Design of an Agile All-Photonic Network

Gregor v. Bochmann
School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada
e-mail: bochmann@site.uottawa.ca

ABSTRACT

"Agile All-Photonic Networks" (AAPN) is the theme of a Canadian research collaboration. An AAPN is a wavelength-division-multiplexed network that consists of several overlaid stars formed by edge nodes that aggregate traffic, interconnected by bufferless optical core nodes that perform fast switching in order to provide bandwidth allocation in sub-wavelength granularity. Specific issues addressed in this context are (a) efficient bandwidth allocation, (b) routing of MPLS flows over the AAPN, (c) allocation of protection paths, and (d) development of a demonstration prototype. This paper high-light research results and design choices related to these issues.

Keywords: Optical networks, transparent switching, optical time division multiplexing, overlaid star architecture, bandwidth allocation, multi-area optimal routing, protection path allocation, segmented protection.

1. INTRODUCTION

Recently, there has been much interest in the development of wavelength-routed optical networks where wavelength division multiplexing (WDM) is used together with agile photonic switches to dynamically establish transparent lightpaths (without E-O-E conversion) between different the different network nodes. These lightpaths that have a typical bandwidth of 10 Gbps, can be used for the transmission of IP packets between powerful Internet routers or switches that located at the different network nodes.

The objective of the research project "Agile All-Photonic Networks" [1, 2] was to explore the possibility of sharing the large bandwidth of a wavelength in the time domain, thus providing transparent optical transmission from end-to-end for effective bandwidths of small fractions of 10 Gpbs, down to the order of 100Mbps. The idea is to extend the transparent optical transmission as close to the end user as possible. For instance, it is thought that an edge node of an agile all-photonic network (AAPN) of the type developed within our project could be located within an office building or a residential complex and provide to a large number of local users a multitude of lightpaths to the other remote edge nodes within the network. It is to be noted that the proposed architecture of an AAPN is suitable for a metropolitan context as well as for a wide-area network spanning a continent. Each edge node of an AAPN would normally be integrated with an IP/MPLS router that provides the link between a local area IP network and the AAPN. Through a local interface, this router would request the establishment of lightpaths with routers located at other edge nodes as a function of the current traffic load.

The sharing of the bandwidth of a wavelength between a large number of lower-capacity flows with different destinations require very fast optical switching, sometimes called optical time division multiplexing (OTDM). For the proposed AAPN architecture, we assume that individual time slots of 10 microseconds duration are routed to their particular destination; this requires a switching speed of the order of one microsecond. Within our AAPN research project, which is a collaboration between five universities, much effort is spend on developing optical components for switching, amplification and transmission impairment compensation that are capable of adjusting their behavior within a time frame comparable with the switching time of one microsecond. The other main effort is related to the development of a suitable network architecture, optimal topological design and dimensioning of AAPNs, and control procedures for bandwidth allocation, fault recovery, and IP/MPLS routing over the network.

In this paper, we first present the main architectural features that have been retained for an AAPN, and then we present some results obtained in our research group at the University of Ottawa in relation with bandwidth allocation, IP routing and allocation of protection paths. Finally, we describe the design of an AAPN prototype that is currently being built at the University of Ottawa.

2. ARCHITECTURE

In order to avoid optical memory and optical header recognition, as required by certain forms of burst switching, we have adopted synchronous slot-by-slot switching which uses fixed-size slots and the arrival of the slots at the input ports of the switch must be synchronized with the slot switching period controlled by the switch. In a mesh network, were a slot transmitted by an edge node must possibly traverse several switching nodes, the propagation delay on the links between several switching nodes must be precisely adjusted to be a multiple of the slot duration. Since this is difficult to realize, we have adopted a star topology with a single core switch (see also the Petaweb proposal [3]). In this case, the only synchronization requirement is that the edge node transmits the next slot at such a time that it arrives at the core switch in time for the beginning of a slot period. This can be realized by a relatively simple synchronization protocol between the edge nodes and the core.

In order to cope with the failure of a core node, and in order to increase the overall capacity of the network, we propose to use an architecture of overlaid stars, as shown in Figure 1(a). Furthermore, since we would like to allow for a large number of edge nodes, say of the order of 1000, we propose that a tree-like architecture could be used, as indicated in Figure 1(b). In this case, the optical wavelengths supported by a given port of the central core would be shared among M edge nodes through a 1xM selector switch. Different versions of this architecture are discussed in more detail in [4]. As a typical configuration, we have considered the use of a 64x64 core switch and 1x16 selector switches supporting 1024 edge nodes. It is to be noted that each supported wavelength requires in general a separate space switch in the core, as shown in Figure 1(c), since the different wavelength would support in general different information flows for various source-destinations pairs.

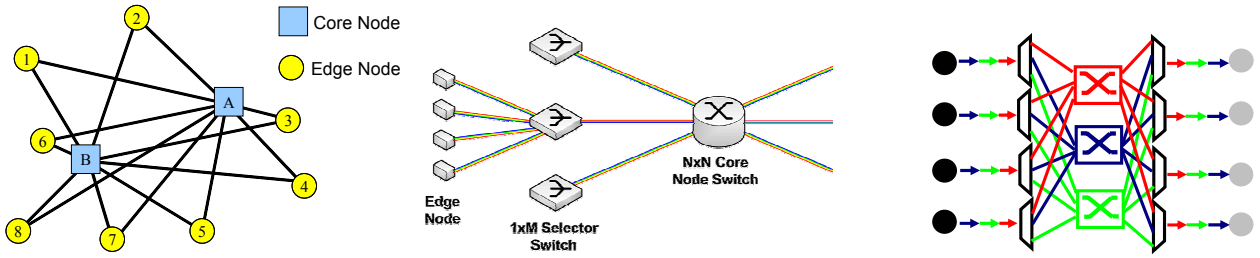


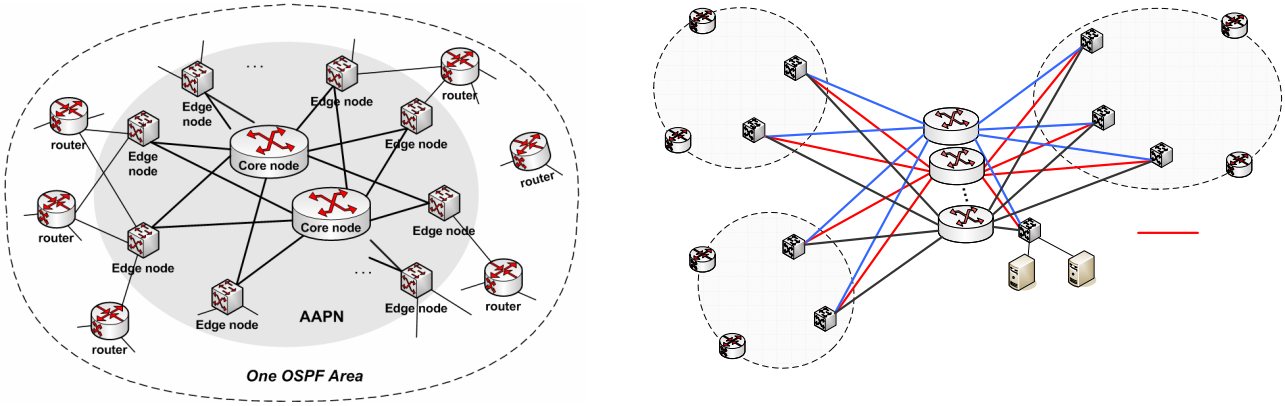
Figure 1: (a) Overlaid star architecture. (b) Tree-like architecture with selector switch. (c) Core switch with three space switches supporting three wavelengths

It is to be noted that the AAPN architecture is relatively scalable. In a small version with a single core, one wavelength and 64 edge nodes, each edge node may simultaneously establish 100 packet flows with an average of 100Mpbs each. This means that such an edge node can support a very large number of users or several servers. In a large configuration involving 16 wavelengths, two 64x64 core switches and 1x16 selector switches, we have 1024 edge nodes, each simultaneously supporting 200 packet flows of 100Mbps.

As mentioned earlier, an edge node provides the interface between the electrical Internet and the transparent optical AAPN. Each edge node includes (electronic) buffers for storing data slots before they can be transmitted through the core node. These are so-called virtual output queues, since there is logically a separate queue of data slots for each destination edge node. The transmission schedule is determined by a master edge node that resides at the site of the core switch and receives bandwidth allocation requests from all the edge nodes. It calculates a suitable transmission schedule and transmits this information in advance to the core nodes, which apply this schedule according to the timing information provided by the synchronization protocol.

The edge node also contains the functionality of slot aggregation, that is, it receives IP and/or MPLS packets, determines the corresponding destination edge nodes and stores them in the next available slot in the output queue of that destination. On reception of a slot from the core, the edge nodes extracts the contained packets and delivers them to the local routing function which forwards the packet through the local Internet. In the case that the AAPN has several core nodes, each edge node will also include a routing function which determines over which core node any given traffic should be routed.

Figure 2(a) shows an AAPN together with some local Internet networks connected to its edge nodes. End users and servers would typically be connected to the AAPN through such a local area network. An AAPN can be used as the backbone area in an OSPF multi-area network scenario as discussed in Section 4. It can also be used as an Internet Exchange, as shown in Figure 2(b) (see [5] for further details).



(a) (b)

Figure 2: (a) Example of an AAPN with two core switches and surrounding Internet routers. (b) Example of an AAPN used as an Internet Exchange; there could be content servers (e.g., IPTV server) directly attached to a "server edge node" to distribute contents efficiently

3. BANDWIDTH ALLOCATION

Bandwidth allocation within the AAPN should be agile, adapting fast to any changing traffic demands. In traditional time-division multiplexing (TDM), a frame structure fixed length is used and time slots are allocated within each frame. Traditionally, the allocation of slots for the different flows within a frame remains constant over a longer period until a new call request comes in. This type of call-by-call allocation has the advantage that the bandwidth allocation algorithm must be executed only when a call starts or terminates. Another scheme, which we call frame-by-frame, calculates the slot allocation for each new frame based on the current traffic demand. Finally, one may consider a slot-by-slot allocation scheme where an edge node would make a request for the transmission of each individual slot.

The slot-by-slot scheme is clearly the most agile bandwidth allocation scheme, but it also requires the largest amount of signaling overhead. In addition, we have to consider that in the long-haul context the propagation delays between the edge nodes and the core are very large compared to the slot duration. A propagation delay of 10 milliseconds (corresponding to a distance of approximately 2000 km) corresponds to 1000 slot durations, and with a frame length between 100 and 1000 slots, is of the same order as the frame period. This means that the agility is never larger than twice this propagation delay. For this reason, we have studied in more details methods for frame-by-frame allocation within a single star network. If the AAPN contains several core nodes, each core node would have its own bandwidth allocation. This is important to obtain independence in case of failure.

We first compared (a) slot transmission without reservation (burst switching) combined with retransmission in case of collisions with (b) simple TDM. Figure 3 (taken from [6]) shows the transmission delays, including queuing, for different traffic loads. Since there were no major differences in traffic delay, we were encouraged to study frame-by-frame allocation in more details.

The problem of bandwidth allocation within a transmission frame has in principle been solved by the Birkhoff-von-Neumann (BvN) algorithm which provides the allocation of slots within a frame when an acceptable traffic matrix is given. In order to apply this approach to the context of an AAPN, we had to solve the following two problems:

- Find a method for obtaining an acceptable traffic matrix from the given traffic demands (obtained from the edge nodes) which may very well oversubscribe the network resources.

- Find an efficient heuristic algorithm to replace the BvN algorithm with its large complexity of $O(N^{4.5})$, where N is the number of edge nodes.

Cheng Peng found solutions to these problems by proposing a projection method for calculating an acceptable traffic matrix [7] and by proposing a heuristic "Quick BvN" algorithm [8] which can be implemented with complexity O(NF) if the word length of the computer is longer than N. (Here F is the length of the frame).

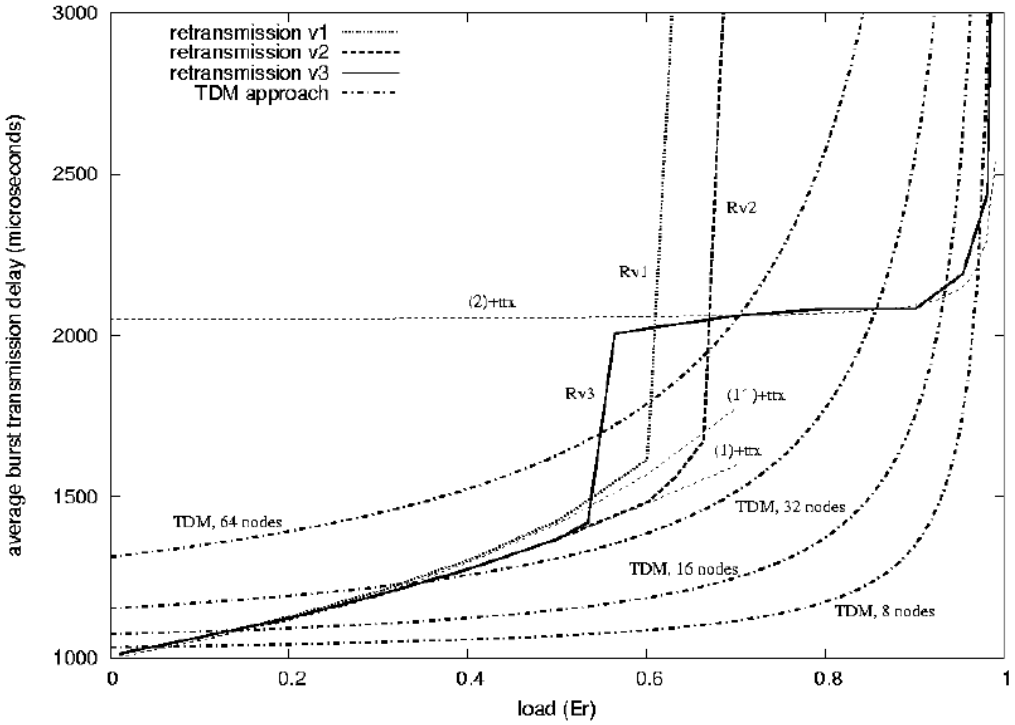
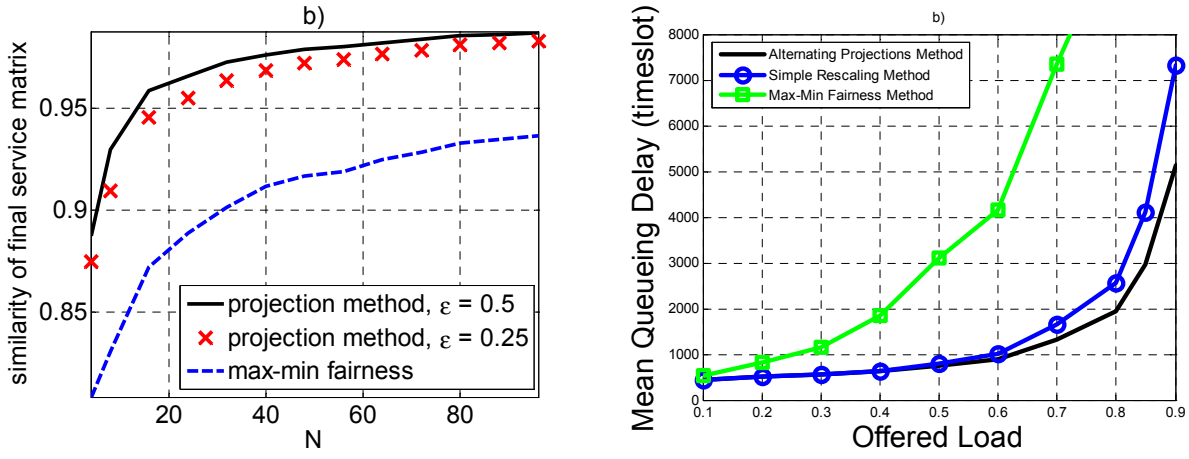


Figure 3: Average slot transmission delay as a function of total offered load to the network per destination; the delays for TDM depend on the frame size (8, 16, 32 or 64) – longer frames imply longer average waiting times; three different procedures for retransmission in case of a loss are shown in the figure (for details see []; the strong increase of the delay for a load around 0.6 is due the fact that the traffic, including the load of retransmission, gets close to 100%; the retransmission scheme v3 goes into the advanced reservation mode when this load is reached.

The performance of the projection method for service matrix construction and the Quick BvN bandwidth allocation algorithm has been studied in detail through simulations (for details, see [7]). The following figures show the major results. In Figure 4, the performance of the matrix construction method is shown. The results show the superiority of our method in respect to the similarity of the resulting service matrix with the given traffic matrix (this relates to the fairness of the approach) and in respect to the queuing delays that occur at higher traffic loads.

A performance comparison of the Quick BvN algorithm with the optimal BvN algorithm and another heuristics (called Greedy Low Jitter Decomposition, GLJD) is shown in Figure 5 (for more details, see [8] and [9]). It shows the mean queuing delay as a function of the offered load. These results indicate that the heuristic Quick BvN provides practically the same performance as the exact BvN algorithm and it is much better at low and medium traffic than the GLJD heuristics. The results in figures 4 and 5 have been obtained for self-similar traffic and an wide-area AAPN configuration with 16 nodes and 1000 km optical links between the edge nodes and the core.



(a) (b)

Figure 4: Performance evaluation of the projection method for the construction of an acceptable service matrix from a given traffic matrix. (a) Similarity of the resulting service matrix with the given traffic matrix as a function of the number of nodes in the AAPN; comparison with the so-called min-max method. (b) Mean queuing delay as a function of traffic load using the optimal BvN bandwidth allocation algorithm; comparison with min-max method and simple rescaling method.

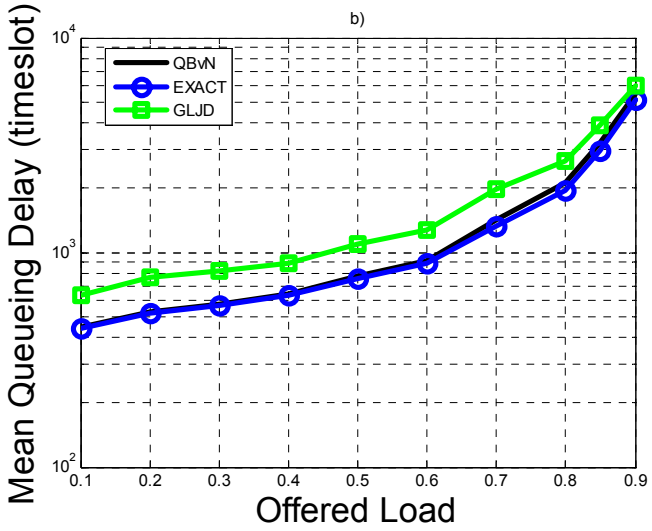


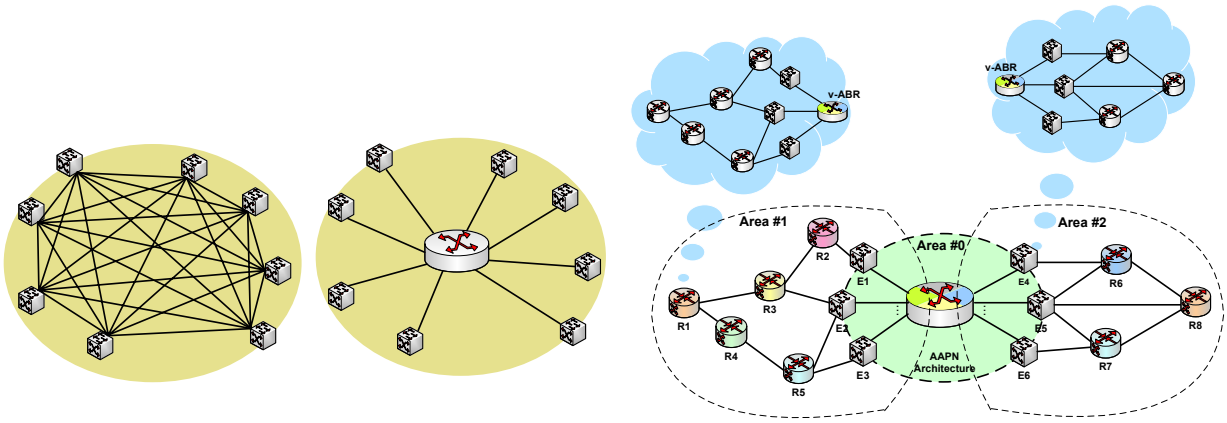
Figure 5: Performance evaluation of the Quick BvN heuristic bandwidth allocation algorithm. The average queuing delay is shown as a function of the offered load; the lines for Quick BvN and BvN coincide; the GLJD heuristic has larger delays.

4. ROUTING OF IP/MPLS PACKETS AND ALLOCATION OF PROTECTION PATHS

As shown in Figure 2, it is expected that an AAPN would be used in a metropolitan context or as a wide-area network for interconnecting many local Internet networks containing many users and servers, and probably also the Internet backbone. Since the AAPN provides transparent optical transmission between all pairs of edge nodes, the architecture, as seen by the Internet routers within the surrounding networks would normally be a completely interconnected mesh, as shown in Figure 6(a). With 1000 edge nodes, this would involve 10^6 links, which is a number much too large to be handled by normal Internet routers.

Peng He has therefore proposed that the AAPN configuration seen by the Internet routers should be a star with a (virtual) router at the place of the core [10], as shown in Figure 6(b). Such a router does not exist at the AAPN core node, but the routers at the edge nodes may project such a vision to the other routers in the connected local Internets. This routing architecture may be implemented by modifying the routers associated with the edge nodes in such a way that they present this virtual router to the other routers in their local environment, and communicate with the routers associated with the other edge nodes by a specially adapted routing protocol that takes into account the bandwidth allocation in the AAPN and other traffic engineering parameters.

In the case that OSPF is used within the whole AAPN environment, the multi-area option of OSPF may be adapted in such a manner that the backbone area (area 0) collapses into the virtual core router and the other independent areas extend up to the virtual core router, as shown in Figure 6(c). This configuration present the advantage that optimal end-to-end routes can be easily established by simply concatenating optimal routes to/from the core, which can be determined by the source and destination sub-areas independently of one another. This problem of finding optimal end-to-end routes can in general only solved by considering global knowledge; in our architecture with a virtual core router, no global knowledge is required, only the local routing information within each non-backbone area.



(a) (b) (c)

Figure 6: Virtual views of the AAPN architecture from the perspective of the IP routing layer. (a) Fully interconnected mesh corresponding to the optical lightpaths that are provided by the AAPN. (b) Proposed view including a virtual router (that does not really exist) at the core of the AAPN. (c) Applying the OSPF multi-area topology in a non-conventional manner to the AAPN, using the virtual view in (b); the local areas associated with the edge nodes extend to include the virtual router at the core; each area only has its local view (as shown above) and an optimal route is obtained by concatenating two routes that are optimal in the source and destination area, respectively.

A scenario of signaling for the establishment of an end-to-end path is shown in Figure 7. Up to the virtual core router, the source area routing information is used to determine an optimal path from the source to the virtual core router. The route establishment request reaches an edge node that is part of the same area (E2) which will communicate with an edge node of the destination area (E4). The latter will find an optimal path from the virtual core node to the destination using the routing information of the destination area. Then this information will be passed back to the originating edge node E2 and also to the edge node on the route within the destination area (E5). The latter will then forward the request to the destination along the calculated route and finally, a confirmation is return to the source along the established path.

This virtual routing IP routing architecture is also useful for establishing protection paths for data flows that require high reliability. Instead of using link or path protection, He [11] pointed out that a protection approach using shared segment protection can take advantage of the multi-area routing architecture with virtual core router. Basically, the idea is illustrated in Figure 8 which shows the working path in black consisting of two half working paths within the two non-backbone areas. Each of the areas has to select a protection path up to the edge node in the other area, consisting a single or several protection segments (Area #1 uses two segments in Figure 8). These protection paths necessarily go through a different edge node of the same area, and within the AAPN they are allocated to a different core node than the working path. As in the case of optimal inter-area path selection, the optimization issues for the protection path can be handled independently in the source and destination areas. Simulation studies have been made to evaluate the efficiency of this

protection path allocation scheme [11]. They confirmed that this scheme leads to low blocking probabilities and efficient sharing of protection bandwidth between multiple working paths.

It is interesting to note that the routing and protection schemes described above can also be adapted to the case where the AAPN interconnects local Internets that belong to different organizations, as in the case of an Internet Exchange [5]. In this case each local Internet would communicate with the outside typically through BGP. The same optimal end-to-end path allocation scheme can be performed by adapted Internet routers in the edge nodes that exchange BGP-specific routing information with one another over the AAPN.

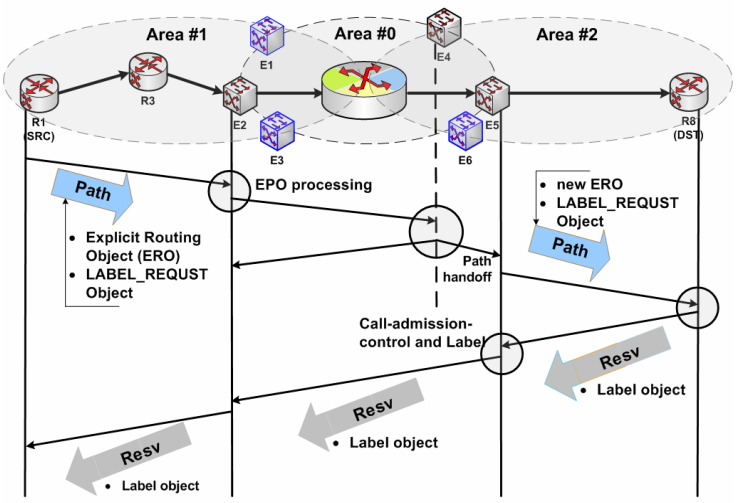


Figure 7: Scenario of signaling messages exchanged during the establishment of an optimal end-to-end path

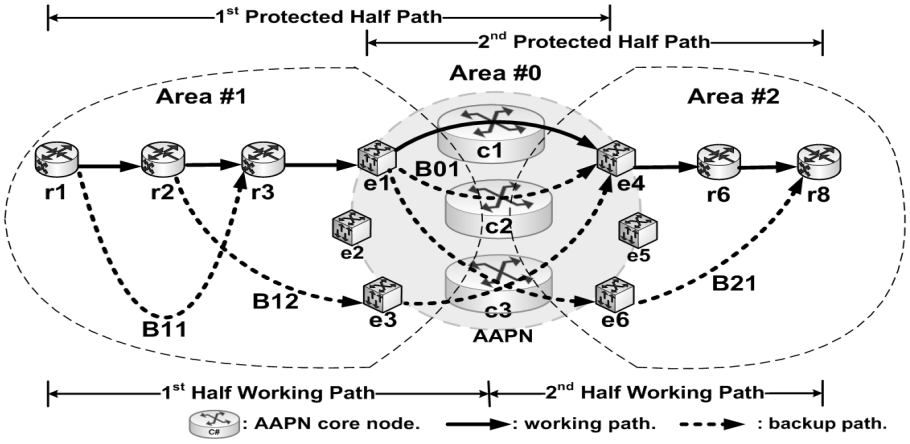


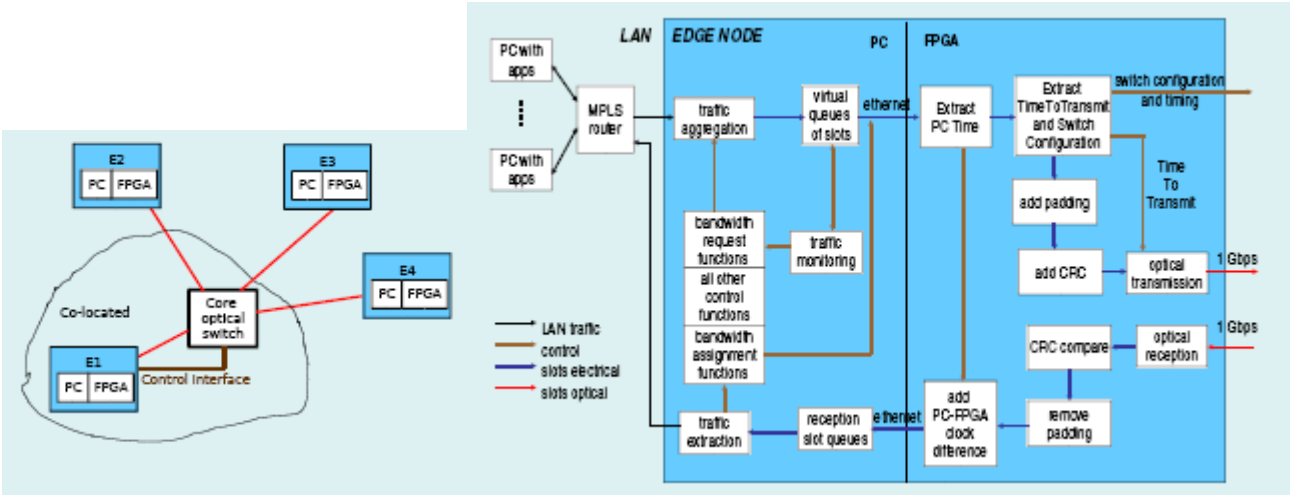
Figure 8: Inter-area shared segment-based protection scheme.

5. PROTOTYPE DEVELOPMENT

In order to demonstrate the feasibility of the construction of an AAPN and for having a context in which we can experiment with the developed AAPN control algorithms, we have embarked in the construction of an AAPN prototype. Our emphasis here was the experimentation with a small AAPN in a metropolitan context, demonstrating transmission, switching and control procedures. The work for this prototype involved hardware and software development.

We first built in software an AAPN control platform that runs in the edge nodes and the master edge node that controls the core switch, assuming that suitable hardware components for transmission and switching would be available. This was tested with a transparent optical switch of very low switching speed and optical Ethernet transmission.

Since then we have developed transmission hardware with burst-mode receivers (very short receive clock synchronization) at 2 Gbps and a 2x2 core switch with a switching speed of less than a microsecond, but a limited switching frequency of 5000 per second (which leads us to a slot period of 200 microseconds). The edge node of this version of AAPN consists of the transmission unit and some of the slot buffering functions implemented on an FPGA and the remaining control functions implemented in software on a connected PC, which also includes the IP/MPLS routing functions and presents the interface to the local Internet. The overall architecture of this prototype is shown in Figure 9(a) and the functional units within an edge node are shown on Figure 9(b).



(a) (b)

Figure 9: AAPN demonstration prototype. (a) Overall architecture including a master edge node E1 and three other edge nodes. (b) Functional units within an edge node.

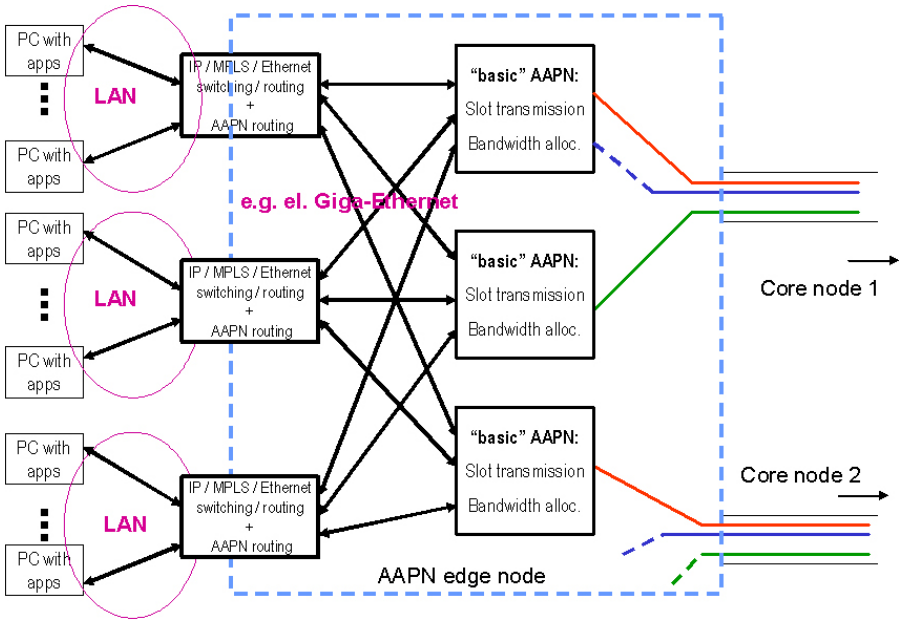


Figure 10: Hypothetical modular architecture of an edge node supporting several wavelength and two core nodes

Figure 10 shows our view of a modular structure for an edge node that supports many wavelengths and two core nodes. It shows the "basic AAPN" functional unit which looks after the control of one wavelength for one core node, and this functional unit is repeated for each additional wavelength and core node. All these "basic AAPN" units exchange IP/MPLS packets with modified Internet routers that select the core node and wavelength for the incoming traffic and perform the routing functions described in Section 4, including protection switching and load balancing between the different core switches [12]. A single modified router may be sufficient, but several instances may be foreseen for capacity reasons.

The software structure of the control platform reflects the different levels of protocols that are required for deploying an AAPN. As already alluded in the previous discussions, the following protocol layers can be identified:

- **IP traffic layer** (Internetwork and application layers): It deals with traffic generators, multimedia applications, MPLS traffic engineering, interconnection of several AAPNs, integration of an AAPN into the Internet, as discussed in Section 4.
- **Global AAPN layer** (Subnetwork layer): It deals with routing, traffic monitoring, load sharing, traffic admission, fault detection and restoration within the AAPN. It is especially important if more than one core node is present.
- **Single-Core AAPN layer** (Link layer): It deals with time slot allocation and switching within a single AAPN star network; and also with the aggregation of packets [13] into bursts that will be transmitted within given time slots.
- **Configuration and Synchronization layer**: It deals with integrating new edge nodes into an active AAPN and conveying the necessary timing information for slotted transmission.
- **Transmission** (Physical layer): It deals with slotted transmission of data bursts at precisely specified transmission times, and burst reception.

The Single-Core AAPN layer of our prototype contains a generic signaling protocol through which the edge nodes exchange traffic requests and allocation responses with the master edge node at the core. We have defined a standard interface at the edge node and the master for exchanging this information in order to facilitate the incorporation of different experimental bandwidth allocation algorithms in the master edge node. Currently, our prototype uses the bandwidth allocation algorithm described in Section 3.

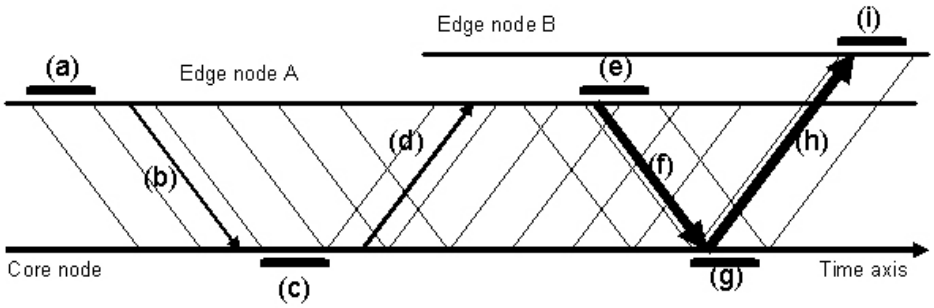


Figure 11: Time sequence diagram of signaling for frame-by-frame bandwidth allocation in a wide-area AAPN. The following sequence is shown: (a) determining traffic demand in edge node, (b) sending a traffic request message to the master edge node at the core, (c) calculating the traffic allocation for a future transmission frame (duration: up to one frame period), (d) transmitting the traffic allocation information to all the edge nodes (node B is located at larger distance than node A), (e) at node A, time has come for transmitting the data slot at the allocated period within the frame in question, (f) the data slot travels to the core node and (h) is switched by the core to the right destination where it is received (i).

Figure 11 shows a time sequence diagram showing the different signaling messages and data slots exchanged between the master and the two other edge nodes. This diagram corresponds to a situation where the propagation delay is approximately 2 frame periods.

6. CONCLUSIONS

The AAPN project has been a very interesting endeavor. By considering a very simple architecture in the form of overlaid stars, we developed several new approaches bandwidth allocation, routing and protection switching which took advantage of this simple architecture. This architecture also simplifies the realization of the synchronization that is required between the different network nodes in order to assure that the transmission slots arrive at the right time at the optical switch. Through the development of a prototype, it was shown, we believe, that such an AAPN is a realistic proposition.

The main feature of an AAPN, as compared with the well-known wavelength-routed optical networks, is the availability of lightpaths of sub-wavelength capacity. This means that an AAPN can extend close to the end-user, either to the building or to the turf. We note that the total capacity required for a residential user would typically be below 100 Mbps, but normally involving several information flows from/to different sources/destinations.

The main competition for an AAPN, on the future networking market would probably be electronic Internet routers or switches. However, we believe that the conceptual simplicity of an AAPN and the power of optical transparent switching without E-O-E conversion will give some advantage to the AAPN approach. We also note that the optical lightpaths provided by an AAPN are protocol and rate-independent at the physical level, which facilitates network evolution.

7. ACKNOWLEDGMENTS

The author would like to thank all members of the AAPN research team, and in particular Cheng Peng, Peng He, Sofia A. Paredes, and Trevor Hall. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and industrial and government partners, through the Agile All Photonic Networks (AAPN) Research Network. The work by Peng He is part of an AAPN Partnered Research Project sponsored by TELUS.

REFERENCES

[1] G. v. Bochmann, et al. “The Agile All-Photonic Network: An architectural outline”, in Proc. 22nd Bien. Symp. on Comm., Kingston, Canada, 2004, pp. 217-218.

[2] The Agile All-Photonic Networks (AAPN) Research Network, http://www.aapn.mcgill.ca/, 2003-2008.

[3] R. Vickers and M. Beshai, “PetaWeb architecture,” 9th Int Telecom. Netw. Planning Symp., Toronto, Canada, 2000.

[4] L.G. Mason, A. Vinokurov, N. Zhao and D. Plant, “Topological design and dimensioning of Agile All- Photonic Networks”, Computer Networks 50 (2006) pp. 268–287.

[5] P. He and G.v. Bochmann, A Novel Internet eXchange (IX) Architecture based on Overlaid-Star All-Optical Networks, submitted for publication.

[6] A. Agustí-Torra, G. v. Bochmann and C. Cervelló-Pastor, “Retransmission schemes for Optical Burst Switching over star networks”, IFIP Int. Conf. Wireless Opt. Commun. Netw., 2004.

[7] C. Peng, S. A. Paredes, T. J. Hall and G. v. Bochmann, Constructing service matrices for agile all-optical cores, 2006 IEEE symposium on Computers and Communications (ISCC 2006).

[8] C. Peng, G. v. Bochmann and T. J. Hall, Quick Birkhoff-von-Neumann decomposition algorithm for agile all-photonic network cores, IEEE International Conference on Communications (ICC 2006).

[9] C. Peng, P. He, G. v. Bochmann and T. J. Hall, Delay performance analysis for an agile all-photonic star network, 2006 IFIP International Conference on Networking (IFIP Networking 2006).

[10] P. He and G. v. Bochmann, Routing of MPLS flows over an agile all-photonic star network, IASTED Intern. Conf. on Communication Systems and Applications (CSA 2006).

[11] P. He and G.v. Bochmann, Inter-Area Shared Segment Protection of MPLS Flows Over Agile All-Photonic Star Networks, IEEE GLOBECOM 2007.

[12] J. Zheng, C. Peng, G. v. Bochmann and T. J. Hall, Load balancing in all-optical overlaid-star TDM networks, 2006 IEEE Sarnoff Symposium.

[13] S. Parveen, R. Radziwilovicz, S.A. Paredes, T.J. Hall. "Evaluation of burst aggregation methods in an optical burst switched agile all-photonic network". SPIE Photonics North 2005, Toronto, 2005.