An Optical Network Architecture With Distributed Switching Inside Node Clusters Features Improved Loss, Efficiency, and Cost
John D. Angelopoulos, Member, IEEE, Konstantinos Kanonakis, George Koukouvakis, Helen C. Leligou, Chris Matrakidis, Theofanis G. Orphanoudakis, Member, IEEE, and Alexandros Stavdas, Member, IEEE

Abstract—The novel core network architecture presented in this paper realizes distributed all-optical switching of payload by partitioning the network into a number of geographically limited domains, where two-way reservations are effective. Thus, inside each domain, loss is eliminated, while traffic from many nodes can be aggregated into single bursts, improving efficiency. Clustered nodes contribute contiguous optical slots, which are marshaled into composite optical frames destined for other clusters, under the guidance of a reservation-based control protocol. The lossless aggregation of traffic from several core nodes allows the use of cost-effective bufferless all-optical transport among the domains with electrical buffers employed at the periphery of the system. The end result is a triple improvement in loss probabilities, efficiency, and cost. This is achieved by exploiting three features of the architecture: the distributed switching functionality (as in early LANs when centralized switching was expensive), localized reservations (avoiding the intolerable delays of end-to-end reservations), and a reduced number of source-destination pairs (by means of node clustering into reservation domains).

Index Terms—All-optical networks, core networks, optical burst switching (OBS), statistical multiplexing, wavelength-division-multiplexing (WDM) networking.

I. INTRODUCTION

RELYING on the solid foundation of the store-and-forward concept, IP routers managed to handle the rapid proliferation of bursty data services with excellent efficiency and multiplexing gain. Endowed with vast and cheap electrical storage, they can resolve output port contention while changing the traffic profiles to fit into the transmission pipes. Their large buffers not only keep loss low but are also the key to high system utilization by storing excess traffic and later using it to keep output lines filled between bursts, molding the temporal properties of flows, to better exploit available-link capacities. However, this successful paradigm shows signs of saturation, as it cannot keep up with fast-improving optical transport, making routers the bottleneck of modern networks. The imbalance manifests itself in the four-orders-of-magnitude difference between the available raw transmission bandwidth, which is achievable via a combination of fiber and wavelength-division-multiplexing (WDM), and the useful bandwidth offered from the current Terabit routers [1]. The weak point is the required per-packet header processing that does not scale well in core networks while it is hardly useful in the context of all-optical switching, due to lack of optical look-up solutions. Approaches that reduce the header-processing burden, such as multiprotocol label switching (MPLS) and generalized MPLS, are providing relief but not a long-term answer to the modern data-service needs.

Despite extensive efforts and significant developments, optical switching lacks storage elements, processing capability, and switching fabrics compatible with the store-and-forward paradigm. It is obvious that barring the emergence of revolutionary advances in fast switching and/or optical buffering that could turn the present situation around, core architectures must seek different concepts that exploit the strong points of optical technology while side-stepping its disadvantages.

Indeed, several attempts in this direction have emerged in recent years, but none can yet fit the bill. Wavelength routed networks [2] present such an architectural paradigm shift, but the achievable light-path reconfiguration speed is too slow to satisfy the fluctuations of data-traffic profiles. Another more dynamic approach, optical burst switching (OBS) [3], [4], can handle bursty traffic with multiplexing gain by exploiting the multitude of WDM channels to handle contention. Still, OBS suffers excessive burst loss, since often, more bursts contend for the same output port than can be accommodated by the WDM channels and the limited (if any) buffers. Loss probabilities grow exponentially with the number of contention points (i.e., the number of nodes that need to be crossed to reach the destination). For typical network sizes of 20–30 nodes needing three to five hops, system loads must be kept quite low (15%–30%) to reduce collision probabilities to acceptable levels [5]. Two-way reservation approaches [6], [7] avoid such losses but are only applicable to domain sizes of tolerable roundtrip propagation delay (i.e., 1000–2000 km). In addition, further delay is required to create long reasonably filled bursts before a reservation request can be sent, as the aggregation delay cannot fully overlap with the reservation delay [8]. Pipelined reservations for only a small fraction of the roundtrip time (RTT) are employed in the probe-and-go concept [9], providing efficiency without loss by finding more collision-free itineraries.
than plain OBS. Still, probe-and-go suffers the range limitations of roundtrip delay and is not applicable to large domains. This led the authors to propose this novel hybrid architectural approach that overcomes this limitation enabling the size of the optical core to be only limited by physical-layer restrictions, which continuously improve with technology, unfettered from control-protocol limitations. Its hybrid approach exploits the best elements of both the one-way and two-way approaches and a smoother migration from current synchronous digital hierarchy (SDH) rings.

This paper is organized as follows. In Section II, the rationale and the operational concept are presented, while in Section III, the control of the system is elaborated upon. The bandwidth-allocation algorithm is presented in Section IV and the intercluster options in Section V. Finally, the performance of the system is evaluated in Section VI, and conclusions are drawn in Section VII.

II. RATIONALE AND CONCEPT OF OPERATION

Given the above limitations of excessive loss for one-way and excessive delay for two-way reservation systems, it is natural to seek to combine the strong points of both schemes in a hybrid architecture that exploits a special kind of collective, coordinated, slot-based, two-way reservations within limited-size domains (created by breaking down the whole network into clusters of neighboring nodes), with open-loop approaches between the clusters. Another strong incentive for clustering nodes is to reduce the number of source–destination pairs. This brings a vast increase in efficiency of burst aggregation, while at the same time reducing the number of switching (hence, contention) points, replacing some of them with a collision-free medium-access protocol, as described in this paper. Thus, collisions are only possible in the very few nodes interconnecting the clusters, diminishing their probability compared to a system permitting contention in all its nodes. Collisions can be even completely eliminated at the penalty of static overprovisioned intercluster links as discussed in Section V.

Keeping the legacy ring topology within the cluster, although not optimal in terms of routing, exploits the well-known advantages of rings in terms of dimensioning, management, multicast, protection and restoration, etc. It also constitutes a natural migration step when starting from today’s SDH rings. The methodology for this approach is similar to protocols developed and implemented for the metropolitan-area-network part in the IST project DAVID [10], as well as those of passive optical networks (PONs) [11]. The adopted strategy is to rely on coordinated collective two-way reservations using the electrical control plane to realize a meticulous accounting of payloads and schedule optical bursts. One of the cluster nodes is selected to act as a master and its role is dual: It acts as the interconnection point with the other clusters, and it governs the operation of its own cluster allocating the bandwidth to the nodes of its cluster in a collision-free manner. Unlike classical reservations that reserve the whole channel for a duration at least equal to a RTT (since they cannot discriminate smaller bursts), the reservations in this paper operate on a slot-by-slot basis and are coordinated and collective using individual requests issued periodically.

To achieve this, the requests from all nodes of a cluster are processed by a special allocation algorithm to produce contiguous collision-free allocations that are combined in an efficient multinode burst covering a longer time window, i.e., a frame time. The data chunk size that is the object of each reservation is fixed (i.e., a data slot) and contains several thousand bytes, which is a practice also followed in GPON standards [11], that can also offer performance advantages in optical switching [10]. Slots are joined to form variable- or fixed-size frames, depending on the optical-switching-technology tradeoffs. Thus, two granularity levels are used: the “slot” and the “frame.” Both time-division multiplexing and WDM are employed. This special and novel form of coordinated two-way reservations creates a control framework which is exploited to turn cluster nodes into ports of a gigantic input-buffered distributed switch obviating several centralized switching points, which would suffer high losses (as happens in OBS) because of the uncoordinated arrival of bursts. The whole framework results in distinct performance advantages that are quantified in the performance section: It eliminates collisions inside the clusters, increases link utilization by reducing voids due to burst coordination, and needs fewer better utilized links between clusters.

In every periphery node, the fixed-size slots are created out of IP packets destined for the same node by a slot-aggregation unit whenever the payload reaches the slot size or a time-out occurs. Slots are queued in edge-node buffers per-destination-cluster and per-quality class. Trains of such slots assembled on the fly under the supervision and guidance of the reservation protocol in a way responding to the bursty traffic fluctuations are formed and then exchanged between clusters. The whole system is slot/frame synchronous but bit asynchronous, requiring burst-mode transceivers but not fully synchronized clocks.

The architecture on which the proposed solution applies is the typical of optical-core networks as exemplified by the network shown in Fig. 1 (left-hand side), consisting of several core nodes, to which a set of (possibly collocated) edge nodes possessing electrical buffers are attached. Electrical buffers exist only in the network periphery, and as soon as a slot has entered the optical network, it will travel to its destination without any further buffering. The next step is to group the nodes into clusters on the basis of mainly vicinity but also administrative domain limits or traffic criteria or the legacy SDH situation which probably coincides with the above criteria already. The nodes inside a cluster are connected in a ring topology (probably along the conduits of an already existing SDH network).

One node inside each cluster is designated as a master node (MN) and it acts as a gateway to the other clusters, to which it is linked via WDM links. It also hosts the reservation protocol that handles the bandwidth allocations to the nodes of its cluster. The MNs are linked in an arbitrary mesh topology as necessary to offer enough connectivity and redundancy. The information transfers in the system take place in four steps.

1) First, nodes announce queued slots trying to reserve transmission slots.
2) Second, using slot allocations granted by the MN on the basis of previously collected reservations, nodes contribute contiguous slots for the same destination cluster (regardless of the specific destination node and traffic
class), creating bursts (in the form of fixed-size frames) inside the source cluster.

3) Third, as the bursts reach the MN in the source cluster, they are forwarded to the destination MN (possibly via one or two other MNs in very large networks).

4) Finally, the bursts arrive at the destination-cluster MN, and as they travel around the (destination) ring, each destination node collects the slots of the frame destined to it as they pass in front.

All the nodes in a cluster use the same frame and wavelength (or wavelengths, should there be sufficient traffic) to transport traffic with the same destination cluster, thus increasing the utilization of each wavelength. The control information travels in a dedicated wavelength, which is the only one converted to the electrical domain and processed in every node. It is pointed out that idle frames are issued when there is no traffic to be transported inside the ring. In brief, clustering aims at improving the aggregation while at the same time obviating contention within the few first nodes (thanks to reservations) and last nodes (thanks to ring multicast).

The aforementioned concept of operation is illustrated in Fig. 1 (right-hand side). The one direction of the wavelengths (for example $\lambda_i$, $\lambda_j$ in the figure) is used for aggregating and sending the traffic and the other for receiving traffic generated in other clusters. Separate wavelengths (for example $\lambda_i$, $\lambda_j$ in the figure) are used for sending and receiving traffic between nodes in different clusters. Local traffic inside the cluster ring also uses separate wavelengths, since there is no synergy to be exploited from mixing it with intercluster traffic. Thus, low-cost transceivers designed for shorter distances can then be used, with regeneration at the MN [12]. Quite mature and low-cost solutions exist for simple WDM rings [10], [12], with or without slot reuse, but our main scope is the transfer of interring traffic so that the transport of local intraring traffic will not be further elaborated upon.

III. CONTROL CHANNEL

The control information is transported in fixed-size frames issued periodically by the MN occupying a tiny portion of the control-wavelength bandwidth. The repetition period is in the order of the RTT. As shown in Fig. 2, it includes an alignment pattern, which is used for the physical-layer bit and byte alignment purposes (see for details the optical nodes in [12]), followed by the slot allocation map (SAM) and the slot reservation map (SRM).

The collection of information for queued traffic is based on the periodic SRM, shown in detail in Fig. 2, which includes a different region for all the possible destination clusters. In each region, the MN sets the destination-cluster address (DCA) and the successive node addresses, while each node accessing it, as it passes by, inserts its queue length ($Q$), expressed in slots, for the destination-cluster and quality-of-service (QoS) class in question as well as the the error-control (EC) field used for robustness reasons. The choice of reporting the full queue length each time, instead of just the new arrivals, is made for robustness reasons. This practice is also followed in similar reservation schemes in metropolitan rings [10] and PON access protocols [11], where it was shown that, in the event of errors resulting in underallocation (whenever the error gave a smaller report—the opposite produces only an insignificant waste of little bandwidth), the problem was automatically corrected in the new round. If new arrival reporting is used, it could lead to queue build-up and overflow. Two QoS classes can be supported and the queue length of each is reported in a separate field. Thus, in each SRM, all the waiting slots for all clusters become known to the MN, allowing dynamic response to traffic fluctuations per destination by proper allocations, as is explained in the next section.

Based on the contents of the SRM, the MN continuously updates a reservation matrix $R$, where element $r_{ij}$ contains the number of slots wishing to go from node $i$ to cluster $j$. Care has been taken to include very simple operations in the algorithm that can be carried out by dedicated hardware logic in one slot time. With two QoS classes supported, two such matrices are needed: one $R_H$ for high-priority traffic and another $R_L$ for low-priority traffic, each based on information from the respective $Q_H$ and $Q_L$ fields of the SRM. The MN also maintains an array $S$ (sum), where each element $s_i$ contains the total number of slots queued at all nodes of the cluster and destined for cluster $i$ calculated by adding the arriving queue lengths from all nodes for cluster $i$ including both priority classes.
The above tools are used to prepare the SAM (the main part of the control frame shown in Fig. 2), which is sent ahead of the actual payload frame by enough time to allow nodes to act on the payload after reading the control channel. For each of the supported data wavelengths, the DCA, where the frame under formation will be eventually sent upon its arrival at the MN, is written. The allocations to each node [specified by the source address (SA) field] provide the exact position in the frame by means of pointers indicating the starting and finishing slots and do not necessarily grant the requested number of slots. This happens only if the total number of requested slots is not greater than the frame duration measured in the slots. The MN also includes a number of destination-address (DA) fields (initially containing the hexadecimal value FF), which is equal to the number of allocated slots, where the node will insert the actual destination node address in the destination cluster, which will be used by the receiving node in the other end to identify which slots are addressed to it. Therefore, the EC field needs to be recalculated in the nodes. Hence, each node allocation consists of 8 bits for the SA field, 12 for the start slot, 12 for the end slot, and 8 for the EC, resulting in 40 bits. To support \( N \) nodes in each cluster with a frame consisting of \( L \) slots, \( 40^*N + 8^*L + 8 \) bits are required in the SAM per wavelength.

As shown at the bottom of Fig. 2, each node transmits a payload burst preceded by a physical-layer preamble, consisting of a guard-band and a pattern for synchronization and power leveling. These are necessary because of the burst-mode-operated transceivers (see [12] for details of the physical layer). After the preamble, the payload slots are inserted as marked by the MN. For example, in \( \lambda_i \), node one inserts a burst starting from slot zero and ending at slot three.

It is worth clarifying two aspects of the slot operations at the node. First, a slot is born with the arrival of the first IP packet that will be accommodated in it, and a request is queued locally until the arrival of the SRM of the next control frame in which it will be reported. Thus, the possibility that the slot may not be actually filled until the time of departure is intentionally ignored to ensure a tight control on delay. However, because of the much larger reservation time, in reality, almost all slots depart fully filled as simulations show. Second, the node can use the allocated slots at will, regarding the QoS of the payload, i.e., will insert high-priority slots irrespective of which queued slots triggered the reservations that resulted in the current allocations. Such high-priority slots will not be of course included in the requests (unless already done), while low-priority slots that still remain in the queue will again be reported in the reservations (\( Q_L \) field).

**IV. FRAME ALLOCATION ALGORITHM**

The objective of the allocation is to aim for high fill-up of the data frame while relying on timers to enforce upper limits to the service delay in each direction. The MN prepares the allocation map using the simple algorithm presented in Fig. 3. It operates on the basis of the information in the array \( S \) (sum of slots for both priorities per-destination cluster), the reservation matrix \( R_H \) (high priority), the reservation matrix \( R_L \) (low priority), and a set of timers per-destination cluster employed to guarantee an upper access delay bound. The basic strategy is that a frame is allocated toward a cluster when enough requests exist to fill it or the relevant timer expired, indicating that further waiting to improve utilization would violate the service-imposed upper delay bound. In that sense, the main difference to the quite similar approach of the frame aggregation unit of any optical core system relying on periphery buffers is the fact that, in this case, the queues are distributed and remotely coordinated. The timers enforce an upper limit of the time from the arrival of a first announcement of a nonempty queue until an allocation is issued to that queue. The timers never run on an empty reservation to a destination cluster.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

ANGELOPOULOS et al.: OPTICAL-NETWORK ARCHITECTURE WITH DISTRIBUTED SWITCHING

Fig. 3. Pseudocode of the frame-allocation algorithm.

The choice of timer values is a tradeoff between efficiency and service delay. The higher values give time for better filled frames, but they also result in higher average delay and upper delay. However, values below the RTT provide progressively less and less improvement to the data delay, which becomes dominated by the sum of the times to send the request and receive the allocations. This sum is always higher than the roundtrip and increases with the total load. On the other hand, the high end of timer values is dictated by the delay tolerance of the services. For high-priority traffic, which is expected to be 20%–30% of the load and carry interactive real-time services, values in the order of few to 30 ms provide an upper limit. In this limit, also dictates the size of the cluster (its RTT), which, therefore, cannot exceed about 6000 km. Thus, a choice of the RTT is optimal for demanding traffic, while best effort traffic can tolerate several hundreds of milliseconds, and the exact choice does not have a significant impact on efficiency or service.

The algorithm provides as its output the destination cluster of the next frame and how the slots within each frame are allocated in the nodes of the (source) cluster. Idle frames are issued when no queued traffic has been reported. The pseudocode of the algorithm, given in Fig. 3, is the instance employed in the simulation model below which employed provisioned wavelengths in the intercluster network and fixed-size frames.

A frame in a wavelength toward a certain cluster is assigned when enough traffic to fill the relevant frame has been announced or the relevant timer has expired, in which case, the queued number of slots is allocated. To prepare the SAM, high-priority traffic is always allocated first, and the second priority is inserted in the frame, only if empty positions still remain, which is, however, more often than not the case due to admission control. The number of slots allocated to each node coincides, if possible, with its previously announced queue size unless the total exceeds the maximum frame size; in this case, each node is allocated a lower number on the basis of the fairness algorithm. The latter distributes bandwidth of node j for cluster i “pro rata” according to weights W[j,i] calculated on the basis of service-level agreements establishing the aggregate-provisioned bandwidth per node. Since only integers are used, the remainders are accumulated, and the fairness algorithm works like a weighted-deficit round robin. If no announced slots remain unallocated, then the timer is not restarted immediately but only when the first report of queued slots to that destination arrives.

V. INTERCLUSTER NETWORK

Once a frame goes around the ring and reaches the MN, it is 2R or 3R regenerated according to the physical-layer constraints and the design rules of the link, and then, it is forwarded toward the destination cluster MN via the intercluster network. The intercluster network, connecting the MNs, is the only part of the system that collisions are possible. However, this is the case only whenever the pursuit of efficiency and further multiplexing gain leads to the choice of shared links. We can distinguish between three possible total network architectures of which only the first two are completely collision-free.

In the initial deployment and for relatively small networks of about 20–30 nodes, where only four to five clusters (with five to six nodes each) need to be connected, a simple obvious
The sending MN has to segregate the elements of the control frame arriving ahead of the burst according to destination cluster and send each individual header toward the destination MN, where it will be needed to allow nodes to detect which slots are destined to them by only monitoring (and translating to the electrical domain) the control channel. A while later, the payloads arrive in each wavelength and they are received by all nodes (due to its virtual-bus architecture inside the ring), but each destination node keeps whatever is addressed to it on the basis of the destination-node-address information (DA field, one for each slot) already inserted in the SAM at the cluster of origin (as shown in Fig. 2). The SA field is used to identify the reassembly engine to which the slot will be directed. The reassembly takes place on the basis of fields inside the slot header not shown in the figures.

For larger networks or for cases making a multihop network necessary, it is obvious that there is the choice of connecting the MNs through light-paths of a switched optical network, which acts as a supercore network since it only works on already aggregated traffic. The MN, in this case, should be an optical cross-connect supporting both space switching and wavelength conversion. Significant synergy exists between the two systems as the intracluster network handles the fast actions creating the multiplexed aggregates, while the intercluster network takes care of slower variations of load. The latter can reassign light-paths whenever a tendency to congestion is observed signaling to the relevant MN to start using the new paths until they are reassigned to another destination as loading patterns vary (at a slower timescale). The end result is that the same total performance and flexibility is reached with a much slower, hence much cheaper, switching among clusters.

Finally, at a later stage, the intercluster network can become completely dynamic adopting any known OBS approach or future improved alternatives. In this case, the information in the SAM will be appropriately adapted at the MN and launched ahead of the payload burst to act as a burst header. This means that the proper lead-time for the total network is set beforehand to the appropriate value. Although now contentions and losses may occur at switching points, their probability will be much lower than that in an OBS with an equal total number of nodes, since contentions will still be eliminated inside the cluster of origin (due to reservations) and the destination cluster (due to its broadcast nature) leaving only a couple of intercluster switching points, even if very large networks become a reality in the future. The choice of the optical-switching technology strongly affects the choice of slot and frame duration. The range of reasonable values could be slots with duration ranging between 10 and 100 µs and frames between 250 µs and 50 ms. Low values would require semiconductor optical amplifier-based switches, while high values allow slow switches based on microelectromechanical-system technology for lower cost.

VI. PERFORMANCE ASSESSMENT

The performance assessment of the proposed architecture was carried out with the help of computer simulation. The slot-assignment algorithm used was the one shown in Fig. 3. The simulation model consisted of five cluster rings with six nodes in each cluster. Slots of 0.1-ms duration were used, while the frame contained 50 slots (5 ms), which is near the RRT of each ring. Four payload wavelengths were used in each cluster, while between the clusters, each λ was provisioned to one of the other clusters with an intercluster link of 2000 km length (10 ms). In each node, there are five pairs of transceivers while four of them are dedicated for interconnecting the five clusters. In all runs, high-priority traffic was 30% of the total. The timer values were 25 ms for high priority and 300 ms for low. The rate at each λ was 10 Gb/s. Both Poisson and self-similar sources (in different scenarios) were used to generate the offered load, which was uniformly distributed for each destination among all nodes and clusters. Self-similar traffic in each node consists of 20 ON--OFF sources per destination cluster (per queue) with Pareto distribution of both the ON and OFF duration: a burstiness of ten and shape parameter of 1.3. As the mean propagation delay is almost deterministic, since, on average, the traffic has to cross one half of the source cluster (i.e., 2.5 ms) and one half of the destination cluster (2.5 ms) plus the intercluster link (10 ms), the real interesting delay parameter of the system to study is the queuing delay and delay variation.

The average queuing delay against the total offered load is shown in Fig. 4(a). Two remarkable effects can be observed in these results. The queuing delay falls with the load (for the high priority, it even goes below the reservation time), and all the instability of overload is suffered by the low priority only, leaving the demanding services to enjoy the required quality of service, even above 100% total offered load. Overall, the system can take up to almost 90% loading, which proves the high efficiency of the approach [the difference between Poisson and self-similar traffic manifests itself at the earlier instability for the latter and the heavier tails visible in the delay probability density function (pdf) shown in Fig. 5(b)].

The delay is higher at low loads, because the frame generation is dominated by the timers. The choice to avoid an algorithm with a contiguous generation of partially filled frames was made to improve efficiency and multiplexing gain in the intercluster network by having fewer better filled frames. In this model, only provisioned links were assumed between clusters; larger networks can resort to one-way OBS among clusters where efficiency would matter.

The mean fill levels of frames as a percentage of the frame versus offered load are shown in Fig. 4(b) separately for the high and low priority. Except at very low loads, below 25%, the frames are practically full and of course the ratio is 30%–70%
as generated by sources. Above 100% (overloaded system), only the high priority is fully serviced, while the low priority suffers loss of the traffic that exceeds capacity. Consequently, the percentage of each in the departing frame is altered.

The way high-priority traffic can beat the reservation delay is by “stealing” positions reserved by low-priority traffic. In other words, a node that has sent the queue information an RTT earlier, and is assigned a number of slots in the next frame, will insert newly arrived high-priority traffic in the place of its low priority while repeating the request for the unlucky low-priority slots (and never presenting any requests for the “furtively” departed high-priority slots). Thus, the system is not fully reservation-based as regards high-priority traffic, for which the reservations act as a contingency plan. This effect can be studied with the help of Fig. 4(b), where the percentage of slots that were assigned to high-priority traffic is shown (the lower curve). Comparing with the curve right above, which shows the departed slots, one can see that the assignments to high priority are always lower than the actual departures of high-priority traffic. In addition, they get even lower as the load rises, reaching zero assignments at 100% load, i.e., at high loads, almost all high-priority traffic departs in unsolicited slots stolen from reservations intended for low-priority traffic, thus avoiding the reservation delay.

For even more demanding services, it is possible to imitate virtual leased line performance by unsolicited periodic semi-permanent slot assignments, but this mechanism was not built in the simulation model because its behavior is deterministic.

The pdf of the delay at an 80% load for both Poisson and self-similar sources is depicted in Fig. 5(a). For high-priority traffic, the framing dominates, explaining the stepwise shape of the pdf. Most traffic departs within one RTT, while some goes for the second, with a few that encountered momentary congestion, necessitating a third. In the self-similar traffic case, the probability inside the first and second frame is more variable, but higher delays than in the Poisson case are also evident. For low priority, the common in many queuing systems bell-shaped pdf is the case with distinctly heavy tail for the self-similar case. However, it is worth noting that the model does not simulate transmission control protocol (TCP) congestion-control algorithms, which would be activated to eventually reduce the low-priority offered load via its closed-loop mechanisms at the transport layer, also improving the experienced delay.

In Fig. 5(b), the loss performance of the proposed system is compared to the “classic” just-in-time one-way OBS system. To reach comparable results in both cases, 16 nodes are used to produce uniform (i.e., destined to any other node with equal probability) Poisson traffic in a single priority level. In the OBS case, the 16 nodes are interconnected in the symmetric 4 × 4 torus [3], [5] topology commonly used in OBS performance evaluations, while in the second case, the four nodes in each of the four columns in the torus are grouped into clusters with the
top node acting as MN. In both cases, all links carry four $\lambda$s and are of the same 250-km distance, giving a 1000-km cluster periphery. Since the clustered architecture does not allow interconnection of nodes in different clusters, except through the MNs, the three “illegal” horizontal links are removed and placed between the MNs to avoid a bottleneck there and produce, to the extent possible, an equivalent system in capacity as any operator adopting the clustered solution would do. Still, the two systems feature differences with OBS needing more ducting and more expensive optics (for example, all OBS nodes employ space switching, while in our approach, only MNs do), but the above setup allows traffic performance comparisons. The results in Fig. 5(b) compare loss probabilities between the two systems under varying load levels. The losses in the latter case are orders of magnitude lower, since the clustered architecture behaves like any normal queuing system and the only losses occur from buffer overflows in the periphery queues. The loss probability is so low that in order to obtain measurable losses, extremely small buffers of 45 slots per queue (i.e., per-destination cluster) were simulated, which correspond to 5625 MB per queue, while an order of magnitude higher electrical buffer would be more cost-effective (cf. with buffers of 512 MB per port used in a typical core network IP router). Even then, losses were detected only at very heavy loads, while OBS suffers significant losses even at low loads due to accidental collisions of bursts, which is something that is absent in the clustered reservation-based system.

The near elimination of the loss in clustering is achieved at the penalty of reservation delay which enables the coordination that makes collision avoidance possible. Still, delay and delay variation remain within quite acceptable limits for both the demanding and the best-effort services. In contrast, one-way OBS suffers such high losses that some form of retransmissions is necessary. Retransmissions are usually left to TCP for best-effort traffic or could be done in the optical layer for real-time traffic. Both cases present serious drawbacks [13]–[15], and although the average one-way delay appears quite small, once retransmission delays are also considered (usually left out of performance studies that do not simulate retransmission mechanisms), the worst-case delay becomes quite high, and bursts that suffered repeated loss have to be dropped [14], [15]. These problems are avoided in the clustered architecture, where the bursts wait in the electronic buffers until a transit path is cleared. Thus, reservations operate as proactive retransmissions, replacing payload retries with control exchanges, thus reducing delay extremes that have a bigger impact on delay-sensitive services than average delay.

VII. CONCLUSION

The clustering approach constitutes a “divide-and-conquer” strategy whereby dividing the core nodes into few clusters based on vicinity allows exploiting two-way medium access control (MAC)-like slot reservations within each cluster. This way, burst loss is avoided inside the clusters, without sacrificing multiplexing gain, while achieving traffic aggregation among several clustered sources and destinations. The reservation mechanism responds to actual fluctuating demand at the required small timescales only by fast laser control obviating expensive fast optical switching. The proposed solution is similar to early Ethernet LANs, where the high cost of switching led to the adoption of a distributed switching function with the stations attached to the bus picking their own frames. Currently, similar cost and other constraints of optical switches can lead to a similar approach, where the nodes inside both the source and destination clusters act as switching ports. The performance of the system allows easy guarantees to demanding services because their delay is fixed by the reservation delay at low loads, decreasing at higher loads thus counteracting the increasing delay in other parts of the network. Best-effort traffic, on the other hand, enjoys high throughput and utilization, while both classes suffer no loss unless ambitious dynamic approaches are adopted among clusters, and even then, loss probabilities remain extremely low due to the big reduction of contention points.

REFERENCES

John D. Angelopoulos (S’92–M’95) received the degree in electrical engineering from the National Technical University of Athens (NTUA), Athens, Greece, in 1973, the M.Sc. degree in automatic control from Nottingham University, Nottingham, U.K., in 1977, and the Ph.D. degree in telecommunications from NTUA in 1992.

He worked for 11 years in the research and development section of the aerospace and telecommunications industries before joining the Technological Institute of Piraeus, Aigaleo, Greece, in 1989, where he is a Professor in the Automation Engineering Department, teaching data-network-related courses. He is also participating in NTUA research activities in the areas of high-speed networks. He has been involved in several research and technological development in the field of communication technologies, ESPRIT, advanced communications technologies and services, and information society technologies projects, with emphasis on broadband access systems (PONs, hybrid fiber/coax (HFC), and wireless). His research interests cover most aspects of high-speed communications with special focus on shared medium-access protocols and implementations.

Dr. Angelopoulos is a member of the Technical Chamber of Greece.

Konstantinos Kanonakis was born in Heraklion Crete, Greece, in 1982. He received the Diploma from the Faculty of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece, in 2004, where he is currently working toward the Ph.D. degree.

His main research interests are in the area of traffic engineering, architectures and control protocols for optical networks and their implementation in hardware.

George Koulouvakis was born in Heraklion Crete, Greece, in 1980. He received the Diploma degree in electrical and computer engineering and the M.Sc. degree in techno-economic systems, both from National Technical University of Athens, Athens, Greece, in 2003 and 2006, respectively.

Currently, he is a Research Associate with the Telecommunication Laboratory, National Technical University of Athens. His main research interests are in the area of computer and optical networks, traffic engineering, and multimedia transmission over IP networks.

Helen C. Leiligou received the Dipl.Eng. and Ph.D. degrees in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 1995 and 2002, respectively. Her Ph.D. dissertation was in the area of access control mechanisms in broadband networks.

She is currently with the Institute of Computer and Communication Systems, National Technical University of Athens. She is currently investigating access mechanisms for high-speed networks for tree-topology hybrid fiber/coax and gigabit passive optical networks systems and wavelength-division-multiplexed metro-ring networks.

Chris Matrakidis received the electrical engineering degree from the University of Patras, Patras, Greece, and the Ph.D. degree in the field of digital communications from University College London (UCL), London, U.K.

From 1995 to 1999, he was a Research Fellow at the Telecommunications Research Group of UCL. His primary research interests are in the areas of simulation and optimization of communication systems. He has also worked on the applications of coding theory, GSM systems, optical networking, satellite communications, and magnetic recording. He is an author or coauthor of several journal and conference publications. He is currently lecturing at the University of Peloponnese, Tripolis, Greece.

Theofanis G. Orphanoudakis (M’98) received the Dipl.Eng. degree in electrical and computer engineering and the Ph.D. degree in telecommunications from the National Technical University of Athens, Athens, Greece, in 1995 and 1998, respectively.

From 1996 to 1999, he was a Research Associate at the Telecommunications Laboratory, National Technical University of Athens. From 1999 to 2005, he lead R&D activities in the area of broadband networking components as a Technical Manager of Ellinmedia Technologies. He is currently with the University of Peloponnese, Tripolis, Greece, as a Visiting Assistant Professor. He has participated in several European Union research programs and has led projects in the areas of broadband networking components and broadband access networks. His research interests include the design and implementation of high-speed networking systems, performance evaluation of MAC protocols for shared broadband systems, traffic engineering, switching, scheduling, and network processor architectures. He has several publications in the aforementioned areas.

Alexandros Stavdas (M’00) received the B.Sc. degree in physics from the University of Athens, Athens, Greece, the M.Sc. degree in optoelectronics and laser devices from Heriot-Watt University/St. Andrews University, Edinburgh, U.K., and the Ph.D. degree in wavelength-routed wavelength-division-multiplexed networks from University College London, London, U.K.

He was recently appointed Associate Professor of optical networking with the Department of Telecommunications Science and Technology, University of Peloponnese, Tripolis, Greece. He is the author or coauthor of more than 50 journal publications and conference articles. His current interests include physical-layer modeling of optical networks, ultrahigh-capacity end-to-end optical networks, optical cross-connect architectures, and optical packet switching.