2014

Balancing Cost and Reliability in the Design of Internet Protocol Backbone Using Agile Optical Networking

Byrav Ramamurthy
University of Nebraska-Lincoln, bramamurthy2@unl.edu

Rakesh K. Sinha
AT&T Labs-Research, Middletown, NJ

K. K. Ramakrishnan
Rutgers University

Ramamurthy, Byrav; Sinha, Rakesh K.; and Ramakrishnan, K. K., "Balancing Cost and Reliability in the Design of Internet Protocol Backbone Using Agile Optical Networking" (2014). *CSE Journal Articles*. 122.
http://digitalcommons.unl.edu/csearticles/122

Pages: 427 - 442, DOI: 10.1109/TR.2014.2315915

Balancing Cost and Reliability in the Design of Internet Protocol Backbone Using Agile Optical Networking

Byrav Ramamurthy, *Member, IEEE*, Rakesh K. Sinha, *Member, IEEE*, and K. K. Ramakrishnan, *Fellow, IEEE*

***Abstract*—To address reliability challenges due to failures and planned outages, Internet Service Providers (ISPs) typically use two backbone routers (BRs) at each central office. Access routers (ARs) are connected to these BRs in a dual-homed configuration. To provide reliability through node and path diversity, redundant backbone routers and redundant transport equipment to interconnect them are deployed. However, deploying such redundant resources increases the overall cost of the network. Hence, to avoid such redundant resources, a fundamental redesign of the backbone network leveraging the capabilities of an agile optical transport network is highly desired. In this paper, we propose a fundamental redesign of IP backbones. Our alternative design uses only a single router at each office. To survive failures or outages of a single local BR, we leverage the agile optical transport layer to carry traffic to remote BRs. Optimal mapping of local ARs to remote BRs is determined by solving an Integer Linear Program (ILP). We describe how our proposed design can be realized using current optical transport technology. We evaluate network designs for *cost* and *performability*, the latter being a metric combining performance and availability. We show significant reduction in cost for approximately the same level of reliability as current designs.**

***Index Terms*—Backbone networks, Internet Protocol over Wavelength Division Multiplexing, multi-layer architecture, network design, performability.**

Abbreviations and Acronyms

AR	Access Router
BR	Backbone Router
BER	Bit-error rate
CapEx	Capital Expenditure
DWDM	Dense Wavelength Division Multiplexing
EDFA	Erbium-Doped Fiber Amplifier
FRR	Fast-Reroute
FXC	Fiber Crossconnect
HSRP	Hot Standby Router Protocol
IGP	Interior Gateway Protocol
ILP	Integer Linear Program
IP	Internet Protocol
ISP	Internet Service Provider
IST	Information Society Technologies
LAG	Link Aggregation Group
LSA	Link State Advertisement
MPLS	Multi-Protocol Label Switching
NOBEL	Next Generation Optical Networks for Broadband European Leadership
O/E/O	optical-to-electronic-to-optical
OpEx	Operational Expenditure
OSPF	Open Shortest Path First
OT	Optical Transponder
OTN	Optical Transport Network
PoP	Point of Presence
ROADM	Reconfigurable Optical Add-Drop Multiplexer
TCP	Transmission Control Protocol
WDM	Wavelength Division Multiplexing

I. Introduction

COMMUNICATIONS traffic on the Internet Protocol (IP) backbones of Internet Service Providers (ISPs) has been continually growing. However, ISP revenue has not kept pace, and there is increasing pressure to reduce costs while maintaining high reliability, or even improving it. Reliability and availability in IP networks are provided in part through redundancy to protect against failures, with restoration mechanisms finding alternate routes for affected communication traffic flows. Note that our use of the term failures includes unplanned outages as well as planned maintenance activities.

Manuscript received November 06, 2013; revised January 26, 2014; accepted March 13, 2014. Date of publication April 10, 2014; date of current version May 29, 2014. The work of B. Ramamurthy was supported by an NSF FIA Grant (CNS-1040765) and the AT&T VURI Program. An earlier version of the paper was presented at the *19th International Workshop on Design of Reliable Communication Networks* (DRCN 2013), Budapest, Hungary, March 4–7, 2013 [13]. Associate Editor: Pin-Han Ho.

B. Ramamurthy is with the University of Nebraska-Lincoln, Lincoln, NE 68588 USA (e-mail: byrav@cse.unl.edu).

R. K. Sinha is with AT&T Labs-Research, Middletown, NJ 07748 USA (e-mail: sinha@research.att.com).

K. K. Ramakrishnan is with WINLAB, Rutgers University, New Brunswick, NJ 08091 USA (e-mail: kkramakrishnan@yahoo.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TR.2014.2315915

In terms of impact on networks, the main difference is that the planned maintenance activities are typically scheduled during off-peak hours to minimize service impact, and are usually not service affecting unless absolutely necessary. But planned events can introduce risk of impact from a subsequent failure. Moreover, before a planned maintenance, operators increase the routing weight on all affected links to gracefully steer traffic away from them. Failures of routers are typically handled by having redundant routers at each point-of-presence (PoP). The typical deployment of dual homing an access router (AR) to a pair of core backbone routers (BRs) to achieve a highly reliable IP backbone is a significant expense, as has been well recognized [10].

Routers, along with their associated linecards, contribute greatly to the overall cost of the network, both due to Capital Expenditure (CapEx), and Operational Expenditure (OpEx). Reducing the overall cost of the network can be achieved through reduction in the amount of equipment deployed. However, there is a lot of additional equipment and complex functionality in an ISP's backbone, beyond just the routers and their line cards. Reduction in overall cost achieved by simplifying the network topology at different layers must ensure a proper tradeoff between cost and reliability. Reduction of equipment and costs at Layer 3 (the standardized network organization category for routers and line cards) should not result in significant additional deployment of components and capacity at a different layer. At the same time, moving to a simpler architecture to keep costs low, where for instance only a single BR exists at each PoP, should not result in unacceptable availability.

The transport equipment in an ISP's network includes reconfigurable optical add-drop multiplexer (ROADM), optical transponder (OT, or simply, a transponder), regenerators, amplifiers and fiber. The cost of transport equipment is a major contributor to the overall cost. We observe that there can be significant opportunities for sharing transport resources provisioned for restoration if the network primarily experiences a single failure at a time. A single failure means planned maintenance or unplanned outage of a single network subsystem. Notice that a single failure can bring down multiple links. For example, failure of a single router fails all its adjacent links. We recognize that there may be situations where multiple failures occur concurrently, but we consider these to be a lower probability event, and also more expensive to protect against. Therefore, we consider the appropriate cost-reliability tradeoff to be one where single failures are handled without impacting reliability adversely. Carriers generally build networks with headroom (overprovisioning) for both failure restoration as well as for future growth. This capacity can be shared across different possible failures.

In the approach we pursue in this paper, we envisage a network with only one BR at each PoP. The ARs homing on that primary BR under normal operation instead home on a remote BR when there is a failure of that primary BR. However, having the ARs home on the remote BRs require transport capacity to be provisioned for this purpose. The novelty in our design approach is to share the capacity in the network across different possible single failures without incurring protocol latencies at the IP layer to recover from a failure. We also propose that the capacity provisioned between the ARs and the remote BR under normal operation is minimal (and the links are assigned a large weight in the Interior Gateway Protocol (IGP), e.g. Open Shortest Path First (OSPF)). Thus, the ARs have an adjacency established both with the local primary BR as well as the remote backup BR. When the local primary BR fails, the transport resources are resized to have sufficient capacity to carry the traffic to and from the ARs homed on the corresponding remote BR. This design avoids the large IGP convergence latency that is often the case when a new adjacency is established, along with all the delays to establish the transport circuit. We envisage an intelligent, agile Layer 1 network that can dynamically resize the transport circuit; we could certainly consider setting up a link-aggregation group that then has additional components added subsequent to detecting a failure.

In related work, Palkopoulou *et al.* [9], [10] compare several variations against a baseline dual router architecture. They evaluate the architecture where ARs are single homed to BRs, as well as architectures with optical switches or a common pool of shared restoration resources. The unit costs in our model have been borrowed from the detailed cost model of Huelsermann *et al.* [4] for multi-layer optical networks. Chiu *et al.* [2] propose reusing inter-office links from a failed BR to a surviving BR, by leveraging the optical layer. They report that this integrated IP-over-Optical layer restoration is 22% more efficient than pure IP based restoration.

In an earlier paper [12], we describe the cost and reliability considerations involved in designing next-generation backbone networks.

Our proposal is to achieve a fundamental redesign of IP backbones that avoids redundant routers by exploiting the capabilities of agile optical transports. We evaluate the alternative backbone designs in terms of cost and performability (a metric combining performance and reliability). Section II includes a detailed description of the operation of the network at the IP and the transport layer. In Section III, we propose alternative backbone network designs which use only a single router at each PoP but use the agile optical transport layer to carry traffic to the remote BRs to survive failures or outages of the single local BR. Section III-A describes a possible realization of the proposed design using current optical transport technology. In Section IV, we describe our evaluation metrics, viz., cost and performability. In Section V, we describe the ILP formulation used to solve the problem of optimally mapping local ARs to remote BRs in the new backbone network design. In Section VI, we describe the results comparing the cost and performability of our alternative design to that of the original design for a network modeled after a Tier-1 ISP backbone network. We then present our conclusions in Section VII.

II. ISP Backbone Architectures: Background

The IP backbone network of a typical ISP is rather complex, comprising multiple layers. Customer equipment, typically a customer edge router, connects to the core network through ARs, which in turn are connected to the core BR. BRs are located at a PoP, often a telecommunications central office

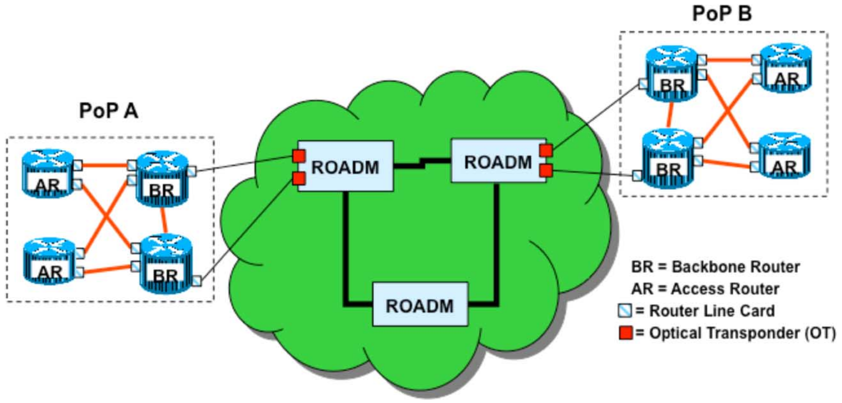

Fig. 1. Legacy backbone network.

or data center. An ISP may have a large number of ARs that aggregate traffic into a BR.

Typically, an ISP's PoPs are located in major metropolitan population centers that are the major source of incoming and outgoing traffic, especially in the United States. Each PoP typically houses a pair of BRs for the purpose of redundancy. The BRs are interconnected within the PoP by high speed short-range fiber links, usually multiple Gigabit/second Ethernet links. The ARs are dual-homed to the BRs within the PoP, again to provide redundancy, and achieve the necessary level of service availability. ARs that are located within the PoP are connected to the two BRs that are in the same PoP. This configuration is also the case for ARs that are close to the PoP. However, ARs that are farther away from a PoP are typically dual-homed to two different PoPs. Fig. 1 shows the configuration within a PoP, where there are two BRs at the PoP that are interconnected using short-range fiber. Further, the two ARs (representative of the multiple that may exist at that PoP) aggregate all the traffic in that metropolitan area, and are dual-homed to these two BRs. The BRs in the backbone at different PoPs are interconnected via a ROADM network. The ROADM network carries the optical signals at a standard wavelength. When each BR is connected to a ROADM, we utilize a transponder for converting the optical signal to an electrical signal (and vice versa), and to perform reshaping, retiming, and retransmitting functions (typically called the 3R functions). While the use of redundant BRs within a PoP is desired to provide the necessary availability, this architecture adds significant cost to the ISP.

In this paper, we explore alternative architectures for the PoP whereby the cost can be significantly reduced without compromising the availability of the ISP's backbone. We start by having only one BR in a PoP, and homing the ARs at that PoP to exactly one BR. However, we expose the backbone to severe degradation when there is a BR outage. All the traffic to and from the ARs single homed to this failed BR would be lost. Therefore, to ensure acceptable availability, we need to provide additional interconnectivity to an alternate BR without increasing cost significantly. This is the key contribution of this paper.

Customers needing higher availability connect to multiple ARs, so that when an AR router goes down, the customer traffic shifts to one of the its (surviving) connected ARs. In this paper, we focus on the outages of BRs, AR-BR links, and BR-BR links only. However the idea of connecting to multiple ARs to protect against AR outage can be applied in conjunction with the methods proposed in this paper.

Routers have become much more reliable over time. Unplanned, complete router failures are rare, except near the edge where routers are simpler, repurposed older, and cheaper. However, hardware and software upgrades continue to cause frequent outages. These are typically planned, but still need to be accounted for in the design of the network topology to provide adequate availability. A few router vendors support in-service software upgrades; but, as argued in [1], there is still a large base of deployed routers without such capability. There are newer mechanisms, such as the Hot Standby Router Protocol (HSRP), but this approach also results in significant additional cost, and may not be supported in legacy routers. The overall effect is that upgrades still have a substantial impact [7], and 1:1 router redundancy remains a prevalent practice in carrier networks.

We need to add redundant links to the topology to provide resiliency to failures at the IP layer. In attempting to reduce the Layer 3 cost of the network (router and its line cards), it is important to understand the impact of that reduction on the associated increase in the cost and complexity at the lower (optical) layers, as well as on overall network availability. Thus, it is useful to examine a key question: should restoration capabilities be provided at Layer 3, or at a lower layer, or should it be a combination?

Providing restoration exclusively at a lower layer is possibly inefficient because of the need to add substantial extra capacity for protection in the absence of statistical multiplexing due to packet switching. Furthermore, one would still have to deal with failures of components at the higher layer (e.g., router line cards) [11]. However, providing restoration at Layer 3 comes at the cost of availability (including the time taken to restore from a failure), because the recovery from a failure is through complex distributed protocols that rely on timers that are set to large values. These considerations have led carriers to add protection at different layers on an ad-hoc basis to compensate for the different failure recovery capabilities at each layer, and the cost to provide this protection. Thus, the overall system has evolved to be both complex and expensive. Carriers have to continually repeat such evaluations, and deploy restoration mechanisms and additional capacity each time the technology at a particular layer changes.

A. IP and Multi-Protocol Label Switching (MPLS) Restoration

The traditional way of providing reliability in the IP network is to provide redundancy, especially at the router level. However, IGP convergence tends to be slow. Production networks rarely shorten their timers to small enough values to allow for failure recovery in the sub-second range due to the potential of false alarms [3]. A common approach to providing fast recovery from single link failures is to use link-based Fast-Reroute (FRR). While some level of shared redundancy is provided to protect against link failures, such as sharing of backup resources for mesh restoration in MPLS FRR, the traditional means for providing protection against a BR failure is still to have a 1:1 redundant configuration of BRs at each PoP.

Typical IP backbone network design provides shared restoration capacity to overcome many different types of failures. However, the notable exception is the treatment of the outage of AR-BR connectivity (whether it is link or router failure), where essentially 1:1 redundancy is provided through the dual-homing approach. Such a conservative approach to deal with BR outages is very expensive, and with the right network design, unnecessary. Moreover, the traditional design of interconnecting an AR to a BR is treated as separate connectivity, not sharing any network resources used for the interconnection among the BRs. Our approach proposed in this paper overcomes both of these issues by having the backbone network capacity be utilized for restoration of both BR-BR failures, as well as AR-BR failures.

Our architecture ensures that the restoration capacity to protect against different BR failures can be shared, instead of being dedicated through dual homing. The capacity needed for protecting AR-BR link failures can also be provided by the backbone network, thus increasing the opportunity for shared restoration capacity. The overall impact therefore is that a shared pool of capacity in the backbone can be used to carry normal traffic as well as provide restoration capacity to protect against all types of failures: BR-BR link failures, BR router failures, AR-BR link failures, etc.

However, in this new architecture, it is important to ensure that the AR-BR links that are provided for restoration need to be set up dynamically, to avoid having dedicated, stranded capacity solely for purposes of restoration. Having such dedicated capacity defeats the goal of achieving sharing, and thereby multiplexing gains. Therefore, understanding the capabilities of the lower layers (physical and optical) to provide this dynamic, shared capacity is essential. We describe their properties next.

B. Optical Transport Layer Considerations

1) Layered Network: We can think of the network as consisting of multiple layers. A link at one layer can be created by a path that spans multiple links at the lower layer (see Fig. 2). For example, each link between two nodes in the ROADM or DWDM layer is a path at the fiber span layer below it.

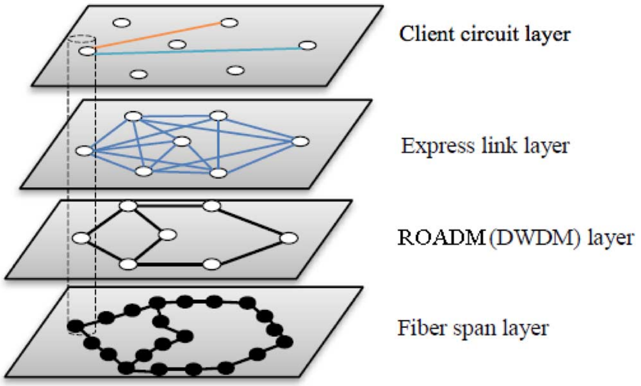

Fig. 2. Multi-layer optical network.

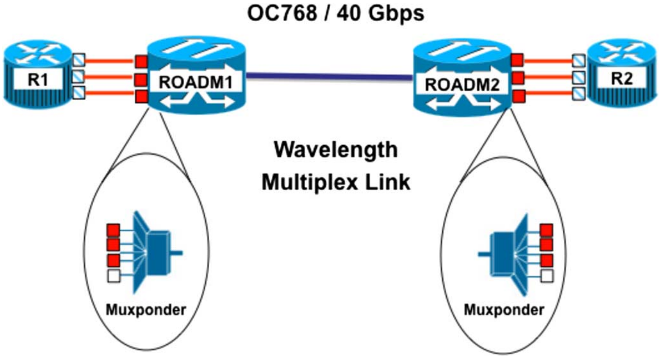

Fig. 3. Routing of the physical links over a multiplex link. Transponders in use are represented by filled squares and those not in use by empty squares.

2) Role of ROADM and Transponders: ROADMs allow optical wavelengths to be added, dropped, or bypassed (switched) in a reconfigurable manner at any network location. The ROADM nodes act as the origination and termination points for each optical circuit (wavelength path) in the network. Usually a ROADM node is co-located with a BR at each PoP. Additional ROADM nodes may be deployed in the network for improved connectivity. A fiber-optic cable carrying a short-reach wavelength is used to connect a router port to the corresponding ROADM.

At each end of an optical circuit, we use an OT adjacent to the router ports. An OT enables the transmission and reception of a client signal over a wavelength in the fiber using optical-to-electronic-to-optical (O/E/O) conversion. The type of transponder (e.g., 10G or 40G) is chosen according on the capacity of the circuit needed.

3) Role of Muxponder: In a given fiber, each optical circuit occupies either a full wavelength (e.g., 40G circuit in a 40Gbps system) or a sub-wavelength (e.g., 10G circuit in a 40Gbps system). Through a mechanism known as traffic grooming, multiple sub-wavelength circuits can be carried over a single wavelength to reduce cost. A special device known as a *muxponder*, which combines the functionality of a multiplexer and a transponder, is used for this purpose. A wavelength path which has been partitioned to carry sub-wavelength circuits is called a *multiplex* link (see Fig. 3). Thus, multiple 10Gs (subwavelengths) are carried over a single wavelength using a muxponder and 40G transport equipment. However, if there is

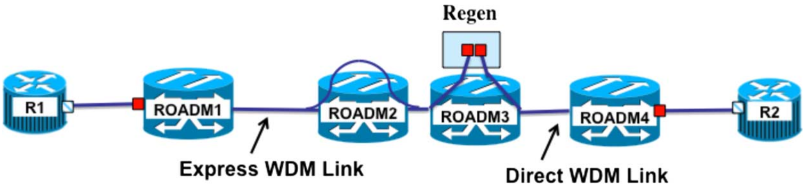

Fig. 4. An express link can bypass regeneration at intermediate ROADM nodes. The express link from ROADM 1 to ROADM 3 bypasses regeneration at ROADM 2. Regenerators on the circuit are represented by two adjacent filled squares.

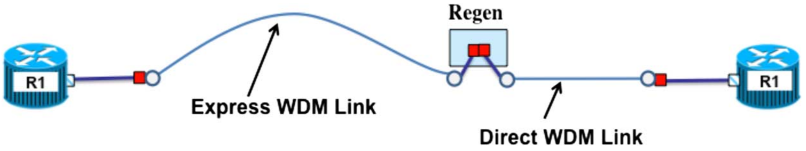

Fig. 5. A physical link can span multiple direct WDM links and express WDM links.

no anticipated capacity growth, it may be cheaper to use 10G transport equipment to carry a 10G circuit. With the advent of 100 Gbps systems, multiplexing would still be possible, with the full wavelength being 100 G, and sub-wavelengths being 10G or 40G.

4) Role of Regenerator: Regeneration is essential to clean up the optical signal to overcome bit-error rate (BER) degradation due to noise and crosstalk. Regeneration is performed on each individual circuit (10G or 40G) as needed, using a *regenerator* (or simply, a regen). Although a regen can be built using two transponders placed back-to-back, it is usually constructed separately in a simpler manner, and at a lower cost. Regeneration usually occurs at a ROADM location where the wavelength can be dropped or demultiplexed or both for this purpose.

5) Direct and Express WDM Links: The decision of where to regenerate the optical signal depends on the optical reach of the underlying transport system. The optical reach is a vendor-specific metric, and is dependent on various physical parameters. A WDM link that spans two adjacent ROADM nodes without any intermediate ROADM is referred to as a Direct WDM link. An Express WDM link, however, can span multiple ROADM nodes without requiring regeneration at intermediate nodes (Fig. 4). An optical circuit can be transported over multiple Direct WDM links or multiple Express WDM links (see Fig. 5). For a sample network, the ROADM or DWDM network layer consisting of only Direct WDM links and the corresponding express link layer consisting of both Direct WDM, and Express WDM links are shown in Figs. 6, and 7 respectively. Each one of the Direct or Express WDM links can be multiplexed to carry sub-wavelength circuits (e.g., 4 × 10G circuits over a 40Gbps wavelength).

6) Role of Amplifiers: An amplifier is a purely optical device which is used to combat signal attenuation by boosting the power of all the wavelengths carried by the optical fiber. Unlike OTs, muxponders, and regenerators which work on a per-wavelength basis, the optical amplifiers operate across all the wavelengths carried on that fiber. The most popular amplifier used in long-haul transmission is the Erbium-Doped Fiber Amplifier (EDFA).

7) Transporting IP Traffic: IP traffic is carried over the client circuits established between router ports across the optical transport layer. Inter-office links connecting BRs establish OSPF Layer 3 adjacencies. A single inter-office link is a logical (or aggregate) link comprising multiple physical links (such as multiple 10G and 40G circuits). In Fig. 8, for example, three 10G circuits between routers R1 and R2 form a logical link with a capacity of 30 Gbps. Logical links reduce the number of OSPF adjacencies, and a local hashing algorithm is used to decide which of the three physical links (circuits) to use for IP packets going over this logical link.

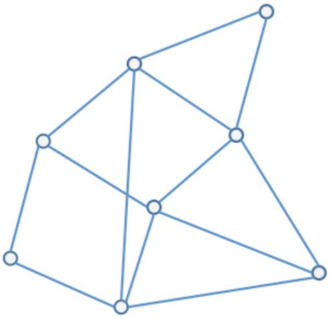

Fig. 6. A network topology consisting of direct WDM links only.

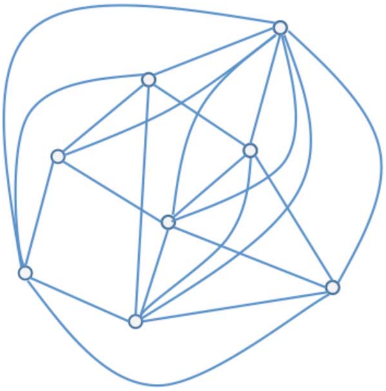

Fig. 7. A network topology consisting of direct WDM links and express WDM links.

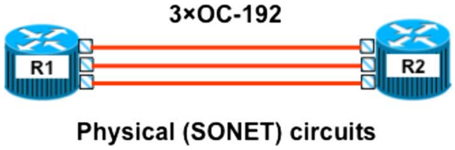

Fig. 8. Physical links that make up an aggregate L3 link.

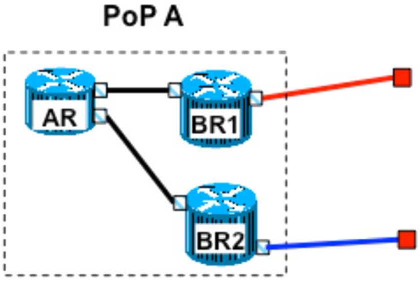

Fig. 9. Current architecture with dual-homed BRs.

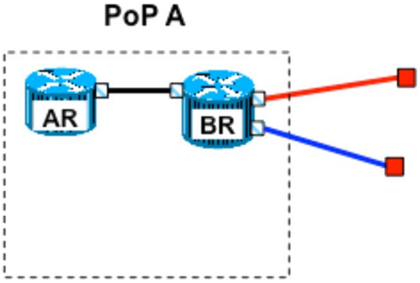

Fig. 10. Option-1: Eliminate one BR and move its links to the other BR.

III. Architecture Alternatives

In this section, we describe the different architecture alternatives that use a single backbone router (BR) at each PoP as a means of reducing cost along with the transport alternatives to carry the traffic to a remote BR.

The first option (Option-1) (see Fig. 10) for reducing the cost of a backbone is to eliminate one BR from each PoP, thus avoiding the cost of the additional BR. While this may be a simple approach, we still need to ensure that this is done in a manner that the availability of the service provided by the backbone network is not adversely impacted. Of the two BR_1 and BR_2 in each office, we eliminate BR_2, and move all of its links to BR_1. The cost reduction comes from eliminating roughly half of the AR-BR links, and all of the $BR_1 - BR_2$ intra-office links. However, this design cannot protect against any BR outage, and our performability evaluation in Section VI shows an unacceptable drop in performability. Option-1 is thus referred to as UR, for unreliable design, in Section VI.

To improve performability, our second option (Option-2) (see Fig. 11) improves on Option-1 by adding a link from each AR to an additional BR, located in a remote PoP (called *remote BR*, in the rest of the paper). While this does save the cost of the $BR_1 - BR_2$ intra-office links, it results in increased transport cost for connecting the ARs to the remote BRs. It also saves the chassis cost of the eliminated BR_2s, but may require extra line cards (with the expectation that this does not result in an additional chassis) on the remote BRs as we have to add more links to the remote BRs. The number of inter-office links, which tends to dominate the Layer 3 cost, does not change substantially as we are effectively replacing each AR—(local, second) BR link with an AR—(remote) BR link.

The final option (Option-3) (see Fig. 11) improves on Option-2 dynamically by setting up an AR—remote BR link upon failure of the local BR. We first eliminate the BR_2 router from each office, and identify a remote BR for each AR. However, instead of setting up permanent full capacity AR-remote BR links (as in Option-2), we size these links dynamically upon failure of the local BR, taking advantage of the agility available in newer optical transport equipment. Because we design for a single BR failure at a time, we need at most one AR—remote BR link at any given time. The cost advantage over Option-2 comes from multiplexing gains achieved by sharing the backup capacity, as we may be able to share transport resources as well as ports on ARs and remote BRs. We illustrate the source of savings in router ports with the following example. Suppose three ARs connect to the same remote BR, and require (respectively) 8, 9, and 7 10G-connections upon failure. In Option-2, this will require $8 + 9 + 7 = 24$ 10G ports. However, with Option-3, we will only need $\max(8, 9, 7) = 9$ 10G ports. We also get multiplexing gains from the sharing of transport resources among AR-(remote) BR connections. Moreover, in Option-2, each AR will need enough ports to connect to its local BR as well as to its remote BR. However, in Option-3, we can reuse the same AR ports to connect to either the local or the remote BR with the use of flexible fiber crossconnects.

We refer to Option-3 as our proposed architecture, and denote it as SR-100 in Section VI. Option-2 is not discussed further in this paper.

A. Realization of the Proposed Architecture

A key motivation for the proposed architecture is the need to efficiently work in the context of the standard network layer (Layer 3) protocols, so as to be able to recover from the failure of a link to the BR, or the entire BR, at an office. This recovery needs to be achieved quickly so that the period of outage is small. The longer the outage, the more packets are lost, leading to an impact on higher layer protocols such as TCP, either to recover from the loss of a large burst of packets, or even the failure of TCP connections.

With link-state routing protocols, each router learns the entire network topology. Each link has an administratively assigned weight. Each router computes the shortest-path tree to all other routers with itself as the root, using the weighted topology graph. Then it computes the next-hop to each possible destination along the shortest-path tree. To have consistent routing, all the routers need to have the same view of the topology. The topology is built in a distributed manner, where each router describes its local connectivity (i.e., the links incident on it and its weight), and reliably floods this information in a Link State Advertisement (LSA) message to all the routers. When a topology change occurs (e.g., a link, or multiple links fail), this information is propagated using LSAs [5, Chapter-11]. There are various timers for ensuring the LSAs are propagated in a timely

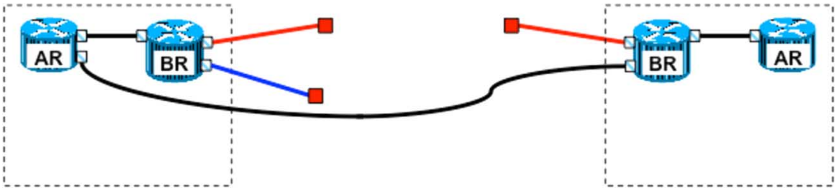

Fig. 11. Proposed architectures—Options 2 and 3.

manner as well as for controlling the amount of LSAs propagated if links flap[1] up and down. Eventually, the time it takes for the topology to converge so that data packets can be forwarded on the network without forming loops is based on the settings of the various timers, as well as the scale of the network (number of routers and links). While the typical time it takes for convergence has come down from several minutes [3] to a few seconds, by setting timers carefully, it can still mean a significant number of packets lost in a high-speed IP backbone upon a failure. More importantly, it is important to note that establishing a new alternate path after a failure can take a significant amount of additional time. When two neighboring routers (with a common link) come up, they form an adjacency to exchange LSAs. The two routers have to ensure that their view of the topology (as reflected in their link-state database) has to be consistent. This consistency ensures that the data packets they forward to each other do not form loops. To do so, the two neighboring routers exchange enough information to ensure that their link-state databases are consistent and synchronized before forwarding packets. Thus, the amount of time it takes to establish a new adjacency can be substantial, especially for a large scale IP backbone. It is exactly this adjacency establishment overhead and latency that we avoid at a time the backbone can least afford it, which is immediately after a failure of a link to a BR or the failure of a BR itself.

Our solution (first proposed in [12]) is instead to set up a permanent AR-remote BR link at a *low rate* to maintain protocol state (e.g., using keep-alive messages). Upon a failure, we dynamically resize this link to the required full rate. Doing so avoids bringing up new router adjacencies as well as propagation of failure information through LSAs. It is the capability of the newer agile optical networks that enable us to make this approach cost-effective by allowing it to be a low rate link under failure-free conditions.

The AR whose local BR has failed can recover connectivity to the rest of the network, through its remote BR adjacency, without the need for the entire network to converge. We recognize the possibility of short-term congestion, while the network is converging, but overall the complete reroute process would be transparent to the routing control plane. It is therefore similar to the case of having two BRs in each PoP, but at a significantly reduced cost.

We propose to use a service platform similar to that utilized by AT&T's GRIPhoN project [6]. A simplified diagram of a PoP is shown in Fig. 12. For simplicity, we show only one AR located in the PoP even though in reality we have several ARs homing on this BR. In some cases, the ARs may be 100s of miles away from this BR. The BR, AR, ROADM, and OTN equipment (not shown) are interconnected by an FXC (fiber crossconnect) switch. Under failure free operation, an AR has several 10G connections to the local BR. In our design, it also has a low-rate connection to a pre-determined remote BR. Upon failure of the local BR, we resize the AR—remote BR connection. One way of achieving this resizing is to set up a Link Aggregation Group (LAG) between the AR and the remote BR, and add additional individual circuits to it as needed. We exploit the OTN layer for sub wavelength circuits (e.g., for setting up the initial low rate 1 G, ODU-0, connection, as in [6]), and the DWDM layer is used for adding wavelength connections, e.g., multiples of 40G. We use an FXC to reuse the ports that are on the AR to the local BR. As shown in Fig. 12, upon failure of the BR at PoP A, all the ports on the access router at PoP A are connected to the BR at the remote PoP B.

IV. Evaluation of Network Designs

We evaluate network designs for *cost* and *performability*. The overall cost of the backbone network includes the cost of the optical transport equipment used for the interconnection of the routers as well as the cost of the routers (chassis, line cards) themselves.

A. Transport Layer Cost

To obtain the transport layer cost for a given optical circuit (10G or 40G) set up between two routers, we first need to know the path used by the circuit, and the list of all the optical transport layer elements encountered along the path. As more and more circuits are established in the network, additional options become available for carrying new circuits (e.g. over a multiplex express link established earlier). In our method, we establish circuits one-by-one in a given order, and compute the cost for each circuit according to the method described below.

As explained in Section II-B, transport equipment includes OTs, regenerators, and muxponders. Because transponders and regenerators are used on a per-circuit basis, the cost of a circuit includes the cost of the corresponding transponders (10G or 40G) at each end, and the cost of any regenerators (10G or 40G). Muxponder costs are incurred only by sub-wavelength circuits (e.g., up to four 10G circuits on a 40G wavelength). In addition, the network includes other optical layer components such amplifiers, ROADMs, and fibers which are pre-deployed. Because this set of transport equipment is *common* to multiple circuits (e.g., one amplifier is used by all wavelengths traversing a fiber), we use an amortized common cost contribution to each circuit on a per wavelength-km basis.

[1] In telecommunications terms, a flap refers to an intermittent failure, characterized by frequent, short duration failure conditions with self recovery.

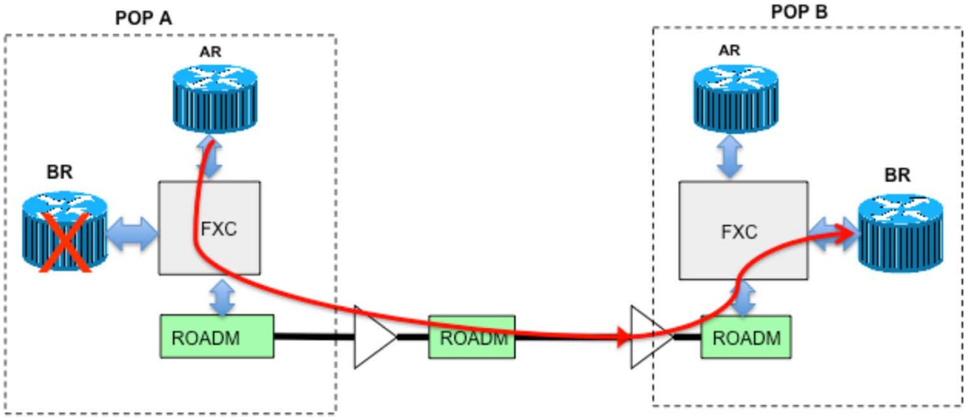

Fig. 12. Re-homing upon BR failure.

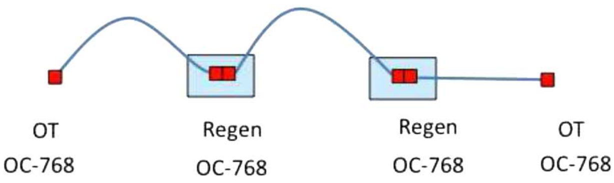

Fig. 13. Cost of a 40G circuit.

1) Cost of a 40G Circuit: The cost of a 40G circuit can be computed as the sum of the costs of the 40G transponders and 40G regenerators used along the WDM links of the circuit's path. Each WDM link in the path can either be a Direct WDM link or an Express WDM link (see Section II-B-5). A 40G transponder is used on each end-point of the circuit, and a 40G regenerator is used for interconnecting two adjacent WDM links in the end-to-end path. For example, in Fig. 13, a new 40G circuit is carried over two Express WDM links (curved lines) followed by a Direct WDM link (straight line). Hence the cost for the circuit is $2 \ * (\text{Cost of 40 G transponder}) + 2 \ * (\text{Cost of 40 G regen})$.

2) Cost of a 10G Circuit: Computing the cost for a 10G circuit is a bit more involved because a 10G circuit is often carried over a multiplex link (see Section II-B-3). Deploying a pair of muxponders to create the first sub-wavelength circuit on a 40G WDM link ensures that additional sub-wavelength circuits can be supported automatically. However, the muxponder cost must be amortized across all the current (and future) sub-wavelength circuits which benefit from it. Thus, for simplicity, we charge each circuit one-fourth the cost of the muxponder. Also, unlike with a 40G circuit, both 10G and 40G regenerators may appear in the path carrying a 10G circuit. A 10G regenerator is used to interconnect two adjacent 10G WDM links in the path while a 40G regenerator is used to interconnect two adjacent 40G WDM links along the path. Thus, the transponder, regenerator, and muxponder costs all contribute to the cost of each 10G circuit.

The cost for a new 10G circuit varies depending on whether a new multiplex link needs to be created in the network or not. A new multiplex link, if one is created, may use a sequence of Direct or Express or both WDM links. Below, we describe four scenarios for carrying a 10G circuit across the transport network. In all of the corresponding figures, new components are shown in darkly shaded portions, existing equipments are shown in lightly shaded portions, Direct links are shown as straight lines, Express links are shown as curved lines, and multiplex links as wavy lines. Also, for comparing different scenarios, we ignore the common cost (cost of ROADMS, fiber, amplifiers etc.).

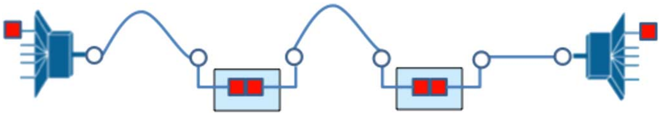

Fig. 14. Scenario 1: Using a new multiplex link routed over two express links and a direct link.

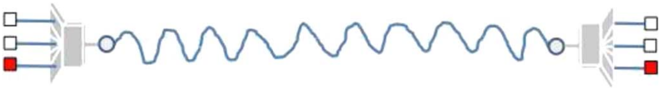

Fig. 15. Scenario 2: Using an existing multiplex link.

In Scenario 1 (see Fig. 14), the new 10G circuit uses a new multiplex link carried over two Express links, and a Direct link using an unused wavelength on each link. The wavelength is operated at 40Gbps, and muxponders are used at both ends to carry the new 10G circuit. The total cost for carrying the 10G circuit is $2 \ * (\text{Cost of 10 G transponder}) + 2 \ * (\text{Cost of 40 G regen}) + 2 * (\text{Amortized cost of muxponder})$. Note that three more 10G circuits can be carried over this multiplex link in the future due to the deployed muxponders.

In Scenario 2 (see Fig. 15), the new 10G circuit is carried over an existing multiplex link. The total cost for carrying the 10G circuit is $2 \ * (\text{Cost of 10 G transponder}) + 2 \ * (\text{Amortized cost of muxponder})$.

In Scenario 3 (see Fig. 16), the new 10G circuit is carried over a pre-existing multiplex link, and a new multiplex link that spans two Express links and a Direct link. An unused wavelength at 40Gbps is used on each WDM link, and muxponders are used at both ends to carry the new 10G circuit (similar to Scenario 1). The total cost for carrying the 10G circuit is $2 \ * (\text{Cost of 10 G transponder}) + 2 \times (\text{Cost of 40 G regen}) + 4 * (\text{Amortized cost of muxponder}) + (\text{Cost of 10 G regen})$. An

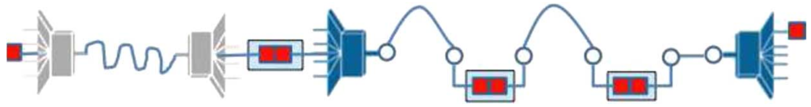

Fig. 16. Scenario 3: Using an existing multiplex link and by creating an additional, new, multiplex link.

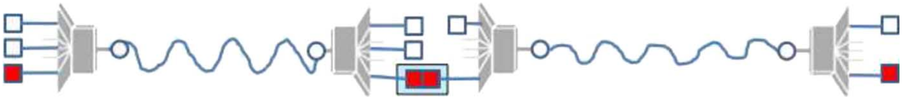

Fig. 17. Scenario 4: Using a sequence of two existing multiplex links.

additional cost incurred here (compared to Scenario 1) is due to a 10G regen which is required to transport the 10G circuit across the old and the new multiplex links.

Finally, in Scenario 4 (see Fig. 17), the new 10G circuit is carried over two pre-existing multiplex links. The total cost for carrying the 10G circuit is $2*(\text{Cost of 10 G transponder})+1*(\text{Cost of 10 G regen})+4*(\text{Amortized cost of muxponder})$.

Through a combination of the above scenarios, additional options are possible for carrying a new 10G circuit over existing and new multiplex links, and their costs should be calculated accordingly.

B. Router Cost

Router equipment includes router ports, line-cards, and chassis. Given the set of circuits in a design, we compute the required number of ports on each router. We assume that all inter-office BR-BR links are 40G or 10G; and all intra-office BR-BR links are 10G. Then we estimate the number of line-cards (and chassis) based on the number of required ports. In this paper, we focus on reducing the cost of the backbone portion of the network, and therefore use a simplified access model where each AR is located inside a PoP. In reality, access networks tend to have a complex hierarchical structure with aggregator switches multiplexing low-rate connections into high rate ports. While our simplified model masks the complexities of interconnecting remotely located ARs, it still provides a good estimate of overall savings as a result of changes *in the backbone portion* of the network.

C. Network Cost

For computing network cost, we used normalized equipment prices reported in [4], which are based on data from Information Society Technologies (IST) Integrated Project on Next Generation Optical Networks for Broadband European Leadership (NOBEL) phase 2. Notice that these prices are different from the equipment price numbers used in our previous paper [12]. These relative costs should be treated as examples only, for illustrating the efficacy of our approach, across a wide range of variation of the relative costs of the various components used in a typical IP backbone network. Equipment prices tend to vary over time, and so, in Section VI, we include a sensitivity analysis of how our estimated savings change with equipment prices.

D. Performability

For evaluating *performability*, we used the *nperf* tool [8]. Our goal is to estimate the expected packet loss from all *failure scenarios* (representing failures of one or more network components). The contribution of any given failure scenario to the expected packet loss metric is the product of its probability and its impact, where we estimate its impact after taking restoration into account. Although it is common in network literature to ignore multiple failures (because of their low probability compared to single failures), certain double failures may in fact contribute more to the overall expected loss metric. This effect can happen, for example, if network restoration method ensures zero or very low loss for a single failure. As a consequence, we consider not only all possible single failures, but also consider a subset of double failures that have a probability of occurrence above a threshold. In our evaluations, we consider 10,000 of the most likely failure scenarios to show that our design achieves our performability target.

The tool considers a set of *failure scenarios* representing failures of one or more network components. For each failure scenario, we first determine the set of failed circuits. A single component failure can bring down multiple circuits. Take, for example, when a router fails, all its incident circuits also fail; an amplifier failure or fiber cut fails all circuits routed over those components. The set of failed circuits in a scenario is the union of failed circuits from the failure of the individual network components.

Next we determine the effect of these failed circuits on *logical* links. Recall that a logical link may be an aggregate of multiple circuits that gives the appearance of a single link to the routers. If only a subset of the member circuits fail, then the net effect is a reduction in this (aggregated) logical link's capacity, but the link does not fail. In this paper, we assume that the network uses OSPF routing. If none of the links fail, then the flows stay on their original routes, but may experience packet loss due to congestion as some of the links in the route may have reduced capacity. If some of the links fail, then OSPF routing recomputes a new set of routes (based on routing weights assigned to each link), and reroutes some of the flows. There are two possible sources of packet loss. For the first source, it is possible that a failure scenario may disconnect the network graph, and thus a flow may not have any possible route. Even if a flow, with failed links in its current route, does have an alternate route, it takes several seconds to detect the failure and reroute this flow, during which time some packets get lost. We broadly categorize this type of packet loss as resulting from *unavailability of routes*. For the second source of packet loss, the amount of flow sent on a link may exceed its capacity. This may happen either because a link lost a subset of its member circuits (and thus has reduced capacity), or because many different flows got rerouted to this link. We categorize this packet loss as resulting from *link congestion*.

For each failure scenario, we determine the amount of traffic loss due to *unavailability of routes*, and *link congestion*. In addition to the loss computation, the tool also computes the probability of this failure scenario, based on vendor and proprietary field tested estimates of mean time between failures (MTBF) and mean time to repair (MTTR) for components. The end results are two probability distributions of traffic losses: (a) loss due to unavailability of routes, and (b) loss due to link congestion. While comparing different designs in Section VI, we report values of one minus the expectations of these distributions, and (respectively) call them *No route*, and *Congestion* performability. For example, a no route performability of 0.999 means that, over a long period of time, we expect $1 - 0.999 = 0.001$ fraction of the traffic to have no route.

E. Cost-Performability Trade-Off

There is an obvious trade-off between cost and performability. Increasing link capacities improves congestion performability, but also increases the cost. So the cost and performability of a design should always be considered together, and not in isolation. In our evaluations, we considered a design goal of surviving all single failures (router ports, complete router, amplifier, OT, etc.) to determine the appropriate capacities on links. Then we ran the *nperf* tool on 10,000 most probable single and multiple failure scenarios to evaluate the performability. Considering single failures is a standard practice because these failures cover a large fraction of the failure probability space. However, we want to emphasize that this design heuristic is one of several possible reasonable choices. We could reduce the capacities a bit to reduce cost at lower performability, or increase capacities to increase cost and performability. Ultimately, the real merit of our results is that we show substantial cost savings while offering acceptable performability. The exact trade-off between cost and performability in our designs can be somewhat adjusted depending on the requirement of the ISP.

F. Baseline Network Design

We used the following iterative method to compute link capacities that are barely sufficient to survive any single failure. We started with a model where each logical link was an aggregate collection of 10G and 40G circuits. (We allow for the possibility that a link may have a *single* circuit.) In each iteration, we increased or reduced capacities on logical links using the following process. We simulated all single failures using the *nperf* tool. For each failure, we computed the circuits that go down, and how those affected flows get rerouted. Then, for each logical link, we obtained the highest utilization across all single failures. If this utilization was more than 100%, we added circuits on that (aggregate) logical link to make the utilization below 100%. We did this by adding the smallest possible capacity in multiples of 10G. For example, if this logical link consisted of two 40G links (a combined 80 G of capacity), and the combined utilization was 140%, then we need the link capacity to be $2 \times 40\ \mathrm{G} \times 1.4 = 112\ \mathrm{G}$ to get the utilization equal to 100%. So we needed another $112\ \mathrm{G} - 80\ \mathrm{G} = 32\ \mathrm{G}$ of capacity. Rounding 32 G to the nearest multiple of 10G suggested adding another 40G link. In certain cases, we added capacity by replacing a 10G link with a 40G link. The resulting new circuits had to be routed over a new set of fiber spans, regenerators, OTs, etc. They also required additional router ports. So our set of single failures changed, and we could not guarantee that the utilization will remain below 100% when one of these newly added components failed. Similarly, if this utilization was less than 100%, we reduced the number of circuits in that logical link (or replaced a 40G link with one, two, or three 10G links), which also changed the set of single failures, and thus changed the highest utilization over the set of single failures.

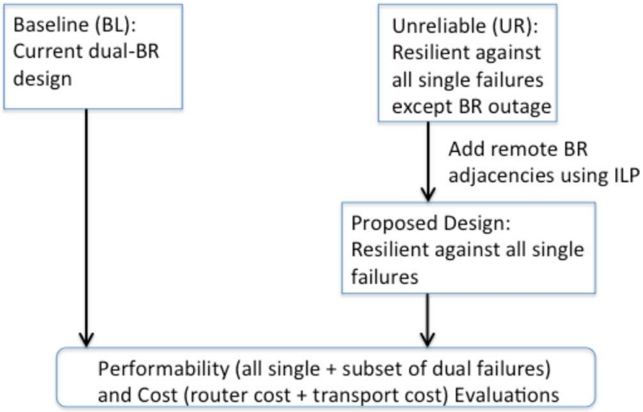
Baseline (BL): Current dual-BR design
Unreliable (UR): Resilient against all single failures except BR outage
Add remote BR adjacencies using ILP
Proposed Design: Resilient against all single failures
Performability (all single + subset of dual failures) and Cost (router cost + transport cost) Evaluations

Fig. 18. BL, UR and proposed designs.

After each iteration, we counted the number of links with maximum utilization less than 90% or greater than 110% upon a single failure. We stopped when either this number became zero (all links had utilization between 90% and 110% upon a single failure), or subsequent iterations stopped reducing this number. We also replaced any set of four 10G circuits (in the same logical link) with a 40G circuit. This network design is referred to as the BL, Baseline design, in Section VI.

Fig. 18 provides a high level overview of the approach we use to design the initial network topology, as well as our approach to evaluate the proposed design.

We have a baseline design (referred to as BL in Section VI) that models the current dual-BR architecture. We use BL to estimate changes in cost and performability of our proposed designs. We first create a simple, single-BR design that can survive all single failures with the exception of complete BR outages (we include failures of links and fiber spans). This design is referred to as Unreliable design, UR, in Section VI. Then we use the ILP in Section V to add links to remote BRs of the Unreliable design to protect against complete BR outages. Thus, the resulting designs are resilient against *all* single failures. Finally, we evaluate these designs on all single failures, and a subset of double failures, to estimate expected packet loss. When the local BR fails in our proposed design, we assume that all packets from the AR directly connected to it get lost for 1 minute while the remote adjacency is getting established. We also estimate the cost of the baseline design as well as our proposed designs by summing up the costs of the router and transport equipment.

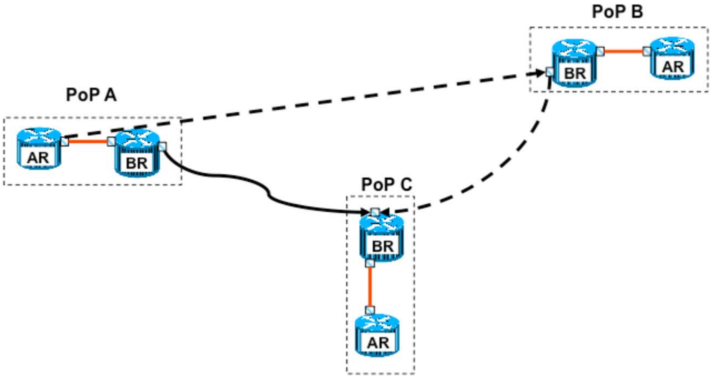

Fig. 19. Recovery scenario, a need for additional capacity.

V. ILP Formulation for Optimal AR-Remote BR Mapping

We start with a design where each PoP has only one BR that all ARs in this PoP are connected to. As outlined earlier, when the local BR fails, the traffic from that office moves to a predetermined remote BR. We need to find the mapping from ARs to remote BRs that minimizes the additional network cost while ensuring that all flows originating at this PoP have a route with sufficient link capacities. We consider a generalized version of the problem where each flow is classified either as *priority*, or as *best-effort*, and we only need to worry about restoring priority traffic. If there is no class-of-service, all traffic is treated as being restorable, as if all were priority traffic.

We find this optimal mapping using an ILP. The ILP formulation assumes that the routing of a circuit only depends on the two end-points of a circuit, e.g, along shortest path on the L1–L2 network.

For any AR, the *locally* best mapping is to assign it to the nearest remote BR to minimize the transport cost. However, notice that we are designing for a *single* BR failure, so we would like to share the resources assigned for AR to remote BR mappings. For example, if AR A_1 is mapped to remote BR B_1, then in some sense we have already paid for these links, and for additional ports on B_1. At this point, if we are trying to find a mapping for AR A_2, we would like to use these resources as much as possible, even if it means mapping A_2 to a far away remote BR. In fact, our optimal solution (with a few exceptions) creates clusters of ARs based on their geographic proximity, and then maps all ARs in a cluster to the same remote BR.

All ARs connected to a given BR are mapped to the same remote BR. So for modeling, we collapse all these ARs into a single (super) AR. Given n PoPs, we number ARs and BRs from $1 \ldots n$, and without loss of generality assume that the (super) AR and BR in the i-th PoP ($1 \leq i \leq n$) are both numbered i. Thus, the i-th AR is connected to its (local) i-th BR, and can be remotely connected to any of the remaining $n-1$ remote BRs. So altogether there are n^{n-1} possible connections. Using a set of precomputed values, we will show that we can pick the best solution among these $\theta(n^n)$ possibilities with an ILP of complexity $O(n^2)$.

The additional cost of protecting BR failures has four different components.

A. Cost Components for AR-Remote BR Mapping

1) At each AR, we need additional (10G) OTs to set up links to remote BRs. We do not need any additional (10G) router ports because we can reuse the router ports used to connect to the local BR. The number of additional OTs is equal to the number of router ports used by priority traffic, and can be precomputed as m_i (corresponding to the number of 10G links) for the i-th AR.
2) At each BR, we need additional (10G) OTs and (10G) router ports to accommodate the link from remote ARs. The number M_j of OTs and ports at the j-th BR is a maximum of m_i across all the ARs mapped to this BR. Notice that we also have a (permanent) low rate connection between ARs and remote BRs that we are not accounting for in the above statement. For example, if two ARs map to the same remote BR, and they each require 50 Gbps of uplink, then we will need two 1 Gbps (ODU-0) connections permanently, and will have to resize one of the 1 Gbps connections to a 50 Gbps connection upon failure of the local BR. So the total additional capacity needed on the remote BR will be 51 Gbps, and not 50 Gbps. A similar statement applies to additional OTs on ARs. Because the cost of these permanent connections is small compared to the the rest of the costs, and to keep the ILP formulation simple, we ignore these small costs in the remainder of this section. They can be added to the final cost once we have determined the AR to remote BR mappings.
3) If the i-th AR is mapped to the jth-remote BR, then we need transport capacity (equal to the amount of priority traffic from i-th AR) to set up this link upon failure of the local BR. Transport cost includes the cost of regenerators, fiber, ROADMs, and amplifiers.
4) We may also need *additional* capacity at certain inter-office BR-BR links. Consider the following scenario (see Fig. 19). Suppose we decide to map the AR at $PoP\ A$ to the remote BR at $PoP\ B$, and let $PoP\ C$ be a different location. When the local BR at $PoP\ A$ fails, all the traffic that was flowing between $PoP\ A$ and $PoP\ C$ now shifts, and is carried between $PoP\ B$ and $PoP\ C$. It is possible that some of the links in the $PoP\ B$ to $PoP\ C$ path do not have enough capacity to carry all this traffic, and would require additional capacity.

TABLE I
Parameters in the ILP Formulation

	Number	Description
n	1	Number of ARs/BRs
N	1	Number of transport resources
m_i	n	Number of 10G OTs at the i-th AR for priority traffic; can be precomputed from the number of ports on the i-th AR
c_k	N	Unit cost of the k-th transport resource
s_k^{ij}	Nn^2	Number of units of the k-th transport resource needed upon remapping of the i-th AR to the j-th remote BR, when the i-th BR fails.

B. Description of Model Parameters and Variables

The output of the ILP is the mapping from ARs to remote BRs. This mapping is defined by indicator variable r_{ij} which is 1, if i-th AR is connected to j-th BR upon failure of its local BR, and 0, otherwise.

In addition, the ILP also computes the number of units of the k-th resource needed, specified by variable S_k, and the number of additional 10G OTs and ports required at the j-th BR for setting up the AR-remote BR links, specified by variable M_j.

There are three sets of parameters that we precompute and pass to the ILP. Parameter m_i is the number of 10G links needed at the i-th AR to carry all its priority traffic. This value can be determined from the number of 10G ports on the i-th AR needed to carry all its priority traffic, and is an input to the ILP. Parameter s_k^{ij} is the number of units of the k-th resource we need for cost items (3) and (4) above upon reconnection of the i-th AR to the j-th BR, when the i-th BR fails. We explain how to precompute s_k^{ij} in the next subsection. Finally, parameter c_k is the unit cost of the k-th resource.

C. Precomputing s_k^{ij}

The key idea behind our ILP's efficiency is that (a) even though there are $\Omega(n^n)$ possible mappings, we can capture the resource usage by $O(n^2)$ s_k^{ij}; and (b) we can precompute these outside of ILP, and thus they become parameters to the ILP. We can precompute s_k^{ij} as follows.

1) Add a link from the i-th AR to the j-th BR in the *nperf* model. This link should have capacity $m_i \times 10$ G, and very high OSPF weight so that it does not carry any traffic unless local BR fails.
2) Use *nperf* to simulate the failure of the (local) i-th BR, and compute the utilization of each edge needed to restore all priority traffic from this AR.
3) For any edge with utilization above 100%, we determine the amount of extra capacity to keep utilization under 100%.
4) Route all additional inter-office links as well as $m_i \times 10$ G capacity from i-th AR to the j-th BR link. The number of k-th resources needed for this set of circuits is s_k^{ij}.

Tables I and II summarize the parameters and variables.

TABLE II
Variables in the ILP Formulation

	Number	Description
M_j	n	Number of 10G OTs/ports at the j-th BR
r_{ij}	n^2	Boolean variable: 1 iff i-th AR gets mapped to j-th remote BR
S_k	N	Number of units of the k-th transport resource needed

D. Objective Function

Our goal is to minimize the total cost of the network. Thus, the objective is

$$\min \left[\sum_{1\leq j\leq n} M_j * (\text{cost of 10G OT and 10G port}) + \sum_{1\leq k\leq N} c_k * S_k \right].$$

To compute the additional network cost, we need to add, to the ILP solution, $m_i * (\text{cost of 10G OTs})$, and the cost of maintaining the AR to remote BR *permanent* connections.

E. Constraints

1) Each AR is connected to exactly one BR:

$$\forall i, \quad \sum_{1\leq j\leq n} r_{ij} = 1.$$

2) Each AR is connected to a remote BR (to remove the degenerate case of AR having two connections to its local BR):

$$\forall i, \quad r_{ii} = 0.$$

3) Each BR needs ports and OTs equal to the maximum number of ports on one of its connected ARs:

$$\forall j, \quad M_j = \max_i \{m_i | r_{ij} = 1\}.$$

Because m_i are input to the ILP (not variables), the above constraint can be equivalently expressed as n^2 linear constraints:

$$\forall i, j, \quad M_j \geq m_i * r_{ij}.$$

When $r_{ij} = 0$, the inequality is vacuously true. When $r_{ij} = 1$, the inequality becomes $M_j \geq m_i$. And because we are minimizing M_j in our objective function, we know that one of these inequalities will be tight, and we will get $M_j = \max\{m_i | r_{ij} = 1\}$.

4) Because we consider at most one BR failure at a time, the additional units of the k-th transport resource is maximum across all AR to remote BR mappings:

$$\forall k, \quad S_k = \max_{i,j} \left\{ s_k^{ij} | r_{ij} = 1 \right\}.$$

As with the previous constraint, this constraint is equivalent to Nn^2 linear constraints:

$$\forall i, j, k, \quad S_k \geq s_k^{ij} * r_{ij}.$$

For a fixed i and k, the n constraints are $\forall j, S_k \geq s_k^{ij} * r_{ij}$. However, we have a separate constraint stating that, for a fixed i, exactly one r_{ij} is one, and the rest are zero. So we can rewrite these n constraints as a single constraint (albeit of n terms) $S_k \geq \sum_j s_k^{ij} * r_{ij}$. Thus, the above set of constraints can be rewritten as Nn constraints:

$$\forall i, k, \quad S_k \geq \sum_j s_k^{ij} * r_{ij}.$$

F. Scalability of the ILP Approach

For the topology we considered (see Section VI below), our ILP formulation resulted in more than 250,000 parameters, more than 1000 variables, and more than 10,000 constraints. The ILP solver obtained a solution in under 10 minutes on a lightly loaded Linux server with a 1.5 GHz Itanium processor.

VI. Results

We started with the topology and traffic matrix modeling a Tier-1 ISP backbone network. This is a *baseline* design to estimate changes in cost and performance of our proposed designs. This model has PoPs in major US cities, where each PoP houses two BRs connected by a set of 10G Ethernet links. Each AR is located in a PoP, and is connected to two BRs in its PoP by a set of 10G Ethernet links. Each inter-city link connecting BRs is an aggregate of 10Gs and 40Gs. As explained in Section IV, we sized each logical link to survive all single failures. Due to long ordering cycles for additional capacity, production networks always have excess capacity for anticipated traffic growth. This additional capacity would have inflated our cost savings as we would be starting with a network of higher cost than necessary. So to create a fair baseline, we resized the capacities on each link to barely survive all single failures of router ports, complete routers, amplifiers, fibers, and OTs. This design is referred to as BL in Table III.

A simplistic option to reducing the cost of a backbone is to eliminate one BR from each PoP, and then move all of the links from the removed BR to the surviving BR. For inter-city BR-BR links, we sized their capacities to survive all single failures except complete router outages. This is referred to as UR in Table III, and Table IV shows its cost and performability. (This design is called Option-1 in Section III.) The cost reduction comes from eliminating roughly half of the AR-BR links, all of $BR_1 - BR_2$ intra-office links, and chassis related to removed BRs. We also save on the inter-city BR-BR links because, with all links concentrated on a single BR instead of being spread out over a pair of BRs, we get better capacity multiplexing. However, this design cannot protect against any BR outage, and has less than three 9s of no route performability, which is an unacceptable threshold in carrier grade networks.

TABLE III
Design Names and Descriptions

Design name	Design description
BL	Baseline design. Each AR is dual homed to two local BRs. Restoration design to protect all traffic upon any single failure
UR	Unreliable design. Each AR homes to a single local BR. Restoration design to protect all traffic upon any single failure except complete router outage. Drop all traffic from ARs when their local BR fails.
SR-100	Each AR homes to a single local BR. Assume 100% of the traffic is priority. Restoration design to protect *all* traffic upon any single failure. Rehome ARs to a remote BR when their local BR fails.
SR-75	Assume 75% of the traffic is priority. Each AR homes to a single local BR. Restoration design to protect *priority* traffic upon any single failure. Rehome ARs to a remote BR when their local BR fails.
SR-50	Assume 50% of the traffic is priority. Each AR homes to a single local BR. Restoration design to protect *priority* traffic upon any single failure. Rehome ARs to a remote BR when their local BR fails.
SR-25	Assume 25% of the traffic is priority. Each AR homes to a single local BR. Restoration design to protect *priority* traffic upon any single failure. Rehome ARs to a remote BR when their local BR fails.

TABLE IV
Cost and Performability

Design	% Savings from BL	Performability	
		No route	Congestion
BL	0	0.99998	0.99996
UR	35.12	0.99896	0.99994
SR-100	30.72	0.99998	0.99998

The last row of Table IV shows our proposed design (referred to as SR-100 in Table III), where any AR, upon failure of its local BR, logically connects (homes) to a remote BR. The rehoming, as well as the additional capacity, is computed by ILP described in Section V starting from UR. For performability evaluation, we assume that, when the local BR fails, traffic originating at that AR *is lost for a brief period (for 1 minute, in our experiments) and then gets rehomed to the remote BR*. As we can see, rehoming adds very little to the overall cost (cost savings from BL reduces from 35.12% to 30.72%), but matches the performability of the baseline design. The reason for such a small additional cost is because, by setting up these remote connections dynamically (instead of permanent connections), we are exploiting statistical multiplexing in use of transport resources. The minor improvement in congestion performability in SR-100

TABLE V
COST AND PERFORMABILITY WITH CLASS OF SERVICE

Design	% Savings from BL	Performability			
		Priority		Best effort	
		No route	Cong	No route	Cong
BL	0	0.99998	0.99996	0.99998	0.99996
UR	35.12	0.99896	0.99994	0.99896	0.99994
SR-100	30.72	0.99998	0.99998	0.99998	0.99998
SR-75	40.59	0.99998	0.99982	0.99998	0.99617
SR-50	48.9	0.99993	0.99997	0.99997	0.99572
SR-25	55.94	0.99997	0.99998	0.99997	0.99510

over BL is incidental. Some of the capacity we added for remote homing happened to help with congestion in multiple failures.

A. Designing for Restoration of Priority Traffic Only

In networks supporting different classes of service (CoS), priority and best effort traffic have different SLAs. We consider network designs where we provide restoration capacity for priority traffic only. Notice that, just because we do not consider best effort traffic in our restoration design, it does not mean that all best effort traffic gets dropped upon a failure. Say link L, upon failure F_1, needs 10 units of additional capacity for priority traffic; and, upon a different failure F_2, the link needs 20 units of additional capacity for priority traffic. Further, suppose that we add $\max(10, 20) = 20$ units of additional capacity on link L. Upon failure F_2, indeed all the additional capacity will be taken by priority traffic, and all affected best-effort traffic will be dropped. However, upon failure F_1, priority traffic only needs 10 units of capacity, and the remaining 10 units will be used to restore best-effort traffic.

Table V lists the performability of the Baseline (BL) and Unreliable (UR) designs and that of our proposed Single Router (SR) designs when 100%, 75%, 50%, and 25% of the traffic is classified as priority. The first two rows repeat the results from Table IV, and (because they do not consider CoS) have the same performability for all traffic. For designs, SR-75, SR-50, and SR-25, we first size their link capacities so that all priority traffic survives any single failure except complete router outage, and then we find the remote BR mapping and additional capacities using the ILP in Section V.

We see substantial improvement in cost savings as the fraction of priority traffic goes down. If half of traffic is best-effort (SR-50), we are getting a savings of nearly 50% where performability of priority traffic nearly matches those in BL. The only drop is in the congestion performability of best effort traffic where the application layer may be able to deal with a small amount of lost packets. The minor differences in no route performability (in the 5th decimal place) is because our design heuristic of getting all link utilizations near 100% is somewhat coarse. As argued at the end of Section IV, with a proper network design tool, we can tweak the performability and costs of these designs.

TABLE VI
COST SENSITIVITY RELATIVE TO ROUTER AND TRANSPORT EQUIPMENT COSTS

Design	Router equipment cost multiplier	% Savings from BL
UR	1.0	35.12
	0.5	33.47
	0.1	28.38
SR-100	1.0	30.72
	0.5	27.58
	0.1	17.92
SR-75	1.0	40.59
	0.5	36.67
	0.1	24.58
SR-50	1.0	48.90
	0.5	46.93
	0.1	40.84
SR-25	1.0	55.94
	0.5	54.37
	0.1	49.56

TABLE VII
COST SENSITIVITY RELATIVE TO TRAFFIC SCALING

Design	Router (unit) cost multiplier	% Savings from BL
SR-100	1.0	30.72
	0.5	27.58
	0.1	17.92
SR-100 with 10x original traffic	1.0	34.94
	0.5	33.1
	0.1	27.09

B. Cost Sensitivity With Respect to Router and Transport Equipment Costs

Our proposed designs have lower transport and router related costs compared to the baseline model, but the percentage savings are lower for transport cost compared to the router related cost. The projected cost savings are dependent on unit equipment costs, and if router equipment costs were to go down (compared to transport equipment costs) then our projected savings will also go down. We estimated our cost savings based on equipment prices reported in [4], but recent trends towards cheaper Ethernet based switching have pushed the router costs down so Table VI shows a sensitivity analysis of our estimated cost savings. Each design has three rows. The top row lists the savings with the equipment cost reported in [4]. If the transport equipment prices go up (relative to router equipment prices), our savings will improve, and we do not show them in the table. However the next two rows shows how our savings go down if router equipment became twice (router cost multiplier is 0.5) or

10 times cheaper (router cost multiplier is 0.1). We see that, even in the case of a 10× reduction in router prices, our cost savings remain nearly 18% for SR-100 to nearly 50% for SR-25.

C. Cost Sensitivity With Respect to Traffic Scaling

Finally, we examine how our savings vary with traffic matrix scaling by increasing traffic 10 fold. This effect has a major impact on the design as the increased traffic nearly eliminates the need for sub-wavelength 10G circuits. As shown in Table VII, our cost savings improve slightly with the higher traffic.

VII. Conclusion

Network service providers continue to see increased pressures to reduce the cost of their IP backbone networks. A significant cost is incurred by the core BRs, and the redundancy of dual routers at each PoP. The increasing reliability for the core IP routers enables ISPs to exploit an elegant design that leverages the strengths of an increasingly agile optical transport to avoid the high cost of redundant core routers while achieving the same level of availability and performance. However, operational aspects in a network still impact router availability, especially with the inability to seamlessly upgrade the hardware and software of these routers.

Our design carefully ensures that connectivity is maintained upon single failures, including that of a complete core router, and also seeks to avoid congestion and packet loss under such failure conditions. We proposed an architecture that dynamically sizes the capacity of the links between the access-routers and a remote BR. We achieve almost the same level of performability as the baseline dual router design, while achieving a cost savings of approximately 30%.

We recognize the current trend among ISPs to provide higher availability to certain classes of traffic (e.g., VPN traffic), rather than all the traffic flowing over their network. When protection and restoration is provided only to high priority traffic, we see a substantial cost reduction.

We also recognize that almost all cost based design decisions are highly affected by the unit costs of routers and optical network components at any given time. To understand this effect, based on the near term trends of which components are experiencing cost reductions as technology evolves, we evaluate the sensitivity of our design to the relative costs of the different components. We examine a range of reductions in the cost of BRs (all the way down to 10% of current costs), and show that we are still able to achieve worthwhile cost reductions while achieving acceptable performability. Finally, our results are robust to projected increases in network traffic. Our results for the cost reduction for the IP backbone makes a compelling case for our architecture.

Our overall approach should point to a new trend in how backbone networks are architected, achieving a suitable tradeoff between cost and reliability while at the same time ensuring that fast restoration is achieved when a BR fails.

Acknowledgment

This work has greatly benefited from our discussions with Jorge Pastor, Martin Birk, Bob Doverspike, Guangzhi Li, Pete Magill and Kostas Oikonomou (all at AT&T Labs-Research).

References

[1] S. Bailey, V. Gopalakrishnan, E. Mavrogiorgis, J. Pastor, and J. Yates, “Seamless access router upgrades through ip/optical integration,” in *Proc. Optical Fiber Commun. Conf. Exposition (OFC/NFOEC) Nat. Fiber Optic Eng. Conf.*, Los Angeles, CA, March 2011, pp. 1–3.

[2] A. L. Chiu, G. Choudhury, M. D. Feuer, J. L. Strand, and S. L. Woodward, “Integrated restoration for next-generation IP-over-Optical networks,” *J. Lightwave Technol.*, vol. 29, no. 6, pp. 916–924, March 2011.

[3] M. Goyal, K. K. Ramakrishnan, and W. Feng, “Achieving faster failure detection in OSPF networks,” in *Proc. IEEE Int. Conf. Commun. (ICC)*, Anchorage, Alaska, May 2003.

[4] R. Huelsermann, M. Gunkel, C. Meusburger, and D. A. Schupke, “Cost modeling and evaluation of capital expenditures in optical multilayer networks,” *J. Opt. Commun. Netw.*, vol. 7, no. 9, pp. 814–833, September 2008.

[5] , C. R. Kalmanek, S. Misra, and R. Yang, Eds., *Guide to Reliable Internet Services and Applications (Computer Communications and Networks)*. Berlin, Germany: Springer-Verlag, May 2010.

[6] A. Mahimkar, A. Chiu, R. Doverspike, M. Feuer, P. Magill, E. Mavrogiorgis, J. Pastor, S. Woodward, and J. Yates, “Bandwidth on demand for inter-data center communication,” in *Proc. 10th ACM Workshop Hot Topics Netw.*, New York, NY, USA, November 2011, pp. 24:1–24:6, HotNets-X.

[7] A. Mahimkar, H. Song, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and J. Emmons, “Detecting the performance impact of upgrades in large operational networks,” in *ACM Special Interest Group on Data Communication (SIGCOMM) Comput. Commun. Rev.*. New York, NY, USA: ACM, August 2010, vol. 41.

[8] K. Oikonomou, R. K. Sinha, and R. Doverspike, “Multi-layer network performance and reliability analysis,” *Int. J. Interdisciplinary Telecommun. Netw.*, vol. 1, no. 3, pp. 1–30, March 2009.

[9] E. Palkopoulou, “Homing Architectures in Multi-Layer Networks: Cost Optimization and Performance Analysis,” Ph.D. dissertation, TU Chemnitz, Chemnitz, Germany, 2012.

[10] E. Palkopoulou, D. Schupke, and T. Bauschert, “Quantifying CAPEX savings of homing architectures enabled by future optical network equipment,” *Telecommun. Syst.* , vol. 52, no. 2, pp. 1–7, August 2011.

[11] S. Phillips, N. Reingold, and R. Doverspike, “Network studies in IP/Optical layer restoration,” in *Proc. Optical Fiber Commun. Conf. Exhibit. (OFC 2002)*, Mar. 2002, pp. 425–427.

[12] B. Ramamurthy, K. K. Ramakrishnan, and R. K. Sinha, “Cost and reliability considerations in designing the next-generation IP over WDM backbone networks,” in *Proc. 20th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Maui, Hi, August 2011, pp. 1–6.

[13] B. Ramamurthy, R. K. Sinha, and K. K. Ramakrishnan, “Multi-layer design of IP over WDM backbone networks: Impact on cost and survivability,” in *Proc. 9th Int. Conf. Design of Rel. Commun. Netw. (DRCN)*, Budapest, Hungary, March 2013, pp. 60–70.

Byrav Ramamurthy (S’97–A’98–M’03) is a Professor and Graduate Chair in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln (UNL). He received his B.Tech. degree in Computer Science from the Indian Institute of Technology-Madras (IIT-M), India in 1993; and his M.S., and Ph.D. degrees in Computer Science from University of California (UC), Davis in 1995, and 1998, respectively. He has held visiting positions at IIT-M and at the AT&T Labs-Research, New Jersey, U.S.A. He is the author of the book “Design of Optical WDM Networks-LAN, MAN and WAN Architectures,” and a co-author of the book “Secure Group Communications over Data Networks” published by Springer in 2000, and 2004 respectively. He has authored over 125 peer-reviewed journal and conference publications. He serves as an Editor-in-Chief for the Springer Photonic Network Communications journal.

Rakesh K. Sinha (M'98) is a Lead Member of Technical Staff in the Network Evolution Research Department of AT&T Labs-Research. He has broad research interests in the areas of network architecture, design, and optimization. Prior to joining AT&T, he worked at the routing and signaling group for Ciena CoreDirector switches, and before that at the Networking Research Department of Lucent Bell Laboratories. He received his B.Tech. (Computer Science) from Indian Institute of Technology (IIT), Kanpur, and his Ph.D. (Computer Science) from University of Washington, Seattle. He won the AT&T Labs President Excellence Award in 2013 and 2010, and Vice-President Excellence Award in 2012. He received the IFIP International conference PSTV-FORTE best paper award in 2000, and Machtey award for best student paper at IEEE FOCS 1994.

K. K. Ramakrishnan (S'83–M'83–SM'04–F'05) is currently at WINLAB, Rutgers University, New Jersey. Until recently, he was a Distinguished Member of Technical Staff at AT&T Labs-Research, Florham Park, NJ. He joined AT&T Bell Labs in 1994, and has been with AT&T Labs-Research since its inception in 1996. Prior to 1994, he was a Technical Director and Consulting Engineer in Networking at Digital Equipment Corporation. Between 2000 and 2002, he was at TeraOptic Networks, Inc., as Founder and Vice President. Dr. Ramakrishnan is an AT&T Fellow, recognized for his fundamental contributions on communication networks, and lasting impact on AT&T and the industry, including his work on congestion control, traffic management, and VPN services. He is an IEEE Fellow, recognized for his work on congestion control. His work on the "DECbit" congestion avoidance protocol has been recognized as one of the significant contributions published in ACM Sigcomm, and received its Test of Time Paper Award in 2006. He has published nearly 200 papers, and has more than 125 patents issued in his name. He received his MS from the Indian Institute of Science (1978), and MS (1981) and Ph.D. (1983) in Computer Science from the University of Maryland, College Park, USA.