Direct-Modulated Optical Networks for Interposer Systems

Mohammad Reza Jokar*
University of Chicago
jokar@uchicago.edu

Lunkai Zhang*†
University of Chicago
lunkai.zhang.1984@gmail.com

John M. Dallesasse
University of Illinois at Urbana–Champaign
jdallesa@illinois.edu

Frederic T. Chong
University of Chicago
chong@cs.uchicago.edu

Yanjing Li
University of Chicago
yanjingl@uchicago.edu

ABSTRACT

We present a new interposer-level optical network based on direct-modulated lasers such as vertical-cavity surface-emitting lasers (VCSELs) or transistor lasers (TLs). Our key observation is that, the physics of these lasers is such that they must transmit significantly more power (21×) than is needed by the receiver. We take advantage of this excess optical power to create a new network architecture called *Rome*, which splits optical signals using passive splitters to allow flexible bandwidth allocation among different transmitter and receiver pairs while imposing minimal power and design costs. Using multi-chip module GPUs (MCM-GPUs) as a case study, we thoroughly evaluate network power and performance, and show that (1) Rome is capable of efficiently scaling up MCM-GPUs with up to 1024 streaming multiprocessors, and (2) Rome outperforms various competing designs in terms of energy efficiency (by up to 4×) and performance (by up to 143%).

1 INTRODUCTION

Interconnects in all scales are increasingly critical in computer systems. In particular, in interposer systems, chip-to-chip interconnects are essential for achieving seamless coupling of all processing and memory elements.

In this paper, we focus on interposer-level optical networks that are based on direct-modulated optical links. Direct-modulated lasers, such as vertical-cavity surface-emitting lasers (VCSELs) [7, 13, 18] and transistor lasers (TLs) [8, 24], provide a promising path to efficient optical networks, thanks to their high energy efficiency, high data rate, and integration and manufacturing advantages (details in Sec. 2). However, existing optical network architectures have various limitations when they are applied to direct-modulated optical links. For example, the fully-connected point-to-point (FC-P2P) network, while is easy to implement and consumes low power, lacks the flexibility to share and allocate bandwidth between different channels, which can lead to large application performance impact. As another example, architectures designed for silicon photonics [15, 16, 25] achieve high performance but impose large power overhead when they are used with direct-modulated optical links (details in Sec. 3). Therefore, investigation of architectures that are optimized for direct-modulated optical links is required.

To this end, we present a new direct-modulated optical network, called *Rome*, which is tailored for direct-modulated optical links in order to meet the high bandwidth, performance, and energy efficiency requirements in interposer systems. This is achieved based on the following *key observation*: the receiver of a direct-modulated optical link requires substantially less optical power (e.g., 21×) than that generated by the transmitter. We salvage this otherwise-wasted excess optical power to achieve efficient bandwidth sharing by making the following changes to a FC-P2P design: (1) splitting each optical signal generated by a single transmitter into multiple signals that are then delivered to different photodetectors; and (2) allowing each receiver to dynamically switch between multiple photodetectors to receive signals from different transmitters. Rome works well in interposer systems because the number of modules (i.e., active chips, which are interposer network nodes) that can be integrated

*These two authors contributed equally.
†Lunkai Zhang is now at Intel.

NOCS '19, October 17–18, 2019, New York, NY, USA

ACM ISBN 978-1-4503-6700-4/19/10…$15.00
https://doi.org/10.1145/3313231.3352368

Table 1: High-Speed Link Technologies (measurement results from industry designs and research prototypes).

Link Technology	Distance	Energy Efficiency	Data Rate	Applicable to interposer?
Long-reach electrical (Intel [4, 9])	50mm to 1m	1.7 pJ/bit to >3.8 pJ/bit	>10Gbps	No
Short-reach electrical (NVIDIA [17])	$\leqslant 4.5mm$	0.54 pJ/bit	20Gbps	Yes for small scale
SiPh with WDM (Oracle [26, 27])	>50cm	1.8 pJ/bit per λ	25Gbps	Yes
VCSEL (IBM [18])	>50cm	1 pJ/bit	25Gbps	Yes
TL (UIUC [24])	>50cm	1 pJ/bit	40Gbps	Yes

Note: Energy-per-bit is for the whole link except for SiPh, which does not include RX power.

on an interposer may be limited (4 - 16; see Sec. 4), and the "free" excess optical power is sufficient to allow signals to be split and delivered to a significant portion (>50%) of all network nodes, which allows the network bandwidth to be flexibly allocated to different source/destination pairs based on traffic patterns to improve network performance. As a result, Rome retains the benefits of FC-P2P (low power and low design complexity) whiles overcomes its major limitation (low performance due to the lack of flexibility to allow bandwidth sharing).

As a case study, we apply Rome to *Multi-Chip Module GPU (MCM-GPU)* systems [2] to evaluate network performance and power. Our results show that Rome can efficiently scale MCM-GPUs with up to 1024 streaming multiprocessors (SMs), which is >16× larger than state-of-the-art commercial GPUs (e.g., NVidia Pascal) and 4× larger than existing research proposals [2]. We also quantitatively compare Rome with different interconnect technologies and architectures. Our results show that, compared to state-of-the-art electrical 2D mesh, Rome reduces network power by up to 62% (and up to 4× with more advanced laser technologies) while improving application performance by up to 143%. Compared to architectures designed for silicon photonics, Rome reduces network power by >2.5× while achieving comparable application performance.

The major contributions of this paper are:

(1) Present a new direct-modulated optical network, Rome, that leverages the novel observation that excess optical power is available to support flexible bandwidth allocation while imposing minimal costs in interposer systems.

(2) Evaluate an example of Rome as applied to a large-scale MCM-GPU system to demonstrate its efficiency and effectiveness, and quantitatively compare it to other interposer network designs. To the best of our knowledge, this is the first detailed study on optical interposer networks, with a focus on direct-modulated optical links.

The rest of the paper is organized as follows. Section 2 provides background and related work. Section 3 discusses the details of Rome. Evaluation methodology and results are presented in Sec. 4, followed by conclusions in Sec. 5.

2 BACKGROUND AND RELATED WORK

In this section, we discuss various link technologies, and present direct-modulated optical links in details.

2.1 High-Speed Link Technologies

Table 1 summarizes key technology parameters and characteristics of various high-speed link technologies. Long-reach electrical links are widely used in bridge and backplane scales, but are not practical at the interposer level due to large transceiver area [4]. Short-reach electrical links achieve high energy efficiency by limiting the distance of data transmission (e.g., to a few mm's) [17]. Therefore, a multi-hop topology (e.g., mesh) is the only viable option, making it inefficient to achieve high-bandwidth and low-latency global communications for large-scale systems.

There are several advantages of optical links over electrical ones, including higher energy efficiency, longer transmission distance, and better bandwidth scalability. Within optical link technologies, *silicon photonics (SiPh)* which supports *wavelength-division multiplexing (WDM)* has been shown to be efficient in chip-to-chip or even on-chip networks [15, 16, 25]. A unique advantage of SiPh links is that multiple wavelengths (e.g., up to 64 in dense WDM) can be carried by a single waveguide. This allows a SiPh network to achieve high bandwidth without the need for dense waveguides. A disadvantage of SiPh is the low energy efficiency of the external lasers. For example, in Table 1, the off-chip laser alone consumes a major portion of the transmitter energy (1.5pJ/bit out of the total 1.8pJ/bit). Another challenge is that SiPh requires separate modulators (ring resonators) to modulate the always-on signals. The optical properties of ring resonators are highly sensitive to waveguide width, refractive index, modal index, sidewall roughness, and temperature [11].

In this paper, we focus on another promising optical technology: direct-modulated optical links based on VCSELs or TLs. VCSEL-based optical networks are already widely adopted in datacenters. Their efficiency for chip-to-chip or on-chip communications has also been demonstrated [13, 18]. TL [8, 24] is an InGaP/GaAs heterojunction bipolar transistor with the addition of quantum wells for photon generation and optical cavity for coherent light output. Measurement data in actual research prototypes demonstrate that TLs are on par with or superior to state-of-the-art VCSELs (Table 1), but TLs are expected to be more scalable than VCSELs as both technologies scale: the power consumption of TLs and VCSELs is similar (due to similar threshold current and voltage which are tied to material properties), but TLs can potentially reach much higher speed and energy efficiency.

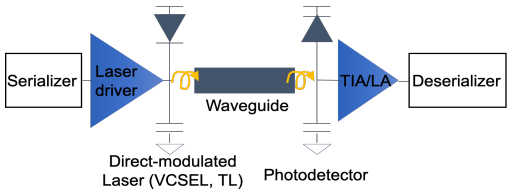


Figure 1: A Direct-Modulated Optical Link.

This is because the modulation bandwidth of TLs is limited by tunneling absorption latency (10fs), which corresponds to a data rate of up to 840Gbps and energy efficiency of up to 20fJ/bit [8, 24]. In contrast, the modulation bandwidth of VCSELs is fundamentally limited by the carrier recombination lifetime (200ps). Therefore, the maximum projected VCSEL data rate and energy efficiency are limited to 160Gbps and 100fJ/bit, respectively [7].

In general, the advantages of direct-modulated optical networks include: (1) Direct-modulated lasers are typically more energy efficient than always-on lasers. (2) Although VCSELs and TLs can be subject to thermal variations, they can still operate reliably because laser drivers can dynamically adjust the input power strength based on thermal sensor information [18]. (3) Separate modulators (ring resonators) are not needed, which reduces network power and design complexity. (4) VCSEL and TL links are ultra high-speed and highly energy-efficient as shown in Table 1. (5) Both VCSELs and TLs can be integrated with silicon [18, 24]. A disadvantage of VCSEL/TL links is that it is not clear how they can efficiently support WDM. However, we will show in Sec. 4 that there is sufficient space in the interposer to fit all waveguides even without WDM. Therefore, we deem VCSEL/TL links a good fit for interposer-level networks.

2.2 Direct-modulated Optical Links

Figure 1 depicts a direct-modulated optical link. On the transmitter side, after serialization, the laser (VCSEL or TL) is modulated by a driver, and generates an optical signal that travels through a waveguide. On the receiver side, the photodetector (PD) converts the optical signal back to electrical, which then passes through the receiver (consisting of TIA, LA, etc.) and deserializer. Complex clock recovery circuits are not likely to be needed at the interposer level [14].

To perform optical switching (so that one source node can reach multiple destination nodes using one output channel), either passive power splitters (which can be used to split one optical signal into multiple ones, either evenly or with different weights) [21] or ring resonators may be used. We opt to perform optical switching using splitters because: (1) they do not impose additional power cost and design complexity; (2) they are easy to manufacture; (3) they function reliably and incur low optical loss (e.g., 0.35dB loss per splitter based on actual measurement data [21]); and (4) given excess optical power (Sec. 2.3), using splitters alone already enables high flexibility and performance in interposer networks. Ring resonators may improve performance (see Sec. 4), but the power and design costs are expected to be higher.

Table 2: Excess Optical Power in VCSEL/TL Links.

	OMA	receiver sensitivity	TX/RX power ratio
state-of-the-art [18]	0.63mW (avg)	0.03mW [20]	21×
projection [7]	5mW	0.12mW	41.5×

2.3 Excess Optical Power in Direct-Modulated Optical Links

The most important and unique observation, which serves as the basis and motivation of Rome, is that the optical power generated by a VCSEL or TL (as measured by *optical modulation amplitude*, or *OMA*) is much higher than what is required by the receiver. The reason is fundamental to optical device physics: it is well known that the data rate increases as the OMA increases, with a square root dependence. Therefore, to support high-speed laser operations, a certain high OMA is required (e.g., at least 0.5mW even with aggressive device optimizations [12]). State-of-the-art low-power and high-speed VCSEL and TL prototypes achieve an average OMA of >0.63mW [18, 24]. In the future, a combination of lower threshold current together with improved conversion efficiency will lead to higher OMA (5mW) even at high temperatures of >200°C (based on simulation results for future VCSELs [7]; similar for future TLs). On the receiver side, low *receiver sensitivity* (i.e., the optical power required for PDs to sense the signals reliably) of 0.03mW has already been demonstrated [20].

As summarized in Table 2, the ratio between OMA and receiver sensitivity is high: > 21×. Note that, the 0.63mW OMA reported in [18] is the average OMA, which is much lower than the max OMA corresponding to the high-state signal. Thus, the 21× ratio is *pessimistic*. In the future, with significant improvement on OMA, even if we assume that receiver sensitivity increases linearly as data rate, the ratio would still be very high.

In essence, due to fundamental physics, it is *not* possible to reduce laser OMA to match receiver sensitivity, so excess optical power is available and abundant in direct-modulated optical links. Such "free" optical power enables the Rome architecture, which provides flexible bandwidth sharing and allocation with minimal overhead.

3 THE ROME NETWORK

3.1 Limitations of Existing Optical Network Architectures

Optical networks favor *single-hop architectures* because optical links are known to be efficient for long-distance communications. Also, single-hop architectures scale better to large

systems, which requires large bandwidth between *any* pair of nodes. Thus, we focus our study on single-hop architectures.

FC-P2P is a simple single-hop architecture that achieves full connectivity between any node pairs. Figure 2(a) shows a two-node example of FC-P2P. A major limitation of FC-P2P is the lack of flexibility to share and allocate bandwidth. For example, suppose there is a large amount of traffic between SRC #2 and DES #1, creating a hot spot in the link between the two nodes. In this case, even if the link between SRC #2 and DES #2 is idle, it cannot be utilized to alleviate the hot spot. Such inflexibility can considerably limit network performance (results in Sec. 4).

We also consider existing single-hop architectures designed for SiPh links, where ring or serpentine-based links are commonly used to connect to all nodes. In such a network topology, global communications can be achieved using various schemes: single-writer, multiple-reader (SWMR), single-reader, multiple-writer (SRMW), and multiple-reader, multiple-writer (MRMW) [15, 16, 25]. If we apply the SWMR scheme [15] to direct-modulated optical links, then each VCSEL/TL link would connect to all network nodes, and the node containing the laser is the writer while all other nodes are readers. There are several limitations of this approach: (1) Ring resonators may be needed to direct optical signals to all destination nodes. If passive splitters are used instead, it would mean that each link must support broadcast operation all the time. Broadcasting requires each VCSEL/TL to output high optical power, which substantially increases power consumption. (2) Arbitration among the readers imposes high control overhead [15]. (3) To achieve the same peak bandwidth as FC-P2P, SWMR require a larger number of receivers: in FC-P2P, each link is connected to one transmitter and one receiver; In SWMR, $n-1$ receivers (where n is the number of network nodes) are needed instead, because each link has a dedicated receiver in each destination node.

The SRMW scheme [25] suffers from a similar set of limitations as SWMR (the only difference is that, instead of requiring a larger number of receivers as in SWMR, it requires a larger number of transmitters compared to FC-P2P, under the same peak bandwidth requirement). MRMW [16] essentially "combines" SWMR and SRMW, and it follows that the combined limitations of SWMR and SRMW apply.

3.2 The Key Idea of Rome

To overcome the limitations of existing architectures, we create the Rome network[1]. Rome is based on FC-P2P, but it takes advantage of the key observation that excess optical power is available (Sec. 2.3) to overcome FC-P2P's inflexibility limitation. In Rome, splitters are used to split each optical signal in FC-P2P into multiple signals connecting to multiple destination nodes, while ensuring that each split signal is still strong enough to be reliably received. Moreover, Rome allows each receiver to dynamically switch among multiple PDs. As shown in Fig. 2(b), each receiver in a destination node is connected to multiple PDs, and each PD is connected to a fast BJT (bipolar junction transistor) switch that is turned on/off dynamically based on information from Rome's control unit. At any given time, only one switch is connected, forming a link from the corresponding source node to the receiver's home node. This way, unlike the SWMR design, Rome does not require a large number of receivers.

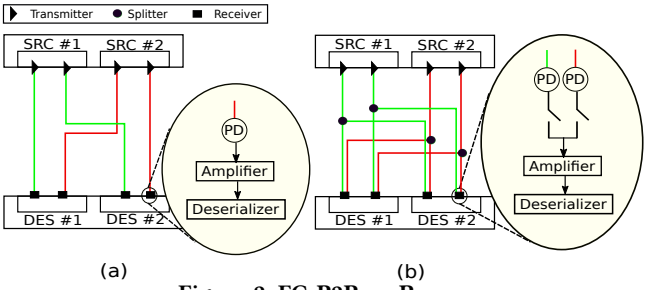


Figure 2: FC-P2P vs. Rome.

To see how Rome enables flexible bandwidth sharing and allocation, let's revisit the hot spot example in FC-P2P. As shown in Fig. 2(b), Rome allows DES #1 to dynamically reconfigure the PD switches in both CH#1 & CH#2 to receive data from SRC #2, effectively alleviating the hot spot, because now a single channel may be shared by multiple receivers in different destination nodes. Moreover, different from the SWMR design where the receiver of a channel can be any network node, in Rome there is a limit on the *splitting degree*, i.e., the number of destination nodes sharing the same data channel, which is constrained by laser OMA, receiver sensitivity, and various sources of optical losses throughout the whole link. Since there is no need to increase laser OMA or use ring resonators to perform broadcast, power overhead and design complexity is minimal. Rome therefore can be thought of a "sweet-spot" between FC-P2P and SWMR, which is capable of achieving balanced network cost and flexibility tradeoffs by varying the splitting degree.

3.3 Physical Implementation of Rome

We describe the physical implementation of Rome in the context of MCM-GPU on a passive optical interposer (containing waveguides, splitters, and couplers); but the general idea is applicable to other systems. As an example, Fig. 3 depicts a 9-module MCM-GPU laid out as a 2D matrix, with a splitting degree of 2. Each GPU module contains a local memory partition. All memory partitions as well as I/O interfaces are accessible by all GPU modules through the Rome network, so the entire interposer system behaves as a large monolithic GPU (similar to [2]). The network is partitioned into two

[1]The name comes from "all roads lead to Rome", because the network enables multiple routes to reach the same destination.

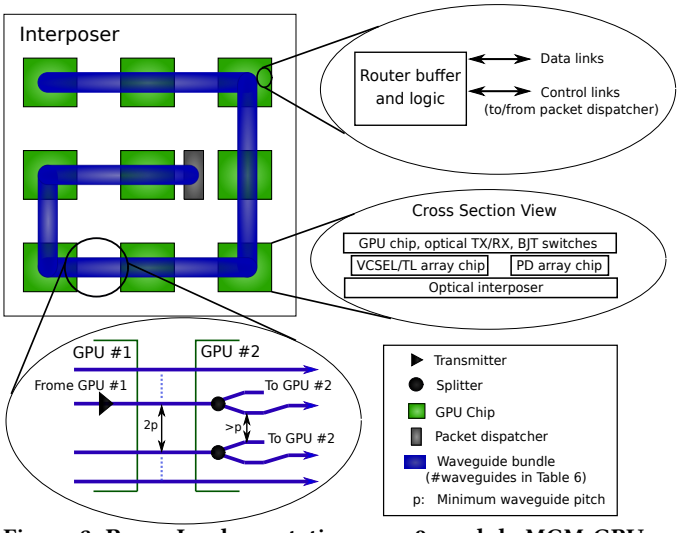


Figure 3: Rome Implementation on a 9-module MCM-GPU (memory and I/O interfaces are not shown).

subnets: one for traffic from SMs to remote memory partitions, and the other in the reverse direction. TL and PD array chips are connected to the transceivers in each GPU module using hybrid integration techniques through TSVs (similar to [13]). Light signals produced by VCSELs/TLs are vertically emitted through the bottom of the chip and coupled into the waveguides. Input light signals are coupled into the PDs that are connected to BJT switches. BJT switches receive control signals from the router, and their outputs are OR'ed together to be used as the input to the optical receivers. Waveguides are routed in a long, serpentine-like manner to minimize optical losses due to waveguide crossings (similar to [10, 25]). The width between two neighboring waveguides is 2× the minimal waveguide pitch to allow space for splitting. The power ratio of the two split signals is variable, such that the power of the signal that is fed to the local GPU module is just sufficient to meet the receiver's sensitivity requirement, which can be estimated statically during the design phase.

3.4 Routing, Arbitration, and Flow Control

Routing is straightforward in Rome, because there is only one route for every packet (even though the route is split to reach multiple destinations). However, we need to correctly control the BJT switches in each destination node to avoid conflicts from multiple source nodes. There are two options to implement this arbitration: centralized and distributed. In the distributed approach, the router in each destination node is responsible for making arbitration decisions, and it requires at least one bi-directional control link between every pair of nodes. The number of control links is larger than that for a centralized approach in most cases, because the number of nodes in an interposer system is limited: a 16-node interposer system is already a very aggressive target (details in Sec. 4). Therefore, although a centralized approach is generally considered to be less scalable, it is adequate in the interposer scale, which leads to lower design complexity.

The centralized control unit in Rome, called the *packet dispatcher*, is connected to all network nodes using VCSEL/TL links. It is responsible for packet scheduling, PD switch control, and credit-based flow control. Under the coordination of the packet dispatcher and routers in network nodes, a network packet's life cycle in Rome consists of four stages:

1. Packet Combining. To amortize scheduling overhead and make efficient use of the high data rate and bandwidth in the optical links, each source router tries to pack multiple network requests with the same destination into one large network packet whenever possible.

2. Packet Registration. The source router registers a network packet by sending a "Registration Request" message to the packet dispatcher (a description of this message is shown in Table 3), when either of the following condition is met: (1) The size of the packet exceeds a *size threshold*; (2) the network packet has been waiting for a time period longer than a *time threshold*. If multiple packets fulfill one of these conditions at the same time, the one with the longest wait time takes priority. These criteria allow Rome to adapt to different traffic patterns while preventing starvation and minimizing scheduling overhead and performance impact. A different policy may be adopted if the network is imposed with other fairness or data criticality requirements.

The control information of each registered packet is stored in a FIFO in the packet dispatcher. The number of registered packets for one (source, destination) pair cannot exceed *Pair_Token_Limit*, which is constrained by the FIFO size.

3. Packet Scheduling. In every clock cycle, for every (source, destination) pair, the packet dispatcher first examines the oldest network packet for which at least one channel that can transfer the packet is unoccupied. Among these "transferable" packets, some of them may share the same source or destination node, so the packet dispatcher avoids any conflict by selecting at most one packet for each (source, destination) pair according to the first-come, first-serve policy. Finally, for each selected packet, the packet dispatcher issues a "Scheduling Request" control message (format specified in Table 3), which contain all the information needed to guarantee correct transmission, to the corresponding source and destination nodes.

4. Packet Transfer. Once a (source, destination) pair receives "Scheduling Request", the destination router decodes the message and generates appropriate control signals to the BJT switches. The source router transmits the packet at the time slot given by the packet dispatcher. The packet is guaranteed to be transmitted successfully at the scheduled time slot, since the packet dispatcher takes into account the latency for control/data communication and BJT/PD switching (plus some guardbands to eliminate synchronization issues).

Table 3: Control Messages in Rome.

Message Type	Fields
Registration Request	valid, DST_ID, packet_id, packet_size
Scheduling Request to SRC	valid, packet_id, delay, channel_id
Scheduling Request to DST	valid, packet_size, delay, channel_id
Almost-full	valid, 1-bit almost-full signal

Table 4: MCM-GPU System Configurations (parameters are derived based on [2]).

configuration	interposer area	GPU module area	network peak bandwidth requirement	projection
4x64	$1500mm^2$	$256mm^2$	2.25TB	7nm-node
16x16	$1500mm^2$	$64mm^2$	2.82TB	7nm-node
9x64	$1500mm^2$	$144mm^2$	6TB	near-future
16x64	$3000mm^2$	$144mm^2$	11.25TB	far-future

To support credit-based flow control, a node will send an "Almost-Full" message to the packet dispatcher when its input buffer size drops below a threshold. To ensure that all types of control messages (as summarized in Table 3) can be sent simultaneously, 4 control links are dedicated to each node in each subnet. The latency to transfer one control message is <= 1 cycle for a 1GHz clock and 25Gbps optical links, if all control messages are <= 25 bits.

4 EVALUATION

We perform thorough evaluation for four *n*x*m* MCM-GPU system configurations (where *n* is the number of GPU modules, and *m* is the number of SMs per module) as summarized in Table 4 to demonstrate the efficiency of Rome.

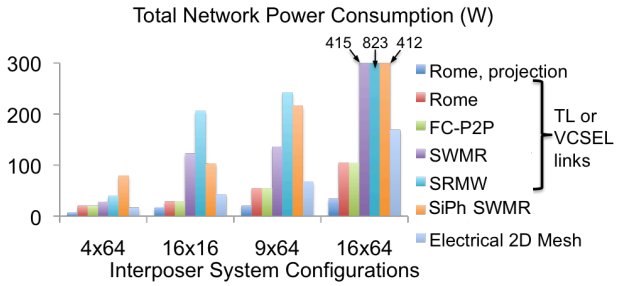
4.1 Power, Area, and Splitting Degree

Based on the area constraints and peak bandwidth requirements in each interposer configuration in Table 4, as well as optical link technology parameters summarized in Table 5, we calculate (1) the area required to hold all optical network resources (including waveguides, TLs, and transceivers), (2) network power, and (3) the max splitting degree that can be achieved. The results are shown in Table 6.

All required network resources fit under the interposer/GPU module area constraints for the near-term configurations. For the aggressive 16x64 configuration, we expect that with more advanced waveguide techniques [23], the width of the waveguide bundle would be reduced by >2×, which fits under the interposer area constraint.

The power of Rome presented in Table 6 is calculated by summing up the power consumption of the following network components: transmitters, receivers, serializer/deserializer (SerDes), BJT switches, and the packet dispatcher. Rome's network power is already small with today's technology, and it is expected to improve further as VCSELs and TLs scale.

In all configurations, the max splitting degree is 8 for OMA=0.63W and receiver sensitivity=0.03mW (after accounting for optical losses including coupling, propagation, bending, and splitter losses). The splitting degree is the same for all configurations because propagation loss of Si_3N_4 waveguides is very small [2]. The projected max splitting degree will be even higher for future VCSELs/TLs based on Table 2.



Note: All networks have the same peak bandwidth for the same interposer configuration. Rome power is taken from Table 6. TL/VCSEL-based SWMR and SRMW results include data/control link power and RX/TX power. SiPh SWMR and electrical mesh power only includes total data link power based on Table 1 data.

Figure 4: Power Comparison of Various Networks.

For Rome, we compare the power consumption of a centralized control approach (using the packet dispatcher as discussed in Sec. 3) with a distributed approach (assuming minimally that 1 control link is allocated per data channel). The centralized approach consumes less power in the 16x16, 9x64, and 16x64 configurations, and is comparable (consumes 0.2W more power) in the 4x64 configuration. This result confirms that a centralized approach is more efficient in Rome.

Figure 4 compares the power consumption of various interposer network technologies and architectures, under the same peak bandwidth constraint. Rome consumes very little additional power (due to package dispatcher overhead) compared to FC-P2P. TL or VCSEL-based SRMW or SWMR schemes consume much higher power than both Rome and FC-P2P because the number of control links and receivers (for SWMR) or transmitters (SRMW) are much higher. Rome outperforms SiPh SWMR by >2.5× for the same reason, and also because Rome's link-level energy efficiency is higher (1pJ/bit vs. 1.8pJ/bit). Electrical 2D mesh consumes the least power in the 4x64 configuration given today's technology because the energy efficiency of the short-reach electrical links is very high (0.54pJ/bit). However, Rome achieves lower (up to 62%) power consumption in all other configurations. Moreover, as TLs/VCSELs scale, which improves their energy efficiency and data rate, the power of Rome reduces further, leading to 1.4× - 3.9× power reduction vs. electrical 2D mesh across all configurations. Note that, in contrast to the scalable TL/VCSEL links, further energy efficiency scaling for electrical links is generally considered difficult, if not impossible, at the interposer scale [14, 17].

[2]We use Si_3N_4 waveguides instead of Si waveguides commonly used in SiPh architectures because they support a wider range of wavelengths, including 980nm or 850nm which are common VCSEL/TL wavelengths that are transparent to Si. Their propagation loss is much lower, but they must sustain a wide pitch than silicon waveguides due to crosstalk issues.

Table 5: Direct-Modulated Link Technology Parameters (based on industry/research prototype results).

data rate	link energy (no SerDes)	SerDes power	link power	Average VCSEL OMA	Receiver sensitivity	TX area & power	RX area & power
25Gbps [18]	1pJ/b[18]	1.3mW [17]	26.3mW*	0.63mW [18]	0.03mW [20]	2128um^2, 13.4mW [18]	1800um^2, 5.3mW per PD [20]

Si_3N_4 waveguide pitch	waveguide layers	waveguide propagation loss	waveguide coupling loss	splitter loss	waveguide bending loss	projected data rate & link power
5um minimum [19, 21]	2 [19]	0.001dB/cm [5]	3dB [22]	0.35dB [21]	0.009dB [1]	100Gbps, 35.2mW** [7]

* link power = data rate × link energy (no SerDes) + SerDes power. ** Assume power of TX, RX, and SerDes scales at the same rate as laser.

Table 6: Rome Network Results.

configuration	number of links (data + control)	total power: network + packet dispatcher	waveguide-bundle width (must be <module height h)	total TX+RX area mm^2 / % of interposer	max splitting degree	projected total power [7]	projected max splitting degree
4x64	744 + 32	20.4W + 0.03W	3.88mm, <h = 16mm	7.1mm^2 / 0.48%	8	6.8W + 0.03W	11
16x16	960 + 128	28.6W + 0.46W	5.44mm, <h = 8mm	16.4mm^2 / 1.1%	8	16.9W + 0.46W	11
9x64	2016 + 72	54.9W + 0.14W	10.44mm, <h = 12mm	33.6mm^2 / 2.3%	8	20.3W + 0.14W	11
16x64	3840 + 128	104.4W + 0.46W	10.1mm (projected), <h = 12mm	64.0mm^2 / 2.2%	8	33.8W + 0.46W	11

Notes: 1. Results are calculated based on parameters in Tables 4 and 5.
2. Packet dispatcher power is obtained from RTL synthesis using Synopsys Design Compiler with a 40nm library, for Pair_Token_Limit=8.

Table 7: Simulation Parameters.

GPGPU Parameters:

frequency	1GHz	L1 cache (per SM) / L2 cache	64KB / 32MB
Max #warps per SM	64	DRAM bandwidth per SM	16GB/s

Rome Parameters (based on physical requirements and sensitivity analysis):

Pair_Token_Limit	8	BJT switch time	1 cycle
TX/RX latency	4ns	link latency	1ns
size threshold	64B	lifetime threshold	4ns

4.2 Performance Evaluation

We conduct detailed architectural simulation using GPGPU-sim [3] to evaluate application performance (GPGPU-sim is adopted for interposer simulation by grouping all SMs that belong to the same GPU module into a single "cluster"). We simulate Rome with different splitting degrees (sd). When sd=1, Rome is equivalent to FC-P2P. An *Ideal* network is Rome with no control overhead and with sd=the total number of nodes. We also simulate the SiPh SWMR and electrical mesh networks for comparison purposes. Table 7 summarizes the simulation parameters.

Similar to previous MCM-GPU studies [2], we use memory-intensive applications (i.e., applications whose memory access per kilo instructions is >40 bytes) from *Rodinia* with the native input set [6]. We are only able to report results obtained from `kmeans` and `pathfinder` for the 16x64 configuration because all other applications exhibit limited parallelism. However, the characteristics of these two applications are representative of various parallel applications, and our results still provide meaningful insights.

Figure 5 shows the instruction per cycle (IPC) results normalized to FC-P2P for each interposer configuration. In all cases, Rome outperforms the electrical mesh (by 20%-143% on average across different interposer configurations) as expected due to the multi-hop nature in the mesh, and the benefit of Rome increases as the number of network nodes (i.e., GPU modules) increases. The performance of SiPh is similar to Rome with sd=the number of nodes in the network (within 5% of each other), which is also expected because

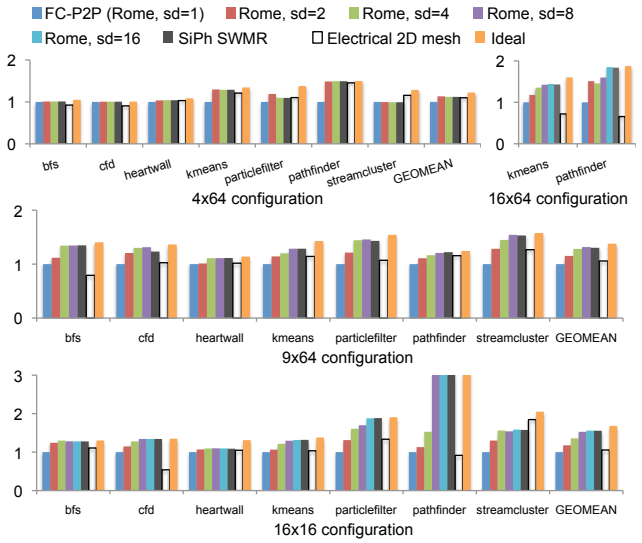


Figure 5: Normalized IPC vs. FC-P2P of Various Interposer Networks.

both networks achieve the same bandwidth and the same degree of flexibility in bandwidth allocation.

Comparing Rome with different sd's, for the 4x64 configuration, sd=2 yields the best performance, improving IPC by 13% on average and up to 49% in `pathfinder` vs. FC-P2P. sd=4 actually results in lower IPC than sd=2 because of higher control overhead. For the 16x16 and 9x64 configurations, application performance increases as sd increases. Even though the network can benefit from a higher sd, sd=8 (the max sd achievable today) already fills 77% and 83% of the gap between FC-P2P and the Ideal network for the 16x16 and 9x64 configurations, respectively (in 16x16, sd=16 fills 82% of this gap, which is only slightly higher than sd=8). The benefit of Rome can be similarly observed for the aggressive 16x64 configuration. sd=8 for `kmeans` achieves very similar performance to sd=16, while `pathfinder` can benefit further from sd=16. However, sd=8 still improves `pathfinder`'s IPC by 60% compared to FC-P2P, which fills 68% of the gap between FC-P2P and the Ideal network. If ring resonators are used in

the place of splitters to perform signal switching, Rome is expected to achieve the performance of that provided by setting sd=number of network nodes, at the price of additional power/design overhead to incorporate ring resonators, as well as additional scheduling and control overhead.

5 CONCLUSIONS

High-performance and energy-efficient networks are essential for enabling future large-scale interposer systems. In this paper, we present Rome, a new network tailored for direct-modulated optical links in the interposer scale. Rome utilizes the key observation that the optical power generated by a laser (TL or VCSEL) is substantially higher than what is needed by the receiver to allow flexible bandwidth allocation by splitting the optical signal (without increasing power consumption or design complexity), which leads to high flexibility and performance with minimal costs.

Our detailed comparison of various architectures and link technologies reveals broad perspectives. While electrical networks can be efficient in small-scale systems, they do not scale beyond a few nodes at the interposer level. We quantitatively demonstrate the energy efficiency benefit of optical networks over electrical ones, and this benefit is expected to increase as optical technologies improve, in contrast to electrical links for which further scaling is difficult.

6 ACKNOWLEDGEMENTS

We thank Prof. Milton Feng, Prof. Lynford Goddard, and Patrick Su of UIUC for their comments. This work is sponsored in part by NSF award number 1640192 and E2CDA-NRI, a funded center of NRI, a Semiconductor Research Corporation (SRC) program sponsored by NERC and NIST.

REFERENCES

[1] A. Arbabi, Y. M. Kang, C.-Y. Lu, E. Chow, and L. L. Goddard. Realization of a narrowband single wavelength microring mirror. *Applied Physics Letters*, 99(9):091105, 2011.

[2] A. Arunkumar et al. Mcm-gpu: Multi-chip-module gpus for continued performance scalability. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ISCA '17, pages 320–332, New York, NY, USA, 2017. ACM.

[3] A. Bakhoda et al. Analyzing cuda workloads using a detailed gpu simulator. In *2009 IEEE International Symposium on Performance Analysis of Systems and Software*, pages 163–174, April 2009.

[4] G. Balamurugan et al. A 5-to-25gb/s 1.6-to-3.8mw/(gb/s) reconfigurable transceiver in 45nm cmos. In *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 372–373, Feb 2010.

[5] J. F. Bauters et al. Planar waveguides with less than 0.1 db/m propagation loss fabricated with wafer bonding. *Opt. Express*, 19(24):24090–24101, Nov 2011.

[6] S. Che et al. Rodinia: A benchmark suite for heterogeneous computing. In *2009 IEEE International Symposium on Workload Characterization (IISWC)*, pages 44–54, Oct 2009.

[7] D. G. Deppe et al. Advanced vcsel technology: Self-heating and intrinsic modulation response. *IEEE Journal of Quantum Electronics*, 54(3):1–9, June 2018.

[8] M. Feng et al. Tunneling modulation of transistor lasers: Theory and experiment. *IEEE Journal of Quantum Electronics*, 54(2):1–14, April 2018.

[9] J. Han et al. 6.2 a 60gb/s 288mw nrz transceiver with adaptive equalization and baud-rate clock and data recovery in 65nm cmos technology. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 112–113, Feb 2017.

[10] R. Ho et al. Silicon photonic interconnects for large-scale computer systems. *IEEE Micro*, 33(1):68–78, Jan 2013.

[11] A. V. Krishnamoorthy et al. Exploiting cmos manufacturing to reduce tuning requirements for resonant optical devices. *IEEE Photonics Journal*, 3(3):567–579, June 2011.

[12] D. M. Kuchta, A. V. Rylyakov, F. E. Doany, C. L. Schow, J. E. Proesel, C. W. Baks, P. Westbergh, J. S. Gustavsson, and A. Larsson. A 71-gb/s nrz modulated 850-nm vcsel-based optical link. *IEEE Photonics Technology Letters*, 27(6):577–580, March 2015.

[13] H. Li et al. Thermal aware design method for vcsel-based on-chip optical interconnect. In *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, pages 1120–1125, March 2015.

[14] D. A. B. Miller. Attojoule optoelectronics for low-energy information processing and communications. *Journal of Lightwave Technology*, 35(3):346–396, Feb 2017.

[15] Y. Pan et al. Firefly: Illuminating future network-on-chip with nanophotonics. *SIGARCH Comput. Archit. News*, 37(3):429–440, June 2009.

[16] Y. Pan, J. Kim, and G. Memik. Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar. In *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*, pages 1–12. IEEE, 2010.

[17] J. W. Poulton et al. A 0.54 pj/b 20 gb/s ground-referenced single-ended short-reach serial link in 28 nm cmos for advanced packaging applications. *IEEE Journal of Solid-State Circuits*, pages 3206–3218, 2013.

[18] J. E. Proesel et al. 35-gb/s vcsel-based optical link using 32-nm soi cmos circuits. In *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, pages 1–3, March 2013.

[19] W. D. Sacher et al. Multilayer silicon nitride-on-silicon integrated photonic platform for 3d photonic circuits. In *2016 Conference on Lasers and Electro-Optics (CLEO)*, pages 1–2, June 2016.

[20] S. Saeedi et al. A 25 gb/s 3d-integrated cmos/silicon-photonic receiver for low-power high-sensitivity optical communication. *Journal of Lightwave Technology*, 34(12):2924–2933, June 2016.

[21] Y. Sakamaki et al. Low-loss y-branch waveguides designed by wavefront matching method and their application to a compact 1 x 32 splitter. *Electronics Letters*, 43(4):217–219, February 2007.

[22] Y. Shani et al. Efficient coupling of a semiconductor laser to an optical fiber by means of a tapered waveguide on silicon. *Applied Physics Letters*, 55(23):2389–2391, 1989.

[23] W. Song et al. High-density waveguide superlattices with low crosstalk. *Nature Communications*, 6(7027), March 2015.

[24] H. W. Then et al. The transistor laser: Theory and experiment. *Proceedings of the IEEE*, 101(10):2271–2298, Oct 2013.

[25] D. Vantrease et al. Corona: System implications of emerging nanophotonic technology. *SIGARCH Comput. Archit. News*, 36(3):153–164, June 2008.

[26] X. Zheng et al. A 33mw 100gbps cmos silicon photonic wdm transmitter using off-chip laser sources. In *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013*, page PDP5C.9, 2013.

[27] X. Zheng et al. Efficient wdm laser sources towards terabyte/s silicon photonic interconnects. *Journal of Lightwave Technology*, 31(24):4142–4154, 2013.