Designing Silicon Photonic Interconnection Networks for Deadline-Driven Applications

Ke Wen, Sébastien Runley and Keren Bergman
Columbia University, New York, United States

Paper Summary

Meeting service deadlines is critical to datacenter applications. This paper proposes a deadline-aware burst switching architecture and an expected reward-based scheduling policy for the implementation of deadline-driven delivery within silicon photonic interconnection networks.

Introduction

Today's datacenters are widely used for Online Data Intensive (OLDI) applications. These applications usually provide real-time services and require the delivery of responses within a specified service deadline in order to provide good user experiences. The capability to meet such deadlines is critical as it directly impacts the operator revenue; for example, an additional 100ms latency cost *Amazon* 1% in sales [1]. Typical implementations of OLDI applications serve a single request by multiple levels of processing nodes in a *partition-aggregate* fashion [2]. Therefore, the service deadline inevitably translates to deadline targets for inter-node communications (Fig. 1) [3].

Previously proposed deadline-driven delivery (D^3) solutions [3, 4], which target at electronic network infrastructures, may struggle to support the ever-growing traffic demands of future datacenters due to the limited bandwidth density and significant power consumption of high-data rate electronic switching. In comparison, silicon photonic (SiP) interconnection networks present a promising solution to the growing need for data switching in datacenters. However, despite the fact that current optical networks are being designed to undertake an increasing amount of functionality from upper network layers, they are generally agnostic of the deadline requirements. As an example, Fig. 2 depicts the arrival patterns of two reply-carrying flows (A and B) from the same origin node but with different deadlines. Also presented are three delivery schemes for these flows: packet-based, time-division-multiplexing (TDM)-based and deadline-aware burst based, respectively. The packet-based scheme transmits each packet according to their arrival order. Due to the lack of deadline awareness and the per-packet reconfiguration overhead, this scheme misses the deadline of Flow B. The TDM scheme avoids frequent reconfiguration overheads (as well as contentions) by assigning fixed slots for every pair of nodes. However, since the deadline requirements are largely unpredictable and can often result in a mismatch with the fixed slot assignments, the TDM scheme is not suitable for D^3 either (Fig. 2c shows missed deadlines for both Flow A and B). The third scheme takes advantage of the high optical bandwidth and assembles packets of the same flow (with typical flow sizes < 50KB [3]) into an optical burst. The burst is stamped with the deadline of the corresponding flow. This scheme avoids the per-packet overhead and the rigidity of TDM switching; more importantly, it leverages the inherent correlation of the packets: packets of the same flow share the same deadline. This knowledge allows *an additional dimension of flexibility* to the traffic scheduling, especially for photonic networks, which typically benefits from some level of traffic aggregation [5].

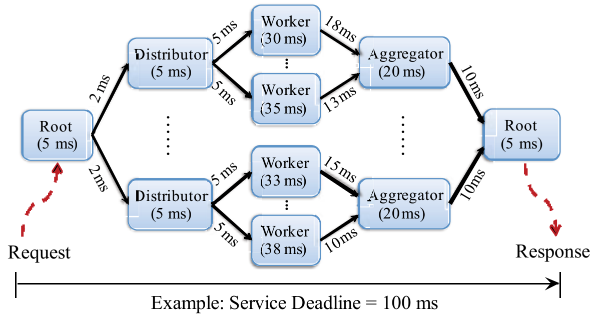

Fig. 1. Internode flow deadlines caused by service deadline.

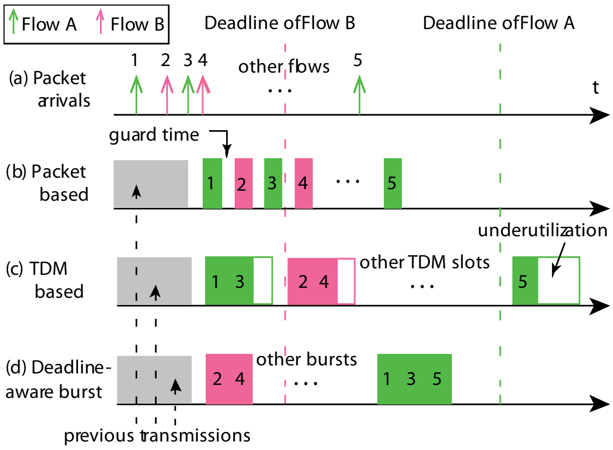

Fig. 2. Comparison of (b) packet-based, (c) TDM-based and (d) deadline-aware burst-based schemes.

Though promising, delivering bursts with a) *variable lengths* and b) *differentiated deadline requirements* is non-trivial for SiP *bufferless* interconnection networks in both hardware and scheduling aspects. This paper addresses the above issues with an architecture/policy co-design approach.

978-1-922107-21-3 © 2014 Engineers Australia

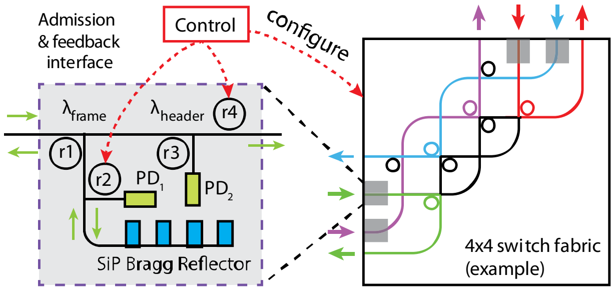

Fig. 3. Example 4×4 SiP switch element incorporated with the admission and feedback interface.

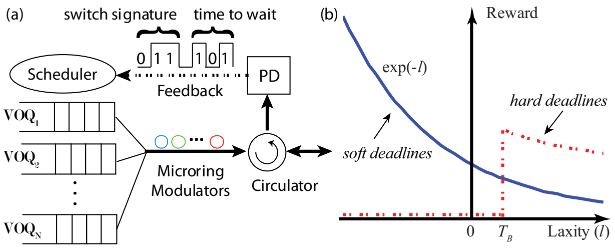

Fig. 4. (a) Client interface and (b) reward functions.

Deadline-Aware Burst Delivery

Deadline-based switching systems previously rely on a centralized arbitrator that typically finds the maximal weight matching based on the imminence of deadlines [6]. However, due to the time complexity of maximal matching algorithms, such centralized arbitrators can hardly scale with the number of nodes, especially for photonic networks, which feature fast switching speeds and ultra-high data rates. Moreover, the length variability of optical bursts adds another level of difficulty to the centralized (and hence often synchronized) arbitration. Instead, optical bursts typically access the network via asynchronous opportunistic injections and contentions can be resolved by retransmission after reception of a NACK notification [7]. However, without knowledge of network states, such retransmission can create head-of-line (HOL) blockings. While frequent retrials (e.g. FastNACK [8]) waste energy and could escalate network congestions, backoff methods might lead to missed deadlines and lower the network utilization. Therefore, to obtain deadline-aware SiP burst-switched networks, several key problems need to be solved: (1) how to implement distributed controls for SiP networks, (2) how to acquire network availability information and (3) how to incorporate such information into the deadline-based scheduling policy design. This work proposes an architecture/policy co-design approach, which includes a novel SiP switch interface design that is capable of *optical availability feedback* and distributed flow control as well as an *expected reward* based scheduling policy using both the availability and deadline information.

A. Switch Interface Design

Depending on the physical limitation, a SiP interconnect can comprise one or multiple SiP chips, with each chip being a switching element (SE) in a multistage topology (e.g. omega, clos, flatterned butterfly or etc). To support variable burst lengths, the proposed architecture uses a frame wavelength to indicate the start and the end of a burst. The frame wavelength (λ_F) sends '1' during the existence of a burst and is used with header wavelengths (λ_H, which precedes the data wavelengths) for fast access grant and switch reconfiguration [9].

Optical availability feedback (OAF) makes use of the frame wavelength and transmit channel availability information from the switch interface back to the client. As Fig. 3 shows, an *admission and feedback* module is added to each input of the switch. Microring **r1** and **r3** filter out λ_F and λ_H, respectively. Microring **r2** is initially tuned to on-resonance state for the detection of λ_F. When there is no contention, the control logic configures the switch fabric and tunes **r4** to pass the input signal to the switch fabric unperturbed. Otherwise, if the packet is not accepted, **r2** will work as an active modulator driven by the control logic for OAF. When **r2** is tuned to the off-resonance state, it passes λ_F to a SiP Bragg reflector [10], which reflects the optical power back to the client as a '1'. Otherwise, **r2** drops the optical power of λ_F to detector $\mathbf{PD_1}$ and the client will see a '0'.

The feedback from a blocking SE to a client includes two parts of information: the signature of the SE and the time to wait (TTW) before the blocking port becomes available (Fig. 4a). The client, with a virtual output queue (VOQ) structure, is then informed of the blocking position. It disables not only the rejected VOQ but also those *whose packets would pass through the same position*. These VOQ are disabled for a time equal to the received TTW value. No disabled VOQ will be scheduled until its TTW timer expires, which avoids unnecessary blockings.

B. Expected Reward Based Scheduling

To prioritize deadline-imminent bursts, we propose an *"Expected Reward"* based algorithm considering both the laxity of the burst as well as the availability feedback. The laxity (*l*) of a burst is defined as the difference between its deadline (*d*) and the current time (*t*), i.e. *l*=*d*-*t*. To provide incentive for delivering small-laxity bursts, the reward for successfully scheduling a burst with laxity *l* is defined by a convex, non-negative and non-increasing function $\Phi(l)$, that is, the smaller the laxity, the higher the reward. A typical example of such reward functions is depicted in Fig. 4b. For bursts with hard deadlines, the reward function $\Phi(l)$ is simply set to 0 when its laxity is smaller than the required transmission time (denoted by T_B in the figure), which reflects the fact there is no use for this burst to arrive at its destination after the hard deadline.

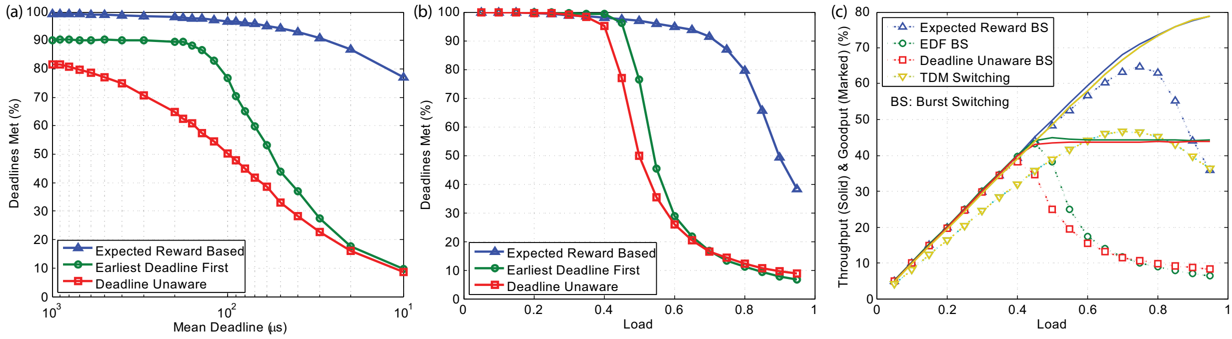

Fig. 5. Simulation results: deadlines met (%) versus (a) deadline mean and (b) traffic load, respectively; (c) throughput and goodput.

If a VOQ with TTW value τ is scheduled at time t ($\tau \leq 0$ as required for scheduling consideration), its chance of successful delivery can be approximated by the probability that no other burst has intervened the channel in the time period $[t-\tau', t]$, where $\tau' = min\{|\tau|, T_{max}\}$ and T_{max} is the maximum burst duration. By Poisson distribution, this probability is

$$P_s(\tau') = e^{-\lambda\tau'}$$

The expected reward of scheduling this burst is then given by

$$E[R] = \Phi(l) \cdot P_s(\tau')$$

Whenever a transmitter is free, it schedules the VOQ whose burst has the highest expected reward.

Simulation Results

The performance of the proposed D^3 co-design is evaluated via simulation against a) the *deadline-unaware* policy, which transmits according to the burst arrival order and b) the *"Earliest Deadline First"* (EDF) policy, which schedules the burst with the earliest deadline. The simulation is based on a 32-node omega topology and assumes an aggregate link bandwidth of 400 Gb/s. The traffic microbenchmark represents typical datacenter flow patterns (e.g. flow sizes, deadline requirements). The flow sizes are uniformly distributed across [2KB, 50KB] and the flow deadlines are exponentially distributed with a given mean value.

Fig. 5a presents the percentage of deadlines met versus the deadline mean under a load of 0.5. The result shows that the expected reward (ER)-based algorithm outperforms the other two policies across the entire simulated range. The advantage becomes more significant as the deadline becomes tighter. For a deadline mean of 60 µs, the ER-based algorithm meets 95% of the deadlines. This shows the capability of SiP interconnects in delivering worker replies within sub-ms requirements. Compared to deadlines on the level of tens of ms in electronic datacenter networks [3, 4], such a much tighter deadline target is invaluable as it allows more time for computation and hence better response quality.

Fig. 5b presents the percentage of deadlines met versus the traffic load. Due to the lack of availability information and congestion collapse, the performance of previous schemes suffer from a cutoff effect: the in-time rate drops drastically as the traffic load grows greater than 0.4. In comparison, the ER-based algorithm avoids early performance cutoff and maintains an in-time rate higher than 90% until a load of 0.7.

Fig. 5c compares the throughput of the burst switching schemes with the TDM scheme. Due to the effectiveness of distributed congestion avoidance, the ER co-design achieves as high throughput as the TDM scheme, while other burst scheduling policies saturate at a level about 35% lower. Also shown is the metric of *"goodput"*, which is the throughput delivered in time. With the deadline awareness, the ER algorithm further delivers as much as 18% more goodput than the TDM scheme, which contributes to more timely services.

Conclusions

This paper addresses the implementation of deadline-driven delivery in silicon photonic interconnection networks with a co-design approach. The proposed deadline-aware burst switching architecture together with the expected reward-based scheduling policy not only shows the potential to support tighter deadlines than previous electronic D^3 solutions, but also effectively improves the throughput and goodput of the network by taking advantage of the additional flexibility enabled by the knowledge of flow deadlines.

Acknowledgment

The authors acknowledge support for this research in part by Sandia National Laboratories under PO 1319001.

References

1. T. Hoff, “Latency is everywhere and it costs you sales - how to crush it,” Jul. 2009.
2. D. Meisner, *SIGARCH Comput. Archit. News*, Jun. 2011.
3. C. Wilson, *ACM SIGCOMM*, 2011.
4. B. Vamanan, *ACM SIGCOMM*, 2012..
5. S. Malik, *Opt. Switching and Netw.*, vol. 2, no. 4, 2005.
6. A. Dua, *GLOBECOM*, 2006.
7. I. Widjaja, *IEE Proc.- Commun*, vol. 142, no. 1, 1995.
8. G. Dongaonkar, *Opt. Interconnects Conference*, 2013.
9. A. Shacham, *IEEE Micro*, vol. 27, no. 4, 2007.
10. I. Giuntoni, *Opt. Express*, vol. 17, no. 21, 2009.