An optical interconnection network with wavelength time slot routing

Ireneusz Szcześniak and Roman Wyrzykowski
Institute of Computer and Information Sciences
Częstochowa University of Technology
ul. Dąbrowskiego 73
42-200 Częstochowa
Poland

***Abstract*—We propose an optical interconnection network with the *wavelength time slot routing*, a novel routing scheme that allows for a simple design of optical interconnection networks. The simulative performance evaluation shows that the proposed scheme demonstrates optimum performance at the maximum uniform exponential network load, and performs well in comparison to store-and-forward routing.**

***Index Terms*—Interconnection network, network on chip, space switch, arrayed waveguide grating, Beneš network, synchronous, simulation.**

I. Introduction

Interconnection networks connect computing and storage nodes, chips on motherboards, and components on a chip. These networks are complicated, require large buffers, cause large latencies, drop packets and perform poorly under heavy loads – exhibiting the opposite of their desired properties.

While interconnection networks are currently electronic and can use optics only for transmission, it is generally agreed that interconnection networks in the long term should become more optical and less electronic to increase performance and reliability, and to decrease power consumption [1]. The following are the key advantages of optics over electronics: optical signal travels faster than electric signal; high data rate transmission is easier to achieve optically than electronically, especially as distance grows; and for data rates above 100 Gb/s the switching energy per bit is lower for optics than for electronics [2], [3].

Optical components continue to replace their electronic counterparts. The production of optical components has become cheap and their integration has made impressive progress [4]. Intensive research on optical networks, both on long-haul single-mode networks and on short-distance multi-mode networks, have resulted in optical components of excellent properties, constantly increasing data rates, and the deployment of wavelength-division multiplexed networks. Intel and IBM already demonstrated working prototypes of optical interconnects [5], [6].

We propose a novel idea of *wavelength time slot routing* (WTSR) for interconnection networks. WTSR is time slot routing augmented with wavelength-division multiplexing (WDM). Time slot routing (TSR), in turn, is a form of time-division multiplexing (TDM) that is used in communication networks to share a transmission link, but we devised TSR as a simple and efficient means of routing packets in an interconnection network.

We propose to switch packets optically. Since optical packet switching (OPS) for long-haul networks has failed [7], [8], [9], it can be argued that interconnection networks with OPS could fail too. However, there are two fundamental differences between these two applications that work in favor of OPS for interconnects. First, optical interconnects, unlike long-haul networks, require a smaller-scale deployment, which make it less costly and less risky. Second, optical impairments in long-haul, multi-hop communication networks are currently hindering the deployment of OPS, while for short-distance interconnect communication with a small number of hops, the optical impairments are much less of a problem.

The article is organized as follows. In the next section, we discuss related work. Then, wavelength time slot routing is introduced, which is followed by its performance evaluation and comparison to store-and-forward routing. The article ends with conclusions.

II. Related work

An interconnection network with prescheduled access, speculative transmissions, and acknowledgments is detailed in [10]. There the authors propose to preschedule access to destination nodes in a similar way as we propose in the paper. They, however, do not employ wavelength-division multiplexing to minimize admission delay and increase throughput, but propose to eagerly (i.e. without waiting for a due time slot) inject packets and wait for their acknowledgments.

In [11] authors demonstrated the experimental validation of the SPINet architecture. In SPINet, nodes can send packets at any time slot. As a packet travels through the network, a path is established for it by configuring switching elements. Once the path is established and the packet reaches its destination, the destination node sends back an acknowledgment (ACK) pulse through the already-established path. If the network blocks the packet, no ACK pulse arrives, the packet has to be retransmitted.

The overview of the optical interconnect architectures based on the arrayed waveguide grating (AWG) can be found in [12], which is of special interest since in our work we also use the AWG.

TABLE I: Sample required permutations for a 4×4 network.

P1	P2	P3
$1 \to 2$	$1 \to 3$	$1 \to 4$
$2 \to 3$	$2 \to 4$	$2 \to 1$
$3 \to 4$	$3 \to 1$	$3 \to 2$
$4 \to 1$	$4 \to 2$	$4 \to 3$

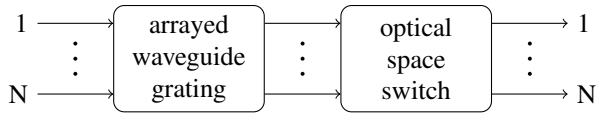


Fig. 2: Interconnection network architecture.

III. Wavelength time slot routing

We first describe TSR and then introduce WDM to produce WTSR. WDM allows for the reduction of the admission time, and for the increase of the network throughput.

To demonstrate WTSR scheme and to evaluate its performance, we chose the Beneš network. Any other network type can be used with WTSR provided it is rearrangeable; the nonblocking property is not required as the network can be rearranged every time slot.

A. Time slot routing

There are N nodes connected by a space switch that exchange data packets synchronously, i.e. according to a single clock shared by the interconnection network with all nodes. The interconnection network does not have buffers, and its configuration, i.e. the configuration of its switching elements, determines the connections between inputs and outputs of the network. A delay suffered by a packet equals only to the delay caused by the light to traverse the switching fabric, which is below a nanosecond.

For each time slot the space switch is configured for a given *permutation*, which is a specific way of connecting inputs to outputs. In a single permutation every source node is connected to a destination node different from the source node. There are $(N-1)$ permutations required, which are repeated periodically, so that each node can send packets to each of the other $(N-1)$ nodes. For a packet sent from source node $n = 0, \ldots, N-1$, time slot number t determines destination node n_{tsr} as given by (1).

$$n_{tsr} = (n + 1 + t \bmod (N-1)) \bmod N \tag{1}$$

Table I reports three sample permutations P1, P2, and P3 required for a 4×4 network. The notation $1 \to 2$ used in the table denotes a connection from node 1 to node 2. Figures 1a, 1b, and 1c show the configurations of the 4×4 Beneš network that implement permutations P1, P2, and P3, respectively.

The link between a node and the interconnection network can be viewed as a link with time-division multiplexing with $(N-1)$ channels, each channel connected to one of the other $(N-1)$ nodes. The throughput of each channel is one packet per $(N-1)$ time slots.

B. Introduction of WDM

In WTSR we extend the TSR by introducing wavelength division-multiplexing (WDM) with W wavelengths. Having the $N \times N$ space switch for TSR, WTSR can be implemented by preceding the space switch with the $N \times N$ arrayed waveguide grating, as shown in Fig. 2, where each waveguide carries W wavelengths. For the special case of $W = 1$, WTSR becomes TSR.

The $N \times N$ AWG is usually designed to route N wavelengths, but here we make use of W wavelengths, where $W < N$. In the $N \times N$ AWG with W wavelengths, wavelength w from input n is routed to output $(n + sw) \bmod N$, where $w = 0, \ldots, W-1$, and step parameter s is an integer constant [13].

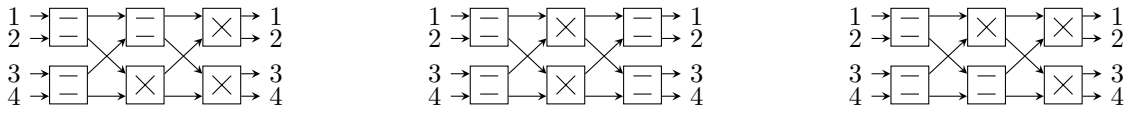
During a time slot, each node can send at most W packets, one packet per wavelength. Destination node n_{wtsr} of a packet sent from node n is determined by wavelength number w and time slot number t as given by (2).

$$n_{wtsr} = (n_{tsr} + sw) \bmod N \tag{2}$$

The optimum value for step parameter s is $s = N/W$, because a source node can send a packet to a given destination node with a minimum average number of time slots waiting. Other values of s cause larger average waiting times and cause unevenly spaced access to destination nodes. For instance, if $s = 1$, a source node can send a packet to a given destination node in W consecutive time slots using consecutive wavelengths, and then has to wait $N - 1 - W$ time slots to be able to send a packet to that destination node again.

The advantages of WTSR are as follows.

- *No packet loss*. Since no packets are lost in the network, there is no need for acknowledgments and retransmissions.
- *No overhead*, since no packet header is required. The network routes packets according to the source node number, the time slot number and the wavelength number.
- *Fit for optics*. The WDM transmission takes advantage of the transparency of the space switch. Furthermore, the cheap and passive AWG implements the distribution stage.
- *Simple design*. WTSR does not require buffers or packet header processing.
- *Simple optical implementation*. The space switch can be implemented with microring resonators or directional couplers as an integrated optical component.
- *Simple electronic implementation*. Since the routing algorithm is simple, it can be implemented with a simple electronic circuit, making it fast and cheap.
- *Optimum throughput* under the maximum uniform load.
- *Constant network latency*. Every packet travels in the network through the same number of nodes and links of equal length.
- *Fairness*. Every source node has a chance to send the same number of packets to the destination nodes as any other node.



(a) (b) (c)

Fig. 1: Sample configurations of the 4×4 Beneš network.

IV. Performance evaluation

The performance of WTSR is compared to the performance of store-and-forward routing (SAFR) for the 64×64 Beneš network ($N = 64$) with 352 switching elements. We assumed the time slot length of 10 ns, since the space switch can be reconfigured in about a nanosecond. There are N nodes connected to the interconnection network with a patch cord of 1 m length that introduces 5 ns delay. Each node has $N - 1$ admission queues, one queue per a destination node, which can grow without a limit. The traffic is uniform and exponential, without and with acknowledgments.

A. Wavelength time slot routing

For WTSR, the analysis of a single admission queue for the exponential traffic with rate λ can be carried out analytically as for a queue of a TDM link with the deterministic service rate of $\mu = W/(N-1)$ packets per time slot. We note that for the exponential input traffic the TDM queue is not the M/D/1 queue, because for an empty TDM queue the service begins when a time slot begins, while for an empty M/D/1 queue the service begins when a new message arrives. The analysis of a TDM queue is complex and uses the Z-transform, since both the continuous and discrete time domains are involved. The maximum throughput of the TDM queue is μ, and the minimum average admission delay is $s/2$ when the queue is empty [14].

Since there are $N(N-1)$ flows with no packet losses, and since the maximum throughput of a flow is μ, then the maximum network throughput T_{max} is given by (3).

$$T_{max} = NW \tag{3}$$

B. Store-and-forward routing

SAFR is used in traditional interconnection networks, and requires buffers at every switching element. A switching element processes a packet at least in a time slot, since it has to buffer the packet and interpret its header. Each output of a switching element has a buffer. When a packet arrives, it is enqueued in the buffer of its preferred output, i.e. an output that yields the shortest path to the packet's destination. If there is no space left in the buffer, the packet is dropped. If the packet prefers either output, the output is chosen at random, which helps to improve the network performance and fairness. If there is no space left in the randomly chosen buffer, the packet is enqueued in the buffer of the other output. If there is no space left in the other buffer, the packet is dropped.

The number g of switching elements a packet has to visit from any source node to any destination node in the $N \times N$ Beneš network is given by (4). Therefore the minimum delay caused by SAFR is g time slots, and the maximum is g times the buffer size.

$$g = 2\log_2 N - 1 \tag{4}$$

C. Simulations

The performance evaluation of WTSR and SAFR was carried out with 4800 simulation runs. Half of the runs was for the traffic without acknowledgments and the other half for the traffic with acknowledgments. The performance of both WTSR and SAFR was evaluated for three numbers of wavelengths: 1, 4 and 16. For SAFR, the performance was evaluated for three sizes of buffers at switching elements: 1, 2 and 3.

For both routing algorithms we used the Beneš network of size $N = 64$ and ten different loads ($l = 0.05, 0.1, \ldots, 1.0$). WTSR had 120 test cases: three numbers of wavelengths, twenty different loads, and the traffic with and without acknowledgments, where each test case was simulated ten times with different pseudo-random values, totaling 1200 simulation runs. SAFR had 360 test cases: three numbers of wavelengths, twenty different loads, three buffer sizes, and the traffic with and without acknowledgments, where each test case was simulated ten times with different pseudo-random values, totaling 3600 simulation runs.

We assume that the network is uniformly loaded: each node sends packets to each of the other $(N-1)$ nodes, and so there are $N(N-1)$ flows in the network. Each flow is of the exponential distribution. For the traffic without acknowledgments, the mean rate $\lambda = l\mu$, where $l \in (0, 1)$ is the network load. The network is loaded to the maximum when $l = 1$. For the traffic with acknowledgments, the mean rate $\lambda' = \lambda/2$, since the returning ACK packets constitute about a half of the traffic.

We evaluated the mean network throughput, the mean number of dropped packets, the mean total delay, the mean admission delay, and the mean admission queue size. We calculated the mean values for a test case based on the results of ten simulation runs for that test case. These simulation results are credible: we calculated the standard errors for the mean values, and found that they were below 1% of the mean values.

D. Comparison

The comparison of WTSR and SAFR is presented in Figures 3, 4, 5, 6, and 7 for the cases without and with acknowledgments. In each of the figures, the subfigures (a) are for $W = 1$, (b) for $W = 4$, and (c) for $W = 16$. The results for the

traffic without acknowledgments are shown with points, and the results for the traffic with acknowledgments are shown with lines. There are no error bars shown for the data points that would express the standard errors, because the standard errors were too small to be drawn.

1) Without ACKs: The results for WTSR are shown with the bullets (•), and the results for SAFR are shown with the pluses (+) for the buffers of size 1, with the crosses (×) for the buffers of size 2, and with the circles (○) for the buffers of size 3.

The mean network throughput is shown in Fig. 3. WTSR performs optimally in that it delivers all the offered load, reaching the maximum throughput at heavy loads as given by (3): at most WN packets per time slot for the $N \times N$ network with W wavelengths. SAFR with buffer sizes of 1, 2, and 3 performs worse because of packet losses.

The mean number of packets dropped per time slot is shown in Fig. 4. Since WTSR does not lose packets, its results are not shown in the figure. SAFR, however, suffers packet loss, even for buffers of size 3. Packet losses are caused by packet contention and shortage of buffer space.

The mean total delay measured in time slots is shown in Fig. 5, which is the sum of the number of time slots spent waiting in the admission queue, and the number of time slots spent in the network. For WTSR, the total delay depends only on the admission delay, since packets stay in the network only a single time slot. Inversely, for SAFR the total delay depends little on the small admission delay, and is dominated by the network delay, which is a function of the number of nodes packets visit, as expressed by (4), and the network load. The network delay grows as the buffer size and the load increase.

The mean admission delay is shown in Fig. 6. WTSR suffers a minimum average admission delay of $s/2 = N/2W$ time slots, and so larger numbers W of wavelengths result in a smaller admission delay. For SAFR the admission delay is smaller than for WTSR, but this advantage can be lost by large network delays.

The mean admission queue size is shown in Fig. 7. For heavy loads an admission queue is occupied on average by a few packets for WTSR, while for SAFR there are almost no packets in the queue.

2) With ACKs: The results for WTSR are shown with the solid lines, and the results for SAFR are shown with the dashed lines for buffers of size 1, with the dash-dot lines for the buffers of size 2, and with the dotted lines for buffers of size 3.

In the tests without acknowledgments, the lost packets were unaccounted for, which cannot be accepted since lost packets must be retransmitted. To evaluate the impact of lost packets on the network performance, we require destination nodes send back ACK packages when they receive packages proper. The source node waits for an ACK package at most $10(g + 1)$ time slots, and retransmits the package proper if that timeout expires. The maximum number of time slots for a packet proper and an ACK packet to traverse a network is $3g + 3g$ for the buffers of size 3, and we allowed some additional time slots for packets to wait in admission queues.

Furthermore, each flow has a window of $10g$ packets that can be "in-flight", i.e. without being acknowledged. A sender stops generating new packets when that window has been exhausted, and waits for ACKs of packets sent before. This method of operation can be referred to as self-throttle or TCP-like (Transmission Control Protocol).

As expected, the performance of WTSR does not deteriorate, while the performance of SAFR does. For SAFR, the network throughput drops for heavy loads especially for a small number of wavelengths. For SAFR the total delay increases drastically for heavy loads, because lost packets are retransmitted. The admission queue size is larger for SAFR in comparison with WTSR for a larger number of wavelengths.

V. Conclusion

We presented wavelength time slot routing, and evaluated its performance for the Beneš network, but other network types can be used too. We report that wavelength time slot routing outperforms the store-and-forward routing. Wavelength time slot routing can be used for optical interconnects in computing and storage centers, but also in networks on chip (NoC). Wavelength time slot routing allows for a simple network design without buffers and header processing.

References

[1] "International technology roadmap for semiconductors: interconnect chapter," 2011. [Online]. Available: www.itrs.net

[2] D. Blumenthal *et al.*, "Integrated photonics for low-power packet networking," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 458–471, March–April 2011.

[3] G. D. Micheli and L. Benini, *On-chip communication architectures: system on chip interconnect*. Morgan Kaufmann Publishers Inc., 2008.

[4] L. Coldren, S. Nicholes, L. Johansson, S. Ristic, R. Guzzon, E. Norberg, and U. Krishnamachari, "High performance InP-based photonic ICs - a tutorial," *Journal of Lightwave Technology*, vol. 29, no. 4, pp. 554–570, February 2011.

[5] M. Taubenblatt, "Optical interconnects for high-performance computing," *Journal of Lightwave Technology*, vol. 30, no. 4, pp. 448–457, February 2012.

[6] R. Leheny, "Molecular engineering to computer science: the role of photonics in the convergence of communications and computing," *Proceedings of the IEEE*, vol. 100, pp. 1475–1485, 2012.

[7] R. Tucker, "The role of optics and electronics in high-capacity routers," *Journal of Lightwave Technology*, vol. 24, no. 12, pp. 4655–4673, December 2006.

[8] ——, "Scalability and energy consumption of optical and electronic packet switching," *IEEE/OSA Journal of Lightwave Technology*, vol. 29, no. 16, pp. 2410–2421, August 2011.

[9] R. Ramaswami, "Optical networking technologies: what worked and what didn't," *IEEE Communications Magazine*, vol. 44, no. 9, pp. 132–139, September 2006.

[10] N. Chrysos, C. Minkenberg, J. Hofrichter, F. Horst, and B. Offrein, "Towards low-cost high-performance all-optical interconnection networks," in *High Performance Switching and Routing (HPSR), 2010 International Conference on*, June 2010, pp. 139–146.

[11] A. Shacham and K. Bergman, "An experimental validation of a wavelength-striped, packet switched, optical interconnection network," *Journal of Lightwave Technology*, vol. 27, no. 7, pp. 841–850, April 2009.

[12] X. Ye, S. Yoo, and V. Akella, "AWGR-based optical topologies for scalable and efficient global communications in large-scale multiprocessor systems," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, no. 9, pp. 651–662, September 2012.

[13] X. J. Leijtens, B. Kuhlow, and M. K. Smit, "Arrayed waveguide gratings," in *Wavelength Filters in Fibre Optics*, ser. Springer Series in Optical Sciences, H. Venghaus, Ed. Springer, 2006, vol. 123, pp. 125–187.

[14] K.-T. Ko and B. Davis, "Delay analysis for a TDMA channel with contiguous output and poisson message arrival," *IEEE Transactions on Communications*, vol. 32, no. 6, pp. 707–709, June 1984.

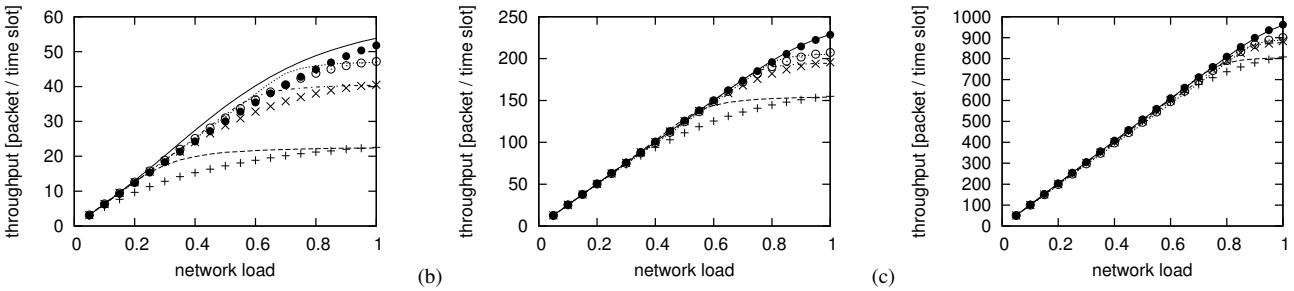


Fig. 3: Comparison of the mean network throughput.

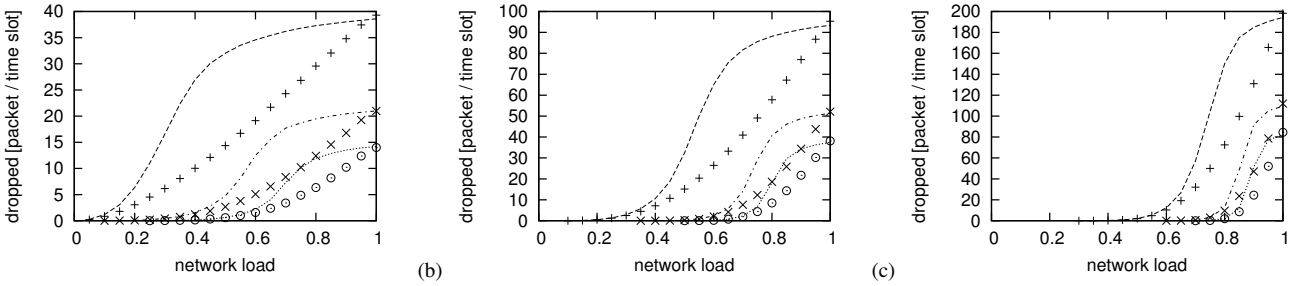


Fig. 4: Comparison of the mean number of dropped packets.

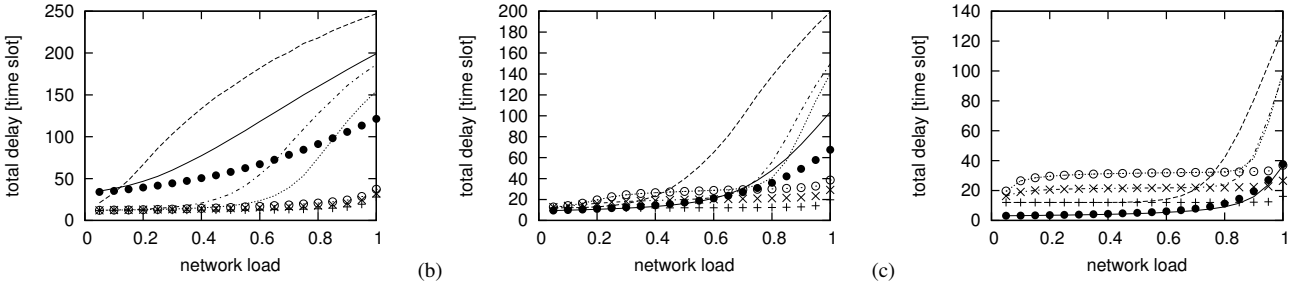


Fig. 5: Comparison of the mean total packet delay.

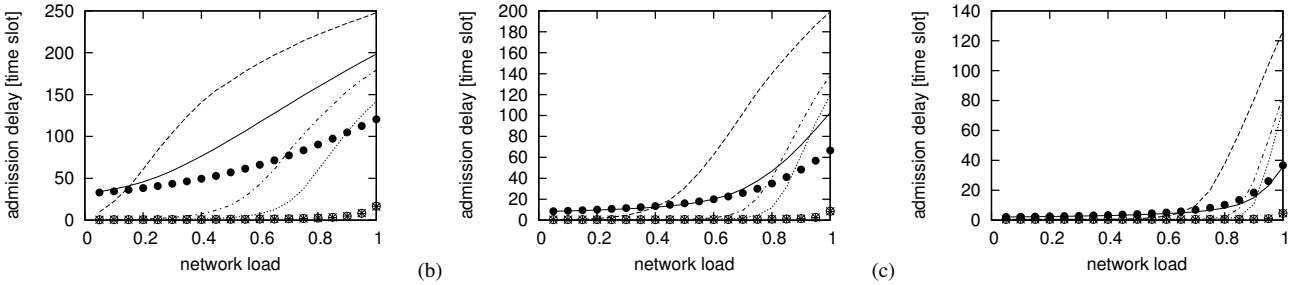


Fig. 6: Comparison of the mean admission delay.

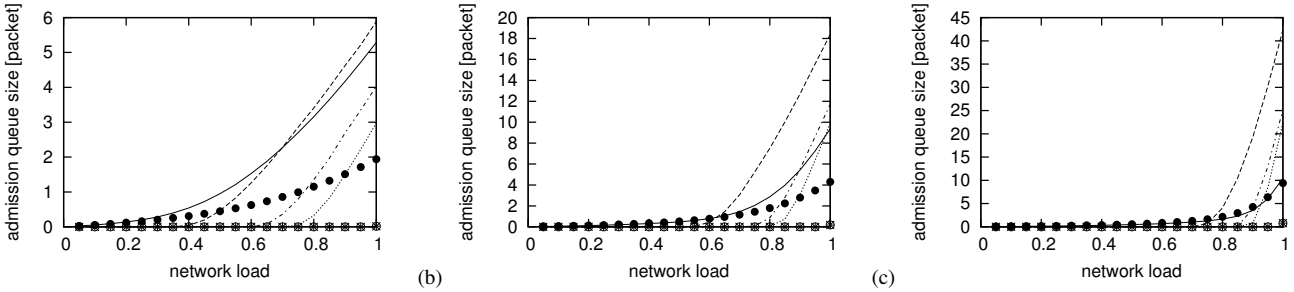


Fig. 7: Comparison of the mean admission queue size.