costly since restoration resources are dedicated to a single connection. Conversely, capacity efficiency can be achieved by sharing restoration resources among multiple non-overlapping connections; however, the time to recover from a failure can be of the order of seconds. The challenge of CORONET is to provide both rapid and cost-effective restoration. The restoration time metric specified by CORONET is RTD + 50 ms, where RTD represents the round-trip fiber transmission delay of the end-to-end restoration path. Rather than explicitly specifying a cost metric, the amount of capacity required to provide restoration serves as a proxy for restoration cost. In CORONET, the ratio of restoration capacity to working capacity for the CONUS portion of the network should be no larger than 75%. This is referred to as the *spare capacity ratio* (SCR). (Note that the 75% metric is specific to CORONET. It should not be treated as a universal benchmark.)

There are three broad classes of outages to consider when allocating restoration resources. First, there are component failures, e.g., amplifiers, switches, and, most notably, transponders. Availability analysis that takes into account the mean time to failure of each component can be used to determine the amount of required redundancy. Second, there are maintenance events, such as a major software upgrade of an IP router, that require bringing down the entire router. Finally, there are catastrophic outages, where an entire node or link, or even several links, goes out of service. This is especially important in a natural disaster or in a government or military network that may come under attack. The CORONET program includes all of these outage types; however, for simulation purposes, the emphasis is on catastrophic network failures. Although the bulk of the traffic requires restoration from a single node or link failure, a small percentage of the overall traffic (10%) requires restoration from up to two concurrent node or link failures, and a smaller percentage (2.5%) requires restoration from up to three concurrent node or link failures. For the double and triple failure scenarios, at most one of the failures can be a node failure, with the remaining failures affecting network links.

C. Quality of Service

Most of the traffic specified in CORONET represents IP services, where it is envisioned, for example, that current

TABLE III
IP DESIGNS WITH INCREASING NUMBER OF EXPRESS LINKS

Network design	Required capacity in units of 1000 wavelength-km	Number of line-side router ports	Normalized total cost
136 links + 0 express links	1565	2700	113
136 links + 45 express links	1713	1985	101
136 links + 73 express links	1810	1839	100
136 links + 107 express links	2005	1785	105

private-line services will eventually migrate to guaranteed bandwidth IP services. Providing a QoS commensurate with the traffic type is mandatory. To ensure that the advances of CORONET were not being achieved at the expense of service quality, the guaranteed bandwidth IP traffic was designated as either loss sensitive or delay/jitter sensitive, with appropriate metrics to measure performance (the metrics are specified in Subsection IV.C).

D. Network Simulation

The CORONET vision outlined above is extremely challenging and requires major innovations in network design and protocols. The traffic classes are a demanding mix of heterogeneous services with stringent QoS requirements. The rapid provisioning and restoration targets within tight capacity constraints are very aggressive. To demonstrate the validity and performance of the network design algorithms and protocols developed to meet these challenges, we constructed complex simulation models on an OPNET platform and performed thousands of hours of simulations of a 100 node global-scale network carrying 20–100 Tb/s of traffic. The simulation models and results were inspected by an independent team from MIT Lincoln Laboratory to verify their accuracy and completeness in meeting all program metrics.

IV. IP SERVICES

In the overall network design performed in Phase 1 of CORONET, capacity was partitioned between IP services and wavelength services. IP services, though comprising 75% of the aggregate network demand, required only about 25% of the network capacity. This is due to the finer granularity of IP traffic, the shorter path lengths of the IP traffic, and the fact that 40% of the IP services (i.e., the best-effort traffic) did not need to be protected against failures. Future work will examine the efficacy of an optical layer that reconfigures capacity in response to the IP layer. This may provide benefits under conditions of severe traffic change, as would occur, for example, if the network were to experience a modified pattern of traffic due to changes in Internet service provider (ISP) peering patterns.

Reference [3] gives an early description of CORONET's approach to IP services, with a focus on network architecture and design. This section provides a summary and update of this design process and focuses on the overall architecture and enabling protocols.

As stated earlier, there are IP routers and all-optical switches at each of the 100 nodes in the CORONET network. IP links were established between all physically adjacent routers (i.e., routers directly connected by a fiber pair). "Express links" between non-adjacent routers were then added, based on a heuristic design algorithm [10]. An express link is a connection between routers that traverses intermediate nodes and bypasses the routers at those nodes. Because the traffic carried on an express link is not processed by the intermediate routers, the total number of required router ports is reduced.

The heuristic design algorithm for express links takes into account two complementary considerations. On one hand, it seeks to add express links along high traffic-volume paths; on the other, it seeks to avoid having an inordinate number of express links carried on the same optical link to minimize the impact of any single optical link failure. The drawback of expressing more traffic is that the required capacity typically increases; i.e., the expressed traffic is entering fewer IP routers, potentially resulting in less efficiently groomed traffic. This is especially true because in CORONET Phase 1, the project requirements specified that capacity only be added in large increments of 40 Gb/s or 100 Gb/s. This was appropriate at the time the project was started (2006); however, recent research and vendor/carrier studies [13] have made the introduction of an OTN-based aggregation layer between the IP and optical layers potentially very attractive. This option is being intensively examined in CORONET Phase 2. Similarly, other recent research areas such as flexgrid and 400 Gb/s transmission were excluded from the 2006 project requirements and hence were not part of CORONET Phase 1.

With express links, a failure in the optical layer may bring down multiple IP links, potentially leading to more required restoration resources. In addition, utilization of the marginal express link drops as the number of express links increases. This somewhat offsets the reduction in router ports expected from express links.

This tradeoff of router ports versus capacity is illustrated in Table III, which shows the various design options that were considered for the 20 Tb/s aggregate traffic scenario. (The designs include the resources needed to restore the IP traffic; however, for the purposes of generating this table, multiple concurrent failures were not considered.) The third design in the table (shaded), with 73 express links, was selected for the final design, yielding an average virtual nodal degree of 4.2 (versus 2.7 for the average physical nodal degree).

Note that in evaluating costs for the various designs, one router port was taken to be the equivalent of 770 wavelength-km of transport. This number may decrease in the future if the relative cost of router ports decreases. Simulations showed that with the 73 express links, 65% of the IP wavelengths that enter a node bypass the IP router, on average.

A. Fast Connection Setup

There are two broad classes of IP traffic included in CORONET: guaranteed bandwidth (GB) services and best-effort (BE) services. Furthermore, the GB services have restoration requirements that range from being restorable after a single node or link failure to restorable after up to three concurrent failures (i.e., GB-1, GB-2, and GB-3). For the purposes of CORONET, the BE traffic affected by a failure does not need to be restored. In actual networks, some effort is typically made to restore BE traffic.

The CORONET setup time requirement for *fast setup* GB traffic is 2 s, which was achieved with MPLS–TE (Multi-Protocol Label Switching–Traffic Engineering) end-to-end tunnels and RSVP–TE (Resource Reservation Protocol–Traffic Engineering) signaling [14]. (The BE traffic was treated as ever-present, though varying, traffic without explicit connection requests.) For each router pair, up to four IP tunnels were designed, one each for the BE, GB-1, GB-2, and GB-3 services. As connection requests arrive, an RSVP–TE *Path* message is sent along the appropriate tunnel from source to destination to determine whether the additional capacity needed for the connection is available. If the required capacity is not available, a *PathErr* message is sent back to the source. If the required capacity is available from source to destination, then the destination sends a *Resv* message back to the source; the connection is established on this pass. When the destination sends the *Resv* message, it also initiates its own connection setup process, sending a *Path* message to the source, with the source responding with a *Resv* message, if appropriate. (Note that while the RSVP–TE protocol is unidirectional in nature, all connections are assumed to be bi-directional in CORONET Phase 1.) A successful connection setup thus takes of the order of three times the one-way fiber transmission delay from source to destination. In the CORONET network, the longest setup time, as measured by simulation, was approximately 700 ms, easily beating the metric of 2 s.

One of the keys to maintaining efficient use of capacity is that bandwidth is not pre-allocated to any of the tunnels. Rather, the tunnels are “elastic” and allowed to grow and shrink in size in response to the traffic demand, with a single router-to-router link potentially carrying numerous tunnels. Because the wavelength line rate (40 Gb/s to 100 Gb/s) is two to three orders of magnitude greater than the average IP GB connection rate (ranging from tens of Mb/s to 25% of the line rate), a very large number of connections are typically multiplexed onto any wavelength. While some of the individual connections may be bursty, the aggregated traffic displays much smoother traffic statistics, allowing the wavelengths to be more tightly packed.

Multi-rate Erlang analysis was used to determine the number of IP wavelengths required on each link (taking into account the required restoration capacity as well, as described below), given the expected traffic intensity matrix. The target link-blocking rate was taken to be no greater than approximately 10^{-4} (a few iterations were performed to more precisely adjust the individual link-blocking rates). This resulted in an overall connection-blocking rate of less than 10^{-3}, in accordance with the CORONET metric. This metric pertains only to the non-failure condition. CORONET does not specify connection-blocking metrics under failure conditions, but it is reasonable to assume that increased connection blocking would be tolerated. Note that the design of the number of IP wavelengths per link is performed up-front. If the traffic intensity matrix were to change significantly, the design would be modified accordingly.

B. Rapid and Efficient Restoration

Two critical aspects of providing efficient and fast restoration against failures are, first, the methodology for routing the working and restoration paths and, second, the restoration mechanism used in response to a failure.

Restoration in the IP layer is inherently a shared process, where the excess capacity that is deployed for restoration is not dedicated to a particular connection. To minimize the amount of restoration resources needed, it is important to maximize the sharing of those resources. In typical IP networks, shortest path routing, e.g., Open Shortest Path First (OSPF) or Intermediate System to Intermediate System (IS–IS), is used for both working and restoration paths. However, always routing over the shortest available path does not provide the maximum amount of sharing that is achievable. In CORONET, therefore, we specify a capacity optimization approach based on bandwidth-aware MPLS–TE with constrained shortest path first (CSPF) routing.

In order to determine feasible capacities for each link under this optimized MPLS–TE approach, we employ the following methodology. As a first step, shortest path routing is used to determine the capacity needed under the no-failure condition; shortest path routing is favored under no failures as it minimizes latency. Next, additional capacity is added in order to be able to restore the GB-2 and GB-3 services, where again shortest path routing is assumed. Finally, we apply a heuristic to determine feasible capacities for the GB-1 traffic under failure. In this heuristic, we start with the capacities needed for normal traffic and for GB-3 and for GB-2 restoration, rounded up to the next integer wavelength. We then consider each possible single failure scenario in turn. For each scenario, we route each MPLS–TE tunnel. In routing a tunnel, we add a penalty to the OSPF weight of links that have inadequate capacity to carry that tunnel. If a penalized link appears in the shortest path, we increment the number of wavelengths on that link. Once all scenarios have been examined, we repeat the process to determine whether any link capacity can be reduced.

Our optimized MPLS–TE approach provides significant cost and capacity benefits as compared to a design where only OSPF is used, as shown in Table IV. (The table corresponds to the 20 Tb/s aggregate demand scenario. For the purposes of generating this table, multiple concurrent failures were not considered.)

Column three of the table, the SCR, is calculated as follows: if W is the amount of capacity needed in the unprotected scenario, and T is the total amount of capacity needed when restoration is taken into account, then the SCR is $(T - W)/W$. Both W and T are measured in terms of wavelength-km. As can be seen from Table IV, the SCR for the optimized TE design

TABLE IV
OSPF VERSUS OPTIMIZED TRAFFIC ENGINEERING (SINGLE FAILURE CASE)

Design technique	Required capacity in units of 1000 wavelength-km	Spare capacity ratio	Number of line-side router ports	Normalized total cost
OSPF routing for all	2281	0.69	2322	126
Optimized TE	1810	0.36	1839	100

is quite low, of the order of 35%. It should be noted that for IP networks in general, regardless of the design technique, some of the wavelengths in the unrestorable design are not completely full; thus, some degree of restoration is attained "for free" using the above SCR definition.

Another important observation that was made in the design process is that providing restoration against multiple concurrent failures for a subset of the traffic required very little extra resources. This is due to the small percentage of traffic falling in this category (10% GB-2, 2.5% GB-3), and the use of optimized traffic engineering to maximize sharing of the restoration resources. The cost of the IP network increased by roughly 3% and the SCR increased by about 10%, as compared to the case where no service required restoration after more than one failure.

Next, we address the restoration mechanism that was used in response to a failure. A two-phase restoration scheme was developed, 2-Phase-Fast-Reroute (2PFRR), that is both fast and efficient. The first phase of the scheme makes use of the standardized Fast Reroute protocol [15]. Fast reroute is a local rather than an end-to-end restoration scheme, and restoration can be very fast. For example, assume the working path is routed over path *A–B–C–D–E*. If link *BC* fails, the traffic going in the *A* to *E* direction is routed from *B* to *D* over an alternative path. (For the traffic going in the *E* to *A* direction, restoration occurs from *C* to *A*.) Note that this next-next-hop restoration tunnel also protects against a router failure at Node C. For the very last link in the path, next-hop restoration is used.

While Fast Reroute is indeed fast due to the localized action, it is not the most efficient means of restoration. Optimized end-to-end restoration tends to require less capacity. Thus, in the second phase of 2PFRR, more efficient end-to-end restoration paths are calculated for the affected traffic. Using make-before-break, the traffic affected by the failure is then shifted from its initial restoration path to the more efficient path. (There may be a very small probability of packet loss during the transition period, depending on the relative lengths of the restoration paths and how the router timers are configured.) The duration of the first phase of 2PFRR is of the order of 10 to 20 s. During this time, BE traffic may be throttled to provide the extra capacity needed for the initial (inefficient) restoration path. The localized restoration in the first phase of 2PFRR easily beats the CORONET metric for restoration time.

Another means of restoration that is being studied as part of CORONET is reconfiguration of the optical layer in response to a failure in order to restore IP traffic. An efficient algorithm for re-assigning IP link bundles from a failed IP router to another IP router in the same location and efficient transponder sharing during optical-layer failures allow significant reduction in the number of IP router ports and optical-layer transponders. An economic study described in [5] projected cost savings in both the IP and the optical layers. Additional architectural and protocol work is required to realize the savings suggested by this economic "proof of concept."

TABLE V
WFQ SETTINGS TO MEET CORONET IP QOS

Traffic class	WFQ bandwidth allocation	WFQ buffer allocation
Best effort	10%	8.0 ms
Loss sensitive	45%	20.0 ms
Latency/jitter sensitive	45%	1.5 ms

C. Quality of Service

The IP traffic is composed of BE and GB services, with the GB traffic further designated as either loss sensitive or latency/jitter sensitive. CORONET specifies appropriate metrics for each traffic type. The allowable packet loss rate for loss-sensitive GB traffic is 10^{-6}; for latency/jitter-sensitive GB traffic or BE traffic, it is 10^{-3}. The allowable latency for latency/jitter-sensitive GB traffic is 25 ms over the one-way fiber transmission delay from source to destination; for BE traffic, it is twice this; for loss-sensitive traffic, the absolute maximum latency is 500 ms. The allowable jitter is 20% of the allowable latency for all three classes of traffic.

To meet the QoS specifications, DiffServ [16] was used, where each packet is identified with a code to indicate to which class of traffic it belongs. The identifier is used to manage the per-node forwarding behavior (e.g., packet priorities), essentially controlling how the network resources are allocated among the traffic types. Weighted Fair Queueing (WFQ) was implemented at the nodes, with the settings shown in Table V.

One of the most challenging aspects of the QoS portion of the CORONET program was determining via simulation whether the metrics were actually met. Due to the tremendous number of packets flowing over each link, measurements could be taken only over a limited period. To address this, the worst bottleneck link in the network was identified, and a time at which this link reached close to 100% utilization was noted. The loss, latency, and jitter of all packets that traversed this link were tracked over a 200 ms time period centered at this peak utilization time. In all cases, the measured traffic statistics were well within the specified metrics.

V. WAVELENGTH SERVICES

Wavelength services comprise 25% of the aggregate network demand in CORONET. 40% of the wavelength services are assumed to be single-wavelength, 20% double-wavelength,

20% quad-wavelength, and 20% octal-wavelength. These percentages are based on bandwidth, not on the number of connections. (In terms of number of connections, single-wavelength connections are 70% of the total wavelength-service connections.)

A. Rapid Connection Setup

There are two classes of non-scheduled dynamic wavelength services in CORONET. First, there are *fast setup* (FS) services similar to the IP FS services. These require a setup time of less than 2 s. Second, there are VFS services that require a setup time of less than RTD + 50 ms. For the longest connections within CONUS, this requirement is of the order of 100 ms. Furthermore, VFS services are assumed to have a very short holding time, ranging from 1 s to 1 min. In the most aggressive scenarios considered, approximately two VFS connection requests arrive network-wide per second on average. The VFS services are all designated as restorable from just one failure.

A key design decision is whether the connection setup mechanism should be more centralized or decentralized. In a pure centralized scheme, all connection requests are directed to one entity, e.g., a path computation element (PCE) [17]. The PCE would determine the working and restoration routes for all connections and assign the wavelengths and transponders to be used by the connection. This setup information would then be disseminated to the network elements along the path so that the connection can be established.

Because the centralized PCE is aware of all resource usage in the network, resource optimization based on complete network state information is possible when assigning resources. However, one drawback of the centralized approach is that the PCE represents a potential bottleneck, especially with the high rate of connection requests specified by CORONET. A second drawback is that it precludes meeting the setup time requirement for VFS services in a global network. Consider an example where the centralized PCE is located in the United States and we wish to establish a connection between Paris and London. The connection request would need to be sent to the PCE, and the setup instructions would need to be sent back to Paris and London. The total delay would not meet the RTD + 50 ms specification (where RTD is between Paris and London in this example).

In a decentralized approach, the nodes determine the paths and assign the resources, allowing for rapid setup. Historically, however, decentralized mechanisms have suffered from problems with contention, where, for example, two concurrent connection requests may attempt to use the same wavelength on a link. This leads to what is known as "backward blocking," necessitating the setup process of at least one of the connections to be restarted.

To maintain the advantages of centralized and decentralized schemes while avoiding the pitfalls, the Three-Way-Handshake (3WHS) distributed signaling protocol, working in concert with a PCE, was developed [6]. The PCE is responsible for calculating candidate diverse working paths for each source/destination pair. For each candidate path, it calculates a figure-of-merit (FOM) based on the distance of the working path plus the amount of *additional* shared restoration capacity that would be needed to restore that working path. This FOM is state-dependent because the amount of restoration resources that are already reserved continually changes. The FOMs are recalculated every 15 min.

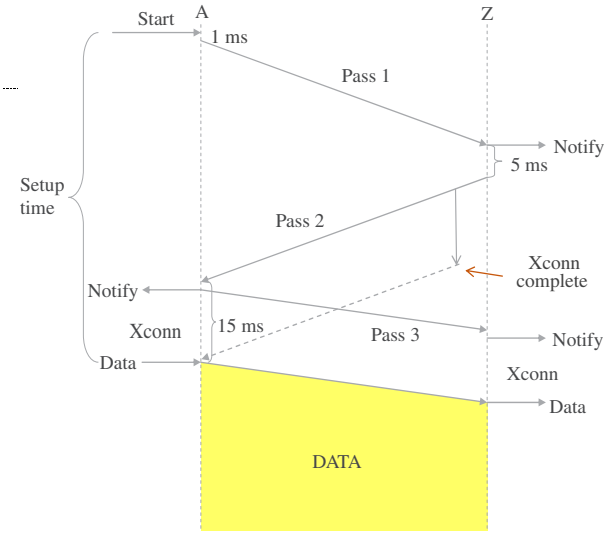

Fig. 3. (Color online) Timing diagram illustrating the 3WHS setup time

To provide near-optimal path choices, the PCE only needs to know the amount of working and restoration resources that are in use on a link or at a node; it does not need to know which connection is using a particular resource. This eliminates the need for time-critical updates to the PCE. The key global state information required is a matrix with one column for each link and one row for each potential failure scenario. An element $A(i,j)$ in this matrix contains the restoration capacity required on link j if scenario i occurs, and the column maximum $\max_i A(i,j)$ is the restoration capacity that must be reserved on link j. The PCE computes the additional shared restoration capacity that would be needed for a candidate service path and diverse restoration path(s) by determining how these column maxima would change. This is straightforward for single failures, but the number of rows is very large (~4.5 million) when double and triple failures are included. This necessitated careful software implementation to ensure acceptable PCE response time.

The actual connection setup, including selecting which resources to use, is performed by the 3WHS signaling protocol, which is a distributed probing mechanism that makes use of the path information from the PCE. The probing is done in real-time so that the actual resources that are available at connection setup time can be determined. The 3WHS mechanism uses a novel methodology to address backward blocking, as detailed below.

We first describe the 3WHS operation as it relates to the setup of VFS services. The 3WHS protocol consists of three passes (a timing diagram illustrating the VFS 3WHS is shown in Fig. 3).

On Pass 1, a signaling message probe is sent along each of the candidate working paths from source (Node A) to

TABLE VI
NUMBER OF EXTRA RESERVED WAVELENGTHS

Connection bandwidth	Number of extra reserved wavelengths
1 wavelength	1
2 wavelengths	1
4 wavelengths	2
8 wavelengths	3

destination (Node Z), collecting information regarding which wavelengths are free on each link and how many transponders are free at each node. Node Z collects the Pass 1 probes,[1] and for each probe it computes the optimal use of resources to meet the connection request (wavelengths to use on each link and transponders required for wavelength conversion and regeneration). It then selects the best path, based primarily on the FOM provided by the PCE and on the number of required transponders. It also takes into account the amount of resources that are available on a path (e.g., it tries to avoid selecting a path that has very few available wavelengths). After selecting the working path, Node Z knows which wavelengths and transponders to use at each link and node on that path. It also knows which "extra" wavelengths and transponders to reserve on Pass 2, as described below. It is assumed that this whole decision process takes of the order of 5 ms.

After Node Z completes its calculations, Pass 2 commences, in which a signaling message is sent from the destination to the source along the selected path, indicating the resources that should be reserved for this connection. In addition to reserving the optimal wavelength(s) as selected by Node Z, Pass 2 also reserves extra wavelengths and transponders, if needed. The number of extra wavelengths is based on the bandwidth of the connection request, as shown in Table VI.

These values were determined based on simulations (but note that the protocol is flexible and can use alternative values).[2] If a link is encountered during Pass 2 where the originally designated wavelength (which had been free during Pass 1) has been grabbed by a competing connection request, the protocol will use one of the extra wavelengths. Simulations showed that reserving a small number of extra wavelengths greatly diminishes backward blocking. The tradeoff is that extra resources are reserved for a very short timeframe. However, this results in the equivalent of about a 1% increase in network load.

As Pass 2 progresses, each node along the path starts the cross-connection process to establish the required connections as soon as it receives the information from the Pass 2 message. The Pass 2 message does not wait for the cross-connections to be made; rather, it continues on toward Node A. When Node A receives the Pass 2 message, it knows which wavelength connections have been successful and makes the final decision as to which resources will be used. Node A then initiates its cross-connect(s) to the client port(s), sends a notification to its client that cross-connects to its ports are being made, and sends a Pass 3 message back to the destination to free the extra reserved resources that are no longer needed. For multi-wavelength connections, the Pass 3 message also informs Node Z of the order in which the client at Node A is using the wavelengths. On receiving the Pass 3 message, Node Z sends a notification to its client indicating that the connection setup is successful, that cross-connects to its ports are in progress, and the order in which the ports should be used in a multi-wavelength connection.

The connection is established at the end of Pass 2 after the cross-connection to the client ports occurs at Node A, and transmission from the client can begin. The client can determine when to begin transmission by using a 15 ms timer after receiving the Notify message after Pass 2. An option would be to have Node A also send a notification to the client after the cross-connects to the client ports are completed. This would provide confirmation to the client that the desired connection has in fact been established. All node cross-connects along the path toward the destination node will be in place when the client signal reaches the node.

As indicated earlier, the setup time requirement for the VFS service is less than RTD + 50 ms. To analyze what is required of the node signaling processing time, α, to meet that requirement, let N be the number of nodes in an A–Z path. From Fig. 3 it is seen that the setup time requirement is

$$21\ \text{ms} + \text{RTD} + 2\alpha(N-2) < \text{RTD} + 50\ \text{ms}$$

or

$$2\alpha(N-2) < 29\ \text{ms}.$$

From our studies of the CORONET topology (Fig. 1), the maximum value for N is 25. This implies that the requirement on signaling processing time is $\alpha < 630$ μs. This is an achievable processing time with today's processing technology.

As indicated earlier, operating a network with a single PCE would be unsatisfactory from a reliability perspective, as well as the need to limit switch-to-PCE signaling delays. In our simulation studies, we used 6 PCEs for the global CORONET network (4 in CONUS, 1 in Europe, and 1 in Asia). Each AOS sends its call setup/takedown messages to its nearest working PCE. It also periodically sends a heartbeat message (e.g., 1/s) to all the PCEs, and each PCE responds to the AOS heartbeat messages. Based on this heartbeat signaling, each AOS can identify which PCEs are "alive" and which is closest in terms of delay.

Each PCE sends call records to all other PCEs when connections are established, and sends notifications when calls terminate. Thus, all PCEs maintain "current" records for all calls in the network, with slight differences in time synchronization. There is no need to maintain precise synchronization among the PCEs because resource allocation is performed by the 3WHS and not the PCEs. AOS failover procedures are defined for detecting PCE failure and switching over to another PCE. PCE recovery procedures are defined to establish a current database before the PCE goes online.

GMPLS [2] is an IETF (Internet Engineering Task Force) standardized distributed connection setup protocol that could be extended to include all the 3WHS capabilities. The 3WHS protocol differs from GMPLS in four principal ways. First, GMPLS selects the resources to use on a segment-by-segment

[1] When the first probe for the call arrives at Node Z, the indicated client for the connection is notified that a connection setup is in progress, which client initiated it, how many wavelengths are requested, etc.

[2] For example, if a particular region is congested, additional "extra wavelengths" can be reserved for paths going through the congested area.

basis, whereas in the 3WHS, the destination selects the resources to use based on all the collected end-to-end path information, thus allowing for an end-to-end optimal selection. One effect of the limited horizon of the current GMPLS definition is that it can result in significantly more wavelength conversion than does the 3WHS, resulting in more required transponders. Second, GMPLS considers just one path; 3WHS probes multiple paths. Third, GMPLS has no standardized way of alleviating backward blocking; 3WHS temporarily reserves extra wavelengths to address this issue, and the number of extra wavelengths can be based on the number of wavelengths in the connection and other factors (e.g., focused overloads). Finally, the current GMPLS standard only establishes single-wavelength connections; 3WHS can set up a connection composed of an arbitrary number of wavelengths.

As described above, 3WHS is used to set up the working path for a VFS connection. After the connection is established, the source then informs the PCE of the path that was selected and requests that an appropriate restoration path be calculated. While it is possible that the PCE cannot find a suitable restoration path, simulations showed that this occurred with less than 0.1% of the VFS connections. In this case, if a failure were to bring down the connection, a new connection request would need to be issued at that time. (It is possible to modify the 3WHS probing mechanism such that both working and restoration paths are found; however, this was not implemented in the simulations.)

Extensive simulations showed that using 3WHS enabled the CORONET setup time metric of RTD + 50 ms to be met for all VFS services. Note that when the PCE determines the candidate paths to probe, excessively long paths are not considered so as not to delay the connection setup process. In addition, the VFS connection-blocking rate was less than 10^{-3}, satisfying the CORONET metric.

The setup process is modified for FS services to take advantage of the more relaxed setup time requirement of 2 s. For FS services that are restorable from a single failure, 3WHS is used as described for VFS services to find a working path (optionally, fewer extra wavelengths can be reserved on Pass 2). As Pass 3 starts, the source informs the PCE of the chosen working path, and requests a restoration path. If a restoration path is found, transmission from the source can begin. If a restoration path cannot be found (or if the 3WHS process itself was not successful) then the PCE attempts to calculate *both* working and restoration paths. The source is informed of the newly calculated working path, and 3WHS is used to establish it. (Any reservations or cross-connections made during the first invocation of 3WHS are undone.)

For FS services requiring restoration from two or three concurrent failures, the FS connection request is immediately sent to the PCE. The PCE communicates to the source the specific working path to use, as well as the associated restoration paths. The source then initiates 3WHS over the selected working path.

Simulations showed that the maximum FS service setup time was roughly 1.5 s, satisfying the CORONET metric of 2 s. Furthermore, the FS connection-blocking rate was less than 10^{-3}, as required by CORONET.

B. Rapid and Efficient Restoration

To provide capacity efficient restoration in the optical layer, it is necessary to implement some type of shared mesh restoration. Furthermore, optical-layer fault isolation can be slow in networks with optical bypass; thus, the rapid restoration requirements of CORONET favor the use of path-based (i.e., end-to-end) restoration. In path-based restoration, the connection end-points are notified that a fault has occurred (e.g., by a loss of signal (LOS) or an alarm indication signal (AIS) in the transport system overhead), and the end-points signal the setup of the restoration path. Faults can be bi-directional or unidirectional. When unidirectional faults occur, only one connection end-point receives an AIS, and that node must signal the setup of the restoration path. When a bi-directional failure occurs, both connection end-points receive an AIS, and there are various options of how to signal the restoration path setup. In the *Robust Optical-Layer End-to-End X-Connection* (ROLEX) restoration signaling methodology that was used (described below), the restoration path setup is signaled from both ends to achieve the fastest possible restoration time.

In shared mesh restoration, link capacity reserved for restoration is shared across a number of diverse working paths. In the CORONET restoration methodology, the PCE determines the best restoration path (e.g., a path that maximizes shared resources), and it identifies how many wavelengths need to be reserved for restoration on each link of the restoration path. The capacity reserved for restoration on a link is measured as a certain number of *R-channels*. Prior to a failure, an R-channel represents capacity on a link that has been reserved for restoration; it does not represent a specific reserved wavelength. As described below, the wavelength used is determined when the restoration path is set up. The PCE and the 3WHS use the number of R-channels assigned to a link to ensure that the combined number of working wavelengths and R-channels does not exceed the link's wavelength capacity.

In the initial incarnations of such schemes, the R-channels are "pre-lit," i.e., there is a transponder located at either end-point of the R-channel and the transponder is always in the ON state [18,19]. This eliminates the need to turn the transponder ON at the time of failure. This is important in networks with optical bypass as rapid power excursions on a link could result in widespread optical transients. However, it is anticipated that in the timeframe that CORONET technology would be deployed, optical-transient control will have progressed sufficiently to eliminate the need to pre-light the R-channels (for example, see [20]). This affords the opportunity to share transponders among R-channels.

With dedicated transponders, the concatenation of R-channels always occurs in the electrical domain. However, with R-channels not demarcated by transponders it is possible to concatenate them all-optically, assuming the rules for regeneration are not violated. Simulations in CORONET showed that roughly 60% of the R-channel cross-connections occurred all-optically.

Given the R-channel-based mechanism, two important accompanying functions are 1) determining how many R-channels are needed for any given traffic set, and 2) concatenating the proper R-channels at the time of failure.

When setting up connections for VFS services and most restorable-1 FS services, the working path is determined by the 3WHS and the restoration path is selected by the PCE. For the restorable-2 and restorable-3 services, the PCE simultaneously calculates both the working and restoration paths. In all cases, the PCE selects the restoration path in order to maximize the sharing of existing reserved restoration resources. Rather than recalculating restoration for the entire optical layer with each new connection arrival, the PCE determines the *incremental* number of R-channels needed on each link of the restoration path required to restore the new service. It informs the source node of the restoration path that has been selected and how many additional R-channels need to be reserved on each link of the restoration path; the source node then sends a signaling message along the restoration path to reserve the required capacity. This signaling message also informs the destination node that a restoration path has been reserved and what links are in the restoration path. The destination node then sends the source node an acknowledgment that the restoration path has been successfully reserved.

Additionally, the PCE checks approximately every minute to see if any R-channels are no longer needed due to connections having been taken down. It then sends a message to each node indicating how many R-channel resources can be released for each of its links.

To track the restoration capacity needed for the wavelength services that require restoration from double or triple failures (10% and 2.5% of the wavelength services, respectively), efficient matrices are maintained so that all combinations of failures are considered (in the case of triple failures, there are over one million combinations). Providing restoration for these services increased the amount of required restoration capacity by about 7%, compared to the case where these services only require restoration against one failure. However, note that the CORONET program assumes that all traffic originating or terminating at a failed node is lost and cannot be restored.

The above process simply reserves restoration capacity. To actually establish the restoration paths at the time of failure, the ROLEX restoration signaling mechanism was extended to work with R-channels [7,21]. To illustrate how the extended ROLEX restoration signaling works, assume that a bi-directional link failure occurs (i.e., the optical fiber in each direction is cut), and consider a restorable wavelength-service connection that is routed over that failed link. Each end of the connection will receive a LOS or AIS indication informing it that the connection has failed. Each end then initiates the restoration signaling process (i.e., two-ended ROLEX is implemented).

The restoration process begins with each end redirecting its transponder(s) from the working path to the restoration path and initiating the ROLEX signaling along the restoration path. This saves two transponders per connection wavelength because it is not necessary to provision additional end-point transponders for restoration. Each end's ROLEX signaling message travels link-by-link along the restoration path, and the two meet somewhere in the middle. It is desired to minimize the use of transponders, so ROLEX will attempt at each node to cross-connect to the same wavelength on the outgoing link that was used on the previous link. The wavelengths used for restoration are hunted from the upper end of the spectrum, and working paths hunt from the lower end of the spectrum. This minimizes contention and increases the probability of having all-optical cross-connections between R-channels.

From a carrier's perspective, it is preferable to use the same wavelength in each direction on a link for a single connection. Thus, to coordinate the choice of wavelengths on a link, one end is designated as the master and the other end as the slave. The master decides which wavelength to use and it informs the slave of its decision.[3] In ROLEX restoration path signaling, there will usually be a ROLEX signaling message coming from each end of the connection, and they will meet at some intermediate link. The master/slave protocol is used to resolve conflicts between the two signaling directions.

Simulations showed the R-channel-based mechanism combined with two-ended ROLEX to be both fast and capacity efficient. The spare capacity ratio (SCR) for the CONUS wavelength services was of the order of 90% to 100%, depending on the demand scenario. When combined with the approximately 40% SCR for CONUS IP services, the overall SCR for the CONUS portion of the network was 50% to 55%, well below the CORONET metric of 75%. Furthermore, simulations showed that restoration could be completed within RTD + 50 ms, as specified by the CORONET metric. This time was met regardless of the number of concurrent failures.

Note that in Phase 1, CORONET required that all wavelengths comprising a multi-wavelength service be routed over the same path for both the working and restoration paths. In Phase 2, this requirement is relaxed to allow the constituent wavelengths to be routed over different paths (e.g., using virtual concatenation). This flexibility is expected to increase the opportunity for sharing restoration resources and reduce the SCR.

VI. Scheduled Services

In addition to the non-scheduled dynamic IP and wavelength services described above, CORONET also includes scheduled IP and wavelength services, where it is assumed that the service request is received by the network at least ten hours prior to the service actually being needed. Network resources are not partitioned between scheduled and non-scheduled connections; however, the scheduled services can use at most 75% of IP link resources (80% of the wavelength link resources), so non-scheduled services are not "starved." At the time a scheduled service request arrives, the network management system decides whether the necessary resources will be available at the scheduled start time, considering the other connections that have already been scheduled. If resources will not be available, the new connection request is blocked. If resources are expected to be available, the connection is accepted, but no actual resources are reserved at that time.

For a scheduled IP service, a B-hour blackout time is implemented when no new IP connections can use the resources that will be needed for the scheduled service for the B

[3] In the case of transoceanic links, the propagation delays are too long to support a master/slave negotiation, so in this case each end of the link determines its direction's wavelength.

hours prior to the scheduled start time. The maximum holding time of IP GB FS connections is ten hours; thus, if the blackout time were ten hours, it would guarantee that the scheduled service would find its required resources free at its target start time. However, a blackout period of this length would result in excessive blocking of new GB connections. Through simulation, a blackout time of two hours was selected. There is a chance that the necessary resources will not be available for the scheduled connection at its target start time; however, this occurred with very low probability. In these cases, the scheduled connection was blocked.

Most of the dynamic wavelength connections fall in the category of VFS, where the holding time is less than a minute. Thus, it was not necessary to establish a blackout period. If the necessary resources are not available at the start time of a scheduled service, then FS and/or VFS services are pre-empted to free up resources. Again, this occurred with very low frequency. The 3WHS protocol is used to set up all scheduled wavelength connections. The CORONET setup time metric for scheduled services is 10 s, which was easily met.

For both IP and wavelength scheduled services, the connection-blocking rate was below 10^{-4}, as required by CORONET.

Additional studies looked at more flexible scheduled services in which acceptable *ranges* were specified for the start time and required bandwidth. This increased flexibility resulted in both lower blocking and increased network utilization.

VII. Traffic Generation

CORONET traffic generation was discussed previously in reference [3]. Here we give additional details about the statistical properties of the generated traffic streams.

Our traffic generator offers four statistical traffic models: TEDB, TE_FULL, UNIFORM, and BIMODAL.

TEDB features a truncated exponential holding time with discrete bandwidth. The holding time CDF (cumulative distribution function) we use for our truncated exponential distribution is

$$\text{Holding time CDF} = \begin{cases} 0, & t < T_{\min}, \\ \left[1 - \exp\left(-\dfrac{a\left(t - T_{\min}\right)}{T_{\max} - T_{\min}}\right)\right] & \\ \quad /[1 - \exp(-a)], & T_{\min} \le t \le T_{\max}, \\ 1, & t > T_{\max}, \end{cases}$$

where $T_{\min}$ and $T_{\max}$ are the minimum and maximum holding times for the traffic class, respectively. The value of the heuristic parameter a was chosen to be 4, which skews the distribution toward shorter hold times ($T_{\text{ave}} < \frac{T_{\min}+T_{\max}}{2}$) while still leaving a significant fraction of the hold times with longer duration. With $a = 4$, the average truncated exponential holding time can be derived as $t_{\text{ave}} = T_{\min} + 0.231 \times \left(T_{\max} - T_{\min}\right)$. The values of $T_{\min}$ and $T_{\max}$ used in our simulations are shown in Table VII.

For the fine, medium, and coarse granularity IP services, there are three discrete bandwidth values used for TEDB that are calculated from the minimum bandwidth $b_{\min}$ and maximum bandwidth $b_{\max}$ as follows:

$$b_1 = b_{\min} + (b_{\max} - b_{\min})/3,$$
$$b_2 = b_{\min} + 2(b_{\max} - b_{\min})/3,$$
$$b_3 = b_{\max}.$$

These three bandwidth values are assigned probabilities 0.6, 0.3, and 0.1, respectively. With these values, the average bandwidth works out to be $b_{\text{ave}} = \frac{b_{\min}+b_{\max}}{2}$. Our IP network design was based on the TEDB statistical model. Optical services, on the other hand, have only a single discrete bandwidth nR, where n is a positive integer and R is the bit rate of an individual wavelength; that is, $b_{\min} = b_{\max} = nR$ for an optical service. The values of $b_{\min}$ and $b_{\max}$ used in our IP design and simulations are shown in Table VIII.

Regarding the other statistical models, TE_FULL uses the truncated exponential distribution for both the hold time and the bandwidth distribution (note that for optical services, TE_FULL is identical to TEDB). UNIFORM uses a uniform distribution between the required minimum and maximum values for both the holding time and the IP bandwidth. BIMODAL uses a discrete bimodal model for both holding time and IP bandwidth with probability 1/2 given to both the minimum and maximum values.

As for other options and capabilities, the demand generator has two modes for the semi-permanent demands. In one mode, a random initial pattern of semi-permanent demands is generated at start-up and is held constant throughout the simulation. In the second mode, semi-permanent demands are allowed to churn during the simulation as other demands do, albeit more slowly, according to the holding times shown in Table VII.

TABLE VII
Holding Times for the Different CORONET Services

Holding time	Very fast	Fast	Scheduled	Semi-permanent
$T_{\min}$	1 s	10 s	1 min	1 month
$T_{\max}$	1 min	10 h	1 month	6 months

TABLE VIII
Bandwidths for the Different CORONET Services

Bandwidth	Fine granularity	Medium	Coarse
$b_{\min}$	1 b/s	400 Mb/s	4 Gb/s
$b_{\max}$	400 Mb/s	4 Gb/s	$R/4$

VIII. Nodal Architecture

The baseline nodal architecture for CORONET Phase 1 assumed that an IP router was in the same location as every AOS. However, note that this is not a CORONET requirement. In an actual deployment, backbone IP routers may be located in only a subset of the nodes. To efficiently support the highly dynamic traffic, the AOS must be both colorless and steerable (steerable is also referred to as directionless or non-directional in the literature [22,23]). These two features together allow

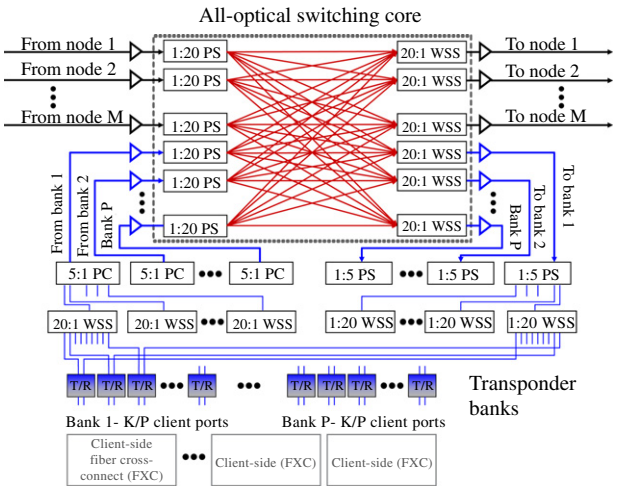

Fig. 4. (Color online) WSS-based all-optical-switch architecture.

any transponder on an AOS to add/drop to/from any network port on any wavelength. These features are typically not supported in currently deployed ROADMs.

One of the features of CORONET was the design of a colorless/steerable transponder-bank architecture using a combination of passive couplers/splitters and WSSs (Wavelength Selective Switches), as illustrated in the lower portion of Fig. 4. The design is both cost effective and scalable; moderate sized WSSs (e.g., 1×20 WSSs) can be added as the amount of add/drop at a node increases. The design optimizes modular cost scaling and minimizes technology risk by relying on components with a well-established track record in commercial deployment. More details of this AOS architecture can be found in [8].

Switching time also plays an important role in achieving rapid connection setup. It was determined that switch timing of roughly 15 ms is needed to meet the CORONET setup time metrics. In the past, vendors have not focused on the switching time of WSSs as carriers have not required rapid reconfiguration; however, discussions with vendors indicate that a switch time of 15 ms is feasible.

Configurability on the client side of the AOS is also important. A fiber cross-connect (FXC) is deployed on the client side of the AOS to 1) allow a connection to be restored in the event of a transponder failure, 2) enable two transponders to be connected together to provide regeneration and wavelength conversion, and 3) facilitate interconnection of terrestrial and submarine links. An FXC also allows the pooling of transponder resources to efficiently meet the add/drop and regeneration needs for both service and restoration. Client-side cross-connects have been shown to mitigate the potentially low transponder utilization caused by contention among add/drop wavelengths. With an equal number of transponder banks to inter-node fiber pairs and a client-side cross-connect, intra-node contention is negligible in this node architecture [24,25]. FXCs with a 10 ms switching time are commercially available [26]. If reliability or scaling objectives drive a modular construction of the FXC, the benefits of transponder sharing could be maintained by module interconnection strategies such as Clos architecture or by constraint-aware transponder management policies.

IX. Sizing Transponder Pools

Transponder costs dominate the cost of the optical layer. CORONET network studies showed that transponders account for roughly 65% of the optical-layer cost, with AOSs, amplifiers, and fiber accounting for the remainder. The transponders are needed at connection end-points to connect client ports to the long haul DWDM system and for regeneration or wavelength conversion. The 3WHS protocol is effective in reducing the amount of transponders needed for these latter two functions along working paths; roughly 50% of the wavelengths that enter a node remain in the optical domain.

In a network with dynamic traffic, pre-deploying transponders is a necessity as there is clearly insufficient time for a truck-roll to deploy the needed equipment. If equipment is not available at the time of connection setup, the connection is blocked. It is important to minimize the number of pre-deployed transponders while still satisfying the connection-blocking requirements. A novel methodology for sizing transponder pools was developed in CORONET, and it is described here with respect to the transponders used for working traffic.

Transponder pools are obviously required at the 40 nodes that support wavelength-service traffic. It is also desirable to provide transponder pools at other nodes so that regeneration and wavelength conversion capability will always be "nearby" when needed. Routing studies were carried out to identify which additional nodes would need transponder pools so that regeneration would always be possible. These studies identified 13 additional nodes, for a total of 53 nodes that would support transponder pools. Initially, simulations were run in which each of the 53 nodes had an unlimited transponder pool. Every 30 min of "network time," the number of transponders actually in use at a node was recorded, and roughly 2500 sample points were taken during the simulations. The generated histogram data for each of these nodes are found to closely follow a chi-squared distribution.[4] For the nodes that source wavelength services, the distributions are best modeled by a chi-squared distribution with from 1 to 15 degrees of freedom; for the other nodes, a chi-squared distribution with one degree of freedom best fits the data. Using these distributions, the transponder pools are sized at the point on the distribution at which its tail has an area of 10^{-4}. This point on the distribution is easily determined from the histogram data, which provides the mean, standard deviation and degrees of freedom of the node's distribution. The pool size is $(\mu + \alpha\sigma)$, where μ and σ are the measured (histogram) mean and standard deviation, and α depends only on the chi-squared distribution degrees of freedom and the desired tail area. Note that if the node histograms do not conform to a known distribution family, extensive simulations are required to obtain a good characterization of the distribution to estimate where a 10^{-4} tail begins.

[4] The "chi-squared distribution" is a one-parameter family of distributions; the parameter is a non-zero integer called the "degrees of freedom."

With the above methodology, one can determine the number of transponders that must be deployed at each node to reduce per-nodal *transponder blocking* below 10^{-4}. Simulations were rerun with the transponder pools sized according to this methodology. As expected, the *connection-blocking* rate increased by about 5×10^{-4}; i.e., this is the additional fraction of dynamic wavelength-service connections that are blocked due to the required transponders not being available at one or more nodes. Using this strategy reduces the number of transponders by 25% to 35%, depending on the aggregate demand level, as compared to the high-water-mark of transponders used at each node when unlimited transponder pools are assumed. More details of this approach to sizing transponder pools can be found in [9].

Similar strategies can be applied to the transponders needed for restoration, although in future work there is likely to be a single shared transponder pool used for both working and restoration. The above methodology can be used to determine the proper statistical distributions. Hybrid scenarios with both 40 Gb/s and 100 Gb/s wavelengths were not considered in Phase 1. Hybrid line rates would necessitate separate 40 Gb/s and 100 Gb/s transponder pools, which would be more difficult to manage and would probably reduce overall transponder utilization.

X. Security

The provision of a secure networking environment is a growing necessity and, in fact, mandatory for government networks. Rather than perform a comprehensive study of all network security mechanisms, CORONET emphasized the security impact of new features, namely, the highly dynamic connection setup. Strategies were investigated to secure the path information provided by the PCEs and to validate the 3WHS signaling without violating the stringent setup time requirements. For example, a Hash Message Authentication Code (HMAC) can be used to protect 3WHS signaling, where each end of a link has a shared private key. The sending end of the link performs a cryptographic hash of the signaling message and encrypts the hash using the private key. The receiving end does the same hash of the message, decrypts the encrypted hash in the received message, and compares the two hashes to authenticate the sender (they have the same private key) and validate the integrity of the message (e.g., there were no replay or data modification attacks). Using standardized cryptographic algorithms, the authentication processing time per link is expected to be a few microseconds.

One advantage of the distributed 3WHS protocol is that it allows the initiator of a connection (Client A) to verify that the established connection goes to the intended destination (Client Z). As part of the Pass 2 3WHS procedures, the destination node sends a connection notification to Client Z, and in response Client Z sends a verification message on the appropriate port(s).[5] As the Pass 2 AOS cross-connects are completed, the verification message is transmitted to the next node toward the source. Thus, when the source node completes its cross-connects to Client A, it can verify that the connection has been set up to the desired Client Z. This verification process adds negligible delay to the setup time. The fact that multiple cross-connects are initially set up at a node due to the extra wavelengths being reserved on Pass 2 can be accommodated by the multicast feature of the AOS at the destination node.

[5] The verification message could be a repeated message (ACK signal) along with an HMAC. When Client A receives the ACK it can verify the HMAC with a shared private or public key. The verification message could be sent in the OTN (G.709) General Communications Channel in the OTU overhead (i.e., GCC0).

XI. Phase 2 Extensions

Many extensions to the architecture and protocols will be studied in Phase 2 of the program. While the focus of Phase 1 was an IP-over-optical layer model, Phase 2 will consider the benefit of a subwavelength layer below the IP layer, e.g., optical transport network (OTN). The rationale for this architecture is that much of the grooming of subwavelength traffic can be offloaded from the IP router to the subwavelength grooming switch, and such switches tend to be more cost effective and more reliable than routers. The efficacy of this approach is still under study.

In terms of protocols, the major development will be extending 3WHS to multi-domain networks, both single-carrier multi-vendor domains and multiple-carrier domains. The impact on connection setup time needs to be determined. The user/network interface will also be addressed in Phase 2, where customer service requests are mapped to the appropriate CORONET services.

The overriding goal of Phase 2 is to transition the developed technology to commercial and government networks. Thus, implementation of the novel protocols of the program on a global scale, 100 node emulation testbed is the centerpiece of the next phase. Additionally, other smaller testbeds will be developed to demonstrate realistic applications.

XII. Conclusion

While the CORONET program specifies aggressive requirements regarding dynamic services and network resiliency, all the program metrics were successfully met in the first phase. Solutions were found that were capacity efficient, scalable, and provided good QoS. Numerous architectural and protocol advancements were developed as part of the program. Transitioning the technology to commercial and government networks in the next few years is the goal of the next phase of the program.

Acknowledgments

The Telcordia and AT&T authors appreciate the support of the DARPA CORONET Program, Contract N00173-08-C-2011, and the U.S. Army RDE Contracting Center, Adelphi Contracting Division, 2800 Powder Mill Rd., Adelphi, MD, under contract W911QX-10-C00094. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for Public Release, Distribution Unlimited.

REFERENCES

[1] A. A. M. Saleh, "Dynamic multi-terabit core optical networks: architecture, protocols, control and management (CORONET)," *DARPA BAA 06-29, Proposer Information Pamphlet*.

[2] E. Mannie, Ed., "Generalized multi-protocol label switching (GMPLS) architecture," *RFC 3945*, Oct. 2004.

[3] A. Chiu, G. Choudhury, G. Clapp, R. Doverspike, J. W. Gannett, J. G. Klincewicz, G. Li, R. A. Skoog, J. Strand, A. Von Lehmen, and D. Xu, "Network design and architectures for highly dynamic next-generation IP-over-optical long distance networks," *J. Lightwave Technol.*, vol. 27, no. 12, pp. 1878–1890, June 2009.

[4] G. Clapp, R. Doverspike, R. A. Skoog, J. Strand, and A. C. Von Lehmen, "Lessons learned from CORONET," in *OFC/NFOEC 2010*, San Diego, CA, Mar. 21–25, 2010, OWH3.

[5] A. Chiu, G. Choudhury, M. Feuer, J. Strand, and S. Woodward, "Integrated restoration for next-generation IP-over-optical networks," *J. Lightwave Technol.*, vol. 29, no. 6, pp. 916–924, 2011.

[6] R. A. Skoog and A. L. Neidhardt, "A fast, robust signaling protocol for enabling highly dynamic optical networks," in *OFC/NFOEC 2009*, San Diego, CA, Mar. 22–26, 2009, NTuB5.

[7] A. Chiu, R. Doverspike, G. Li, and J. Strand, "Restoration signaling protocol design for next-generation optical network," in *OFC/NFOEC 2009*, San Diego, CA, Mar. 22-26, 2009, NTuC2.

[8] S. L. Woodward, M. D. Feuer, J. Jackel, and A. Agarwal, "Massively scaleable node design for a highly-dynamic core network," in *OFC/NFOEC 2010*, San Diego, CA, Mar. 21–25, 2010, JThA18.

[9] R. A. Skoog and B. J. Wilson, "Transponder pool sizing in highly dynamic translucent WDM optical networks," in *OFC/NFOEC 2010*, San Diego, CA, Mar. 21–25, 2010, NTuA3.

[10] G. L. Choudhury and J. G. Klincewicz, "Survivable IP link topology design in an IP-over-WDM architecture," in *7th Int. Workshop on the Design of Reliable Communication Networks (DRCN 2009)*, Washington, DC, Oct. 25–28, 2009.

[11] Sample Optical Network Topology Files available at http://www.monarchna.com/topology.html.

[12] "AT&T Optical Mesh Service – OMS," *AT&T Product Brief*, May 13, 2008 [Online]. Available: www.business.att.com/content/productbrochures/PB-OMS_16312_V01_05-13.pdf.

[13] S. Gringeri, B. Basch, V. Shukla, and R. Egorov, "Flexible architectures for optical transport nodes and networks," *IEEE Commun. Mag.*, vol. 48, no. 7, pp. 40–50, July 2010.

[14] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP tunnels," *IETF RFC 3209*, Dec. 2001.

[15] P. Pan, G. Swallow, and A. Atlas, Eds., "Fast reroute extensions to RSVP-TE for LSP tunnels," *IETF RFC 4090*, May 2005.

[16] K. Nichols, S. Blake, F. Baker, and D. Black, "Definition of the differentiated services field (DS Field) in the IPv4 and IPv6 headers," *IETF RFC 2474*, Dec. 1998.

[17] A. Farrel, J. P. Vasseur, and J. Ash, "A path computation element (PCE)-based architecture," *IETF RFC 4655*, Aug. 2006.

[18] G. Li, A. L. Chiu, and J. Strand, "Failure recovery in all-optical ULH networks," in *5th Int. Workshop on Design of Reliable Communication Networks (DRCN'05)*, Island of Ischia, Italy, Oct. 16–19, 2005.

[19] J. M. Simmons, "Cost vs. capacity tradeoff with shared mesh protection in optical-bypass-enabled backbone networks," in *OFC/NFOEC 2007*, Anaheim, CA, Mar. 25–29, 2007, NThC2.

[20] X. Zhou, M. Feuer, and M. Birk, "Fast control of inter-channel SRS and residual EDFA transients using a multiple-wavelength forward-pumped discrete Raman amplifier," in *OFC/NFOEC 2007*, Anaheim, CA, Mar. 25–29, 2007, OMN4.

[21] R. Doverspike, G. Sahin, J. Strand, and R. Tkach, "Fast restoration in a mesh network of optical cross-connects," in *OFC 1999*, San Diego, CA, Feb. 21–26, 1999.

[22] S. L. Woodward, M. D. Feuer, J. Calvitti, K. Falta, and J. M. Verdiell, "A high-degree photonic cross-connect for transparent networking, flexible provisioning & capacity growth," in *Proc. European Conf. Optical Communications*, 2006, Th1.2.2.

[23] M. D. Feuer, D. C. Kilper, and S. L. Woodward, "ROADMs and their system applications," in *Optical Fiber Telecommunications, volume B: Systems and Networks*. Elsevier Inc., London, UK, 2008.

[24] S. L. Woodward, M. D. Feuer, P. Palacharla, X. Wang, I. Kim, and D. Bihon, "Intra-node contention in a dynamic, colorless, non-directional ROADM," in *OFC/NFOEC 2010*, San Diego, CA, Mar. 21–25, 2010, PDPC8.

[25] M. D. Feuer, S. L. Woodward, P. Palacharla, X. Wang, I. Kim, and D. Bihon, "Intra-node contention in dynamic photonic networks," *J. Lightwave Technol.*, vol. 29, pp. 529–535, 2011.

[26] X. J. Zhang, M. Birk, A. Chiu, R. Doverspike, M. D. Feuer, P. Magill, E. Mavrogiorgis, J. Pastor, S. L. Woodward, and J. Yates, "Bridge-and-roll demonstration in GRIPhoN (globally reconfigurable intelligent photonic network)," in *OFC/NFOEC 2010*, San Diego, CA, Mar. 21–25, 2010, NThA1.