Dynamic bandwidth allocation for all-optical wide-area networks

Davide Cuda, Raluca-Maria Indre, Esther Le Rouzic, James Roberts

Dynamic bandwidth allocation for all-optical wide-area networks †

Davide Cuda[1] and Raluca-Maria Indre[1] and Esther Le Rouzic[1] and James Roberts[2]

[1] *Orange Labs Issy-les-Moulineaux, France*
[2] *INRIA Paris-Rocquencourt, France*

Nous proposons une architecture pour un réseau WAN tout-optique basée sur la notion de connexions optiques multipoint à multipoint, que nous appelons *multipaths*. L'ensemble des noeuds d'accès est partitionné en groupements pour l'émission et la réception. Une (ou plusieurs) longueur d'onde est allouée à chaque paire groupement-source groupement-destination. Les noeuds d'un même groupement-source se partagent la capacité de cette longuer d'onde selon un protocole MAC adapté. Les données transmises sur une longueur d'onde sont diffusées à tous les noeuds du groupement-destination, et chaque récepteur extrait alors les données qui lui sont destinées à partir du flux reçu. Le réseau que nous proposons ne nécessite que des composants existants et se compare favorablement en termes de complexité et d'efficacité énergétique à des solutions alternatives comme la commutation optique par paquet (Optical Packet Switching - OPS) ou la commutation optique par rafale (Optical Burst Switching - OBS). Nous présentons d'abord l'architecture multipath et comparons sa consommation d'énergie celle d'un réseau classique à base de routeurs. Nous proposons ensuite un algorithme d'allocation dynamique de la bande passante. Nous évaluons la performance de l'algorithme proposé à l'aide de simulations et nous montrons que notre solution permet d'atteindre d'excellentes performances en terme de délai et temps de réponse.

Keywords: all-optical network, wide area network, dynamic bandwidth allocation, flow-aware networking.

1 Introduction

In today's networks, optical technologies such as Wavelength Division Multiplexing (WDM) are used to provide high-capacity point-to-point lightpaths consisting of a wavelength channel carried over a succession of fibres. As each wavelength has 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 networks, traffic is usually concentrated through a hierarchy of electronic transit routers. In this paper we propose a solution for performing lightpath sharing in the optical domain. Previous proposals for lightpath sharing include both point-to-multipoint and multipoint-to-point sharing. Point-to-multipoint lightpath sharing can be realized using splitters and optical time divison multiplexing as proposed in [1]. Multipoint-to-point lightpaths are obtained by merging signals from two or more incoming fibres on to a single outgoing fibre. This principle is used notably for upstream transmission in passive optical access networks (PONs). The capacity of the upstream channel is shared by means of a dynamic bandwidth allocation (DBA) algorithm.

We propose to use DBA to control sharing of multipoint-to-multipoint lightpaths, called *multipaths* for short, interconnecting edge nodes in a wide-area network. The proposed network uses presently feasible optical technology, namely tunable transmitters, burst mode receivers and wavelength selective cross-connects, to replace electronic routers in a country-wide ISP network.

We present the envisaged multipath infrastructure before proposing an EPON-inspired MAC protocol to control bandwidth sharing. We describe the scope for flexibly defining DBA algorithms within this framework and present some performance evaluations illustrating the attractiveness of a preferred realization.

† The research leading to these results has been partially funded by the FP 7 TREND Network of Excellence.

2 From PON to WAN : extending the reach of dynamic bandwidth allocation

Proposed DBA algorithms are realized using either EPON or GPON standards. Our proposal derives mainly from EPON solutions where the OLT attributes to its ONUs non-conflicting time slots defined flexibly in terms of a start time and a duration. We believe this is sufficient to meet requirements and is preferred to frame-based GPON solutions for reasons of simplicity.

The distributed control implementation of TWIN proposed by Saniè and Widjaja [2] employs a form of dynamic bandwidth allocation. Each destination has its own wavelength and receives light signals from source nodes via a tree network created over an optical infrastructure. The destination grants source nodes time slots on its particular wavelength tree in much the same way an OLT grants time slots to its ONUs. The algorithm ensures bursts from remote sources do not collide within the network or at the destination. An EPON-like MAC protocol for TWIN and an associated DBA algorithm are described in [3].

TWIN is not suitable for a wide-area network. It has limited scalability since a distinct wavelength must be assigned to each destination. Its geographical span must be relatively small to limit delays due to report/grant signaling. While time slot collisions are avoided within each destination tree, destinations act independently and blocking can occur (i.e., grants cannot be completely fulfilled) when time slots overlap at a source, significantly reducing traffic capacity. On the other hand, a single destination may not have sufficient traffic to justify a dedicated multipoint-to-point lightpath.

Our proposed network of multipoint-to-multipoint lightpaths, or *multipaths*, avoids these problems [4]. A cluster of sources shares a wavelength tree as in TWIN except that the root of the tree is not the destination but a *controller* that attributes slots to each source. The tree is extended beyond the controller towards a cluster of destinations. Light in this part of the multipath is split and amplified as necessary to ensure each destination receives a copy of the combined signal from the source cluster. Each destination converts this signal to electronic form and retains its own packets (as downstream transmission is performed in a PON). Only the control signals are converted to electrical form at the controller.

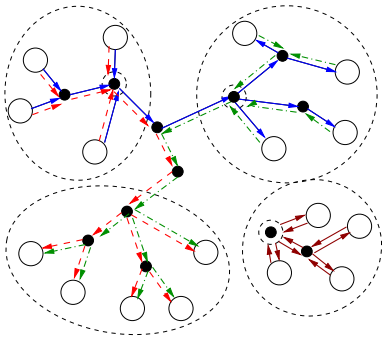

FIGURE 1: Network of multipaths - showing just 4 multipaths.

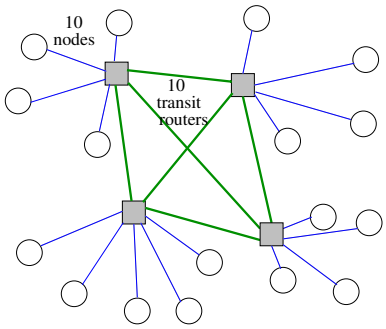

FIGURE 2: Traditional network using transit routers.

Figure 1 illustrates a sample of the multipaths needed to interconnect a toy network of 100 edge nodes. Each node is assumed to send an average of 9 Gb/s distributed uniformly over all the other nodes. We form clusters of 10 nodes and create multipaths to interconnect them. Intra-cluster multipaths are also created, as in the lower righthand corner of the figure. Multipaths are assumed here to have a capacity of 10 Gb/s, sufficient to handle inter- and intra-cluster demand. The controller in each source cluster coordinates sharing on all its multipaths and can thus avoid blocking. It is close to the source nodes ensuring negligible signalling delays. Each edge node is equipped with one receiver for each of 10 incoming multipaths and one tunable transmitter.

Figure 2 represents a traditional network interconnecting the same 100 edge nodes using 10 electronic transit routers to concentrate traffic onto high capacity point-to-point optical links. The multipath network avoids this transit layer leading to a significant economy in equipment and energy. A more realistic example

is worked out in [4], demonstrating that the multipath approach readily extends to a country-wide ISP network of 420 edge nodes. It realizes a power saving of 1.7 MW by eliminating a total of 92 metro, regional and core transit routers.

3 Slot allocation : a MAC protocol ensuring maximal utilization

All multipaths accessed by a source cluster share the same fibres up to the controller. A further wavelength on these fibres is used to constitute upstream and downstream control channels for report and grant signalling, respectively. The control channels are also used for synchronization and ranging. The EPON time stamp exchange protocol is used, allowing the controller to measure the round trip time RTT_i between itself and each source i of its cluster and synchronizing each source clock to the controller clock time *plus* an unknown one-way propagation delay (see [3]). These data are used in the following grant timing algorithm.

Grant timing must account for a guard time Δ_g between burst emissions and for grant signalling delay that we suppose is bounded by some value τ. In evaluations reported below we assume $\Delta_g = 100$ ns and $\tau = 1$ ms. The process of grants emitted by the controller to the source nodes of a given multipath j is specified by the functions $g_j(.)$, $s_j(.)$ and $d_j(.)$ 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.

$$g_j(n+1) = g_j(n) + d_j(n) + \Delta_g, \quad (1)$$

$$s_j(n+1) = g_j(n+1) + \Delta_O - \text{RTT}_i, \quad (2)$$

where Δ_O is an offset satisfying $\Delta_O \geq \max_i(\text{RTT}_i) + \tau$. It can be proved that this defines a feasible schedule ensuring maximal lightpath utilization (see [3]).

The choice of source i for the $(n+1)^{th}$ grant is not specified in the above recursion. If the choice were made independently for each multipath, grants might overlap leading to capacity loss due to transmitter blocking. This can be avoided here since, as proved in [4], the controller can always select a source i for the $(n+1)^{th}$ grant that is known to have an available transmitter at the calculated start time $s_j(n+1)$.

4 Reporting and granting : a flexible choice of DBA algorithm

As for a PON, it is possible to design a wide variety of DBA algorithms using the above MAC protocol. These are distinguished by the content of reports and the manner in which grants are attributed. We briefly outline the scope for choice before presenting the algorithm we recommend for a wide-area ISP network.

Reports for all multipaths of a source cluster are delivered using a common signalling channel. Management of this channel is flexible beyond the requirement that the arrival of new traffic must be signalled in a timely manner to the controller. Each report indicates the current status of its packet queues, distinguishing flows of diverse granularity and/or classes of service, as required by the implemented service model.

Grants are calculated at the instants g_j defined above using the latest reports received for the considered source and multipath. The source may be selected in simple round-robin order or by using a more complex state dependent algorithm, accounting for a given priority policy, say (while avoiding blocking between multipaths, as previously discussed). It is also possible that grants be attributed without prior reports, as envisaged in the GPON "non-status reporting" option.

Our preferred approach for the wide-area network is a flow-aware DBA. We assume application flows can be reliably identified and that sources report the current number of active flows, i.e., flows having at least one packet in the multipath buffers. The controller calculates grant time slot durations d_j to accommodate one "quantum" for each such flow. Sources are given grants in round-robin order. They use the grants to serve active flows also in round-robin order. To give priority to low rate flows, any flow that newly becomes active is placed at the head of the round-robin schedule. This scheme realizes network-wide, per-flow fair bandwidth sharing with well-known traffic control advantages (e.g., see [5]). It minimizes the delay of low rate flows ensuring streaming audio and video flows, that naturally fall into this category, see negligible packet latency.

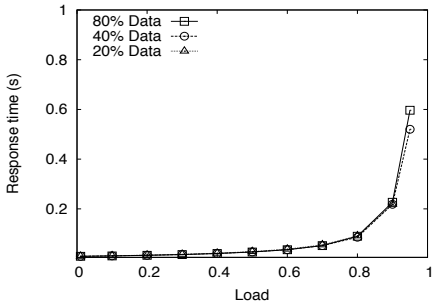

FIGURE 3: Mean response time of data flows .

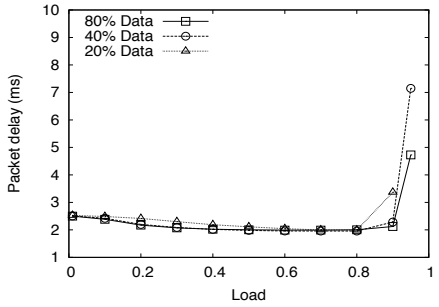

FIGURE 4: Mean packet delay of video flows.

5 Performance of the flow-aware DBA

We have simulated the toy network of Figure 1 under the following traffic model. Flows arrive according to a Poisson process and are one of two types : high speed data flows able to transmit at up to the 10 Gb/s multipath rate or video streams emitted at the constant rate of 2 Mb/s. The size of data flows is 10 MB on average and has an exponential distribution. The duration of video flows is also exponential with a mean of 30 s. The flows share multipath bandwidth according to the flow-aware DBA scheme outlined above. Each source has a single tunable transmitter. Network parameters are $\tau = 1$ ms, $\Delta_g = .1\mu$s, $\text{RTT}_i = i \times 100\mu$s for $i = 1, \ldots, 10$.

Figure 3 shows the mean response time of data flows for a traffic mix with 20%, 40% or 80% of data. We see that this parameter has limited impact on performance which is excellent until demand attains multipath capacity (load$\rightarrow$1). Figure 4 plots the delay experienced by packets of the video flows. These packets suffer an average delay of 2.5 ms at very low load due to the signalling exchange (1-2) and the assumed control channel implementation. Delay then decreases due to the priority mechanism before exploding when demand attains capacity.

6 Conclusion

The present proposal demonstrates that is possible to do without costly and power hungry electronic transit routers in a wide-area network using presently available optical technology. Specifically, we propose to use available technology such as rapidly tunable transmitters, burst mode receivers and cross-connects to create multipoint-to-multipoint lightpaths called multipaths. Multipath bandwidth is shared between source-destination traffic flows under the control of a MAC protocol designed to avoid collisions at merge points and destinations. The use of DBA in this context is an original approach to all-optical networking that we believe has considerable potential. We are currently working on applying this approach to interconnect servers in a datacenter network.

Références

[1] P. Petracca, M. Melia, E. Leonardi, F. Neri Design of WDM networks exploiting OTDM and light-splitters In *Proc of QoS-IP*, 2003.

[2] I. Saniee, I. Widjaja. Design and performance of randomized schedules for time-domain wavelength interleaved networks. *Bell Labs Technical Journal*, 14(2), 2009.

[3] P. Robert and J. Roberts, "A flow-aware MAC protocol for a passive optical metropolitan area network," ITC 23, 2011.

[4] D. Cuda, R-M. Indre, E. Le Rouzic, J. Roberts, "Getting routers out of the core : Building an all-optical network with 'multipaths'," In *Proc of ONDM*, 2012.

[5] S. Oueslati, J. Roberts A new direction for quality of service : Flow aware networking. In *Proc of NGI*, 2005.