Seamlessly transformable hybrid packet and circuit switching for efficient optical networks

Weiqiang Sun (孙卫强)*, Pingqing Li (李平青), Chao Li (李 超), and Weisheng Hu (胡卫生)

State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University,
Shanghai 200240, China

**Corresponding author: sunwq@sjtu.edu.cn*

Received October 31, 2012; accepted December 7, 2012; posted online January 6, 2013

Efforts in realizing all-optical packet switching are overwhelming in the past decade. While optical packet switching remains an attractive switching paradigm in the long run, technical challenges prohibit it from becoming a practical solution for the ever-growing bandwidth hunger during the next few years. Finding a technically viable way to meet the increasing capacity requirement with good scalability and flexibility becomes a clear pursue for the community. Hybrid packet and circuit switching is considered to be one promising technique in realizing high performance switching at low cost and less energy consumption, by taking the advantage of both packet switching and circuit switching. In this paper, we review existing work in hybrid optical packet and circuit switching. We discuss the key technical challenges in realizing hybrid optical packet and circuit switching. We further introduce our ongoing efforts in building a seamlessly transformable packet/circuit-switching node with hybrid optical and electronic components. We show that in a hybrid node, the scheduling complexity with typical scheduling algorithms may be reduced to half of a node running in full packet switching mode.

OCIS codes: 060.4256, 060.2330, 060.4250.
doi: 10.3788/COL201311.010601.

1. Introduction

Optical packet switching (OPS) is considered to be the ultimate solution for future networks, since it brings together the high transmission rate of wavelength channels and the versatile per-packet processing intelligence. Over the years, tremendous efforts have been made, significantly extending the frontiers of OPS related technologies[1]. At the mean time, alternative ways of realizing high performance switching with good scalability and low energy consumption are constantly being explored. Such efforts are particularly important when a number of technical challenges are preventing OPS from becoming a practical solution during the next few years.

Circuit/packet hybrid switching is considered to be a good alternative paradigm of OPS[2]. Although ways in realizing hybrid circuit/packet switching vary in detail in different research efforts, the essential idea is straightforward. Circuit switching provides guaranteed end-to-end packet delivery service at constant bitrates, best suited for stable traffic demand. Packet switching provides best effort forwarding by means of statistical multiplexing, and is highly flexible for bursty traffic. Intuitively, it would be attractive to serve the "un-changing" part in the overall traffic with circuit switching, and the remaining fluctuating part with packet switching.

In this paper, we first review existing efforts in hybrid circuit/packet switching design and implementation. We try to identify the challenges in hybrid circuit/packet switching, mostly from a control plane's point of view. We then introduce our ongoing effort in building a hybrid switching testbed with both electrical and optical components. We highlight the seamlessly transformable switching feature in our design and show that the scheduling complexity in the hybrid node can be as low as 50% of full packet switching nodes.

2. Hybrid circuit/packet switching-a review

In a review article by Gauger *et al.*, hybrid optical networks were classified into three categories: client-server, parallel, and integrated[2]. In client-server hybrid switching networks, client layers are usually packet or burst switching networks, connected through the virtual topology provided by the server layer. From the traffic point of view, all packets must enter and depart the network through the client layer switching nodes. It is also possible that the server layer is capable of providing wavelength services as well[3]. Examples of this type of hybrid switching include optical burst switching (OBS) over optical circuit switching (OCS), and electronic packet switching over OCS. Izmailov *et al.*[4] showed that by applying both type of cross-connection, e.g., O-E-O and all-optical, significant capital expenditure reduction might be realized. The energy consumption and footprint of the network devices may also be reduced considerably with hybrid structures. It is also demonstrated that the benefit may be extended by applying hierarchical node architecture with non-uniform waveband size. In the envisioned hybrid nodes the O-E-O part is responsible for relatively expensive grooming and adding/dropping local traffic. And the all-optical part is used for transparent forwarding. A similar hybrid-switching paradigm was introduced in Ref. [5]. An electronic frame switching with buffering and E-O conversion is combined with all-optical slot switching. The authors argue that the proposed scheme offers efficiency, flexibility, and robustness for near term deployment. Chen *et al.*[6] showed that the delay performance of OBS could be improved by adopting lightpath switching for creating the virtual topology. Similar results are also shown in Ref. [7].

In parallel hybrid switching networks, the two types of switching operate in parallel. Special edge nodes are needed to direct traffic either to the packet switching, or to the circuit switching networks. An OBS/OCS hybrid node was proposed[8]. In the hybrid node, OBS and OCS share a common set of resources (e.g., time slots, or wavelengths). Incoming IP flows are classified into short lived flows or long lived flows, and will be served with OBS and OCS respectively. Performance of such a switching paradigm is evaluated and results show that nodal delay increases dramatically at load higher than 70%. Zervas *et al.*[9] proposed yet another hybrid switching architecture called MG-OXC, in which fast switch component (such as SOA-MZI-based switch) and slow switch component (such as MEMS-based switch) were put together in a node, either in parallel or sequential manner. Again, as in Ref. [8], flows are classified into long and short bursts, and will be switched by the slow switch and fast switch respectively. Simulation results show that MG-OXC with a limited number of fast ports has similar performance to a design with only fast ports. Similar results are obtained in the research done by Leenheer *et al.*[10].

In integrated hybrid switching networks, each network element is capable of transport incoming packet streams in either packet switched mode, or circuit switched mode. The switching regime may be selected on a per-packet basis. It is easy to see that this type of hybrid networks is ideal and may achieve optimal resource utilization. However, in practice, it is difficult to make decisions on a per-packet basis, especially with line-rates beyond 40 Gb/s. A multi-wavelength OPS (MW-OPS) and OCS hybrid node was developed and demonstrated[11]. In the design, dynamic resource allocation between OPS and OCS was realized. Blocking performance of different routing and wavelength assignment (RWA) algorithms in hybrid MW-OPS and OCS node is investigated[12,13]. The performance of a hybrid OBS/OCS node, in which OBS and OCS were performed on the same switch matrix (and hence called integrated mode), was evaluated[14]. Blocking performance approximations are derived for burst/circuit traffic with or without priority differentiation. It is also demonstrated that hybrid switching may help realize a significant multiplexing gain.

More recently, designing hybrid packet/circuit switching networks for data centers becomes a topic of common interest[15−17]. Parallel mode of hybrid operation is used in all in reported designs. Researchers show that by incorporating all-optical circuit switching into existing electronic data center networks, one can reduce the network construction cost to half with moderate number of server racks. The power consumption in a hybrid network may be reduced to as low as 1/5 as in traditional networks. We believe hybrid packet/circuit networks will be a promising solution for future large-scale data center networks with high traffic volume.

3. Challenges in resource partitioning in hybrid circuit/packet switching systems

In a hybrid switching system, determining which part of traffic should be switched with which type of component is of primary importance and has crucial influence on system performance such as delay and loss rate. In the existing researches, it is often assumed that the network node has enough intelligence to figure out the characteristic of the incoming traffic, whether it being a large/small flow, or long/short flow. Of equal importance is the resource allocation among the different switching components.

For the client-server mode, difficulty lies in the traditional virtual-topology provisioning problem. In static cases, one must take into account the traffic matrix between all client nodes in planning the virtual topology. It is even more challenging in dynamic cases, when traffic demand between edge nodes varies with time. In such cases, the dynamic circuit provisioning capability of the server layer must be taken into account. For the parallel mode, one has to determine the amount of resource being allocated in each switching plane. Resource partitioning in this case is static at network construction stage and cannot be modified during run-time. Intuitively, allocating more resource for the circuit switching plane results in network best suited for stable traffic patterns, and tends to provide better packet delivery performance in terms of packet loss, delay and jitter. While allocating more resource for the packet switching plane allows for more dynamic traffic patterns and tends to have less attractive statistical packet delivery performance. As traffic pattern between network nodes evolves, the resource partitioning may have to be manually adjusted. Ideally resource partitioning in the integrated mode is adaptive to traffic pattern. The difficulty lies in the characterization/detection of traffic pattern, and dynamic resource partitioning itself. Circuit provisioning delay has important implications on the overall system performance. It is also worth noting that to avoid out-of-order delivery, per-flow decision-making is more desirable than per-packet ones, but may result in sub-optimal resource usage. The actual performance of the integrated mode warrants intensive further study.

It must be pointed out that although in hybrid systems the packet switching regime is only responsible for handling part of the traffic, all challenges in traditional OPS remains relevant[1]. For instance, burst mode transmitter and receiver are crucial components in OPS and so are in hybrid switching systems. Network performance such as throughput and packet loss rate relies heavily on burst capable interfaces. The absence of usable random access memory still remains a major obstacle in realizing high performance hybrid switching nodes. In the short run, the best way to walk around these difficulties is using the relatively mature electronic counterparts, e.g., electronic RAM and logic processing. Orphanoudakis *et al.*[18] showed that introducing electronic RAM into OPS systems could help increase network utilization and have limited impact on average latency. In the remainder of this paper, we will introduce our ongoing effort in building a hybrid switching system with both electronic and

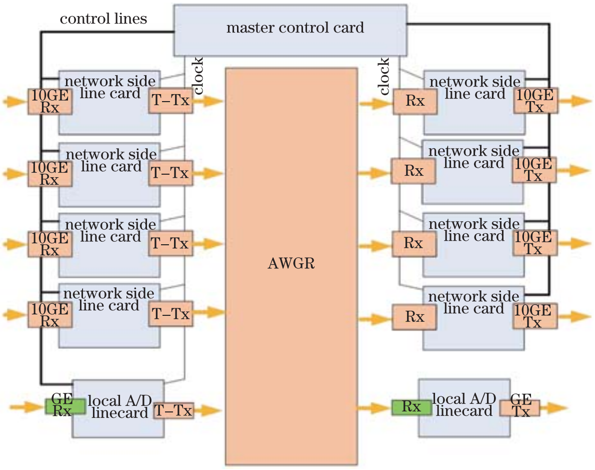

Fig. 1. Hybrid switching node with AWGR switching core and electronic line-cards.

optical components. As will be shown below, our design features an all-optical switching matrix, connecting a number of interface cards with electronic packet processing capability. This combination can on the first hand relax the strict burst mode receiving capability. On the other hand, it also facilitates the design of complex logic processing and hence allows for a seamless integration of circuit and packet switching regimes.

4. Design of a seamlessly transformable hybrid circuit/packet switching node

4.1 Node design

The lack of optical memory makes it difficult to build an all-optical network node with per-packet forwarding capability. Electronic buffer is by far the only and best way to implement packet storage in a packet switching node. Figure 1 shows our node design. The switching fabric is made up of an N×N arrayed waveguide grating router (AWGR) plus N fast tunable transmitters (T-Tx). As reported in Ref. [19], the switching speed of a few tens of nano-seconds can be achieved with sampled grating distributed bragg reflector (SGDBR) laser. By tuning the wavelength of each T-Tx, packets can be routed from incoming linecards to their desired output ports. It is easy to see that this switching fabric is capable of packet switching and circuit switching. As will be shown below, with proper control and management, it is also capable support integrated packet/circuit switching.

Ironically, the really interesting part in the design is in fact on the electronic linecards and master control card. Packets arriving at the network edge (or A/D lineards) will be aggregated into fixed size frames of a few hundreds of microseconds. Before leaving the network from any of the A/D linecards, fixed size frames will be disassembled into their original form (i.e., Ethernet frames).

At the incoming side of each network side linecards, there is a receiver capable of receiving burst signals. We use off-the-shelf optical and electronic components to realize the receiver (BM-Rx). From our preliminary testing, we find that the CDR circuitry in the BM-Rx is able to recover data within 150–200 ns, under the condition that the interval between data bursts are less than a few hundreds of microseconds. In typical packet switching systems, inter-packet delay can be arbitrarily large hence such burst mode receiving performance is not acceptable without modifying the system design. In systems with electronic buffer, one can use “keep-alive” packets to maintain the recovery state of the CDR circuitry. In our design, we monitor the status of each outgoing interface and send out keep-alive packets once it becomes idle.

As in any conventional router, fixed size frames in the linecards are routed by the master control card. A switching operation is performed by tuning the T-Tx on each linecards and sending the serialized frames on them. The master control card is also responsible for monitoring the status of each output port. It instructs linecards with idle T-Tx to generate and send keep-alive packets to idle output ports. Since we have plenty of logic processing capability on the linecards and the master control card, complex control schemes can be implemented and verified.

It is worth noting that the design in Fig. 1 does not allow multicasting/broadcasting. To realize this, one can implement another type of linecard with multiple fixed (or tunable) wavelength transmitters. Upon receiving a multicast/broadcast packet, the master control card first routes it to the special linecard. The special linecard duplicates this packet and transmit copies to the desired output ports. It is easy to see that supporting multicasting/broadcasting in an AWGR based switch matrix is expensive, especially when the number of input/output ports is large. To reduce cost, one can choose another implementation in which duplicated packets are sent one by one, instead of simultaneously. This will eliminate the need for a special linecard, at the cost of non-uniform packet delay among different recipients.

4.2 Integrated network control

The use of electronic linecards and master control card greatly facilitates network control and management. With the frame format in Fig. 2, we realize payload delivery and network control on the same data plane.

Ethernet frames are assembled into fixed size frames on ingress A/D linecards. With the next packet pointer (NPP), we can easily find client frame boundaries and dissemble them on egress A/D linecards. Eight bits are allocated to indicate whether this is a control packet or data packet, and whether it should be duplicated. A flow ID field is defined in the header to implement flow based control schemes.

The switching regime in the system is adaptive to traffic pattern. By adjusting time-slot allocation between all input-output pairs, a node may operate in 100% packet switching mode, or 100% circuit switching

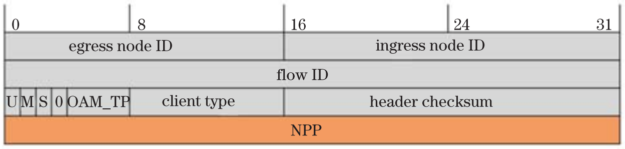

Fig. 2. Frame header definition in the hybrid-switching network.

mode, or anything in between. Traffic measurement and monitoring functions can be easily implemented in the linecards and the control card. In the current design, resource reservation is done independently on each node, hence circuits in the network traverse only one hop.

5. Results on node transformability and scheduling complexity

In the current design, buffer occupancy is used as indications of input traffic load. We monitor the virtual output buffer occupancy and determine how much bandwidth should be allocated to circuit switching. A simulation is performed to verify our design at this stage. Figure 3 shows how the slots are allocated on input port 10 as time progresses. As the buffer occupancy (in bytes) changes, the part of resource that is allocated to circuit switching changes accordingly. It demonstrates that with our resource allocation/transformation heuristics, the system is able to "transform" between the two switching paradigms. Figure 4 shows the scheduling complexity against time when different scheduling algorithms are applied. We compare the scheduling complexity among iterative longest port first (iLPF), iterative round robin matching with SLIP (iSLIP), MUCS, maximum weight matching (MWM), and maximum size matching (MSM), together with time slot assignment (TSA) algorithm. Scheduling complexity is defined as the computation needed to allocate time slot resources. It is closely related with the number of overall time slots in the system. Upon initialization, the system works in full packet switching mode, hence all the time slots are subject to scheduling, resulting in the highest scheduling complexity. As time progresses and some slots are allocated to circuit switching, the scheduling complexity decreases and stabilizes on a certain level. It can be seen that as high as 50% reduction can be achieved in a hybrid node.

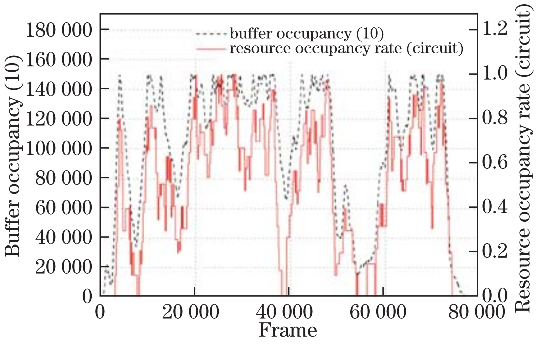

Fig. 3. Resource occupied by circuits and virtual output buffer occupancy (in bytes) versus time.

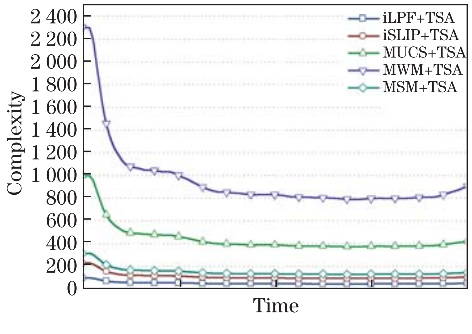

Fig. 4. Scheduling complexity with different scheduling algorithms.

6. Discussion and future work

Our design features a fast reconfigurable switching core that is capable of packet switching and circuit switching. It can be seen that the use of electronic buffer is the key to achieve good system throughput and to implement complex system intelligence. With electronic buffers and control, we can realize hybrid packet/circuit switching and the system is capable of transforming between 100% packet switching and 100% circuit switching. We believe this is a perfect testing ground for resource partitioning algorithms in hybrid switching systems.

With higher speed linecards, the system can be easily upgraded to support 40 Gb/s signals. Distributed control, especially the support of multiple-hop circuits in a hybrid system, is subject of further study.

This work was supported by the National Natural Science Foundation of China (Nos. 60825103 and 61271217), the Fok Ying Tung Education Foundation, the National "863" Program of China (No. 2012AA011301), the National "973" Program of China (No. 2010CB328204-5), and the Foundation of Ministry of Education (No. 20110073130006).

References

1. S. J. B. Yoo, in *Proceedings of Photonics in Switching 2008* 1 (2008).
2. C. M. Gauger, P. J. Kuhn, E. V. Breusegem, M. Pickavet, and P. Demeester, IEEE Commun. Mag. **44,** 36 (2006).
3. G. J. Eilenberger, in *Proceedings of OSA/OFC/NFOEC 2011* OTuP1 (2011).
4. R. Izmailov, S. Ganguly, T. Wang, Y. Suemura, Y. Maeno, and S. Araki, IEEE Commun. Mag. **40,** 88 (2002).
5. H. Leligou, A. Stavdas, and J. Angelopoulos, Proc. SPIE **6388,** 63880C (2006).
6. B. Chen and J. Wang, IEEE J. Sel. Areas Commun. **21,** 1071 (2003).
7. Y. Wang, S. Wang, S. Xu, and X. Wu, in *Proceedings of ICACT 2009* **3,** 1873 (2009).
8. G. M. Lee, B. Wydrowski, M. Zukerman, J. K. Choi, and C. H. Foh, IEEE Globecom **5,** 2508 (2005).
9. G. S. Zervas, M. de Leenheer, L. Sadeghioon, D. Klonidis, Y. Qin, R. Nejabati, D. Simeonidou, C. Develder, B. Dhoedt, and P. Demeester, J. Opt. Commun. Netw. **1,** 69 (2009).
10. M. De Leenheer, C. Develder, J. Vermeir, J. Buysse, F. De Turck, B. Dhoedt, and P. Demeester, in *Proceedings of ONDM 2008* 1 (2008).
11. T. Miyazawa, H. Furukawa, and K. Fujikawa, J. Opt. Commun. Netw. **4,** 25 (2012).
12. K. Machida, H. Imaizumi, H. Morikawa, and J. Murai, in *Proceedings of ONDM 2009* 1 (2009).
13. M. Takagi, H. Li, K. Watabe, H. Imaizumi, T. Tanemura, Y. Nakano, and H. Morikawa, in P*roceedings of OFC 2009* OTuA6 (2009).
14. H. Le Vu, A. Zalesky, E. W. M. Wong, Z. Rosberg, S. M. H. Bilgrami, M. Zukerman, and R. S. Tucker, J. Lightwave Technol. **23,** 2961 (2005).

15. G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. S. E. Ng, M. Kozuch, and M. Ryan, in *Proceedings of the ACM SIGCOMM 2010* (2010).
16. N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, A. Vahdat, N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, in *Proceedings of the ACM SIGCOMM 2010* 339 (2010).
17. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. Eugene Ng, D. G. Andersen, M. Kaminsky, M. Kozuch, and A. Vahdat, in *Proceedings of the ACM SIGCOMM 2011* 30 (2011)
18. T. G. Orphanoudakis, A. Drakos, C. Matrakidis, C. Politi, and A. Stavdas, in *Proceedings of 9th International Conference on Transparent Optical Networks* 222 (2007).
19. J. E. Simsarian, A. Bhardwaj, J. Gripp, K. Sherman, Yikai Su, C. Webb, L. Zhang, and M. Zirngibl, IEEE Photon. Technol. Lett. **15,** 1038 (2003).