Latency, Bandwidth and Power Benefits of the SuperCHIPS Integration Scheme

SivaChandra Jangam, Saptadeep Pal, Adeel Bajwa, Sudhakar Pamarti, Puneet Gupta and Subramanian S. Iyer
Center for Heterogeneous Integration and Performance Scaling (CHIPS),
Electrical Engineering Department, University of California Los Angeles
{sivchand, saptadeep, s.s.iyer}@ucla.edu

***Abstract*— In this paper, we describe the performance and power benefits of our Fine Pitch integration scheme on a Silicon Interconnect Fabric (Si-IF). Here we propose a Simple Universal Parallel intERface (SuperCHIPS) protocol enabled by fine pitch dielet to interconnect fabric assembly. We show the dramatic improvements in bandwidth, latency, and power are achievable through our integration scheme where small dielets (1-25 mm^2) are attached to a rigid Silicon Interconnect Fabric (Si-IF) at fine interconnect pitch (2-10 µm) and short inter-die distance (50-500 µm) using solderless metal-to-metal thermal compression bonding (TCB). Our simulations show that links in the Si-IF with short wire-lengths (<500 µm) have excellent signal transfer characteristics with low channel loss (<-2 dB) and low cross-talk (<-15 dB). With fine interconnect pitches (<10 µm), our scheme can achieve >5-25x improvement in data bandwidth. This can improve system performance (>20x) when compared to PCB-style integration and may even approach single die SoC metrics in some cases. Furthermore, our protocol is simple and non-proprietary. We show that this scheme enables heterogeneous system integration using a dielet based assembly method and provides significant reduction in design and validation cost. System-level analysis of heterogeneous integration scheme promises power benefits of more than 15% even for very small systems.**

Keywords- Silicon Interconnect Fabric; Thermal Compression Bonding; Fine Pitch Interconnect, SuperCHIPS

I. INTRODUCTION

Mainstream system integration technologies use PCB based substrates to build systems, from server blades, to Internet-of-Things (IoT) systems. Packaged dies are assembled on organic boards using solder based Ball Grid Array (BGA) connections. Fig. 1(a), shows an example of system-on-board with packaged dies placed on a PCB. The dimensions of the solder balls have reached their minimum limits due to factors such as solder extrusion, bridging and warpage of substrate etc. This constrains the number of connections the packages can have, which in turn limits achievable bandwidth. This also restricts the size of the package, which needs to accommodate all the I/O and power links. The package to silicon die area ratio can be large (2x-10x) for systems with large pin count. These traditional packaging schemes of individual dielets constrain the minimum inter-dielet spacing on a substrate. Also, due to minimum wiring feature size and signal integrity issues, PCBs are designed to have many wiring levels. Consequently, traces between separately packaged dies run from a few to several centimeters leading to increased communication latency and channel loss. To increase the bandwidth through such links, serialization and deserialization circuits, commonly known as SerDes are implemented. These circuits have complex high-speed transmitter and receivers to ensure signal integrity over such long data links. They not only occupy a substantial portion of real estate on die but also consume significant power, which can be as high 30% of the total chip power.

The System-on-Chip (SoC) approach offers solutions to these problems by designing and fabricating an entire system with different IP blocks on a single silicon die. Availability of fine pitch wiring and short inter-block distances provides low latency and high bandwidth. However, the time, complexity, and cost for designing such systems are very high. Today's systems demand SoC-like high performance interconnections for inter-die communication. Our approach of simple universal parallel interconnections with high performance (SuperCHIPS) integration can realize such SoC-like performance by assembly of individual dielets in close proximity (<100 µm) and interconnected at SoC-like wiring pitches. Fig. 1(b), shows a floor-plan of the same system realized using a dielet based assembly on Silicon-Interconnect Fabric (Si-IF) using SuperCHIPS protocol.

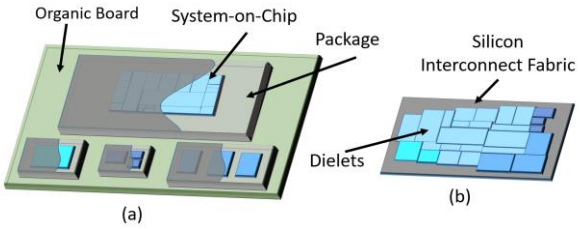

Figure 1. (a) Conventional integration scheme on organic board with individual packages. (b) Integration scheme on Silicon Interconnect Fabric.

Our analysis shows that the SuperCHIPS protocol can result in 50x improvement in interconnect energy efficiency (pJ/bit), 13x latency decrement and 5x-30x increment in bandwidth/millimeter compared to PCB based systems. In comparison with SoCs, the latency and energy numbers of SuperCHIPS are only 2x and 5x or even lower compared to large SoCs. Thus, our fine pitch integration scheme with the SuperCHIPS protocol finds a sweet spot between traditional interconnect solutions and SoC based designs. The Si-IF accommodates large number of data links, thus increasing the inter-dielet bandwidth. The low interconnect channel loss eliminates the need for complex transceiver circuitry, reducing power consumption, design complexity and the latency hit incurred at the transceivers. The number of layers required to route inter-die connections is also reduced by 2-5x due to fine wiring feature size in Si-IF. Our study predicts the latency of a link to be around 50-100ps, dominated by the Electro-Static Discharge (ESD) capacitance. With simple

AUTHORS' COPY MARCH 2017
To appear at IEEE Electronic Components and Technology Conference (ECTC), Florida, USA, May 30 – June 2, 2017

transceiver drivers, energy per bit of <0.4pJ/b can be achieved even for data rates >10Gbps per link. The bandwidth in SuperCHIPS systems can reach several terabits per second at total power <2.5W.

This technology also provides the flexibility to decompose an SoC into sets of constituent components, where each set is implemented on a different dielet, by providing a solution to tightly reintegrate them. Because different sub-components may be optimized differently, dielet-based assembly provides opportunities to optimize overall power, performance, reliability, and cost of a system. The rest of the paper is organized as follows: Section II discusses related work and concepts and outlines the features of the Si IF integration scheme. In Section III, we present detailed interconnect modeling and performance analysis. Section IV compares SuperCHIPS protocol with traditional interconnect technologies. Several key parameters such as power, latency and bandwidth are compared. Discussion regarding system partitioning and benefits of heterogeneous dielet integration is presented in Section V. Conclusion is given in Section VI.

II. RELATED WORK AND CONCEPTS

This section discusses the various related work on high performance interconnects, heterogeneous integration, and wafer scale integration.

A. Interconnect Technology

As mentioned earlier, the PCB/BGA pitch has become the bottleneck for data bandwidth. Typical BGA pitch is ~400 µm and typical C4 bump technology is ~100 µm. Several interconnect technologies have been proposed in the recent past with targets to achieve high performance system level integration. Silicon interposer technology has been presented as a solution for high interconnect density with a thinned silicon as a redistribution layer (RDL) between dielets [16-18,30-32]. However, the interposer size and the Through Silicon Via (TSV) cost restrict the wide spread applicability of the technology. Also, the interposer is finally connected to an organic board with solder adding an additional level of packaging. Fine pitch interconnects using copper pillars with solder cap at 20 µm pitch were developed and presented in [1]. Solid Liquid Inter-diffusion (SLID) process between metal (Cu) and solder (Sn) for bonding was used which forms Cu-Sn intermetallic compounds that can cause thermal, mechanical, and electrical reliability concerns. Solderless interconnects using thermal compression bonding (TCB) have been proposed to overcome these problems. The elimination of solder by direct metal-metal bonding also provides opportunities for ultra-fine pitch interconnects. 6 µm interconnect pitch with 3 µm copper pillars were demonstrated in [21] using direct Cu-Cu TCB process.

These interconnect technologies aim at providing flexibility to system designers and enable modular designs. Modular CHIps or MoChi [3] is an integration scheme where SoCs can be split into multiple smaller cost-optimized modules and reintegrated without compromising on system performance. Fine pitch interconnects make such integration schemes possible. However, today's growing IO density, bandwidth demand for finer interconnection pitches <10 µm.

We developed our Si-IF technology to offer an alternative platform for system scaling. Our technology aims at elimination of the use of solder by direct metal-to-metal thermal compression bonding (TCB) [2] between metal pillars on substrate, to metal pads on the dielets. This allows us to scale down the interconnect pitch down to 2 -10 µm as the solder extrusion is no longer a limitation. We also remove packaging of individual dielets and place the dies directly on the Si IF with inter-dielet spacing of less than 100 µm. Thus, our data links can be much shorter (50- 500 µm). Our substrate is rigid Si with SiO_2 dielectric layer, which acts as the interconnection platform for entire system. The wiring in the dielectric layer is compatible with the mature Back End of the Line (BEOL) technology in CMOS fabrication. The feature size of wires in BEOL can be as fine as 0.5 µm which is a 10-20x reduction from feature size on organic boards. Also, fine features decrease the wiring congestion, which in turn reduces the number of wiring levels required for connections. In our Si-IF technology, up to four wiring levels are possible, though this is not a fundamental limit. This technology provides better chip-package interaction (CPI), elimination of under bump metallurgy (UBM) and decrease in fabrication cost. We demonstrated in [35], the continuity of 10 µm pitch interconnects with inter-dielet spacing of ~100 µm and alignment accuracy of <2 µm.

B. Heterogenous Integration:

As SoCs get more complex, design and manufacturing becomes challenging. SoCs are inherently made in one technology node, which is not always optimal from a system level integration point of view [5]. Authors in [5] discuss that high-performance interconnect fabrics could be used to integrate the processor and the L3 cache tightly incurring minimum latency while providing desirable bandwidth. Heterogeneous integration can also have impact on yield [6, 7], for e.g., yield of processor and cache is coupled and they both influence each other. Having the cache separately on another dielet would alleviate this problem. Several past works [8,9,10] have focused on integrating components from different materials onto a chip, however it remains a very costly and tedious challenge. Using our high-performance interconnect fabric [5,11] can help attain massive heterogeneity in a system while retaining the benefits of a SoC in terms of performance and energy efficiency.

C. Wafer Scale Integration

Wafer scale integration (WSI) is a way to build very large wafer scale systems [15, 34]. For massively parallel system, WSI shares a similar goal to dielet based assembly on fine pitch interconnects. The goal is to integrate large systems on a single wafer to reduce interconnect energy and latency. This helps in realizing better performance and reduced cost of packaging. Despite significant efforts, wafer scale chip integration has not been practically realized [34].

Low yield of manufacturing a massive chip, interconnect reliability, timing correctness due to across-wafer variation etc. are the major issues. In dielet-based assembly, each individual die is small and thus yield and performance can be tightly controlled. Also, the interconnect fabric is a simple wiring fabric, which can be manufactured reliably.

III. Si-IF Interconnect Modelling

A. Interconnect Model

As mentioned earlier, Si-IF interconnects use BEOL wiring technology. The interconnect trace and pitch dimensions are comparable to the top metal layers of a dielet (2-10 µm). Compared to PCB wire trace widths which maybe of the order of ~100 µm, Si-IF wire widths range from 0.5µm to 5 µm and the pitch from 1 µm to 10 µm. Trace lengths in PCBs can be several cm to tens of cms and those in interposers can be few to several mm. In Si-IF technology, we can realize dielet to dielet interconnects of lengths less than 500 µm more typically 100 µm. The insertion loss in our scheme is low due to reduced parasitics. Crosstalk is also small due to low coupling parasitics resulting from the fine dimensions of wires. These are all characteristics on on-chip wiring as well.

We simulated 3-D models of our Si-IF interconnect links in Electromagnetic (EM) solvers like ANSYS HFSS to study the signal transfer characteristics. For our models, we assumed direct Cu-Cu bonding with no additional metal layers and no intermetallic compounds at the interface. Also, we assume perfect bonding at the interface with no voids and thus we can apply bulk properties of copper across the interface. We also placed the dielets in near proximity (~100 µm), so that wirelengths of 100 – 500 µm are realizable. The simulated Si-IF structure is shown in Fig. 2. The bottom substrate is Si with SiO_2 dielectric layer. For our analysis, we assumed a single copper metal layer for data links inside the dielectric layer. However, for a real system, four or more wiring levels are possible and the characteristics should not deviate too much from the simulated structure. The top layer of Si-IF is terminated with copper pillars that protrude out of the surface. The top dielets also consist of Si and SiO_2 dielectric layer. The top layer of dielets are terminated with copper pads openings that are flip chip TCB bonded to copper pillars. We designed different models with varying lengths, pitches, and configurations to analyze insertion loss and cross-talk trends. The dimensions of the layers used in our simulations is shown in Table. I. The Si layer thickness is lower than expected for ease of simulation.

The copper pads in the top dielet act as the terminal for the EM wave excitation. We investigated different terminal configurations for insertion loss and cross-talk estimation. In this paper, we discuss the results of three configurations (i) Ground-Signal-Ground (GSG), (ii) Ground-Signal-Signal-Ground (GSSG) and (iii) Ground-Signal-Signal-Signal-Signal-Ground (GSSSSG) configurations shown in Fig. 3. The ground (return path) link structure is also a wire instead of traditional planes in PCB. Both the signal (forward path) and ground (return path) wires are of same dimensions as mentioned in Table. I. The bottom of the Si substrate is grounded.

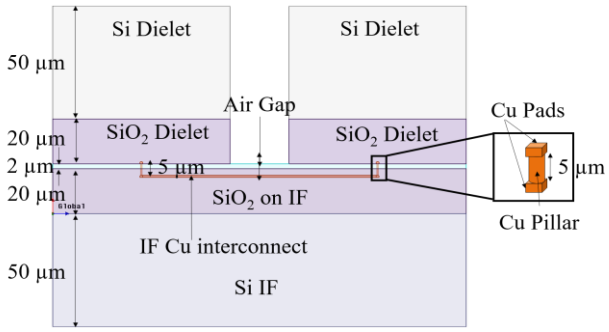

Figure 2. Structure of the model used to simulate link characteristics

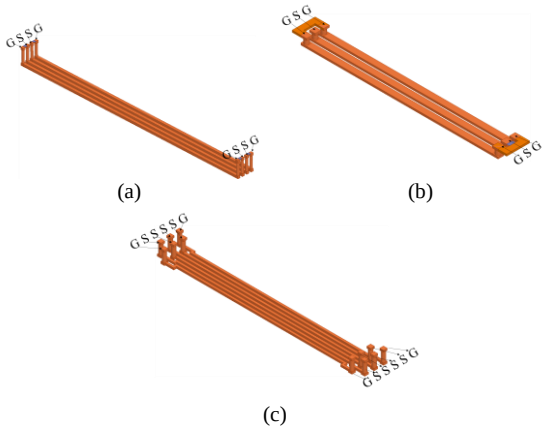

Figure 3. (a) GSSG wire configuration. (b) GSG wire configuration. (c) GSSSSG wire configuration.

TABLE I. Simulated Model Dimensions

Component	*Thickness (µm)*	*Width (µm)*	*Pitch (µm)*
Copper Pillar	5	1, 5	2, 10
Copper Data Link	1	1	1.5, 2, 10
Si Substrate	50	50	N.A
SiO_2 dielectric layer	20	50	N.A
Air Gap	2	50	N.A

B. Insertion Loss and Cross Talk

Parasitic inductance and capacitance of the long wire traces determine the transfer characteristics of links on PCB or interposer limiting their bandwidth. Transmission line models are used to characterize the behavior of such links. The inductance becomes significant for wires with wire lengths greater than $1/10^{th}$ of the wavelength (λ) of the propagating EM wave. For 100 µm ($\lambda/10$) wires of copper in SiO_2, the inductance becomes significant only at 100GHz. In our scheme, the short wire lengths (<500 µm) correspond to very low inductances but the fine dimensions lead to higher

resistances. Therefore, the resistance and capacitance of these links are the only significant contributors to the frequency response. Consequently, simple lumped RC circuit can be used to model these links.

Fig. 4(a) shows the insertion loss of the links with pillar diameter and wire width of 1 µm with pitch of 2 µm for different wire lengths. Fig. 4(b) shows the insertion loss of the links with wire width of 1 µm, pillar diameter of 5 µm and pitch of 10 µm. The plots indicate that the 100 µm and 500 µm wires behave as a RC circuit for most of the frequencies of interest (0.1-100 GHz), and the 1 mm wires deviate from RC behavior due to larger inductances at high frequencies. The Si-IF data links show excellent signal transfer with insertion loss of less than -2dB for 500 µm wires even for frequencies up to 100 GHz. The insertion losses are significantly lower compared to other technologies [16-20]. The excellent characteristics signify the importance of short wire lengths for signal transfer. The low loss reduces the need for complex transceiver or receiver circuits. We propose that simple tapered buffers be used for the transceiver circuits. Due to RC characteristics of the link, there is no inter-symbol-interference, which occurs due to reflections in a transmission line, thus eliminating the use of complex equalizers.

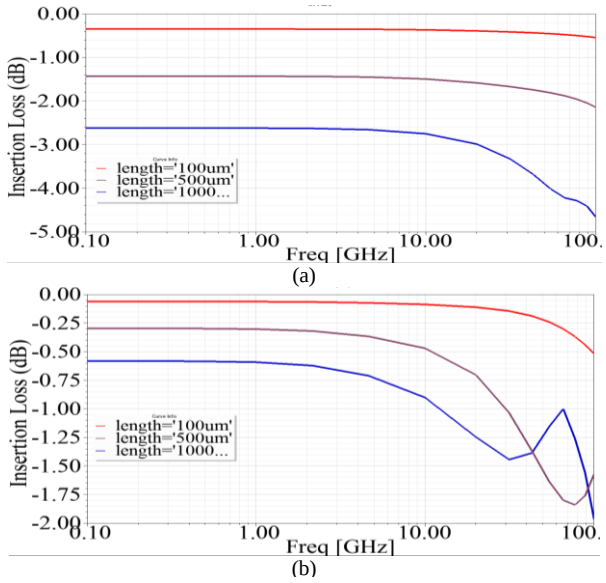

Figure 4. (a) Insertion Loss for 2 µm interconnect pitch. (b) Insertion Loss for 10 µm interconnect pitch.

For analog RF signals frequencies (10GHz – 1THz), we modeled GSG configuration links. At very high frequencies (~50GHz) depending on the wire length, the wires may behave like transmission lines due to self-inductance. The characteristic impedance is not clearly defined for short Si-IF links and is a valid concept only for longer links. We designed models with characteristics impedances of the coplanar GSG link traces to be 50 Ω and 100 Ω. The wire width was calculated to be 6 µm and the wire spacing was 3 µm and 7 µm respectively [28]. The insertion loss for different terminations and wire lengths are shown in Fig. 5. The simulated insertion loss is less than -3dB even for THz signals and termination of 100 Ω may also be realized.

The crosstalk between these links is due to the capacitive coupling and mutual inductance of the coplanar traces. Our objective was to estimate the crosstalk between parallel links in the same layer. For crosstalk analysis, we modified our pillar array arrangement to staggered array in the GSSSSG configuration with different pitches as shown in Fig. 3(c). The width of the wire is 1 µm. The near-end cross-talk (NEXT) between links with non-shared grounds is shown in Fig. 6(a), and the NEXT between links with shared grounds is shown in Fig. 6(b). The far-end cross-talk (FEXT) between links with non-shared grounds is shown in Fig. 7(a), and the FEXT between links with shared ground is shown in Fig. 7(b). The simulations show that the NEXT and FEXT between links with non-shared ground is less than -20dB for all pitches even at very high frequencies. The worst case NEXT between links with shared ground is less than -15dB at 10GHz and -5dB at 100GHz. The worst-case FEXT is less than -20dB at 10Ghz and less than -12.5dB at 100GHz. The NEXT and FEXT between links with shared ground are higher due to ground bounce effect. These values are lower than the typical acceptable crosstalk of -12dB.

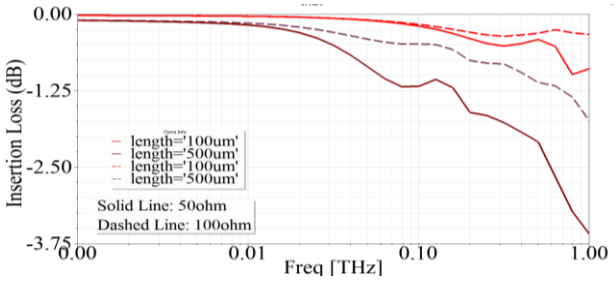

Figure 5. Insertion Loss at different characteristics impedance and length.

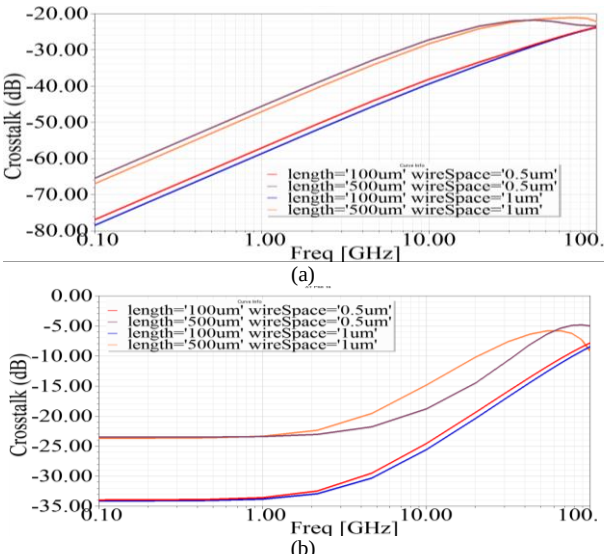

Figure 6. (a) NEXT for signals without shared ground (b) NEXT for signals with shared ground.

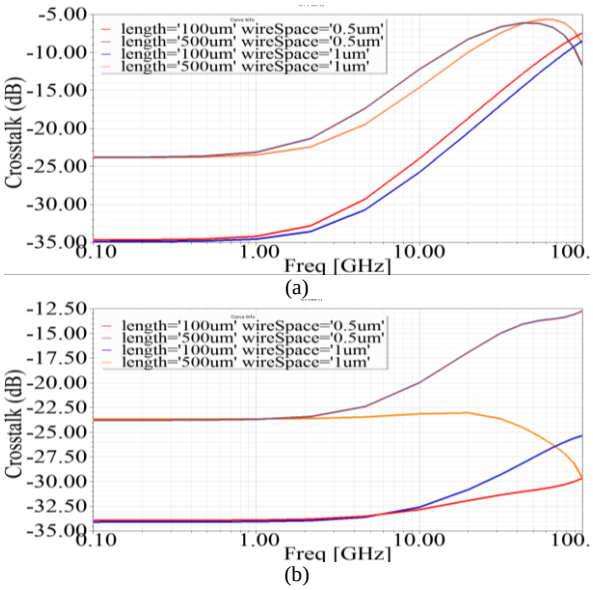

Figure 7. (a) FEXT for signals without shared ground. (b) FEXT for signals with shared ground.

C. *Signal Integrity Analysis*

The lumped RLGC of the Si-IF interconnects were extracted using the ANSYS Q3D extractor software shown in Table II. To make a meaningful estimate of the parasitics, we assumed a fixed wire width of 1 µm and interconnect pitches of 2 µm and 10 µm with pillar diameter being half the pitch. As shown, the parasitic capacitance and resistance of our Si-IF links are much lower than those discussed in [23]. We used the lumped equivalent circuit model to simulate an end-to-end link with transceivers as shown in Fig. 8(a). We used a commercial 45 nm technology library to design the transceivers and HSPICE L-2016.06 for circuit-level simulations. A tapered buffer is used to design the transmitter with the last stage having an NMOS width of 1um. This ensures good signal slew while providing required drive strength. The receiver circuit is a simple buffer. Due to the low contact area per interconnect and dielet handling in our technology, we predict the ESD protection capacitors needed for our dielets to be lower than those in traditional packages. Since our pad openings are much smaller than in traditional packages, the parasitic pad capacitance is also low. We assumed a total ESD protection capacitance to be 50fF in our simulations, which is the additional load at the transceiver output, and receiver input terminals. We simulated two different pitches with wire length of 2 µm pitch being 100 µm; and the 10 µm pitch being 500 µm. We show the eye diagram at the output of receiver buffer at operating frequency of 10GHz. The rise and fall time were assumed to be 10% of Unit Interval (UI) and the duty cycle distortion to be 10% of UI. The eye diagrams are shown in Fig. 8(b), 8(c). The eye-opening height is 997 mV for both 2 µm and 10 µm pitch channel. The eye-opening width is 68.4 ps and 59.81 ps for 2 µm pitch channel and 10 µm pitch channel respectively.

TABLE II. RLGC EXTRACTION

Interconnect pitch	*Wire-length*	*R @1GHz (Ω)*	*L (nH)*	*C (fF)*
2 µm	100 µm	2.09	0.10	17.30
2 µm	500 µm	9.33	0.68	79.23
10 µm	100 µm	1.89	0.10	8.54
10 µm	500 µm	8.85	0.54	34.10

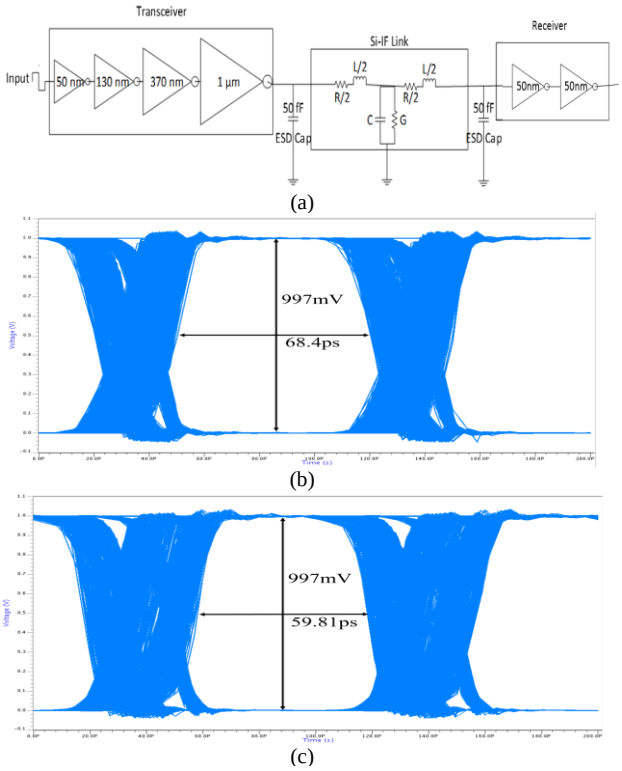

Figure 8. (a) Schematic of circuit use for signal integrity analysis. (b) Eye-diagram of 2 µm pitch interconnect at 10GHz input frequency. (c) Eye-diagram of 10 µm pitch interconnect at 10GHz input frequency.

IV. SUPERCHIPS BENEFITS

In this section, we compare the latency, bandwidth, and power benefits of SuperCHIPS integration approach with the traditional PCB packaging, interposer technology and SoC design. Analysis were done for 2 µm and 10 µm interconnect pitch with pillar diameter being half the pitch and trace width of 1 µm. SuperCHIPS provides a protocol based on fine pitch fine integration of system where the inter-dielet spacing is ~10-20x smaller than the conventional packaged systems on PCB. The fine pitch interconnects provide ~15-80x more number of I/O pins compared to BGA interconnects and ~2-10x more compared to copper micro-bumps [36].

For a first order estimate of latency in our data links, we calculated the Elmore delay from the last stage of the driver to the receiver input. For our calculations, we assumed 32nm technology, with the last stage driver being 3.2 µm wide. The

driver resistance and capacitance values are given in [29]. The 2 µm interconnect pitch links were assumed to be of 100 µm trace width while the 10 µm interconnect pitch links were assumed to be of 500 µm trace width. The overall latency depending on the technology and driver design in SuperCHIPS systems 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. The comparison of latencies is presented in Fig. 9(a). From the extracted parasitics, we calculated the maximum data-rate achievable with simple buffer stages for 6τ settlement. The data-rate per link can range from 4-10 Gbps with ESD capacitance. The data-rate per link in SuperCHIPS is expected to be lower than those in SerDes links. However, the bandwidth per millimeter of chip edge is higher due to increased density of connections. We assumed for bandwidth estimations, two rows of staggered pins with half of them being signal and rest are ground. The predicted bandwidth is shown in Fig. 9(c). The energy per bit is significantly lower (<0.4pJ/bit) using SuperCHIPS compared to traditional systems. This is attributed to the reduced driver complexity and elimination of power hungry SerDes transceiver and receivers. The comparison of energy per bit with other technologies is shown in Fig. 9(b). Table III presents the overall comparison of key benefits from SuperCHIPS approach with conventional packaging. The bandwidth can further be increased by using stronger drivers at the cost of higher energy per bit.

TABLE III. COMPARISON OF INTERCONNECT SCHEMES

Interconnect pitch/protocol	***2 µm on Si IF Super-CHIPS***	***10 µm on Si IF Super-CHIPS***	***50 µm on Si Interposer DDR3***	***400 µm on FR4 PCB/ SerDes***
Dielet Size (mm^2)	1-25	10-100	25-600	25-625
No of signal links	1,000-5,000	600-2,000	100-1,000	100-500
Inter-die distance (µm)	<100	<500	<5,000	10,000
Link Latency (ps)	5.5[a] 24.3[b]	8.7[a] 27.3[b]	N.A	N.A
Overall Latency (ps)	37[a] 55.75[b]	40.22[a] 58.8[b]	300[23]	~1,000
Max data-rate/link (Gbps)	20.5[a] 4.76[b]	13[a] 4.21[b]	1.6[24]	40[37]
Energy per bit (pJ/b)	<0.3[b]	<0.4[b]	9.48[24]	23.2[37]
Max Bandwidth per mm (Gbps/mm)	10,250[a] 2,380[b]	1,300[a] 421[b]	32	100
Total I/O power (W)	2.82-14.28	2.13-6.74	6-15	46-230

[a] Without ESD capacitance

[b] With ESD capacitance

V. SYSTEM LEVEL EVALUATION

Our design approach is to partition the system into dielets that can be heterogeneously integrated on the Si-IF. This methodology helps to design and build a large system using a "divide and conquer" method at minimum cost and maximum IP reuse. Multiple candidate dielet sets satisfying the power, performance, reliability and cost (PPRC)

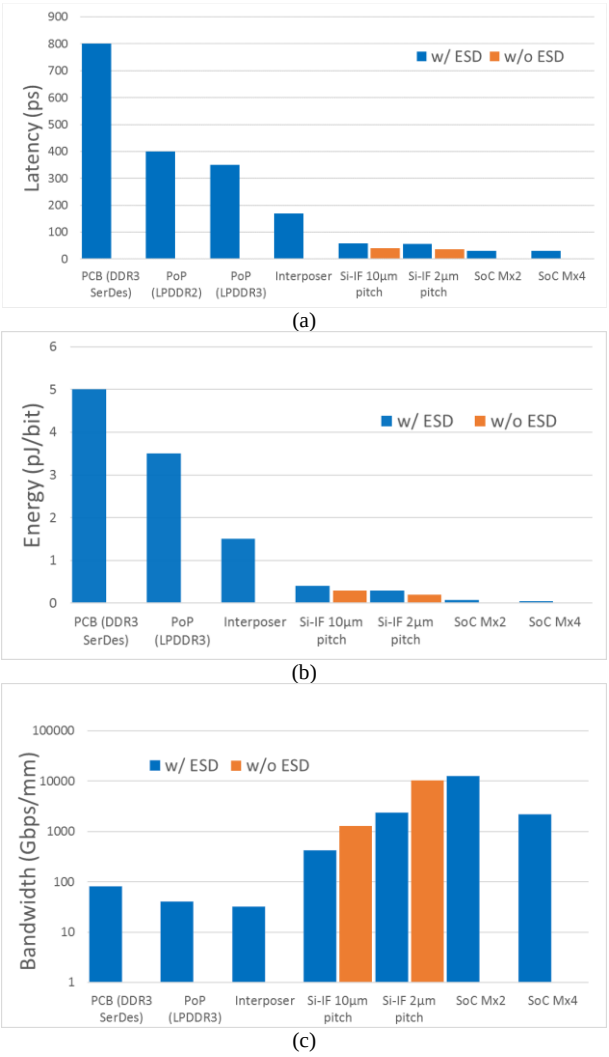

Figure 9. Comparison of (a) Latencies. (b) Energy per bit. (c) Bandwidth/mm.

constraints are possible. The dielet assembly approach allows us to choose heterogeneous dielets from different technologies, nodes and materials leading to optimal PPRC benefits with high probability of dielet and IP reuse. Partitioning starts from an initial system specification described as a set of sub components, inter-connections and the PPRC and area constraints of the system. Each component contains tightly coupled latency sensitive micro-architectural units. The objective of the partitioning scheme is to find those partitions, which can be manufactured as separate dielets. This approach of heterogeneous integration of a system with small dielet size, interconnect pitch and inter-dielet spacing on an Si-IF allows a new dimension in system scaling.

Here we discuss system level benefits of heterogeneous technology integration in a dielet based assembly method on the Si-IF using an example of a hexa-core CORTEX-M0 [14] processor system in a big.LITTLE [11,12] configuration. Configuration of the hexa-core processor

included 2 big cores, intended for higher throughput and 4 LITTLE cores tuned for higher energy efficiency. By implementing components in different technologies, a designer can hit better power-performance targets. For example, technologies such as LPE (low power early) can provide much lower power at the cost of reduced performance. In a monolithic approach, a technology would need to be selected for the entire processor. By enabling the selection of specific technologies for individual cores on a dielet-based multi-core processor, a wider range as well as a finer granularity of control over the power-performance curves for the system can be achieved. Many processors today implement some form of core heterogeneity, but this is normally achieved through changes in micro-architectural implementation as well as dynamic voltage and frequency scaling (DVFS). Both approaches have significant overhead in terms of design and verification cost. A dielet-based architecture can achieve the benefits of heterogeneity without incurring the cost of designing multiple cores and maintaining a separate, low-performance instruction set architecture (ISA). Table IV shows the system power of the hexa-core processor system at different activity factors and heterogeneous technology choice for the big and LITTLE cores. Activity factor refers to the switching activity. Also, power for different frequency of operation is reported. Power numbers are normalized to nominal operating frequency for GP process.

TABLE IV. THE POWER AND PERFPRMANCE FOR VARIOUS COMBINATIONS OF LPE AND GP IN A HEXA-CORE PROCESSOR SYSTEM BIG.LITTLE CONFIGURATION AT DIFFERENT ACTIVITY FACTORS AND OPERATING FREQUENCIES

Design: CortexM0	**Power in mW**		
Activity Factor	0.001	0.01	0.1
Config: GP + GP			
nominal/nominal	0.262	0.526	3.8
0.5 nominal/ 0.5 nominal	0.06	1.68	1.45
0.1 nominal/0.1 nominal	0.032	0.5	0.7
nominal/0.5 nominal	0.161	1.103	2.56
nominal/0.1 nominal	0.147	0.514	2.25
Config: GP + LPE			
nominal/nominal	0.174	0.546	7.44
0.5 nominal/ 0.5 nominal	0.038	1.55	1.38
0.1 nominal/0.1 nominal	0.0175	0.39	0.46
nominal/0.5 nominal	0.139	0.536	2.53
nominal/0.1 nominal	0.1325	0.403	2.09
Config: LPE+LPE			
nominal/nominal	0.086	0.564	11.8
0.5 nominal/ 0.5 nominal	0.016	1.42	1.36
0.1 nominal/0.1 nominal	0.003	0.28	0.29
nominal/0.5 nominal	0.051	0.992	6.22
nominal/0.1 nominal	0.0445	0.422	5.685

We analyzed the power and performance of Cortex-M0 for two state-of-the-art commercial 65 nm technology libraries: general purpose (GP) and low-power early (LPE). GP libraries are targeted towards energy-delay optimized designs while LPE libraries are targeted towards low performance but energy efficient designs. We used PROCEED [33], a circuit level emulator to find the optimal power-delay pareto curve for CORTEX-M0. A wide operation region (MHz to GHz) was assessed for both GP and LPE technologies. Fig. 10. shows the power of the processor system in three different implementation scenarios. We explored the cases of iso-performance activity optimized cores, where the operating frequency is same for big and LITTLE cores, however, activity is adjusted as per the required performance. For heterogeneous process chip-multi-processor (CMP), activity factor is constant for big and LITTLE cores, while operating frequency is higher for big cores while it's lower for the LITTLE cores. The results indicate that the configuration where the LITTLE cores are realized using LPE technology and big cores using GP technology, provides energy savings of up to 15% and 37% over the homogeneous implementation cases where all the cores were in GP and LPE technologies respectively. Such a benefit while using same processors for heterogeneous workloads can only be obtained using heterogenous dielets assembled on a high-performance interconnect fabric, such as the Si-IF.

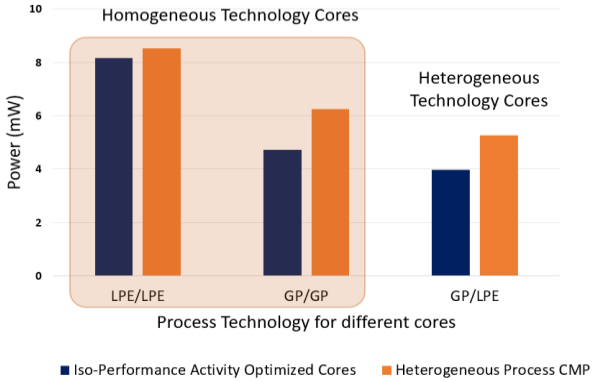

Figure 10. Power of an all-CORTEX-M0 hexa-core processor system in big.LITTLE configuration implemented in heterogeneous technologies.

VI. CONCLUSION

In this paper, we introduced a simple universal parallel interface integration protocol, SuperCHIPS for heterogeneous system scaling. SuperCHIPS promises SoC-like performance and flexibility for system-level heterogeneous integration. We show that the close inter-dielet assembly and fine pitch interconnects on Si-IF ensures low channel loss and link latencies. Our simulations show the channel loss to be less than -2dB and the cross talk less than -15dB. With SuperCHIPS approach, we achieve total interconnect bandwidths up to few Tbps with massive number of parallel links instead of the traditional SerDes techniques. This results in 50x improvement in energy efficiency compared to PCB based integration schemes. System-level evaluation of heterogeneous integration scheme promises significant power benefits while providing reduction in design and validation cost.

VII. ACKNOWLEDGEMENT

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 Professor Rakesh Kumar and Daniel Petrisko from UIUC for helpful discussions.

REFERENCES

[1] C. Honrao, T. C. Huang, M. Kobayashi, V. Smet, P. M. Raj and R. Tummala, "Accelerated SLID bonding using thin multi-layer copper-solder stack for fine-pitch interconnections," 2014 IEEE 64th Electronic Components and Technology Conference (ECTC), Orlando, FL, 2014, pp. 1160-1165.

[2] J. Fan and C. Tan (2012), " Low temperature wafer-level metal thermo-compression bonding technology for 3-d integration" in Metallurgy- Advances in Materials and Process, Dr. Yogiraj Pardhi (Ed.), InTech. doi: 10.5772/48216

[3] Mochi architecture. http://www.marvell.com/architecture/mochi/

[4] R. Haring et al., "The IBM Blue Gene/Q Compute Chip," in IEEE Micro, vol. 32, no. 2, pp. 48-60, March-April 2012.

[5] S. S. Iyer, "Heterogeneous Integration for Performance and Scaling," in *IEEE Transactions on CPMT*, vol. 6, no. 7, pp. 973-982.

[6] K. Agarwal, R. Rao, D. Sylvester, and R. Brown., "Parametric yield analysis and optimization in leakage dominated technologies", *IEEETransactions on Very Large Scale Integration (VLSI) Systems*, pp. 613–623, June 2007.

[7] J. Zhang and M. Nakhla. "Yield analysis and optimization of vlsi interconnects in multichip modules", in *IEEE Proceedings of Multi-Chip Module Conference*, pp. 160–163, Mar 1993

[8] Z. Wang, M. Pantouvaki, G. Morthier, C. Merckling, J. vanCampenhout, D. van Thourhout, and G. Roelkens, "Heterogeneous Integration of InP devices on silicon", *In 2016 Compound Semiconductor Week (CSW)*, June2016

[9] O. Moutanabbir and U. Gösele. "Heterogeneous integration ofcompound semiconductors" in *Annual Review of Materials Research*, pp. 469–500, 2010.

[10] A. Gutierrez-Aitken, P. Chang-Chien, D. Scott, K. Hennig,E. Kaneshiro, P. Nam, N. Cohen, D. Ching, K. Thai, B. Oyama,J. Zhou, C. Geiger, B. Poust, M. Parlee, R. Sandhu, W. Phan, A. Oki,and R. Kagiwada., "Advanced Heterogeneous Integration of InP HBT and CMOS Si Technologies", In *2010 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, pages 1–4, Oct 2010

[11] Heterogeneous Integration Roadmap. *http://cpmt.ieee.org/technology/heterogeneous-integration-roadmap.html*.

[12] ARM Big.LITTLE Architecture, *https://www.arm.com/products/processors/technologies/biglittleprocessing.php*

[13] A. Butko et al., "Full-System Simulation of big.LITTLE Multicore Architecture for Performance and Energy Exploration," 2016 *IEEE 10th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSOC)*, Lyon, 2016, pp. 201-208

[14] CORTEX M0 processor, https://www.arm.com/products/processors/cortex-m/cortex-m0.php

[15] R. Dettmer. "Brighter prospects for wafer-scale integration", *Electronics and Power, pp.283–288, April 1986*

[16] H. Lee, et al., "Design and signal integrity analysis of high bandwidth memory (HBM) interposer in 2.5D terabyte/s bandwidth graphics module," 2015 IEEE 24th Electrical Performance of Electronic Packaging and Systems (EPEPS), San Jose, CA, 2015, pp. 145-148.

[17] K. Cho et al., "Design optimization of high bandwidth memory (HBM) interposer considering signal integrity," 2015 IEEE Electrical Design of Advanced Packaging and Systems Symposium (EDAPS), Seoul, 2015, pp. 15-18.

[18] K. Cho, H. Lee, J. Kim, "Signal and power integrity design of 2.5D HBM (High bandwidth memory module) on SI interposer", Pan Pacific Microelectronics Symposium (Pan Pacific), Jan. 2016, pp. 1-5.

[19] G. Kumar, T. Bandyopadhyay, V. Sukumaran, V. Sundaram, S. K. Lim and R. Tummala, "Ultra-high I/O density glass/silicon interposers for high bandwidth smart mobile applications," 2011 IEEE 61st Electronic Components and Technology Conference (ECTC), Lake Buena Vista, FL, 2011, pp. 217-223.

[20] R. Weerasekera, J. R. Cubillo and G. Katti, "Analysis of signal integrity(SI) robustness in through-silicon interposer (TSI) interconnects," 2012 IEEE 14th Electronics Packaging Technology Conference (EPTC), Singapore, 2012, pp. 397-398.

[21] L. Xie, S. Wickramanayaka, S. C. Chong, V. N. Sekhar, D. Ismeal and Y. L. Ye, "6um Pitch High Density Cu-Cu Bonding for 3D IC Stacking," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, 2016, pp. 2126-2133.

[22] Y. Eo and W. R. Eisenstadt, "High-speed VLSI interconnect modeling based on S-parameter measurements," in IEEE Transactions on Components, Hybrids, and Manufacturing Technology, vol. 16, no. 5, pp. 555-562, Aug 1993.

[23] H. Kalargaris and V. F. Pavlidis, "Interconnect design tradeoffs for silicon and glass interposers," 2014 IEEE 12th International New Circuits and Systems Conference (NEWCAS), Trois-Rivieres, QC, 2014, pp. 77-80.

[24] M. A. Karim, P. D. Franzon and A. Kumar, "Power comparison of 2D, 3D and 2.5D interconnect solutions and power optimization of interposer interconnects," 2013 IEEE 63rd Electronic Components and Technology Conference, Las Vegas, NV, 2013, pp. 860-866.

[25] X. Ye, J. Fan, B. Chen, J. L. Drewniak and Q. B. Chen, "Accurate characterization of PCB transmission lines for high speed interconnect," 2015 Asia-Pacific Symposium on Electromagnetic Compatibility (APEMC), Taipei, 2015, pp. 16-19.

[26] Y. Zhang et al., "Design methodology of high performance on-chip global interconnect using terminated transmission-line," 2009 10th International Symposium on Quality Electronic Design, San Jose, CA, 2009, pp. 451-458.

[27] B. Kleveland, T. H. Lee and S. S. Wong, "50-GHz interconnect design in standard silicon technology," 1998 IEEE MTT-S International Microwave Symposium Digest (Cat. No.98CH36192), Baltimore, MD, USA, 1998, pp. 1913-1916 vol.3.

[28] G. Ghione and C. Naldi, "Analytical formulas for coplanar lines in hybrid and monolithic MICs," in Electronics Letters, vol. 20, no. 4, pp. 179-181, February 16 1984.

[29] S. X. Shian and D. Z. Pan, "Wire sizing with scattering effect for nanoscale interconnection," Asia and South Pacific Conference on Design Automation, 2006., Yokohama, 2006, pp. 6 pp.-.

[30] B. Sawyer et al., "Design and Demonstration of 2.5D Glass Interposers as a Superior Alternative to Silicon Interposers for 28 Gbps Signal Transmission," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, 2016, pp. 972-977.

[31] B. Sawyer, B. C. Chou, S. Gandhi, J. Mateosky, V. Sundaram and R. Tummala, "Modeling, design, and demonstration of 2.5D glass interposers for 16-channel 28 Gbps signaling applications," 2015 IEEE 65th Electronic Components and Technology Conference (ECTC), San Diego, CA, 2015, pp. 2188-2192.

[32] Y. Kim, J. Cho, K. Kim, V. Sundaram, R. Tummala and J. Kim, "Signal and power integrity analysis in 2.5D integrated circuits (ICs) with glass, silicon and organic interposer," 2015 IEEE 65th Electronic Components and Technology Conference (ECTC), San Diego, CA, 2015, pp. 738-743.

[33] S. Wang, A. Pan, C. O. Chui, and P. Gupta. "Proceed: A paretooptimization-based circuit-level evaluator for emerging

devices", *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pp. 192–205, Jan 2016

[34] 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," in *IEEE Spectrum*, vol. 21, no. 10, pp.32-39,Oct.1984

[35] A. A. Bajwa, S. Jangam, S. Pal, N. Marathe, M. Goorsky, T. Fukushima, S. S. Iyer, "Fine Pitch Die-to-Si Interconnections using Thermal Compression Bonding", 2017 IEEE 67th Electronic Components and Technology Conference (ECTC), Orlando, FL, 2017.

[36] L. J. Bum, J. A. J. Li and D. R. M. Woo, "Process development of multi-die stacking using 20 um pitch micro bumps on large scale dies," 2014 IEEE 16th Electronics Packaging Technology Conference (EPTC), Singapore, 2014, pp. 318-321.

[37] R. Navid et al., "A 40 Gb/s Serial Link Transceiver in 28 nm CMOS Technology," in IEEE Journal of Solid-State Circuits, vol. 50, no. 4, pp. 814-827, April 2015.