NUPLet: A Photonics Based Multi-Chip NUCA Architecture

Janibul Bashir and Smruti R. Sarangi
Department of Computer Science and Engineering
Indian Institute of Technology, New Delhi, India 110016
Email: {janibbashir,srsarangi}@cse.iitd.ac.in

***Abstract*—Area, manufacturing yield and lack of scalable interconnects restrict single chip designs to a small number of cores (16-32). However, multi-chip designs with the help of silicon photonics can overcome area and yield constraints and make it possible to design a virtual chip, which can scale to a large number of cores. Sadly, the scalability of such designs is limited by the high percentage of inter-chip messages and relatively lower hit rate in remote cache banks.**

In this paper, we propose *NUPLet*, a multi-chip architecture that tries to remove these limitations by separating the intra and inter chip networks. It proposes to use a non-uniform cache architecture (NUCA) scheme on top of a virtual chip in order to decrease inter chip communication and increase the hit rate in the last level cache. In addition, we propose a prediction mechanism for predicting the number of inter chip messages in the network. This is used to modulate the laser accordingly, and reduce static power consumption.

We simulated a four chip based *NUPLet* design with each chip containing 32 cores. For a suite of Splash2 and Parsec benchmarks, *NUPLet* increased the last level cache hit rate by 70% as compared to other state of the art proposals. Furthermore, *NUPLet* improved performance by 28%, reduced power consumption by 39%, and reduced ED^2 by 41%.

I. Introduction

With the impending death of Moore's law, and the lack of non-silicon alternatives (as of today), sustaining traditional rates of scaling computational throughput in manycore processors is a challenging problem. Since it will not be possible to scale the number of cores per chip by a factor of two every few years, innovation in computer architecture has moved to other areas. One such promising area is the design of ultra-fast and high bandwidth on-chip and off-chip communication networks. In specific, optical communication networks have gained a lot of traction in this area in terms of functioning prototypes made by the research community [1]–[3], and devices made by major semiconductor vendors [4]. Optical devices and networks have started to appear in the roadmaps of major processor companies such as Intel [5], and IBM [6].

The natural advantages of optical networks – low latency, high bandwidth, and low power – have been used to create on-board networks at a commercial scale (E.g: Intel's optical PCI Xpress). The field of on-chip networks is being actively worked on, and there are strong predictions that optical networks will begin to be used in on-chip networks by the end of this decade. A lot of work has been done in the field of on-chip optical networks [3], [7] by the computer architecture community. However, the problem of co-designing off-chip and on-chip networks, and creating scalable shared memory based board level systems is relatively new.

The main idea in related work is to co-locate multiple chips(known as *chiplets*) and connect these chiplets together to form a large virtual chip (Galaxy [8]). However, as we shall show in Section II-A, Galaxy has significant issues with respect to scalability, performance, and the number of optical fibers that are required.

To remove these limitations, we propose *NUPLet*, a multi-chip based virtual chip design, which is both scalable and power efficient. *NUPLet* separates the inter chiplet network from the intra chiplet network such that even if we increase the number of chiplets, the intra chiplet network is not affected. Moreover, we propose to implement a non-uniform cache architecture (NUCA) on top of a virtual chip in order to restrict most of the messages to a core's local chiplet and simultaneously increase the overall hit rate of the LLC. For an intra chiplet network we use the *ColdBus* [3] network and the inter chiplet network is based on a Multiple Writer Multiple Reader(MWMR) crossbar. The MWMR based crossbar reduces the number of optical channels by allowing the chiplets to share the available optical channels at the cost of a larger number of ring resonators. In addition, *NUPLet* reduces the static power consumption by predicting the number of inter chip messages and then modulating the laser accordingly.

The main contributions of *NUPLet* are :

- Separation of intra and inter chiplet networks, resulting in a scalable architecture for multi-chip designs.
- Implementation of a NUCA scheme on top of a virtual chip, restricting most of the messages to a single chip and increasing the hit rate of the LLC and thereby increasing the performance of a virtual chip.
- A novel prediction mechanism to reduce power consumption of multi-chip networks.

The rest of the paper is organized as follows. Section II-A describes the background related to silicon photonics, and related work. We motivate the design in Section III, describe the architecture in Section IV, and in Section V we evaluate our proposal. We compare our scheme with *Galaxy* and a scheme proposed by Peter et al. [9]. For a suite of Splash2 [10] and Parsec [11] benchmarks, we demonstrate a speedup of 28% and a 39% reduction in power consumption as compared to state of the art proposals.

II. Background and Related Work

A. Basics of Silicon Photonics

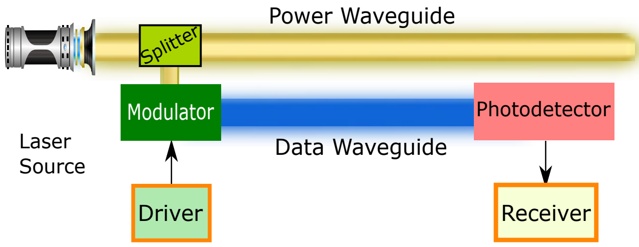

Fig. 1: Basic optical communication framework

1) Optical Components: Figure 1 shows the basic structure of an on-chip optical network. The network requires a light source(laser), a transmitter, an optical channel(waveguide) and a receiver(photodetector).

The **light source** in our *NUPLet* architecture is an off-chip directly modulated laser(DML). These lasers can switch at GHz frequencies [12], are commercially available and have high thermal stability [13]. Each DML laser can generate 180mW of optical power. *NUPLet* uses $\mathscr{N}$ arrays of DML lasers, where $\mathscr{N}$ is the number of chiplets in the design. Each array contains 32 DML lasers (cumulative optical power: 5.75W). We can treat an array with 32 lasers as a virtual power source with 32 power levels (lasers can be individually turned on/off). We further observe that using separate light sources (laser arrays) for each chiplet reduces the power loss and the complexity in delivering the light to different chips as demonstrated in [14], [15].

In *NUPLet*, we use silicon based optical channels known as **waveguides** (for both power and data). In order to divert the power from the silicon waveguides, beam **splitters** are placed at every optical station(transmitter+receiver). Each splitter is associated with a small optical loss, which is a function of its split ratio. In order to decrease the power loss, we need to optimally compute the split ratio of the power splitters. This is because for a large network like *NUPLet*, cascading such splitters in series will result in an exponential increase in the power loss [16]. In order to decrease the loss, *NUPLet* uses the near optimal approach suggested by Peter et al. [16]. This work assumes tunable beam splitters, which can change their split ratio at GHz frequencies. We use one such splitter based on ring resonators [17]. It supports 16 different split ratios, and requires 400 ps to retune.

Furthermore, in order to meet the demands of high bandwidth, *NUPLet* sends multiple wavelengths together through a single waveguide using dense wavelength division multiplexing (DWDM). In such a case, we require a wavelength selective filter at both ends. We use the fast and area-efficient modulator proposed by Xu et al. [18].

2) Network Topologies: There are several ways of creating optical networks. Let us consider an important subset – crossbars. In an SWMR [19] (single writer multiple reader) based crossbar, each optical station (as a writer) is connected to the rest of the stations with its dedicated waveguide. The rest of the stations read a part of the optical signal using beam splitters. This approach is easy to implement; however, there is a chance of excessive power loss in beam splitters (connected in series). It is not necessary to have receivers on all the time. We can have a dedicated waveguide (capacity: 1 bit) to turn a receiver on, and then switch it off after the message is received [19] (known as reservation assisted SWMR). In contrast, in an MWSR [1] crossbar (multiple writer single reader), every receiver has a dedicated waveguide. The rest of the senders can write on that waveguide; however, if two stations wish to transmit simultaneously, then there is a need for arbitration. The combination of SWMR and MWSR – MWMR [2] – allows a station to read and write data from any one of a set of optical channels. This method also requires arbitration. Given the fact that it combines the SWMR and MWSR paradigms, it is considered to be a very flexible option (at the cost of complexity).

3) Non-Uniform Cache Architecture (NUCA): Last level caches (LLCs) are fairly large as of today and the access time is a function of the relative positions of the cache bank and the requesting core. To improve performance, researchers have proposed NUCA schemes [20], where we treat a set of cache banks as a *bank set*. The core first requests one of the banks in the bank set for a line (known as the *home bank*). If the home bank does not have the line, we search the rest of the banks in the bank set for a copy of the line. Subsequently, if the line is found, we return it to the requesting core, and then migrate the line within the bank set to a bank that is *closer* to the requesting core. This reduces the access time for subsequent requests. This is the basic idea. Current implementations of NUCA extend this paradigm to far more complex protocols.

B. Related Work

Let us briefly discuss the *ColdBus* [3], *Galaxy* [8] and *Optical NUCA* [9] projects, which are the closest in terms of related work.

ColdBus [3] is based on the SWMR crossbar. It divides the execution time into fixed size durations called *epochs* and in each epoch it predicts the laser power requirement for the subsequent epoch, and then modulates the laser accordingly. The prediction mechanism is based on the PC of the memory instruction. The predictor tries to predict the incidence of private data cache misses. It uses the history of the memory instruction to predict the epoch in which the actual miss or hit will occur and determines the laser power for that *epoch* accordingly. Moreover, it uses some extra power waveguides that are shared between the optical stations. They are used as a contingency power delivery mechanism.

Galaxy [8] is a chiplet based design, which connects the various chiplets together using optical links. It uses a MWSR based optical crossbar to connect these chiplets. Each chiplet has a set of clusters, which are connected together using silicon waveguides. The number of stations in a cluster depends upon the number of chiplets. In every cluster, there is one representative station for sending messages to a different chiplet.

Since it uses the MWSR based topology, scaling *Galaxy* to a larger number of chiplets will result in a linear increase in the number of optical channels. Moreover, in *Galaxy*, the cluster size inside a single chiplet is based on the number of chiplets connected together. For example, for a 5 chiplet based virtual chip, *Galaxy* requires 320 optical fibers and a cluster of size 4. However, scaling this to 16 chiplets will require 3840 optical fibers and a cluster of size 15. With the increase in cluster size, the Galaxy architecture becomes complex and results in a less efficient intra-chiplet network.

The *Optical NUCA* protocol proposed by Peter et al. [9] creates bank sets (*overlays*) out of banks based on their access frequency. The banks are not necessarily proximate. The authors propose three different protocols – *TSI*, *Broadcast*, and *OP_BCAST* – to search and migrate blocks in an overlay. The main limitations of the proposed protocols is that they are not efficient in the case of multi-chip architectures. The reason is that the cache banks in an overlay may belong to various chiplets and a miss in one bank shall result in sending messages to other banks in the overlay. This will increase the inter chiplet communication and thereby reduce the performance and concomitantly increase power consumption.

NUPLet solves the problems of *Galaxy* by separating the intra and inter chiplet networks. It uses a novel scheme to efficiently utilize the inter chiplet network in order to limit the number of optical channels and increase their power efficiency. In addition, it proposes a NUCA scheme that will be at least as efficient as Optical NUCA [9] with reduced inter chiplet communication.

III. Motivation

In this section, we characterize the behavior of a suite of Splash [10] and Parsec [11] workloads, and motivate the design of *NUPLet* based on the observed patterns. We use the same configurations and the same set of workloads as used in the most related prior work [8], [21].

A. Unbalanced Inter-Chiplet Traffic → MWMR Design

We ran the simulations on a four chiplet based design by assuming 32 4-issue out-of-order cores (see Table I for architectural parameters) per chiplet. We simulated the Galaxy protocol [8], which we treat as a baseline in this paper.

Figure 2 shows the relative number of messages injected by each chiplet into the inter-chiplet network. It is clear from the graph that the chiplets inject varying amount of traffic into the network. As a result, some chiplets require more bandwidth as compared to others. In order to handle this imbalance, we need to allow the stations to share the available optical bandwidth. Thus, we would like to use the more flexible [2] MWMR topology as a baseline for our inter-chiplet network (**Insight 1**).

B. Low Hit Rate in Remote Banks → Use NUCA

A miss in the L1 level (coherence restricted to a single chiplet) results in sending a request to one of the cache banks in the last level cache(LLC). The cache bank that contains the block may lie in the same chiplet from which the request originated or in some other chiplet. If the cache bank is in the same chiplet, we call it the home bank, otherwise it is called a non-home(remote) bank. In Figure 3, we show the relative distribution of LLC access requests due to misses in the private data caches. Nearly, 57% of these requests are sent to non-home banks. In addition, only 7%(mark on the graph) of these result in a hit. Moreover, accessing remote cache banks requires more cycles. Thus, for a smaller hit rate we are sending a very large number of inter chiplet messages resulting in increased power consumption and decreased performance.

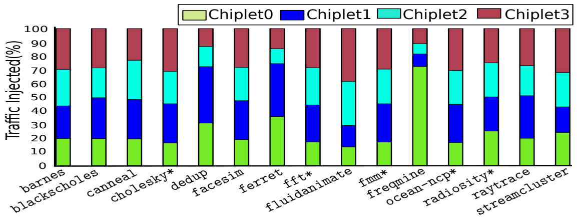

Fig. 2: Chiplet load distribution (* Splash benchmarks)

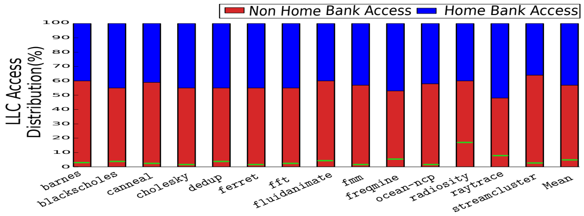

Fig. 3: Comparison of number of accesses and hit rate in the LLC

Due to so many inter chiplet messages and concomitant lower hit rates in non-home banks, the overall performance of the multi-chip design gets restricted. We can remove this disadvantage by maximizing the number of LLC access messages to local (within the same chiplet) cache banks. This can be done by using NUCA schemes, where we can migrate cache blocks to banks that are on the same chiplet as the requesting cores. **Insight 2**

C. Expensive Inter-Chiplet Messages → Predict

The power required to send an inter chiplet message is nearly 6X (calculated using data from Table II) more than the power required for an intra chiplet message. The reason is that the message has to travel through a long inter chiplet waveguide. In addition, there are large input/output coupling losses between inter and intra chiplet waveguides. Let us interpret this result in the broader context of power consumption in optical networks.

In any optical network, we wish to minimize the laser power. This is because if laser power is not used, it will be wasted (there is no way to store optical power). As a result, most of the work [3], [22] in power reduction has focused on predicting the traffic (and resultant laser power) for the next epoch (interval of time) given the behavior of the application in the last few epochs. However, such algorithms were not designed with multiple chips in mind. When we have multi-chip networks such as *NUPLet*, it is necessary to have prediction mechanisms such that we can predict the power required to send inter chiplet messages as well. Since these messages are expensive in terms of power, we need a more accurate predictor, otherwise we might end up wasting a lot of additional power.

Hence, **Insight 3:** we should use a prediction scheme for predicting the number of intra and inter chiplet messages (and the consequent power required) in every *epoch*.

IV. Architecture of *NUPLet*

A. Basic Setup

NUPLet is an electro-optical communication network for a multi-chip design. In this paper, we consider a system with 4 chips (4 sockets on a server), where each chip is called a chiplet. Please note that we consider a specific example in this paper; however, the scope of the idea is generic.

Each chiplet in *NUPLet* is composed of 32 cores. It has an 8 MB last level cache divided into 32 cache banks. Inside each chiplet, we use the ColdBus communication network (see Section II-B). We have a power waveguide that distributes monochromatic light at 1550 nm to each station. When a station wishes to send data, it diverts some portion of the light from the power waveguide. A comb splitter [23] splits this light into 64 different equidistant wavelengths. These wavelengths are modulated and inserted into the data waveguides using ring resonators. The data waveguides carry these wavelengths to the destination station. Moreover, we assume a double pumping strategy in which the data is sent on both edges of the clock [1].

B. On-Chip and Off-Chip Networks

1) Intra Chiplet Network Topology: Each chiplet contains 32 cores and 32 cache banks. We divide the architecture into 16 tiles (clusters), where each tile contains 2 cores, 2 cache banks, and 1 optical station. The stations at the periphery of the chiplet are used for intra chiplet as well as inter chiplet communication and are called inter chiplet optical stations(ICOS). The ICOSs contain two separate message queues for holding intra chiplet and inter chiplet messages. In our design, there are 4 ICOSs per chiplet.

2) Inter chiplet Network Topology: Figure 4(a) shows the four chiplet based example of a *NUPLet* design. The inter chiplet network is based on an MWMR topology(**Insight 1**) in which the stations are allowed to share the available optical channels. We have a total of 16 ICOSs (4 per chiplet) that can use the MWMR network. The inter chiplet MWMR network has 8 data waveguides, and 8 ICOS power waveguides (IPW), where there is a one-to-one mapping between an MWMR data waveguide and an IPW waveguide. An ICOS needs to first get access to a data-IPW waveguide pair, and then it needs to divert power from the IPW waveguide, use a comb splitter (see Section IV-A) to split it into 64 wavelengths, modulate the wavelengths, and transmit the signal on the data waveguide. For arbitration, we have a two pass token stream based arbitration scheme [2], which is fast, efficient, and fair. Figures 4(b) and (c) show the design of a single chiplet and the layout of waveguides (at an ICOS) respectively. As compared to other contemporary architectures [1], [3] our overheads are very modest.

C. NUCA Schemes

In our four chiplet based *NUPLet* design, the LLC (L2 in this case) has 128 cache banks. We propose a non-uniform cache access protocol (NUCA) in order to increase the number of home bank hits, and decrease the number of inter chiplet messages(**Insight 2**). We first start with two baseline schemes called global static NUCA(GS_NUCA) and GD_NUCA (D→ dynamic). The latter is an adaptation of a scheme proposed by Peter et.al. [9]. These two schemes provide some benefits by increasing the hit rate in the LLC, but they have many other overheads, which restrict these schemes to single chip designs. Let us elaborate.

GS_NUCA: We create 32 bank sets with each set containing 4 cache banks, one from each chiplet. Now, when an L1 miss occurs in some chiplet, we extract the 5 LSBs (least significant bits) of the block address and search its corresponding bank in the chiplet. This bank is called the home bank. If there is a miss in the home bank, then the request is broadcast to the other three banks in its bank set (in other chiplets). For sending a broadcast, we send a message to one of the ICOSs. The ICOS forwards these messages to the other chiplets. If there is a miss in all of these banks, then we forward the request to main memory. However, if there is a hit in any of these banks, then the block is migrated to the home bank.

GD_NUCA: This is an adaptation of the scheme proposed by Peter et al. [9] for our setting. In this scheme, we make bank sets called *overlays* based on the access frequencies of the banks. Whenever, the program detects a phase change, the overlays are re-arranged based on the updated access frequencies. There are elaborate protocols to manage the overheads of this process. In our adaptation, we create and manage overlays in a similar manner. We first locate the home bank based on the 7 LSBs of the block address, and if there is a miss, we search the other banks in its overlay.

In the original paper, this scheme was shown to be much more efficient than state of the art schemes that are used for electrical networks. However, our main issues with this scheme is that it has not been designed with multiple chips in mind. As a result inter chip latency, bandwidth, and power become important bottlenecks.

D. NUPLet

1) Overview: We have two main contributions: a novel NUCA scheme, and a novel prediction scheme for predicting the power usage of our inter chiplet network.

2) LLC Access Scheme: To describe the LLC access scheme used in *NUPLet*, let us first briefly explain two important tables used to implement this scheme.

Block History Table(BHT): The BHT is indexed using the last 20 bits of the block address. Given a block we map it to an entry in the BHT based on its address. Each entry contains 2 bits, which indicate the id of the chiplet that accessed the block the last time. Note that for 4 chiplets, we require 2 bits to uniquely identify them. Now, for the BHT, we require an

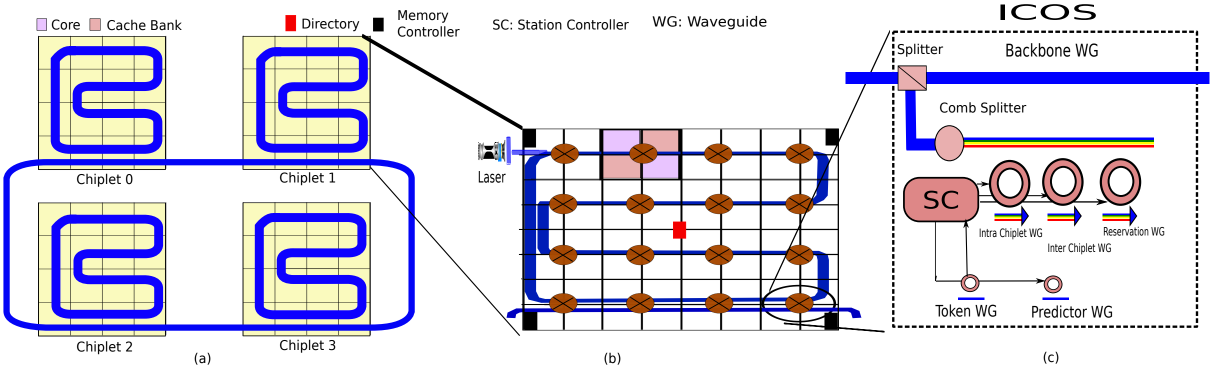

Fig. 4: NUPLet Architecture

additional 256 KB of space (64KB per chiplet). The table is distributed across the four chiplets (address space of blocks divided uniformly). The area overhead is less than 0.5% and can be accessed in a single cycle (if the respective BHT portion is in the same chiplet) (area and access time found using Cacti-5.3 [24]).

Chiplet Affinity Table(CAT): The CAT table is indexed by the last 10 bits of the block address akin to the BHT. Each entry has 8 bits. We have 2 bits per chiplet. These 2 bits are indicative of the probability (level of confidence) that the block is in that chiplet. Every optical station has a CAT table, which requires an additional 1KB of space.

Let us now elaborate on the flow of actions, after a miss at the L1 level, which uses directory based cache coherence (see Figure 5), the home bank is found out using the 5 LSBs of the block address. Then we send a request to the optical station of the cluster containing the home bank (home bank cluster), and a request to simultaneously access the CAT.

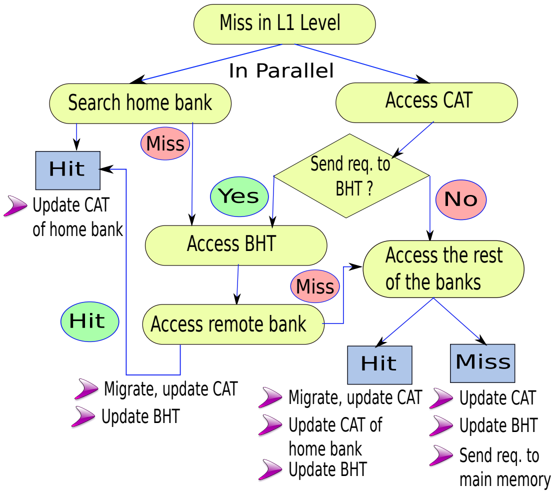

Fig. 5: LLC access flowchart

We read the value of the confidence bits from the CAT for the corresponding entry. These bits tell us if we need to access the BHT to predict the chiplet that might have a copy of the block. In addition, they are used to choose the chiplet in which the evicted cache block should be migrated. Let us suppose that there is a hit for a cache block in a bank in chiplet 1, then the value of the confidence bits corresponding to chiplet 1 for the respective cache block is incremented at the respective cache bank's CAT table. In addition, during eviction and migration of a cache block, we decrement the value of the confidence bits corresponding to the chiplet from which the cache block is migrated (or evicted) and the value of the confidence bits is incremented for the destination chiplet.

A request is not sent to the BHT if the value of the confidence bits for the current chiplet is non-zero and is zero for other chiplets. Otherwise, we send a request to the BHT. Subsequently, we wait for a hit/miss decision. If there is a cache hit, we send the block to the core, otherwise we initiate an inter chiplet message. On a miss, we wait for the value from the BHT. Let us say that we have a miss in chiplet 0, and the BHT indicates that the block might be in chiplet 1. We send a message to chiplet 1. If there is a miss, we send messages to chiplets 2 and 3. Next, the block migrates from a remote chiplet to the current chiplet to increase locality. The BHT is simultaneously updated.

During evictions, the cache block is sent to main memory only if the value of the confidence bits is 00 for every chiplet in the CAT table entry, otherwise the evicted block is migrated to a cache bank in the chiplet with the largest confidence. Moreover, during the migration or eviction of a cache block, the confidence values related to that block are sent to the destination bank's CAT table, which replaces its own entry.

3) Prediction Scheme: In *NUPLet*, we send messages – destined for other chiplets – to ICOS stations, which subsequently send them to other chiplets after arbitrating for an MWMR waveguide. The power is sourced from IPW waveguides (see Section IV-B). In this context, note that the power from the off-chip laser is split into two parts – one part goes to the power waveguides used by ColdBus and the other part is distributed to the IPWs. In addition, we have two other waveguides – inter chiplet token waveguide(ICTW) and inter chiplet prediction waveguide(ICPW), which run in parallel to the intra chiplet data waveguides. The number of tokens carried by the ICTW is the same as the number of IPWs carrying optical power. It is necessary to grab a token before an ICOS can send data. Finally, at the end of every epoch, the optical stations send their usage data to the laser controller via the ICPW.

We try to predict the number of inter chiplet messages that will be sent in the next epoch and accordingly increase or decrease the number of tokens in the ICTW by modulating the off-chip laser. The whole scheme works as follows.

At the end of every epoch, each optical station sends a 2-bit value to the laser controller indicating the number of requests sent by a station to an ICOS in the current epoch. In addition, the ICOSs send the mean number of pending messages (divided by 2) in their message queues to the laser controller (4 bits). The laser controller adds 16 2-bit values from 16 different stations to calculate the sum(C). In the meanwhile, the four 4-bit values from ICOSs are added to get the total number of pending messages(P).

We have a 64 entry table called the Power Request Table(PRT), which is indexed by the 6-bit calculated sum(C) and initialized by the $Maximum\_Number\_of\_Tokens/2$. Each entry stores a 4-bit number. In addition, the prediction mechanism starts with the second epoch. Let the sum for the previous epoch be C, and for the current epoch be C'. We need to compute the number of tokens (T) for the next epoch, as a function of C, $PRT[C]$, C', and P (sum of pending messages). The intuition here is that we increase the number of tokens when either the sum is high or there are too many pending messages (and vice versa).

$$T = \begin{cases} PRT[C]-1 & (C' < C) \wedge (16 \leq P < 32) \\ PRT[C] & (C' \geq C) \wedge (16 \leq P < 32) \\ PRT[C]-1 & (P < 16), PRT[C] \leftarrow PRT[C]-1 \\ PRT[C]+1 & (P \geq 32), PRT[C] \leftarrow PRT[C]+1 \end{cases}$$

After adding the power requirements (*NUPLet* + ColdBus), the laser controller calculates the total power requirement, and split ratios of the splitters, which partition the incoming optical power among power and IPW waveguides. The splitters and the lasers are subsequently retuned. We assume similar latencies as Peter et al. did for ColdBus [3]: 2 cycles for collating the results at each station, 3 cycles for sending it to the on-chip laser controller, 2 cycles for computing the new configuration (using fast lookup tables), 5 cycles for broadcasting new split ratios of splitters and retuning, and (in parallel) 6 cycles for sending a message to the laser and retuning it.

V. Evaluation

A. Experimental Setup

We compared the performance (inversely proportional to simulated execution time) and energy consumption of *NUPLet* vis-a-vis state of the art proposals: Galaxy and GD_NUCA. We also consider a more simplistic baseline, Non-NUCA, where we assume that we are not using any NUCA proposal. The address space is equally partitioned across all the LLC banks (across the chiplets) based on the 7 LSB bits of the block address.

The architectural parameters are shown in Table I. We use four 32-core chiplets in all our designs. We evaluate all our designs on a cycle architectural simulator, Tejas [26]. It has been rigorously validated against native hardware.

Parameter	**Value**	**Parameter**	**Value**
No. of Chiplets	4	Cores per chiplet	32
Frequency	2.5 GHz	Technology	22 nm
Processor Core			
Pipeline	Four-issue out-of-order	IW size	54
iTLB	128 entry	dTLB	128 entry
Private L1 i-cache, d-cache			
Write-mode	Write-back	Block size	64 bytes
Associativity	4	Size	32 kB
Latency	2 cycles	MSHRs	32
Directory	fully mapped distributed MOESI, 2MB		
Shared L2(Per Chiplet)			
Write-mode	Write-back	Block size	64 bytes
Associativity	4	# banks	32
Latency (per bank)	8 cycles	Bank size	256 KB
Main Memory			
Latency	80 cycles (optically connected)	Mem. controllers	4
Queue Sizes			
Station Queue	16	ICOS Queue	32

TABLE I: Simulation parameters (also see [21])

Optical Parameters	
Wavelength (λ)	$1.55\mu m$
Width of waveguide (W_g)	$3\mu m$
Slab height	$1\mu m$
Rib height	$3\mu m$
Refractive Index of $SiO_2(n_r)$	1.46
Refractive Index of Si (n_c)	3.45
Input Driver Power	76 μW
Insertion Coupling Loss	50%
Photodetector minimum power	36 μW
Combined transmitter and receiver delay	180-270 ps
Optical propagation delay	7 ps/mm
Electrical propagation delay	35 ps/mm
Bending Loss	1 dB
Waveguide Loss	1 dB
Coupler Loss	1 dB
Photodetector	0.1 dB
Splitter Loss	0.36 dB
Laser wall plug efficiency	30%
Micro-heater power	21.6 μW/°C

TABLE II: Optical simulation parameters [14], [25]

Tejas uses the Orion-2 [27] and McPat [28] tools, for generating the power consumption values for the NoC and cores/caches respectively. We use programs from the Splash2 [10] and Parsec [11] benchmark suites to evaluate our design (same workloads and configurations as [9]). For all our experiments, we use an epoch size of 100 cycles.

Table II shows the various optical parameters considered in our design. The trimming power is calculated on the basis of thermal simulations carried out using Ansys Icepak, where we assume that each ring resonator needs to have a temperature of 80°C.

B. Hit Rate Comparison and Non-Home Bank Accesses

Columns A and B of Table III show the increase in home bank hits and the mean hit rates of GS_NUCA, GD_NUCA and *NUPLet* as compared to Non-NUCA. To summarize, *NUPLet* resulted in an 81.53% increase in home bank hits as compared to Non-NUCA. In addition, the proposed NUCA scheme resulted in an overall increase in hit rate of the LLC by 69.31%. However, GD_NUCA is the best scheme and increases the hit rate by 10% as compared to *NUPLet* (mainly because of the extra messages + overhead).

Column D of Table III shows the percentage reduction in non-home bank accesses due to *NUPLet* as compared to

Benchmarks	Increase in Home Bank Hits (%) (A)			Increase in Total Hit Rate (%) (B)			Prediction Accuracy (%) (C)	Decrease in Non-home Bank Accesses(%)(D)	Avg. # of inter chiplet msgs/epoch due to cache miss(E)		Decrease in wait time due to NUPLet(F)
	GS_NUCA	GD_NUCA	NUPLet	GS_NUCA	GD_NUCA	NUPLet			NUPLet	GD_NUCA	
barnes	62.4	80.32	68.1	73.42	121.31	68.56	94.97	43	41.26	72.4	16.32
blackscholes	120.5	145.74	119.4	72.34	149.47	74.15	91.41	32	64.86	95.38	12.65
canneal	49.53	60.17	43.86	53.27	69.23	51.45	93.82	38	75.1	121.13	24.31
cholesky*	128.65	157.43	128.57	48.32	69.87	64.38	99.25	40	69.86	116.43	22.91
dedup	110.22	61.34	108.32	78.45	32.68	73.44	91.38	16	74.6	88.8	5.32
facesim	87.43	102.31	83.76	83.57	104.11	92.67	98.64	43	59.13	103.74	27.38
ferret	67.31	79.23	68.72	88.54	102.54	82.42	94.23	47	6.27	11.83	8.48
fft*	35.54	48.34	33.46	5.3	7.4	5.1	97.26	47	39.84	72.5	27.92
fluidanimate	28.23	31.67	25.76	2.5	8.3	1.6	82.45	18	22.95	28	9.25
fmm*	60.51	91.38	53.87	53.65	68.44	50.48	99.22	40	86.56	144.27	7.95
freqmine	102.45	112.32	103.54	79.27	86.54	73.33	92.43	53	0.12	0.26	16.47
ocean_ncp*	110.34	157.72	105.87	65.76	79.21	98.48	99.23	45	58.05	105.54	12.84
radiosity*	132.65	152.34	127.85	50.46	86.13	92.99	92.65	29	29.59	41.67	3.11
raytrace	8.1	26.81	7.62	62.51	70.92	73.12	77.43	37	47.91	76.05	4.25
streamcluster	145.38	196.32	144.21	102.34	133.45	137.47	94.59	30	19.99	28.56	27.25
Mean	83.28	100.23	81.53	61.31	79.31	69.31	93.26	37.2	46.41	73.8	15.1

TABLE III: Analysis of LLC hit rates and prediction accuracies (* Splash benchmarks)

GD_NUCA. Using *NUPLet* results in a 37.2% reduction in non-home bank accesses, which results in decreased static power consumption(see Section V-F), and increased performance.

C. Prediction Accuracy

Column C of Table III shows the accuracy of our prediction mechanism in predicting the number of inter chiplet messages. We infer a correct prediction, when we can send all our messages in a given epoch without requiring extra power. The accuracy is high (93%) for all the benchmarks other than *raytrace*, mainly because of the irregular nature of its memory accesses.

D. Contention at ICOSs

Column E of Table III shows the average number of inter chiplet messages per epoch due to cache misses in home banks. *NUPLet* has 37% less inter-chiplet messages as compared to GD_NUCA. Column F shows the percentage reduction in wait time at the ICOSs due to the use of the *NUPLet* scheme as compared to the GD_NUCA scheme. The reason for reduction in wait times is the lower number of inter chiplet messages as compared to GD_NUCA.

E. Performance Comparison

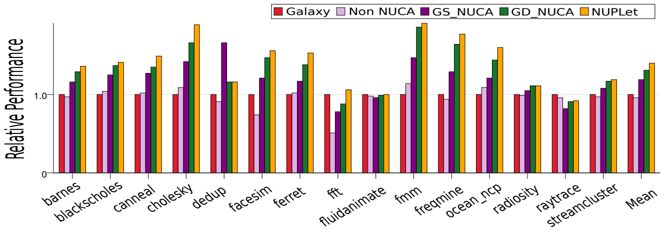

Fig. 6: Performance comparison

Figure 6 shows the relative performance(reciprocal of simulated execution time) of different designs. The values are normalized to the Galaxy scheme. It is clear from the graph that our scheme, *NUPLet*, is the best scheme with 6% better performance as compared to the nearest competitor, GD_NUCA. GD_NUCA is the second best and performs 23% better as compared to *Galaxy*. The better performance of *NUPLet* is due to the increase in hit rate in the LLC, decrease in contention, and the ability of ICOSs to send multiple messages at a time because of the shared inter chiplet interconnect (Columns A and F in Table III). In addition, the performance improvement in *NUPLet* as compared to GD_NUCA is because of 31% reduction in non home bank accesses, and a decrease in inter chiplet traffic. *Galaxy* performs 4% better than Non-NUCA because of the availability of optical power all the time.

In the case of benchmarks such as *cholesky*, *fmm*, and *ocean_ncp*, the high speedup is due to higher hit rates. However, the lower prediction accuracy is responsible for the decreased performance in the case of *raytrace*. Moreover, the dip in performance in the case of *fft* and *fluidanimate* is due to the lower hit rate in the LLC in these two benchmarks (results less sensitive to NUCA and prediction schemes).

F. Laser Power Consumption

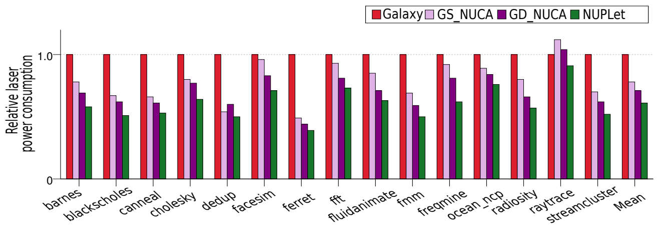

Fig. 7: Relative laser power comparison

This section compares the laser power consumed by various configurations. In the case of *Galaxy* the laser supplies optical power all the time making it the most power consuming design. In the case of *NUPLet*, we are modulating the laser based on our prediction scheme described in Section IV-D3. Due to laser modulation, we decrease the total laser power consumption. Figure 7 compares the relative laser power consumed by different configurations. *NUPLet* is the best scheme consuming 39% less power as compared to *Galaxy*.

The decrease in laser power consumption in the case of *NUPLet* is because of our shared inter chiplet network, higher prediction accuracies leading to better modulation of off-chip lasers, and a decrease in non-home bank accesses (Column

D of Table III). Our method of sharing MWMR waveguides, allows the ICOS to share the available optical power resulting in an increase in overall power utilization.

G. ED^2 Comparison

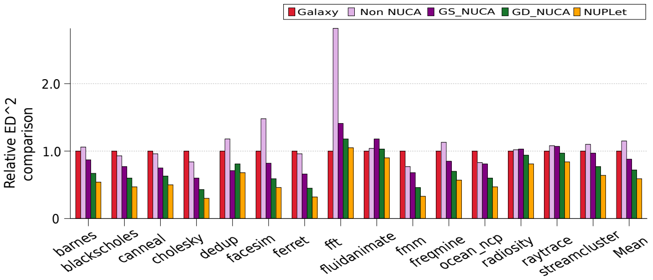

Fig. 8: ED^2 comparison

The *energy_delay*^{2} product is the metric of choice for comparing designs that consume different amounts of power and have varying performance. Here, the term *energy* refers to the full system energy that includes the energy consumed by cores and lasers. The ED^2 values of different configurations are plotted in Figure 8. The values are normalized to *Galaxy*. *NUPLet* is the best scheme with a 41% reduction in ED^2 as compared to *Galaxy*. The higher reduction in ED^2 values in the case of *NUPLet* is attributed to its higher performance and concomitant lower laser power consumption.

In *fft* and *fluidanimate*, the relatively higher values of ED^2 are due to the lower performance with *NUPLet*. However, in all other cases, the ED^2 values in *NUPLet* are much lower than those in *Galaxy*. GS_NUCA has higher ED^2 values as compared to GD_NUCA because of its lower performance (the only exception being *dedup*).

VI. Conclusion

In this paper, we proposed a multi-chip architecture, *NUPLet*, which tries to address the issue of scalability in multi-chip designs by separating the intra and inter chiplet networks. We decrease inter chiplet communication by creating spatial locality using a novel NUCA scheme. The proposed NUCA scheme increases the hit rate in home banks, and drastically reduces non-home bank accesses. This reduces inter chiplet traffic to a large extent. We tackle the problem of high power consuming inter chip messages by predicting inter chiplet traffic accurately, and using this information to modulate off-chip lasers. Using such novel techniques, we were able to decrease the static power consumption by 39% with a 28% increase in performance as compared to state of the art proposals. Moreover, we reduced ED^2 by 16% as compared to the nearest competitor, GD_NUCA.

References

[1] 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 *ISCA*, 2008.

[2] Y. Pan, J. Kim, and G. Memik, "Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar," in *HPCA*, 2010.

[3] E. Peter, A. Thomas, A. Dhawan, and S. R. Sarangi, "Coldbus: A near-optimal power efficient optical bus," in *HiPC*, 2015.

[4] Sicoya. Fully integrated silicon photonics transceiver chips. Http://sicoya.com/.

[5] M. T. Review. (2008) A record-breaking optical chip. Https://www.technologyreview.com/s/410383/a-record-breaking-optical-chip/.

[6] IBM. (2012) Ibm lights up silicon chips to tackle big data. Http://www-03.ibm.com/press/us/en/pressrelease/39641.wss.

[7] I. O'Connor, "Optical solutions for system-level interconnect," in *SLIP*, 2004.

[8] Y. Demir, Y. Pan, S. Song, N. Hardavellas, J. Kim, and G. Memik, "Galaxy: A high-performance energy-efficient multi-chip architecture using photonic interconnects," in *ICS*, 2014.

[9] E. Peter, A. Arora, J. Bashir, A. Bagaria, and S. R. Sarangi, "Optical overlay nuca: A high-speed substrate for shared l2 caches," *J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 4, May 2017.

[10] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The splash-2 programs: characterization and methodological considerations," *SIGARCH Comput. Archit. News*, vol. 23, pp. 24–36, May 1995.

[11] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: characterization and architectural implications," in *PACT*, 2008.

[12] M. Faugeron, M. Tran, F. Lelarge, M. Chtioui, Y. Robert, E. Vinet, A. Enard, J. Jacquet, and F. Van Dijk, "High-power, low rin 1.55-directly modulated dfb lasers for analog signal transmission," *Photonics Technology Letters*, vol. 24, no. 2, pp. 116–118, 2012.

[13] J.-S. Huang, H. Lu, and H. Su, "Ultra-high power, low rin and narrow linewidth lasers for 1550nm dwdm 100km long-haul fiber optic link," in *IEEE Lasers and Electro-Optics Society, 2008. LEOS 2008. 21st Annual Meeting of the*, Nov 2008, pp. 894–895.

[14] R. Morris, E. Jolley, and A. K. Kodi, "Extending the performance and energy-efficiency of shared memory multicores with nanophotonic technology," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 25, no. 1, pp. 83–92, 2014.

[15] R. Morris and A. K. Kodi, "Exploring the design of 64-and 256-core power efficient nanophotonic interconnect," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 16, no. 5, pp. 1386–1393, 2010.

[16] E. Peter and S. R. Sarangi, "Optimal power efficient photonic swmr buses," in *Silicon Photonics (with HiPEAC)*, 2015.

[17] E. Peter, A. Thomas, A. Dhawan, and S. R. Sarangi, "Active microring based tunable optical power splitters," *Optics Communications*, 2016.

[18] S. P. M. L. Qianfan Xu, Bradley Schmidt, "Micrometre-scale silicon electro-optic modulator," *Nature*, May 2005.

[19] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary, "Firefly: illuminating future network-on-chip with nanophotonics," in *ACM SIGARCH Computer Architecture News*. ACM, 2009.

[20] K. Changkyu, D. Burger, and S. Keckler, "Nonuniform cache architectures for wire-delay dominated on-chip caches," *Micro, IEEE*, 2003.

[21] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal, "Atac: a 1000-core cache-coherent processor with on-chip optical network," in *PACT*, 2010.

[22] L. Zhou and A. K. Kodi, "Probe: Prediction-based optical bandwidth scaling for energy-efficient nocs," in *NOCS*, 2013.

[23] J. S. Levy, Y. Okawachi, M. Lipson, A. L. Gaeta, and K. Saha, "High-performance silicon-based multiple wavelength source," in *CLEO: Science and Innovations*, 2011.

[24] S. Thoziyoor, N. Muralimanohar, J. Ahn, and N. Jouppi, "Cacti 5.3," *HP Laboratories, Palo Alto, CA*, 2008.

[25] G. T. Reed, *Silicon Photonics: The State of the Art.* John Wiley & Sons, 2008.

[26] S. R. Sarangi, K. Rajshekar, K. Prathmesh, G. Seep, and P. Eldhose, "Tejas: A java based versatile micro-architectural simulator,," in *PATMOS*, 2015.

[27] A. B. Kahng, B. Li, L.-S. Peh, and K. Samadi, "Orion 2.0: A fast and accurate noc power and area model for early-stage design space exploration," in *DATE*, 2009.

[28] 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 *MICRO*, 2009.