NORTHWESTERN
UNIVERSITY

Electrical Engineering and Computer Science Department

Technical Report
Number: NU-EECS-13-08

July, 2013

Galaxy: A High-Performance Energy-Efficient Multi-Chip Architecture Using Photonic Interconnects

Y. Demir, Y. Pan, S. Song, N. Hardavellas, J. Kim, and G. Memik

Abstract

The scalability trends of modern semiconductor technology lead to increasingly dense multicore chips. Unfortunately, physical limitations in area, power, off-chip bandwidth, and yield constrain single-chip designs to a relatively small number of cores, beyond which scaling becomes impractical. Multi-chip designs overcome these constraints, and can reach scales impossible to realize with conventional single-chip architectures. However, to deliver commensurate performance, multi-chip architectures require a cross-chip interconnect with bandwidth, latency, and energy consumption well beyond the reach of electrical signaling. We propose Galaxy, an architecture that enables the construction of a many-core "virtual chip" by connecting multiple smaller chiplets through optical fibers. The low optical loss of fibers allows the flexible placement of chiplets, and offers simpler packaging, power, and heat requirements. At the same time, the low latency and high bandwidth density of optical signaling maintain the tight coupling of cores, allowing the virtual chip to match the performance of a single chip that is not subject to area, power, and bandwidth limitations. Our results indicate that Galaxy attains speedup of 2.2x over the best realistic single-chip alternatives with electrical or photonic interconnects (3.4x maximum), and 2.6x smaller energy-delay product (6.8x maximum). We show that Galaxy scales to 4K cores and attains 2.5x speedup at 6x lower laser power compared to Oracle Macrochip.

Keywords

Galaxy: A High-Performance Energy-Efficient Multi-Chip Architecture Using Photonic Interconnects

Yigit Demir[†], Yan Pan[¥], Sukwoo Song[‡], Nikos Hardavellas[†], John Kim[‡], and Gokhan Memik[†]

[†]Northwestern University
Dept. of Electrical Eng. and Computer Science
Evanston, IL, USA
yigit@u.northwestern.edu
{nikos, g-memik}@northwestern.edu

[¥]Globalfoundries Inc.
Malta, NY, USA
panyan@gmail.com

[‡]KAIST
Dept. of Computer Science
Daejeon, Korea
jjk12@kaist.edu
sukwoo24@gmail.com

ABSTRACT

The scalability trends of modern semiconductor technology lead to increasingly dense multicore chips. Unfortunately, physical limitations in area, power, off-chip bandwidth, and yield constrain single-chip designs to a relatively small number of cores, beyond which scaling becomes impractical. Multi-chip designs overcome these constraints, and can reach scales impossible to realize with conventional single-chip architectures. However, to deliver commensurate performance, multi-chip architectures require a cross-chip interconnect with bandwidth, latency, and energy consumption well beyond the reach of electrical signaling.

We propose Galaxy, an architecture that enables the construction of a many-core "virtual chip" by connecting multiple smaller chiplets through optical fibers. The low optical loss of fibers allows the flexible placement of chiplets, and offers simpler packaging, power, and heat requirements. At the same time, the low latency and high bandwidth density of optical signaling maintain the tight coupling of cores, allowing the virtual chip to match the performance of a single chip that is not subject to area, power, and bandwidth limitations. Our results indicate that Galaxy attains speedup of 2.2x over the best realistic single-chip alternatives with electrical or photonic interconnects (3.4x maximum), and 2.6x smaller energy-delay product (6.8x maximum). We show that Galaxy scales to 4K cores and attains 2.5x speedup at 6x lower laser power compared to a Macrochip with silicon waveguides.

1. INTRODUCTION

Advanced silicon fabrication allows for exponentially increasing transistor counts, pointing to increasingly dense multicore chips. However, physical limitations in area, yield, off-chip bandwidth, and power limit the scalability of single chip designs. Area and yield considerations push for small die sizes, and the latest ITRS models reflect the competitive requirements for affordability by targeting flat chip-size trends for both high-performance and cost-performance processors (lowered to 260 mm^2 and 140 mm^2 respectively [10]). At the same time, while transistor counts grow exponentially, voltage scaling has slowed. This has lead to a dramatic increase in power density with decreasing feature size [13], creating chips that require a power budget beyond what is practical today to operate and leading to "dark silicon" [11,9,19]. Moreover, the limited pin count and low efficiency in off-chip communication severely limit the off-chip bandwidth [25], rendering it increasingly difficult to feed all cores with data fast enough to keep them busy. This bandwidth wall hampers the scalability of future CMPs and their performance, even for highly-parallel workloads [11].

As a result, multicore scalability is being rapidly pushed to an end. Physical constraints limit single chip designs to either a relatively small number of cores, beyond which scaling becomes impractical, or to designs that trade single-core performance for high aggregate instruction throughput, which can only be achieved if all cores are simultaneously employed by the executing workload. For example, a single core in Intel i7-3960X has a peak theoretical performance of 187 *GFLOPS*, but only 6 such cores fit in the chip's area and power budget. In contrast, Intel Phi 5110P features 60 cores, but at only 17 *GFLOPS* per core, and NVIDIA GTX-680 features 1536 CUDA cores but at a paltry 2 *GFLOPS* each.

Alternative designs can break free of some physical limitations, but not all. Aggregating together several discrete smaller dies instead of having a large one (*disintegration*) overcomes the area and yield limitations [7], as only few dies need to be replaced if they are faulty [3,7]. At the same time the total silicon area of the aggregate chip can scale beyond reticle size limits, allowing the aggregate chip to reach scales impossible to realize with a monolithic design (*macrochip integration*). 3D-die stacking can realize these benefits by vertically connecting several smaller dies in a package with through-silicon-vias (TSVs). However, 3D-die stacking incurs significant challenges in power delivery and heat removal, and is best employed when the additional dies implement low-power applications (e.g., DRAM). By contrast, high-power applications (e.g., high-performance processors), are ideally spread out as an array of chips, allowing for power delivery to and heat removal from each individual die directly. Unfortunately, connecting a large array of chips at high bandwidth presents unique challenges.

Limitations in the density of chip I/O and package routes dramatically constrain the number of links that can be routed across chips, and severely constrain bandwidth. A 580 mm^2 die can have 25600 pins to the package substrate at a pitch of 150 μm, but the substrate-to-board pitch is 0.8 mm which allows only 3844 pins to the board from a 5 cm x 5 cm package [10]. This forces the use of over-clocked and high power serial links for chip-to-chip communication. Thus, using SerDes links [24] on an FR-4 board incurs significant energy

consumption or long delays (20 *pJ/bit* typically, and at best 2.5 *pJ/bit* and 2.5 *ns* latency over 4 inches of electrical strip [24]) as the designers have to trade energy for performance or vise-versa. Silicon interposers (i.e., 2.5D integration) allow chips to connect laterally within the same package through "bridge" silicon chips, thus exploiting the high density of die-to-package and on-chip wires. However, this enables only modest-sized arrays of chips, and their scalability is further limited by the low speed of on-chip wires, especially over distances longer than 10 *mm* [15,16].

With the introduction of nanophotonics, systems can break free of all these limitations. The low latency and high bandwidth density of optical signaling can facilitate efficient off-chip communication and bring physically distant chips effectively close together. This makes it possible to build a physically large but logically dense many-core "virtual chip" by optically connecting several chiplets together [7,15,21].

To integrate chiplets into a larger system, Nanophotonic System in Package (NSiP) [7] uses silicon-nitride waveguides across chiplets within a package, and the Oracle Macrochip [15] uses silicon waveguides etched on a wafer. While these proposals mitigate the area, yield, and memory bandwidth limitations of conventional designs, they do not address the power constraints. The high optical loss of silicon waveguides (typically 0.1-0.3 *dB/cm* [4]) makes routing long cross-chiplet optical channels impractical from a power standpoint. Thereby, designs utilizing waveguides are confined to a small physical space (e.g., a wafer [15] or a package [7]). This increases the thermal density to the point where liquid cooling is required to avoid thermal runaways [15,16], or confines the aggregate "virtual chip" to power limitations not much different from a monolithic design [7]. Aggressive technology can produce low-loss waveguides (0.05 *dB/cm* [16]) which could enable the wide separation of discrete chiplets, but only at the expense of performance. These waveguides are 20 times wider than conventional ones, and the high area occupancy forces the design of exceedingly narrow data path links between sites (e.g., 2-bit links for an 8x8 chiplet array [15,16]) which in turn imposes significant serialization delays that degrade performance.

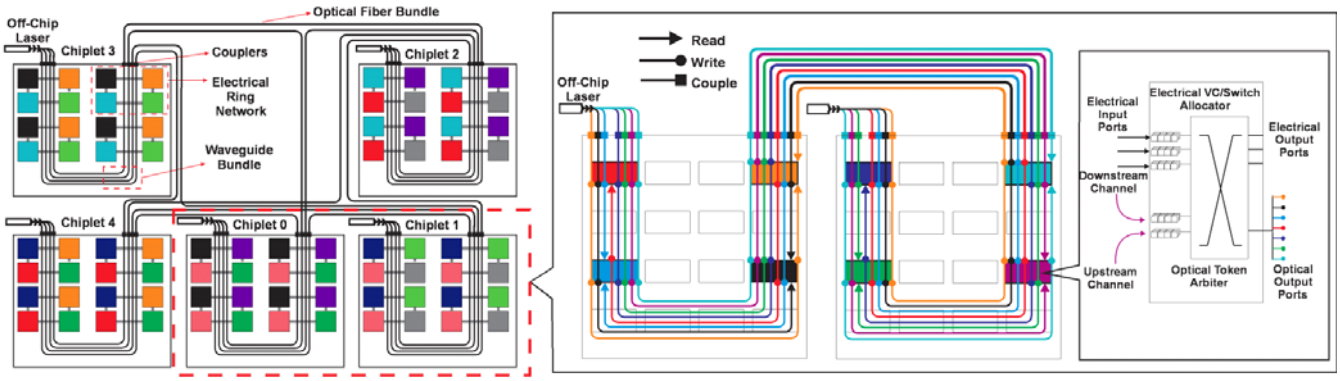
In contrast, Galaxy is designed to push back the power constraints, in addition to overcoming the area, yield, and bandwidth limitations, while matching the high performance of tightly-coupled chips. Optical fibers have tremendously low optical loss that is measured in kilometers rather than centimeters (0.2 *dB/km*), so very long channels can be drawn at very low power. Galaxy uses fibers for cross-chiplet communication, and also guarantees that each optical path employs only a small fixed number of couplers, keeping the optical loss and the corresponding laser power low. These two design choices allow spreading discrete chiplets far apart in physical space to minimize heat transfer and lower the power density of the virtual chip, which in turn enables each chiplet to operate at a higher frequency/voltage than power-limited designs. At the same time, the propagation speed of light in fibers (0.676 *c)* is considerably higher than in silicon waveguides (0.286 *c*), or electrical transmission lines on FR-4 boards (*0.5 c*), allowing for low-latency communication over long distances. In comparison to electrical lines, fibers transmit at about 33 times lower energy per bit [2].

Previous research [15] dismissed the use of optical fibers for cross-chiplet communication under the assumption that chips connect to fibers at a relatively large 250 μm core pitch, not the 20 μm pitch of optical proximity couplers that silicon waveguides use. Hence, the chip-to-chip bandwidth over fibers would not improve much over area solder balls connected to package routes. Galaxy overcomes this consideration by exploiting new tapered coupler technologies that couple an array of fibers at 250 μm pitch into an array of waveguides with 20 μm pitch at the edge of the chip [17]. Our results indicate that fibers can provide sufficient bandwidth for communication to chiplets and to memory, allowing for much wider data paths than low-loss but slow silicon waveguides, and in turn boost both the performance and the energy efficiency of the multi-chip system by several times.

In summary, optical fibers are faster, impose lower optical loss, and require lower energy than available alternatives for chiplet communication. They are cheap (a few cents per foot) and flexible, allowing for arbitrary placement of chiplets (e.g., across boards within a rack) without the need for additional coupling. Thus, fibers are especially suitable for long, inter-chiplet optical channels, as they are easy to route, and can even go off the plane or off the board. Galaxy utilizes optical fibers for cross-chiplet communication and offers simple packaging, power, and heat requirements, yet provides the performance advantages of a tightly-coupled system. While prior works have touched upon some of these issues in the context of multi-chip architectures [2, 3, 7, 15, 16, 21], to the best of our knowledge, this is the first work that extends the impact of disintegration and multi-chip integration on power limitations, and provides a comprehensive analysis of the performance, power, energy, and thermal characteristics of multi-chip architecture alternatives.

More specifically, the contributions of this paper are:

1. We quantify the performance and energy overheads that power and bandwidth constraints impose on monolithic single-chip designs, and the limitations of electrical links and SOI waveguides when used for chip communication.
2. We propose Galaxy, an architecture that allows both processor disintegration and macrochip integration. Galaxy builds a many-core "virtual chip" by connecting multiple smaller chiplets through optical fibers.
3. We thoroughly evaluate the performance, power, energy, and thermal characteristics of Galaxy, and compare it against both single-chip designs (processor disintegration) and alternative multi-chip designs (macrochip inte-



(a) **(b)** **(c)**

FIGURE 1. Galaxy layout, MWSR optical crossbar and router architecture.

gration). Our results show that Galaxy attains speedups of 1.8-2.2x over the best single-chip alternatives with electrical, photonic, or hybrid interconnects (3.4x maximum), and 2.6x smaller energy-delay product (6.8x maximum). We show that Galaxy scales to 4K cores and attains 2.5x speedup at 6x lower laser power compared to a design with silicon waveguides.

It is important to note that Galaxy is just one design that can support processor disintegration and macrochip integration. Other topologies and designs are possible. Our focus in this paper is to demonstrate the benefits of breaking free from the limitations of single-chip monolithic designs, and show that macrochip integration can achieve both high performance and low power consumption without expensive cooling solutions.

2. THE GALAXY ARCHITECTURE

Galaxy builds a physically large but logically dense many-core "virtual chip" by optically connecting several discrete chiplets together. Each chiplet consists of two dies stacked in 3D: one logic die with cores, caches, and support circuits, and one die with the photonic devices and waveguides. The two dies are connected through TSVs. Electrical signals from the logic die travel vertically to the photonic die, where they are converted to optical signals, and vise-versa.

Silicon waveguides are compatible with CMOS processes [6] and they are more efficient for long-distance on-chip communication than electrical signaling [23], leaving global on-chip wires redundant. Galaxy utilizes electrical signaling for nearest neighbor communication within a chiplet, and SOI waveguides for long-distance communication within a chiplet. The on-chip photonic interconnect extends across chiplets by coupling light to an optical fiber at the edge of the chip [17]. A photonic link in Galaxy consists of an off-chip laser source, optical fibers, fiber to on-chip waveguide couplers, SOI waveguides on the chip, a laser splitter, ring modulators, drop filters, and Germanium-based photodetectors.

2.1 Network Topology

The hybrid electrical/photonic interconnect is based on Firefly [23], which is extended to support cross-chiplet communication at low power by minimizing coupler crossings and the number of sharers of each optical path. Figure 1(a), depicts a 5-chiplet Galaxy design. The colored squares within each chiplet represent routers. The routers within a chiplet are divided into local clusters. Each cluster contains exactly one router per remote chiplet. In our example, there are 4 clusters per chiplet, with 4 routers per cluster. A local cluster in Chiplet 3 consists of neighboring black, orange, blue, and green routers (red outline in Chiplet 3, Figure 1(a)). Each cluster supports a number of cores based on a concentration factor. The cores and routers in a cluster are electrically connected. In our example, we use concentration 1 and an electrical ring within the cluster, but other topologies are possible. A source-destination pair within the same cluster uses only electrical links.

Clusters communicate with each other through optical crossbars. Every optical crossbar is represented by coloring routers with the same color. For example, the pink routers in Chiplet 0 and the pink routers in Chiplet 1 belong to the same optical crossbar. Each optical crossbar extends between two chiplets. In our example, the optical crossbar between Chiplet 0 and Chiplet 1 consists of the pink routers in Chiplets 0 and 1, the U-shaped waveguides that connect these routers together in each chiplet, and the fibers that connect the two chiplets together. Figure 1(b) shows a close-up of that crossbar, where the pink routers have been re-colored to assist a detailed explanation later in the section.

Routing a packet from Chiplet 0 to Chiplet 1 is carried by traversing the corresponding optical crossbar. This is done in 3 steps: (1) Route electrically within the source cluster in Chiplet 0 to a pink router; (2) Take the optical link and arrive at the pink router of the destination cluster in Chiplet 1; (3) Route electrically within the destination cluster to the final destination. Communication between any two clusters is performed similarly. Source-destination clusters within the same chiplet use only the silicon waveguides in that chiplet. If the clusters are at different chiplets, the packet will traverse the waveguides within the source chiplet, the fiber connecting the two chiplets, and the waveguides in the destination chiplet.

In Galaxy, every cluster has as many routers as remote chiplets, and every router in a cluster is connected to a different

optical crossbar. Thus Galaxy forms a point-to-point network between chiplets. Also, every crossbar extends across all clusters of the two chiplets it connects. Thus, each cluster has a direct connection to every cluster of every chiplet. A packet that traverses an optical link will directly reach a router in the destination cluster which is very close to the final destination, and every packet traverses the optical link only once. This minimizes coupler crossings and optical loss, as every optical path is short because it extends across only two chiplets, and has at most 3 couplers (including the laser coupling).

In general, if each chiplet has *X* clusters, each with *Y* routers, and a concentration of *c*, the proposed Galaxy architecture can connect *(Y+1)* chiplets, using radix-*(2X)* optical crossbars, supporting a total of *c*Y*X*(Y+1)* cores. The example in Figure 1 is a case with *X=Y=4*, *c=1*, for a total of 80 cores.

Firefly [23] uses Single Writer Multiple Reader (SWMR) optical crossbars, which use global broadcast channels to send messages or to reserve a channel, thereby increasing power consumption. Galaxy adopts a modified Firefly topology with Multiple Writer Single Reader (MWSR) optical crossbars. In MWSR crossbars, each router "listens" on a dedicated channel and sends flits on the listening channels of all the other routers in the crossbar. Figure 1(b) illustrates an MWSR crossbar that extends over chiplets 0 and 1, with 8 senders and 8 receivers. Every router is shown with a distinct color. Every router receives data from its own channel, which is shown with the same color as the receiver router, and writes 7 other channels which are the listening channels of the other routers in the crossbar. Galaxy adopts FairQuota [22] to guarantee that only a single router transmits on a channel at any moment, avoid starvation, and provide QoS support.

Figure 1(c) shows a hybrid electrical/optical router in Galaxy. Routers store the flits received from the electrical or optical networks in electrical buffers, after optical to electrical (O/E) conversion if needed. Two electrical input and output ports route packets on the electrical local cluster ring. The third electrical input and output port is used for data injection. Each router has a pair of dedicated optical receiving channels, the upstream and downstream channels. The dark blue and green routers in Figure 1(b) send messages to the purple router through its upstream channel, while the rest send messages to the purple router through its downstream channel. Thus, 2 extra ports are added on the input side of the router to receive packets from the dedicated optical receiving channels from both directions. On the output side, 7 additional output ports switch outgoing packets to different optical channels.

2.2 Switch Arbitration and Flow Control

The electrical switch within each router is arbitrated using conventional electrical arbiters, and uses conventional credit-based flow control. The optical crossbars require arbitration of the optical channels and the buffers at the optical receiving ports. The optical channel arbitration is equivalent to a global switch allocation, and is achieved using a 1-pass optical token stream [30] that extends across two chiplets.

Because the optical links are traversed at most once, at most two Virtual Channels (VCs) are needed for the optical channels. The buffers of each optical VC channel are arbitrated using a separate optical VC token stream. Every router keeps a count of the available buffer space for each VC, and distributes an optical VC token every cycle as long as there is still available space. A sender acquires a VC token of its intended VC before entering the arbitration for the data channel. An acquired VC token is held even if the sender fails the subsequent channel arbitration. To keep the balance of VC tokens, the tokens perform a double traversal. The receiver router of a channel first sends the VC tokens in the direction opposite to the data channel (*back-traversal*), all the way to the origin of the laser injection point, skipping all the senders on the way. Then, the VC token goes through O/E and E/O conversion, and is re-modulated onto a VC token stream in the same direction as the data channel (*forward-traversal*). The unused VC tokens eventually arrive back at the receiver and are re-collected to ensure that the receiver always knows how many VC tokens are consumed by the senders. The extra OE/EO conversion at the origin of the data channel ensures that only short optical waveguides are used.

2.3 Inter-Chiplet Optical Connection

Galaxy employs optical fibers to connect chiplets, rather than silicon waveguides. While silicon waveguides offer high bandwidth density [4], more than one order of magnitude higher than electrical interconnects [5], and have been shown to outperform SerDes links or silicon interposer chips for inter-chiplet communication [16], they have limited optical performance. Fibers have 15000x lower optical loss than typical silicon waveguides [4] and are 2.4x faster. Extremely low-loss waveguides (0.05 *dB/cm* [16]) reduce the difference to 2500x, but they are much wider than conventional waveguides (20x). This forces the design of narrow datapaths (e.g., 2-bit chiplet-to-chiplet links for an 8x8 chiplet array [15,16]) which degrades performance. Thus, fibers are especially suitable for long, inter-chiplet optical channels.

Fibers connect to chiplets through a coupler that tapers an array of fibers at 250 μm pitch down to 20 μm pitch channels, and couples them into an array of SOI waveguides at the edge of the chip [17]. The measured coupling loss caused by the refraction index change from fibers to the waveguides including misalignment is 0.8 *dB*, and the internal loss of the coupler caused by tapering the channels is 3 *dB*. Misalignment within 0.7 μm, 0.4 μm, and 0.7 μm in the lateral, vertical, and optical axes produces losses under 1 *dB* [17]. The performance of the tapered coupler is comparable to that of an optical proximity coupler (3.5 *dB* coupler loss, plus 0.5 *dB* per 1 μm misalignment in the y-axis, plus less than 1 *dB* loss due to misalignment within 2.5 μm in the x- and z-axis [34]).

TABLE 1. Nanophotonic Parameters

	per Unit	Total
Splitters	0.2 *dB*	0.2 *dB*
Waveguide Loss	0.3 *dB/cm*	1.5 *dB*
Fiber Loss	0.2 *dB/Km*	~0 *dB*
Nonlinearity	1 *dB*	1 *dB*
Coupler Loss	3.8 *dB*	7.6 *dB*
Modulator Insertion	0.5 *dB*	0.5 *dB*
Ring Through	0.01 *dB*	1.28 *dB*
Filter Drop	1.5 *dB*	1.5 *dB*
Photodetector	0.1 *dB*	0.1 *dB*
Total Loss		**13.68 *dB***
Detector Sensitivity		**-20 *dBm***
Laser Power per Wavelength		**0.233 *mW***
Total Laser Power		**1.195 *W***

2.4 Nanophotonic Parameters and Power Budget

On-chip lasers dissipate a lot of power and heat up the chip, thus Galaxy adopts off-chip WDM-compatible lasers. The laser is brought on chip via optical fibers connected to tapered couplers [17], and a splitter distributes it to low-loss on-chip waveguides [4]. Tapered couplers [17] also transfer the laser from on-chip waveguides to the off-chip optical fibers and vise-versa. Galaxy uses the modulators, demodulators, drop filters, splitters, and detectors introduced in [1]. The modulation and demodulation energy is 150 *fJ/bit* at 10 *GHz* [1]. The optical parameters assumed in Galaxy are detailed in Table 1.

Galaxy consists of 10 radix-8 MWSR crossbars that transfer 64-bit flits. We assume a modest 16-way DWDM, thus Galaxy uses a total of 320 fibers (128 fibers attached to each chiplet) and 40960 ring resonators (8192 per chiplet). Because every optical channel requires a 1-token-pass arbitration mechanism, a total of 20 additional fibers and 3840 rings are used for arbitration. Another 80 rings and 10 fibers are used for forward clock signal distribution [16].

To calculate the total ring heating power we extend the method by Nitta *et al.* [20] by incorporating the heat generated by the cores. The cores heat up the photonic layer, and the ring heaters provide the remaining heat necessary to bring

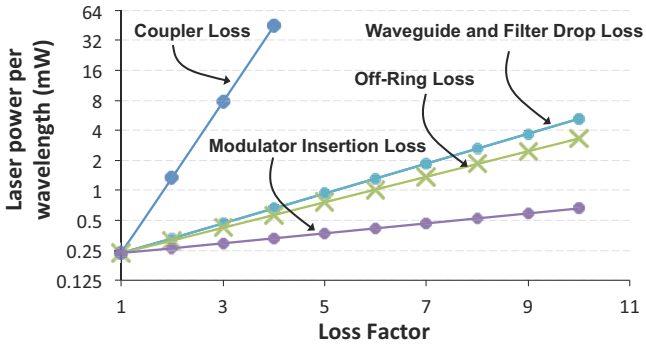


FIGURE 2. Laser power sensitivity to optical parameters

TABLE 2. Architectural Parameters.

CMP Size	80-cores, 580mm^2
Processing Cores	ULTRASPARC III ISA, max 5 Ghz, OoO, 8-stage pipeline, 4-wide dispatch/retirement, 96-entry ROB
L1 Cache	split I/D, 64KB 2-way, 2-cycle load-to-use, 2 ports, 64-byte blocks, 32 MSHRs, 16-entry victim cache
L2 Cache	512 KB per core, 16 way, 64-byte blocks, 14 cycle-hit, 32 MSHRs, 16-entry victim cache
Memory Controllers	One per 4 cores, or 4 MCs per chip. 1 channel/MC Round-robin page interleaving;
Main Memory	DDR3, 80GB, 8K pages, 20 ns access latency Interfaces: (a) Conventional pins, (b) Optically-connected memory (OCM) [1], (c) 3D-stacked [15]
Networks	CMesh, Corona, Firefly, Galaxy, Oracle Macrochip

the photonic layer within the ring tuning range. As current injection may cause a thermal runaway [20], we only consider trimming by heating. Section 3.2 details the model. While Galaxy may benefit from trimming power saving methods [20], they are out of the scope of this paper.

Figure 2 demonstrates the sensitivity of Galaxy's laser power to the nanophotonic parameters. The laser power is sensitive to the coupler loss, but relatively insensitive to the other parameters, indicating that our results will likely hold under a wide range of nanophotonic device technologies.

When evaluating electrical links (SerDes) for off-chip communication, existing literature typically omits inefficiencies in the generation and delivery of the electrical power. By analogy we didn't include the generation and delivery cost in the laser power calculations presented throughout the paper. For completeness, however, in this section we calculate the laser power including these overheads. The additional coupling loss increases the laser power to 2.9*W*. Assuming 25% efficiency for the off-chip laser source (typical range is 25-50% [34]), the wall-socket laser power consumption is 12*W*.

3. EXPERIMENTAL METHODOLOGY

We evaluate the performance of a 5-chiplet 80-core Galaxy design on a full-system cycle-accurate simulation infrastructure using Flexus 4.0 [12,32] integrated with Booksim 2.0 [8] and DRAMSim 2.0 [26]. Table 2 details the architectural modeling parameters. We target a 16nm technology, and have updated our tool chain accordingly based on ITRS projections [10]. We follow the SimFlex sampling methodology [32] with 95% confidence intervals. We model performance as the number of user instructions committed per unit of time [32]. The simulated system executes a selection of SPLASH benchmarks and scientific workloads.

We compare Galaxy against three single-Chip CMPs, all of which implement the architecture described in Table 2. The first CMP uses an all-electrical 2D Concentrated Mesh on-chip interconnect with express links [8] and concentration of

4 (CMeshExp). Concentrated mesh is often chosen for on-chip networks as it maps well to a 2D-VLSI planar layout with low complexity. We evaluated a regular 2D-Mesh and a 2D Concentrated Mesh without express links, and found that CMeshExp outperforms the other designs on all metrics (performance, power, and energy). Thus, we only show results for CMeshExp. We model routers with 8 input and output ports and a 3-cycle routing delay. Routers are connected through 166-bit bi-directional links with a 1-cycle link delay.

The second CMP uses an all-optical MWSR crossbar (Corona [31]), implemented with 256-bit data channels creating 80 MWSR crossbars. We model global switch arbitration using an optical token ring. A token for each node, which represents the right to modulate on the node's wavelength, continuously passes around all nodes on a dedicated arbitration waveguide. A node grabs and absorbs a token to transmit a packet, and then releases the token to allow other nodes to obtain it. We estimate 16cm long waveguides for the Corona chip, resulting in 8 cycles token round-trip time at 5 *Ghz*.

The third CMP implements a hybrid interconnect where clusters of electrically-connected cores are connected through an on-chip optical SWMR crossbar (Firefly [23]). The topology we model resembles Galaxy, but it is entirely on-chip.

We model Galaxy with 1-cycle latency for processing an optical token request [30]. Each Galaxy router can initiate a maximum of 8 token requests per cycle, but can utilize at most 2 acquired tokens [30]. Galaxy uses 1-pass token stream arbitration for combined VC and channel arbitration. We estimate that the round-trip time of a token is 8 also cycles. The input buffers are implemented as a DAMQ [29], with packets queued separately based on their destination. A data packet contains 512 bits, which are divided into eight 64-bit flits.

3.1 Power and Temperature Modeling

All systems we model employ Dynamic Voltage and Frequency Scaling (DVFS) to lower the voltage and frequency of a chip or chiplet when it reaches the limits of safe operational temperature (90^oC). Figure 3 shows the flow diagram of our simulation tool chain. We collect runtime statistics from full-system simulations, and use them to calculate the power consumption of compute cores, caches, and memory controllers using McPAT [18], and the power consumption of the electrical and optical networks using DSENT [28] and the analytical model by Joshi *et al.* [14] respectively. Based on these power estimates, we calculate the temperature of the chip and chiplet assemblies using HotSpot 5.0 [27] and FloTherm [33], a computational fluid dynamics tool that models the heat transfer between chiplets through air flow and convection. The estimated temperature is then used to refine the leakage power estimate, and we iteratively calculate the power and temperature profiles until the system reaches a stable state. We use the stable-state power and temperature estimates to adjust DVFS, and repeat the process until we identify a DVFS setting for which the chip or chiplet stays just below 90^oC.

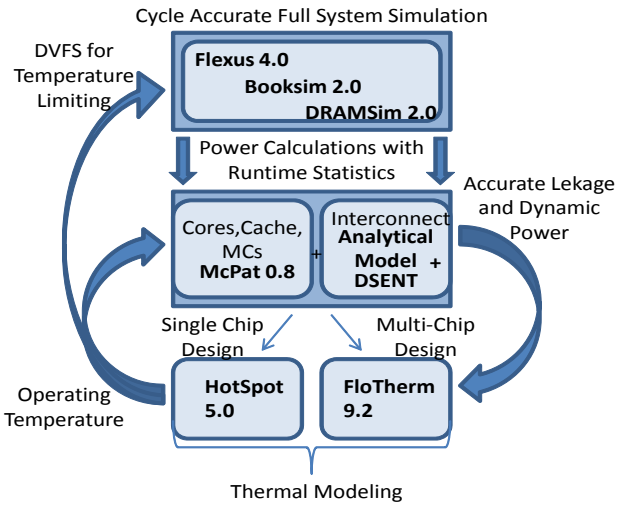


FIGURE 3. Simulation flow chart

3.2 Resonant Ring Heater Modeling

To calculate the total ring heating power for Galaxy, Corona, and Firefly, we extend the method by Nitta *et al.* [20] by additionally accounting for the heating of the photonic die by the operation of the cores. We model the thermal characteristics of a 3D-stacked architecture where the photonic die sits underneath the logic die using the 3D-chip extension of HotSpot [27]. For each target architecture (Corona, Firefly, and Galaxy) we measure the maximum temperature of the logic die during the execution of each one of the workloads. Then, we tune the micro-rings to the maximum of all the observed temperatures that the logic layer reaches across all benchmarks executing on the target architecture, plus a small margin. When a workload executes, we calculate the ring heating power required to maintain the entire photonic die at the micro-ring trimming temperature during the entire execution.

3.3 Modeling Memory and Physical Constrains

For systems with conventional DDR3 memory, ITRS [10] pin projections limit single-chip designs to four memory controllers (MCs). In contrast to single chips, Galaxy can employ 20 MCs (5 chiplets with 4 MCs each). Emerging memory technologies such as optically-connected memory (OCM) [1] or 3D-stacked memory [15] lift this constraint and allow 20 MCs per chip for all designs. Thus, we separately evaluate the performance of Galaxy against single-chip CMPs with emerging memory technologies. We model a 10 *ns* access latency for OCM and a 2 *ns* access latency for 3D-Memory.

To show the effect of physical constraints on single-chip CMPs, we evaluate Galaxy against CMPs that operate beyond physical constraints: CMeshExp_4MC, Corona_4MC, and Firefly_4MC model a CMeshExp, Corona, and Firefly interconnect, respectively. The interconnects operate within the bandwidth limits imposed by technology projections for 16nm [10], but run at the maximum speed allowed by the design (5 *GHz*), by disregarding power and thermal limitations. CMeshExp_20MC and Corona_20MC additionally dis-

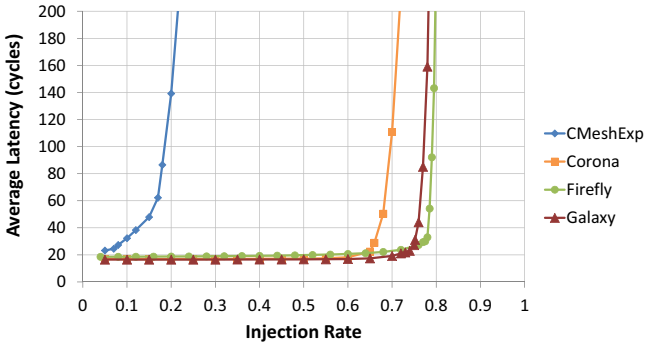


FIGURE 4. Load latency curves for uniform random traffic.

regard bandwidth limitations and operate with 20 MCs for conventional memory (i.e., DDR3 memory connected to the processor through the off-chip pins). We refer to such architectures as "*Unrealistic*", as they can operate beyond the power, thermal, or bandwidth limits. In contrast, "*Realistic*" architectures employ 4 MCs on a chip (20MCs with OCM and 3D-Memory) and operate under 90^oC with DVFS. While we compare Galaxy to both "Realistic" and "Unrealistic" single-chip CMPs, Galaxy is always modeled as "Realistic".

3.4 Modeling Large-Scale Designs

Galaxy can scale up to 1088 cores with 17 chiplets (64 cores each with concentration 4), and 4160 cores with 65 chiplets. When increasing the number of chiplets, we decrease the width of chip-to-chip links to keep the network power consumption and component count within reasonable levels, and we faithfully model the serialization delay due to narrower datapaths, and increased link latency due to longer links. We evaluate the scalability of Galaxy by comparing it against (a) Galaxy with SOI waveguides and optical proximity (OPC) couplers [34], (b) Galaxy with electrical links (SerDes), and (c) the Oracle Macrochip [15]. For fairness, we adjust the datapath width of Galaxy alternatives so they fit into similar power envelopes, and then calculate the latency overhead. The Oracle Macrochip model closely follows [15,34]. Table 3 details the characteristics of each design. To keep the simulations tractable, we estimate the performance of the scaled-out designs by imposing the latency overheads of each scaled-out system from Table 3 on an 80-core 5-chiplet model. As we impose the scaling overheads onto same-size designs in all cases (80 cores, 5 chiplets), the higher core count of Galaxy compared to the Oracle Macrochip does not affect the results.

4. EXPERIMENTAL RESULTS

4.1 Network Performance

Figure 4 analyzes the load-latency of CMeshExp, Corona, Firefly, and Galaxy. CMeshExp saturates quickly, which is indicative of its relatively low bandwidth. Corona saturates at a little less than 0.7 injection rate, while Firefly reaches an injection rate of almost 0.8 before saturating. Galaxy trails Firefly closely, but falls slightly short in performance as it saturates at an injection rate of 0.75. This is expected because

TABLE 3. Galaxy scalability.

# of Cores	Multi-Chip Architecture		Bandwidth per Chip (TB/s)	Laser Power (W)	Serialization Overhead (cycles)	Link Latency (cycles)
320	Galaxy	Fibers	10	4.0	1	2
		Waveguides	5	4.9	2	10
		SerDes	0.320	3.9	32	12
1088		Fibers	20	27.0	2	10
		Waveguides	5	26.0	8	20
		SerDes	0.640	26.8	64	12
4160		Fibers	40	47.6	4	10
		Waveguides	10	44.9	16	20
		SerDes	0.320	47.9	512	12
4096	Oracle MacroChip		0.630	~40.0	64	20

Galaxy is similar to a 2-level Firefly design that creates a single datapath between two clusters, while packets in Firefly can take several alternate routes, thereby utilizing more of the available bandwidth. Nonetheless, the difference is small, indicating that Galaxy is a competitive interconnect design.

4.2 Comparison to Single-Chip Designs

To assess the effect of bandwidth and power limitations on single chip CMPs, we compare Galaxy to "Unrealistic" architectures (Section 3.3). Figure 5 shows the speedup achieved by (a) CMeshExp_4MC, Corona_4MC, and Firefly_4MC (i.e., CMPs that are not subject to power limitations, but operate under realistic off-chip memory bandwidth constraints), (b) CMeshExp_20MC and Corona_20MC (i.e., CMPs that are not subject to power nor off-chip bandwidth constraints), and (c) Galaxy (fully-constrained in both power and off-chip bandwidth). All designs are supported by conventional memory. The speedups are normalized to CMeshExp_Realistic with conventional memory.

The memory-intensive workloads (appbt, em3d, ocean, tomcatv) highly utilize the memory and the on-chip interconnect. When memory bandwidth is not limited, the average message latency largely determines their performance: Galaxy and Corona_20MC (16-cycle average message latency for both) show similar speedups, whereas CMeshExp_20MC (23-cycle

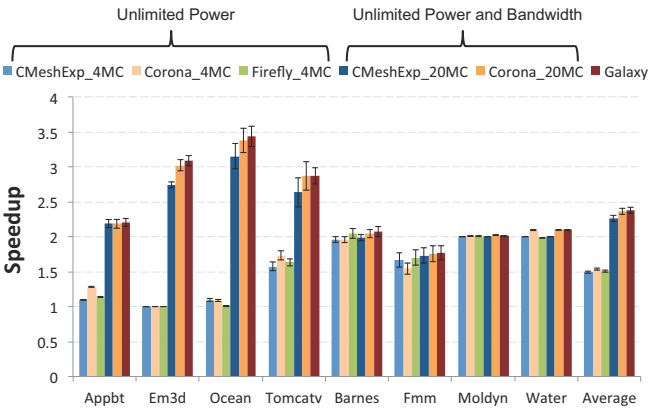


FIGURE 5. Speedup of unrealistic architectures

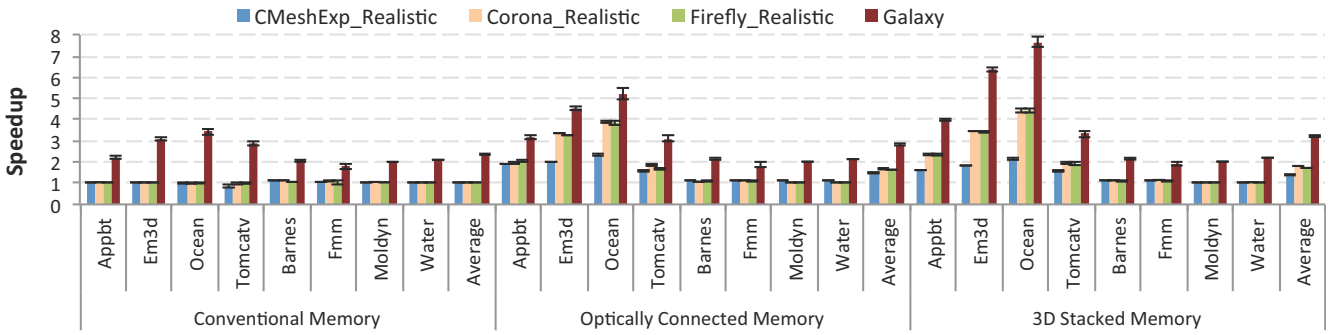


FIGURE 6. Speedup of realistic architectures with various memory technologies (normalized to CMeshExp_Realistic with DDR3).

average message latency) is slower. However, when memory bandwidth is limited, it overthrows message latency and becomes the main bottleneck: Galaxy is 3.4x faster than Corona_4MC running ocean, while CMeshExp_4MC and Firefly_4MC perform almost as well as Corona_4MC. As expected, compute-intensive workloads (barnes, fmm, moldyn, water) are insensitive to memory bandwidth limitations. Overall, Galaxy matches or exceeds the performance of single-chip architectures, even when they are not limited by power or off-chip bandwidth.

"Realistic" architectures (Figure 6) employ DVFS to keep the chips below 90^oC. In the process of doing so, DVFS slows down the compute-intensive workloads the most, as they have high core utilization which in turn dissipates more power. For example, Corona_Realistic with conventional memory runs barnes at only 2.25 *GHz* from a nominal frequency of 5 *GHz*. Firefly_4MC exhibits a similar slowdown. In comparison, Galaxy never exceeds 70^oC, and thus it can run at the full 5 *GHz* and outperform all single-chip alternatives. Memory-intensive workloads on designs with conventional memory show degraded performance mainly due to the off-chip bandwidth limitations, while the slowdown due to DVFS is secondary. For example, CMeshExp_Realistic runs em3d at 4.25 *GHz*, but Galaxy still demonstrates 3x speedup. Because of this dual slowdown, Galaxy achieves the maximum speedup over realistic CMPs on memory-intensive workloads (2.3x on average, and up to 3.46x when running ocean).

Optically-connected memory (OCM) [1] is able to overcome the bandwidth limitations and decrease the memory latency. Corona_Realistic with OCM outperforms Corona_Realistic with conventional memory by 3-4x on memory intensive workloads. Firefly_Realistic and CMeshExp_Realistic show similar trends. However, Galaxy still outperforms Corona, Firefly, and CMeshExp by 1.8x on average, as Galaxy runs at the full 5 *GHz* while DVFS limits the single-chip alternatives (for example, Corona_Realistic with OCM runs em3d at 3.25 *GHz*). 3D-stacked memory lowers the memory access latency and increases the speedup for Galaxy, while Corona, Firefly, and CMeshExp do not get any faster as they are still power limited. Overall, Galaxy outperforms alternative designs by up to 2.95x (2x on average). We conclude that Galaxy can leverage the emerging memory technologies to the fullest extent, while single-chip CMPs are still limited by the single-chip power envelope and fail to utilize fully the new memory technologies.

Figure 7 shows the breakdown of the normalized energy-delay product (EDP) and the average energy per instruction of CMeshExp_Realistic, Corona_Realistic, Firefly_Realistic, and Galaxy with conventional memory. The dynamic energy consumption of cores and caches for Galaxy is higher as it achieves 2.3x speedup on average over single-chip designs. This effect is more pronounced for compute-intensive workloads. However, the chiplets in Galaxy run at only 70^oC and dissipate $55W$ each, compared to 90^oC and $130W$ for realistic CMeshExp-, Corona-, and Firefly-based chips. As a result, Galaxy lowers leakage to just over 10% of energy, while single-chip designs waste 36-40% of their energy on leakage. Overall, realistic CMPs consume 1.12-1.2x more energy per instruction on average than Galaxy (Figure 7(b)). Galaxy

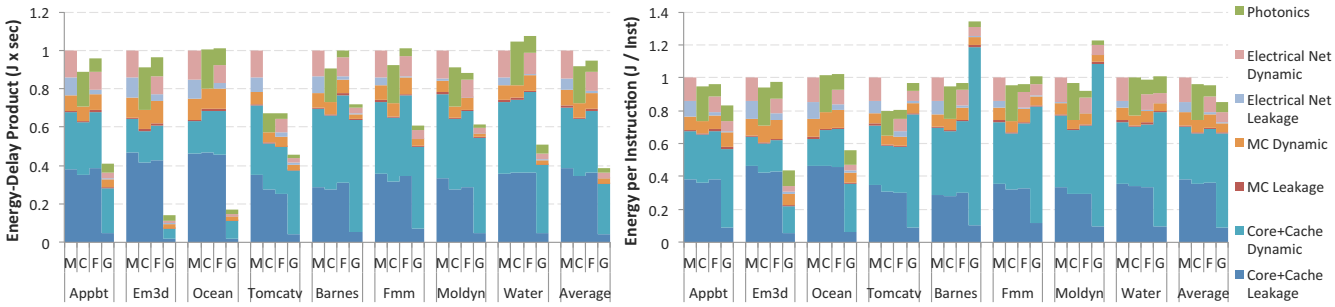


FIGURE 7. (a) Energy x Delay product for realistic architectures. (b) Average energy per instruction for realistic architectures. Key: M=CMeshExp_Realistic, C=Corona_Realistic, F=Firefly_Realistic, G=Galaxy

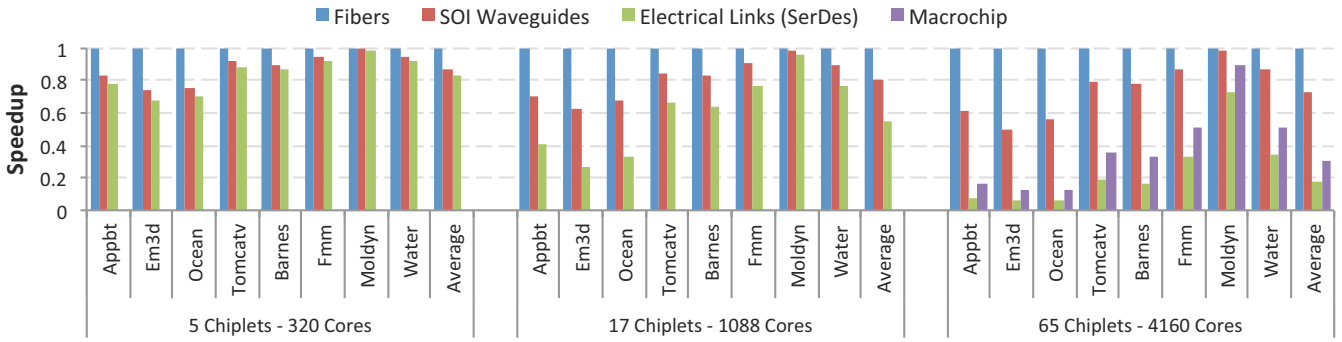


FIGURE 8. Comparison of Galaxy with different chiplet-to-chiplet interconnect technologies, and the Oracle Macrochip.

reaches its highest energy efficiency increase on memory-bound workloads (2-2.3x), as it achieves over 3x speedup and the chiplets dissipate less power waiting for memory. Galaxy attains up to 6.8x lower EDP than single-chip CMPs (2.8x on average; Figure 7(a)).

Because Galaxy chiplets run cooler when running memory intensive workloads, the energy consumption of the photonic network (including laser power, modulation/demodulation, and ring heating) is higher, as the ring heaters dissipate more power to keep the photonics layer at the trimming temperature. The energy consumption of photonics is lower with compute intensive workloads, because cores dissipate more power and heat the photonic die, so ring heaters work less.

4.3 Comparison to Multi-Chip Designs

Galaxy can scale up to 1088 cores with 17 chiplets, and 4160 cores with 65 chiplets (Section 3.4). Table 3 details the power, bandwidth, and latency characteristics of the scaled out designs, and compares Galaxy with fibers to designs that utilize SOI waveguides or electrical (SerDes) links for chiplet-to-chiplet communication, as well as the Oracle Macrochip. Figure 8 compares the performance of these alternatives by modeling the effect of link latency and serialization on performance, following the methodology in Section 3.4.

The power-hungry SerDes links cannot provide enough bandwidth within the power envelope, resulting in high serialization delay that increasingly hurts performance as the system scales up. Similarly, SOI waveguides require higher laser power than fibers, as the optical loss in SOI waveguides increases rapidly as their length grows, and at the same time they are 2.3x slower than fibers due to different light propagation speeds between the two materials. As a result, fibers increasingly outperform SOI waveguides as the system scales up. The performance gap is higher for memory-intensive workloads which stress the interconnect more. A 65-chiplet Galaxy with fibers outperforms Galaxy with SOI waveguides by up to 1.44x (1.24x on average), and Galaxy with electrical links (SerDes) by up to 9.53x (4.58x on average).

The Oracle Macrochip [15,16] uses SOI waveguides and OPCs [34] to create point-to-point photonic links across chips. Galaxy outperforms the Oracle Macrochip by 2.5x on average (Figure 8) because the Macrochip employs 2-bit-wide data channels which impose high serialization delay, and SOI waveguides are slower that optical fibers

Because the coupler loss is the biggest contributor to the laser power consumption, we evaluate the sensitivity of laser power to the coupler loss for the Oracle Macrochip and Galaxy (Figure 9). In the figure we indicate the laser power consumption of the Oracle Macrochip with measured coupler losses for passive-aligned and active-aligned OPCs [34], as well as under aggressive OPC loss predictions [15,16]. For Galaxy, we indicate the laser power consumption under SION and SU8 tapered couplers using loss measurements of existing prototypes [17]. Because macrochip links have to pass through 3 couplers to go from one chiplet to another (vs. 2 for Galaxy), the slope of the laser power is higher indicating that it is more sensitive to coupler loss. The Macrochip with actively-aligned OPCs requires 6x more laser power than Galaxy. Even if the predicted OPC loss is achieved, Galaxy with existing couplers would still require less laser power.

4.4 Thermal Evaluation

To effectively push back the power wall while still employing conventional forced air cooling solutions and cheap packaging appropriate for high-volume markets, a disintegrated design requires the chiplets to be physically far enough from each other to minimize heat transfer. Our thermal modeling using CFD tools [33] and HotSpot [27] indicates that a Galaxy architecture with active heatsinks on each chiplet allows the chiplets to operate at $66.2^{o}C$, sufficiently cool for most applications. In fact, even cheaper cooling solutions seem sufficient. Figure 10(a) shows a Galaxy design with 5 chi-

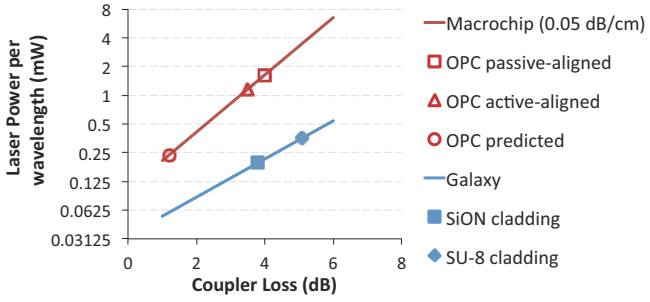


FIGURE 9. Laser power sensitivity to coupler loss.

plets. The chiplets use passive heatsinks and are spaced 8 *cm* apart, with a global fan blowing air horizontally in 45^oC ambient air temperature in a box shell. The fanless (passive) heatsinks cool chiplets down to 88.2^oC, and deliver low packaging and cooling costs and increased lifetime. Thus, even very simple and cheap cooling solutions (fanless heatsinks and a global fan) are adequate to cool an 80-core 5-chiplet Galaxy.

Optical fibers allow Galaxy to spread chiplets far apart for better cooling, while SOI waveguides and electrical SerDes links can not. As the Oracle Macrochip utilizes SOI waveguides for intra-chiplet communication, it is confined to a single wafer [15] and requires specialized liquid cooling solutions, which are too expensive for most market segments. We compare the thermal characteristics of a Macrochip-like dense design to an equal-size Galaxy by modeling a 3x3 Macrochip architecture and a 9-chiplet Galaxy. Both designs use the same heatsinks. Based on the details of the Macrochip architecture [15,16], we estimate that the heatsinks will almost touch each other resulting in the layout shown at Figure 10(b). We observe that the sites that are further away from the fan reach 249^oC, and hence cannot be cooled with conventional forced air solutions.

In comparison, a 9-chiplet Galaxy design which dissipates the same amount of dynamic power as the Macrochip can be cooled with forced air and passive heatsinks. The thermal-aware placement of chiplets on a 2D-plane shown in Figure 10(c) increases the x-dimension of the board from 12 *cm* in the Macrochip layout to 28 *cm*, while the y-dimension remains the same. In return for the larger board, the Galaxy design achieves a maximum temperature of 110^oC, which is a full 139^oC lower than Macrochip. Furthermore, using optical fibers for cross-chiplet communication allows Galaxy to utilize multiple boards. Figure 10(d) shows that Galaxy can bring a 9-chiplet design down to a cool 87^oC using only conventional forced air and a 3D layout. This freedom of placement gives a significant advantage to Galaxy compared to silicon-waveguide-based designs, and allows it to spread the volume enough to cool even large-scale designs.

5. LIMITATIONS AND CHALLENGES

5.1 Misalignment and Fiber Density Considerations

The use of fibers for chiplet-to-chiplet communication in Galaxy brings two new challenges: coupling the fibers on chip and attaching enough fibers to achieve the highest performance or lowest EDP, depending on the optimization target.

Fibers connect to chiplets through a coupler that tapers an array of fibers at 250 μm pitch down to 20 μm pitch channels, and couples them into an array of SOI waveguides at the edge of the chip [17]. Characterization on fabricated tapered couplers has measured total coupling loss as low as 3.8 *dB* [17]. Part of this loss comes from misalignment. Misalignment within 0.7 μm, 0.4 μm, and 0.7 μm in the lateral, vertical, and

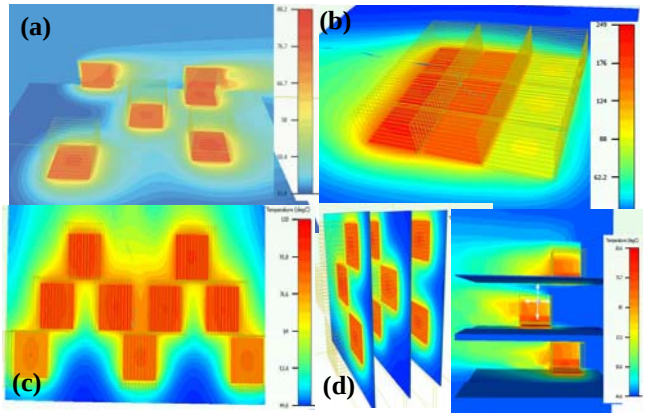


FIGURE 10. Thermal effects of chiplet placement.

optical axes produces losses under 1 *dB* [17]. In comparison, optical proximity couplers have been measured to achieve as low as 3.5 *dB* optical loss [34]. The optical loss of OPC couplers increases by 0.5 *dB* per 1 μm misalignment in the y dimension, plus less than 1 *dB* loss due to misalignment within 2.5 μm in the x and z dimensions [34].

Overall, the performance of the tapered coupler is comparable to that of an optical proximity coupler. OPC coupling is more forgiving of misalignment, allowing three times higher misalignment than tapered couplers in the x- and z-axis for similar loss. Without a large volume of characterization experiments, however, it is hard to distill statistically significant results for either technology. In addition, tapered couplers are more amenable to active alignment (albeit at a higher manufacturing cost), as each time only a subset of the fibers is aligned, while OPC couplers need to be aligned all together. Despite the misalignment hurdles, tapered couplers allow the use of fibers which exhibit simultaneously both negligible optical loss and high bandwidth density, more than making up for the higher misalignment loss (Figure 9).

Galaxy requires enough length along the periphery of a chiplet to attach the fibers. Galaxy's 116 mm^2 chiplets provide over 43 *mm* in total length along the edge of a chip, allowing up to 172 fibers at a 250 μm pitch. The design we have evaluated assumes 128 fibers per chiplet with 16 DWDM on 64-bit-wide datapaths. Figure 11 indicates that having 512 fibers (i.e., 4x the fiber density) will increase the performance by only 3%, while dissipating 4x more laser power, so this is not a desirable design point. On the other hand, using 64 fibers would reduce performance by only 2.4% over the Galaxy design we evaluated, and consume half the laser power, so this is also a viable solution that requires fewer fibers per chiplet. Employing a less dense fiber array, however, causes evident performance degradation. Galaxy with 32 fibers per chiplet is 7% slower, and Galaxy with 16 fibers is 15.5% slower than the design we evaluated in this paper. While these design points will still provide a performance and EDP benefit over electrical SerDes links and SOI waveguides, the bandwidth limitations quickly reduce the performance of the

system. Thus, applications that require significant chiplet-to-chiplet bandwidth, but allow only a few fibers to be attached to a chiplet due to practical or economic reasons, may not benefit as much as the workloads we evaluated in this paper.

5.2 Board-Level Effects

Spreading the chiplets far apart to decrease the thermal density of the design and allow forced-air cooling requires larger boards (Section 4.4). This may be an impediment to designs that strive for compute density. However, it is important to note here that any cooling solution applicable to multi-chip or multi-socket systems is readily applicable to Galaxy. The additional advantage that Galaxy offers is that the system designer can choose how close the chiplets should be to realize a given cooling solution. Thus, Galaxy allows higher design flexibility, and the ability to explore all cooling solutions and their economic trade-offs, from forced air to liquid cooling and beyond.

Similarly, by allowing the chiplets to run at full speed, Galaxy consumes more power at the board level which may stress the board-level power delivery system. However, fibers allow the chiplets to be spread in 3D-space and occupy multiple boards, while still behaving like a large virtual chip (Section 4.4). Thus, Galaxy can utilize as much power as can be safely delivered to each board, and improve performance (Figure 6) while minimizing waste (Figure 7). Overall, the flexibility to spread the design over multiple boards allows the system designer greater flexibility in deciding how many boards to employ and how much power to deliver to each one, based on the other physical, financial, and engineering constraints that the system must satisfy.

5.3 Yield and Financial Considerations

Galaxy relies on the manufacturing of a photonic die, 3D integration of the photonic and the logic dies, and the manufacturing of tapered couplers and fibers. Each one of these steps carries its own inefficiencies and costs, which are likely to be higher (at least initially) than the cost of the mature CMOS processes. Of all these components, fibers have been manufactured at high volumes and they have become very cheap (a few cents per foot). To assist in calculating the cost of the system, Section 2.4 provides component counts for the nanophotonic devices. While the absence of yield and manufacturing data for nanophotonic systems do not allow us to make quantitative arguments, we expect that the additional manufacturing steps will increase the overall cost of the system.

However, processor disintegration allows Galaxy to recover the additional overhead or even achieve lower overall cost than conventional monolithic single-chip designs. By breaking a monolithic chip into multiple smaller chiplets, one can increase yield and lower non-recurring and marginal costs by a significant factor, especially for low and medium volume markets, as only the defective chiplets need to be replaced rather than an entire large chip ([7]). As technology matures, nanophotonic devices and 3D integration are likely to enjoy higher yields and be competitive to CMOS processes, tilting the balance more in favor of disintegrated architectures.

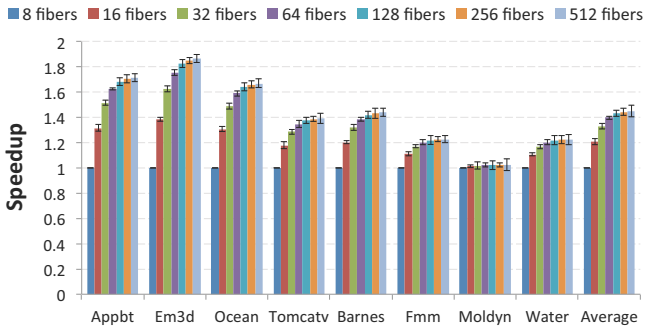


FIGURE 11. Sensitivity to fiber density per chiplet.

6. RELATED WORK

Several on-chip interconnect networks exploiting optical signaling have been proposed. The Corona [31] architecture implements a monolithic crossbar topology to support on-chip as well as off-chip communication. Joshi *et al*. [14] propose a nanophotonic clos network. The hierarchical Firefly architecture [23] advocates the use of partitioned nanophotonic crossbars to connect clusters of electrically-connected nodes, improving power efficiency, and providing uniform global bandwidth between all clusters.

Previously, Beamer *et al*. [3] explained how multi-socket systems can provide higher hardware parallelism while using smaller dies with high production yield. Batten *et al*. [1] proposed to connect a many-core processor to the DRAM memory using monolithic silicon. Koka *et al*. [15] discuss the design and implementation of a silicon-photonic network for a large multi-die “macrochip” system. In contrast to these architectures, Galaxy leverages optical fibers to create a high-bandwidth, scalable, low-latency photonic interconnect that can support both processor disintegration and multi-chip integration, and at the same time enable cheap cooling solutions.

7. CONCLUSIONS

In this paper we propose Galaxy, a multi-chip architecture which builds a many-core “virtual chip” by connecting multiple smaller chiplets through optical fibers. Galaxy is designed to push back the power constraints, in addition to overcoming the area and bandwidth limitations, while matching the high performance of tightly-coupled chips. We demonstrate that Galaxy achieves 1.8-3.4x average speedup over competing single-chip designs, and achieves 2.6x lower energy-delay product (6.8x maximum). The careful design of optical paths in Galaxy minimize coupler crossings and allows it to scale beyond 4K cores, showing significant promise as the foundation of practical large-scale virtual chip designs. Finally, we show that a scaled-out Galaxy attains significant speedup and energy efficiency advantages over competing designs such as the Oracle Macrochip as it achieves at least 2.5x speedup with 6x more power-efficient optical links.

8. REFERENCES

[1] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C. W. Holzwarth, M. A. Popovic, H. Li, H. I. Smith, J. L. Hoyt, F. X. Kartner, R. J. Ram, V. Stojanovic, and K. Asanovic. Building many-core processor-to-DRAM networks with monolithic CMOS silicon photonics. *IEEE Micro*, 29:8–21, 2009.

[2] S. Beamer. *Designing Multisocket Systems with Silicon Photonics*. PhD thesis, University of California at Berkeley, 2009.

[3] S. Beamer, K. Asanovic, C. Batten, A. Joshi, and V. Stojanovic. Designing multi-socket systems using silicon photonics. In *Proceedings of the Annual International Conference on Supercomputing (ICS)*, pages 521–522, Yorktown Heights, NY, 2009.

[4] J. Cardenas, C. Poitras, J. Robinson, K. Preston, L. Chen, and M. Lipson. Low loss etchless silicon photonic waveguides. *Optics Express*, 17(6):4752–4757, 2009.

[5] M. Chang, J. Cong, A. Kaplan, M. Naik, G. Reinman, E. Socher, and S.-W. Tam. CMP network-on-chip overlaid with multi-band RF-interconnect. In *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 191–202, Feb. 2008.

[6] G. Chen, H. Chen, M. Haurylau, N. Nelson, P. M. Fauchet, E. Friedman, and D. Albonesi. Predictions of CMOS compatible on-chip optical interconnect. In *7th International Workshop on System-Level Interconnect Prediction (SLIP)*, pages 13–20, 2005.

[7] M. Cianchetti, N. Sherwood-Droz, and C. Batten. Implementing System-in-Package with Nanophotonic Interconnect. *Workshop on the Interaction between Nanophotonic Devices and Systems (in conj. with MICRO-43)*, December 2010.

[8] W. J. Dally and T. B. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann Publishing Inc., 2004.

[9] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger. Dark silicon and the end of multicore scaling. In *Proceedings of the 38th Annual International Symposium on Computer Architecture*, ISCA '11, pages 365–376, 2011.

[10] European Semiconductor Industry Association (ESIA), Japan Electronics and Information Technology Industries Association (JEITA), Korean Semiconductor Industry Association (KSIA), Taiwan Semiconductor Industry Association (TSIA), and United States Semiconductor Industry Association (SIA). *International Technology Roadmap for Semiconductors (ITRS), 2012 edition*.

[11] N. Hardavellas, M. Ferdman, B. Falsafi, and A. Ailamaki. Toward dark silicon in servers. *IEEE Micro*, 31(4):6–15, July-August 2011.

[12] N. Hardavellas, S. Somogyi, T. F. Wenisch, R. E. Wunderlich, S. Chen, J. Kim, B. Falsafi, J. C. Hoe, and A. G. Nowatzyk. SimFlex: a fast, accurate, flexible full-system simulation framework for performance evaluation of server architecture. *SIGMETRICS Performance Evaluation Review, Special Issue on Tools for Computer Architecture Research*, 31(4):31–35, April 2004.

[13] M. Horowitz. Scaling, power and the future of cmos. In *Proceedings of the 20th International Conference on VLSI Design*, page 23, 2007.

[14] A. Joshi, C. Batten, Y.-J. Kwon, S. Beamer, I. Shamim, K. Asanovic, and V. Stojanovic. Silicon-photonic clos networks for global on-chip communication. In *Proceedings of the IEEE International Symposium on Networks-on-Chip (NOCS)*, pages 124–133, 2009.

[15] P. Koka, M. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. Krishnamoorthy. Silicon-photonic network architectures for scalable, power-efficient multi-chip systems. In *Proceedings of the 37th Annual International Symposium on Computer Architecture*, ISCA '10, pages 117–128, Saint-Malo, France, 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*, 97(7):1337 –1361, july 2009.

[17] B. Lee, F. Doany, S. Assefa, W. Green, M. Yang, C. Schow, C. Jahnes, S. Zhang, J. Singer, V. Kopp, J. Kash, and Y. Vlasov. 20um-pitch eight-channel monolithic fiber array coupling 160 Gb/s/channel to silicon nanophotonic chip. In *Conference on Optical Fiber Communications and National Fiber Optic Engineers Conference (OFC/NFOEC)*, pages 1–3, March 2010.

[18] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi. Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In *Proceedings of the 42nd IEEE/ACM Annual International Symposium on Microarchitecture*, MICRO-42, pages 469–480, 2009.

[19] R. Merritt. ARM CTO: Power surge could create dark silicon. http://www.eetimes.com/electronics-news/4085396/ARM-CTO-power-surge-could-create-dark-silicon-, Oct. 2009.

[20] C. Nitta, M. Farrens, and V. Akella. Addressing system-level trimming issues in on-chip nanophotonic networks. In *IEEE 17th International Symposium on High Performance Computer Architecture (HPCA)*, pages 122–131, Feb. 2011.

[21] Y. Pan, Y. Demir, N. Hardavellas, J. Kim, and G. Memik. Exploring benefits and designs of optically connected disintegrated processor architecture. *Workshop on the Interaction between Nanophotonic Devices and Systems (in conj. with MICRO-43)*, December 2010.

[22] Y. Pan, J. Kim, and G. Memik. Featherweight: low-cost optical arbitration with QoS support. In *Proceedings of the 44th IEEE/ACM Annual International Symposium on Microarchitecture*, MICRO-44, pages 105–116, 2011.

[23] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *Proceedings of the 36th Annual International Symposium on Computer Architecture*, ISCA '09, Austin, TX, 2009.

[24] J. Poulton, R. Palmer, A. Fuller, T. Greer, J. Eyles, W. Dally, and M. Horowitz. A 14-mW 6.25-Gb/s transceiver in 90-nm CMOS. *IEEE Journal of Solid-State Circuits*, 42(12):2745 –2757, Dec. 2007.

[25] B. M. Rogers, A. Krishna, G. B. Bell, K. Vu, X. Jiang, and Y. Solihin. Scaling the bandwidth wall: challenges in and avenues for CMP scaling. In *Proceedings of the 36th Annual International Symposium on Computer Architecture*, ISCA '09, pages 371–382, 2009.

[26] P. Rosenfeld, E. Cooper-Balis, and B. Jacob. DRAMSim2: A cycle accurate memory system simulator. *Computer Architecture Letters*, 10(1):16 –19, Jan.-June 2011.

[27] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-aware microarchitecture. In *Proceedings of the Annual International Symposium on Computer Architecture (ISCA)*, ISCA '03, pages 2–13, 2003.

[28] C. Sun, C.-H. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic. DSENT - a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling. In *Sixth IEEE/ACM International Symposium on Networks on Chip (NoCS)*, pages 201 –210, May 2012.

[29] Y. Tamir and G. Frazier. Dynamically-allocated multi-queue buffers for VLSI communication switches. *IEEE Transactions on Computers*, pages 725–737, 1992.

[30] D. Vantrease, N. L. Binkert, R. Schreiber, and M. H. Lipasti. Light speed arbitration and flow control for nanophotonic interconnects. In *Proceedings of the IEEE/ACM Annual International Symposium on Microarchitecture (MICRO)*, pages 304–315, New York, NY, 2009.

[31] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. G. Beausoleil, and J. H. Ahn. Corona: System implications of emerging nanophotonic technology. In *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ISCA '08, pages 153–164, 2008.

[32] T. F. Wenisch, R. E. Wunderlich, M. Ferdman, A. Ailamaki, B. Falsafi, and J. C. Hoe. SimFlex: statistical sampling of computer system simulation. *IEEE Micro*, 26(4):18–31, Jul-Aug 2006.

[33] M. Yang. A comparison of using icepak and flotherm in electronic cooling. In *Proceedings of the 7th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM)*, volume 1, pages 240 –246, May 2000.

[34] X. Zheng, J. E. Cunningham, I. Shubin, J. Simons, M. Asghari, D. Feng, H. Lei, D. Zheng, H. Liang, C. chih Kung, J. Luff, T. Sze, D. Cohen, and A. V. Krishnamoorthy. Optical proximity communication using reflective mirrors. *Optics Express*, 16(19):15052–15058, Sep 2008.