Title
Advanced packaging and heterogeneous integration to reboot computing

Authors
Pal, Saptadeep
Iyer, Subramanian S.
Gupta, Puneet

Publication Date
2017-11-01

Peer reviewed

Advanced Packaging and Heterogeneous Integration to Reboot Computing

Saptadeep Pal, Subramanian S. Iyer and Puneet Gupta
Center for Heterogeneous Integration and Performance Scaling (CHIPS),
Electrical and Computer Engineering Department, University of California Los Angeles,
{saptadeep, s.s.iyer, puneetg}@ucla.edu

***Abstract*—In the past several decades on-chip dimensions have scaled over 2000X, while dimensions on printed circuit board have scaled 4-5X. This modest scaling of packaging dimensions has severely limited system scaling. To address this, we have proposed a disruptive package-free integration scheme. We replace the traditional organic printed circuit board (PCB) with silicon interconnect fabric (SiIF) and replace the traditional package by directly mounting bare chiplets on to the SiIF. Fine pitch solderless copper pillar connections increase IO density by 20-80X and the inter-chiplet spacing is reduced by 10-20X. This enables highly parallel communication instead of serialized links. This achieves higher bandwidth/mm (~100X) and lower latency (~25X) and lower communication energy per bit (~200X). This integration technology allows us to challenge the conventional communication-limited architectures in a substantial way. The ability to heterogeneously integrate diverse dies with arbitrarily fine granularity, but on a wafer scale, reduces the cost of processor-memory communication energy opening new compute paradigms. In addition, the superior heat spreading properties of the SiIF compared to organic PCBs allows us to run the cores harder. The heterogeneous integration property of our scheme, allows for an intimate mingling of heterogeneous processor cores, FPGAs and memory types opening new avenues to reboot computing.**

Keywords—Silicon Interconnect Fabric (SiIF); Thermal Compression Bonding; Fine Pitch Interconnect, Wafer-scale Integration, Semiconductor Packaging

I. INTRODUCTION

Aggressive scaling of minimum silicon features has resulted in increased transistor density every technology generation [1]. Alongside, with the development of design [2, 3] and test infrastructure [4, 5], increased levels of integration are made possible using the system-on-chip (SoC) technology where multiple system components/ modules are built and integrated on a single monolithic piece of silicon. However, demand for ever-larger systems with intimately connected, diverse components has just been increasing. Moreover, incompatibility of efficient implementation of many of these components with silicon (passives, power regulators, oscillators, etc.) have prevented full-system scaling.

Traditionally, individual chips are packaged, and multiple packages are integrated using a printed circuit board (PCB) [6]. The packages are mounted on the PCB using either pin grid arrays, ball-grid arrays or land grid arrays [7, 8, 9]. The dimension of these connections on the package have only scaled by about a factor of 4 while the minimum feature size on chip has scaled by over a 2000x [10]. This constrains the total number of signal input/output and power delivery connections that can be accommodated on to a single package, thus limiting the bandwidth and total power that can be delivered to a chip. As a result, larger packages need to be built to accommodate the required IOs. The package-to-silicon die area ratio can be as large as up to 20x [11]. The interconnect connecting multiple dies in separate packages need to traverse the packages and the board level traces. As a result, these interconnect traces run a few millimeters to a few centimeters leading to increased latency and energy of communication. Moreover, since PCBs are manufactured using organic materials, the traces on the PCBs have larger pitch as compared to on-chip interconnects. To increase the bandwidth, these links are driven using serialization and deserialization circuits, commonly known as SerDes. However, such serialized communication techniques increase the energy per bit dramatically, as much as up to 30% of total chip power.

Today's high-performance systems demand on-chip like bandwidth and latency for system level interconnects. Though several past efforts have been made to build very large wafer scale systems to realize better system performance and reduce cost of packaging, manufacturing yield of such massive systems, interconnect reliability and across the wafer variation have made such large systems impractical [12]. Other modern approaches of multi-chip assembly such as interposers allow integration of multiple bare chiplets on a silicon sub carrier [13, 14]. However, the maximum size is limited due to the fragile nature and yield issues of thinned interposers and thus can only hold a few chiplets. A full system implementation using interposer still requires the interposer assembly to be packaged and integrated to other components using PCB.

Our approach is to replace the PCB with a thick silicon substrate (we call it a silicon interconnect fabric, SiIF) and mount bare chiplets directly on the substrate using fine-pitch interconnects. This helps eliminate packages

and bring down the inter-chiplet distance significantly to less than a hundred microns. Moreover, global on-chip interconnect like wiring pitches can be easily achieved on the SiIF using mature back-end-of-line (BEOL) fabrication technologies. Thus, high speed serialized links to communicate between different chiplets can be now replaced with parallel communication interface where each wire runs at a lower speed. Our analysis shows that such an approach can result in dramatic improvements in interconnect energy, latency of communication and available bandwidth/millimeter of die edge compared to traditional PCB based systems. Elimination of packages helps reduce the system footprint and weight significantly. Moreover, since the PCB is now replaced with silicon substrate which has better thermal conductivity properties, heat sinking can be potentially done from both top of the chiplets as well as back side of the SiIF. This potentially improves sustainable thermal design power (TDP) by up to 60% which can be leveraged to increase the power and performance of a system. This technology also allows decomposition of an SoC into multiple component chiplets, possibly in different technologies, and reintegrate them tightly without significant loss in performance. This has significant implication on cost, as smaller chiplets yield better than the large SoCs and chiplet assembly allows the use of older, cheaper processes for non-critical portion of a system. Further, we expect chiplet assembly to substantially increase hard, pre-fabricated IP reuse as the chiplet ecosystem evolves.

II. SiIF Technology

Silicon Interconnect Fabric technology allows integration of multiple bare silicon dies on a silicon substrate using very fine pitch copper pillar based IO. The SiIF is fabricated using a conventional and mature BEOL process which can have up to four levels of conventional copper dual damascene based interconnect [15]. The wire pitch on the SiIF can be as low as few hundred nanometers to a few microns. The copper pillars manufactured using a damascene process are 2 - 5 μm diameter on the SiIF. The bonding of copper pillars on the SiIF with the IO pads on the chiplet is accomplished using a direct metal-to-metal thermal compression bonding process. Unlike solder based balls/bumps which suffers from extrusion, copper pillars don't have the extrusion problem and thus dense pillar spacing is possible. Since, we use a thick silicon wafer (unlike thinned interposers), chemical-mechanical polishing (CMP) process can be done on the SiIF which is essential to obtain very smooth copper surface required for a good contact. In addition, thermal expansion mismatches between multiple material involved in die-package-board connections are largely absent in SiIF resulting in reliability improvements.

We designed and developed test chiplets and SiIF to measure the electrical and mechanical quality of the bonds. An array of pillars with 5 μm diameter copper pillars was built on each chiplet and pillars on each pillar row was connected using a serpentine structure through the SiIF (shown in Figure 1(c)) to check for electrical continuity across the chiplet. Electrical tests on early test prototypes have shown >97% yield with low contact resistance of about 42 mΩ for 5 μm diameter copper pillars [16]. The average sheer strength of our metal-metal bonds was found to be higher than that of state-of-the-art Sn-capped copper pillar bonds with annealing time of ~6 min. More details regarding fabrication process steps of the SiIF substrate and metal-to-metal bonding used can be found in [16].

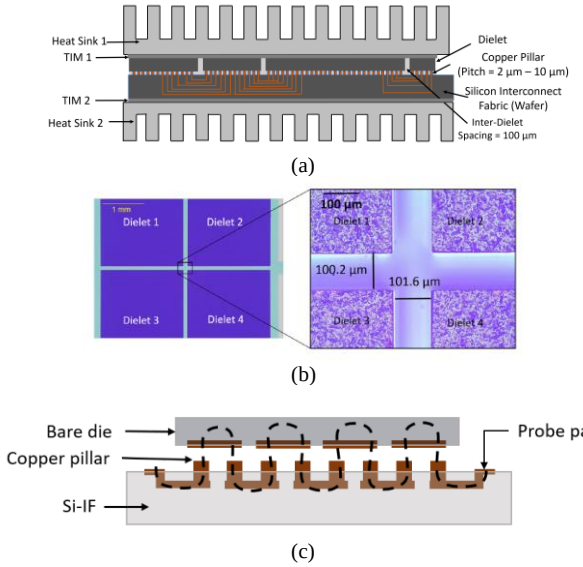

Figure 1: (a) Schematic cross-section of a system assembly on SiIF. (b) Micrograph of multiple chiplets on SiIF with ~100μm spacing [16] (c) Serpentine test structure with 5 μm diameter pillar

Figure 1(a) shows a schematic cross-section of SiIF based system assembly. In Figure 1(b), we show the micro graph of multiple chiplets bonded on to an SiIF, where the chiplets are paced with inter-chiplet distance of 100 μm. Different chiplets can come from different technologies, for e.g. Si, GaN, InP, SiC etc. In fact, an SoC can be disintegrated and multiple component chiplets can come from different technology nodes to optimize for cost and power. For e.g., a processor core can be in 32nm technology while L2 cache chiplet can be implemented in 22nm. Thus, SiIF allows for easy integration of heterogeneous technologies unlike in SoC where the whole die needs to be in a single technology node.

III. Benefits of SiIF Technology

In this section, we compare the bandwidth, latency and energy of communication, thermal, area form factor and weight benefits of the SiIF technology with traditional PCB based approaches.

A. Inter-Chiplet Communication Bandwidth, Latency and Energy Benefits

SiIF can accommodate up to 80x and 10x more number of IO pin connections compared to BGA based interconnections and solder based micro-bumps respectively. Since the length of the interconnect traces between the chiplets are about 100 μm, the wires on the

SiIF resembles closely the global interconnect wiring levels on the SoCs. The achievable data rate per wire can be up to 10 Gbps. Though the data rate per link is smaller than that of the SerDes implementation in PCBs, number of links available per millimeter edge of a chiplet is large. This provides a significant gain in bandwidth over the PCB based systems as shown in Figure 2.

Since the interconnect traces are very short and are driven at much lower frequencies than SerDes links, simple transceiver circuits such as buffer based drivers can be used. This reduces the circuit level overhead which is otherwise required to push data at much higher frequencies in serialized links. Elimination of the complex communication circuitry also reduces the effective latency of data transfer between two chiplets.

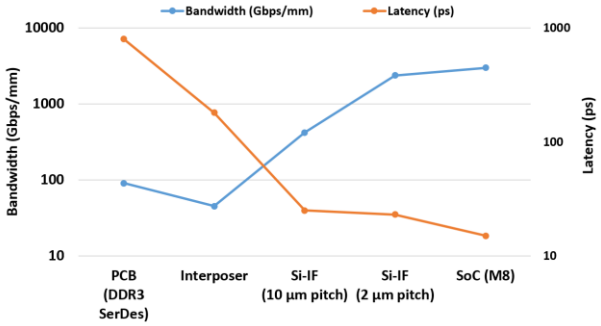

Figure 2: Bandwidth and Latency Comparisons

As shown in Figure 2., both SiIF bandwidth and latency numbers are very similar to the SoC ones and thus SiIF provides a platform for very high-performance system level integration. The overall latency depending on the driver design range from 50-100ps dominated by the external ESD protection capacitance assumed to be 50fF on each pad terminal. Without ESD, the latencies can go as low as 30-40ps. Detailed modelling and analysis of the interconnect characteristic is provided in [17].

Simple buffer based drivers and smaller capacitive load on each interconnect wire is expected to lower the energy per bit of communication as well. Our simulation analysis shows that <0.3 pJ/bit can be achieved on SiIF versus >5 pJ/bit and >10 pJ/bit required for interposer links and PCB links respectively [17].

B. Thermal Benefits

Silicon is a good conductor of heat unlike a PCB or a package. In case of PCB based system integration, about 80-90% of the heat is extracted through the top of the package and the rest gets dissipated through the PCB [18]. In case of SiIF, the thermal resistance from the chiplet to the heat sink is smaller than that of the chiplet to heat sink via package case by about 30%. Moreover, significantly more amount of heat can now be dissipated through the bottom silicon substrate as well. Thus, an additional backside heat sink (or silicon fins on the thick SiIF) can help efficiently extract the heat from the bottom side as well. In fact, the secondary bottom side heat would also act as a mechanical support/ protection for the SiIF substrate. In Figure 3, we compare the relative sustainable TDPs for a processor system of size ~600 mm^2 for a chip junction temperature of ~80^o C with 40^o C ambient temperature. The benefit in TDP from removing the package and conventional one heat sink approach can be about 20% while with two heat sinks, the sustainable TDP for the same junction temperature increases by about 70%.

C. Form-factor Benefits

The large package to die area ratio means that packages take up significantly large amount of area footprint on the PCBs. In SiIF based system integration, removal of package would help decrease the area footprint.

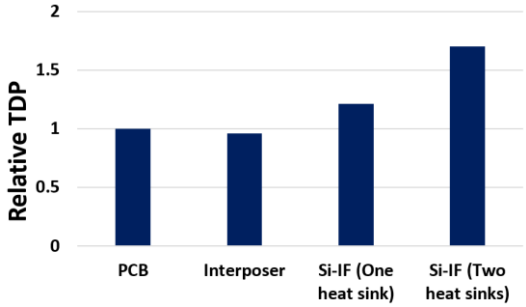

Figure 3: Relative TDP for different integration schemes

A 600 mm^2 chip with ~200W TDP would require a package of minimum size of ~30 cm^2 to accommodate the power and signal IOs using LGA connections. Removing the package gives us an area slack of about 80% of the package area. This area slack can be either used to decrease the area footprint of the system or the package area can be used to accommodate additional processor dies to increase compute density.

Moreover, since the traces delivering power on PCBs are long, they exhibit inductive behavior. Decoupling capacitors is used to nullify the inductive *Ldi/dt* noise. These decoupling capacitors often take up a significant portion of the PCB surface area. Since the traces would be much shorter because of overall area footprint and close proximity chiplet placement, the inductive noise is expected to be much smaller which implies that smaller number of decoupling capacitors would be required. Thus, an SiIF based system is expected to save overall system area by more than 50% and thus increase volumetric compute density. For large datacenters, this can potentially reduce the total cost of ownership per unit compute power.

D. Weight Benefits

System integration on SiIF helps eliminate the package, PCB and a significant number of passives, primarily decoupling capacitors. This helps reduce the overall system weight. For weight critical applications, for e.g. space systems, micro-robots, drones etc., where reduction of each gram of weight is essential for increasing fuel efficiency, total flight time, available payload, etc., SiIF based integration can be immensely useful. We analyzed two commercial embedded class microprocessor system platforms (Beagle bone black [19] and MBED [20]) and compared the weights of the PCB based integration and prospective SiIF based systems. The weight distribution of different components in the PCB and SiIF integration schemes are shown in Figure 4. For this comparison, we considered that the

passive components are conventional surface mount devices and assumed an upper bound on the passive's requirement which gives us an approximate lower bound of the gain in the total overall system weight. In practice, SiIF should give even more benefits in terms of weight reduction.

As shown in Figure 4, based on the two testcases we analyzed, SiIF based integration can save more than 60% of the total system weight.

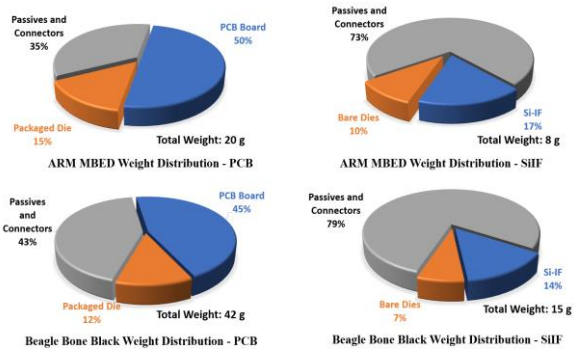

Figure 4: Comparison of PCB and SiIF based integration's weight distribution shown for ARM MBED and Beagle Bone Black

For larger systems with bigger PCBs, the weight benefit would be even larger. Thus, for weight-critical applications, SiIF based system assembly provides a platform to decrease system weight while retaining similar system level functionality and performance.

IV. Chiplet Assembly Design Methodology

Constructing systems using chiplet assembly comes with its own set of challenges. To realize full benefits of large system assembly, issues of physical design, clock delivery, power delivery, etc. need to be addressed. Consider the example of connecting different chiplets together on SiIF. SiIF is a passive substrate (necessary to ensure low-cost, high-yield integration) and therefore, unlike an SoC, cannot have buffers. Therefore, inter-chiplet wiring can become a timing bottleneck (especially so, if we want to retain simple on-chip like signaling strategies and assemble large number of small chiplets). In this section, we discuss a possible solution to this physical design problem: pin assignment and feedthrough buffer insertion.

In SoCs, a single monolithic chip is designed using multiple IP blocks, while in our approach, pre-fabricated hard-IPs in the form of chiplets would be mounted on the SiIF to build a system. The goal of SiIF design is to interconnect the different chiplets and ensure timing closure of the entire design. Since SiIF is a passive substrate and simple transceivers (buffers) drive the signals through the interconnect, minimizing the lengths of the timing critical nets on the SiIF is extremely critical for timing closure of the design. Given a particular floorplan on the SiIF, lengths of the interconnect wires are determined by the pin assignment on the chiplets. In the SoC methodology, pin assignment of different blocks is usually flexible and based on a particular floorplan, most design tools find the optimal pin assignment for these different blocks. However, in SiIF based assemblies, the component blocks are hard-IPs and thus the pin assignment is fixed and floorplan agnostic. Moreover, unlike in the SoC case, where buffers can be added between blocks, SiIF buffers are only inside the chiplets. We solve this problem using a mathematical programming based hierarchical pin-assignment methodology and the results for two test cases are shown in Table I.

Table 1: Wirelength Comparison of Hierarchical vs Non-Hierarchical Pin Assignment

		Test Case 1	Test Case 2
Hierarchical	Mean	77.5 μm	300.2 μm
	Max	378.2 μm	1049.2 μm
Non-Hierarchical	Mean	66.7 μm	81.7 μm
	Max	373.5 μm	1124.3 μm

We compare the mean and maximum wirelengths for our hierarchical pin assignment strategy versus a non-hierarchical approach. Note that non-hierarchical assignment is the optimal pin assignment while considering wirelength minimization, however, it would result in multiple physical chiplet realizations for the same function. We notice that using our hierarchical approach, the maximum wirelength remains almost similar to the non-hierarchical approach while the mean wirelength increases. Mean wirelength usually isn't critical for timing correctness and SiIF's abundant wiring resources can accommodate the increase in the total wirelength.

To further increase reuse, the hard IPs i.e., the chiplets are expected to be used in multitude of different systems with different floorplans. The issue here is that a pin assignment tailored towards a particular floorplan may not be suitable for other floorplans. To alleviate this problem, we utilize the abundance of IO pins that are available per chiplet with fine pitch copper pillar based IOs by making multiple copies of the same pin (for e.g. Figure 5).

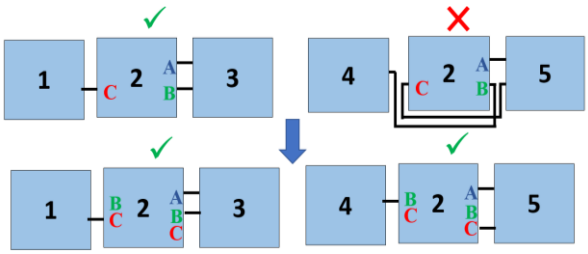

Figure 5: Multiple physical pin for the same logical pin on a chiplet helps achieve shorter interconnect length on SiIF

Since only one physical copy per logical pin is going to be used and the unconnected pins wouldn't drive any links or gates, with smart redundant pin planning the latency and energy overhead per additional copy of a pin can be managed to be negligible. Eventually, physical design approaches which integrate pin planning with chiplet floorplanning would be essential.

The passive nature of SiIF means that the nets interconnecting non-neighboring chiplets need to

traverse >1mm without buffering. These long interconnects would result in poor signal slew and unacceptable delays. We propose that every chiplet allocate standardized feedthrough buffer banks as shown in Figure 6 which can help buffer the signal on these long interconnects.

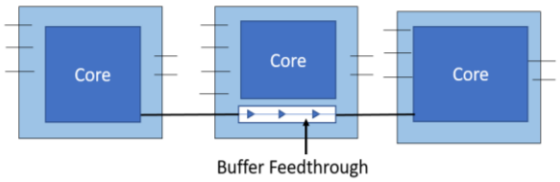

Figure 6: Buffer feedthrough scheme for communication between non-neighboring chiplets

.

V. Architectural Implications of SiIF

As mentioned earlier, removing the packages and replacing the PCB substrate with SiIF could potentially improve bandwidth by up to two orders of magnitude, sustainable TDP by up to 70% and increase the compute density significantly. Here, we discuss few examples of substantial architectural benefits of SiIF.

Increasing the processor-memory bandwidth and number of memory channels can improve the main memory performance significantly. In some recent products [21, 22], high-bandwidth memory (HBM) [23] or hybrid memory cube (HMC) [24] has been used to boost processor performance. These products use interposer or EMIB [25] solutions to integrate a few HBM modules close to the processor using high bandwidth interface inside a single package. However, majority of the main memory still sits outside the package and is connected using conventional DDR4 interface on PCB. This limits the bandwidth to the off-package memory modules. On the other hand, in SiIF a much larger number of dies can be integrated around the processors and the full memory subsystem can be provided with high bandwidth interface. In fact, number of memory channels can also be increased beyond what is supported by the limited IO count in packaged systems, which could potentially reduce memory access delays and improve performance significantly.

Thermal Slack in terms of TDP can be leveraged by increasing the number of active computational units or the frequency of operation. Both the approaches could potentially provide performance gains for different applications. Single thread performance dominated applications would benefit from higher frequency of operation while highly parallel and multi-threaded application would gain heavily from additional cores.

Chiplet based integration allows easier system customization by mixing and matching chiplets without incurring the cost of manufacturing multiple SoCs. One average best system design cannot serve different diverse applications with good efficiency. Thus, different applications are targeted using different processors and not a single processor. However, processor design cost, verification and manufacturing costs are increasing, thus chiplet based assembly provides a viable platform to build a large number of systems using a handful of dielets tailored towards specific application needs without the cost of building multiple full SoCs. Our preliminary experiments with 35 diverse set of applications representing different application classes (server, desktop, embedded) show that by using 4-5 chiplets and building systems out of them, one can achieve within 5% efficiency that can be provided by 14-20 systems. We also observed that *chiplets, unlike SoCs, are shared heavily across multiple applications and application suites, which helps increase chiplet reuse*. This has major implications on the cost of design and manufacturing. Chiplet design and manufacturing costs are much smaller than full systems due to their smaller size. *The fact that fewer number of chiplets can still serve a wide array of applications bodes well for an era of customizable systems using a chiplet-assembly approach.* When cost minimization is an objective, optimal selection of chiplets helps cut cost of design and manufacturing by 20-50% over multiple optimal SoC option while proving similar performance and energy efficiency.

Wafer Scale Computing

Typical PCB based high performance systems contains 2-4 processor sockets. Some HPC applications running on clusters often span multiple sockets and sometimes span multiple PCBs. Communication latency and bandwidth bottleneck arises when data needs to be migrated across different sockets and PCBs. Typical memory accesses using QPI is ~50-100 ns, while a 4KB transfer over a 100 Gbps HPC fabric takes ~1 us [26]. Moreover, inter-socket communication energy is about 20-50 pJ/bit. Wafer scale processor systems can accommodate multiple many-core processor dies on a single SiIF wafer. High bandwidth, low latency and high energy efficiency of the SiIF assembly can be leveraged to achieve significant performance improvement. Moreover, the increased volumetric computational density has economic benefits in terms of total cost of ownership for datacenter operators. Architecting such large systems comes with its own challenges but SiIF mitigates the yield issues of monolithic wafer scale system of past and communication issues of PCB-based large systems of present.

VI. Conclusion

Conventional PCB based integration scheme with packaged dies creates system level performance bottleneck due to poor (~4x) scaling over the past few decades while silicon features have scaled by more than a 2000x. We propose a novel system level integration scheme where bare silicon chiplets are directly bonded on to a silicon interconnect fabric (SiIF). SiIF based integration scheme can deliver dramatic improvements in communication latency/energy/bandwidth, reduce system footprint and allow for mcuh better thermal characteristics. Chiplet assembly on SiIF can enable

inexpensive system customization as well as make wafer-scale computing systems viable.

VII. Acknowledgements

The Defense Advanced Research Projects Agency (DARPA) through ONR grant N00014-16-1-263 and the UCLA CHIPS Consortium supported this work. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We thank Adeel Ahmed Bajwa, SivaChandra Jangam for discussions and SiIF technology development. We thank Daniel Petrisko and Rakesh Kumar at UIUC for discussions and ongoing collaboration on architectural implications of SiIF.

References

[1] M. Bohr, "A 30 Year Retrospective on Dennard's MOSFET Scaling Paper," *IEEE Solid-State Circuits Society Newsletter*, vol. 12, pp. 11–13, Winter 2007.

[2] E. Worthman. (Apr. 14, 2014). *A Guide to Advanced Process Design Kits*. [Online]. Available: http://semiengineering.com/a-guide-toadvanced-process-design-kits

[3] N. A. Sherwani, *Algorithms for VLSI Physical Design Automation*. Boston, MA, USA: Kluwer, 1995.

[4] R. Arnold, S. M. 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*, pp. 374–382, Apr 1998.

[5] R. H. Parker, "Bare die test," in Proceedings 1992 IEEE Multi-Chip Module Conference MCMC-92, pp. 24–27, Mar 1992.

[6] (Sep. 8, 2017). *Printed Circuit Board*. [Online]. Available: https://en.wikipedia.org/wiki/Printed_circuit_board.

[7] J. H. Lau, *Flip Chip Technologies*. NY, USA: McGraw-Hill, 1996.

[8] F. L. Miller, "Controlled collapse reflow chip joining," *IBM J. Res. Develop.*, vol. 13, no. 3, pp. 239–250, May 1969.

[9] Intel Corp., Land Grid Array (LGA) Socket and Package Technology, Sept 2009

[10] S. S. Iyer, "Heterogeneous Integration for Performance and Scaling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, pp. 973–982, July 2016.

[11] "Atom - Intel." https://en.wikichip.org/wiki/intel/atom.

[12] 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, pp. 32–39, Oct 1984.

[13] 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, Dec 2013.

[14] D. Malta, E. Vick, S. Goodwin, C. Gregory, M. Lueck, A. Huffman, and D. Temple, "Fabrication of TSV-based silicon interposers," in *2010 IEEE International 3D Systems Integration Conference (3DIC)*, pp. 1–6, Nov 2010.

[15] J. Heidenreich *et al.*, "Copper dual damascene wiring for sub-0.25 μm CMOS technology," in *Proc. IEEE Int. Interconnect Technol. Conf.*, Jun. 1998, pp. 151–153.

[16] A. A. Bajwa et al., "Heterogeneous Integration at Fine Pitch (≤ 10 µm) Using Thermal Compression Bonding," in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, Orlando, FL, 2017, pp. 1276-1284.

[17] S. Jangam, S. Pal, A. Bajwa, S. Pamarti, P. Gupta and S. S. Iyer, "Latency, Bandwidth and Power Benefits of the SuperCHIPS Integration Scheme," *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, Orlando, FL, 2017, pp. 86-94.

[18] R. T. Howard, *Thermal Management Concepts in Microelectronic Packaging: From Component to System* (ISHM Technical Monograph Series). Boca Raton, FL, USA: CRC press, 1984.

[19] "BeagleBone Black." https://beagleboard.org/black

[20] "mbed Microcontrollers." https://developer.mbed.org/handbook/mbed-Microcontrollers

[21] Intel, "Intel Xeon Phi," (2014).

[22] AMD, "AMD Radeon R9," (2015).

[23] *High Bandwidth Memory (HBM) DRAM*, JEDEC Standard 235, 2013.

[24] "Hybrid memory cube." (accessed July 31, 2017), https://en.wikipedia.org/wiki/Hybrid_Memory_Cube

[25] R. Mahajan, R. Sankman, N. Patel, D. W. 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, May 2016.

[26] Eliot Eshelman, "HPC-oriented Latency Numbers Every Programmer Should Know," [Online] Available: https://gist.github.com/eshelman