Design and Analysis of Optical Flow-Switched Networks

Guy Weichenberg, Vincent W. S. Chan, and Muriel Médard

***Abstract*—In our previous work [Chan *et al.*, "Optical flow switching," in *BROADNETS 2006*, pp. 1–8; Weichenberg *et al.*, "Cost-efficient optical network architectures," in *ECOC 2006*, pp. 1–2; Weichenberg *et al.*, "On the throughput-cost tradeoff of multi-tiered optical network architectures," GLOBECOM '06, pp. 1–6], we presented optical flow switching (OFS) as a key enabler of scalable future optical networks. We now address the design and analysis of OFS networks in a more comprehensive fashion. The contributions of this work, in particular, are in providing partial answers to the questions of how OFS networks can be implemented, how well they perform, and how their economics compare with those of other architectures. With respect to implementation, we present a sensible scheduling algorithm for inter-metropolitan-area-network (inter-MAN) OFS communication. Our performance study builds upon our work in *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 84–101, 2007 and Weichenberg *et al.*, "Performance analysis of optical flow switching," presented at the IEEE International Conference on Communications, Dresden, Germany, June 14–18, 2009, and includes a comparative capacity analysis for the wide area, as well as an analytical approximation of the throughput-delay trade-off offered by OFS for inter-MAN communication. Last, with regard to the economics of OFS, we extend our previous work from ECOC 2006 and GLOBECOM '06 in carrying out an *optimized* throughput-cost comparison of OFS with other prominent candidate architectures. Our conclusions indicate that OFS offers a significant advantage over other architectures in economic scalability. In particular, for sufficiently heavy traffic, OFS handles large transactions at far lower cost than other optical network architectures. In light of the increasing importance of large transactions to communication networks, we conclude that OFS may be crucial to the future viability of optical networking.**

***Index Terms*—Optical network; Network architecture; Optical communications**.

I. Introduction

Optical networking is unfolding as a two-generation story. The initial impetus behind optical networking was the prospect of tapping the vast usable bandwidth of optical fiber—roughly 30 THz—to meet increasing telephony traffic demands. First-generation optical networks of the 1970s and 1980s thus employed optical fibers as replacements for copper links, but otherwise maintained traditional architectures that were tailored to the use of electronic networking components. Informational bottlenecks were thereby preempted at transmission links, but still loomed at network nodes whose operations were constrained by the speed of electronics.

Second-generation optical networks, which began emerging in the 1990s, employ optical networking devices—in addition to fiber—in novel network architectures to mitigate electronic bottlenecks arising from steep growth in data application traffic. Data traffic served by these networks, in addition to being characterized by orders of magnitude more aggregate volume, is characterized by detailed statistics (e.g., governing transaction length and burstiness) and quality of service demands that are quite different from those of the telephony traffic served by first-generation optical networks. Among these is the self-similarity of traffic arising (most likely) from the heavy-tailed nature of data transactions [1–5]. This changing nature of network traffic, coupled with the novel properties and cost structures of optical networking devices vis-à-vis electronic networking devices, have called for a fundamental rethinking of optical network architecture. In response, measured architectural advancements in the wide-area network (WAN) environment occurred, reducing the cost per transmitted bit by exploiting wavelength division multiplexing (WDM) in conjunction with optical am-

Manuscript received November 30, 2008; revised April 1, 2009; accepted June 6, 2009; published July 31, 2009 (Doc. ID 113585).

The authors are with the Claude E. Shannon Communication and Network Group Research Laboratory of Electronics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA (e-mail: gew@mit.edu).

Digital Object Identifier 10.1364/JOCN.1.000B81

1943-0620/09/030B81-17/$15.00

plification and switching. Nevertheless, the end-user's access to the vast core network bandwidth has been restrained by the lag in architectural innovation in the metropolitan-area network (MAN) and access environments. To a significant extent, then, the future economic viability of optical networking hinges on cost-effective access to core network bandwidth.

In this work [6,7], we further develop optical flow switching (OFS) [8,9] as an attractive candidate network architecture that will help support future traffic growth by providing this desired cost-effective access for end users with large bandwidth demands. OFS is an end-to-end transport service, in that it directly connects source and destination end users through the access network, MAN, and WAN. It is, moreover, intended for users with large transactions (i.e., those that can fully utilize a wavelength channel for hundreds of milliseconds or longer), which are expected to contribute increasingly to future traffic volume. Furthermore, OFS can be readily implemented with today's device technology [10], since it possesses a simple, all-optical data plane that is separated from its electronic control plane. In addition to improving the quality of service for its direct end users, OFS has the additional important benefit of lowering access costs for all users by relieving WAN routers of the onerous burden of serving large transactions.

This work is organized as follows. In the remainder of this section, we provide an overview of OFS. In Section II, we address OFS in the context of the wide area via a comparative analysis of network capacity with other prominent transport architectures. In Section III, we outline a simple and sensible scheduling algorithm for OFS networks and then proceed with an approximate throughput-delay performance analysis of OFS networks. In Section IV, we introduce the notion of cost via an approximate capital expenditure (CapEx) model. This enables us to carry out an approximate throughput-cost comparison of OFS with other prominent candidate architectures in the context of both homogeneous and hybrid architectures. We conclude this work in Section V.

A. Overview of Optical Flow Switching

OFS is an end-to-end, all-optical transport service that provides end users with cost-effective access to the core network bandwidth by means of exploiting the complementary strengths of optics and electronics [8–11]. In a sense, OFS may be viewed as a generalization to end users of the optical bypass feature of generalized multiprotocol label switching (GMPLS).[1] This generalization, however, presents a host of new challenges at multiple layers of network architecture. Moreover, as we shall see, the extension of optical bypass renders OFS a sensible architecture only for large transactions, for otherwise the network management burden and cost of end-user equipment required to set up an end-to-end, all-optical connection for a small transaction would outweigh the benefits of OFS. Thus, OFS is an architecture that is envisioned to serve large bandwidth transactions exclusively, while comparatively small bandwidth transactions will more efficiently be served by architectures such as electronic packet switching (EPS) or GMPLS.

In OFS, end users request long-duration (i.e., hundreds of milliseconds or longer), end-to-end lightpaths by communicating, via an EPS control plane, with scheduling nodes assigned to their respective MANs. These scheduling nodes, in turn, coordinate transmission of data across the WAN in the EPS control plane. Owing to the intractable complexity of coordinating transmission among many OFS users, the control plane for OFS is envisioned to have a mixture of centralized processes that occur on coarse time scales, and distributed processes that occur on fine time scales. In particular, centrally computed candidate routes for WAN reconfiguration will be disseminated to scheduling nodes in of the order of seconds or minutes (or even longer). The scheduling of individual transactions may thus occur in a distributed fashion with exchange of little information for coordination and with physical layer reconfiguration times in the metro area of the order of milliseconds. The access environment, in contrast, is envisioned to have a broadcast structure, with access to it granted by the end-to-end scheduling algorithm.

In OFS, it is assumed that the smallest granularity of bandwidth that can be reserved across the core is a wavelength. Motivated by the minimization of network management and switch complexity in the network core, transactions are served as indivisible entities. That is, data cells comprising a transaction traverse the network contiguously in time, along the same wavelength channel (assuming no wavelength conversion), and along the same spatial network path. This is in contrast to EPS, where transactions are broken up into cells, and these cells are switched and routed through the network independently. In the event that several end users have moderately-sized transactions that are not sufficiently large to warrant their own wavelength channels, they may concatenate their data for transmission across the WAN via dynamic broadcast group formation, albeit at the delay of aggregating these transactions.

Note that in OFS networks, unlike packet-switched networks, all queuing of data (in addition to admission control) occurs at the end users, thereby obviating the need for buffering in the network core. A core

[1]Note that, in addition to optical bypass, GMPLS includes electronic processing capability at routers.

node is thus equipped with a bufferless optical cross connect (OXC). The elimination of buffering at OFS nodes presents a significant scalability advantage, since the queueing subsystem of EPS routers is becoming their major bottleneck as the number of ports and line rates increase [12]. This absence of buffering, however, coupled with the requirement of serving transactions as indivisible blocks of data, renders the efficient utilization of network resources more difficult. Mixing transactions with different quality of service requirements—for example, best-effort traffic that can be preempted by higher-priority traffic—may provide significant advantages in this regard.

II. Optical Flow Switching in the Wide Area: Comparative Capacity Analysis

In this section, we briefly review our work in [13], in which we employed a framework based on network capacity to analyze the performance of optical network architectures in the wide area. Roughly speaking, we define network capacity as the set of exogenous traffic rates that can be stably supported by a network under its operational constraints. Our metric of network capacity is particularly relevant to the wide area because, owing to the high cost of supporting transport traffic, capacity is a precious commodity in the WAN.

Our work constitutes a best-case throughput comparison among optical network architectures, in that (i) a tolerance to unbounded delay is implied, and (ii) any capacity inefficiencies arising from coupling with MAN architectures are neglected. Along with our results for OFS networks, we address the capacity of packet-switching (i.e., EPS and optical packet switching), optical circuit switching (OCS), and optical burst switching (OBS) networks. Each transport architecture's physical and operational properties impose constraints on its capability for logical topology reconfiguration, naturally leading to different capacity performances.

A. Summary of Results

1) Packet-Switched Networks: A packet-switched network is an interconnection of routers, which we model as an interconnection of cell-based, input-queued switches that make scheduling decisions in a distributed fashion. Transport along links is carried out in optical fiber by using WDM, and switching/routing functions at network nodes are carried out in the electronic domain (i.e., EPS) or optical domain (i.e., optical packet switching).[2] Packet-switched architectures were found to be optimal from a capacity perspective, in that the capacity region of such a network is given simply by its admissibility constraints ([13], Theorem 2.1), which ensure that each channel in the network is offered a rate of traffic less than its channel capacity.

2) Optical-Flow and Optical-Circuit Switching: OCS is a wide-area optical network architecture in which WAN ingress routers assemble packets into bursts and centralized, advanced scheduling sets up lightpaths for these bursts between edge routers. This is in contrast to OFS, where lightpaths are established from end user to end user. Nevertheless, from a wide-area perspective, OFS and OCS networks are similar in that they do not employ buffering in the network core and are therefore amenable to the same approach to characterizing their capacity regions. The capacity region of an OFS or OCS network is obtained simply by time sharing over all feasible configurations of the network ([13], Theorems 2.2 and 2.3). Our analysis in arriving at this result was constructive in that algorithms were outlined that achieve the capacity limits. Unfortunately, these algorithms are generally NP-complete, and will therefore be difficult to implement for inter-MAN OFS communication.

3) Optical-Burst Switching: Like OCS, OBS is a wide-area optical network architecture in which WAN ingress routers assemble packets into bursts that are destined for WAN egress routers. From the perspective of the wide area, most OBS networks can be viewed as incarnations of OFS networks in that they lack buffering capability in the core of the WAN and in that they require bursts to be served as indivisible entities. However, because they employ random access instead of scheduling,[3] OBS networks are generally characterized by nonzero burst blocking probabilities. Specifically, the fact that bursts may require retransmission can lead to instability on an individual link, even if the offered traffic is admissible. Furthermore, the lack of coordination among core links implies that resources are wasted if they are consumed by bursts that are eventually discarded. For these two reasons, OBS networks are generally incapable of achieving rate stability within the OFS capacity region ([13], Corollary 2.1). The degree to which the capacity of an OBS network differs from that of the analogous OFS network depends on traffic statistics and on network architecture parameters such as burst aggregation and retransmission policies.

In summary, under the assumption of an identical physical topology, the capacity region of packet-switched architectures dominates that of OFS/OCS, and the capacity region of OFS/OCS dominates that of OBS. These differences in capacity performance arose

[2]Although at present the capabilities of electronic logic greatly exceed those of optical logic, we do not draw a distinction between networks employing electronics versus those employing optics for logic because we are interested in the fundamental capacity limits of packet-switched networks.

[3]This is true of all variations of OBS, except for wavelength-routed OBS [14]. This version of OBS is essentially OCS, in that advanced scheduling sets up lightpaths for bursts.

because of the benefits of core buffering and scheduling. Ultimately, however, a network should not be judged on performance alone, but rather on the performance–cost trade-off it presents to the end user. Thus, the most useful comparison would include a cost model for each of the candidate networks—the subject of Section IV.

III. Optical-Flow Switching Scheduling and Performance Evaluation

In this section, we focus our attention on the algorithmic implementation and performance evaluation of inter-MAN OFS communication. (The work of [15] addresses intra-MAN OFS communication.) We begin by addressing the scheduling algorithm, which arbitrates access to resources in OFS networks. The underpinnings of our proposed algorithm are that WAN wavelength channels are a precious resource and should therefore be efficiently utilized and that traffic in the WAN will be sufficiently heavy and smooth that a quasi-static logical topology is reasonable. Using our proposed algorithm, we then conduct an approximate throughput-delay analysis of OFS networks.

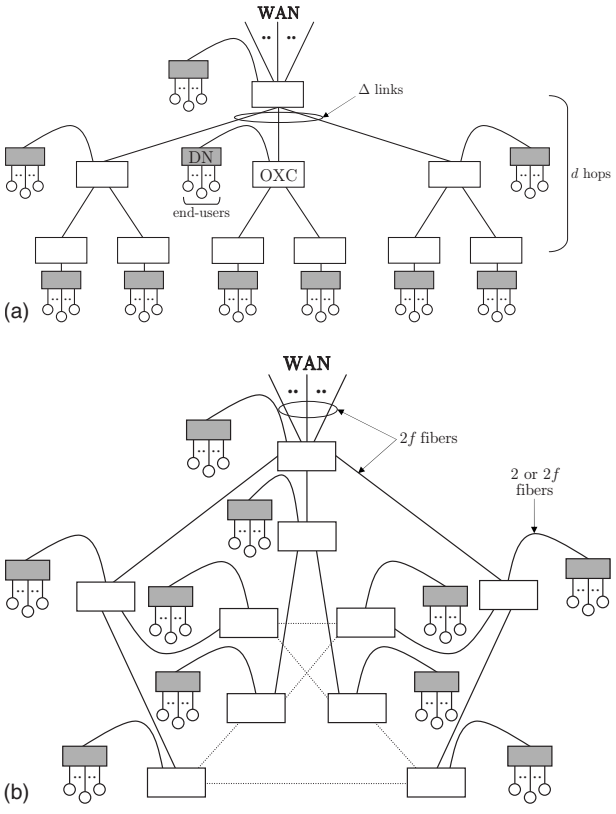


Fig. 1. Example of an OFS MAN based on a Moore graph (Petersen graph) with $\Delta=3$ and $d=2$. A MAN node (white box) comprises an OXC with one or more access DNs (gray boxes) connected. Optical amplifiers are not shown. (a) Embedded tree portion of the OFS MAN. (b) Mesh OFS MAN based on the Petersen graph. Note that the tree topology in (a) is embedded with this topology. Fiber links *not* in the embedded tree are shown as dotted lines.

A. Modeling Assumptions

1) Network Topology and Other Physical Layer Issues: In our network model, a single WAN connects n_w MANs, all of which employ OFS. As drawn in Fig. 1, a MAN is connected to the WAN via a MAN node residing at the MAN–WAN interface. The wavelength channels provisioned for inter-MAN OFS communication reside within f fibers in each direction connecting this node to the rest of the WAN. Note that these fibers may carry traffic from other transport mechanisms besides OFS (e.g., EPS).

An OFS MAN node comprises an OXC with direct connections to adjacent MAN nodes as well as one or more access networks based on optical distribution network (DN) architectures.[4] We let $\tilde{n}_a$ denote the total number of such DNs per MAN . The bidirectional links forming these connections are actually implemented with two contradirectional fiber links, as is done in practice.

While, in principle, the mesh topologies underlying such MANs may be arbitrary, we shall assume that they are based on Moore graphs[5] [e.g., Fig. 1(b)], because topologies based on this family of graphs lend to cost-effective MAN architectures [20]. In reference to Fig. 1, we let Δ denote the number of bidirectional links connecting the root node to other MAN nodes.

Within the physical topology of each MAN, we assume the existence of an embedded regular tree topology with the root node located at the WAN edge [see Fig. 1(a)]. Under normal operating conditions, inter-MAN traffic is assumed to be carried solely on the fibers of the links in the embedded tree, whereas the fibers on the links outside the embedded tree are assumed to carry only intra-MAN traffic. However, in the event of network element failures in the MAN, or significant deviations from expected traffic—considerations beyond the scope of this work—it may be necessary to reroute inter-MAN traffic outside the embedded tree.

Since intra- and inter-MAN OFS traffic could coexist on the same fiber in the embedded tree, we shall assume that w_t channels on each fiber are (quasi-statically) allocated for inter-MAN communication. This separation between wavelength channels for inter- and intra-MAN communication is made for analytical tractability, but may also prove to be sensible in implementing real networks, since it enables simpler resource scheduling, albeit with a potential performance penalty.

Since WAN wavelength channels are precious net-

[4]The DN architectures that we envision for OFS employ optical amplification *within* the DN via multiple segments of erbium-doped fiber remotely pumped by a common laser. We refer the reader to [16], chap. 3, for further details.

[5](Generalized) Moore graphs are discussed in [16–19].

work resources, the scheduling algorithm employed should ensure that they are efficiently utilized. For our scheduling algorithm to do this in a manner that is not computationally prohibitive, we shall require that, for each WAN wavelength channel provisioned for inter-MAN OFS communication, there exists a dedicated wavelength channel in each link of the embedded tree in both source and destination MANs. In the absence of wavelength conversion capability at the MAN–WAN interface, wavelength continuity between WAN and MAN wavelength channels must be respected.

With respect to the number of fibers connecting each DN to its parent MAN node, we consider two extreme cases: (i) two fibers, one in each direction, and (ii) $2f$ fibers, f in each direction, such that there is one-to-one correspondence to the $2f$ fibers connecting the MAN to the WAN. Clearly, the performance corresponding to the second case will be better; but, as we shall see, the performance margin is not great under expected future network dimensions (i.e., number of nodes, volume of traffic, etc.). Moreover, the case of two fibers per DN is attractive in that it is a simpler and more scalable design: less hardware at end users is required, and no modifications to this hardware are required as the number of fibers f in the MAN increases. We will also consider the role of wavelength conversion between each DN and its parent MAN node.

2) Traffic: We will confine our attention in this work to serving the inter-MAN traffic demand. We will assume uniformity in inter-MAN traffic demands, as well as uniformity in DN traffic demands. With w_m denoting the number of wavelength channels on which each MAN communicates with every other MAN, the former assumption implies that

$$w_m \leqslant \frac{fw_t}{n_w - 1},$$

with the exact value of w_m dependent on the efficiency of wavelength reuse achieved by the routing and wavelength assignment in the WAN.

OFS traffic is assumed to be generated at end users such that the aggregate traffic generated for a particular source–destination MAN pair arrives according to a Poisson process with rate λ_m. This Poisson assumption is reasonable for commercial networks, since the superposition of many stationary, identical, and independent point processes—which are reasonable models themselves for individual flow sources—is well known to converge to a Poisson point process. In the event that there do not exist sufficiently many flow sources to multiplex, OFS may not be an appropriate architecture.

The duration of flows are modeled as identical and independently distributed random variables with probability density function $p_L(\cdot)$ and kth moment $\overline{L^k}$. Last, we consider only unicast transactions, although we recognize the increasing importance of multicast transactions, particularly for video content distribution.

B. Scheduling Algorithm for Inter-MAN Communication

As discussed previously, we focus on networks for which there exists significant multiplexing of flows in each MAN, resulting in appreciable statistical smoothing on wavelength channels [21]. Such smoothing of aggregated traffic renders a quasi-static WAN logical topology sensible for serving this traffic. That is, changes to the WAN logical topology will be required on time scales that are of the order of many flows. Consequently, in the algorithm design for the scheduling of individual flows it may be assumed that the wavelength channels provisioned for inter-MAN communication are static. *This allows us to significantly decouple the scheduling resources among pairs of communicating MANs*. Our simple scheduling algorithm, in particular, reserves end-to-end optical paths for flow transmission via a sequential reservation process in which wavelength channels in the MAN and WAN are reserved first, followed by parallel reservations of the source and destination DN wavelength channels. Because wavelength channels in DNs are not heavily loaded, the latter reservation step should entail very little contention, thereby ensuring efficient end-to-end scheduling.

1) Scheduling Algorithm Description: In the following description of the scheduling algorithm, we assume that each DN is connected to its MAN by two fibers. (The case of $2f$ fibers is a straightforward simplification of the following algorithm.) In addition, we neglect the possibility of transmitter and receiver collisions that arise when two or more flows simultaneously require an end user's transmitter or receiver, respectively. This is a reasonable assumption when each end user transmits flows only occasionally, which we assume to be the case. We now illustrate the scheduling algorithm by stepping through the process by which an end-to-end all-optical path is established for the transmission of a particular flow.

Consider a flow that is generated at an end user residing in DN D_s within MAN M_s and that is destined for an end user residing in DN D_d within MAN M_d. As soon as this flow is ready for transmission, the source end user sends a primary request r_w to the scheduling node associated with M_s, requesting an end-to-end all-optical path for its flow transmission.

At a MAN's scheduling node, there exist $n_w - 1$ first-

in first-out queues, one queue corresponding to every possible MAN destination. Each queue can be thought of as the queue for an $M/G/w_m$ queueing system, in that the w_m wavelength channels dedicated to transmission from M_s to M_d eventually serve the primary requests waiting in it. After it arrives at M_s's scheduling node, r_w is placed at the end of the queue associated with M_d, which we denote $Q_{M_s}^{M_d}$. Once r_w reaches the head of $Q_{M_s}^{M_d}$, r_w's flow is assigned to wavelength channel ω, the first of the w_m wavelength channels dedicated to transmission from M_s to M_d to have the primary request it is serving depart. After this wavelength channel assignment is made, an all-optical path is established on wavelength channel ω from the edge of D_s to the edge of D_d, passing through M_s, the WAN, and M_d. (Such a path is guaranteed to exist since there are $2f$ fibers within each link on this path, one of which has a dedicated ω wavelength channel for communication from M_s to M_d.) Now, in order to reserve the single outgoing ω wavelength channel in D_s and the single incoming ω wavelength channel in D_d, two additional secondary requests, r_s and r_d, respectively, are sent. During this process, r_w remains at the head of $Q_{M_s}^{M_d}$.

Secondary request r_s joins the end of the source secondary queue associated with D_s's ω wavelength channel, denoted $\hat{Q}_{M_s}^{D_s}(\omega)$, which is physically located in M_s's scheduling node; and secondary request r_d joins the end of the destination secondary queue associated with D_d's ω wavelength channel, denoted $\bar{Q}_{M_d}^{D_d}(\omega)$, which is physically located in M_d's scheduling node. These queues contain secondary requests to use the ω wavelength channel on D_s's outgoing fiber and D_d's incoming fiber, respectively. When r_s and r_d each reach the heads of their respective secondary queues (which we assume to be first-in first-out), they each notify both M_s's and M_d's scheduling nodes. As soon as M_s's and M_d's scheduling nodes have received *both* notifications, they instruct the source and the destination end users, respectively, to begin flow transmission immediately. After the flow transmission is complete, r_w, r_s, and r_d depart their queues.

C. Performance Analysis

In the following performance analysis, we neglect propagation delay of requests and transactions, as such delays are small compared with the duration of flows. For brevity, we omit derivations of expressions and refer the interested reader to [16], chap. 4.

In our description of the scheduling algorithm for inter-MAN communication in Subsection III.B.1, we mentioned that primary requests for a source–destination MAN pair may be modeled as customers of an $M/G/w_m$ queueing system with arrival rate λ_m. The service time in this model is the time spent by a primary request r_w at the head of its primary queue, which includes flow transmission time in addition to the time spent reserving wavelength channels in the source and destination DNs (in the case of two fibers per DN). Unfortunately, there is no exact solution for the $M/G/w_m$, except in the special cases of exponential service times, $w_m=1$, or $w_m=\infty$. We must therefore resort to an approximate performance analysis of an OFS network.

To this end, we randomly split the Poisson process of intensity λ_m representing the flow arrivals of each source–destination MAN pair into w_m child Poisson processes of intensity

$$\lambda_c = \frac{\lambda_m}{w_m},$$

one child Poisson arrival process for each wavelength channel dedicated to the source–destination MAN pair. We also replace queue $D_{M_s}^{M_d}$ in MAN M_s by w_m independent, parallel queues:

$$Q_{M_s}^{M_d}(\omega_1), Q_{M_s}^{M_d}(\omega_2), \ldots, Q_{M_s}^{M_d}(\omega_{w_m}),$$

each corresponding to a wavelength channel dedicated to the source–destination MAN pair and accepting primary requests from the corresponding child Poisson process.

To compute the expected queueing delay of this system we shall eventually apply the Pollaczek–Khinchin (P-K) formula to the single server queue $Q_{M_s}^{M_d}(\omega)$ accepting primary requests from flows generated in M_s, destined for M_d, and employing wavelength channel ω. The arrival rate to this queue is λ_c, and we shall let X denote the service time of a primary request in this queue with the kth moment $\overline{X^k}$. In the case of $2f$ fibers per DN, we simply have $X=L+\tau$, where L is the flow transmission time and τ is the hardware reconfiguration time, since no additional time is required for reserving resources at the source and destination DNs. However, in the case of two fibers per DN, X includes not only the sum of the flow transmission time L and hardware configuration time τ, but also any time spent at the head of queue $Q_{M_s}^{M_d}(\omega)$ reserving source and destination DN resources. We therefore generally express the average service time $\bar{X}$ as

$$\bar{X} = \bar{L} + \bar{Y} + \tau \approx \bar{L} + \bar{Y}, \tag{1}$$

where $\bar{Y}$ is the average time spent at the head of the primary queue reserving resources in both the source and destination DNs and is equal to zero in the case of $2f$ fibers per DN. The approximation for the average service time neglects the hardware reconfiguration time, because in OFS's most sensible regime of operation, as characterized in the next section, $\bar{L}+\bar{Y} \gg \tau$. More concretely, $\bar{L}+\bar{Y}$ is, at the very least, of the order

of hundreds of milliseconds when OFS is economically viable, whereas τ is likely to be of the order of ten milliseconds.

Recall that, in the case of two fibers per DN, after a primary request reaches the head of $Q_{M_s}^{M_d}(\omega)$, secondary requests are sent to each of $\hat{Q}_{M_s}^{D_s}(\omega)$ and $\bar{Q}_{M_d}^{D_d}(\omega)$, where D_s and D_d are the source and the destination DNs of the flow, respectively. At each of these two secondary queues there could be up to $f-1$ other secondary requests associated with other destination and source MANs, respectively. However, since flows are equally likely to be generated at or destined for each of the DNs in a MAN, the probability that a primary request at $Q_{M_s}^{M_d}(\omega)$ generates a secondary request to a particular $\hat{Q}_{M_s}^{i}(\omega)$ or $\bar{Q}_{M_d}^{j}(\omega)$ is $1/\tilde{n}_a$. The arrival rate of secondary requests to a secondary queue contributed from each of the f contending primary queues is $\lambda_c/\tilde{n}_a$, for an aggregate arrival rate at each secondary queue of $f\lambda_c/\tilde{n}_a \ll 1$. Now, for a fixed aggregate arrival rate of $f\lambda_c/\tilde{n}_a$, as f and $\tilde{n}_a$ become proportionately large—corresponding to a large WAN and large MANs, respectively—the arrival process of secondary requests to each secondary queue is known to converge to a Poisson process. We thus model the arrival process to each secondary queue as a Poisson process of rate $f\lambda_c/\tilde{n}_a$, with the approximation becoming increasingly accurate as f and $\tilde{n}_a$ become large.

We now address the calculation of the moments of Y. Recall that Y is the time spent at the head of the primary queue reserving resources in both the source and destination DNs and is thus the maximum of the two queueing delays experienced by the two peer secondary requests. Since the service time of each secondary request already enqueued at a secondary queue is itself coupled to the state of the queue in which its peer secondary resides, the characterization of Y is difficult. Thus, given the Poisson assumption of secondary request arrivals to secondary queues, we derive upper and lower bounds for $\bar{Y}$ and $\overline{Y^2}$, which will be used in conjunction with the P-K formula to obtain optimistic and pessimistic approximations, respectively, for the expected queueing delay experienced by a flow. These approximations do not serve as strict bounds, since the Poisson assumption is an approximation.

1) Optimistic Approximation for Primary Request Service Time: We may compute simple lower bounds for $\bar{Y}$ and $\overline{Y^2}$ by viewing the time spent to reserve the source and destination DNs as equal to the time spent by a single secondary request in a single queue with service time drawn from $p_L(\cdot)$. These bounds can be employed to yield the following approximations for the first two moments of the service time in the primary request queue:

$$\bar{X} \approx \bar{L} + \frac{f\lambda_c\overline{L^2}}{2(\tilde{n}_a - f\lambda_c\bar{L})}, \tag{2}$$

$$\overline{X^2} \approx \frac{\tilde{n}_a\overline{L^2}}{\tilde{n}_a - f\lambda_c\bar{L}} + \frac{(f\lambda_c\overline{L^2})^2}{2(\tilde{n}_a - f\lambda_c\bar{L})^2} + \frac{f\lambda_c\overline{L^3}}{3(\tilde{n}_a - f\lambda_c\bar{L})}. \tag{3}$$

We note that the above expressions for $\bar{X}$ and $\overline{X^2}$ include contributions from $\overline{L^2}$ and $\overline{L^3}$, respectively. This (undesirable) proportionality to higher-order moments of L arises from the service time X containing a queueing delay term, itself proportional to the second moment of L. However, the contributions from these higher-order moments of L can be made arbitrarily small by designing the network such that $\tilde{n}_a \gg f\lambda_c\bar{L}$. Since $f\lambda_c\bar{L}$ is interpreted as the load offered to each wavelength channel color in a MAN, designing the network in this way ensures that there is little contention for each wavelength channel in each DN.

2) Pessimistic Approximation for Primary Request Service Time: In order to compute upper bounds for $\bar{Y}$ and $\overline{Y^2}$, we imagine that secondary requests, instead of being sent simultaneously after a primary request reaches the head of its queue, are sent sequentially. Specifically, we assume that the secondary request reserving the destination DN is sent only *after* the secondary request responsible for reserving the source DN reaches the head of its queue in effect reserving the source DN. Thus we have

$$\bar{Y} < \overline{Z_s} + \overline{Z_d},$$

where Z_s is the time spent by a secondary source request in its queue prior to reaching the head of the queue, and Z_d is the time spent by a secondary destination request in its queue prior to reaching the head of its queue. By invoking the Poisson approximation for the arrival process to the secondary destination queue, we may thus treat the queue as an $M/G/1$ queueing system. The first three moments of Z_d may be obtained in a straightforward manner from the Takács recurrence formula. To compute the first two moments of Z_s, we invoke the Poisson approximation for the arrival process of secondary requests to the secondary source queue, and thus treat the queue as an $M/G/1$ queueing system. The pessimistic approximations for the first two moments of the service time in the primary request queue are then given by

$$\bar{X} \approx \bar{L} + \overline{Z_s} + \overline{Z_d}, \tag{4}$$

$$\overline{X^2} \approx \overline{L^2} + \overline{Z_s^2} + \overline{Z_d^2} + 2\bar{L} \times \overline{Z_s} + 2\bar{L} \times \overline{Z_d} + 2\overline{Z_s} \times \overline{Z_d}. \tag{5}$$

As in the optimistic approximation for primary request service time, these expressions for $\bar{X}$ and $\overline{X^2}$ in-

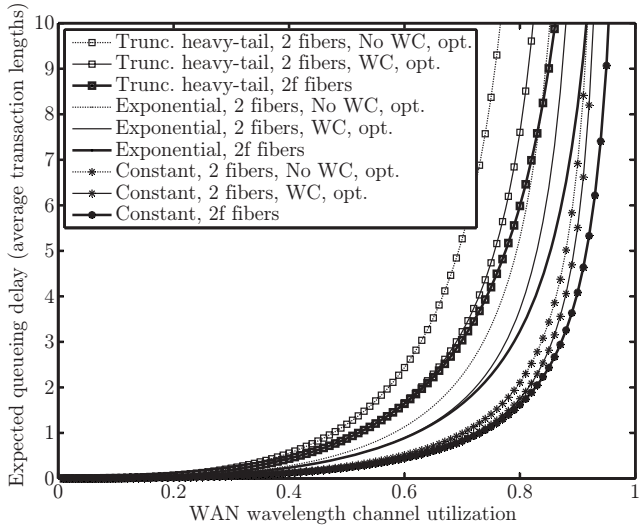


Fig. 2. Expected queueing delay versus throughput for DN with two fibers and $2f$ fibers per DN under three flow length distributions.

clude contributions from higher-order moments of L, which can be made arbitrarily small by designing the network such that $\tilde{n}_a \gg f\lambda_c \bar{L}$.

3) Wavelength Conversion in Distribution Networks: In the presence of wavelength conversion capability in each DN, there is only one secondary source queue and one secondary destination queue associated with each DN. Each of these queues is served by w_t wavelength channels akin to an $M/G/w_t$ queueing system. In this case, the maximum number of contending secondary requests at a queue would be $(n_w - 1)w_m$ rather than f as in the case of no wavelength conversion. However, since there are w_t candidate wavelength channels serving this queue instead of just one, the normalized arrival rate of secondary requests to a secondary queue is again $f\lambda_c/\tilde{n}_a \ll 1$. The convergence of this arrival process to Poisson is faster in this case than in the case of no wavelength conversion owing to the fact that $(n_w - 1)w_m > f$. Furthermore, the expected delay experienced by secondary requests at these secondary queues will be less than in the case of no wavelength conversion owing to the statistical smoothing gain. We refer the interested reader to [16], chap. 4, for further details.

4) Total Expected Queueing Delay: Given the above optimistic and pessimistic approximations for the first two moments of X, we now turn to computing the total expected queueing delay seen by a transaction. To do this, we first invoke the P-K formula with respect to the primary request queue with an additional term reflecting the expected queueing delay experienced by a primary request while at the head of its queue. We then invoke the previous $M/G/k$ approximation to obtain the following approximation of the total expected queueing delay experienced by a flow, including the time spent reserving DN wavelength channels just prior to flow transmission:

$$W \approx \left[\bar{Y} + \frac{\lambda_c \overline{X^2}}{2(1-\lambda_c \bar{X})}\right]\left[\frac{\hat{W}_{M,w_m}(\lambda_m \bar{X}, p_X)}{\hat{W}_{M,1}(\lambda_c \bar{X}, p_X)}\right]. \quad (6)$$

Equations (2) and (3) or (4) and (5), or their wavelength conversion counterparts, may be substituted into Eq. (6), ultimately yielding optimistic and pessimistic approximations of W, respectively.

D. Numerical Results

In Fig. 2, the approximations of the expected queueing delay derived from Eq. (6) are plotted versus WAN wavelength channel utilization for three flow length distributions: constant, exponential, and truncated heavy-tailed, respectively. For each set of flow length distributions, the approximations, which differ only in the way DN reservations are carried out, yield similar performances at low loads, since there is very little contention for DN resources. The performances diverge with increasing traffic load—especially between the case of no wavelength conversion and the other cases—owing to the increasing role of DN reservation time. In comparing performance across distributions, we observe that constant length flows offer the best delay-throughput trade-off, followed by exponentially distributed flows, and then truncated heavy-tailed distributed flows. These results indicate very large flows impeding subsequent flows to the greatest extent in the truncated heavy-tailed distribution and to the least extent in the constant distribution.

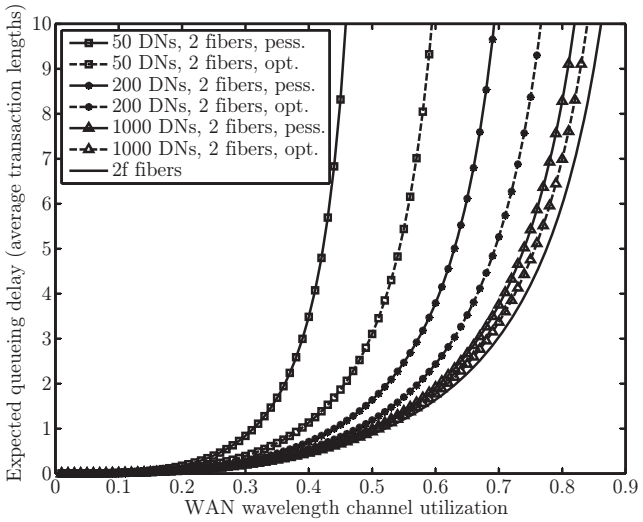


Fig. 3. Expected queueing delay versus throughput for a truncated heavy-tailed flow distribution with different numbers of DNs ($\tilde{n}_a$) per MAN with two fibers per DN and no wavelength conversion. The case of $2f$ fibers per DN is also shown as a performance benchmark.

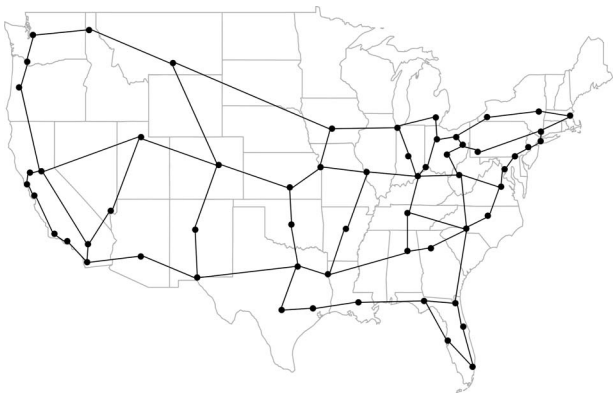


Fig. 4. WAN topology of the U.S. considered throughout this section. Based on [24], Fig. 8.1.

In Fig. 3, we illustrate the impact of the number of DNs per MAN on the delay–throughput trade-off. As expected, the performance of the two fibers per DN case converges to that of the $2f$ fibers per DN case as the number of DNs per MAN increases. This, of course, is because as $\tilde{n}_a$ increases, the amount of traffic per DN decreases, ultimately resulting in less contention for DN resources. We also observe the expected result that the gap between the optimistic and pessimistic approximations for the same $\tilde{n}_a$ narrows as the number of DNs per MAN increases.

The previous numerical results and discussion provide us with justification for narrowing the space of design alternatives in the design of OFS networks. First, equipping each DN with $2f$ fibers for bidirectional communication provides little performance benefit for a large MAN in which there are hundreds of DNs per MAN. This observation, coupled with the aforementioned practical concerns regarding network upgrades, render this design alternative less attractive than the case of two fibers per DN. Wavelength conversion was found to provide moderate performance benefit, with the benefit decreasing for a large MAN. In spite of this performance benefit, use of this technology in individual DNs may be imprudent, as the present-day costs of the relevant technologies are prohibitive.

IV. Throughput-Cost Comparison of OFS With Other Architectures

In this section, we carry out an approximate throughput-cost study to substantiate our claim that OFS is an economically attractive candidate architecture for end users wishing to send large transactions. Our performance-cost study compares OFS to the following prominent optical network architectures: EPS, OCS, GMPLS, and OBS. Since some architectures are applicable only in the wide area (e.g., OCS), we must employ them in conjunction with other architectures in the metro-area and access (i.e., EPS) to enable an end-to-end comparison with OFS. This work represents an extension of our previous work [22,23] in the following two ways: (i) we formulate a more comprehensive cost model employing additional sources of CapEx (although we emphasize that these cost assumptions are approximate); (ii) we optimize the physical layers of the architectures that we compare, in order to obtain a best-case scenario for each architecture.

A. Topology and Traffic Assumptions

In our cost study, we shall assume that all of the architectures considered operate on the same WAN fiber plant topology drawn in Fig. 4. This 60-node network was introduced in [24] as a representative U.S. carrier backbone network. Relevant attributes of this fiber plant topology are listed in Table I. The assumption of a preexisting fiber plant is a good assumption for countries, such as the U.S., that have established telecommunication infrastructures. The WAN source–destination average traffic demands that we consider are uniformly scaled versions of the set employed in [24], chap. 8. This traffic set reflects actual U.S. backbone network traffic and is therefore not uniform all-to-all in nature.

In the metro-area, unlike in the wide area, we do not assume a fixed fiber plant topology over which all architectures operate. Instead, we analytically optimize the fiber topology in accordance with the switching and fiber deployment costs particular to each architecture. We restrict our consideration of fiber plant

TABLE I
WAN Parameters and Values for the Numerical Studies[a]

Parameter	Value
No. of nodes	60
No. of links	77
Average node degree	2.6
Largest node degree	5
Average link length	450 km
Longest link length	1200 km
Optical amplifier spacing	80 km
Number of wavelength channels per fiber link	200
Average length of an end-to-end connection	1950 km
Average no. of hops of an end-to-end connection	4
Average no. of nonnodal regenerations along an end-to-end connection in EPS	0.1
Average no. of regenerations along an end-to-end connection in OCS or OFS	0.3
Capacity efficiency scaling factor	0.95
Line rate	40 Gbits/s

[a]Discussed in Subsections IV.F and IV.G (adapted from [24], Tables 8.1 and 8.2).

TABLE II
MAN PARAMETERS AND VALUES FOR THE NUMERICAL STUDIES (SUBSECTIONS IV.F AND IV.G)

Parameter	Value
No. of nodes	30
Average link length	10 km
OFS hardware reconfiguration time	10 ms
Line rate	40 Gbits/s

topologies to those that are based on regular graphs with nodal symmetry, since such topologies are reasonable models of a real MAN and are more analytically tractable. We found that the family of generalized Moore graphs minimizes the MAN cost, albeit with different dimensions for different architectures ([16], chap. 5). With respect to traffic in the MAN, we assume that intra-MAN traffic is uniform all-to-all, whereas inter-MAN traffic is uniform all-to-one (and one-to-all) to (from) the root node of the embedded MAN tree. Typical values for relevant MAN parameters, which we invoke later in this section, are listed in Table II.

In the access environment, we employ the DN designs detailed in [16], chap. 3, as the basis for both our non-OFS passive optical networks (PONs) and OFS DN. With regard to traffic generated or sunk at access networks, we assume uniformity. However, with respect to individual end-user data rate requirements, we consider both homogeneous and heterogeneous requirements. Typical values for the relevant access network parameters are listed in Table III.

B. Cost Modeling Approach and Assumptions

The relative costs employed in this section, collected from a variety of sources, are organized in Tables IV and V. In this subsection, we address the underlying assumptions in these tables. We reiterate that these cost assumptions are approximate as of today, and are, moreover, subject to some fluctuation with time. They are thus employed to support only high-level claims about the economic viability of OFS.

TABLE III
ACCESS NETWORK PARAMETER VALUES FOR THE NUMERICAL STUDIES (SUBSECTIONS IV.F AND IV.G)

Parameter	Value
Average end-user link length	35 m
End-user duty cycle	0.001
Ratio of PON line rate to WAN line rate	0.25
OFS DN line rate	40 Gbits/s
PON line rate	10 Gbits/s

TABLE IV
RELATIVE COSTS OF NETWORK ELEMENTS FOR BOTH 10 AND 40 GBIT/S LINE RATES[a]

Network Element	Relative Cost 10 Gbits/s	Relative Cost 40 Gbits/s
Tunable medium-reach transceiver	0.3*x*	0.75*x*
Tunable long-reach WDM transceiver	0.4*x*	*x*
Tunable WDM transponder	40*x*	100*x*
Tunable WDM regenerator (with optical line terminal)	56*x*	140*x*
Optical terminal chassis (per wavelength)	2.5*x*	2.5*x*
WAN amplifier and dispersion compensation (per wavelength)	2*x*	3.1*x*
WAN OXC port with amplification and dispersion compensation	8*x*	9.1*x*
WAN router port	120*x*	300*x*
MAN OXC port with amplification	5*x*	5.1*x*
MAN grooming/router port	60*x*	150*x*
OFS scheduler	2*x*	2*x*
Access fiber deployment (per km wavelength)	0.2*x*	0.2*x*
MAN fiber deployment (per km wavelength)	0.2*x*	0.2*x*
WAN fiber deployment (per km wavelength)	0.1*x*	0.1*x*
Access network erbium-doped fiber amplifier pump power (per 100 mW)	2*x*	2*x*

[a]From [24–27]. An optical reach of 2500 km, 200 wavelengths per fiber, and a bidirectional element function are assumed.

1) CapEx Cost Components: Our model focuses on CapEx costs and neglects ongoing operating expense costs, which admittedly constitute a significant portion of a network's cost, and, moreover, differ across architectures. Our cost model addresses CapEx insofar that *major* cost components that differ from architecture to architecture are captured. A potential shortcoming of this cost model—in addition to the omission of operating expense—is the possible neglect of significant sources of cost that are roughly constant across architectures, resulting in an overemphasis on the cost differences among architectures.

TABLE V
COST SCALING PARAMETERS AND VALUES FOR THE NUMERICAL STUDIES (SUBSECTIONS IV.F AND IV.G)

Parameter	Value
Cost ratio of 10 to 40 Gbit/s amplifier and dispersion compensation equipment	0.64
Cost ratio of 10 to 40 Gbit/s electronics	0.4
Cost ratio of 600 to 2500 km optical reach equipment	0.63

2) Fiber Deployment: The cost of deploying fiber depends largely on whether the fiber plant of the network preexists or needs to be augmented. For greenfield networks, where the fiber plant does not preexist, cables, each containing tens of fibers, need to be installed underground. The cost of a fiber link would reflect the material cost of the fiber strand and the aforementioned installation cost—including right-of-way cost—amortized over the number of fiber strands installed. Installation costs also arise in augmenting existing fiber plants; and these costs could be significant for urban networks such as MAN and access networks, as indicated in Table IV. For networks in which the fiber plant preexists and requires no augmentation, however, much of the fiber installation cost is a sunk cost. In either case, the cost of deploying a link c_f is well modeled as a linear function of its length, with a proportionality constant reflecting the various factors discussed above.

3) Line Rate: In our cost study, we have chosen to focus on networks with WAN line rates (i.e., the bit rate of a wavelength channel) of 40 Gbits/s. While we assume that MAN and OFS DN operate at the same line-rate as the WAN, our PON model (i.e., for non-OFS architectures) employs 10 Gbit/s equipment as a means of lowering access network cost. In Table IV, we include relative costs of both 10 and 40 Gbit/s equipment. For fixed optical reach, the costs of amplifiers and dispersion compensation equipment scale with maximum fiber capacity (i.e., number of wavelengths per fiber times the line rate). Specifically, for every doubling in fiber capacity, the costs of amplifiers and dispersion compensation equipment have empirically increased by approximately 25% [24].

4) Optical Reach: In our present study, for architectures employing optical bypass at nodes, we shall assume an optical reach of 2500 km, which was shown to be close to optimal for the network parameters assumed thus far [24]. For the EPS architecture, which entails optical-electrical-optical conversions at each node, we shall assume an optical reach of 600 km, which is equivalent to one third longer than the average link length in Fig. 4.

5) Linearity of Device Costs: Implicit in our cost model is the assumption that the cost of a device with multiple wavelength channel ports scales linearly with the number of such ports. In addition, we shall assume that the cost of a laser pump scales linearly with its output power.[6]

[6]For laser pump powers in excess of hundreds of milliwatts, the cost scaling for a single device may be superlinear. However, an equivalent output power with linear cost may be engineered by cascading several smaller sources.

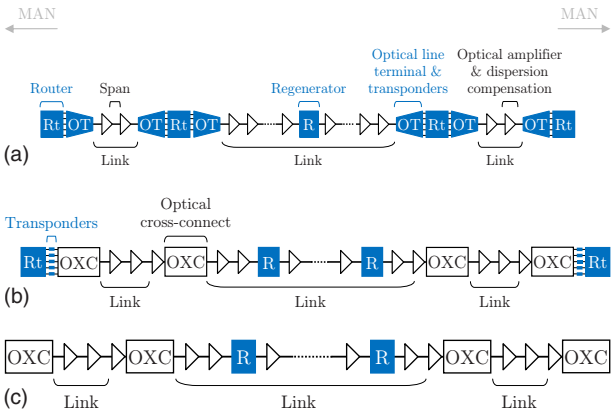


Fig. 5. (Color online) Sample end-to-end WAN connection under EPS, OCS/OBS, and OFS. Electronic networking devices are shown in blue; optical networking devices are shown in black and white.

C. Wide-Area-Network Cost Model

The major CapEx components in the WAN are fiber (i.e., material, trenching, and right-of-way costs), optical amplification, chromatic dispersion compensation, polarization-mode dispersion compensation, and regeneration, as well as switching, routing, and grooming at nodes. In accounting for these costs, we consider a typical bidirectional end-to-end connection in the WAN of wavelength granularity under each architecture (see Fig. 5).

A fair comparison among the architectures must account for their different WAN capacities, as discussed in Section II. Recall that packet-switched architectures (i.e., EPS, optical packet switching) maximize network capacity, whereas architectures employing optical bypass (i.e., OCS, OFS) do not. However, the difference in network capacity among architectures is sensitive to the underlying fiber plant topology. We capture the capacity efficiency of EPS by scaling the total cost of EPS by a (topology-dependent) factor. However, as mentioned earlier in this section, the cost structures of the architectures considered are sufficiently different that our conclusions in this section are relatively insensitive to reasonable values of this parameter.

Beyond fundamental differences in capacity efficiency, implementation issues can impact the throughput achieved under each architecture. Indeed, in Section III, we proposed a *practical* scheduling algorithm for inter-MAN communication that, even in the absence of delay constraints, results in a throughput penalty relative to an optimal, but *infeasible*, scheduling algorithm. Similarly, there are implementation issues at routers in EPS and OCS that, in reality, result in a throughput penalty. Algorithms achieving maximum throughput are too computationally intensive to be executed at routers, and therefore less complex algorithms that perform suboptimally are employed

(e.g., [28]). Our model does not account for the penalty that this entails, so the results of our numerical studies in Subsections IV.F and IV.G are conservative with respect to OFS.

We refer the reader to [16], chap. 5, for the expressions governing the average cost of a bidirectional end-to-end wavelength-granular WAN connection under each architecture.

D. Metropolitan-Area-Network Cost Model

In optimizing the MAN topology, we assume that the number of MAN nodes n_m is a fixed parameter. We assume that intra-MAN OFS traffic is uniform all-to-all of w_u wavelengths between each node pair; and that inter-MAN OFS traffic is represented as all-to-one (and one-to-all) traffic of w_a wavelengths to (from) the root node on the augmented tree portion of the MAN topology. Under this traffic scenario, generalized Moore graphs were shown to minimize network cost ([16], chap. 5).

The major CapEx components in the MAN that we consider are fiber, optical amplification (at optical nodes), and switching, routing, and grooming at nodes. For simplicity, we shall assume a constant MAN fiber link length of l_m, resulting in uniform MAN fiber link costs α. Node infrastructure (e.g., huts) are considered to be a significant expense, should they need to be built, but we again assume this cost to be roughly the same across the different architectures and do not include this in our cost model.

In [16], chap. 5, we show that the node degrees that minimize the total MAN cost are given by

$$\Delta^*_{\mathrm{EPS}} \approx \frac{k \ln n_m [w_u n_m + 2w_a]}{4\alpha l_m} \times \left[\mathcal{W}\left(\sqrt{\frac{k \ln n_m [w_u n_m + 2w_a]}{4\alpha}} \right) \right]^{-2} \quad (7)$$

for the case of electronic (dis)aggregation and by

$$\Delta^*_{\mathrm{OFS}} \approx \frac{k w_u n_m \ln n_m}{4\alpha l_m} \left[\mathcal{W}\left(\sqrt{\frac{k w_u n_w \ln n_m}{4\alpha l_m}} \right) \right]^{-2} \quad (8)$$

for the case of optical (dis)aggregation, provided that

$$2 \leq \{\Delta^*_{\mathrm{EPS}}, \Delta^*_{\mathrm{OFS}}\} \leq n_m - 1,$$

where k is the cost per port. The general behavior of these curves is illustrated in Fig. 6, indicating that, as switching becomes more expensive relative to fiber deployment, it is economically advantageous to route traffic along fewer hops by increasing the node degree versus routing traffic along more hops (corresponding to lower node degree), which requires more switching resources.

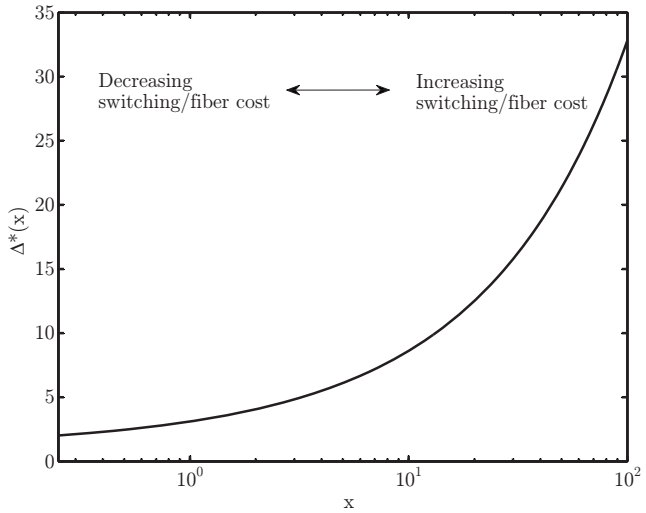


Fig. 6. Generic optimal node degree for generalized Moore graphs: $\Delta^*(x) = x/\mathcal{W}(\sqrt{x})^2$.

We refer the reader to [16], chap. 5, for the expressions capturing the total MAN cost for electronic and optical (dis)aggregation.

E. Access Network Cost Model

In the access environment, the major cost components that we consider are optical amplifier pumps, optical line terminals, transceivers, and fiber. We expect that the cost of shared passive components, including erbium-doped fiber segments, to be insignificant compared with the above cost components, and we may lump the costs of passive components required for each end user (e.g., coupler-based tap) into the transceiver cost.

For PONs, maximizing the number of end users per DN results in the most economical design for PONs. We can optimize PON design in this isolated manner because a PON is only loosely coupled with its adjoining MAN as a result of optical-electrical-optical conversion and buffering at the head-end optical line terminal. We refer the reader to [16], chap. 5, for the details on computing the maximum number of end users per DN.

On the other hand, the optimization of OFS DN is more complicated because of the tight coupling of the access and metro environments in OFS. One of our important observations in Section III was that, for a fixed aggregate amount of inter-MAN OFS traffic, a larger number of DNs over which to distribute this traffic results in better performance. In particular, a larger number of DNs per MAN results in less contention for resources in each DN, thereby resulting in higher WAN wavelength channel utilization, as illustrated in Fig. 3. This ultimately requires fewer provisioned WAN wavelength channels than an otherwise

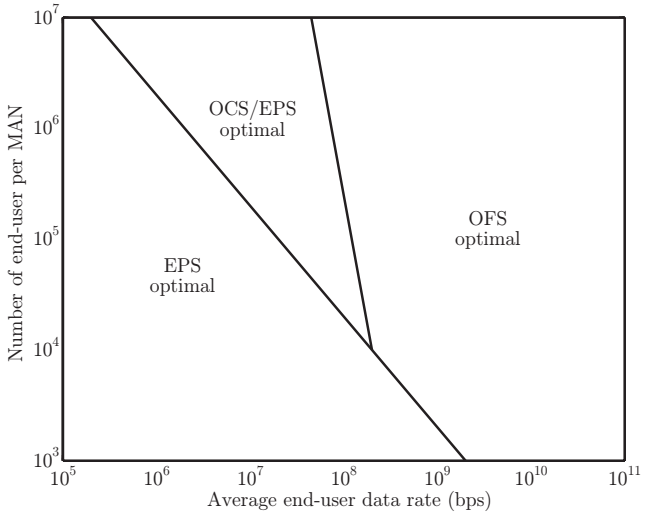


Fig. 7. Minimum-cost architecture as a function of MAN size and average end-user data rate. It is assumed that transactions have a truncated heavy-tailed distribution and that DNs have two fibers and no wavelength conversion.

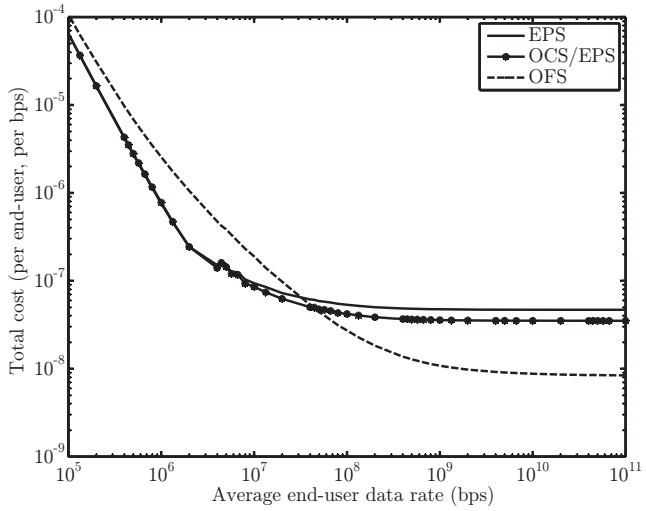


Fig. 8. Normalized total network cost (in the units of x used in Table IV) versus average end-user data rate. It is assumed that each MAN has an end-user population of 10^6, transactions have a truncated heavy-tailed distribution, and DNs have two fibers and no wavelength conversion.

identical scenario with fewer DNs. Therefore, in OFS, it can be economically advantageous to employ more than the minimum number of DNs per MAN, each supporting less than the maximum number of users. This optimization, as well as with the expressions for total DN costs, are contained in [16], chap. 5.

F. Throughput-Cost Comparison of Architectures

We now carry out a throughput-cost comparison of the following three end-to-end network architectures:

- EPS: EPS is used in the wide area, electronic (dis)aggregation is used in the metro area, and PONs are used in the access.
- OCS/EPS: OCS is used in the wide area, electronic (dis)aggregation is used in metro area, and PONs are used in the access.
- OFS: OFS employs the scheduling algorithm proposed in Section III.

These three architectures represent an evolution from electronic to optical switching, from the network core toward the end users at the network edge. To be sure, these three architectures are not exhaustive of the space of architecture alternatives. For instance, in the previous subsections we also described a cost model for OBS, which we do not consider any further in this subsection. The purpose of formulating a CapEx cost model for this architecture was to highlight the fact that it is identical to that of OCS. The critical difference between OBS and OCS lies in how network resources are allocated. OBS generally employs random-access for resource reservation, which is known to result in inferior throughput performance relative to scheduled approaches (Subsection II.A.3). As a result, OBS is guaranteed to appear inferior to OCS in a throughput-CapEx study, and we therefore do not consider this architecture any further.

1) Modeling Assumptions: We shall assume that transaction lengths are drawn from a (truncated) heavy-tailed distribution. Moreover, we shall assume that the average transaction length $\bar{L}$ scales linearly with an end user's average data rate. This is a reasonable assumption, given that the number of transactions transmitted per unit time by a user is roughly constant. Last, we shall assume that equal proportions of traffic generated in a MAN are intra- and inter-MAN in nature.

When comparing the costs of supporting a fixed traffic demand under each architecture, we assume that the underlying networks have been optimized in the manner outlined in the last several subsections. The WAN, however, as we discussed in Subsection IV.A, is assumed to be fixed and based on the fiber plant topology drawn in Fig. 4.

2) Numerical Results: In Fig. 7, we indicate the minimum-cost architecture as a function of number of end users per MAN and average end-user data rate. When aggregate MAN bandwidth demand—given by the product of abscissa and ordinate values—is relatively low, EPS is seen to be the most sensible architecture. Electronic switches and routers, to be sure, are less economically scalable technologies than OXC, but they operate at finer data granularity than OXC. Thus, when aggregate traffic is low, it is wasteful to provision entire wavelength-granular OXC ports that are poorly utilized—which is why EPS is the

minimum-cost architecture in this regime of operation. However, when bandwidth demand between each MAN pair is of the order of multiple wavelengths, optical switching in the WAN is sensible, rendering OCS/EPS the minimum-cost architecture. In fact, under heavy aggregate traffic, the cost difference between EPS and OCS/EPS scales approximately linearly with the product of this aggregate traffic and the difference in cost between a router and OXC port. As aggregate traffic grows even larger, such that the traffic carried on MAN links is of the order of wavelengths, optical switching in the MAN and at the access boundary is most economical, rendering OFS the minimum-cost architecture. Before moving on, a comment on the very large abscissa values in Fig. 7 (and in the other figures in this section) is warranted. Owing to our tolerance to unbounded delay, the abscissa values in our figures are very large. In the presence of realistic delay constraints, the abscissa values at which transitions between optimal architectures occur could be an order of magnitude or more lower.

In Fig. 8, we depict a horizontal cross section of Fig. 7 at a MAN population of 10^6 end users. On the ordinate, we plot total network cost normalized by the number of end users and by the average end-user data rate. The figure indicates that when end users have average data rates below 5×10^7 bits/s, the EPS and OCS/EPS architectures have the lowest normalized cost. Intuitively, this, again, is because relatively little expensive electronic equipment is necessary to support the aggregate traffic in these architectures, whereas in OFS, wavelength-granular optical equipment is wastefully overprovisioned in the MAN—and to a lesser extent in the WAN—along with expensive long-haul transceivers at end users operating at the WAN line rate. Beyond data rates of 5×10^7 bits/s, however, we see that OFS is the most cost-efficient architecture because (i) the aforementioned optical equipment in the WAN, MAN, and DN equipment is better utilized and (ii) this equipment is more economically scalable than analogous electronic equipment.

In summary, we conclude that *OFS is the most cost-scalable architecture of all*, in that its asymptotic normalized cost is several times lower—approximately a factor of four in Fig. 8—than that of competing architectures. As a final remark, we point out that, while the precise values at which transitions in the optimal architecture occur are sensitive to the exact parameter values assumed, the general trends observed are manifestations of present-day cost structures of architectures and their building blocks. Thus, in the absence of disruptive technologies with radically different cost structures, we expect the trends observed in these figures to hold for a reasonable range of parameter values.

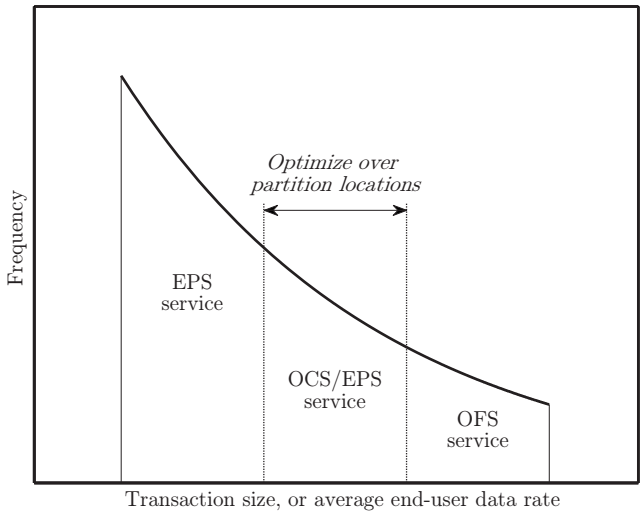


Fig. 9. Partitioning of the truncated heavy-tail distribution into architecture service regions.

G. Hybrid Network Architectures

In the previous subsection, we investigated the throughput-cost trade-offs offered by different *homogeneous* network architectures. In our study, we assumed that a MAN supports a uniform base of end users, each sending transactions with average length $\bar{L}$ proportional to the long-term average end-user data rate. Given that the size of a transaction greatly impacts the efficiency with which it is served by an architecture, hybrid architectures—architectures comprising two or more of the aforementioned homogeneous network architectures—may be economically advantageous to homogeneous structures. We devote our attention in this subsection to investigating this hypothesis.

1) Modeling Assumptions: In this subsection, we shall focus on hybrid architectures in which component subarchitectures operate in parallel with varying degrees of interaction. Since the metro-area and access designs are identical for EPS and OCS/EPS, we shall allow end users belonging to these two architectures to share resources (i.e., wavelength channels, switches, or router ports) in these environments. In the wide area, however, their transport mechanisms differ, but the hybrid architecture they form together resembles GMPLS. Owing to the significant differences between these two architectures and OFS in all three geographic network tiers, OFS end users are allocated their own network resources from end to end. The design of more integrated hybrid architectures may provide better performance-cost trade-offs than those considered here—but not before their many outstanding physical layer and protocol issues are resolved.

Consistently with our discussion of network traffic

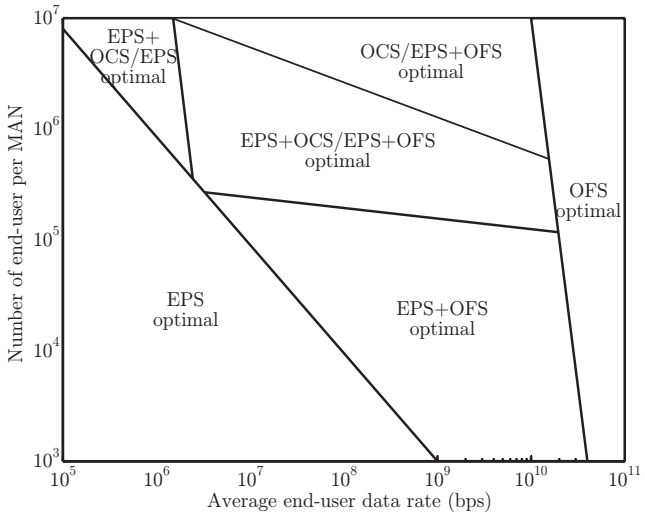


Fig. 10. Minimum-cost hybrid architecture as a function of MAN size and average end-user data rate. It is assumed that average end-user data rates are drawn from a truncated heavy-tailed distribution with initial lower limit 10^3 bits/s and width 10^4 bits/s and that DNs have two fibers and no wavelength conversion.

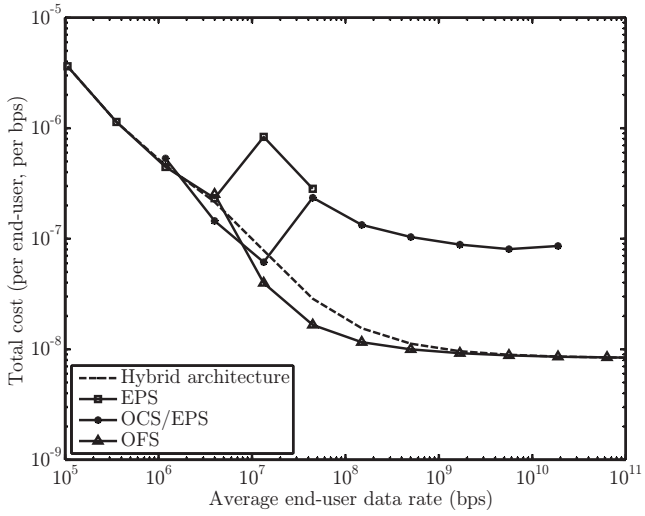


Fig. 11. Normalized cost components of the minimum-cost hybrid architecture (in the units of x used in Table IV) versus average end-user data rate. It is assumed that average end-user data rates are drawn from a truncated heavy-tailed distribution with initial lower limit 10^3 bits/s and width 10^4 bits/s, that each MAN has an end-user population of 10^6, and that DNs have two fibers and no wavelength conversion.

in Section I, we shall assume that the lengths of transactions generated or sunk in a MAN are drawn from a (truncated) heavy-tailed distribution. Motivated by our results in Subsection IV.F, we confine our attention to hybrid architectures in which transactions—or equivalently end users—are partitioned for service as shown in Fig. 9. Transactions (equivalently, end users) are partitioned into three contiguous regions such that the transactions (end users) in each partition are served exclusively by the indicated architecture. The optimal hybrid architecture is defined as the architecture that minimizes total network cost by judicious positioning of the inner two dotted boundary lines in Fig. 9. We point out that this is an idealized formulation, in that a given end user's transactions may, in reality, be more efficiently served with more than one subarchitecture.

2) Numerical Results: In Fig. 10, we indicate the minimum-cost hybrid architecture as a function of the number of end users per MAN and average end-user data rate. As in Fig. 7—and for the same reasons discussed in Subsection IV.F.2—when aggregate MAN traffic is relatively low the homogeneous EPS architecture is optimal, and when aggregate MAN traffic is relatively high the homogeneous OFS architecture is optimal. However, for intermediate aggregate traffic, we observe that hybrid architectures become preferable to homogeneous architectures. When the number of end users per MAN is small or moderate and the average end-user data rate is moderate, aggregate MAN traffic is in the lower range of intermediate values, and EPS is a component of the optimal hybrid architecture—the other component being OFS—since it serves intermediate and low-end end users most economically. When the number of end users per MAN is moderate or large and the average end-user data rate is low or moderate, a hybrid architecture employing GMPLS (i.e., EPS and OCS/EPS) is optimal, with OFS possibly serving the high-end end users. When the number of end users per MAN is large and average end-user data rates are moderate, there is sufficient traffic generated from intermediate to low-end end users (i.e., multiples of wavelengths) such that electronic switching in the WAN is no longer economically viable, and OCS/EPS together with OFS constitutes the minimum-cost hybrid architecture.

In Fig. 11, we depict a horizontal cross section of the minimum-cost hybrid architecture in Fig. 10 at a MAN population of 10^6 end users. On the ordinate, we plot (sub)architecture cost normalized by the number of end users served by the (sub)architecture and the average data rate of end users served by the (sub)architecture. Consistent with our results in Subsection IV.F, the asymptotic normalized costs are lowest for OFS, followed by OCS/EPS, and then EPS. The dashed curve, which represents the normalized cost of the entire hybrid architecture, is essentially a weighted average of the three solid subarchitecture curves. At low average end-user data rates the (dashed) hybrid architecture curve follows the (squares) EPS curve, and for high average end-user data rates the (dashed) hybrid architecture curve follows the (dots) OFS curve, indicating the dominance of these architectures at these two extremes.[7]

[7]As in the case of homogeneous architectures, the curves in our plots of this section exhibit sharp bumps. This arises from the inte-

V. Conclusion

In this work, the OFS architecture was shown to be a cost-effective way of serving large transactions, which are increasing in importance with every passing day. Moreover, only modest technological hurdles exist before OFS can be implemented: the required device technology is currently available, and the remaining algorithmic and protocol challenges should be manageable.

To be sure, the work presented here does not constitute an exhaustive examination of the OFS architecture. Beyond immediate extensions to this work, and the concurrent work carried out in [15,29], there exist avenues of research in further developing and analyzing OFS. In terms of further developing the OFS architecture, mechanism(s) ensuring reliable end-to-end communication with OFS are needed. Since OFS is a scheduled, flow-based transport architecture, there is no need for the congestion and flow control mechanisms of TCP, which are sure to impede the efficiency with which data is transmitted through the network. Instead, a lightweight transport layer protocol, akin to UDP, should be used in conjunction with an error-checking mechanism, such as LDPC. Another interesting avenue of future research is less related to how OFS itself may be implemented than to how it may efficiently coexist with other subarchitectures in a hybrid network setting. Naturally, it is preferable to integrate component subarchitectures as tightly as possible, as this enables better utilization of network resources. In practice, however, the challenge will be to design algorithms and protocols that are capable of computing and disseminating network resource allocation information within the time frames required for agile resource reconfiguration at the architecture level.

Acknowledgments

The authors thank Eric A. Swanson for his contributions in improving this paper.

References and Notes

[1] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 835–846, 1997.

[2] M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Comp. Commun. Rev.*, vol. 27, pp. 5–23, 1997.

[3] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Trans. Netw.*, vol. 5, no. 1, pp. 71–86, 1997.

grality constraints on certain network parameters: occasionally, for small perturbations in parameter values, the integrality constraints induce sudden changes in the composition of the optimal hybrid architecture.

[4] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Netw.*, vol. 2, no. 1, pp. 1–15, 1994.

[5] V. Paxson and S. Floyd, "Wide-area traffic: the failure of Poisson modeling," *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, 1995.

[6] Part of this work appeared in , "Performance analysis of optical flow switching," presented at IEEE International Conference on Communications (ICC), Dresden, Germany, June 14–18, 2009, and also in Ref. [7].

[7] G. Weichenberg, V. W. S. Chan, and M. Médard, "Throughput-cost analysis of optical flow switching," in *Optical Fiber Communication Conf.*, OSA Technical Digest (CD), Washington, DC: Optical Society of America, San Diego, CA, March 22, 2009, paper OMQ5.

[8] V. W. S. Chan, G. Weichenberg, and M. Médard, "Optical flow switching," in *3rd Int. Conf. on Broadband Communications, Networks and Systems, 2006. BROADNETS 2006*, San Jose, CA, Oct. 1–5, 2006, pp. 1–8.

[9] N. M. Froberg, S. R. Henion, H. G. Rao, B. K. Hazzard, S. Parikh, B. R. Romkey, and M. Kuznetsov, "The NGI ONRAMP test bed: reconfigurable WDM technology for next generation regional access networks," *J. Lightwave Technol.*, vol. 18, no. 12, pp. 1697–1708, 2000.

[10] B. Ganguly and V. W. S. Chan, "A scheduled approach to optical flow switching in the ONRAMP optical access network testbed," *Optical Fiber Communications Conf.*, A. Sawchuk, ed., vol. 70 of OSA Trends in Optics and Photonics, Washington, DC: Optical Society of America, 2002, paper WG2.

[11] V. W. S. Chan, K. L. Hall, E. Modiano, and K. A. Rauschenbach, "Architectures and technologies for high-speed optical data networks," *J. Lightwave Technol.*, vol. 16, no. 12, pp. 2146–2168, 1998.

[12] S. Kumar, J. Turner, and P. Crowley, "Addressing queuing bottlenecks at high speeds," in *13th Symp. on High Performance Interconnects*, Stanford, CA, Aug. 17–19, 2005, pp. 107–113.

[13] G. Weichenberg, V. W. S. Chan, and M. Médard, "On the capacity of optical networks: a framework for comparing different transport architectures," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 84–101, 2007.

[14] E. Kozlovski, M. Düser, A. Zapata, and P. Bayvel, "Service differentiation in wavelength-routed optical burst switched networks," in *Optical Fiber Communications Conf.*, A. Sawchuk, ed., vol. 70 of OSA Trends in Optics and Photonics, Washington, DC: Optical Society of America, 2002, paper ThGG114.

[15] B. Ganguly, "Implementation and modeling of a scheduled optical flow switching (OFS) network," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2008.

[16] G. Weichenberg, "Design and analysis of optical flow switched networks," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2009.

[17] A. J. Hoffman and R. R. Singleton, "On Moore graphs with diameters 2 and 3," *IBM J. Res. Dev.*, vol. 4, pp. 497–504, 1960.

[18] R. R. Singleton, "On minimal graphs of maximum even girth," *J. Comb. Theory*, vol. 1, pp. 306–322, 1966.

[19] M. Sampels, "Vertex-symmetric generalized Moore graphs," *Discrete Appl. Math.*, vol. 138, pp. 195–202, 2004.

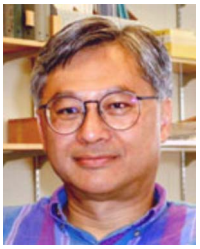
[20] C. Guan, "Cost-effective optical network architecture—a joint optimization of topology, switching, routing and wavelength assignment," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2007.

[21] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "The effect of statistical multiplexing on the long-range dependence of Internet packet traffic," Bell Laboratories Tech. Rep., 2002, http://cm.bell-labs.com/stat/doc/multiplex.pdf.

[22] G. Weichenberg, V. W. S. Chan, and M. Médard, "Cost-efficient optical network architectures," in *European Conf. on Optical*

Communications, 2006. ECOC 2006, Cannes, France, Sept. 24–28, 2006, pp. 1–2.
[23] G. Weichenberg, V. W. S. Chan, and M. Médard, "On the throughput-cost tradeoff of multi-tiered optical network architectures," in *IEEE Global Telecommunications Conference, 2006. GLOBECOM '06*, San Francisco, CA, Nov. 27–Dec. 1, 2006, pp. 1–6.
[24] J. M. Simmons, *Optical Network Design and Planning*, New York, NY: Springer Science + Business Media, 2008.
[25] S. Sengupta, V. Kumar, and D. Saha, "Switched optical backbone for cost-effective scalable core IP networks," *IEEE Commun. Mag.*, vol. 41, no. 6, pp. 60–70, 2003.
[26] N. S. Patel, "Optical networking: historical perspectives and future trends," MIT Lecture Notes, 6.442 Optical Networks, 2008.
[27] E. A. Swanson, MIT Lincoln Laboratory, Lexington, MA 02173, personal communication, May 2008.
[28] S. T. Chuang, S. Iyer, and N. McKeown, "Practical algorithms for performance guarantees in buffered crossbars," *Proc. IEEE INFOCOM 2005. 24th Annual Joint Conf. of the IEEE Computer and Communications Societies*, Miami, FL, March 13–17, 2005, vol. 2, pp. 981–991.
[29] A. R. Ganguly, "Optical flow switching architectures for ultra-high performance applications," M.Eng. dissertation, Massachusetts Institute of Technology, Cambridge, MA, in preparation.

Guy Weichenberg recently (2009) received the Ph.D. from MIT, where his research interests included network theory and optical communications. He also received the S.M. from MIT, and the B.A.Sc. from the University of Toronto. He is currently with the RAND Corporation.



Vincent W. S. Chan, the Joan and Irwin Jacobs Professor of EECS, MIT, received his B.S. (71), M.S. (71), E.E. (72), and Ph.D. (74) degrees in electrical engineering from MIT. From 1974 to 1977, he was an assistant professor of electrical engineering at Cornell University. He joined MIT Lincoln Laboratory in 1977 and had been Division Head of the Communications and Information Technology Division until becoming the Director of the Laboratory for Information and Decision Systems (1999–2007). This year he helped formed and is currently a member of the Claude E. Shannon Communication and Network Group at the Research Laboratory of Electronics of MIT. In July 1983, he initiated the Laser Intersatellite Transmission Experiment Program and in 1997, the follow-on GeoLITE Program. In 1989, he formed the All-Optical-Network Consortium among MIT, AT&T, and DEC. He also formed and served as Principal Investigator for the Next Generation Internet Consortium, ONRAMP among AT&T, Cabletron, MIT, Nortel, and JDS, and a Satellite Networking Research Consortium formed between MIT, Motorola, Teledesic, and Globalstar. He has been serving as the Editor-in-Chief of the IEEE Optical Communications and Networking Series, JSAC Part II that has transitioned to a new IEEE/OSA journal, the *Journal of Optical Communications and Networking*, in March 2009. He is a Member of the Corporation of Draper Laboratory, member of Eta-Kappa-Nu, Tau-Beta-Pi, and Sigma-Xi, a Fellow of the IEEE, and the Optical Society of America.

Muriel Médard is a Professor in electrical engineering and computer science at MIT. She was previously an Assistant Professor in the Electrical and Computer Engineering Department and a member of the Coordinated Science Laboratory at the University of Illinois Urbana–Champaign. From 1995 to 1998, she was a Staff Member at MIT Lincoln Laboratory in the Optical Communications and the Advanced Networking Groups. Professor Médard received B.S. degrees in electrical engineering and computer science and in mathematics in 1989, a B.S. degree in humanities in 1990, an M.S. degree in electrical engineering 1991, and a Sc.D. degree in electrical engineering in 1995, all from the Massachusetts Institute of Technology (MIT), Cambridge. She has served as an editor of several IEEE journals. She received the IEEE Leon K. Kirchmayer Prize Paper Award 2002, the Best Paper Award at the Fourth International Workshop on the Design of Reliable Communication Networks (DRCN 2003), the Information Theory Society/ Communications Society Joint Best Paper Award 2009, and the William R. Bennett Prize in the Field of Communications Networking 2009. She received an NSF Career Award in 2001 and was cowinner of the MIT 2004 Harold E. Edgerton Faculty Achievement Award. She was named a 2007 Gilbreth Lecturer by the National Academy of Engineering. Professor Médard is a member of the Board of Governors of the IEEE Information Theory Society and a Fellow of IEEE.