Waferscale Network Switches

Shuangliang Chen
UIUC

Saptadeep Pal
Etched AI

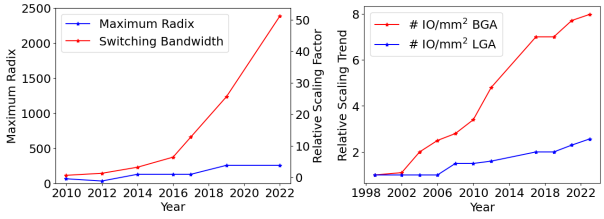
Rakesh Kumar
UIUC

Abstract*—In spite of being a key determinant of latency, cost, power, space, and capability of modern computer systems, network switch radix has not seen much growth over the years due to poor scaling of off-chip IO pitches and switch die sizes. We consider waferscale integration (WSI) as a way to increase the size of the switch substrate to be much bigger than a single die and ask the question: can we use WSI to enable network switches that have dramatically higher radix than today's switches? We show that while a waferscale network switch can support up to 32x higher radix than state-of-the-art network switches when only area constraints are considered, the actual radix of a waferscale network switch is not area-limited. Rather, it is limited by a combination of internal bandwidth, external bandwidth, and power density. In fact, without optimizations, benefits of a waferscale network switch are minimal. To address the scalability bottlenecks, we propose a heterogeneous network switch design that reduces switch power by 30.8%-33.5% which, in turn, allows an increase in radix (by up to 4x) by increasing internal I/O bandwidth at the expense of energy efficiency. We also propose subswitch deradixing that increases the overall radix by 2x by decreasing the radix of the subswitches to alleviate the internal I/O bottleneck. We use Area I/O and Optical I/O schemes to alleviate the external I/O bandwidth bottlenecks of conventional SerDes-based external connectivity. In addition to scalability optimization, we present optimizations such as low latency buffering and proprietary routing that improve the performance of waferscale switches. Finally, we present a system architecture for a waferscale network switch that supports its port count, power delivery, and cooling requirements in a compact form factor. We show that the switch can be used to enable new computing systems such as single-switch datacenters and massive-scale singular GPUs. It can also lead to a dramatic reduction in datacenter network costs. Overall, this is the first work quantifying the benefits of waferscale switches and identifying and addressing the unique challenges and opportunities in building them.

I. Introduction

The importance of network switch radix - the number of bidirectional ports soldered onto a switch ASIC - in modern computing infrastructure cannot be overstated. In the context of datacenters, switch radix dictates the depth of the network topology and, therefore, the latency of communication between nodes [38]. It also dictates the number of switching devices and optical links required for connections and, therefore, the cost, power, and space requirements of the network [39]. In the context of high-performance computing systems, the switch radix can dictate the size (amount of memory and compute) and performance of the system that can be built. In the context of dedicated AI/ML training clusters (e.g., DGX GH200 [8]), it can even dictate the size of the largest models that can be trained (by dictating the number of GPUs and their interconnection bandwidth in a cluster, for example).

Despite this, network switch radix has not seen much growth over the years - the maximum radix has increased only by 8x over the last 12 years (Figure 1.a). This relative anemic growth (compared to total bandwidth, for example) is due to poor scaling of off-chip IO pitches [48] as well as poor scaling of switch die sizes [58]. Figure 1.b shows that the density of connections (number of I/Os/mm^2) for the ball grid array (BGA) and land grid array (LGA) technologies has increased only by 8x and 2.6x respectively over the last 24 years. Switch die sizes scale even slower since the maximum chip size is dictated by reticle limits during lithography. The biggest monolithic chip for the $5\mathrm{nm}$ node, for example, does not exceed 858 mm^2 [53].



(a) (b)

Fig. 1: (a) Scaling of radix and total switching bandwidth during 2010-2022. (b) Number of IO pins per mm^2 from 1999 to 2023 for BGA and LGA packaging technologies [15].

Since I/O pitches are notoriously difficult to scale [48], any large increase in network switch radix must come from a new approach to increasing the size of the switch substrate to be much bigger than a single die.

The last few years have seen the introduction and commercialization of waferscale integration (WSI) technologies that can support computing systems to be built on substrates as large as a wafer [26], [33], [40], [59] (Table I). This manifold increase in system size over a reticle-limited die has enabled orders of magnitude higher computing performance [40], [49], [59]. However, no prior work has looked at building a network switch at waferscale. In this work, we ask the question: can we use WSI to enable network switches that have dramatically higher radix than today's switches?

Technology	**Silicon Interposer [42]**	**Si-IF [33]**	**InFO-SoW [26]**
I/O Pitch ($\mu\mathrm{m}$)	3-10	2-10	80-150
Interconnect Wire Pitch ($\mu\mathrm{m}$)	2-10	1-10	20
Maximum Sizes/Dies	8.5 cm^2	Full Wafer	Full Wafer
Inter-die Distance (mm)	50	1-100	10-50
BW Density ($\mathrm{Gbps/mm/layer}$)	1000	800-1600	1000-3200
Energy/b ($\mathrm{pJ/bit}$)	0.25	0.06-4	1.5-3
Latency (ns)	0.1	0.04-10	10-15

TABLE I: Technologies that can enable chiplet-based waferscale integration (vs a silicon interposer).

To keep yield and design complexity manageable, we study waferscale integration of Tomahawk 5 (TH-5)-like chips in-

Total Power (W)	500 [13]	**Configuration 1**	256x200Gbps
w/o I/O Power (W)	400	**Configuration 2**	128x400Gbps
Area (mm^2)	800 [13]	**Configuration 3**	64x800Gbps

TABLE II: Tomahawk 5 (TH-5) parameters used in this work.

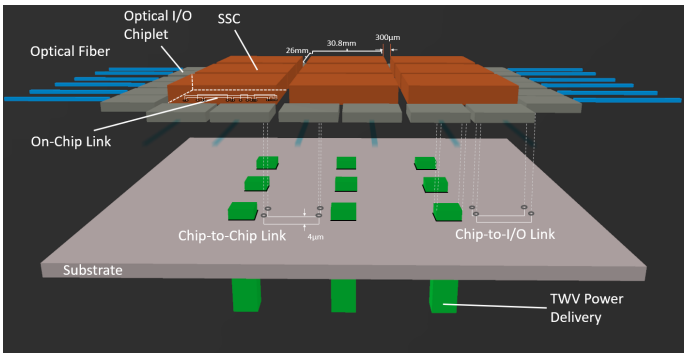


Fig. 2: A high-level diagram of a 3x3 waferscale network switch architecture with Optical I/O chiplets. Chip-to-chip links and chip-to-I/O links are passive interposer traces embedded in the substrate with a pitch size of 4 µm. Optical signals from optical fibers are converted to electrical signals and forwarded to SSCs through chip-to-I/O links. Connections between SSCs are accomplished using on-chip links and chip-to-chip links. Power delivery to each SSC is done with through-wafer-vias.

terconnected in a Clos topology to enable a higher radix switch. TH-5 [13] is a state-of-the-art high radix network switch from Broadcom that supports configurations listed in Table II and has a reported power (area) of 500W ($800mm^2$). We derive the non-I/O power of TH-5 by assuming an I/O energy consumption of 2 $\mathrm{pJ/bit}$ [10]. A large number of $800\mathrm{mm}^2$ TH-5-like chips (*sub-switch chiplets or SSCs*) can fit on a wafer, potentially enabling a high radix switch. Figure 2 shows a high-level diagram of such device.

We find very quickly that the radix of a waferscale network switch is limited not by the area of the substrate, but rather by the internal bandwidth of connections between the SSCs (Section IV). A large number of logical connections is needed to support a Clos topology, especially at higher levels, saturating the physical bandwidth density available between adjacent SSCs. We also find that external bandwidth may become a bottleneck when using the conventional SerDes-based external connectivity using wafer periphery. Finally, power density may become a bottleneck and limit the radix that can be supported for a given cooling mechanism.

To address the power density bottleneck, we propose a *heterogeneous network switch* (Section V). Instead of all-identical SSCs, we replace some of the SSCs in the Clos with lower-radix SSCs. This can lead to significant power reduction since the power of an SSC decreases superlinearly with its radix. This power reduction allows an increase in radix by increasing internal I/O bandwidth at the expense of energy efficiency. To address the internal bandwidth bottleneck, we propose *subswitch deradixing*. In this optimization, we *decrease* the radix supported by each sub-switch without reducing the area of the SSC. This increases the number of I/Os per port for the subswitch and, therefore, increases the overall switch radix. To address the external bandwidth bottleneck, we use Area I/O (where a mezzanine PCB allows external bandwidth I/O to scale with substrate area) and Optical I/O (which uses on-substrate chiplets for electrical-to-optical and optical-to-electrical I/O conversion) schemes (Section III). We quantify the impact of our optimizations on switch radix size and power. We also quantify the performance improvements possible from a waferscale switch as well the new opportunities for switch microarchitecture that it enables (Section VI).

A waferscale network switch also presents a unique system architecture challenge. Compactly supporting a switch with one to two orders of magnitude higher port count while meeting the extraordinary power delivery and cooling requirements require careful design. We present a system architecture of a waferscale network switch (Section VIII) that takes up no more than 20RU of space while supporting the power delivery and cooling needs for the switch. We then use this system architecture to demonstrate the end-to-end benefits of a waferscale network switch. We show that such a switch can enable *single-switch* data centers with up to 8192 servers. It can also enable a singular GPU that is eight times larger than the largest singular GPU today built using an NVSwitch network. It can also significantly reduce the number of switches and optical cables in a hyperscale datacenter network by 3x-10x.

This paper makes the following key contributions:

- We show that a waferscale network switch can support up to 32x higher radix than state-of-the-art network switches when only area constraints are considered.
- We show that the actual radix of a waferscale network switch is not area-limited. Rather, it is limited by a combination of internal bandwidth, external bandwidth, and power density. Without optimizations, the benefits of a waferscale network switch are minimal.
- We propose a heterogeneous network switch design that reduces switch power by 30.8%-33.5% which, in turn, allows an increase in radix (by up to 4x) by increasing internal I/O bandwidth at the expense of energy efficiency. We propose subswitch deradixing that increases the overall radix by 2x by decreasing the radix of the subswitches to alleviate the internal I/O bottleneck. We quantify the effectiveness of Area I/O and Optical I/O schemes in alleviating the external I/O bandwidth bottleneck caused when the wafer periphery is used for conventional SerDes-based external connectivity.
- We present a system architecture for a waferscale network switch that supports its port count, power delivery, and cooling requirements in a compact form factor. We show that the switch can be used to enable new computing systems such as single-switch datacenters and massive-scale singular GPUs. It can also lead to a dramatic reduction in datacenter network costs.
- We show that a waferscale switch opens up opportunities for performance and microarchitectural optimizations, including low latency buffering and proprietary routing. Overall, this is the first work quantifying the benefits of waferscale switches and identifying and addressing the

unique challenges and opportunities in building them.

II. Related Work

There has been a large body of work on enabling higher radix through scalable switch microarchitectures [18], [19], [22], [38], [43]. The most related work constructs high-radix switches using lower radix-switches [19], [38]. Ahn *et al.* [19], for example, describe a network-within-network approach to create high-radix switches by interconnecting subswitches. Chrysos *et al.* implement a Scalable Clos-On-Chip (SCOC) that achieves a radix of 136 [25]. Liang *et al.* [43] demonstrate a switch with an overall radix of 144 by utilizing 12 radix-36 switch chiplets.

Our work also builds higher radix switches using lower radix switches. However, the scale of our switches is very different. We use waferscale integration to enable radix that is up to 32x higher than previous works. Also, the large scale of our switch creates new bottlenecks for us to address - internal bandwidth, external bandwidth, and power density. We propose new optimizations such as a heterogeneous network switch, subswitch de-radixing, and Area and Optical I/Os to address these bottlenecks.

Router Name	[17]	[12]	[7]	WS (300mm)	WS (200mm)
Space (RU)	16	21	15.8	20	11
Total Bandwidth (Tb/s)	115.2	230.4	115.2	1638.4	819.2
Port Count (w/ 200Gbps)	576	1152	576	8192	4096
Total Power (kW)	11.2	25.9	11.0	50	25
Power / Port (W)	19.4	22.5	19.1	6.1	6.1
Capacity Density (Tbps/RU)	7.2	11	7.5	81.9	74.5

TABLE III: Modular switches vs waferscale switches (WS).

It is possible to support very high radix using optical switches [44], [50], [55]. However, our work achieves very high radix using an alternative approach that is arguably much lower cost since it leverages the economies of scale of the existing semiconductor supply chains. Also, we can easily support packet switching and, therefore, can be deployed widely, unlike optical switches which mostly do not support packet switching [23] and, therefore, see limited deployment [14], [50].

It is also possible to support high radix through modular switches [7], [12], [17] with commercial products already supporting 576-1152 ports at 200Gbps per port. Modular switches typically consist of several 1U (rack unit) line cards, switched internally by fabric modules. These line cards and fabric modules are essentially switch boxes, subject to the same radix scaling constraints as discussed in Section I. They also have high power consumption per port, making it increasingly difficult for their radix to scale [47]. Waferscale integration allows our switches to have 7.1x-14.2x more ports in a 300mm substrate or 3.6x-7.1x more ports in a 200mm substrate, while being 68%-72.9% more energy efficient for both substrate sizes, and having much higher capacity density (7.5x - 11.4x denser with 300mm waferscale switch, 6.79x - 10.3x denser with 200mm waferscale switch). Table III provides a comparison[1].

[1]Multiple configurations exist for commercial modular switches; 200Gbps per port was chosen for comparisons.

III. Exploiting Waferscale Integration to Build Very High Radix Switches

A. Waferscale Integration (WSI)

Waferscale integration (WSI) allows us to build integrated circuits at the scale of an entire silicon wafer. Unlike conventional approaches where larger systems are built by interconnecting separately packaged chips using relatively low bandwidth and high energy interconnects, WSI allows us to build extremely large and tightly integrated systems [40], [45], [49], [59]. Today's WSI technologies can be classified into two categories: monolithic and chiplet-based.

Monolithic waferscale integration uses the entire traditionally manufactured wafer as a chip instead of dicing the wafer into dies. 1970's Trilogy systems' architecture [45] and 2020's Cerebras' waferscale engine (WSE) [40] are examples of monolithic WSI systems. In monolithic WSI, the wafer is manufactured using a conventional reticle step-and-repeat process, but unlike a traditional chip-building process where the reticles are diced into smaller dies, here a post-processing step is performed where inter-reticle interconnects are built on the wafer. This allows one to architect a large tightly coupled waferscale system. There are a few challenges this technology faces: (a) since wafer manufacturing encounters defects and parametric variation, monolithic waferscale architectures need to carefully build in redundancy to minimize yield loss [40], (b) since a wafer is built using a given technology process/node and also uses step-and-repeat process, it only allows for homogeneous architectures (e.g., DRAM and logic or transistors from different nodes which are optimized for different functionalities, logic vs I/O are not allowed).

Alternatively, **chiplet-based** waferscale integration uses a waferscale substrate to tightly integrate and interconnect pre-verified dies/chiplets. Silicon interconnect fabric (Si-IF) [33] and TSMC's integrated fan-out system-on-wafer (InFo-SoW) [26] are examples of chiplet-based waferscale technologies. These technologies allow one to select pre-tested known-good chiplets/dies (KGD) [21] and integrate these chiplets on a waferscale interconnect substrate. Previous works [48], [49] have shown that the KGDs can be integrated/bonded onto the substrate at very high yields (>99.9% [48]), thereby allowing one to achieve high system-level manufacturing yield. Also chiplets from disparate technologies can be integrated into a waferscale system, which in turn allows architectures that are not feasible with monolithic WSI (e.g., Tesla Dojo [59], Waferscale-GPU [49]).

B. Waferscale technology for a Network Switch

For this work, we choose to focus on chiplet-based waferscale integration because of its ability to achieve high yield [48] and integrate heterogeneous chiplets on the same substrate [34] (e.g., high bandwidth I/O chiplets at the edge alongside networking logic chiplets at the center of the substrate). We assume that all the chiplets communicate with their neighboring chiplets in a mesh-like configuration. For communication between non-adjacent chiplets, multi-hop connections are built by using the intermediate chiplets as repeaters. Depending

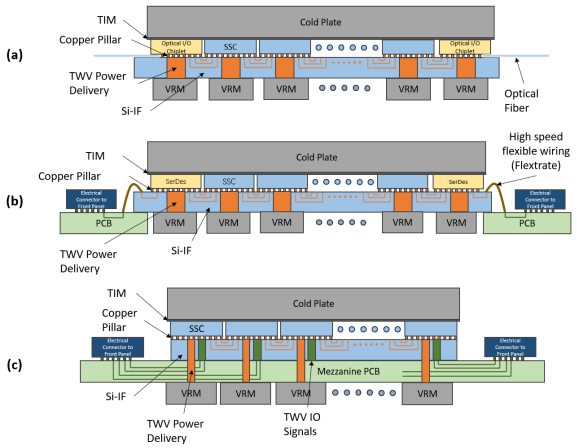


Fig. 3: Schematic cross-section of a waferscale network switch with (a) Optical I/O, (b) SerDes, (c) Area I/O.

on the integration scheme (Table I), power consumption and bandwidth density of the internal interconnections will be different.

A waferscale network switch also needs to connect to and communicate with the external world at high bandwidth. One way to do this is to place chiplets responsible for external I/O on the perimeter of the substrate (*Periphery IO*). These I/O chiplets have been SerDes-based [41] in today's waferscale systems. However, they can also be Optical I/O-based [16] where the chiplets directly support optical transceivers. In both schemes, I/O chiplets are connected to SSCs the same way using on-wafer interconnects, as shown in Figure 3.a and 3.b. Note that only the chiplets sitting on the perimeter of the substrate will have a physical connection with the I/O chiplets. If an internal chiplet (non-periphery chiplet) requires I/O connections, intermediate chiplets can repeat the signals to form a logical connection as discussed previously. An alternative way to bring external I/O signals onto the substrate is *Area I/O*. As the name suggests, Area I/O allows the I/O signals to reach any chiplets on the substrate directly, i.e., the external I/O transceivers are interspersed in the chiplets on the substrate. A cross-section schematic for Area I/O is shown in Figure 3.c. Note that TWVs (through-wafer-vias) allow the I/O signals and power to traverse and transit through the substrate wafer. The carrier PCB (we call it Mezzanine PCB) onto which the substrate wafer is mounted acts as a redistribution layer (RDL) layer for the Area I/O signals to escape the wafer and reach external I/O port modules. Table IV shows typical characteristics of SerDes, Optical I/O and Area I/O.

Technology	**SerDes** [41]	**Optical I/O** [16]	**Area I/O** [9]
type	periphery	periphery	area
bandwidth density / layer	512Gbps/mm	800Gbps/mm	16Gbps/mm^2
number of layers	1	4	1
energy / bit (pJ/bit)	8.0	5.0	8.0

TABLE IV: External I/O technologies considered in this work.

C. Building a Waferscale Network Switch

Our physical implementation of the waferscale network switch is a waferscale mesh network of sub-switches, where a state-of-the-art single-die network switch serves as the sub-switch. Mesh allows breaking longer interconnects into smaller

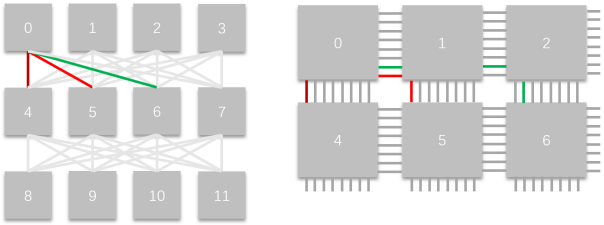


Fig. 4: Mapping a logical Clos topology onto the wafer.

segments improving energy, latency, and yield [49]. Previous commercial waferscale systems [40], [59] also implement a physical mesh. For the majority of this work, we will use the characteristics of Tomahawk 5 (TH-5), a state-of-the-art 256-port, and 200Gbps line rate switch ASIC built by Broadcom, as the subswitch chiplet (SSC), with a couple of modifications. One modification we assume is that the high-speed SerDes at the edge of a TH-5 chip would be replaced with fine-pitch inter-chiplet I/Os. The inter-chiplet wires connect one SSC to its four adjacent neighboring chiplets in the physical mesh. Second, since mesh topology has low saturation throughput, low bisection bandwidth, and high latency which is undesirable for a network switch [54], we need to be able to map more desirable logical topologies such as Clos that have high saturation throughput and non-blocking switching [27] on this physical mesh. To enable this mapping, we assume that a subset of the inter-chiplet I/Os and a small fraction of the chiplet area ($<2\%$) on the TH-5 chips would be used for feedthrough channels (with repeaters). With these modifications, each SSC can be used as a repeater to form a long logical link that connects two non-adjacent chiplets. The latency of the the connection between two adjacent SSCs is 1ns. Therefore, the worst-case latency between two remote SSCs on the wafer will be $2N$ ns where the number of SSCs on the wafer is N^2. Our TH-5-like SSC consumes 500W (400W without I/O power) and occupies 800mm^2.

Figure 4 shows an example mapping of a portion of a Clos onto the physical mesh. SSCs 0 and 4 have a logical link in the Clos. Because SSC 0 and SSC 4 are adjacent, they will have a direct connection on the substrate. SSCs 0 and 5 also have a logical link in the Clos, but they are not adjacent. Therefore, SSC 1 is used as an intermediate chiplet to create the physical link between SSCs 0 and 5. Similarly, SSC 0 and 6 are connected logically and are not adjacent. We can use SSCs 1 and 2 as repeaters to form the physical link. Note that in this case, two logical links of the Clos are mapped to the physical implementation such that both links pass through SSC 1.

More generally, when we map any non-mesh topology onto the substrate, there will be some SSCs that have multiple logical links passing through them. But because all the chiplets are identical and have the same off-chip bandwidth connectivity, some SSCs will have higher link utilization than others.

IV. Quantifying Radix Size Benefits of Waferscale Network Switches

A waferscale network switch can be two orders of magnitude larger in silicon area available and one to two orders of mag-

nitude larger I/O count than today's single chip switches. To quantify the resulting radix size benefits, we consider building a high radix switch using SSCs connected in a Clos topology. We focus on Clos because of its superior properties compared to several other topologies [61]. Since the physical layout of the interconnection is still a mesh topology, mapping a logical Clos onto it means that the link latencies between different Clos switches are now non-uniform. This will not affect the performance of the waferscale switch, however, as each SSC is a TH-5-like chip that has input buffers to handle non-uniform latency [20]. In addition, one might think that the non-blocking property of a Clos may get affected when mapped onto a mesh. However, we don't share wiring resources to multiple logical links. This means that every logical link is guaranteed to have at least a bandwidth of 200Gbps.

A. Mapping a Clos Topology

How one maps a logical Clos topology on a physical mesh would dictate the characteristics of the waferscale network switch. The goal of the mapping problem is to maximize the radix and the bisection bandwidth of the switch. The optimal mapping problem can be formulated as the following: Let G be a given topology, M_i be a floorplan of G, $C(M)$ calculates the maximum number of logic links between any adjacent chips in F. The objective of the optimization problem becomes:

$$\min_{M_i} C(M_i) \quad \forall i$$

Since each mapping is discrete and not differentiable, there is no analytic solution. In addition, for N chiplets, there will be $N!$ different mappings for placing these chiplets onto the wafer, so brute force search is not possible for large N. Therefore, we choose the pairwise exchange method [11], shown in Algorithm 1, that helps us explore potential local optimal. Starting from an initial mapping, at each iteration we swap the positions of a pair of chiplets. If this decreases the maximum number of logical links between any pair of SSC, the swapping is kept, and vice versa. The algorithm stops until no more swapping can be made anywhere on the wafer.

Algorithm 1 Pairwise Exchange Algorithm For Optimizing Mapping

```
R ← [r_i] list of SSCs
P ← [p_i] list of position on the wafer
M ← M(r_i) = p_i initial mapping
C ← C(M) = maximum number of logic links between any pair of SSCs
C_prev ← C(M)
while M Not Converged do
    for all SSC pair (r_i, r_j) in R do
        swap(M(r_i), M(r_j))
        C_new = C(M)
        if C_new < C_prev then
            C_prev = C_new
        else
            swap(M(r_i), M(r_j))
        end if
    end for
end while
```

We run the above algorithm 1000 times with different random initial mappings and report the best results (though we find that difference across trials yielded less than 1% difference in all cases). Our algorithm is effective (Figure 5). Compared to an unoptimized random initialization, our heuristic increases the worst-case internal I/O bandwidth per port by 147.6%.

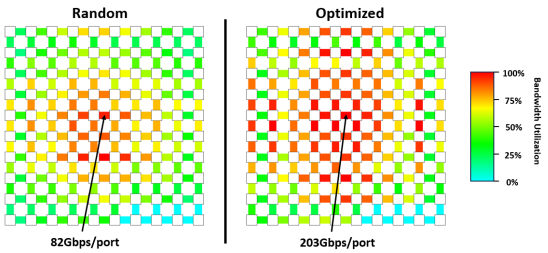


Fig. 5: Random mapping vs mapping using Algorithm 1.

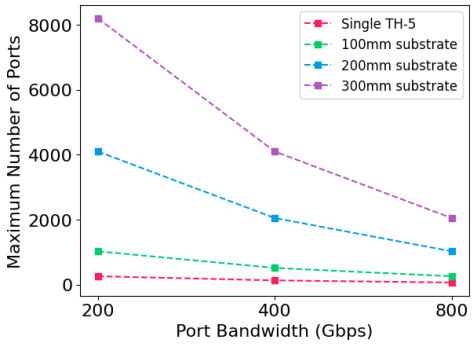


Fig. 6: Maximum number of ports achieved using WSI without considering any constraints.

B. The Ideal Case

We use the above mapping heuristic to determine the maximum number of ports attainable in a waferscale switch (Figure 6) for different TH-5 port bandwidth configurations when substrate area is the only constraint. We assume a square substrate - so, 100mm corresponds to a square with a side of 100mm. The three TH-5 configurations are highlighted as red squares in the graph.

We see that chiplet-based waferscale integration enables up to 32x higher number of ports compared to a single state-of-the-art Tomahawk 5 for a 300mm substrate. For 200mm and 100mm substrates, the benefits over TH-5 are 16x and 4x respectively. These large benefits can lead to significantly reduced hop count and latency, as well as direct all-to-all connection within a large cluster of terminals. The maximum number of achievable ports decreases when demand for port bandwidth increases and the substrate size remains the same. Even for high port bandwidth, however, 2-8x benefits are possible.

C. The reality

Figure 7 shows the maximum number of ports achievable at different substrate sizes for a realistic internal bandwidth of 3200Gbps/mm (Si-IF-like technology) for three external I/O schemes - periphery-based SerDes (512Gbps/mm) and Optical I/O (3200Gbps/mm), and mezzanine PCB-based Area I/O (16Gbps/mm^2). First, consider periphery-based SerDes - the external connectivity scheme used in all recent waferscale systems [40], [49], [59]. We see that this scheme only manages to double the number of ports (512) even when going to a substrate (300mm) that has over two orders of magnitude higher

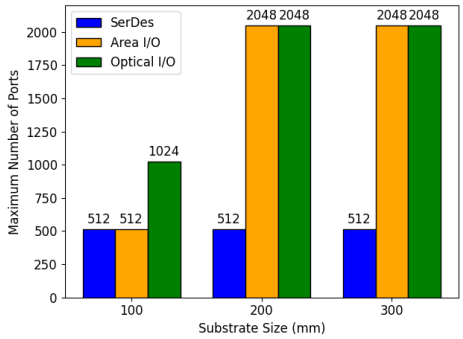


Fig. 7: The maximum number of 200Gbps ports achievable for different external I/O technologies at internal bandwidth density of 3200Gbps/mm.

area than a TH-5. Compared to the ideal case, this is 16x fewer ports. This big gap is due to the limited bandwidth it provides. It is clear that waferscale switches need newer interfaces such as Optical I/O and Area I/O that provide higher bandwidth external connectivity. In fact, we see that the number of ports for any given substrate size is up to 4x higher for Optical I/O and Area I/O, though still a 50-75% reduction from the ideal case for 200mm and 300mm.

The continued gap between the ideal and Optical I/O and Area I/O is explained by a combination of continued internal and external IO bandwidth bottlenecks. Note, for example, that the maximum number of ports achieved for Optical and Area I/O stays the same as the substrate size increases from 200mm to 300mm in spite of the higher external bandwidth available with increased substrate size. To understand this, we calculate the maximum radix after doubling the internal IO bandwidth density to 6400Gbps/mm (Figure 9). We see that maximum number of ports now increases for Optical I/O by 2-4x for all cases matching the ideal. This shows that the baseline inter-SSC I/O bandwidth density of 3200Gbps/mm was the bottleneck. Also, the maximum number of ports scales up for Optical I/O when going from 200mm to 300mm at 6400Gbps/mm. This shows that the internal I/O bandwidth was fully saturated at 200mm in the 3200Gbps/mm case, causing the maximum number of ports to stay the same at 300mm as well. Unlike Optical I/O, Area I/O does not increase radix size even at 6400Gbps/mm due to the lower bandwidth it provides compared to Optical I/O at these substrate sizes.

Figure 8 shows how the internal and external bandwidth utilization and bottlenecks change when the external I/O technology is switched from SerDes to Optical I/O and when the internal I/O bandwidth is doubled.

The above results do not consider power delivery or cooling constraints. Figure 10 shows the power consumption of the switches. For 200mm and 300mm substrates, power exceeds 14KW when Optical I/O and Area I/O are used. When power delivery and cooling constraints are considered (e.g., air cooling vs water cooling vs multi-phase cooling), the maximum number of ports may get limited even further.

V. Scalable Waferscale Network Switches

A. Improving I/O bandwidth by trading off energy efficiency

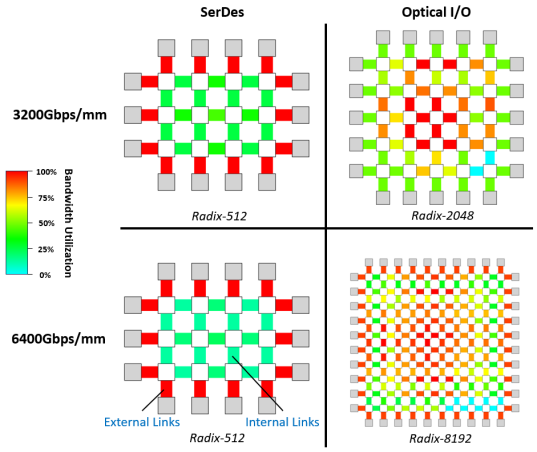


Fig. 8: Bandwidth utilization of internal and external I/Os for SerDes and Optical I/O at their maximum feasible radix, at internal I/O bandwidth density of 3200Gbps/mm and 6400Gbps/mm respectively. Grey squares are I/O chiplets.

There exist WSI technologies that afford higher I/O bandwidth at higher power. TSMC's InFO-SoW, for example, can provide four times the bandwidth (Table I), albeit at much less energy efficiency (1.5pJ/bit vs 0.06pJ/bit). Similarly, Si-IF bandwidth can be increased by scaling up link frequency and voltage at the expense of energy efficiency. Since our goal is to maximize radix while still being within a reasonable power budget, we target a bandwidth of 6400Gbps/mm (1600Gbps/mm/layer for 4 layers) - double the bandwidth used in the previous section - by doubling the frequency of the Si-IF-like I/O links while scaling up V_{dd} of the links accordingly. If the bandwidth of a physical wire is B and the energy per bit of it is P, the following relationship holds for supply voltage V_{dd} and threshold voltage V_{th} [51]:

$$P \propto V_{dd}^2$$
$$B \propto \frac{(V_{dd} - V_{th})^2}{V_{dd}}$$

We use the above relationship to model the internal I/O power at 6400Gbps/mm (four layers in total).

Figure 9 shows the achievable maximum radix at 6400Gbps/mm. We see that the maximum number of ports at 300mm now increases to 8192 - a 4x increase over the 3200Gbps/mm case. There is a 2x increase in the maximum radix at 200mm. The maximum radix at 100mm remains unchanged since it was already the same as the ideal case for Optical I/O.

Of course, the above large increase in switch radix comes at the expense of power. Figure 11 shows the power consumption of the switch at different substrate sizes. Power consumption can be as high as 62kW (for an 8192-radix switch). This is up to 3.5x larger than the power for 3200Gbps/mm. A large fraction of overall power (33%-43.8%) is internal I/O power + external I/O power.

We performed the same analysis with InFO_SoW WSI technology, which provides much higher I/O bandwidth density (12.8Tbps/mm), but has higher power consumption

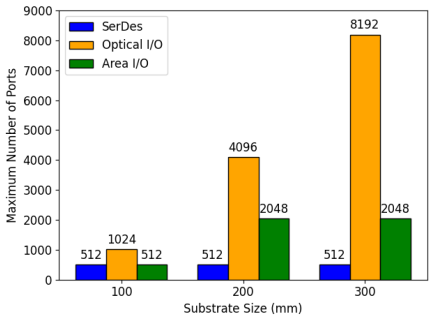


Fig. 9: The maximum number of 200Gbps ports achievable for different external I/O technologies at 6400Gbps/mm internal I/O bandwidth density.

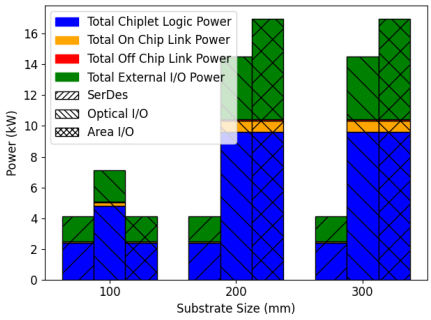


Fig. 10: Power breakdown for different external I/O technologies when internal I/O bandwidth is 3200Gbps/mm.

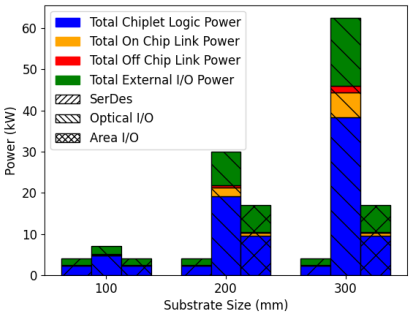


Fig. 11: Power breakdown for different external I/O technologies when internal I/O bandwidth is 6400Gbps/mm.

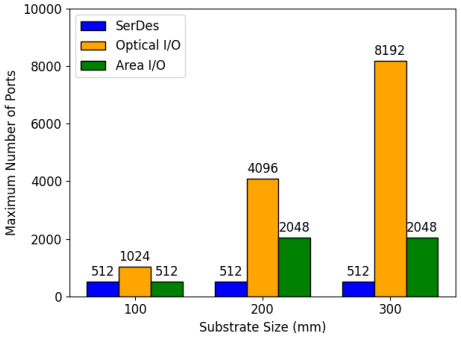


Fig. 12: The maximum number of 200Gbps ports achievable for different external I/O technologies at 12.8Tbps/mm internal I/O bandwidth density with InFO_SoW.

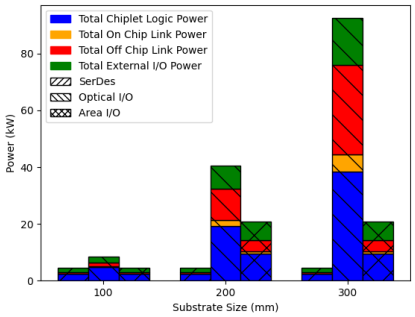


Fig. 13: Power breakdown for different external I/O technologies when internal I/O bandwidth is 12.8Tbps/mm with InFO_SoW technology.

(1.5pJ/bit). Figure 12 shows that InFO_SoW can achieve the same number of ports as 6400Gbps/mm Si-IF, but the overall package now draws 92.5kW of power (Figure 13). Because of the high power consumption of InFO_SoW, we will focus on Si-IF as our primary WSI technology for the remainder of this work.

B. Heterogeneous network switch

Thus far, we have mapped a homogeneous Clos topology to the substrate where all the SSCs - both leaf SSCs (ingress/egress SSCs) and spine SSCs (root-level SSCs that switch between leaves) - have the same radix. Here, we leverage an insight about the Clos topology that the leaf SSCs in a Clos can be disaggregated into multiple smaller leaf SSCs without change in overall radix as long as the connections to the spine SSCs are kept. This allows us to use 2 half-radix SSCs, for example, to replace the original leaf SSC, without altering the system-wide radix. Note that this disaggregation does increase the average hop latency of the switch by roughly 1%.

We observed that various commodity high-radix switches show super-linear (near quadratic) scaling of *normalized* power consumption with respect to the switch radix. Figure 15 shows the normalized reported consumption of various radix switches from Broadcom Tomahawk series [13] and Marvell TeraLynx series [1]. The raw power values are scaled using the methodology described in [57]. The scaling tracks well the quadratic scaling suggested by Ahn *et al.* [19] for both monolithic crossbars and hierarchical crossbars. The super-linear scaling suggests that the total power consumption of two radix-$k/2$ switches will be lower than a single radix-k switch upon replacement.

Figure 16 shows the overall power reduction of the wafer-scale switch after applying the heterogeneity optimization at different substrate sizes. At 300mm, we see an overall power reduction of 30.8%, when using scaled Tomahawk 3 dies as leaf nodes. This power reduction is significant enough that the new power density can be handled by water cooling [40] (0.69kW/mm^2 vs 0.48kW/mm^2 - water cooling can sustain 0.5kW/mm^2, we discuss power supply and cooling design in detail in Section VIII). Notably, this optimization reduces only the overall power consumption of the SSCs – it has minimal effect on the internal I/O power. Therefore, the power reduction decreases with increasing switch scale (substrate size) because internal I/O power is a more significant portion of total power at large substrate sizes.

C. Subswitch Deradixing

One somewhat counter-intuitive technique to relax the internal I/O bandwidth density requirement is to intentionally *decrease* the radix of SSCs while keeping the area of each SSC the same. Since the inter-chiplet I/Os are shared between feedthrough and non-feedthrough (port) I/Os, decreasing the

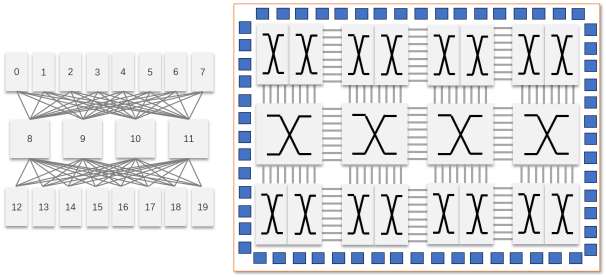


Fig. 14: Topology and placement illustration for a heterogeneous switch design with scaled TH-4 as leaf SSCs.

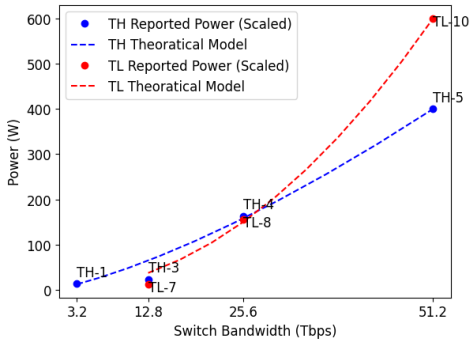


Fig. 15: Reported power consumption of Tomahawk series switches (TH-1, TH-3, TH-4, TH-5) and TeraLynx series switches (TeraLynx-7, TeraLynx-8, TeraLynx-10), normalized to 5nm process node (only non-IO power is shown). Theoretical quadratic power scaling models for both series are shown as well.

radix allows for a larger number of feedthrough I/Os through an SSC. This, in turn, allows us to connect more SSCs on the substrate at higher bandwidths, thus meeting the overall internal bandwidth requirement of the Clos topology. This results in higher overall radix in certain cases. Figure 19 shows the available internal I/O bandwidth per SSC port when constructing a 300mm system with both original 256-port SSCs and deradixed 128-ports SSCs, keeping the per-SSC chip area the same. In the baseline 256-port SSC case, we could achieve 8192 and 4096 ports for the switch, but their respective available per port bandwidth does not meet the required 200Gbps/port. So we are stuck with a smaller overall radix (2048) configuration. On the other hand, when we use deradixed SSCs, the per port bandwidth exceeds the required 200Gbps/port and an overall radix of 4096 is achieved.

Figure 17 shows the maximum number of ports achieved by the switch if we decrease the radix of each SSC. We see that, for the 300mm substrate, reducing the radix to half doubles the achievable number of ports from 2048 to 4096. Similarly, halving or quartering the radix per SSC for the 200 and 300mm substrates using SerDes also increases the achievable radix.

Of course, excessive deradixing may lead to an overall reduction in the number of ports on the switch since fewer ports overall can be packed in the same substrate area. This behavior is more pronounced at 6400 Gbps/mm internal I/O bandwidth density (Figure 18) where the internal I/O bandwidth is already sufficient.

VI. Performance and Microarchitectural Implications

Since $B = RTT \times BW/\sqrt{n}$ where B is the buffer size, RTT is the round trip time of the link, BW is the bandwidth of the link, and n is the number of flows on that link [20], the lower latency of on-wafer interconnect (Table V) will translate to lower buffering requirements inside the SSCs. This is confirmed by Figure 21 which shows link delay vs buffer size generated using Booksim2 [35], a cycle-accurate detailed network simulator that was used to model all four stages of the switch microarchitecture shown in Figure 20 - route computation, virtual channel allocation, switch allocation, and switch traversal, for a router with 64 virtual channels (VCs), an equivalent delay of 200ns, and a shared buffer policy for all the input ports. For simulations to end within a reasonable amount of time, we define a simulation cycle within Booksim2 to be 20ns (so 200ns corresponds to 10 simulation cycles), and the flit length is adjusted to match the line rate of TH-5. Smaller buffers not only save on area and power, they also have lower latency since they can be implemented using fast SRAM instead of slower DRAM that is used to implement large buffers [56]. Therefore, an SSC on a a waferscale network switch can have lower latency than the commodity switch (e.g., TH-5) due to lower buffering requirements.

Another opportunity is replacing slow Layer-3 route computation in the non-ingress SSCs of a waferscale network swich with custom low-latency route computation. In a Layer-3 route computation, the best matching prefix of the IP address needs to be calculated which increases the IP table lookup time [52]. For a waferscale switch, the network topology is fixed, providing a chance to eliminate the IP table lookup in the non-ingress SSCs. Normal IP table lookup is performed is still performed at ingress SSCs and the destination port of the entire switch is calculated. The destination port information is added to the front of the packet header and the packet gets routed to the next SSC. The spine SSC and the egress SSC will not perform any Layer-3 routing. Instead, they will rely on the destination port to perform route computation. The egress switch will undo the modification to the header before sending the packet out. We estimate that the route computation (RC) without IP lookup will take $\frac{1}{4}$th of the baseline time [30].

We used Booksim2 to quantify the benefit of not needing IP table lookup at non-ingress SSCs. We consider a 2-level Clos network with 96 radix-256 SSCs, forming an overall radix of 8192. The switch latency is set to 16 simulation cycles and the routing delay to 4 simulation cycles. A shared buffer of 128 flits per port is used. The number of VCs is set to 64. No internal speedup is used. We used random uniform traffic patterns. We also configured the routing delay to 2 simulation cycles and 1 simulation cycle respectively when the packet is going through ingress and non-ingress SSCs. Figure 22 shows the results. We see that not only the zero-load latency is reduced when using proprietary routing, but the saturation throughput is better as well, with 14.5% and 11% increase for 200mm and 300mm substrate switch respectively.

To estimate potential application-level performance benefits

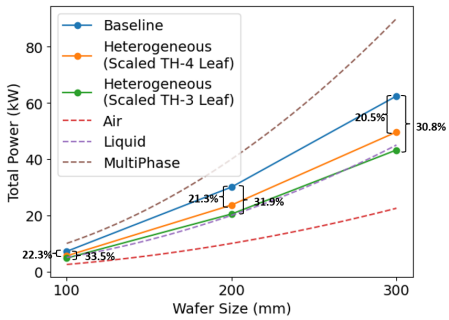


Fig. 16: Power reduction from a heterogeneous switch design. The power envelopes allowed by air cooling [46], liquid cooling [40], and multiphase cooling [36] are shown as well.

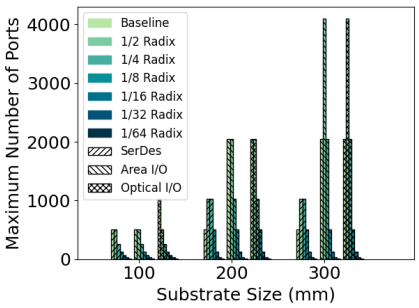


Fig. 17: Maximum number of ports when reducing SSC radix size (3200Gbps/mm internal I/O bandwidth).

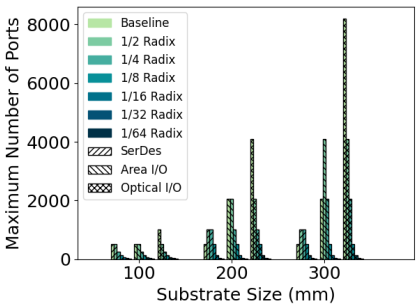


Fig. 18: Maximum number of ports when reducing SSC radix size (6400Gbps/mm internal I/O bandwidth).

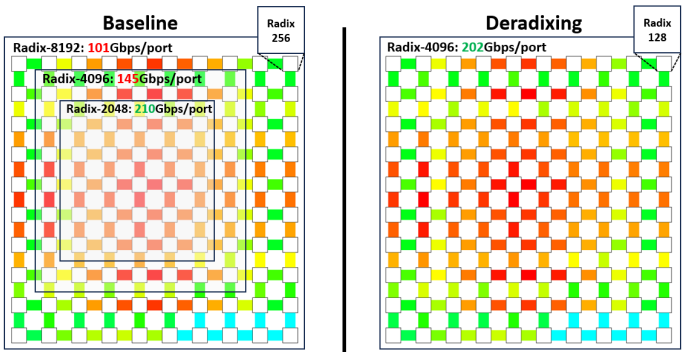


Fig. 19: An illustration of deradixing at 300mm substrate size. When radix-256 SSC is used (left), the bandwidth requirement is met only when the system radix is 2048 (labeled green) and violated for larger system radix of 4096 and 8192 (labeled red). By reducing the radix of each SSC from 256 to 128 (right), we are able to pack more SSCs onto the wafer because of the increase in available internal I/O bandwidth per SSC port.

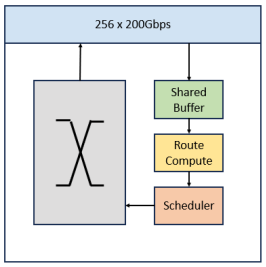


Fig. 20: Simulated SSC microarchitecture.

Type of connection	Latency (ns)
On-wafer connection [33]	10-20
In Rack PCB connection [60]	100-200
100m Optical link [2]	350

TABLE V: Latency values for different types of connections between two switching ASICs.

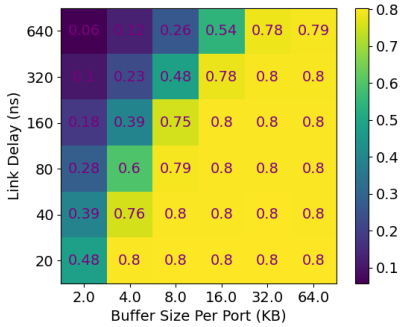


Fig. 21: Saturation throughput for various buffer sizes and link latency.

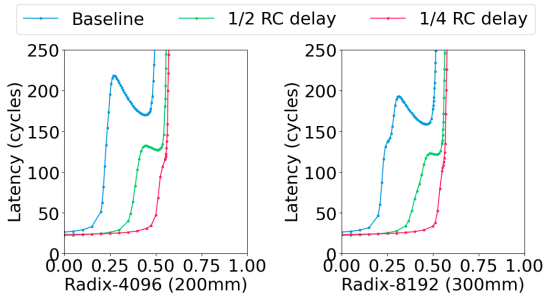


Fig. 22: Latency versus load for when removing IP table lookup in the RC unit of non-ingress SSCs.

of a waferscale network switch (both from faster interconnects and faster SSCs), we compared using Booksim2 the performance of a 2048-port 800Gbps waferscale switch against its equivalent switch network. We set the number of VCs to 16 with a buffer size of 32 flits per input port, SSC delay to 11 simulation cycles, switch box delay in the baseline switch network to 15 simulation cycles, the I/O delay (host to the ingress switch) to 8 for both the waferscale switch and the baseline, the latency between each SSC to 1 simulation cycle, and the latency between each switch in the baseline to 8 simulation cycles. We not only observe (Figure 23) equal or higher saturation throughput for the waferscale switch but also that the zero-load latency of a waferscale switch is 38% lower than a TH-5 switch network, at 37 simulation cycles and 60 simulation cycles respectively for all the traffic patterns except asymmetric traffic.

We also evaluated performance on four HPC NERSC traces [3]. We parsed these traces and fed them into Booksim2

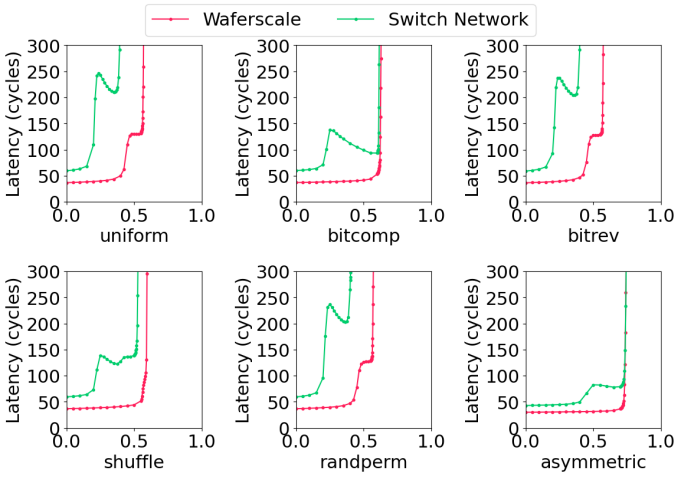


Fig. 23: Load vs. Latency graph for a 2048 port waferscale switch and an equivalent 2048 host switch network for different synthetic traffic patterns. One simulation cycle is equivalent of 20

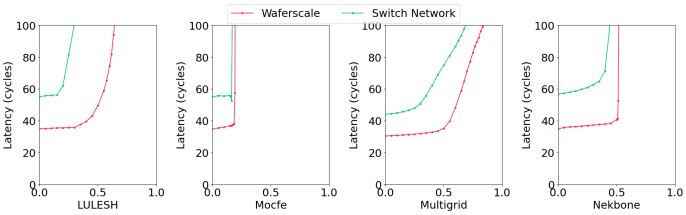


Fig. 24: Load vs. Latency graph for a 2048 port waferscale switch (vs 2048 host switch network) for NSERC benchmarks.

simulator. Since these traces were generated using 512 or 1024 nodes, we duplicated the packets twice or four times in order to evaluate our 2048 node network. The results are shown in Figure 24. For LULESH, Mocfe, Multigrid, and Nekbone, the saturation throughput of a waferscale network switch is 116.7%, 16.7%, 21.4%, and 15.2% higher than the TH-5 switch network baseline.

VII. Discussion

Non-Clos topologies. Waferscale integration enables orders of magnitude more ports than a monolithic switch in the ideal case (Figure 25.a) even for non-Clos topologies. A 300mm WSI switch constructed with butterfly topology achieves 44x, 31x, 19x, and 44x more ports compared to a single Tomahawk 5, using Butterfly, DragonFly, Flatten Butterfly, and Mesh topologies respectively. Once power density and per port bandwidth constraints are considered, the radix benefits that we see reduce dramatically (Figure 25.b). A significant fraction of the ideal benefits is reclaimed once we consider optimizations such as deradixing and heterogeneous design (Figure 25.c).

Mesh provides 10% higher radix benefit than Clos due to ease of 2D layout, but suffers from poor bisection bandwidth and being highly blocking. Butterfly provides 10% higher radix in the optimized cases than Clos, but has lower bisection bandwidth and path diversity. Dragonfly and Flatten Butterfly provide 1.7x-3.2x lower radix than Clos since they are direct topologies which increases their external bandwidth requirement and makes them hard to lay out.

Constructing a physical Clos. In this work, we focused on mapping topologies to a physical mesh. It is possible to construct a physical Clos with repeaters for long wires (Figure 26.a, Figure 26.b, Figure 26.c). We see that physical Clos always has lower radix than Mapped Clos since the significantly increased number of links for physical Clos cuts into the area that can be used to place TH5s. This is true even when we allow the interposer links to traverse underneath the SSC (not often possible in reality as the center region of the chiplet is often used for power delivery). The relatively large number of off-chip links also causes a 10% power overhead vs Mapped Clos (iso-radix).

Alternative microarchitectures. We created a (Clos) network of TH-5s to improve the switch radix. Other possible approaches to build a high-radix switch include building a hierarchical crossbar [38] and modular crossbar [22]. However, these approaches scale a lot worse. Table VI shows the number of chiplets required to construct a given radix waferscale switch. The area, power, and monetary costs are prohibitively expensive for hierarchical crossbars and modular crossbars.

	Clos	**HC**	**MC**
# chiplets	3(N/k)	$(N/k)^2$	$(N/k)^2$
# chiplets (N=2048, k=256)	24	64	64
# chiplets (N=8192, k=256)	96	1024	1024

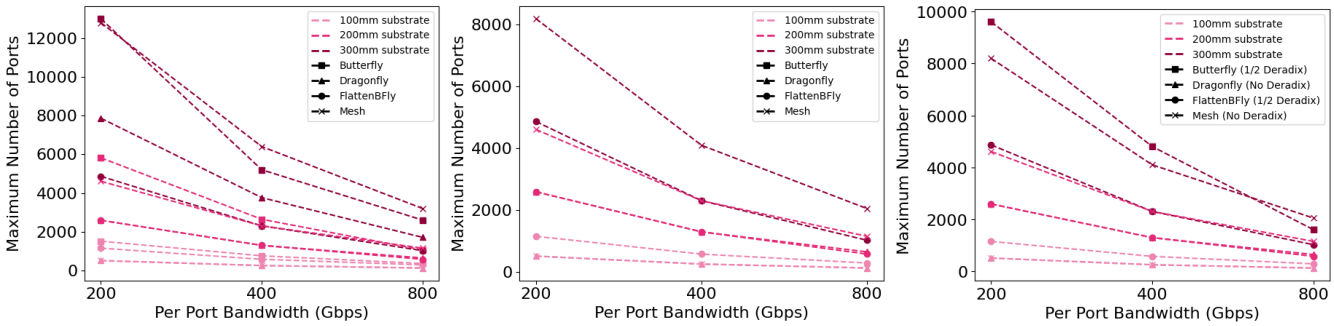
TABLE VI: The number of chiplets required by Clos, hierarchical crossbar (HC), and modular crossbar (MC). N is the network size (total number of ports), and k is the radix of SSC.

Sensitivity to internal bandwidth density. Our previous analysis mostly assumed a four-micron pitch and a maximum of four metal layers for SSC communication. To reduce crosstalk, we assume that each communication metal layer alternates with a power/ground metal layer for a total of eight layers. This is already an aggressive assumption for an interposer - the current generation of TSMC CoWoS-S supports five metal layers [32]; earlier generations supported only three metal layers. The limited number of metal layers for any interposer is due to the loss of yield with every additional layer due to the increased number of processing steps. Nevertheless, we perform a sweep to understand the sensitivity of radix with respect to internal bandwidth density (increases linearly with number of metal layers). Figure 27 confirms that the available area on the wafer will become the bottleneck if the internal bandwidth density can be increased by multiple factors (unlikely, at least in the short to medium term).

VIII. System-level Architecture and Use Cases

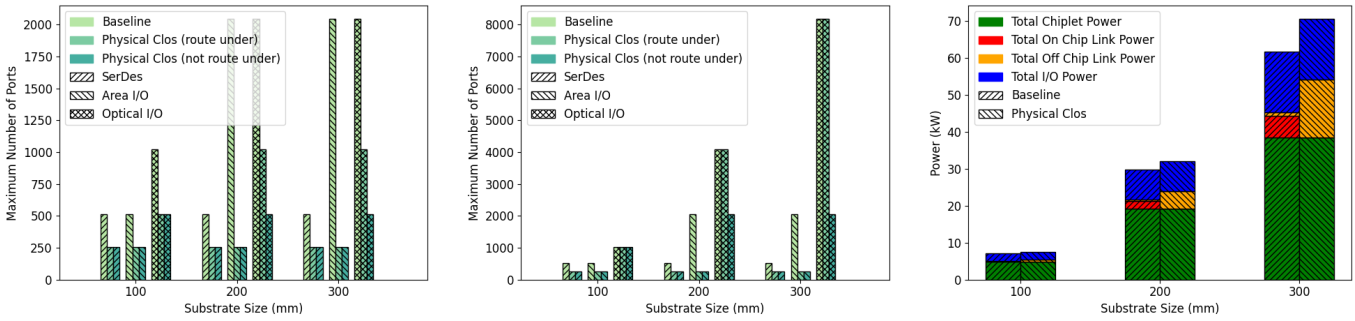
A. System-level Architecture of Waferscale Network Switches

The system-level architecture of a waferscale network switch must accommodate the port count, power delivery, and cooling requirements in a compact form factor. Figure 29 shows the proposed enclosure structure for a 300mm substrate waferscale network switch. This switch allows 8192 ports at 200Gbps per port, 4096 ports at 400Gbps per port, or 2048 ports



(a) (b) (c)

Fig. 25: The maximum number of 200Gbps ports achieved when considering non-Clos topologies, (a) when no constraints are considered, (b) when area/bandwidth/power constraints are considered, and (c) when optimizations are applied.



(a) (b) (c)

Fig. 26: The maximum number of 200Gbps ports achieved at (a) 3200Gbps/mm and (b) 12800Gbps/mm when considering Clos-map-to-Mesh vs physical Clos. (c) The power comparison and breakdown for Clos-map-to-Mesh vs physical Clos at iso-radix.

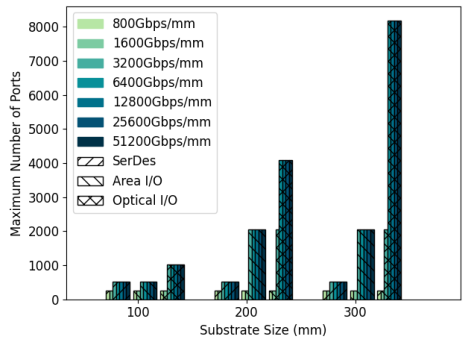


Fig. 27: The maximum number of 200Gbps ports achieved when considering different internal bandwidth densities.

at 800Gbps. After applying the heterogeneity optimization, a 300mm substrate draws around 45kW of power. To supply this power, we consider using 25 high-density server power supply units (PSUs), each capable of delivering 4kW [5]. PSUs are used to provide 50kW + 50kW of power (5kW for non-ASIC components), with N+N redundancy in mind. Note that the PSUs will down-convert 3-phase 240V AC to 48V DC. Then we use a set of DC-DC converters (48V-12V) and voltage regulation modules (VRMs, 12V-<2V) [4] to supply current at 0.75V-1.2V to the SSCs. Each DC-DC converter is about 27mm x 18mm and can provide 1kW+ power; each VRM is 10mm x 9mm and can supply about 130A of current. All the 50 DC-DC converters and the 420 VRMs (10% redundancy) easily fit under the area of the wafer (with $1/3^{rd}$ the space left for other passive devices). The placement of the VRMs is critical to minimize I^2R power loss, and attaching them directly on the back-side of the wafer helps achieve lower losses as well as better voltage regulation response [49] (see Figure 3). To further enhance voltage regulation, in-wafer deep-trench capacitors [37] (similar to CoWoS DT-Cap [31]) would be used.

Cooling is a challenge for a 45kW waferscale switch. Recall, however, that heterogeneous design has lowered the power density to 0.48W/mm^2, which is lower than Cerebras' WSE-2 power density of 0.4976W/mm^2. As such, a liquid cooling design similar to the one used for WSE-2 can be used here as well. Our largest waferscale switch system contains a 12x12 array of switching and external I/O chiplets. A passive cold plate loop (PCL) (Figure 29) copper spreader is used to cover 2x2 chiplets, totaling 36 PCLs, where each PCL dissipates 1.6kW of power. Three consecutive PCLs share the same set of supply channels. Therefore, 12 supply channels leave the wafer. Each channel connects to a pump inlet and outlet. We calculated that the pump needs to deliver 10-12 linear feet per minute (LFM) of deionized (DI) water at 10psi as per OCP guidelines [24] for the required dissipation. Our simulations show that with 20°C inlet temperature, the junction temperature will be 70-80°C.

Figure 28 shows the maximum radix supportable for each cooling solution (after heterogeneous optimization). Even traditional cooling technologies such as air cooling and water cooling support radix improvements by 8x and 32x compared

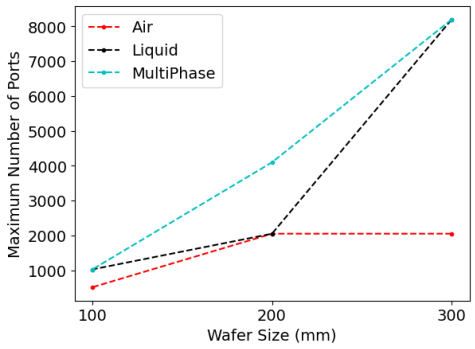


Fig. 28: The maximum number of ports allowed by different cooling solutions at different wafer sizes, after applying heterogeneous optimization to reduce total power.

to a single TH-5, though multi-phase cooling is recommended to derive full benefits from a waferscale switch at all wafer sizes.

This system architecture utilizes the back panel of the chassis for power delivery and cooling infrastructure. The entire front panel is therefore available for network ports.

We use optical I/O such that the optical fibers can directly come out of the substrate and go to adapters located at the front panel. If we use mezzanine connectors (Area I/O scheme), the optical transducers placed on the mezzanine PCBs are connected to the front panel using optical fiber links. In both these schemes, O/E/O conversion happens on the wafer/PCB plane. This allows us to avoid QSFP modules at the front panel, allowing us to pack more connectors at the front panel with the use of optical adapters (also called optical couplers).

Each RU (rack unit) can fit 108 CS optical adapters [6], so 19U allows us to fit 2052 adapters on the front panel. The top 1U space is dedicated to a management server. If the system is configured to have higher radix, say 8192 ports or 4096 ports, the solution is to use splitter cables to bifurcate a single 800 port to multiple ports. Through this design decision, we can make the entire system fit within 20U, as opposed to 80U.

System-level architecture for a 200mm substrate waferscale network switch can be derived directly from above and is shown in Figure 30.

B. Use cases

A waferscale network switch can enable new computer systems due to its very high port count. One use case is that it can be the only switch for small to moderately-sized data centers (Figure 31). The waferscale switch is placed at the center rack of the data center while optical splitter cables travel below the ceiling to reach every rack in the datacenter. A single-switch datacenter can deliver significant benefits (Table VII) in terms of space (e.g., there is a 90% reduction in rack space when we integrate all the TH-5 switches into a waferscale switch - TH-5s sit in a 2U switch box), cost (since we can remove all the optical cables and pluggable modules in between all the switches in the original TH-5 switch network), and latency (since packets will go through just 1 integrated switch as opposed to 3 switches). Actual benefits may be higher through optimizations (Section VI).

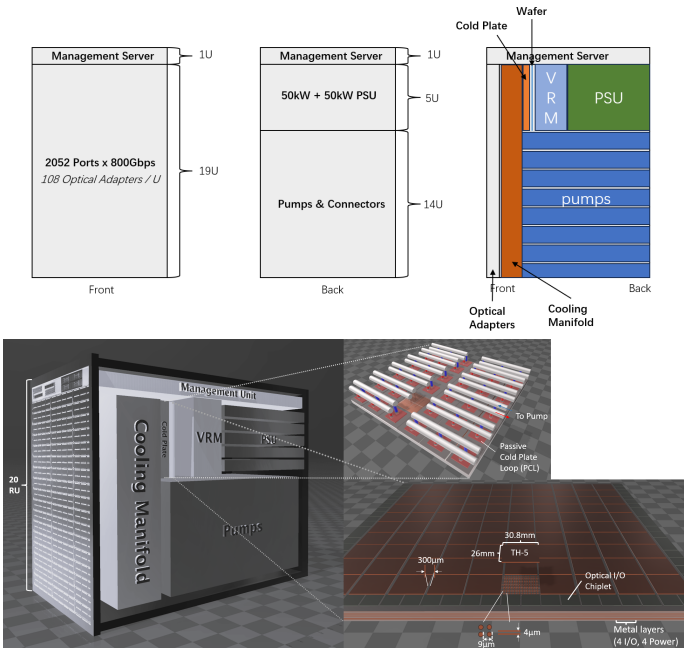


Fig. 29: The system architecture of the waferscale network switch using 300mm substrate. 3D view at the bottom.

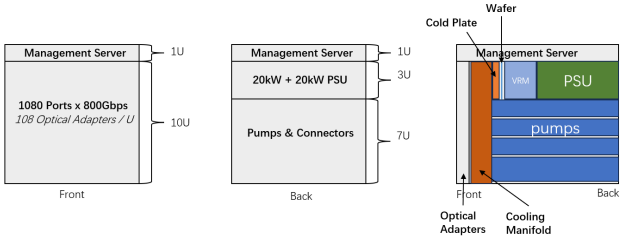


Fig. 30: The system architecture of the waferscale network switch using 200mm substrate.

System	Waferscale switch	TH-5 Clos Network
# of servers	8192 (4096)	8192 (4096)
# of switches	1	96 (48)
# of cables	8192 (4096)	16384 (8192)
worst case hop count	1	3
size (RU)	20 (11)	192 (96)
port bandwidth (Gbps)	200	200
bisection bandwidth (Tbps)	800 (400)	800 (400)

TABLE VII: A datacenter built with a 300mm waferscale switch (values for 200mm shown in parenthesis) vs an equivalent Clos network built with TH-5.

A waferscale network switch can also enable a massive-scale singular GPU. With the rapid development of Large Language

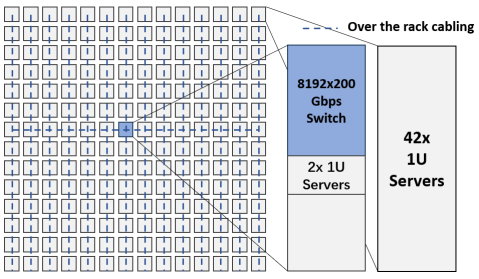


Fig. 31: The floorplan of a single-switch datacenter enabled by a waferscale switch.

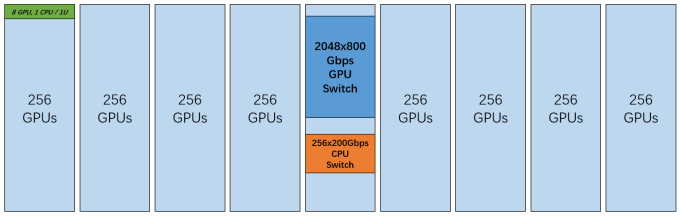


Fig. 32: The rack architecture of a singular GPU enabled by a waferscale switch.

Models (LLM) and large recommendation models, the need for larger GPU clusters with large amounts of memory also grows to enable distributed training of these models. DGX GH200, for example, uses 132 NVswitches to support 256 GPUs with 144TB of shared VRAM at the data rate of 900Gbps per GPU [8]. Using a 300mm substrate waferscale switch in 2048 $\times$ 800Gbps configuration, we can support 2048 GPUs reaching an astonishing 1.152 PB of shared memory at a single hop count (Table VIII). Figure 32 shows the rack architecture of the 2048-GPU cluster. The cluster has eight compute racks - a single compute rack has 32 server boxes–8 GPUs and 1 CPU per box, and a center switch rack. The GPU switch box has the same architecture as Figure 29 - simply configured as a 2048 $\times$ 800Gbps switch). Since each GPU requires a large amount of bandwidth going into it, no TOR switch is used. Instead, optical cables go into each GPU directly. A regular Tomahawk-5 switch box is used for connecting all the hosts (CPU) together. The host (CPU) network will not require a large bandwidth because all the computational data flows from GPU to GPU directly; the CPUs simply act as controllers.

System	Waferscale switch	NVswitch network
# of GPUs	2048 (1024)	256
# of switches	1	132
# of cables	2048 (1024)	2304
hop count	1	3
size (RU)	20 (11)	195
port bandwidth (Gbps)	800	900
bisection bandwidth (Tbps)	819.2 (409.6)	115.2

TABLE VIII: A singular GPU cluster with a 300mm waferscale switch (values for 200mm shown in parenthesis) vs a 2-layer NVswitch network.

In another use case, multiple waferscale switches can be connected together to implement a large-scale DCN with less rack space, fewer optical cables, and fewer switches (Table IX). The rack layout is shown in Figure 33 for the 300mm substrate case. 48 waferscale switches form the spine network of the data center. Each waferscale switch is 2048x800Gbps, connected in a DCN-level Clos topology. The TOR switch of each rack will be connected to the spine switches with two 800Gbps links, giving a per-rack throughput of 1600Gbps.

We reduce the number of optical links needed by 66% - for context, the cost of optical fiber is around $400 per km, and a single 800Gbps QSFP-DD Transceiver Module costs $5000 [29]. The rack space allocated to spine switches is reduced by 94% - the colocation cost per 1U server is around $75-$300 [28]. Together, this could result in millions of dollars of savings - *hundreds of millions* for the biggest datacenters.

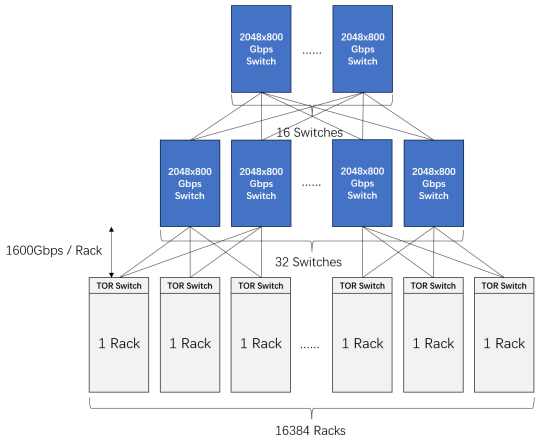


Fig. 33: A DCN with 48 300mm waferscale switches.

System	Waferscale switch	TH-5 Clos Network
# of racks	16384 (8192)	16384 (8192)
# of switches	48 (48)	4608 (2304)
# of cables	65536 (32768)	163840 (114688)
worst case hop count	3	5
size (RU)	960 (528)	18432 (9216)
Per Rack BW (Gbps)	1600	1600
bisection bandwidth (Tbps)	13107.2 (6553.6)	13107.2 (6553.6)

TABLE IX: A DCN with conventional switches vs 48 300mm waferscale switches (values for 200mm shown in parenthesis).

IX. Summary and Conclusions

In this work, we asked the question: can we use WSI to enable network switches that have dramatically higher radix than today's switches? While a 32x higher radix switch can be built within the area constraints of a wafer, practical limitations from internal bandwidth, external bandwidth, and power density limit the realizable radix. To address scalability challenges, we introduced a heterogeneous switch design, reducing power consumption by 30.8%-33.5%, thus enabling a 4x increase in radix by selectively trading off energy efficiency for link bandwidth. We also proposed subswitch deradixing to alleviate internal I/O bottlenecks, achieving a 2x increase in the overall radix. To mitigate external I/O bandwidth bottlenecks, we utilized Area I/O and Optical I/O schemes for external connectivity. Alongside scalability optimizations, we presented enhancements like low-latency buffering and proprietary routing to boost wafer-scale switch performance. Finally, we outlined a compact system architecture supporting port count, power delivery, and cooling requirements, demonstrating the potential for new computing systems and substantial reductions in datacenter network costs. This study marks the first comprehensive exploration of the benefits, challenges, and opportunities associated with waferscale switches.

X. Acknowledgments

We would like to thank the reviewers, Matthew Tomei, and Aravind Srikumar for their feedback. We would also like to thank Pavan Hanumolu and his students - they asked good, hard questions during this research that helped refine this work.

REFERENCES

[1] [Online]. Available: https://www.marvell.com/products/switching/datacenter.html

[2] [Online]. Available: https://www.commscope.com/globalassets/digizuite/2799-latency-in-optical-fiber-systems-wp-111432-en.pdf?r=1

[3] [Online]. Available: https://portal.nersc.gov/project/CAL/doe-miniapps.htm

[4] "48v power modules." [Online]. Available: https://www.monolithicpower.com/en/products/power-modules/48v-modules.html

[5] "bel power solutions tet4000 titanium efficiency 4 kw power supply." [Online]. Available: https://www.avnet.com/wps/portal/abacus/products/new-products/npi/bel-power-solutions-tet4000-series/

[6] "Cs connectors high density fiber cs connector technology for 200g and 400g."

[7] "netengine 8000 series routers." [Online]. Available: https://e.huawei.com/en/products/routers/ne8000

[8] "Nvidia dgx gh200 massive memory supercomputing for emerging ai." [Online]. Available: https://www.nvidia.com/en-us/data-center/dgx-gh200/

[9] "Ocp mezzanine card 2.0 design specification." [Online]. Available: https://www.opencompute.org/documents/facebook-ocp-mezzanine-20-specification

[10] "part 2: getting there faster: the evolution of serdes and high-speed data links — silicon creations technical article." [Online]. Available: https://www.chipestimate.com/Part-2-Getting-there-faster-The-evolution-of-SERDES-and-high-speed-data-links/Silicon-Creations/Technical-Article/2020/12/15

[11] "Placement - iit kgp." [Online]. Available: http://www.facweb.iitkgp.ac.in/~isg/CAD/SLIDES/09-placement.pdf

[12] "ptx10000 line of modular routers datasheet — juniper networks us." [Online]. Available: https://www.juniper.net/us/en/products/routers/ptx-series/ptx10000-line-of-packet-transport-routers-datasheet.html

[13] "Tomahawk 5 - ethernet network switches." [Online]. Available: https://www.broadcom.com/products/ethernet-connectivity/switching/strataxgs/bcm78900-series

[14] "Will optical replace electronic packet switching," May 2007. [Online]. Available: https://spie.org/news/0687-will-optical-replace-electronic-packet-switching

[15] "Product specifications," 2018. [Online]. Available: https://ark.intel.com/content/www/us/en/ark.html

[16] "Technical brief: Optical i/o chiplets eliminate bottlenecks to unleash innovation_2021," Jun 2021. [Online]. Available: https://ayarlabs.com/technical-brief-optical-i-o-chiplets-eliminate-bottlenecks-to-unleash-innovation/

[17] "Cisco nexus 9800 series switches white paper," August 2023. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus-9800-series-switches-wp.pdf

[18] "The inevitable transition_2023," Oct 2023. [Online]. Available: https://community.juniper.net/blogs/sharada-yeluri/2023/10/13/chiplets-the-inevitable-transition

[19] J. H. Ahn, S. Choo, and J. Kim, "Network within a network approach to create a scalable high-radix router microarchitecture," in *IEEE International Symposium on High-Performance Comp Architecture*, 2012, pp. 1–12.

[20] G. Appenzeller, I. Keslassy, and N. McKeown, "Sizing router buffers," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, p. 281–292, aug 2004. [Online]. Available: https://doi.org/10.1145/1030194.1015499

[21] R. Arnold, S. Menon, B. Brackett, and R. Richmond, "Test methods used to produce highly reliable known good die (kgd)," in *Proceedings. 1998 International Conference on Multichip Modules and High Density Packaging (Cat. No.98EX154)*, 1998, pp. 374–382.

[22] C. Cakir, R. Ho, J. Lexau, and K. Mai, "Scalable high-radix modular crossbar switches," in *2016 IEEE 24th Annual Symposium on High-Performance Interconnects (HOTI)*, 2016, pp. 37–44.

[23] F. Callegati, D. Careglio, L. H. Bonani, M. Pickavet, and J. Solé-Pareta, "Why optical packet switching failed and can elastic optical networks take its place?" *Optical Switching and Networking*, vol. 44, p. 100664, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1573427721000618

[24] C. Chen, D. Trieu, T. Shah, A. Guo, J. Cheng, C. Chapman, S. Kang, E. Dagan, H. Labs, A. Dinstag, and J. Yao, *OCP OAI SYSTEM LIQUID COOLING GUIDELINES*. [Online]. Available: https://www.opencompute.org/documents/oai-system-liquid-cooling-guidelines-in-ocp-template-mar-3-2023-update-pdf

[25] N. Chrysos, C. Minkenberg, M. Rudquist, C. Basso, and B. Vanderpool, "Scoc: High-radix switches made of bufferless clos networks," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, 2015, pp. 402–414.

[26] S.-R. Chun, T.-H. Kuo, H.-Y. Tsai, C.-S. Liu, C.-T. Wang, J.-S. Hsieh, T.-S. Lin, T. Ku, and D. Yu, "Info_sow (system-on-wafer) for high performance computing," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020, pp. 1–6.

[27] C. Clos, "A study of non-blocking switching networks," *The Bell System Technical Journal*, vol. 32, no. 2, pp. 406–424, 1953.

[28] B. Dobran, "The definitive guide to colocation pricing," Nov 2023. [Online]. Available: https://phoenixnap.com/blog/colocation-pricing-guide-to-costs

[29] Fs, "Generic compatible 800gbase-dr8 qsfp-dd pam4 1310nm 500m dom mtp/mpo-16 smf optical transceiver module." [Online]. Available: https://www.fs.com/products/150363.html

[30] D. Gibson, H. Hariharan, E. Lance, M. McLaren, B. Montazeri, A. Singh, S. Wang, H. M. Wassel, Z. Wu, S. Yoo *et al.*, "Aquila: A unified, low-latency fabric for datacenter networks," in *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, 2022, pp. 1249–1266.

[31] S. Hou, H. Hsia, C. Tsai, K. Ting, T. Yu, Y. Lee, F. Chen, W. Chiou, C. Wang, C. Wu, and D. Yu, "Integrated deep trench capacitor in si interposer for cowos heterogeneous integration," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019, pp. 19.5.1–19.5.4.

[32] P. K. Huang, C. Y. Lu, W. H. Wei, C. Chiu, K. C. Ting, C. Hu, C. Tsai, S. Y. Hou, W. C. Chiou, C. T. Wang, and D. Yu, "Wafer level system integration of the fifth generation cowos®-s with high performance si interposer at 2500 mm2," in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, 2021, pp. 101–104.

[33] S. S. Iyer, S. Jangam, and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for ai systems," *IBM Journal of Research and Development*, vol. 63, no. 6, pp. 5:1–5:16, 2019.

[34] S. Jangam and S. S. Iyer, "Silicon-interconnect fabric for fine-pitch (10 m) heterogeneous integration," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 5, pp. 727–738, 2021.

[35] N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J. Kim, and W. J. Dally, "A detailed and flexible cycle-accurate network-on-chip simulator," in *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2013, pp. 86–96.

[36] Y. Joshi and Z. Wan, *Single- and Multiphase Flow for Electronic Cooling*. Cham: Springer International Publishing, 2017, pp. 1–58. [Online]. Available: https://doi.org/10.1007/978-3-319-32003-8_49-1

[37] K. T. Kannan and S. S. Iyer, "Deep trench capacitors in silicon interconnect fabric," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020, pp. 2295–2301.

[38] J. Kim, W. Dally, B. Towles, and A. Gupta, "Microarchitecture of a high radix router," in *32nd International Symposium on Computer Architecture (ISCA'05)*, 2005, pp. 420–431.

[39] J. Kim, W. J. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *2008 International Symposium on Computer Architecture*, 2008, pp. 77–88.

[40] G. Lauterbach, "The path to successful wafer-scale integration: The cerebras story," *IEEE Micro*, vol. 41, no. 6, pp. 52–57, 2021.

[41] J. Lee, P.-C. Chiang, P.-J. Peng, L.-Y. Chen, and C.-C. Weng, "Design of 56 gb/s nrz and pam4 serdes transceivers in cmos technologies," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 9, pp. 2061–2073, 2015.

[42] T. G. Lenihan, L. Matthew, and E. J. Vardaman, "Developments in 2.5d: The role of silicon interposers," in *2013 IEEE 15th Electronics Packaging Technology Conference (EPTC 2013)*, 2013, pp. 53–55.

[43] C. Liang, Y. Dai, W. Xu, and Q. Li, "Advanced architecture design of high-radix router based on chiplet integration and ip reusability," in *2021 IEEE 23rd Int Conf on High Performance Computing Communications; 7th Int Conf on Data Science Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud Big Data Systems Application (HPCC/DSS/SmartCity/DependSys)*, 2021, pp. 123–132.

[44] O. Liboiron-Ladouceur, A. Shacham, B. A. Small, B. G. Lee, H. Wang, C. P. Lai, A. Biberman, and K. Bergman, "The data vortex optical packet switched interconnection network," *Journal of Lightwave Technology*, vol. 26, no. 13, pp. 1777–1789, 2008.

[45] J. F. McDonald, E. H. Rogers, K. Rose, and A. J. Steckl, "The trials of wafer-scale integration: Although major technical problems have been

overcome since wsi was first tried in the 1960s, commercial companies can't yet make it fly," *IEEE Spectrum*, vol. 21, no. 10, pp. 32–39, 1984.
[46] W. Nakayama, "Exploring the limits of air cooling," Aug 2006. [Online]. Available: https://www.electronics-cooling.com/2006/08/exploring-the-limits-of-air-cooling/
[47] A. Ovrashko, *Cisco Switching Portfolio Update*, 2021. [Online]. Available: https://www.cisco.com/c/dam/m/ru_ua/training-events/2021/cisco-tech-talks/pdf/switching.pdf
[48] S. Pal, D. Petrisko, A. A. Bajwa, P. Gupta, S. S. Iyer, and R. Kumar, "A case for packageless processors," in *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2018, pp. 466–479.
[49] S. Pal, D. Petrisko, M. Tomei, P. Gupta, S. S. Iyer, and R. Kumar, "Architecting waferscale processors - a gpu case study," in *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2019, pp. 250–263.
[50] L. Poutievski, O. Mashayekhi, J. Ong, A. Singh, M. Tariq, R. Wang, J. Zhang, V. Beauregard, P. Conner, S. Gribble, R. Kapoor, S. Kratzer, N. Li, H. Liu, K. Nagaraj, J. Ornstein, S. Sawhney, R. Urata, L. Vicisano, K. Yasumura, S. Zhang, J. Zhou, and A. Vahdat, "Jupiter evolving: Transforming google's datacenter network via optical circuit switches and software-defined networking," in *Proceedings of ACM SIGCOMM 2022*, 2022.
[51] J. M. Rabaey, *Digital Integrated Circuits: A Design Perspective*. USA: Prentice-Hall, Inc., 1996.
[52] M. Ruiz-Sanchez, E. Biersack, and W. Dabbous, "Survey and taxonomy of ip address lookup algorithms," *IEEE Network*, vol. 15, no. 2, pp. 8–23, 2001.
[53] D. Schor, "Tsmc announces 2x reticle cowos for next-gen 5nm hpc applications," May 2021. [Online]. Available: https://fuse.wikichip.org/news/3377/tsmc-announces-2x-reticle-cowos-for-next-gen-5nm-hpc-applications/
[54] D. Seo, A. Ali, W.-T. Lim, and N. Rafique, "Near-optimal worst-case throughput routing for two-dimensional mesh networks," in *32nd International Symposium on Computer Architecture (ISCA'05)*, 2005, pp. 432–443.
[55] T. J. Seok, K. Kwon, J. Henriksson, J. Luo, and M. C. Wu, "Wafer-scale silicon photonic switches beyond die size limit," *Optica*, vol. 6, no. 4, pp. 490–494, Apr 2019. [Online]. Available: https://opg.optica.org/optica/abstract.cfm?URI=optica-6-4-490
[56] A. Shpiner and E. Zahavi, "Race cars vs. trailer trucks: Switch buffers sizing vs. latency trade-offs in data center networks," in *2016 IEEE 24th Annual Symposium on High-Performance Interconnects (HOTI)*, 2016, pp. 53–59.
[57] A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of cmos device performance from 180nm to 7nm," *Integration*, vol. 58, pp. 74–81, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167926017300755
[58] L. T. Su, S. Naffziger, and M. Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017, pp. 1.1.1–1.1.8.
[59] E. Talpes, D. Williams, and D. D. Sarma, "Dojo: The microarchitecture of tesla's exa-scale computer," in *2022 IEEE Hot Chips 34 Symposium (HCS)*, 2022, pp. 1–28.
[60] T. S. Team, "What is signal propagation delay in a pcb?" Nov 2020. [Online]. Available: https://www.protoexpress.com/blog/signal-propagation-delay-pcb/
[61] F. Yao, J. Wu, G. Venkataramani, and S. Subramaniam, "A comparative analysis of data center network architectures," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 3106–3111.