

I/O Architecture, Substrate Design, and Bonding Process for a Heterogeneous Dielet-Assembly based Waferscale Processor

Saptadeep Pal*, Irina Alam*, Krutikesh Sahoo*, Haris Suhail*, Rakesh Kumar†, Sudhakar Pamarti*, Puneet Gupta* and Subramanian S. Iyer*

*Electrical and Computer Engineering, University of California, Los Angeles, CA 90095, USA

†Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL 61820, USA

Email: saptadeep@ucla.edu

Abstract—Demand for large amounts of parallelism is growing rapidly for today's computing systems. This is due to the proliferation of applications such as graph processing, data analytics, machine learning, etc. which require a large number of processing cores and a large amount of memory bandwidth. Often systems comprising of many individual packaged chips are employed to run these applications. However, inter-package communication has not scaled well and this bottleneck threatens the performance scaling of these applications. One way to alleviate this bottleneck is to build waferscale processors where many compute cores and memory blocks can communicate efficiently at very high bandwidths. In this work, we attempt to build a many-core waferscale processor using heterogeneous dielet assembly on the Silicon Interconnect Fabric (Si-IF) technology. The design and implementation of a dielet based waferscale processor have their own set of challenges. Some of the challenges include (1) design of area and energy efficient highly parallel I/O cells, (2) Si-IF substrate design and its impact on signaling and power delivery, and (3) reliable and efficient dielet-to-wafer bonding process. In this work, we will discuss the solutions to these three challenges that we employed in our dielet and Si-IF substrate design. Our custom-designed I/O cell is only $157.8\mu m^2$, which is 95% smaller than the standard cell I/Os and consumes only about 0.075pJ/bit. We co-designed the dielets with the Si-IF substrate to ensure that we can achieve on-chip like communication characteristics for inter-dielet communication. This helps us to seamlessly partition a large design into fine-grained dielets. For delivering power to the dielets across an entire wafer, we use power delivery from the edge of the wafer. This scheme results in large resistive power loss, and as a result, we designed a novel power management unit on each tile to provide reliable power to the core circuitry in the dielets. Lastly, we briefly discuss the copper-gold bonding process and the heterogeneous dielet assembly scheme we developed for an efficient and reliable assembly process. Shear tests show that the bond strength achieved with this process is ~ 113.3 MPa which is $>5\times$ compared to the previously reported bond strength for gold-gold bonding.

Keywords—Waferscale Processors, Silicon Interconnect Fabric, Dielet Assembly

I. INTRODUCTION

In recent times, there has been a rapid proliferation of highly parallel workloads such as graph processing, data analytics, machine-learning, etc. that are driving the need for a large number of processing cores, large memory capacity and high bandwidth in today's high performance computing systems [1],

[2]. These applications are often run on systems comprising of multiple discrete packaged processors connected using conventional off-package communication links through PCBs and between PCBs. The inter-package communication links are one of the major bottlenecks in today's systems due to their much poorer energy efficiency and bandwidth compared to that of the on-die links that is limiting the performance scaling of these applications [3]. This is because though Moore's Law has helped shrink the on-chip features by $>1000\times$ over the last four decades, off-chip package components have scaled by merely about $4\times$ [4]. To push higher bandwidth between packaged components where interconnect wiring is sparse, data rate per wire/link needs to be increased. This is done using high-speed I/O circuitry in combination with serialization and de-serialization (SerDes) schemes [5]. The SerDes circuitry is used to convert low frequency parallel data interfaces inside the dies to high-speed serialized interfaces required for high bandwidth communication between the packages.

Such a SerDes based scheme comes with its own challenges. First, the area taken up by the complex I/O circuitry to support chip-to-chip communication is often large and already exceeds 25% on some of today's processors and power overhead of such I/O can often exceed 30% [4]. Moreover, large communication latency is incurred which often results in significant bottlenecks to multi-chip performance scaling. While consuming large amount of power, area and latency, the off-chip bandwidth still lags on-chip bandwidth by up to 50x. Recent advances in multi-chip module (MCM) [6], [7] and interposer technology [8] have targeted this mismatch and these technologies can integrate multiple processor and memory dies tightly inside a package by inserting a new level of inter-dielet interconnection which provides high-bandwidth and low-latency. Examples of such technologies includes TSMC CoWoS [9] and Intel's EMIB [10]. Though these technologies alleviate some of the issues of conventional single-die packages, they are still constrained by the size limit and can accommodate only a few dies within one package. A scale-out high performance system today therefore needs to integrate many multi-die packages on a PCB or across multiple PCBs to satisfy the compute and memory needs of modern workloads. There again, the off-package intra-PCB and

between-PCB links become significant bottlenecks to system performance scaling.

Waferscale processor systems can significantly reduce this communication overhead and can satisfy the bandwidth demand by tightly integrating a large number of processor and memory dies on a wafer [11], [12]. This has led to the renewed interest in waferscale integration. Recently, monolithic waferscale systems have shown promising performance and energy benefits [11], [13]. However, there are two main short-comings with monolithic waferscale systems: (1) They are homogeneous in nature and, hence, cannot integrate heterogeneous non-CMOS technologies like DRAM that would provide better cost-performance trade-off opportunities, and (2) a single silicon chip as large as a wafer suffers from yield issues and, hence, their architecture needs to be designed to tolerate defects and often non-negligible system area has to be devoted towards redundant cores and other resiliency schemes to obtain reasonable yield. For example, Cerebras' architecture comprises a swarm of tiny cores connected using a mesh network, and redundancy is employed in both core count and network links. Though such an architecture is well suited to a certain class of workloads, other architectures where each micro-architectural unit may be large would not yield well when using monolithic waferscale integration and redundancy will come at a higher cost.

An alternate approach to building waferscale systems is to use a dielet-based approach where many pre-tested bare dielets, both compute and memory, are integrated on a passive silicon interconnect wafer like the Silicon Interconnect Fabric (Si-IF) technology [14]. Si-IF allows the bare un-packaged dielets to be tightly integrated and communicate using fine-pitch ($10\mu\text{m}$ pitch), high density parallel links. Use of highly parallel links can eliminate the need for SerDes circuitry and these links can be run at lower frequencies. As a result, area and power hungry high speed I/Os can be replaced with efficient low speed I/Os. Overall, such a scheme can dramatically cut down the latency of inter-chip communication and can provide close to on-chip latencies for inter-chip communication, thus enabling better performance scalability. Also, such an interconnect scheme allows us to partition a large design in to fine-grained dielets without dramatic hit in performance. And, the small dielets can be individually tested before bonding. Since the Si-IF technology uses a purely passive interconnect-only substrate, it can be built at very high yields. Coupled with high yield of bonding [14], [12], dielet based waferscale systems can be manufactured at very high yield. Overall, the dielet based waferscale approach allows us to achieve heterogeneous integration and has the potential to improve performance scalability and cost benefits of a many-core system.

However, building such a large dielet-based waferscale system has several challenges and requires several design decisions to be taken ranging from customizing the I/O architecture, delivering power across the entire wafer to finally reliably bonding and assembling the different dielets on the Si-IF wafer. In this paper, we will discuss some of the details and

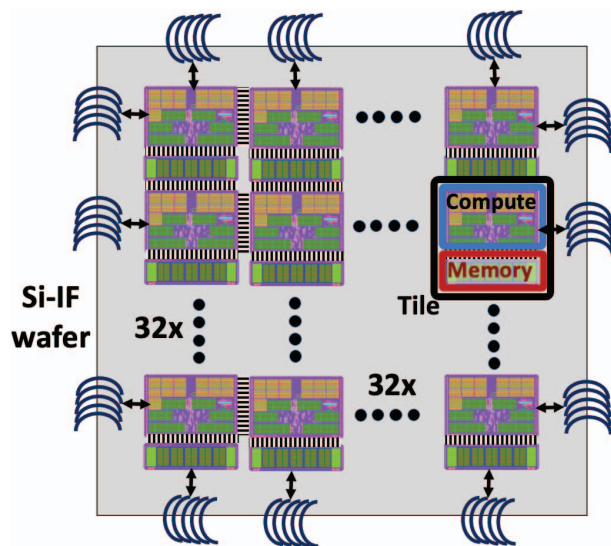


Fig. 1: Overview of the waferscale processor system is shown. A compute and a memory dielet form a tile and an array of 32×32 tiles are assembled on an Si-IF wafer.

challenges we faced during building a dielet based waferscale processor prototype. The waferscale processor system is a tiled array of processors, where each tile comprises a compute dielet and a memory dielet. We briefly discuss the overall waferscale processor system architecture in Section II and the details of the compute and memory dielets that we designed. In Section III, we will cover the motivation behind designing the I/O cells from ground-up in the compute and memory dielets used for inter-dielet communication and the details of the final transmitter/receiver cells. Section IV covers the details of the Si-IF substrate design, the challenges of reliably delivering power through the substrate in such a large system and the power management unit that was designed to circumvent the issues. In Section V, we discuss the novel copper-gold bonding scheme that will be used to assemble the large waferscale processor system and other optimizations that were done to the overall assembly process. Section VI concludes the work.

II. WAFERSCALE PROCESSOR SYSTEM OVERVIEW

The waferscale architecture is comprised of an array of logical compute-memory tiles. Each tile comprises a compute dielet and a memory dielet manufactured on the TSMC 40nm-LP process node. The compute dielet has an area of 7.56mm^2 and consists of 14 ARM-Cortex M3 cores [15], a private SRAM scratchpad memory for each core, network routers, clock distribution circuitry, and a power management unit. The memory die has an area of 3.47mm^2 and consists of SRAM based shared memory blocks that are accessible by any processing core on the entire wafer.

The components inside the compute die are connected using an intra-tile crossbar interconnect while the array of tiles across the wafer is connected through a mesh network. The node routers of the inter-tile mesh network reside on the compute die. Hence, each compute die has parallel links escaping from

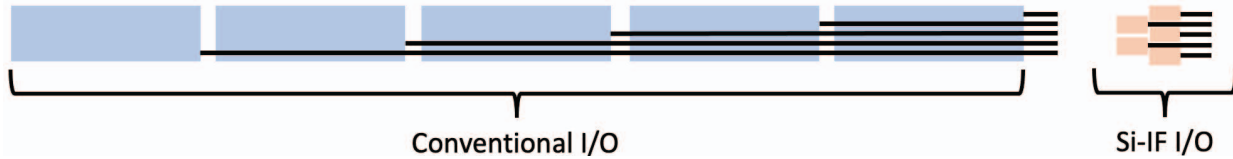


Fig. 2: The difference between big and small I/O cells is shown. To attain high interconnect density while attaining small interconnect length, I/O size needs to be small.

all four sides that allow connection to the compute dies in the four neighboring tiles. The memory dielet provides buffered feedthroughs for the north-south interconnect links only and, therefore, has links escaping only from the top and the bottom edges of the dielet. The communication between the dies on the wafer is accomplished using the SuperCHIPS-like [16] highly parallel interface which uses fine pitch interconnects on the Si-IF substrate. This allowed us to design a waferscale processor using small-sized high-yielding dielets which are tightly connected using a large-sized passive silicon substrate.

Figure 1 shows the schematic of the overall waferscale processor system. The 32×32 tile prototype comprising a total of 14336 cores, 64KB of private memory per core and 512MB of globally shared memory occupies a total area of 15100mm^2 and consumes total peak power of 725W. In the subsequent sections we will discuss the some of the design challenges and their solutions in detail.

III. I/O DESIGN

Si-IF technology allows us to build highly parallel and dense inter-dielet communication interfaces where interconnects of length as small as $100\mu\text{m}$ need to be driven. As a result, low latency and high bandwidth communication is possible. However, to fully utilize the benefits that the Si-IF technology provides, we had to design the I/Os from the ground-up for energy and area efficiency. In this section, we discuss the design decisions we took and the justification behind the decisions.

On-die like low-latency and high-bandwidth communication interface between dielets allows us to partition a large design in to fine grained dielets. Therefore, we wanted the dielets to be small in order to reduce design complexity and achieve high die yields. As a result, we needed to integrate a large number of I/O cells (2020 I/Os in the compute dielet and 1250 I/Os in the memory dielet) in each small-sized dielet. This requires us to minimize the aggregate area overhead of these I/O cells. The I/O cells in the standard cell library available to us were of the size of $120\mu\text{m} \times 25\mu\text{m}$. Using these standard I/O cells would result in two issues: (a) The total area that would need to be dedicated to the I/Os per dielet would be very large, 6.06mm^2 and 3.75mm^2 in the compute and memory dielets respectively. This would have resulted in unacceptable area overhead ($> 80\%$). (b) To support fine pitch wiring, five I/O rows would need to be supported (as shown in Figure 2) which would have significantly increased the inter-dielet interconnect lengths, and would have negatively impacted the I/O energy efficiency which is promised by the Si-IF technology.

An I/O usually is comprised of two components: (1) Transmitter/Receiver (Tx/Rx) circuitry (2) ESD diodes. The Tx/Rx

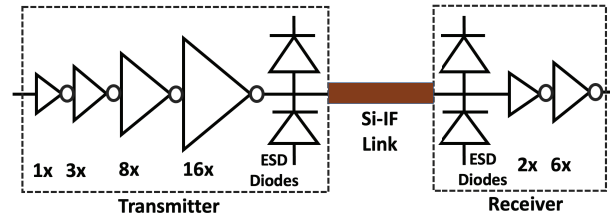


Fig. 3: Schematic of the transmitter and the receiver cells are shown.

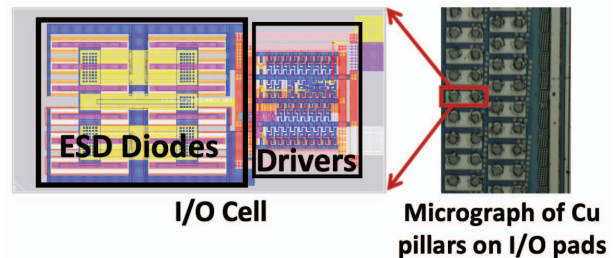


Fig. 4: Internal layout of the I/O cell and the placement of the cells along the edge of the dielet are shown. The micrograph on the right also shows two copper pillars attached per I/O pad.

circuitry is usually designed to drive long interconnect distances (usually 10s of millimeters) which results in large driver sizes. Moreover, ESD diodes are designed so that the chip can withstand ESD events which correspond to high voltage (2-8kV) human body model (HBM). Both these factors are responsible for large I/O sizes. Si-IF technology allows dielets to be placed with an inter-dielet spacing of $50\text{-}100\mu\text{m}$ [16] and therefore, if the I/O cell size can be reduced dramatically, the interconnect distance that needs to be driven can be as small as $100\mu\text{m}$. As a result, the drivers can be miniaturized. Also, since the ESD requirements for bare-dielet to waferscale substrate bonding only need to tolerate less stringent ESD events, such as 100V HBM or machine model (MM), the ESD circuitry can be significantly stripped down.

Figure 3 shows the schematic of the fine-pitch I/O cells we designed. The I/O cell is designed such that it can fit under a single I/O pad ($157.8\mu\text{m}^2$ only) which connects to two copper pillars on the Si-IF side, leading to a 95% reduction in I/O cell area compared to that of the cells in the standard library. The driver circuitry is designed using appropriately sized cascaded buffers from the standard cell library. Each I/O cell contains 12 ESD diodes in parallel. Overall, the ESD diodes add only about 18fF of capacitive load, and as such, this I/O cell can drive links of length $300\mu\text{m}$ or more at 1GHz with eye-width

opening of >0.95 UI. Using SPICE simulation of the I/O cells with pseudo-random bit stream as input (PRBS), we estimated that the energy per bit of communication would be roughly 0.075pJ/bit . To maximize I/O density along the chip periphery, we designed the I/O pad such that two copper pillars (in each pad) are placed orthogonal to the die edge as shown in Figure 4. The interconnect wiring on the Si-IF connecting two dies has a $5\mu\text{m}$ pitch and the I/Os on the chip have a $10\mu\text{m}$ pitch. Therefore, we are able to hit the maximum wire density per layer using just two staggered columns of I/O cells. Also, the smaller I/O cells help to reduce the worst case inter-dielet link length by $\sim 566\mu\text{m}$ compared to the conventional I/O cells that would have required 5 staggered columns to hit the maximum wire density (see Figure 2). This reduction helps to significantly improve the I/O energy efficiency.

IV. SUBSTRATE DESIGN AND POWER DELIVERY MECHANISM

The Si-IF technology currently supports up to four metal layers. Therefore, both inter-dielet signaling and power distribution have to be done using these four metal layers only. The architecture was designed such that no serialization is required for inter-dielet communication. This is to ensure that similar to on-die network hop latency is achievable in our design where the network is split across the dielets. As a result, a large number of interconnect wires (2020 for compute dielet and 1250 for memory dielet) need to fit within the periphery of the dielets. The signal wire density is highest between a tile's compute and memory dielets and about 800 wires needs to be accommodated on the southern edge of the compute dielet. Since the width of the compute dielet is 2.4 mm and the wiring density per layer is 200 wires per mm, two layers are required to sustain the ~ 800 wires at that interface. As a result, only the bottom two layers are primarily available for power delivery. This has implications on the power delivery regulation for the waferscale processor.

There are two ways to deliver power to all the dielets in the waferscale processor. One is to deliver power from the backside of the wafer using through wafer vias (TWVs) [17], and the other is to deliver power from the edge of the wafer. Using a backside power delivery allows one to provide regulated core voltage directly to the dielets. However, backside power delivery requires a complex TWV process which we decided to avoid for this prototype. Since this is a sub-kW system, we decided to reduce complexity and supply power from the edge of the wafer. The current therefore traverses through the bottom metal layers from the edge to the center of the wafer which leads to unwanted voltage droops and results in resistive power loss. In order to minimize the voltage droop, the power planes are designed as dense slotted planes such that the resistive power loss is minimized. Even after dedicating two entire thick ($2\mu\text{m}$) metal layers and some portions of the second signal wiring layer, our calculations show that the central dies will get a maximum rail-to-rail voltage droop of ~ 1 V compared to the edge dies (see Figure 5). The standard cells in the 40nm LP library require a nominal operating voltage of 1.1V and can

operate reliably at up to a maximum of 1.3V. In order to ensure that the central dielets receive power at least at 1.4V, power at the edge has to be supplied at 2.5V and a power regulator would be required in each dielet to regulate the input voltage and output an operating voltage in the range of 1.0V - 1.2V for the core devices. Since the same compute and memory dielets would be used across the entire wafer, we could not design custom regulators for the dielets based on their location in the wafer. Hence, the power regulator had to be designed to support a very wide input voltage range of 1.4V to 2.5V.

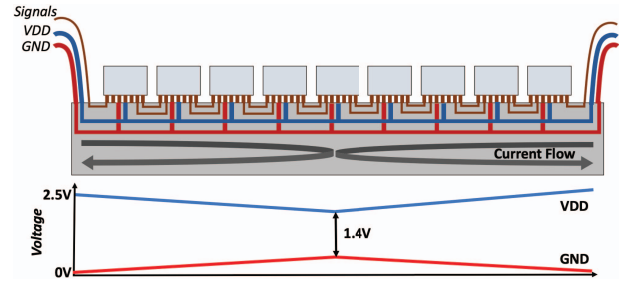


Fig. 5: Edge power delivery scheme is shown. The dielets towards the center of the wafer get power at a lower rail-to-rail voltage than the dielets at the edge because of resistive power loss in the waferscale power plane.

We designed a custom low-dropout (LDO) regulator based power management unit for this purpose. The schematic of the LDO that was incorporated inside the compute dielet is shown in Figure 6. The second stage of the LDO is a pFET buffer, the gate of which tracks the supply voltage. Traditional differential amplifiers would not be able to track this wide input voltage variation, causing the FETs to go into triode state at lower supply voltages as they try to produce an output voltage required by the buffer. We resolved this by ensuring that the output of the first stage amplifier also tracks the supply voltage similarly to the buffer. This was done by adding a scaled replica of the buffer into the amplifier. This allowed the amplifier to track the supply, and the LDO to work over a wide supply voltage range.

The digital load in the dielets consumes 350mW of peak power and the PMU has to sustain up to 200mA current demand change within a few tens of cycles. This requirement led to the second big challenge we faced during the LDO design. Lack of large decoupling capacitors on the wafer, especially for the dielets away from the edge meant that sudden current spikes would result in voltage instability. The capacitance from the digital load was insufficient and therefore, we devoted roughly 30% area of the dielets to additional on-chip decoupling capacitors built using metal-oxide-metal (MoM) capacitors on metal-oxide-semiconductor (MoS) for a total of 20nF per tile. This additional capacitance however wasn't enough and because the dominant pole of the LDO was at the output it was difficult to achieve the required stability and accuracy using a conventional LDO design. We considered a miller capacitance compensation on an internal node, but this would reduce the

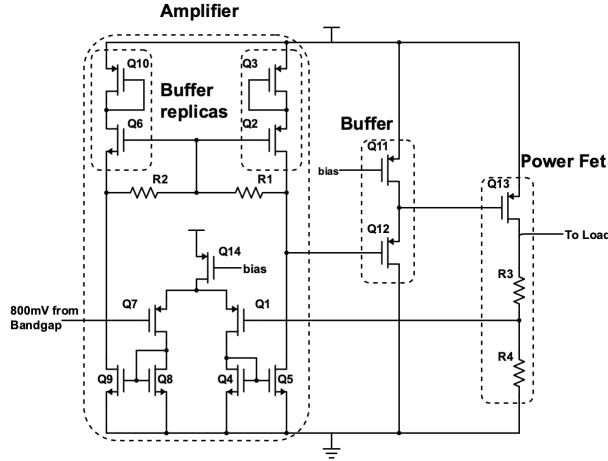


Fig. 6: Schematic of the LDO based power management unit (PMU)

bandwidth of the LDO below the requirements. Instead, we resolved it by reducing the gain of the first stage differential amplifier of the LDO by using additional poly-resistors across its terminals. The reduction in DC gain causes a reduction in the regulation accuracy and increased power consumption of the LDO, however it ensures that across all process-voltage-temperature (PVT) corners, the gain is always enough to keep the accuracy between $\pm 10\%$. Overall, this ensured that the regulated voltage is always between 1.0V and 1.2V, and the LDO could sustain the current variation requirements. In the future, this problem can be alleviated by introducing deep-trench capacitors in the Si-IF wafer itself. It would not only provide $>10\times$ decoupling capacitance [18] which would help with better power regulation, it would also help reduce the area overhead of integrating on-die decoupling capacitance.

V. SI-IF BONDING AND ASSEMBLY

In this section, we discuss the dielet-to-Si-IF wafer bonding and assembly process that we employ for this processor system. To connect the dielets to the wafer, fine-pitch copper pillars are formed on the Si-IF wafer that attaches to the I/O pads on the dielets. Unlike conventional dielets where the back-end of the dielets are built up to the Aluminium I/O pads, for direct metal-to-metal bonding, we terminate the back-end of our dielets at the top most copper metal layer. As shown in Figure 4, each I/O pad was designed to accommodate two copper pillars. The I/O pad pitch along the edge of the dielet was kept at $10\mu m$ and has a width of $7.2\mu m$.

On the Si-IF side, we use $3\mu m$ wide pillars (at $10\mu m$ pitch) instead of $5\mu m$ wide pillars that has been proposed previously [19]. The reduction in pillar width has two primary benefits: (1) It allows greater tolerance to alignment error that can cause pillars to short to mis-matched pads, and (2) It allows us to minimize bond force while retaining the pressure required for successful thermo-compression bonding (TCB). This is because smaller area results in larger pressure for a given force.



Fig. 7: Micrograph of sheared-off dielet shows excellent alignment accuracy during bonding.

Unlike previous works which either used gold-to-gold (Au-Au) [20] or copper-to-copper (Cu-Cu) [19] bonding, bonding many dielets to a waferscale Si-IF substrate reliably across the entire substrate area needs a different approach. Au-Au bonding isn't possible in our case since Au termination of the dielet pads was not available with the TSMC N40LP process we used. Since the dielets were terminated with Cu-pads, one option is to use Cu-Cu bonding. During the bonding process, the Si-IF substrate needs to be kept at elevated temperature. The dielets are bonded sequentially on to the Si-IF substrate and each dielet may take 10-30 sec [19] depending on the bonding process used. Since a large number of dielets need to be assembled and bonded, the Si-IF substrate would remain at the elevated temperature for a long duration. This results in Cu oxidation of the Cu-pillars. Though prior work in dielet-to-wafer Cu-Cu bonding used localized reduction to eliminate the oxide that forms on the copper pads and pillars, the long exposure of the Cu-pillars to the ambient conditions and elevated temperature leads to large amount of oxidation and the reduction process falls short. As a result, this can then lead to poor bonds and may even lead to complete bonding failure of the dielets which are bonded later in sequence.

To circumvent this issue, we developed a process where the Si-IF side copper pillars are capped with a thin layer of gold (200nm) to prevent copper oxidation. We therefore developed a novel Cu-Au TCB bonding process. Overall, we achieved excellent bond shear strength of 113.3 MPa which is $>5\times$ compared to the previously reported shear strength in [20] which used Au-Au bonding and is comparable to the bond strength achieved using Cu-Cu bonding [19]. Further details about the novel Cu-Au bonding process and its characterization can be found in [21].

The other issue with this assembly process arose from the fact that each tile had two dielets of different sizes. The bonding tool we used can handle only a single sized die at once and the bonding alignment process needs to be tuned for each die size. We, therefore, first bond all the dies of the same size

on the wafer before bonding the dies of the other size. This mechanism requires only a single swap of the bond head tool and allows us to tune the bonding force only twice in the entire assembly process. Figure 7 shows the micro-graph of a post-bonding sheared dielet. We were able to achieve excellent alignment accuracy within $\pm 1\mu\text{m}$ for both the compute dielets and the memory dielets.

VI. CONCLUSION

Dielet assembly based waferscale processor systems have the potential to provide greater performance and energy scalability than today's processors built using conventional packaging technologies. In this work, we attempt to build a large dielet based waferscale system prototype that has an area of 15000mm^2 and is comprised of 14336 cores, 64KB of private memory per core and 512MB of SRAM-based globally shared memory. The dielets are manufactured in TSMC 40nm-LP process node and are tightly integrated using the passive Silicon Interconnect Fabric (Si-IF) technology that allows the dielets to communicate through high density, fine-pitch parallel links. We discussed, in detail, the I/O cell architecture that we designed from ground-up to reduce the area overhead of the I/O cells by 95%, improve the I/O energy efficiency and reduce the inter-dielet link length compared to the I/O cells provided in the 40nm-LP standard cell library. We also discussed the challenges of delivering power in such a large system through the Si-IF substrate and the circuit level details of the low-dropout regulator based power management unit that we designed to reliably power all the dielets across the wafer. Finally, we discussed the bonding and assembly scheme which allows us to reliably and efficiently bond the dielets on to the large Si-IF wafer. Our ongoing work aims at assembling and characterizing the waferscale processor prototype.

ACKNOWLEDGEMENT

This work was supported in part by DARPA award N6600-12-C-4001, UCOP MRP-17-454999, the UCLA CHIPS consortium, CDEN and 2019-2020 Qualcomm Innovation Fellowship. The authors would like to thank Matthew Tomei, Nick Cebry, Jingyang Liu, and Shi Bu for their contribution to the architecture and design of the dielets, Alexander Graening for help with the test infrastructure, and TSMC for their support with tape-out and dielet manufacturing.

REFERENCES

- [1] "Workload Analysis of Blue Waters." <https://arxiv.org/ftp/arxiv/papers/1703/1703.00924.pdf>, (accessed Nov 23, 2020).
- [2] K. Shirahata, H. Sato, T. Suzumura, and S. Matsuoka, "A scalable implementation of a mapreduce-based graph processing algorithm for large-scale heterogeneous supercomputers," in *13th International Symposium on Cluster, Cloud, and Grid Computing*, pp. 277–284, 2013.
- [3] 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)*, pp. 466–479, 2018.
- [4] S. S. Iyer, "Heterogeneous integration for performance and scaling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 7, pp. 973–982, 2016.
- [5] D. R. Stauffer, J. T. Mechler, M. A. Sorna, K. Dramstad, C. R. Ogilvie, A. Mohammad, and J. D. Rockrohr, *High Speed Serdes Devices and Applications*. Boston, MA: Springer, 2008.
- [6] M. Ishida, "APX (Advanced Package X) - Advanced Organic Technology for 2.5D Interposer," 2014.
- [7] A. Arunkumar, E. Bolotin, B. Cho, U. Milic, E. Ebrahimi, O. Villa, A. Jaleel, C. Wu, and D. Nellans, "Mcm-gpu: Multi-chip-module gpus for continued performance scalability," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pp. 320–332, 2017.
- [8] 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)*, pp. 53–55, 2013.
- [9] Y. Chuang, C. Yuan, J. Chen, C. Chen, C. Yang, W. Changchien, C. C. C. Liu, and F. Lee, "Unified methodology for heterogeneous integration with cowos technology," in *2013 IEEE 63rd Electronic Components and Technology Conference*, pp. 852–859, 2013.
- [10] R. Mahajan, R. Sankman, N. Patel, D. Kim, K. Aygun, Z. Qian, Y. Mekonnen, I. Salama, S. Sharan, D. Iyengar, and D. Mallik, "Embedded multi-die interconnect bridge (emib) – a high density, high bandwidth packaging interconnect," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pp. 557–565, 2016.
- [11] K. Rocki, D. V. Essendelft, I. Sharapov, R. Schreiber, M. Morrison, V. Kibardin, A. Portnoy, J. F. Dietiker, M. Syamlal, and M. James, "Fast stencil-code computation on a wafer-scale processor," 2020.
- [12] S. Pal, D. Petrisko, M. Tomei, P. Gupta, S. S. Iyer, and R. Kumar, "Architecting waferscale processors - a gpu case study," in *IEEE International Symposium on High Performance Computer Architecture*, pp. 250–263, 2019.
- [13] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *International Symposium on Circuits and Systems*, pp. 1947–1950, 2010.
- [14] A. A. Bajwa, S. Jangam, S. Pal, B. Vaisband, R. Irwin, M. Goorsky, and S. S. Iyer, "Demonstration of a heterogeneously integrated system-on-wafer (sow) assembly," in *68th Electronic Components and Technology Conference*, pp. 1926–1930, 2018.
- [15] "ARM CPU Cortex M3." <https://www.arm.com/products/silicon-ip-cpu/cortex-m/cortex-m3>, (accessed February 27, 2021).
- [16] S. Jangam, S. Pal, A. A. Bajwa, S. Pamarti, P. Gupta, and S. S. Iyer, "Latency, bandwidth and power benefits of the superchips integration scheme," in *IEEE Electronic Components and Technology Conference (ECTC)*, May 2017.
- [17] M. Liu, B. Vaisband, A. Hanna, Y. Luo, Z. Wan, and S. S. Iyer, "Process development of power delivery through wafer vias for silicon interconnect fabric," in *69th Electronic Components and Technology Conference*, pp. 579–586, 2019.
- [18] K. T. Kannan and S. S. Iyer, "Deep trench capacitors in silicon interconnect fabric," in *IEEE 70th Electronic Components and Technology Conference*, pp. 2295–2301, 2020.
- [19] S. Jangam, A. A. Bajwa, U. Mogera, P. Ambhore, T. Colosimo, B. Chylak, and S. Iyer, "Fine-pitch ($\leq 10\mu\text{m}$) direct cu-cu interconnects using in-situ formic acid vapor treatment," in *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pp. 620–627, 2019.
- [20] A. Bajwa and S. Jangam and S. Pal and N. Marathe and T. Bai and T. Fukushima and M. Goorsky and S. S. Iyer, "Heterogeneous Integration at Fine Pitch ($\leq 10\mu\text{m}$) Using Thermal Compression Bonding," in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, pp. 1276–1284, May 2017.
- [21] K. Sahoo, S. Pal, N. Shakoorzadeh, Y.-T. Yang, and S. S. Iyer, "Copper to gold thermal compression bonding in heterogenous wafer-scale systems," in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, 2021.