Abstract—We propose an all-optical networking solution for an Internet core network based on the notion of multipoint-to-multipoint lightpaths that, for short, we call multipaths. The multipath MAC protocol concentrates the traffic of a cluster of source nodes on a wavelength channel and multicasts this traffic to a cluster of destination nodes. The proposed network can be built using existing components and is shown to bring significant energy savings compared to classical, router-based core networks. A flow-aware DBA algorithm is proposed and shown to have excellent performance.

Index Terms—all-optical network, core network, dynamic bandwidth allocation, flow-aware networking.

I. INTRODUCTION

As Internet traffic continues its exponential growth, there is increasing concern that the current networking paradigm based on large electronic routers will not scale. Scalability in capacity is mainly limited by the power consumption of the core networking systems (see [1] and other papers in same journal issue). While researchers are increasingly looking towards dynamic optical switching technologies for the solution to this problem, most proposals are either futuristic or ill-suited to the burstiness and fine granularity of Internet traffic [2].

In this paper we propose an original approach to building a transparent country-wide optical core using technology that is available today. Our proposal is to share the bandwidth of multipoint-to-multipoint lightpaths, that by contraction we call multipaths, under the control of an adapted medium access control (MAC) protocol. It derives from related proposals for optical networking in access and metropolitan area networks (MAN) and has similarities to alternative optical solutions for the core.

Our proposal for controlled lightpath sharing is largely inspired by the dynamic bandwidth allocation (DBA) algorithms used in passive optical access networks (PONs) [3], [4]. Such algorithms are also used to manage multipoint-to-point lightpath sharing in time-domain wavelength interleaved networking (TWIN), from which we borrow and enhance the distributed MAC layer [5], [6]. To extend the scalability and geographic scope of TWIN, we create clusters of source and destination nodes, as in the proposed CANON architecture [7], [8]. Our proposal is actually quite similar in principle to the network with static inter-cluster wavelength assignments discussed in [7]. The realization is quite different, however, and we would argue that our approach is simpler, more flexible and has better performance.

In the following sections we successively present the three main contributions of the paper. In Section II we define the multipath and its control, clarifying its relation to prior work on lightpath sharing and discussing required components. Section III shows how these components can be assembled to build the core network of a typical European country-wide ISP. We notably evaluate the significant energy savings that would derive from eliminating electronic transit routers. The proposed MAC layer is flexible and able to support a range of possible DBA algorithms. In Section IV we briefly present the MAC protocol and a preferred flow-aware DBA algorithm. Simulation results illustrate the excellent performance achieved in terms of both expected flow bit rate and mean packet delay.

II. THE MULTIPATH

We first discuss some related work on lightpath sharing before presenting the multipath and its control mechanisms.

A. Sharing lightpaths

WDM is currently used to provide high-capacity point-to-point lightpaths consisting of a wavelength channel carried over a succession of fibres interconnected by wavelength selective optical cross connects (OXC). As wavelengths have a capacity of 10 Gb/s or more, a lightpath is only used efficiently when its end points concentrate a large amount of traffic. In current ISP networks, traffic is usually concentrated through a hierarchy of electronic transit routers. In this paper we envisage the alternative of performing lightpath sharing in the optical domain.

Point-to-multipoint sharing of lightpaths is realized using splitters to disseminate the wavelength signal over multiple outgoing fibres. Sharing could be performed using optical time division multiplexing with each destination extracting its signals from dedicated time slots [9]. A more flexible alternative is that implemented in PONs where the composite downstream signal is received and converted to electronic form by all destinations [3], [4]. Each node recognizes its own traffic and discards the rest.

Multipoint-to-point lightpaths are realized by merging signals from two or more incoming fibres onto a single outgoing fibre. This principle is used notably for upstream transmission in PONs or for sharing an optical bus (e.g., [10], [11]). A MAC
protocol is necessary to ensure signals from distinct sources do not collide.

Of particular relevance to the present work is time-domain wavelength interleaved networking (TWIN) [12] and the distributed MAC proposed by Saniée and Widjaja [5]. Each destination in TWIN is the root of a multipoint-to-point lightpath and implements the MAC protocol to manage its incoming traffic. The alternative distributed MAC for TWIN proposed by Robert and Roberts [6] is actually more relevant to the present approach.

TWIN is well-suited to building a MAN. Extension to a country-wide network is not possible, however, notably for reasons of scalability, since each destination requires a distinct wavelength, and latency, since destinations must be close to sources to limit the duration of signalling exchanges. The proposed multipath structure avoids these problems and also brings improved spectral efficiency, as explained later.

Though the multipath structure can be seen as an extension to TWIN, it also comparable in some respects to CANON [7], [8]. The main differences are that multipaths are assemblies of trees rather than rings and that their MAC protocol does not impose a rigid frame and slot structure. We believe these differences endow our proposal with greater flexibility and bring improved performance thanks to lower signalling and burst assembly delays.

B. Multipoint-to-multipoint lightpaths (multipaths)

A multipath adjoins a multipoint-to-point lightpath to a point-to-multipoint lightpath, as illustrated in Figure 1.

In the left hand figure, the multipath interconnects a cluster of source nodes on the left to a cluster of destinations on the right. Access to the multipath is governed by a controller C located at the root of the concentration tree. Optical signals are transmitted transparently from source to destination. Each destination receives a composite signal and extracts its own packets after conversion to electronic form. Destinations may be a long way from the source cluster and signal amplification and regeneration may be required.

The right hand figure depicts a multipath that interconnects the nodes of the same cluster. Each node is assumed to be equipped with at least one incoming and one outgoing fibre so that the same wavelength can be used in both directions. This intra-cluster multipath in effect constitutes a MAN.

Each source is equipped with one or more tunable transmitters enabling optical bursts to be sent successively on every multipath to which it has access. Each destination must be equipped with a dedicated receiver for every multipath it terminates. To maintain optical transparency these would need to be burst mode receivers. An alternative, where bursts are converted to continuous mode at the controller, is discussed later in Section II-D.

Cluster size is mainly limited by the lightpath traffic capacity and by the distance between edge nodes and the controller. To ensure sufficiently prompt signalling exchanges, the source nodes are assumed to be geographically close to their controller (within 100 km, say). This assumption is valid in particular for most European countries where Internet access is concentrated in large urban areas. A “cluster” may consist of a single source or destination, if its traffic is sufficiently high.

C. Multipath control

Controller C is responsible for allocating multipath time slots to sources in order to avoid collisions. Bursts in these slots cannot collide at any OXC in the multipath if they are timed not to collide at C. A MAC protocol that avoids collisions while fully using the lightpath capacity is described and evaluated in Section IV.

While data are transmitted transparently from end to end, to perform control operations it is necessary to convert signals to electronic form before processing at C. We therefore constitute dedicated upstream and downstream control channels between controller and source nodes. The upstream channel allows source nodes to communicate their requirements. The downstream control channel is used to inform sources of their allocated time slots.

The same control channel can be used for all multipaths having the same controller. This motivates the decision to constitute sets of multipaths whose sources are all contained in the same cluster. Multipaths of the same cluster and their upstream control channel must use exactly the same fibres. This ensures the control channel has the same propagation time as the multipaths it controls and can be used for synchronization and ranging operations.

Scheduling transmissions of several multipaths at a common controller also has the following significant advantage. It is possible to coordinate time slot assignments to avoid a phenomenon of transmitter blocking. This arises in TWIN when time slots allocated to a source by two or more destinations overlap but cannot all be used because the source has only one transmitter. Loss in achievable utilization can attain 37% [5], [6]. In the present structure, the controller is aware of allocations for all multipaths in a cluster and can avoid both lightpath collisions and transmitter blocking.

D. Continuous mode transmission

Sources emit a succession of bursts on different multipath wavelengths leading to strong requirements on the receivers that must convert the optical signals to electronic form.
limit complexity of destination nodes, we propose to convert transmission from burst mode to continuous mode at the OXC hosting the controller.

Bursts received at this OXC are converted to electronic form and retransmitted in continuous optical mode without any logical processing of the transmitted data. The OXC must be equipped with one burst mode receiver for each wavelength but all destinations then only require a continuous mode receiver for each of their incoming multipaths.

As well as bringing a potentially significant cost saving, this has the advantage of facilitating optical amplification and multiplexing and demultiplexing operations performed at OXC within the long-haul transmission network.

E. Managing a multipath network

Multipaths can be rapidly set up and torn down in the optical infrastructure using a control plane like GMPLS [13]. Reliability can thus be assured using familiar routing and wavelength assignment techniques although additional provision would be required to ensure adequate controller redundancy. A control plane like GMPLS would make it relatively easy to reconfigure the multipaths so as to meet changing traffic requirements.

III. REALIZING A LARGE ISP MULTIPATH NETWORK

We demonstrate how multipaths can be assembled to build a country-wide ISP network with considerably smaller energy consumption than the classical router based solution.

A. An ISP reference network

The reference network handles some 4 Tb/s of traffic and is intended to be representative of a large national ISP. It interconnects 420 edge nodes and provides access to the Internet via 4 gateways. Additionally, 2 peering points provide interconnection to peers or data centres. Each edge node sends a total of 1 Gb/s to the other edge nodes within the WAN. This is assumed to be distributed uniformly over all other edge nodes. Edge nodes also send 1 Gb/s to the Internet via one of the 4 equivalent gateways. Traffic outgoing via the 2 peering points is less than 100 Mb/s per edge node. Nodes receive 4 Gb/s from the Internet via one of the 4 gateways and 2 Gb/s of traffic from the 2 peering points combined.

The reference network is illustrated in Figure 2. Edge routers are interconnected via three hierarchical layers of high capacity transit routers. Internet gateways are connected at the level of regional transit nodes. Peering is realized via the top level core routers. For reasons of reliability, each transit node, gateway and peering point is duplicated (the figure shows only one member of each pair). The routers are interconnected by point-to-point links as follows:

- 10 Gb/s from edge node to metro node,
- 80 Gb/s from metro node regional node,
- 200 Gb/s from regional node to core node,
- 200 Gb/s from core node to core node,
- 280 Gb/s from gateway to regional node,
- 240 Gb/s from peering point to core node.

Links are sized to support the total traffic originating from or destined to the edge nodes with redundant dual connections to ensure reliability. Note that a total of 92 routers (8 core, 14 regional and 70 metro routers) are required to provide full connectivity to the 420 edge nodes.

B. Multipath network

We now replace the transit routers by a set of 10 Gb/s multipaths designed to interconnect edge nodes, gateways and peering points. Each edge node is connected to the optical infrastructure by one incoming and one outgoing fibre. It is equipped with one 10 Gb/s tunable transmitter and one receiver for each multipath it terminates.

Traffic assumptions are as above. We assume the performance of a multipath is satisfactory as long as its load is less than 90%. This is justified in Section IV. Figure 3 shows a small sample of the multipaths required.

For internal traffic, seven clusters are constituted with 60 nodes each. Destination clusters coincide with source clusters. The traffic from one cluster to another is then 8.5 Gb/s and can be handled by one multipath. Intra-cluster multipaths have slightly less traffic than 8.5 Gb/s.

Incoming Internet traffic is handled by point-to-multipoint multipaths from a gateway or peering point to a cluster of edge nodes. Since each edge node receives 4 Gb/s of traffic from the Internet, clusters of 2 nodes result in a traffic of 8 Gb/s.
per multipath. Each peering point sends 1 Gb/s of traffic to each node which results in clusters of 9 nodes with a traffic of 9 Gb/s per multipath.

Multipaths for outgoing Internet traffic connect each 60-node source cluster to a gateway. The 60 Gb/s of outgoing cluster traffic can be handled by 7 such multipaths, distributed over the 4 gateways. The small amount of traffic towards each peering point can be handled by one further multipath from each source cluster.

As in the ISP network, gateways and peering points are duplicated for reliability. The same multipaths can be used to interconnect clusters with each pair, only one gateway or peering point being active at any time. Overall network reliability is assured at the optical infrastructure level. This is possible since multipaths can be reconfigured rapidly in the event of failure. We do not account explicitly for redundant fibre and cross-connects since this is more or less common to both transit router and multipath architectures and consumes relatively little energy.

Edge nodes only generate 2 Gb/s of outgoing traffic so a single tunable transmitter is sufficient for 16 multipaths (6 inter-cluster, 1 intra-cluster, 7 gateway and 2 peering point paths) and the control channel. They must be equipped to receive a total of 9 multipaths (6 inter-cluster, 1 intra-cluster, 1 gateway and 2 peering point paths) and one control channel, i.e., 10 receivers in all. Each gateway or peering point must be able to emit on 53 different channels and to receive up to 13 different multipaths.

C. Comparing energy consumption

To estimate the power consumption of the ISP network, we use the model proposed in [14], taking data from vendor-supplied specifications. IP routers are modeled according to the structure of the Cisco CRS-1 router. Specifically, each router is equipped with line cards (short-reach interfaces) connected to transponders, which perform E/O and O/E conversion. The number of ports, interfaces, and transponders depends on the number of lightpaths each router needs to switch. Depending on the link, we assume that a lightpath has capacity C = 10 or C = 40 Gbps and therefore consider a Cisco CRS Single-Port STM-256c POS Interface Module with the Modular Services Card as line card and the Tellabs 7100 Optical (Ten or Forty Gbps) Transponder Module as transponder.

Table I gives our power consumption estimates for the different network nodes. The indicated power consumption includes air conditioning which we assume contributes 33% of the device consumption.

Table II summarizes the consumption of multipath components. Overall, the increased consumption of the equipment required for the multipath network is less than 123 kW.

In conclusion, the multipath architecture reduces consumption in the considered ISP network by some 27%. The gain is clearly much more significant if we compare only the part that changes: 123 kW instead of 1.7 MW.

D. A future proof technology

Consider now the potential of multipaths to cope with growing demand in the Internet. One could increase the capacity of each wavelength channel or increase the number of channels per fibre.

Suppose a tenfold increase of the traffic generated by each edge node of the ISP network. The same set of multipaths would be sufficient if each operates at 100 Gb/s instead of 10 Gb/s [19].

An alternative approach would be to retain the same 10 Gb/s transmitters and receivers and to create smaller clusters. One
might alternatively allocate more than one wavelength channel to the same source-destination cluster pair. Despite the higher number of wavelength channels required, this solution appears perfectly feasible with present day WDM systems which can deliver up to 160 10 Gb/s wavelength channels per fiber [20].

IV. DYNAMIC ALLOCATION OF MULTIPATH BANDWIDTH

We propose a flexible MAC protocol and demonstrate the excellent performance of a preferred DBA algorithm.

A. Reports and grants

The MAC protocol relies on report–grant exchanges inspired by EPON standards. Each source periodically transmits reports to the controller on the upstream control channel. The controller computes non-conflicting grants enabling the sources to transmit their reported traffic to the multipath destinations and distributes these grants via the downstream control channel. Each grant specifies a start time and a duration.

Grant timing is designed to avoid collisions in the concentration tree bringing traffic from the sources to the controller node and beyond. This is possible using the algorithm described in the next section as long as the controller has precise knowledge of propagation times between itself and the sources. This can be obtained by applying the synchronization and ranging procedure defined in EPON standards (e.g., [6]). This procedure ensures local source clocks are precisely one controller-to-source delay slow with respect to the controller clock and enables the controller to measure the round trip time \(RTT_i\) for each source \(i\).

B. A flexible MAC layer

Grant timing must account for a guard time \(\Delta_g\) between burst emissions necessary notably to allow laser re-tuning. It is also necessary to account for delays in sending grants on the downstream control channel. We suppose the sum of such delays is less than a tolerance \(\tau\).

The process of grants emitted by the controller to the source nodes of a given multipath \(j\) is specified by the functions \(g_j(\cdot), s_j(\cdot)\) and \(d_j(\cdot)\) defined as follows. The \(n^{th}\) grant sent for multipath \(j\) to some source is formulated by the controller at time \(g_j(n)\) and instructs the source to transmit for duration \(d_j(n)\) starting at source local time \(s_j(n)\). Assume the \((n+1)^{th}\) grant is issued to source \(i\). Epochs \(g_j\) and \(s_j\) are calculated recursively, as follows.

\[
g_j(n + 1) = g_j(n) + d_j(n) + \Delta_g, \quad (1) \]
\[
s_j(n + 1) = g_j(n + 1) + \Delta_O - RTT_i, \quad (2) \]

where \(\Delta_O\) is an offset satisfying \(\Delta_O \geq \max_i(RTT_i) + \tau\). It is proven in [6] that this recursion indeed avoids collisions and guarantees the multipath is fully utilized.

Note that the choice of source \(i\) for the \((n+1)^{th}\) grant is so far not specified. To avoid the phenomenon of transmitter blocking that significantly reduces the capacity of TWIN (cf. [5], [6]), the controller ensures the grantee source for one multipath has not previously been scheduled to transmit at that time on another multipath. This is possible since all concerned multipaths are managed by the same controller. When, as normally, several sources are eligible to receive a grant, the controller applies a fair selection algorithm like round-robin.

C. A flow-aware DBA

It is possible to design different DBA algorithms exploiting the above MAC. For instance, one could distinguish classes of service and report queue contents for each class. Grants could then be allocated preferentially to high priority classes. We propose an alternative flow-aware DBA that is arguably preferable for the type of network considered in Section III. Broadly the same algorithm was applied to TWIN in [6].

In this DBA, reports transmit the number of currently active flows for each source and each multipath. Grants are calculated to provide a “quantum” of service for each flow. In addition, at the instants when grants are fulfilled (i.e., at local times specified by \(s_j(n)\)), the source preferentially serves flows that have become active (i.e., they have received a new packet following an inter-packet silence) since the last grant was fulfilled. As explained in [6], this ensures low rate flows, including conversational and streaming flows, are handled with priority and therefore have very low packet latency. High rate data transfers, on the other hand, are selected according to a round-robin scheme, thus receiving a max-min fair share of multipath bandwidth.

D. Performance

We report some simulation results for a multipath network implementing the above flow-aware DBA. We consider a cluster of 60 source nodes sharing 16 multipaths, as in the network of Section III. The guard time is \(\Delta_g = 100\) ns, delay tolerance is \(\tau = 1\) ms, multipath bandwidth is 10 Gb/s and round trip times are evenly distributed between 0 and 1 ms \((\leq 100\) km between source and controller). Reports are issued with a period of 1 ms.

Traffic is symmetric for all sources and all multipaths and is made up of Poisson arrivals of flows of two types. “High rate flows” are assumed to place all their packets in a (virtual) per-flow queue on arrival\(^2\). “Low rate flows” emit packets individually at a rate of 2 Mb/s. Flow sizes are 10 MB and 7.5 MB, respectively, and have an exponential distribution. We presents results when 80% or 40% of traffic comes from the high rate flows. The per-flow quantum is 1000 bytes.

Figure 4 plots the mean packet delay of low rate flows as a function of multipath load (flow arrival rate \(\times\) mean flow size / multipath rate). The figure shows that delay is dominated by the signalling time until load is very close to one\(^3\). Delay decreases with increasing load because newly active low rate flows “steal” transmission time previously allocated for high rate flows (cf. Sec. IV-C).

\(^2\)In practice, the rate of these flows is controlled by a protocol like TCP that ensures the flow maintains enough packets in the buffer to fully use its fair share.

\(^3\)This time is composed as follows: a packet arrives; this arrival is signalled in the next report (average time \(5\) ms); the report arrives after a one-way propagation time (average \(25\) ms); a grant is formulated immediately and in view of (2) the sending time is set \(2\) ms offset – one-way propagation time = \(1.75\) ms on average.)
We believe the present approach has considerable potential beyond the context considered here and, in future work, will consider the adaptations needed to meet the respective networking requirements of data centres and large tier-1 providers.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the EU 7th Framework Programme under grant no. 257740 (Network of Excellence TREND).

REFERENCES