Wavelength Stealing: An Opportunistic Approach to Channel Sharing in Multi-chip Photonic Interconnects

Arslan Zulfiqar
University of Wisconsin-Madison
zulfiqar@wisc.edu

Pranay Koka
Oracle Labs
pranay.koka@oracle.com

Herb Schwetman
Oracle Labs
herb.schwetman@oracle.com

Mikko Lipasti
University of Wisconsin-Madison
mikko@engr.wisc.edu

Xuezhe Zheng
Oracle Labs
xuezhe.zheng@oracle.com

Ashok Krishnamoorthy
Oracle Labs
ashok.krishnamoorthy@oracle.com

ABSTRACT

Silicon photonic technology offers seamless integration of multiple chips with high bandwidth density and lower energy-per-bit consumption compared to electrical interconnects. The topology of a photonic interconnect impacts both its performance and laser power requirements. The point-to-point (P2P) topology offers arbitration-free connectivity with low energy-per-bit consumption, but suffers from low node-to-node bandwidth. Topologies with channel-sharing improve inter-node bandwidth but incur higher laser power consumption in addition to the performance costs associated with arbitration and contention.

In this paper, we analytically demonstrate the limits of channel-sharing under a fixed laser power budget and quantify its maximum benefits with realistic device loss characteristics. Based on this analysis, we propose a novel photonic interconnect architecture that uses opportunistic channel-sharing. The network does not incur any arbitration overheads and guarantees fairness.

We evaluate this interconnect architecture using detailed simulation in the context of a 64-node photonically interconnected message passing multichip system. We show that this new approach achieves up to 28% better energy-delay-product (EDP) compared to the P2P network for HPC applications. Furthermore, we show that when applied to a cluster partitioned into multiple virtual machines (VM), this interconnect provides a guaranteed 1.27× higher node-to-node bandwidth regardless of the traffic patterns within each VM.

Categories and Subject Descriptors

C.1.2 [**Computer Systems Organization**]: Multiprocessors—*Interconnection architectures*; B.4.3 [**Hardware**]: Interconnections—*Topology*

MICRO-46, December 07-11, 2013, Davis, CA, USA

ACM 978-1-4503-2638-4/13/12 ...$15.00.

General Terms

Design, Performance

Keywords

Interconnection Networks, Nanophotonics

1. INTRODUCTION

The trend towards many-core systems continues to grow [26, 17]. Scaling single chip systems for higher performance leads to increasing fabrication costs and low process yields [24, 20, 15]. Multi-chip systems can alleviate these concerns but require substantial chip-to-chip bandwidth to provide sustained performance. However, due to the packaging limitations of chip I/O pins and excessive power consumption of high-speed serial links, silicon-photonic technology has been proposed as an alternative for networking multi-chip systems [15, 16]. Optical interconnects offer "speed-of-light" communication at high-bandwidth density enabled by wavelength-division-multiplexing (WDM) that allows multiplexing of many parallel streams of information into a single waveguide or fiber. This performance gain is envisioned with lower energy consumed-per-bit requirements compared to electrical interconnects [16, 2].

To make optical communication a reality in multi-chip computing systems, two types of challenges need to be addressed: device-level and architectural. Device-level challenges involve design and fabrication of optical devices that are low-loss and high speed. Such optical devices include components such as modulators, drop-filters, couplers, waveguides, etc. and constitute the building blocks of a silicon photonic network. Fabrication of these devices is under extensive on-going development and many components have been demonstrated in the literature [33, 31].

From an architectural standpoint, the main challenge is to design an interconnect that is energy efficient at realistic device loss parameters and yields the best performance on the target applications. This architectural challenge is exacerbated by the fact that there is no clear roadmap or consensus on loss assumptions of photonic devices. Hence, many interconnect designs have been proposed in the literature ranging from the simplest point-to-point (P2P) network [15, 14] to numerous designs based on wavelength sharing [28, 21, 22].

In photonic networks, a channel (logical connection) between a sender and a destination is formed using one or more waveguides. Each waveguide can support multiple wavelengths (links) using WDM, and these wavelengths carry bit

information in the form of modulated light. A simple P2P network statically partitions the total network bandwidth (wavelengths) between the sender-destination pairs leading to relatively low bandwidth (narrow) node-to-node channels. On the other hand, a network that enables sharing combines wavelengths to form a single logical high bandwidth (wide) shared channel. Thus, sharing-based networks can potentially provide higher node-to-node bandwidths compared to a P2P network, albeit at the cost of arbitration delays in accessing the shared channel. The peak node-to-node BW is proportional to the following terms:

$$\text{Node-to-Node BW} \propto s \times Eff(s) \times \frac{Total\ wavelengths}{N^2} \quad (1)$$

where, N: total network nodes; s: sharing degree (≥ 1), and $Eff(s)$: efficiency of sharing $[0, 1]$. The fractional term $Eff(s)$ captures the costs associated with sharing, e.g. overheads of arbitration, fairness, etc. $Eff(s)$ is inversely proportional to the sharing degree s due to higher overheads (e.g. contention). Sharing ($s > 1$) can provide higher bandwidths compared to a P2P network ($s = 1$) as long as the costs do not outweigh the benefits i.e. $s \times Eff(s) > 1$. In addition to the efficiency penalty, high static power consumption is another significant cost associated with sharing in photonic networks.

Photonic networks based on ring resonators are static power dominated because of laser power and ring tuning power [21, 14, 12]. A higher degree of sharing requires more devices along a wavelength, thereby increasing the required input laser power (optical) and the device tuning power (electrical). Efficiencies of commercially available WDM lasers are $1 - 5\%$, and may be expected to exceed 10% in the next decade [34, 5, 18]. When laser efficiency is considered, laser power becomes the dominant contributer to static power dissipation. Thus, optimizing for laser power must be considered a first-order design constraint. The laser power consumption is proportional to the following terms:

$$\text{Laser Power Consumption} \propto Total\,wavelengths \times \overbrace{\underbrace{\frac{\#\ devices}{wavelength}}_{\text{Increases with sharing } s} \times \frac{loss}{device}}^{\text{Avg. loss per wavelength}} \quad (2)$$

This paper uses the power-constrained design approach described in [14] which assumes a *fixed* input laser power budget for all designs under consideration. This constraint ensures that any performance gains that arise from sharing do not come with the costs of increased laser power consumption. Equating the laser power consumption of a sharing design to a P2P network using Eq.(2) leads to the observation that:

$$Total\ wavelengths_{\text{sharing}} < Total\ wavelengths_{\text{P2P}}$$

Thus it is clear that the total peak bandwidth of a network with wavelength sharing will be lower than that of an energy-equivalent point-to-point network. If this sharing design can still provide higher node-to-node bandwidth (Eq.(1)) even with fewer total wavelengths, then it may be the preferred design choice over a P2P network depending on the target applications. Thus, a sharing design can win on performance (BW) *and* power (laser) only when:

$$\underbrace{s \times Eff(s)}_{(>1)} \times \underbrace{\frac{Total\ wavelengths_{\text{sharing}}}{Total\ wavelengths_{\text{P2P}}}}_{(<1)} > 1 \quad (3)$$

Most prior sharing-based proposals have assumed very *aggressive* values for device losses. This has led to designs in which sharing has had negligible impact on the average loss per wavelength in Eq.(2) leading to $\frac{Total\ wavelengths_{\text{sharing}}}{Total\ wavelengths_{\text{P2P}}} \approx 1$. With no penalty from this ratio term in Eq.(3), these designs have pushed sharing to very high levels (e.g. $s = 64$) and have shown significant performance gains with minimal impact on laser power consumption.

This paper models the impacts of *conservative* loss assumptions on photonic network design and makes the following contributions:

- An analytical model to determine limits on sharing degree and the ideal gains of sharing,
- The design of a novel arbitration-free, energy efficient shared channel network architecture, called *wavelength stealing*,
- Detailed performance evaluation of the wavelength stealing architecture implementation on a single-layer wafer-scale multi-chip system, and
- Application of the wavelength stealing architecture to improve the network throughput of a partitioned multi-chip cluster using a smart hypervisor.

The rest of the paper is organized as follows. Section 2 covers background and related work. Section 3 discusses the additional losses that arise due to sharing and quantifies the ideal performance gains achievable by a sharing design. Section 4 presents a novel sharing-based design called wavelength stealing. Implementation of wavelength stealing on a multi-chip system is discussed in section 5, and the application of wavelength stealing architecture to support multiple virtual machines is presented in section 6. Section 7 discusses the evaluation methodology and results, and section 8 concludes the paper.

2. BACKGROUND AND RELATED WORK

In recent years, a number of silicon-photonic interconnect designs have been proposed ranging from the simplest P2P network [4, 15, 14, 16] to numerous sharing-based designs [28, 23, 10]. In most cases, the assumed device losses have varied greatly. Some designs have made aggressive (low) loss assumptions and, as a result, have been able to show significant performance gains. Yet, other papers have assumed more conservative device loss parameters and have argued for simpler designs and topologies to keep the laser power consumption low. In this section we discuss the various categories of shared channel topologies and show how the wavelength stealing architecture fits into the broader landscape.

2.1 Classification of Shared Channel Topologies

Optical crossbars are often used to implement channel sharing. Such designs fall into the following four general categories:

1) SWSR (Single-Writer, Single-Reader) - Each communication channel has only one source and one destination. This architecture is essentially a statically allocated point-to-point (P2P) topology [15, 14]. An optical point-to-point

topology has the least device complexity and link loss. However, it suffers from low node-to-node bandwidth.

2) SWMR (Single-Writer, Mutiple-Reader) - Multiple reader channels are typically implemented using broadband switches or tunable microrings to selectively divert all the optical energy to one destination. Due to active and pass-through losses of these devices, multiple reader channels have significant link losses that increase with the sharing degree. In addition, SWMR networks require a broadcast-based mechanism to notify the target destination to tune in and the other destinations on the channel to tune out. The Firefly architecture [22] implements a SWMR based interconnect across a 64-node network.

3) MWSR (Multiple-Writer, Single-Reader) - Network architectures in this category require switches or microrings to enable shared access to a waveguide or selective wavelengths in a waveguide from multiple sources. This category of interconnects have a similar link loss as the SWMR networks and require an arbitration mechanism at the source to resolve access conflicts to the shared channel. The Corona network [28] implements a MWSR architecture with a high degree of sharing using ring modulators and token arbitration mechanism across the entire system. Vantrease *et al.*[27] build on [28] and advocate using time-slotted channels with continuous token arbitration to improve the channel utilization. The shared-source-row architecture [15] implements a variation of MWSR architecture using MZI broadband switches and a two-phase arbitration mechanism.

4) MWMR (Multiple-Writer, Multiple-Reader) - MWMR networks require switches/rings both at the source and the destination leading to the highest link loss compared to the other three categories. These networks require both an arbitration mechanism at the source and a mechanism to select the appropriate destination. MWMR channels were first considered by [21] where both senders and receivers share the channels. This design proposed a "two-pass" continuous token arbitration scheme between senders and required receiver side arbitration as well on its MWMR channels. The "Channel Borrowing" scheme [30] argues for simplifying the two-pass token arbitration proposed in [21] by restricting the number of senders on shared channels to two.

All shared channel topologies require some form of arbitration at the source and/or at the destination depending on how the sharing is implemented. A number of techniques to incorporate sharing [6, 23, 11, 9, 13, 21, 27, 28, 15, 14] have been proposed. Each of these techniques suit different topologies and differ in complexity and latency overheads.

2.2 Sharing in the Wavelength Stealing Architecture

The shared network architectures described above have good performance characteristics but suffer from the following issues. First, with conservative device losses, SWMR and MWSR architectures [28, 21, 15] suffer from high link loss at high sharing degrees due to the requirement for a large number of rings/switches. Such architectures are hard to implement and are energy inefficient [14]. Second, MWMR topologies such as [30, 21] have high link losses even at low sharing degrees due to the requirement for rings/switches at both the destination and the source. Third, even if the sharing degrees are reduced significantly to control link loss, the arbitration overheads can negatively affect performance.

An ideal shared channel architecture should provide higher node-to-node bandwidth than the SWSR networks with low optical link loss and minimal arbitration overheads. In order to achieve this, the wavelength stealing architecture introduced in this paper implements an MWSR-type sharing over a fully-connected point-to-point (P2P) topology and avoids arbitration completely by using a novel aggressive channel-stealing mechanism with graceful recovery from collisions. In section 7, we show that our arbitration-free architecture exhibits lower latency and better throughput performance compared to traditional arbitration-based architectures.

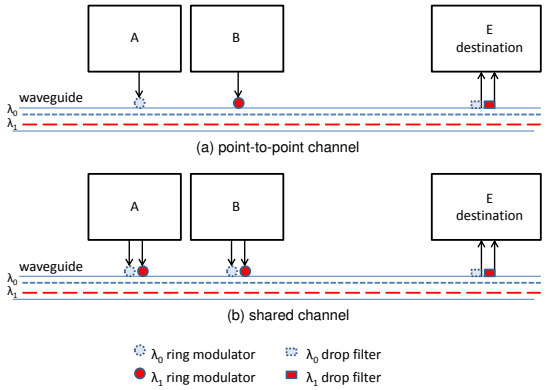


Figure 1: A point-to-point (P2P) versus shared channel. Due to extra modulator rings, light on a shared wavelength suffers from higher losses.

3. SHARING IN PHOTONIC NETWORKS

3.1 Ring Modulator Losses

Fig. 1a shows a waveguide carrying two wavelengths in a point-to-point topology where source nodes 'A' and 'B' modulate different wavelengths to destination node 'E'. Each modulator ring placed along the waveguide is tuned to a specific wavelength and modulates light on that wavelength. Modulation is controlled by electrically biasing the ring using the data stream to either pass light (transmit a '1') or absorb light (transmit a '0'). An active ring (that modulates a wavelength) causes a significant insertion loss of $4.0dB$ to the wavelength. As shown in the figure, the wavelength of light also passes by rings that are tuned to other wavelengths of a waveguide. These rings cause a smaller passive through-loss of 0.05dB per ring.

3.2 Wavelength Sharing

Fig. 1b shows a waveguide carrying two wavelengths that are shared by two senders 'A' and 'B' to a destination node 'E'. Each node sharing a wavelength has a ring along the waveguide tuned to that wavelength. Thus in Fig. 1b, each wavelength passes by twice as many rings compared to a wavelength in the point-to-point channel. Multiple active rings on a wavelength will significantly increase the loss even though only one of them would be transmitting data. To achieve lower loss, a ring can be detuned dynamically away from the target wavelength as long as it is not transmitting data. However, due to the fast response times required, it is not feasible to detune a ring far enough from the target wavelength to make the loss negligible. Even with aggressive device techniques, we can expect a loss of $0.5dB$ per detuned (inactive) ring. In this work, we assume that tuning or de-

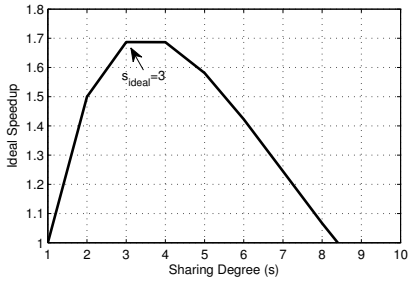


Figure 2: Ideal speedup versus sharing degree s assuming $w = W_{sharing} = 16$ and $T_{prop} = 0$.

tuning the microrings will occur in one bit time. This is an aggressive device technology goal and is under investigation.

3.3 Ideal Sharing Gains

From Fig. 1b, it is evident that wavelength sharing increases the link loss. This section explores the limits on sharing imposed by these additional losses. By extending the topology shown in Fig. 1b to sharing degree s and WDM factor w, the additional optical power loss of a shared wavelength compared to the P2P wavelength becomes:

$$\begin{aligned}\triangle L_{dB}(\lambda) &= Loss_{sharing} - Loss_{P2P} \\ &= (s-1)[\underbrace{0.5dB}_{\text{inactive rings}} + \underbrace{(w-1)0.05dB}_{\text{other } \lambda \text{ rings}}] \end{aligned} \quad (4)$$

Now, the amount of laser power consumed by $W_{sharing}$ wavelengths in a shared design and W_{P2P} unshared wavelengths in the P2P design is given by:

$$P_{sharing} = W_{sharing} \times 10^{(P_{rx}+\triangle L_{dB}(\lambda)+Loss_{P2P})/10}$$

$$P_{P2P} = W_{P2P} \times 10^{(P_{rx}+Loss_{P2P})/10}$$

By equating these two equations, the number of unshared wavelengths that consume *equivalent laser power* to a given number of shared wavelengths can be expressed as:

$$W_{P2P} = W_{sharing} \times 10^{\triangle L_{dB}(\lambda)/10} \quad (5)$$

This equation clearly shows that under the equivalent laser power constraint, the unshared P2P network can support higher number of wavelengths and hence offers higher total bandwidth (capacity) than a shared design. However, sharing can lead to higher node-to-node bandwidths over the P2P network provided there is no contention on the shared channel. We quantify these node-to-node bandwidth gains below.

We define $Speedup_{ideal}$ to be the ratio of time taken by a message of size $messagesize$ to be delivered to a destination on a P2P (unshared) channel versus time taken on a shared channel. It can be computed as:

$$Speedup_{ideal} = \frac{\left[\frac{messagesize}{W_{P2P}} + T_{prop}\right]}{\left[\frac{messagesize}{s \times W_{sharing}} + T_{prop}\right]} \quad (6)$$

where T_{prop} is the propagation time between the sender and destination. This definition of speedup is called "ideal" because it does not associate any overheads (in terms of time or wavelengths) with sharing.

Fig. 2 shows the ideal sharing gains achievable as a function of sharing s assuming 16-way WDM waveguides. From Fig. 2 and Eq.(6), the following observations can be made:

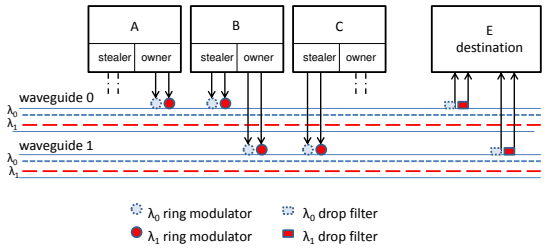


Figure 3: A 2-way wavelength stealing design example showing sender B's channels to destination E. Sender B can send $2bits/cycle$ guaranteed on its (owned) channel to E, and can opportunistically steal bandwidth on A's channel to send 2 extra $bits/cycle$ provided A is not using its channel. Note that this figure does not show the stealing channel of sender A and the owned channel of sender C to destination E.

- The ideal achievable speedup is independent of message size assuming $T_{prop} = 0$. This is because the $messagesize$ term in the numerator and denominator simply cancel each other out in Eq.(6).
- Wavelength sharing is only effective at low sharing degrees. In fact, ignoring all overheads of sharing, the optimal[1] sharing degree is just 3 (s_{ideal}).
- Beyond the optimal point, the number of wavelengths in the shared channels decreases significantly leading to a drop in the achievable speedup.

4. WAVELENGTH STEALING ARCHITECTURE

This section presents a novel interconnection architecture for multi-chip systems called *wavelength stealing*.

4.1 Design Overview

The topology of the wavelength stealing interconnect is similar to that of a point-to-point (P2P) network. Each node in the system has a dedicated channel (one or more waveguides) to every other node in the system and is called the 'owner' of that channel. The owner has non-blocking access to send information to a destination using its dedicated channel and is always guaranteed service on that channel. In addition to its dedicated channel, the sender can also steal access to channels owned by other senders to that destination. However, access to this additional (stolen) bandwidth is not guaranteed. Fig. 3 shows an example where node 'B' has a dedicated (owned) channel to destination 'E' and can also steal on the channel owned by node 'A'. Similarly node B's dedicated channel to E can be stolen by another node 'C'. Hence every channel in the system is owned by one node and can be stolen by one or more other nodes. Stealing is performed arbitration-free (without notification to the owner or other stealers). Any errors (collisions) that arise from stealing are corrected at the destination using mechanisms described in later sections. Stealing is a form of wavelength sharing and is accomplished by placing additional modulator rings along the shared waveguide as shown in Fig. 3. These additional rings cause higher wavelength losses (as described in Sec. 3). Hence to match the laser

[1]Lowest sharing degree with the highest speedup value.

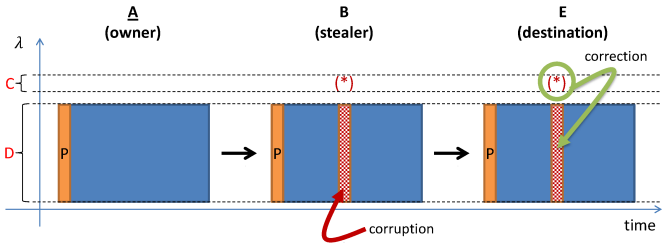


Figure 4: Erasure coding example. Corruption in A's message due to a collision from B gets marked $(*)$ in the control wavelengths. This location information is used to perform erasure correction at the destination.

power budget of a P2P network, the wavelength stealing architecture will have to use fewer wavelengths per channel than the P2P network. However, it can still provide higher node-to-node bandwidths than the P2P network provided stealing access on other channels is successful.

4.2 Implementation Details

For correct operation, an implementation of the wavelength stealing design should satisfy some strict requirements:

1. The owner must be guaranteed non-blocking access without any arbitration delays.
2. A stealer can steal bandwidth without arbitration (no prior notification to the owner or other stealers) and should be notified if it needs to stop stealing.
3. The destination must be notified if a received phit is corrupted due to collision and must be able to correct the bit errors. On receiving a valid phit, the destination must be able to identify the sender of the phit.

To meet the above requirements, the wavelength stealing architecture employs erasure coding [25] and special control wavelengths per channel. For simplicity, the rest of this section assumes only one stealer per channel.

4.2.1 *Erasure Coding*

In the wavelength stealing architecture, a stealer is allowed to steal (use wavelengths) on a channel without prior notification to the owner (i.e., it is arbitration-free). In this case, whenever a stealer steals on a channel on which the owner is actively sending data, a collision occurs, causing errors in the owner's message. These errors are corrected at the destination using erasure coding. When a collision occurs, a stealer is notified by the control wavelengths to stop stealing, preventing further errors in the owner's message. This ensures that an owner's message is never corrupted beyond the point of recovery. Erasure codes rely on location information of potential errors to provide better correction capability than codes that correct random bit errors [8]. For example, with location information, a parity code is capable of *correcting* a single bit error. In the case of multi-bit errors, stronger erasure codes can be employed [8].

Fig. 4 shows a channel in the wavelength stealing architecture with associated data wavelengths (indicated by D on the y-axis) and control wavelengths (indicated by C on the y-axis). The owner's message (A) has a parity column appended to it. As this message goes past the stealer (B), B steals on the owner's channel leading to an error. This error is automatically marked (discussed below) in the control wavelengths $(*)$. A stealer detects collisions with the help of the control wavelengths and stops stealing to prevent further errors. The corrupted message arrives at the destination (E) where the computed parities are compared with the parity column in the message. If there is a parity mismatch, the corresponding bits at the marked location are inverted to correct the bits in error.

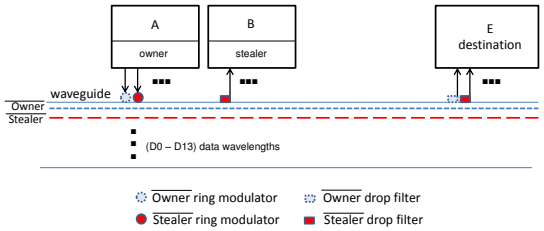


Figure 5: Abort control wavelengths.

It is important to emphasize that if no errors are marked in the control wavelengths then a received message is completely error-free and the destination doesn't need to wait for the subsequently arriving parity bits. Thus, in the absence of contention at low loads, the latency overhead of accessing the shared channel is completely hidden and messages only experience minimal latencies[2] which is not possible in a design based on arbitration.

4.2.2 *Control Wavelengths - Two Designs*

The control mechanism for wavelength stealing can be implemented using one of two designs, called abort and sense. These designs exhibit different trade-offs but provide the following functionality:

1. Mark the location of corrupted bits for erasure correction at the destination.
2. Inform stealer to stop stealing when the owner becomes active to limit the corruption to a single bit collision.
3. Inform destination of the ID (owner's, stealer's, or corrupted) of the received communication (phit).

Abort Design.

Fig. 5 shows a channel consisting of one waveguide to destination 'E' owned by sender 'A' with a stealer 'B'. $\overline{Owner}$ and $\overline{Stealer}$ are the control wavelengths and $D0 - D13$ are the data wavelengths in the waveguide. The behavior of the control wavelengths in the abort design is given in Table 1. When the owner (A) is not using the channel, it transmits a continuous 10 on the control wavelengths $\overline{Owner}$ and $\overline{Stealer}$ respectively. If the owner (A) uses the channel, it transmits a continuous 01 on the two control wavelengths. When the stealer (B) needs to transmit data to E it begins data transmission on its dedicated channel to E and steals the channel owned by A. Sender B also turns on the drop filter on the $\overline{Stealer}$ wavelength. The drop filter pulls out all light (bits) traveling on the control wavelength. If a value of 0 is read by the drop filter, then the stealer (B) knows that there has not been a collision with the owner. If the drop

[2]There are no latency overheads beyond message serialization delay and propagation delay.

Active Sender	A		B	E		Received
	$\overline{Own.}$	$\overline{St.}$	$\overline{St.}$	$\overline{Own.}$	$\overline{St.}$	
A	0	1	–	0	1	A
B	1	0	0	1	0	B
A, B	0	1	1	0	0	$Collision$
(Invalid)	1	1	–	1	1	(Invalid)

Table 1: Abort design functionality for owner (A), stealer (B) and destination (E). (The values 11 should not arise during normal system operation.)

filter reads a value of 1, then the stealer (B) knows that a collision has just occurred. It then suspends stealing, but continues to use its dedicated channel to E. At the destination side, a 01 indicates owner's (A) phit, a 10 indicates stealer's (B) phit and a 00 represents a corrupted (collided) phit. The destination tracks the control wavelength information to perform the protocol steps discussed in Sec. 4.2.3.

Sense Design.

The sense design requires separate waveguides for control and data. The control waveguides of two owner channels 'A' and 'B' are as shown in Fig. 6. The need for separate waveguides arises because this design uses optical splitters which are fabricated as broadband devices that split all wavelengths in a waveguide. Since the splitting functionality is only required for the control wavelengths, they are placed in waveguides that are separate from the data wavelengths. There is only one control wavelength per control waveguide, called $Owner$ and abbreviated as "OW" in the rest of the discussion. The control wavelengths for the owner A's channel and owner B's channel are denoted by $OW(A)$ and $OW(B)$ respectively. Some useful terminology is defined in Fig. 7.

In the sense design, the control functionality of the owner (A), stealer (B) and destination (E) depends on both the current and previous values (state) of the control wavelengths (OW) as shown in the state machine diagrams in Fig. 7. The state machine diagram for owner (A) shows that whenever A uses its channel, it puts a continuous 1 on $OW(A)$. The operation of the stealer (B) then depends on the value of $OW(A)$. From the stealer's (B) state machine, it is clear that it can be in one of two states when it has a message to send: STEAL or SENSE. In the STEAL state, the stealer (B) can actively steal on the owner's (A) channel. Now, if the owner becomes active ($OW(A) == 1$), then the stealer (B) transitions to the SENSE state. While in this state, the stealer does not steal and simply waits for an opening on the owner's (A) channel so that it can revert to stealing.

Note that the destination state machine needs to monitor the control wavelength of both the owner (A) and the stealer (B) to function properly. If the destination observes both $OW(A) == 1$ and $OW(B) == 1$, it knows that a collision has occurred. The destination then transitions into the SENSE state. While in the SENSE state, the only valid phit that is received is from the owner (A). The rest of the functionality in these state machines (Fig. 7) is self explanatory.

Abort vs. Sense Trade-offs.

Device-Level Trade-offs: The control wavelengths of the abort design can be accommodated with the data wavelengths of a channel in a single waveguide. The sense design requires separate waveguides for the control wavelengths.

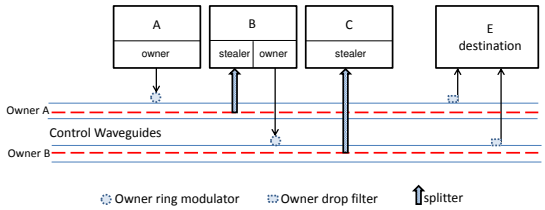


Figure 6: Sense control waveguides.

However, the sense design requires fewer modulator rings than the abort design, and hence is more energy-efficient.

Performance Trade-offs: The sense design can potentially provide better performance gains than the abort design because of its 'sensing' capability. That is, the sense design does not require the stealer to abort stealing upon collision of its message; instead it temporarily halts stealing and waits for an opening to revert to stealing. The abort design does not have the sense capability and thus has to operate more conservatively.

4.2.3 Protocol Operation

When a sender node needs to transmit a flit, it performs several steps. These steps are explained according to the example channels shown in Fig. 3 where the sender 'B' has a flit to send to destination 'E':

1. B's flit has T phits (value of T is known at design time).
2. Split the flit occupying T cycles into two chunks each of length $T/2$ phits: 'owner chunk' and 'stealer chunk'.
3. Parity protect the owner chunk and send it on B's channel.
4. Send the stealer chunk on A's channel.
5. If a collision occurs:
   - Abort design: Terminate stealing. The unsent phits are parity protected and sent on B's channel after the owner's chunk is sent.
   - Sense design: Halt stealing. Resume if an opening is sensed. If the owner chunk completes before the stealer chunk, then send the remaining stealer chunk phits (with parity protection) on B's owned channel.
6. The destination uses the information on the control wavelengths to perform erasure correction and correctly reassemble the received phits into the original flit.

The protocol operation described above assumes a basic flit size of T phits. It can be extended to support flits of multiple sizes (number of phits). For example, to support flits of two sizes - data flits and control flits - just two control bits are needed at special locations in the flit to identify its size at the destination. For the data flit, the sender can set the first bit of the first two phits in the owner chunk to 0. Now, even if one of the phits gets corrupted, the destination can look at the duplicated value to know which size flit this is. For control flits, the sender can use the value 1. For large messages, these two bits will amount to negligible overhead.

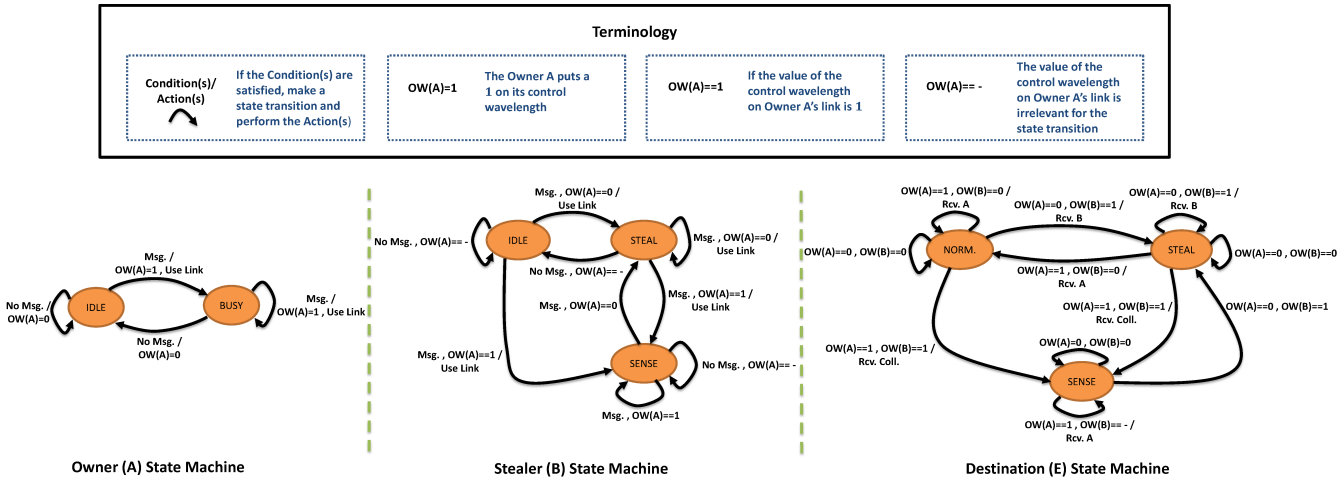


Figure 7: Sense design functionality for owner (A), stealer (B) and destination (E).

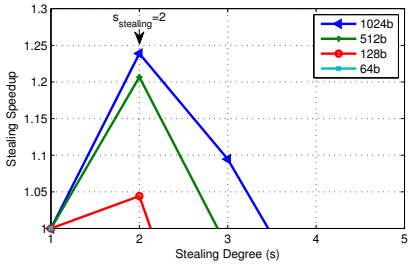


Figure 8: Wavelength stealing gains versus stealing degree s for different message sizes assuming $w = W_{sharing} = 16$ and $T_{prop} = 0$. The small $64b$ message does not exhibit a speedup.

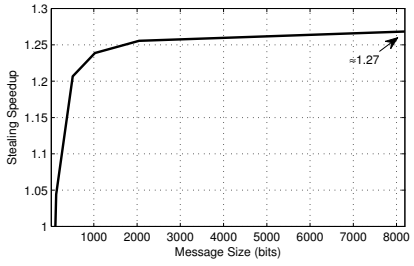


Figure 9: Wavelength stealing speedup as a function of message sizes for $s = 2$.

4.3 Wavelength Stealing Gains

In section 3, we analyzed the ideal case benefits and limits of a wavelength sharing network [3]. This section extends the analysis to the wavelength stealing architecture taking into account the overheads of control wavelengths and erasure coding.

The achievable speedup of the wavelength stealing architecture as a function of the stealing degree s can be expressed as:

$$Speedup_{stealing} = \frac{\left[\frac{messagesize}{W_{P2P}} + T_{prop}\right]}{\left[\frac{messagesize}{s \times \{W_{sharing} - c(s)\}} + e(s) + T_{prop}\right]}\ ;\ (s \geq 2) \quad (7)$$

where, s: stealing degree, $c(s)$: control wavelength overheads; and, $e(s)$: erasure coding overheads. For 2-way ($s = 2$) stealing, $c(2) = 2$ (two control wavelengths per channel), and $e(2) = 1$ (single parity bit). For any arbitrary $s \geq 3$, the number of stealers on a channel is $(s-1)$. This requires control overheads $c(s)$ that scale linearly with s. In addition, the minimum number of check bits $e(s)$ required to correct up to $(s-1)$ erasures can be estimated from the Hamming bound [19].

Fig. 8 plots the speedup gains of the wavelength stealing architecture as a function of the stealing degree s. From this figure, the following observations can be made:

- Ignoring the overheads of sharing, the ideal sharing degree is $s_{ideal} = 3$ (shown in Sec. 3.3). However, due to overheads, the wavelength stealing architecture yields maximum speedup at a stealing degree of $s = 2$ (2-way stealing).
- Contrary to an ideal wavelength shared network, the speedup in the wavelength stealing architecture is dependent on the message (flit) size. This dependency is due to the overheads of erasure correction coding which get amortized better at larger message sizes. Fig. 9 plots Eq.(7) as a function of message sizes for 2-way stealing. This figure clearly shows higher speedups for large messages with saturation at a speedup of 1.27.

The wavelength stealing architecture implements dedicated all-to-all connectivity similar to a P2P but is able to achieve higher node-to-node bandwidth in the presence of idle channels (for stealing to be successful) while consuming equivalent optical power. From the speedup analysis, it is also clear that the performance gains of the wavelength stealing architecture are more pronounced for larger messages. This makes the architecture more suitable to message passing applications that exhibit large-messages and low "fan-out" communication patterns [14, 32].

[3]The speedup discussion presented in this section assumes the abort design. The sense design will experience similar speedups because it uses the same erasure coding technique and its single broadcast control wavelength consumes (approximately) the same laser power as the abort design's two control wavelengths.

5. "MACROCHIP" - A MESSAGE-PASSING MULTI-CHIP SYSTEM

This section presents the architecture of a photonic multi-chip system called the Macrochip [16] on which the wavelength stealing techniques are evaluated. The macrochip architecture consists of an array of sites (also called nodes). These sites are interconnected using a high-bandwidth silicon-photonic communication substrate. Sites in the macrochip can be processor chips (with multiple cores), memory chips or some other components. This paper uses a configuration in which all sites have processors and memory that generate messages directed to other sites in the array [14]. This paper does not cover the details of the site architecture; instead, it focuses on the site-to-site optical interconnect.

5.1 System Layout

The layout of a 64-site macrochip system is shown in Fig. 10. Each site has an optical bridge chip on the top layer and communicates with the other sites using data waveguides in the bottom substrate layer. The optical bridges house the optical devices and circuitry to support them. Optical (laser) power is generated by external lasers and delivered to the macrochip using edge connected fibers. This laser light is then forwarded to the sites using power waveguides (shown in red) for modulation. The data waveguides carry modulated light for inter-site communication.

Fig. 10 shows a fully connected point-to-point layout composed of data waveguides shown as a blue loop. When implemented, the data path is composed of multiple waveguide segments where each segment begins at a sender site and terminates at a destination site and does not form a loop. Between any two nodes, there are two possible paths for laying out a channel between them: a clockwise and a counter-clockwise path. With these two choices, channels in this layout are designed such that the propagation distance between a sender and its destination is minimized. Thus, sender 7 has a counter-clockwise channel to destination 0 and a clock-wise channel to destination 44 as shown in Fig. 10. Consequently, for a given destination, half of its senders will have channels that go in the clockwise direction towards that destination, and the other half will have channels that travel in the counter-clockwise direction.

Implementing the wavelength stealing interconnect on the macrochip requires placement of some modulator rings (for the stealers) in the bottom communication (waveguide) substrate. This can make the fabrication process more complex compared to a simple point-to-point interconnect. Interlayer couplers can be used to avoid having rings in the substrate layer; the trade-off however is higher link losses[4].

5.2 Stealing Pattern and Collision-Free Subsets

In the wavelength stealing architecture, a sender node uses its dedicated point-to-point channel to communicate with a destination but can also steal access on a channel to the same destination owned by another node. For a given destination, the static mapping between a sender and the node it steals from specifies the "stealing pattern" of a wavelength stealing topology. The wavelength stealing architecture for the macrochip uses a stealing pattern in which *a sender steals on channels owned by its two immediate bridge chip neighbors*

[4]The device loss of an inter-layer coupler is $\approx 2-3dB$ [14].

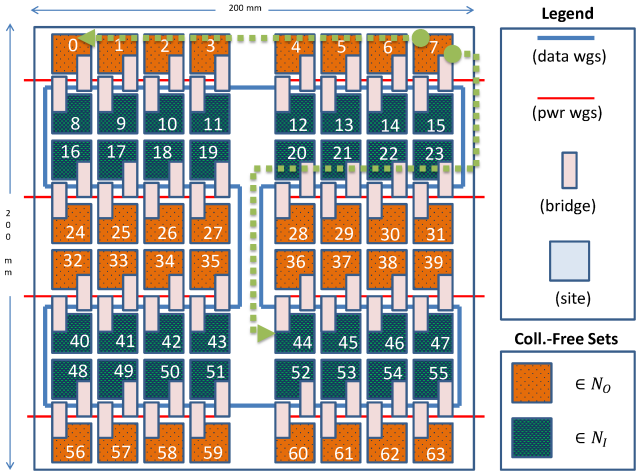


Figure 10: 8×8 single-layer (planar) macrochip layout.

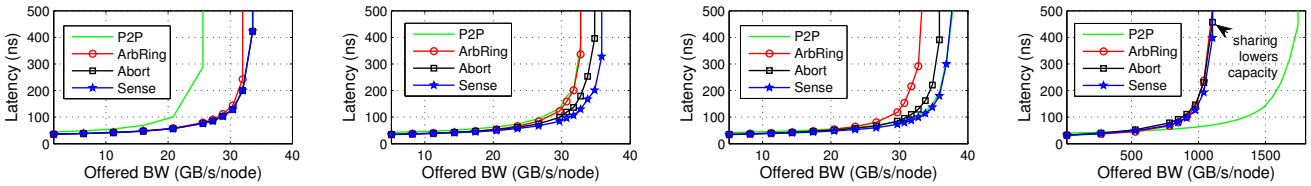
along the waveguide loop. Thus, in Fig. 10, sender 7 steals from 15 and 23 because they are its immediate two neighbors along the blue waveguide loop. To communicate with destination 0, sender 7 steals on node 23's channel to node 0. Similarly, to communicate with destination 44, sender 7 steals on node 15's channel to node 44. Thus, for a given destination, a sender steals from its immediate "upstream" neighbor along the waveguide loop. This upstream neighbor stealing pattern leads to a partitioning of the macrochip into two node sets, N_O and N_I, with the property that all nodes in one set steal only from nodes in the other set. These two sets are highlighted in Fig. 10 and are called *collision-free sets*. They are called collision-free because as long as nodes in one set do not communicate with destinations in the other set and vice versa, collisions never occur. This is because under this scenario, there is never a case where both the owner and the stealer of a channel talk to the same destination. Restating this more formally: *the two sets N_I and N_O are collision-free because when members of a set restrict their communication to nodes within the set, there are no collisions*. The collision-free property is valid regardless of the communication pattern and the number of active senders within the sets, as long as the restriction on *no* inter-set communication is observed. This property is used extensively in the next section. For a N-node layout, the maximum number of nodes in each of the collision-free sets is $N/2$.

Pairing a node with an upstream neighbor to form an owner-stealer relationship leads to the farthest two senders of a destination being devoid of any channels to steal on. Only $\approx 3\%$ of the total source-destination pairs in the network fall into this category [5]. In order to maintain bandwidth symmetry, these few sender-destination pairs are provisioned with some additional wavelengths. The energy required for these is accounted for in the power budget.

6. GUARANTEED GAINS ON VIRTUAL MACHINES

An architectural implication of the collision-free sets is that the cluster of nodes on the macrochip can be partitioned into multiple Virtual Machines (VMs) such that

[5]For an N-node layout, this fraction is $2N/(N \times (N-1))$.



(a) Bit Compliment (b) Asymmetric20 (c) Asymmetric30 (d) Uniform Random

Figure 11: Synthetic traffic simulations depicting latency versus offered load for the three network architectures: wavelength stealing (Abort/Sense), token-ring arbitration (ArbRing) and point-to-point (P2P).

nodes within a VM always steal from nodes outside a VM. With no inter-VM communication, this architecture provides higher node-to-node bandwidth (because stealing is guaranteed to be successful with no collisions), and lower message latencies compared to a P2P network. To realize these VM gains, a hypervisor scheduling layer can be designed that schedules the VM jobs on the appropriate sites of the macrochip to ensure a collision-free operation.

To explain further, denote a VM job as $VM(np)$ where np is the number of processor chips (network nodes) required by this virtual machine for execution. In the 64-node macrochip system shown in Fig. 10, the two collision-free sets N_O and N_I contain 32 nodes each. This means that, two independent 32-processor virtual machines each hosting a multi-process or multi-threaded application can be scheduled *concurrently* and the intra application communication will not suffer *any* collisions in the network. With this scheduling, the wavelength stealing architecture will guarantee a 1.27× higher node-node bandwidth over the P2P network. Since any subset of a collision-free set (N_O or N_I) is also collision-free, multiple VMs that require fewer processors than 32 can be scheduled together on a single collision-free set and take advantage of the guaranteed bandwidth gains over the P2P network. Thus, a set of VMs $\{VM_0(16), VM_1(16)\}$ can be scheduled on N_I and an independent set $\{VM_2(16), VM_3(16)\}$ can be scheduled on N_O so that all of them execute concurrently without collisions.

In general, suppose there are $(m+n)$ VM that need to be scheduled on an N-node macrochip. A hypervisor scheduling layer can be constructed that maps each of these $(m+n)$ VMs to the appropriate collision-free subsets. This hypervisor simply partitions the total VMs into two sets such that each of them can be scheduled on an $N/2$ sized collision-free set. Formally put, the hypervisor scheduling layer is able to schedule the $(m+n)$ VMs if it can separate them into two sets, $S_m = \{VM_0(np_0), \ldots, VM_{m-1}(np_{m-1})\}$ and $S_n = \{VM_0(np_0), \ldots, VM_{n-1}(np_{n-1})\}$ such that they satisfy the following conditions:

$$\underbrace{S_m : \sum_{i=0}^{m-1} np_i \leq \frac{N}{2}}_{m \text{ VMs}} \quad ; \quad \underbrace{S_n : \sum_{j=0}^{n-1} np_j \leq \frac{N}{2}}_{n \text{ VMs}} \qquad (8)$$

7. RESULTS AND DISCUSSION

7.1 Evaluation Methodology

We performed detailed evaluation of the wavelength stealing architecture against two baseline designs: the unshared P2P network and the classic token-ring arbitration scheme [1] that has inspired many recent photonic network implementations [27, 28]. A detailed cycle-accurate network simulator was developed that models complete functionality of these interconnect architectures.

All designs were evaluated on the 64-node macrochip layout shown in Fig. 10. We used both synthetic and application-derived traffic from message-passing applications to evaluate the networks. These workloads are summarized in Table 2. Performance of the applications running in single cluster and partitioned cluster configurations was also analyzed.

Pattern		Description
Synthetic	High-Radix	Uniform Random
	Low-Radix	Permutation/ Asymmetric
Application	NAS BT	Block Tridiagonal Solver
	NAS CG	Conjugate Gradient Kernel
	NAS DT WH	"White Hole" Graph Analysis
	NAS DT BH	"Black Hole" Graph Analysis
	NAS DT SH	"Shuffle" Graph Analysis

Table 2: Workload descriptions.

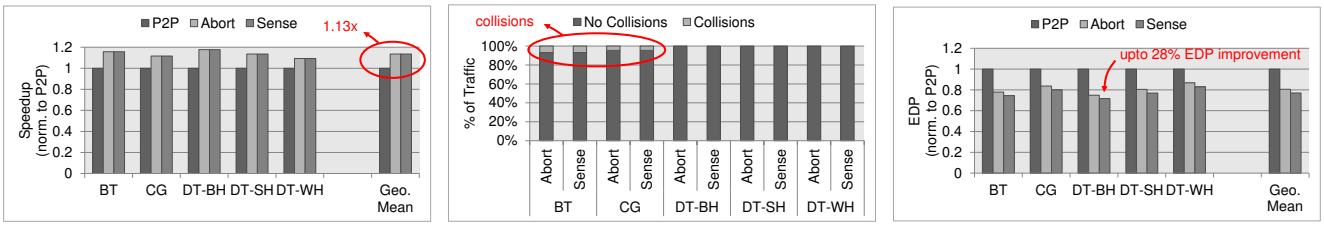
7.2 Synthetic Workload Evaluation

For synthetic workload evaluation, two categories of traffic patterns were simulated: high-radix and low-radix. We categorize a traffic pattern as having high-radix (low-radix) if a sender node communicates with a large (small) number of destination nodes. All synthetic patterns use a fixed message size of $1KB$. Synthetic simulation results are shown in Fig. 11.

7.2.1 *Wavelength Stealing vs. Arbitration*

In token-ring arbitration, a single token is circulated per shared channel. This token represents the exclusive right of a sender to use the shared channel. It is well-known that the token-ring design does not scale well to highly-shared channels owing to high token-rotation latencies [15]. However current device loss constraints restrict sharing to just two senders per shared channel. With limited sharing, the rotation latency of the token-ring design is very small making this scheme a competitive point of comparison for our designs.

We used three types of traffic patterns to compare performance. The bit-complement [7] (low-radix) traffic pattern causes no contention in the shared networks. To evaluate performance under various levels of contention, we used a new permutation pattern called Asymmetric k. In this traffic pattern, given an offered load, one of the two senders on the shared channel is active (on-average) k% of the time while the other is active $100-k$% of the time (note that bit-complement traffic represents $k = 100$%). Finally, the



(a) Execution time speedup (b) Traffic breakdown (Coll. vs. No Coll.) (c) *Energy* × *Delay* (*EDP*)

Figure 12: Application benchmark simulations.

uniform-random [7] traffic pattern represents all-to-all (high-radix) communication that causes uniform contention on the shared channels.

From Fig. 11, we see that as contention on the shared channel is increased, the throughput of the arbitration design drops significantly compared to the proposed stealing approaches. In addition the latency of the wavelength stealing designs is lower than the arbitration network, making it a good design-choice for latency-sensitive applications as well. Since the wavelength-stealing architecture performs either as well or better than a classical arbitration-based network, the rest of our evaluation will focus on the arbitration-free stealing architecture.

7.2.2 *Wavelength Stealing vs. Point-to-Point (P2P)*

As discussed in Sec. 3, sharing based networks have fewer wavelengths per channel and hence lower total bandwidth (capacity) compared to the P2P network. The effect of this can be observed from the uniform random ("all-to-all") traffic pattern in Fig. 11. The P2P network has has higher total bandwidth and hence exhibits higher sustained throughput on this pattern.

From Fig. 11, it can be observed that the wavelength stealing schemes yield 1.27× higher throughput than the P2P network on the contention-free bit complement traffic pattern as quantified in Sec. 4. As contention in the traffic increases (see asymmetric patterns in Fig. 11) the performance of the P2P network increases due to better utilization of the channels. In addition, the sense design gives better performance than the abort design at higher contention (see Sec. 4).

These simulations clearly show that the P2P network is ideally suited for high-contention traffic patterns while the sharing-based wavelength stealing architecture gives excellent performance under low-contention traffic. This fundamental design trade-off should be carefully considered when choosing a network implementation for a target application.

7.3 Application Workload Evaluation

For application-traffic simulations, five benchmarks listed in Table 2 were chosen from the NAS parallel benchmark suite [3]. Traces collected from the MPI versions of these benchmarks using Scalasca [29] were used to drive the network simulator.

7.3.1 *Performance Analysis*

To evaluate benchmark performance, we measured the execution time of the application traces on the P2P topology and the abort/sense designs of the wavelength stealing architecture. Fig. 12a shows the speedup as the execution time of the wavelength stealing designs relative to that of the P2P network. The wavelength stealing designs achieve up to 1.17× speedup on some benchmarks and a geometric-mean speedup of 1.13× over the P2P network. These benefits come from the low-contention traffic behavior of these applications. Fig. 12b shows that over 90% of the traffic in these applications does not suffer from any collisions and is able to utilize higher site-to-site bandwidths by successfully stealing idle channels. The variations in the achieved speedups between benchmarks arise due to the differences in their traffic patterns (collisions), message sizes and frequency of messages. Since much of the stealing is performed without contention, the conservative abort design performs on par with the sense design.

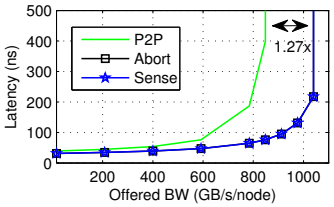
Parameter	Assumption
Mod. (Insertion) Ring Loss	$4dB$
Inactive Mod. Ring Loss	$0.5dB$
Active Drop-Filter Ring Loss	$1dB$
Passive Ring Loss	$0.05dB$
Waveguide Loss	$0.05dB/cm$
Bridge Chip Waveguide Loss	$1dB$
Coupler Loss	$2dB$
Receiver Sensitivity Margin	$4dB$
Receiver Sensitivity Level	$-21dBm$
Ring Tuning Power	$0.3mW/ring$
Mod. Driver	$35fJ/bit$
Detector Driver	$65fJ/bit$
Max. Fiber WDM-Factor	32
Max. Waveguide WDM-Factor	16
Max. Port Fibers	2500
Power per Fiber	$32mW$

Table 3: Optical device parameters.

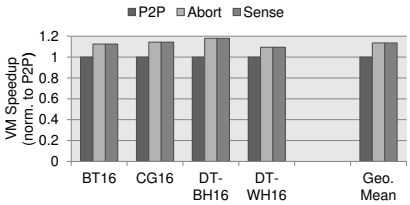
7.3.2 *Energy-Delay Analysis*

This section discusses the performance and energy trade-offs of the simulated networks. We use *Energy* × *Delay* (*EDP*) as our metric to compare the different network architectures. The static and dynamic energy for the networks were calculated using device parameters given in Table 3. The energy calculation for the wavelength stealing architecture takes into account the additional dynamic energy expended on the parity bits. However, this has negligible impact on the total energy because the dynamic energy consumed is just a small fraction compared to the static energy consumption of these networks.

Fig. 12c shows the *EDP* of the networks for each workload. This graph is normalized to the P2P network. The wavelength stealing architectures achieve up to 28% lower *EDP* than the point-to-point network in the best case. The abort and sense designs achieve on average (geometric mean)



(a) Domain uniform synthetic traffic pattern



(b) Application execution time speedup

Figure 13: Virtual machine performance gains. (a) Domain uniform synthetic traffic pattern depicting the collision-free subset property of the wavelength stealing architecture. (b) Four VMs are mapped into collision-free subsets to realize speedup gains.

20% and 23% lower EDP respectively over the P2P network. The sense design uses fewer rings than the abort design leading to a slight reduction in the static tuning power and hence a marginally better EDP.

7.4 Virtual Machine (VM) Evaluation

Sec. 6 discussed leveraging the collision-free subset property of the wavelength stealing architecture to partition the macrochip into multiple VMs where each VM can execute an application and realize guaranteed bandwidth gains over the P2P network irrespective of the traffic pattern. To highlight the collision-free property of the subsets, we used a variant of the uniform random traffic pattern called "Domain Uniform Random". This communication pattern is the same as the uniform random pattern in Table 2 except that senders belonging to a collision-free set only pick other nodes within the set as their random destinations. Fig. 13a shows the latency curve for this synthetic pattern. Because no collisions are encountered in the system and stealing is successful 100% of the time, the wavelength stealing architecture is able to achieve the theoretical 1.27× bandwidth advantage over the P2P network.

To explore the VM scheduling gains on applications, 16-node traces were collected for four NAS benchmarks listed in Table 2. The macrochip was partitioned into four clusters and the four applications were scheduled concurrently using the algorithm presented in Sec. 6. Fig. 13b shows the execution speedups observed on these four application derived traffic patterns. All four applications achieved positive speedups and experienced no collisions. These results show the potential applicability of the wavelength stealing interconnect on a wide range of cluster configurations.

8. CONCLUSION

Interconnects with shared optical channels overcome the low node-to-node bandwidth limitation of a simple P2P network but suffer from high optical losses. In this paper we developed analytical models to quantify the limits on shared channels and found that channel sharing with realistic photonics device losses does not scale beyond a sharing degree of three.

Based on this analysis, we proposed a novel interconnect architecture called *wavelength stealing* that enables arbitration-free optimistic access to shared optical channels and uses simple erasure coding to recover from collisions. Analytically, we showed that the maximum performance benefits of this architecture occurs with two sharers on every channel. We presented the design and implementation of such an architecture using the same input optical power budget as that of a P2P network.

We simulated the P2P network and the wavelength stealing interconnect in the context of a 64-node multichip system using synthetic and application derived traffic patterns. Using detailed performance and power analysis we have demonstrated that the wavelength stealing architecture exhibits up to 28% better EDP than the P2P network on applications with low-radix traffic. Furthermore we showed that the wavelength stealing architecture can be leveraged to partition a multi-chip cluster into multiple VMs with guaranteed bandwidth gains over a P2P network under certain constraints.

9. ACKNOWLEDGMENTS

This material is based upon work supported, in part, by DARPA under Agreement No. HR0011-08-09-0001. The authors thank Dr. Jag Shah of DARPA MTO for his inspiration and support of this program. The views, opinions, and/or findings contained in this paper are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for public release, distribution unlimited.

10. REFERENCES

[1] IEEE standards for local area networks: Token ring access method and physical layer specifications. *IEEE Std* 802.5 – 1989, 1989.

[2] M. Asghari and A. V. Krishnamoorthy. Silicon photonics: Energy-efficient communication. *Nature Photonics*, May 2011.

[3] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. C̃arter, L. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. S̃chreiber, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga. The NAS parallel benchmarks – summary and preliminary results. In *Proc. of the ACM/IEEE conference on Supercomputing*, Supercomputing '91, New York, NY, USA, 1991.

[4] S. Beamer, K. Asanović, C. Batten, A. Joshi, and V. Stojanović. Designing multi-socket systems using silicon photonics. In *Proc. of the International Conference on Supercomputing*, ICS '09, New York, NY, USA, 2009. ACM.

[5] B. Ben Bakir, A. Descos, N. Olivier, D. Bordel, P. Grosse, J. Gentner, F. Lelarge, and J.-M. Fedeli. Hybrid si/III-V lasers with adiabatic coupling. In

Group IV Photonics (GFP), 2011 8th IEEE International Conference on, Sept. 2011.

[6] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi. Phastlane: a rapid transit optical routing network. In *Proc. of the International Symposium on Computer Architecture*, ISCA '09, New York, NY, USA, 2009.

[7] W. Dally and B. Towles. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

[8] E. Fujiwara. *Code Design for Dependable Systems: Theory and Practical Application*. Wiley-Interscience, 2006.

[9] J.-H. Ha and T. M. Pinkston. A new token-based channel access protocol for wavelength division multiplexed multiprocessor interconnects. *J. Parallel Distrib. Comput.*, Feb. 2000.

[10] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanović, and V. Stojanović. Silicon-photonic clos networks for global on-chip communication. NOCS '09, New York, NY, USA, 2009. ACM.

[11] N. Kirman and J. F. Martínez. A power-efficient all-optical on-chip interconnect using wavelength-based oblivious routing. ASPLOS '10, New York, NY, USA, 2010. ACM.

[12] A. Kirshnamoorthy, X. Zheng, G. Li, J. Yao, T. Pinguet, A. Mekis, H. Thacker, I. Shubin, L. Ying, K. Raj, and J. Cunningham. Exploiting CMOS manufacturing to reduce tuning requirements for resonant optical devices. *IEEE Photonics Journal*, 3, June 2011.

[13] A. K. Kodi and A. Louri. Design of a high-speed optical interconnect for scalable shared-memory multiprocessors. *IEEE Micro*, Jan. 2005.

[14] P. Koka, M. McCracken, H. Schwetman, C.-H. Chen, X. Zheng, R. Ho, K. Raj, and A. Krishnamoorthy. A micro-architectural analysis of switched photonic multi-chip interconnects. ISCA '12, 2012.

[15] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. ISCA '10, 2010.

[16] A. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. Cunningham. Computer systems based on silicon photonic interconnects. *Proceedings of the IEEE*, July 2009.

[17] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal. ATAC: a 1000-core cache-coherent processor with on-chip optical network. PACT '10, New York, NY, USA, 2010. ACM.

[18] M. Lamponi, S. Keyvaninia, C. Jany, F. Poingt, F. L̃elarge, G. de Valicourt, G. Roelkens, D. Van Thourhout, S. M̃essaoudene, J.-M. Fedeli, and G. Duan. Low-threshold heterogeneously integrated InP/SOI lasers with a double adiabatic taper coupler. *Photonics Technology Letters, IEEE*, Jan. 2012.

[19] S. Lin and D. J. Costello. *Error Control Coding, Second Edition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2004.

[20] B. Murphy. Cost-size optima of monolithic integrated circuits. *Proceedings of the IEEE*, Dec. 1964.

[21] Y. Pan, J. Kim, and G. Memik. Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar. In *High Perf. Computer Architecture (HPCA), 2010*.

[22] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *Proc. of the Int'l Symposium on Comp. Architecture (ISCA)*, 2009.

[23] A. Shacham, K. Bergman, and L. P. Carloni. On the design of a photonic network-on-chip. NOCS '07, Washington, DC, USA, 2007. IEEE Computer Society.

[24] C. Stapper. On murphy's yield integral [IC manufacture]. *Semiconductor Manufacturing, IEEE Transactions on*, Nov. 1991.

[25] C.-E. Sundberg. Erasure and error decoding for semiconductor memories. *Computers, IEEE Transactions on*, Aug. 1978.

[26] S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. F̃inan, P. Iyer, A. Singh, T. Jacob, S. J̃ain, S. Venkataraman, Y. Hoskote, and N. Borkar. An 80-tile 1.28TFLOPS network-on-chip in 65nm CMOS. In *Solid-State Circuits Conference, 2007. ISSCC 2007.*, Feb. 2007.

[27] D. Vantrease, N. Binkert, R. Schreiber, and M. Lipasti. Light speed arbitration and flow control for nanophotonic interconnects. In *Proc. of Microarchitecture (MICRO)*, Dec. 2009.

[28] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. . Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. Beausoleil, and J. Ahn. Corona: System implications of emerging nanophotonic technology. ISCA '08, June 2008.

[29] F. Wolf. Scalasca. In D. Padua, editor, *Encyclopedia of Parallel Computing*. Springer, 2011.

[30] Y. Xu, J. Yang, and R. Melhem. Channel borrowing: an energy-efficient nanophotonic crossbar architecture with light-weight arbitration. In *Proc. of the International Conference on Supercomputing*, ICS '12, New York, NY, USA, 2012. ACM.

[31] J. Yao, X. Zheng, G. Li, I. Shubin, H. Thacker, Y. Luo, K. Raj, J. Cunningham, and A. Krishnamoorthy. Grating-coupler based low-loss optical interlayer coupling. In *Group IV Photonics (GFP)*, Sept. 2011.

[32] R. Zamani and A. Afsahi. Communication characteristics of message-passing scientific and engineering applications. In *IASTED PDCS'05*.

[33] X. Zheng, J. Lexau, Y. Luo, H. Thacker, . T. Pinguet, A. Mekis, G. Li, J. Shi, . P. Amberg, N. Pinckney, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy. Ultra-low-energy all-CMOS modulator integrated with driver. *Opt. Express*, Feb. 2010.

[34] A. J. Zilkie, P. Seddighian, B. J. Bijlani, W. Qian, D. C. Lee, . S. Fathololoumi, J. Fong, R. Shafiiha, D. Feng, B. J. Luff, X. Zheng, . J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari. Power-efficient III-V/Silicon external cavity DBR lasers. *Opt. Express*, Oct. 2012.