Optical Flow Switching

Vincent W. S. Chan, *Fellow, IEEE, OSA,* Guy Weichenberg, *Student Member, IEEE,*
and Muriel Médard, *Senior Member, IEEE*
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

(*Invited paper*)

***Abstract* – In this work, we evaluate an attractive candidate for optical network data transport: Optical Flow Switching (OFS). We describe the operation and implementation of the architecture, characterize its capacity region and capacity-cost tradeoff, and compare it to other prominent optical network architectures.**

I. Introduction

In the early days of the Internet, the most precious resource was long-haul transmission capacity. Thus, the packet switching architecture was designed to use this resource as efficiently as possible (Figure 1). With the ubiquitous deployment of WDM in core networks in the last few years, long-haul transmission is now cheaper than router switching at core nodes; this is because the rate of decrease in cost per unit of optical fiber capacity is well below that of Moore's Law, which governs the cost of electronic ICs from which electronic routers and switches are built. To keep reaping the economic benefits of optical networking technologies, long duration (quasi-static) optical circuit switching – GMPLS – will be used in the next cycle of build-up in the core. In access networks, electronic routers and switches are currently used for aggregation of bursty computer communications data. With the inexorable increase in communications bandwidth demand, such architectures will soon impede the trend of cost reductions, at which point optical access and some form of optical switching and routing for large transactions will be required. In this paper, we propose a transport architecture – Optical Flow Switching (OFS) – that will exploit optical switching, routing, and transport technologies to continue the lowering of costs faster than Moore's Law. This will ultimately allow access of high-rate services to the masses sooner than current trends would otherwise allow.

II. The OFS Architecture

A. Overview of OFS

In OFS, users request end-to-end lightpaths for long duration (>100 msec) transactions (Figure 2). If possible, each transaction is scheduled to meet a time delay requirement over a finite time horizon. If the lightpath request is granted, the resources dedicated to the transaction are relinquished for other users once the transfer is complete. Since it may not be possible to honor every lightpath request, there will be a trade-off among three network performance metrics: delay, blocking probability, and wavelength utilization. This trade-off lends to multiple service qualities coexisting in the same network, which may be implemented by simply incorporating proper cost functions into the scheduling protocol.

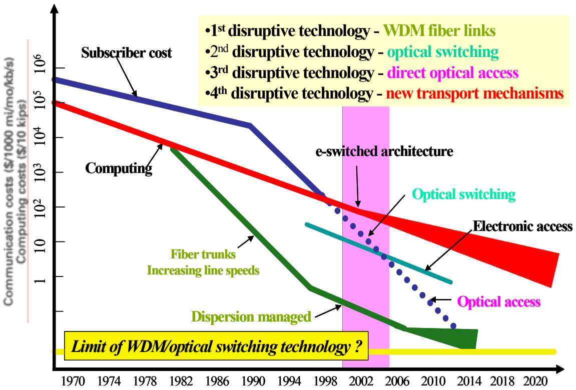

Figure 1. Network cost reduction as results of disruptive optical network technologies.

In order to schedule data transmission across the WAN, users communicate via an electronic control plane with the scheduling processors assigned to their respective MANs. These scheduling processors, in turn, coordinate transmission of data across the WAN in an electronic control plane. In OFS, it is assumed that the smallest granularity of bandwidth that can be reserved across the core is a wavelength. For implementation simplicity, we further assume that wavelength conversion is not used in the network, although this assumption may be relaxed. In the event that several single users have transactions which are not sufficiently large to warrant their own wavelength channels, they may multiplex their data for transmission across the WAN via dynamic broadcast group formation.

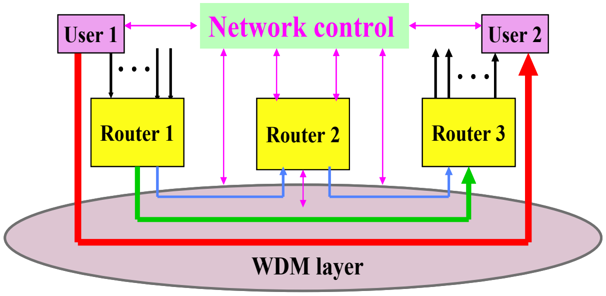

Figure 2. Transport mechanisms (EPS in blue, GMPLS in green, OFS is from user to user in red).

Motivated by the minimization of network management and switch complexity in the network core, flows are serviced as indivisible entities. That is, data cells comprising a flow traverse the network contiguously in time, along the same wavelength channel, and along the same spatial network path. This is in contrast to packet switched networks, where transactions are broken up into constituent cells, and these cells are switched and routed through the network independently. Note that in OFS networks, unlike packet switched networks, all queuing of data occurs at the end users, thereby obviating the need for buffering in the network core. Core nodes are thus equipped with bufferless optical cross-connects (OXCs). OFS is a centralized transport architecture in that coordination is required for logical topology reconfiguration. However, OFS traffic in the core will likely be efficiently aggregated and sufficiently intense to warrant a quasi-static logical topology that changes on coarse time scales. Hence, the centralized management and control required for OFS is not expected to be onerous. The network management and control carried out on finer time-scales will be distributed in nature in that only the relevant ingress and egress access networks will need to communicate.

B. Detailed description of OFS

The key to high utilization of backbone wavelength channels – the costly part of the network due to the need for optical amplifiers and dispersion management – is statistical multiplexing of large flows from many users in a scheduled fashion. In a large access network, there is less aggregation closer to the user and hence there is less cost sharing of expensive (usually active) optical components. Figure 3 shows a hierarchical access network comprising a feeder network and many distribution networks, passive optical LANs, fed by the feeder network.

Physical topology and other physical layer design

Whereas many users share the feeder network, distribution networks will serve relatively few users.

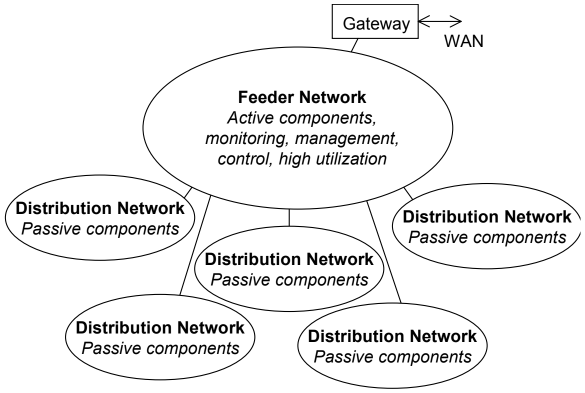

Figure 3. Hierarchical optical access network.

Thus, motivated by cost, one may reasonably take the approach of allowing costly active components in the feeder network, but avoiding these components in the distribution networks. Figure 4 shows a feeder network with active access nodes and bus/tree type distribution networks with passive optical components.

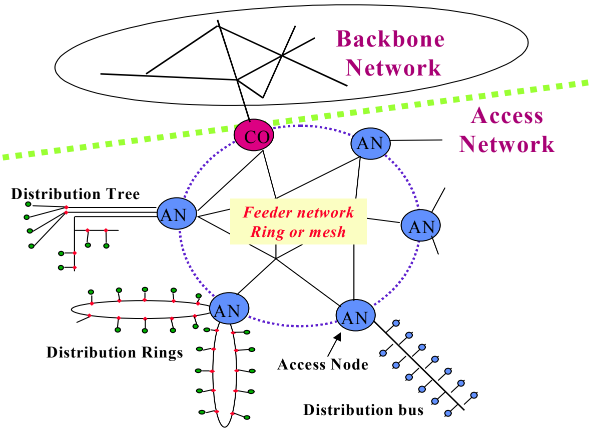

Figure 4. Physical architecture of access network.

The feeder network itself can be a ring or mesh. It has active optical and electrical components and supports lightpath reconfigurability and dynamic wavelength broadcast group formation in millisecond time frames. Since there is already significant traffic aggregation in the feeder, important core network technology building blocks such as OXCs, tunable filters, optical amplifiers, and add-drop multiplexers may economically be used.

The distribution network serves as the interface between the network and its users. Since fiber at the end user is less of a shared medium, bandwidth is in great abundance and may therefore be prudently traded for a low cost distribution network. For the physical topology of the distribution network, all-optical rings are not viable since feedback generates oscillations. The topologies that can be used are buses and trees (and stars which are degenerate trees). The number of users supported by a single distribution network should be the largest number possible without active amplification in this part of the network. Passive buses and trees, however, impose severe limits on the number of users that can be supported by these distribution networks. Furthermore, fiber nonlinearities impose an upper limit on the amount of optical power that can be launched into a distribution network from the access node. The number of users at ~2.5 Gbps that can be supported by a passive optical network is theoretically between 10 and 800, but realistically, the range will be approximately 20-100. In some applications, such as large computer file transfer, users can be active with less than 1% duty cycle when they are in session. If only 10% of users on the network are in session at any given time, then to gain some degree of efficient statistical multiplexing in a wavelength channel in a fiber, the number of users accessing the wavelength channel should be on the order of 1000. Thus, we need an access network architecture that is better than that of conventional passive optical networks.

Remotely-pumped EDFAs offer the possibility of greatly increasing the number of users supportable. The pump sources can be maintained at an access node on the feeder network simplifying deployment, management and control, and also conforming to the desire to have no component in the distribution network that requires electrical supply. In the design of such a distribution network, sections of erbium-doped fiber are inserted at suitable locations along the length of undoped single-mode fiber, as shown in Figure 5. The architecture consists of a bus network with a chain of remotely pumped EDFAs. A tributary passive star or tree network is fed by a tap after the gain stage. This has the additional advantage that the EDFAs are concentrated near the access node, and the shorter fiber length to them helps reduce absorption of the pump power by the fiber. Towards the middle to the end run of the bus when amplified spontaneous emission (ASE) is the dominant noise, the tapped power must increase (linearly) with the tap number. This is why the amplification process cannot be used to extend the range indefinitely. However, the maximum number of users supportable is over 10,000 and thus more than adequately meets the aforementioned requirement of ~1000 users for efficient statistical multiplexing.

Dynamically arranged broadcast groups within the access network are an additional key feature of our envisioned network architecture. They allow for arbitrary connectivity of users, efficient sharing of wavelength channels over time, flexibility in setting up OFS connections, and ease of multicast and broadcast services. Dynamic broadcast groups can be formed via assignment of specific wavelengths to a group of users. Access to these wavelengths is supported in the distribution network via broadcast, but switching and wavelength conversion in the feeder network can be used to form the group. To further localize power distribution for high-rate applications and a high degree of aggregation, remotely optically switched MEMs can be used to channel wavelengths in passive tributaries (Figure 6).

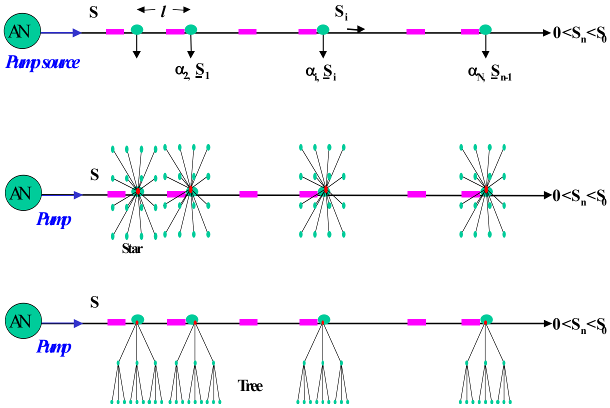

Figure 5. Remotely-pumped distribution network feeding end-users, passive stars, and trees.

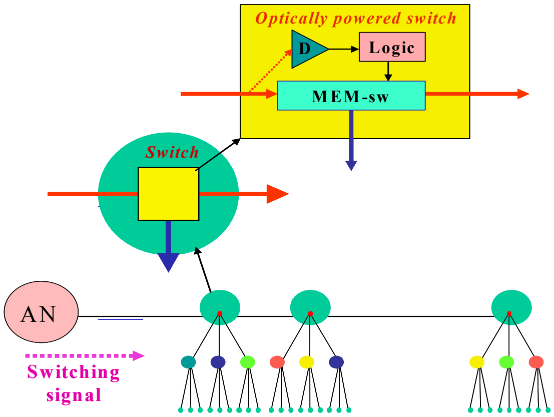

Figure 6. Remotely optically switched access network.

Higher layer design

Figure 7 shows how end-to-end, bursty, all-optical flows can be routed from the distribution network across a WAN to a peer access network. As discussed previously, access nodes can be fully reconfigurable, with programmable ADMs and OXCs.

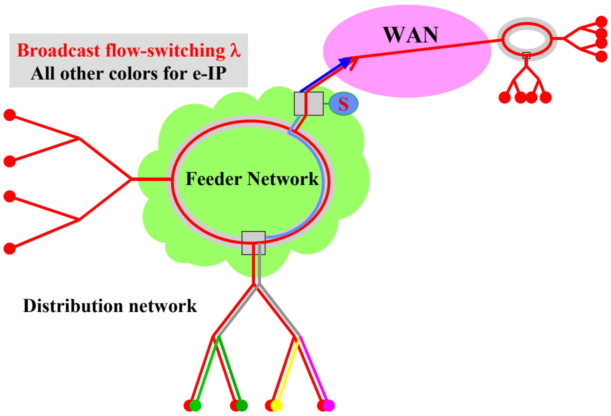

Figure 7. Routing of flows across a multi-tiered network.

Each access node has three entities that perform three functions, but they can be incorporated into one processor unit instead of being physically separated (Figure 8). The first is an entity that runs a media access control (MAC) protocol for users in the distribution network to gain access to resources in the access node, the feeder network, and the distribution network. In a hybrid network architecture, the user can choose between IP packet service or OFS service and this entity coordinates this choice. The second entity is a router that accepts IP packets and acts as an aggregator for IP packet traffic. It also conveys network resource requests from users to the first and third functional processor blocks. This third block is a network configuration manager and an OFS scheduler which manages optical network functions such as formation of dynamic broadcast groups to gain access to WAN wavelengths for OFS, as well as path set-up and tear-down for OFS flows.

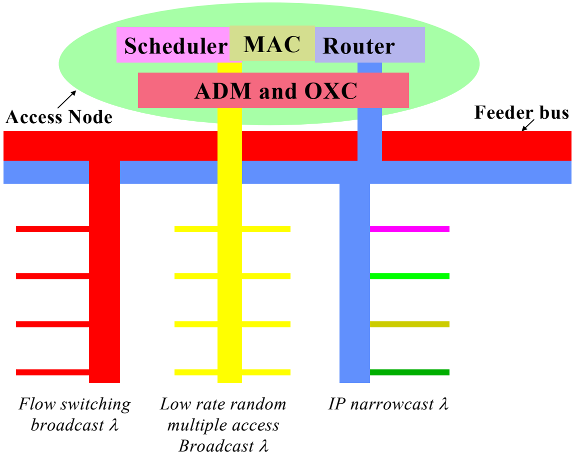
Scheduler
MAC
Router
ADM and OXC
Access Node
Feeder bus
Flow switching broadcast λ
Low rate random multiple access Broadcast λ
IP narrowcast λ

Figure 8. Access node architecture.

Since OFS is a scheduled flow-based transport architecture, there is no reason to use TCP for congestion and flow control. A lightweight transport layer protocol should be used instead to reduce cost. Good candidates for checking errors in OFS transactions are Low Density Parity Check Codes (Figure 9). If any errors are detected, a request for retransmission of the whole file is done via feedback to the transmitter. With the low error rates of current optical transmission technology, this is not inefficient in terms of throughput, but does require an additional delay of at least one transaction transmission time. For delay critical transactions, forward error correcting codes can be used, albeit at the expense of more computational resources for decoding.

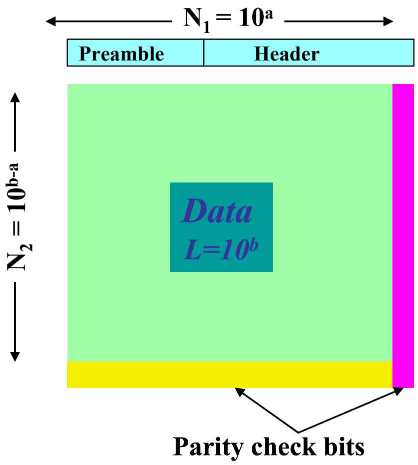

Figure 9. Parity check code for OFS transactions.

III. Other Optical Network Architectures

A. Optical Burst Switching and Tell-and-Go

There is a great deal of variability in how Optical Burst Switched (OBS) networks are designed. The particular OBS model that we focus on in this work, though simple, captures the spirit of the general OBS transport philosophy [3-6]. One variation of OBS that we consider – although only in Section V – is based on the Tell-and-Go (TaG) protocol for ATM networks, in which end users act as sources for bursts of data [7, 8]. In the more common OBS architectures, data cells are assembled at access nodes into collections of cells known as bursts. We assume, as in OFS, that wavelength conversion is not used in the network core. Burst contention, owing to the absence of scheduling in the core, is resolved by burst discardment. In more sophisticated OBS networks, burst discardment is a last resort, should other ways of contention resolution, such as segmentation or deflection routing, fail [9, 10].

B. Electronic Packet Switching and Optical Packet Switching

A packet switched network is an interconnection of routers which we model as an interconnection of cell-based, input queued (IQ) switches which make scheduling decisions in a distributed fashion at every time slot. Transport along links is carried out in optical fiber using WDM, and switching and routing functions at network nodes are carried out in the electronic domain or optical domain, corresponding to Electronic Packet Switching (EPS) or Optical Packet Switching (OPS), respectively. Note that employing electronics to carry out operations at network nodes implicitly equips nodes with wavelength conversion capability, whereas the use of optics does not. Since optical buffers are not readily available in OPS networks, fiber delay lines and even deflection routing are sometimes used as substitutes. However, the performance of these schemes will deviate from what is given below for classical packet switching.

C. Electronic Packet Switching / Generalized Multi-protocol Lambda Switching

Electronic Packet Switching / Generalized Multi-protocol Lambda Switching (EPS/GMPLS) is a network architecture that is conceptually intermediate to OFS and EPS. Specifically, the MAN design in EPS/GMPLS is identical to that of EPS, while the WAN design is similar to OFS in that all-optical transmission along dedicated wavelength channels passing through OXCs is employed in order to circumvent electronic processing at intermediate nodes. At the interface of the MAN and the WAN there exist ingress and egress routers that are responsible for assembling and disassembling large blocks of data, respectively.

IV. Capacity analysis

In this section, we present a framework for comparing transport architectures for core networks on the basis of network capacity, which we define as the set of exogenous traffic rates that can be stably supported by a network under its operational constraints. We are able to suppress access network architecture in our *capacity analyses*, as they are independent of detailed traffic statistics (except for OBS, as discussed below). Our performance metric of network capacity is particularly relevant to core networks because, owing to the high cost of supporting long-haul traffic, capacity is a precious commodity in the core (but not necessarily in the access network). In addition to OFS, we examine two other prominent candidate core architectures: OPS/EPS and OBS.

In our model, each fiber can support a maximum of w unit capacity active wavelength channels, and we assume that each node is equipped with $t \le w$ tunable transceivers per fiber. In the case of OPS/EPS, note that $t = w$ necessarily; that is, there exists a (fixed) transceiver for each of the w wavelength channels. We neglect propagation delay, and we only consider unicast transmission. Finally, we assume that time is slotted.

A *routing* is a list of transaction types and possible paths, along with a set of probabilities that data cells follow these paths in the network. If each transaction type is restricted to follow exactly one path from source to destination, then the routing is a *simple* routing. A system of queues is *rate-stable* if, with probability 1, the long-term rate at which traffic enters the network's queues equals the long-term rate at which traffic departs the network's queues. The *capacity region* of a network is the closure of the set of exogenous traffic rate vectors for which the system of queues in the network is rate-stable for some routing and for some scheduling policy. We emphasize that the capacity region of a network is not tied to a particular routing. Rather, it is the closure of the collection of achievable traffic rates taken over the set of all routings. A set of exogenous traffic rates is *admissible* if a routing exists for which every channel in the network is offered a rate of traffic which is less than or equal to its channel capacity.

B. Capacity of OFS networks

In OFS networks, data is scheduled to traverse the network from source to destination without being buffered at intermediate nodes. A *feasible network state* is a set of transaction or flow types that can be simultaneously served, while respecting the connectivity and operational constraints of the underlying network. A *flow incidence vector* is a binary vector $(e_1, e_2, ..., e_F)$, where F is the number of flow types in the network, in which $e_i = 1$ if flow i is being serviced and $e_i = 0$ otherwise. The OFS capacity region is then characterized as follows [11]:

Theorem 1 *The capacity region of an OFS network is the convex hull of the union (over all simple routings) of the flow incidence vectors corresponding to feasible network states.*

C. Capacity of OBS and EPS/OPS

<u>OBS</u>

OBS networks can be viewed as incarnations of OFS networks in that they lack buffering capability in the core, and that they require bursts to be served as indivisible entities. However, owing to the fact that they employ random-access instead of scheduling, OBS networks are generally characterized by nonzero burst blocking probabilities. Furthermore, the lack of coordination among core links implies that resources are wasted if they are consumed by bursts that are eventually discarded. This leads to the following:

Corollary 1 *The capacity region of an OBS network is* *bounded by the convex hull of the union (over all simple routings) of the flow incidence vectors corresponding to feasible network states.*

The degree to which the capacity of an OBS network differs from the capacity region of the analogous OFS network depends upon the traffic statistics, and on network architecture parameters such as burst aggregation and retransmission policies.

<u>EPS/OPS</u>

In [11], the EPS/OPS capacity region is characterized as follows:

Theorem 2 *The capacity region of an OPS network, with or without wavelength conversion, is the admissible rate region for the network. That is, OPS networks achieve the maximum capacity region.*

The capacity regions for EPS/OPS, OFS, and OBS, are notionally shown in Figure 10. Note that the higher capacity transport mechanism may not be the best since, if cost is included in the comparison, lower capacity mechanisms may actually be preferred.

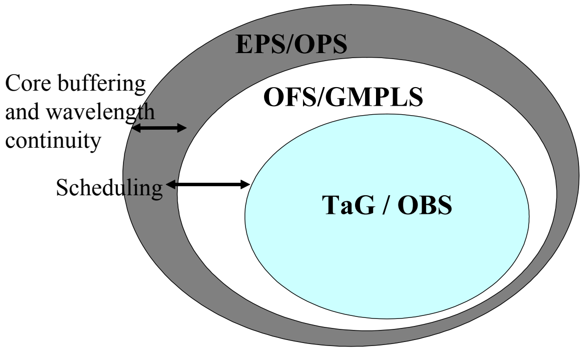

Figure 10. Capacities of different transport mechanism.

D. Capacity comparison in bidirectional rings

As an example of the above results, we investigate here the EPS/OPS, OFS, and OBS capacity properties in an *N* node bidirectional ring with uniform all-to-all traffic of magnitude *r*, under the special case where the number of transceivers equals to the number of wavelengths ($t = w$). Under shortest path routing in EPS/OPS for the case of *N* odd, there is a unique shortest path connecting each node pair, and each link's load is exactly equal to the average link load $r\overline{L}_{odd}$. In the case of *N* even, owing to the existence of two shortest paths for nodes which are diametrically spaced, we ensure that each link's load is exactly equal to $r\overline{L}_{even}$ by routing $r/2$ units of traffic along each of the two paths connecting each diametrically spaced node pair. Since the maximum link load in the network must be less than *t* and $r\overline{L}$ is the exact load on each link, Theorem 2 allows us to conclude that any *r*, such that $r < t/\overline{L}$ is achievable. One can show that the maximum achievable rate *r* by an OFS architecture is actually the same as in EPS/OPS. We omit the details for brevity, but mention that one way of showing this is to propose a set of feasible network states for each of the *t* wavelength channels over which time is shared equally; and then conclude that doing so yields the same maximum *r* as in EPS/OPS. In order to determine the maximum achievable *r* in an OBS network, we employ the two OBS models derived in [11] and numerically maximize their corresponding expressions. Figure 11 illustrates the maximum session rate performance as a function of network size for the different switching architectures. The most immediate observation from the figure is that the EPS/OPS and OFS architectures significantly outperform the OBS architecture.

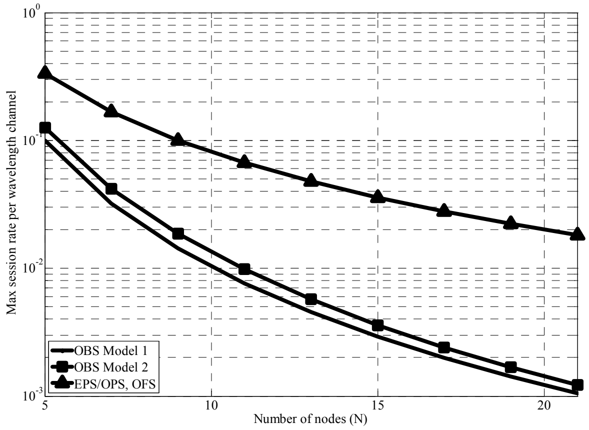

Figure 11: Maximum session rate *r* per wavelength channel versus number of network nodes for the bidirectional ring. Uniform all-to-all traffic and $t = \lfloor N/2 \rfloor$ wavelength channels is assumed.

V. COST-CAPACITY ANALYSIS

In this section, we conduct a capacity-cost comparison of OFS with other network architectures that incorporate varying degrees of electronics and optics: TaG, EPS, and EPS/GMPLS. The context in which we compare these different architectures is a simple, multi-tiered network comprising two MANs, each with a large number of subscribing users of which only a fraction active at any given time (see inset of Figure 12).

In this network, end users are equipped with multiple transceivers, each of which can transmit at the wavelength channel rate in the network. End users are grouped into LANs, distribution networks, or dynamically grouped into subnetworks for the session as described above, which in the cases of the all-optical

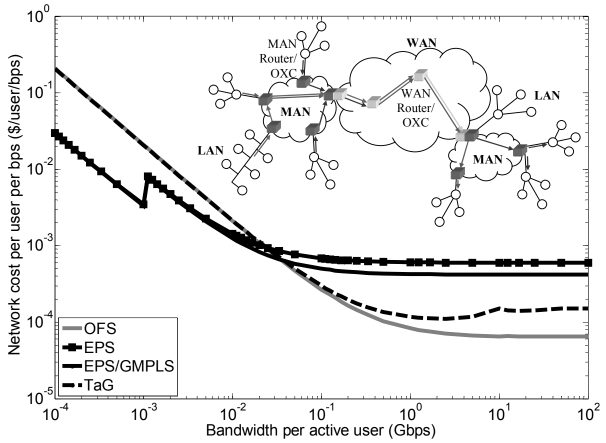

Figure 12: Normalized network cost vs. average user data rate for network with 10^4 users/MAN. (Inset: network model.)

OFS and TaG architectures, are passive, optical broadcast networks. For the MAN, we assume that the physical topology connecting the nodes is arbitrary, but identical for each architecture. The MAN node design (i.e., use of OXC or router), however, is dependent upon the network architecture. Although our network model is restricted to have just two MANs, this simple model may represent a subset of a more complex network. Moreover, we believe that for flow switching to be an economical transport mechanism, there should be enough traffic between two MANs interconnected by a WAN to have sufficient traffic for at least one wavelength channel, and that the assignment of WAN wavelength channels is quasi-static. Thus, the analysis of even such a simple case may serve as a useful building block for more complex network analyses.

In our cost model, we address the transceiver, switching, scheduling, routing, and amplification costs entailed by communication among the MAN users. We omit fiber plant costs – digging, cabling, leasing, and right-of-way costs – as we assume that they are approximately the same for all of the architectures considered. Furthermore, we omit operational costs of the network, which we recognize to be an important cost component.

C. Comparison results

A throughput-cost comparison of the four network architectures using cost and architectural parameters which reflect the state of present-day networks indicates that each architecture is optimal for a range of user rates (Figures 12-14) [12, 13]. Figure 12 illustrates the normalized network cost for each of the four possible homogeneous architectures, given a particular network size; and Figure 13 illustrates the cost-optimal homogeneous network architecture as a function of the number of users per MAN and the average user data rate. EPS and EPS/GMPLS dominate for lower rates because a small amount of electronic equipment is necessary to support the aggregate traffic; whereas, for OFS and TaG, expensive tunable, long-haul transceivers are always required at each end user. On the other hand, at high data rates, regardless of the number of users, OFS always dominates, implying that *OFS is the most scalable architecture of all*. In the high user data rate regime, aggregate traffic is always high, so requiring electronic equipment to support this traffic in the network – even if only in the MAN – is expensive. We note that TaG's regime of optimality is relatively small, since the low cost of scheduling in OFS usually yields great performance benefit relative to the otherwise identical TaG architecture. Figure 13 also suggests that a hybrid network architecture may be sensible when user demands are heterogeneous. Indeed, Figure 14 illustrates that if user demands are heterogeneous (with small to large transactions), then a hybrid network architecture (e.g. EPS+OFS) can be more cost-efficient than a homogeneous architecture. Figure 14 also implies that designing subnetworks for different classes of service may result in a more cost-efficient architecture than a homogeneous architecture.

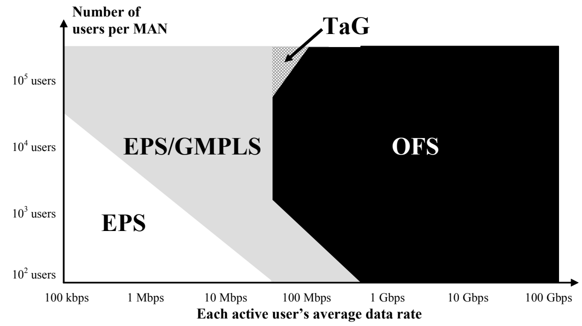

Figure 13: Homogeneous cost-optimal architecture.

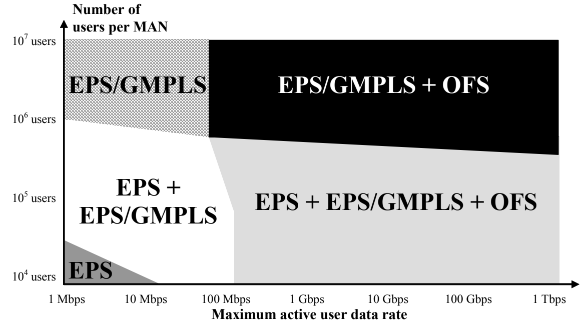

Figure 14: Hybrid cost-optimal architectures.

D. Delay

For delay sensitive applications, there is a delay-utilization-blocking probability trade-off for OFS and OBS, as mentioned earlier. For OFS, utilization and blocking performance can be improved at the expense of delay, as shown in Figure 15. For OBS, delay performance is expected to be poor owing to retransmission protocols after collisions at high utilization. Classical ‘tree algorithms’ can be use to improve the throughput of OBS, but this would be at the expense of the high complexity of segmentation of the transaction. Figure 16 illustrates the benefit of scheduling in OFS [2]. As can be seen, scheduling even just a few sessions *M* into the future improves throughput by a significant amount.

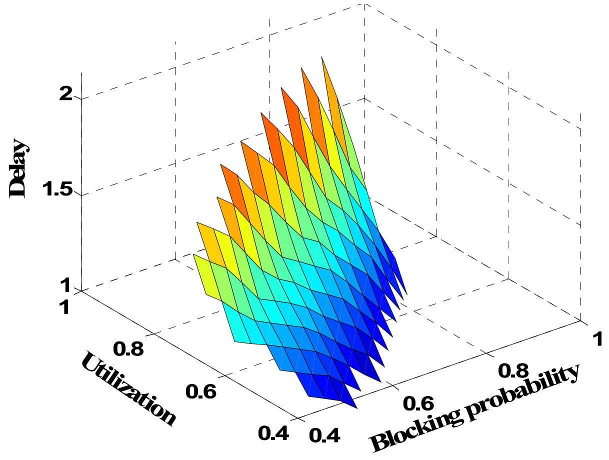

Figure 15. Delay-utilization-blocking probability trade-off for OFS.

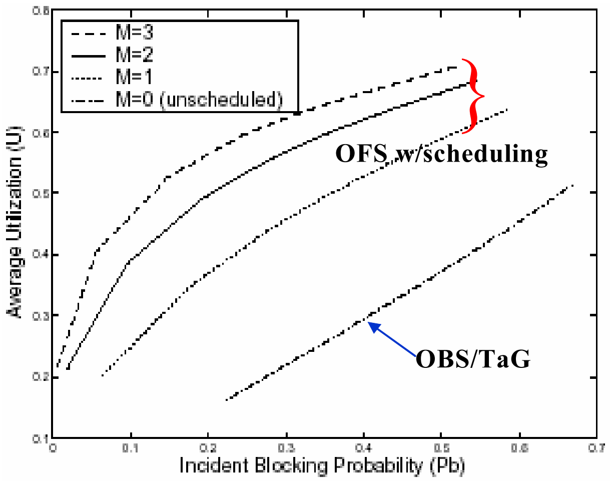

Figure 16. Utilization-blocking probability trade-off of scheduling time horizon for OFS.

VI. Conclusion

If WDM optical networks are to fulfill their promise of a (~3) orders of magnitude increase in data rate to the end user, economical transport for large transactions (> 1 Gbyte) must be used. We believe that OFS will enable such cost-effective transport. To support OFS efficiently, the Physical Layer network architecture must be well designed to support statistical multiplexing via dynamically configured broadcast groups using a MAC protocol and optical switches that act at millisecond time scales. Transport layer protocols must also be simplified for these large transactions. In addition, OFS must be designed to co-exist with lower rate EPS services. A simple cost study has indicated that OFS may provide orders of magnitude decrease in cost for large transactions compared to EPS and other contending transport mechanisms.

References

[1] V. W. S. Chan, K. L. Hall, E. Modiano, and K. A. Rauschenbach, “Architectures and technologies for high-speed optical data networks,” IEEE/OSA Journal of Lightwave Technology, 1998.

[2] B. Ganguly and V. W. S. Chan, “A scheduled approach to optical flow switching in the ONRAMP optical access network testbed,” OFC 2002.

[3] M. Yoo, M. Jeong, and C. Qiao, “A high speed protocol for bursty traffic in optical networks,” SPIE All-Optical Communication Systems, vol. 3230, pp. 79–90, 1997.

[4] C. Qiao and M. Yoo, “Optical burst switching OBS – a new paradigm for an optical internet,” Journal of High Speed Networks, vol. 8, no. 1, pp. 69–84, Jan. 1999.

[5] S. Verma, H. Chaskar, and R. Ravikanth, “Optical burst switching: A viable solution for terabit IP backbone,” IEEE Network, vol. 14, no. 6, pp. 48–53, Nov. 2000.

[6] T. Battestilli and H. Perros, “An introduction to optical burst switching,” IEEE Optical Communications, vol. 41, no. 8, pp. S10–S15, Aug. 2003.

[7] I. Widjaja, “Performance analysis of burst admission-control protocols,” IEE Proceedings Communications, vol. 142, pp. 7–14, 1995.

[8] J. Turner, “Terabit burst switching,” Journal of High Speed Networks, vol. 8, pp. 3–16, 1999.

[9] V. M. Vokkarane, J. P. Jue, and S. Sitaraman, “Burst segmentation: An approach for reducing packet loss in optical burst switched networks,” Proceedings of IEEE International Conference on Communications, vol. 5, pp. 2673–2677, Apr. 2002.

[10] Y. Chen, H. Wu, D. Xu, and C. Qiao, “Performance analysis of optical burst switched node with deflection routing,” Proceedings of IEEE Infocom, vol. 2, pp. 1355–1359, May 2004.

[11] G. Weichenberg, V. W. S. Chan, and M. Médard, “On the capacity of optical networks: A framework for comparing different transport architectures,” Infocom 2006.

[12] G. Weichenberg, V. W. S. Chan, and M. Médard, “Cost-Efficient Optical Network Architectures,” to appear in ECOC 2006.

[13] S. Sengupta, V. Kumar, and D. Saha, “Switched optical backbone for cost effective scalable core IP networks,” IEEE Communications Magazine, pp. 60–70, June 2003.