Pathfinding for 2.5D Interconnect Technologies

Saptadeep Pal
saptadeep@ucla.edu
University of California, Los Angeles
Los Angeles, California

Puneet Gupta
puneetg@ucla.edu
University of California, Los Angeles
Los Angeles, California

ABSTRACT

As conventional technology scaling becomes harder, 2.5D integration provides a viable pathway to building larger systems at lower cost. Therefore recently, there has been a proliferation of multiple 2.5D integration technologies that offer different interconnect characteristics such as wiring pitch, bump/pad pitch, inter-die distance, etc. All these factors affect the interconnect metrics of bandwidth, latency and energy-per-bit which ultimately determine system performance. There are other factors such as the choice of ESD circuitry, dicing technology and signaling voltage that also influence these interconnect metrics. In this work, we propose a novel pathfinding methodology for 2.5D interconnect technologies, which seeks to identify the trade-offs among the different factors which affect the performance metrics. We show that incessant scaling of the critical dimensions of the interconnect is not very useful. We emphasize the importance of managing ESD and dicing in improving energy efficiency of these interconnects. We also show that a heterogeneous chiplet ecosystem comes with significant I/O energy penalties. Overall, we demonstrate that a holistic approach considering features of 2.5D integration technology, chiplet technology and various other factors need to be considered and optimized simultaneously to maximize the performance and cost benefits of these integration solutions.

CCS CONCEPTS

• **Hardware → Buses and high-speed links**; **Multi-chip modules**; **Metallic interconnect**.

KEYWORDS

2.5D interconnects, μBumps, chiplet assembly, interposer

ACM Reference Format:
Saptadeep Pal and Puneet Gupta. 2020. Pathfinding for 2.5D Interconnect Technologies. In *System-Level Interconnect - Problems and Pathfinding Workshop (SLIP '20), November 5, 2020, San Diego, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3414622.3431906

1 INTRODUCTION

On one hand, transistor scaling is becoming more difficult and costly; on the other hand, the demand for larger System-on-Chips (SoCs) is growing rapidly as a result of growing need for performance. Large monolithic implementation of SoCs in advanced technology nodes often suffer from yield issues and are costly to design and manufacture. As an alternative, instead of building large monolithic dies, the components of an SoC can be separately manufactured in to disparate chiplets and integrated on a separate interconnect substrate shown in Fig. 1. These chiplets would be smaller, thus resulting in better yield of manufacturing, and the chiplets can be manufactured in suitable technology nodes for additional cost and performance optimization opportunities [1, 25].

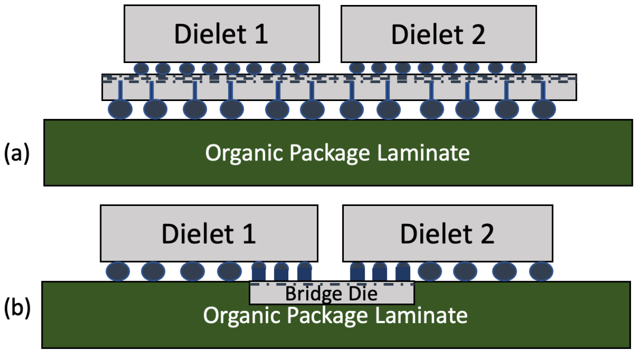


Figure 1: Cross-section view of two 2.5D substrates: (a) Silicon Interposer [9, 23], (b) EMIB [22]

However, to enable SoC like performance and energy efficiency, the interconnects on the substrate should closely resemble those of on-chip interconnects. Unlike conventional multi-chip module (MCM) substrates [8, 31] or PCB based interconnects which have coarse interconnect bump and wiring pitch, recent advancement in 2.5D technologies (such as silicon interposer [23]) allow communication between chiplets at high bandwidth, energy efficiency and low latency. This is achieved by: (1) manufacturing the substrates using mature semiconductor back-end-of-the line (BEOL) technology, such as 65*nm* or 90*nm* process node and (2) using fine-pitch μbumps[1] (pitch of below 70μm) to connect the flip-chip bare dies to the integration substrate for larger I/O density. Therefore, these technologies enable high performance multi-chiplet systems without the traditional off-chip communication bottlenecks.

Several 2.5D integration technologies have already been commercialized and many others are under active development. Examples include TSMC's CoWoS [9, 12], InFO [21], Intel's EMIB [22], Samsung I-Cube [2], Amkor's CoS, CoW, HDFO technologies [19], Silicon Interconnect Fabric (Si-IF) [6, 7], etc. These technologies

[1] In this paper we use μbump to refer to copper pillars, solder bumps or other bonding interfaces.

SLIP '20, November 5, 2020, San Diego, CA, USA

ACM ISBN 978-1-4503-8106-2/20/11.
https://doi.org/10.1145/3414622.3431906

offer minimum μbump pitch in the range of $10\mu m$ - $65\mu m$, minimum wire pitch in the range of $0.4\mu m$ - $4\mu m$, and 2-4 layers of metal routing.

The number of links between dies and the characteristics (bandwidth, energy, latency) of these links depend on multiple factors such as μbump/copper pillar sizing, wire sizing, inter-die spacing (length of the links), number of metal layers available for routing, ESD circuitry, etc. Past research [17, 26] has focused on one or few of these factors and discussed their scaling impacts. In this work, we focus on silicon based 2.5D interconnects such as silicon interposer, EMIB and Si-IF and comprehensively investigate all the multiple factors that affect the energy-bandwidth-latency scaling of inter-die links and highlight the trade-offs that exist between them. We develop a 2.5D interconnect pathfinding framework that takes all the design parameters as inputs, along with technology constraints and system design requirements (e.g. perimeter bandwidth density) and ranks all possible substrate designs in descending order of parameter set dimensions and energy-per-bit. We then analyze the link energy-per-bit and perimeter bandwidth density of these design points to understand and evaluate the trade-offs that come with scaling the critical dimensions of the multiple design parameters. Knowing this would help understand the exact parameters that need to be scaled in order to obtain substantial link energy, bandwidth and latency gains. For example, for a fixed number of routing metal layer, ESD capacitance and minimum inter-die link length, what is the optimal wire and μbump pitch beyond which scaling down only incurs additional manufacturing cost overhead while providing negligible benefits?

The rest of the paper is organized as follows: Section 2 discusses the interconnect pathfinding framework flow and the different components of our interconnect model that we used in our analysis. Section 3 covers our detailed analysis and highlights the main takeaways from our study. Section 4 discusses the trade-offs between designing a 2.5D interconnect substrate for homogeneous vs. heterogeneous chiplet ecosystems. Section 5 concludes the paper.

2 2.5D INTERCONNECT PATHFINDING FRAMEWORK

In this work, we propose a framework that evaluates trade-offs that come from scaling and varying physical link and interconnect substrate parameters. The framework helps assess return on (technology) investment in inter-chiplet interconnect and integration substrate.

2.1 Pathfinding Flow

Fig. 2 shows the 2.5D interconnect pathfinding framework flow. The framework takes system design parameters (wire and μbump dimensions, number of metal routing layers, minimum link length), technology constraints (ESD capacitance, inter-layer dielectric[ILD] thickness, wire thickness, inter-die spacing, etc.), and system constraints (e.g., perimeter bandwidth density) as inputs. Based on the inputs, the 2.5D interconnect design space is enumerated. For each interconnect design (a set of parameter values), we apply the interconnect length model and wire parasitic model to compute the maximum inter-die link length and the link parasitics. Based on this, we calculate the maximum load that needs to be driven by the transmitter. We assume that homogeneous transceiver circuitry would be used for all the neighboring inter-die communication links and therefore, we appropriately size the transceiver circuitry to support the maximum inter-die capacitive load. The models and transceiver circuit are then provided as inputs to our HSPICE [3] based simulation framework to calculate the latency and energy-per-bit for each design. Once the link characteristics are enumerated for all designs that meet the system constraints, the framework ranks the designs in descending order of parameter set dimensions and energy-per-bit.

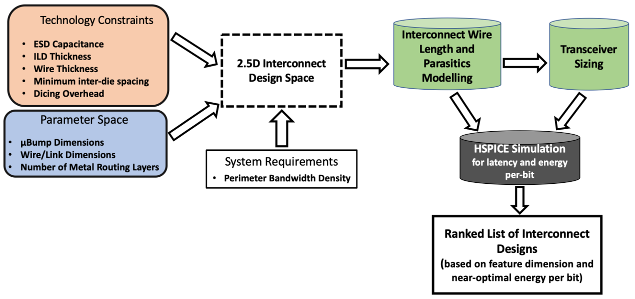

Technology Constraints
ESD Capacitance
ILD Thickness
Wire Thickness
Minimum inter-die spacing
Dicing Overhead
Parameter Space
µBump Dimensions
Wire/Link Dimensions
Number of Metal Routing Layers
2.5D Interconnect Design Space
System Requirements
Perimeter Bandwidth Density
Interconnect Wire Length and Parasitics Modelling
Transceiver Sizing
HSPICE Simulation for latency and energy per-bit
Ranked List of Interconnect Designs (based on feature dimension and near-optimal energy per bit)

Figure 2: 2.5D interconnect pathfinding framework

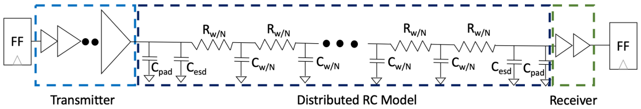


Figure 3: Distributed wire model for interconnect link. C_{pad} **is the pad/μbump capacitance,** C_{esd} **is the capacitance introduced by the ESD protection circuitry,** $R_{w/N}$ **and** $C_{w/N}$ **are the wire segment resistance and capacitance respectively.**

2.2 2.5D Interconnect Modeling

In 2.5D integration, bare chiplets are directly bonded on the interconnect substrate. Since the dies are un-packaged, inter-die spacing is small and the link lengths can be as small as $100\mu m$ and usually the maximum length of the inter-die wires is about $5mm$. Moreover, since abundant interconnect wiring resources are available in the 2.5D substrates, they are operated at a few GHz and the interfaces are usually designed as parallel interfaces instead of serialized/deserialized interfaces (SerDes [33]) that are used in conventional coarse-grained interconnect substrates. As a result, the transmitters and receivers can be designed using simple appropriately-sized cascaded inverters. We build link and μbump models to calculate the inter-die link parasitics and maximum length, and μbump parasitics based on the input parameter dimensions. We also appropriately size the transmitter circuitry based on the load it is driving. Next, we describe the components of our model in detail.

Modeling Wire Parasitics: In this work, we model repeaterless interconnect links that are found in today's passive integration substrates using a multi-segment Π model as shown in Fig. 3. We explore multiple wire and I/O pad parameters such as width, length,

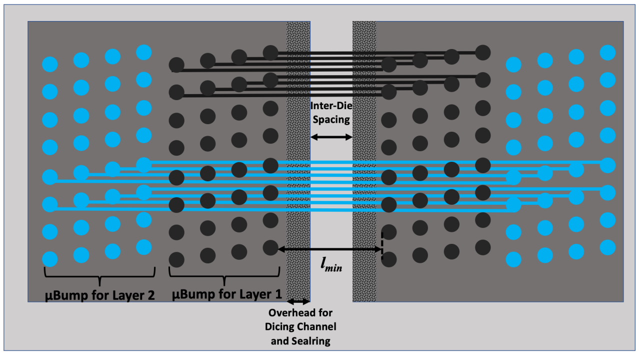


Figure 4: Top-down view of two dies on a 2.5D substrate with μbumps and interconnect wiring on multiple layers.

spacing. For each combination of these parameters, we calculate the link parasitics using the model proposed in [32] and validate the results against experimental results in [14]. We take in to account both neighboring wire coupling capacitance as well as the substrate ground capacitance.

The typical wire lengths in these interconnect technologies do not exceed a few millimeters. Therefore, the inductance effect is negligible [14] and the links behave as RC links. We further verified this effect by including inductance in a subset of our experiments.

Modeling Interconnect Link Length: Multiple columns of staggered I/O μbumps are used to support the interconnect wires that escape the periphery of the die per routing layer. In Fig. 4, we show an example with μbump pitch that is 4x of the wire pitch. To support the maximum possible wire density that can escape in one layer, four columns of I/O μbumps are required. As the ratio of μbump pitch to wire pitch increases, the number of I/O columns also increases. In addition to this, as the number of routing layers increase, the number of columns of μbumps increase as well. Therefore, the maximum length of the interconnect link increases as the columns grow orthogonal to the edge of the die. We calculate the worst case link length (l_{max}) using equation 1. It can be seen from equation 1 that scaling down the μbump pitch not only reduces the number of columns, it also reduces the additional link length per I/O column as shown in Fig. 5. As l_{max} increases, the capacitive load as well as resistance of the link increases.

$$l_{max} = l_{min} + \left(\frac{IO_{pitch}}{wire_{pitch}}\right) \times IO_{pitch} \times (2 \times N_{layers} - 1) - IO_{pitch} \quad (1)$$

where, l_{min} is the minimum distance between the μbumps in the neighboring dies, N_{layers} is the number of routing layers as shown in Fig. 4. l_{min} primarily depends on two factors: (1) inter-die spacing, and (2) distance of the first column from the edge of the die. Inter-die spacing can range from as low as 50μm (requires precise die placement and low die edge roughness) [7, 12, 22] to usually a few millimeters [29]. The first I/O column is placed at a distance from the edge of the die to accommodate dicing channel and sealring. This distance varies across foundries and processes and usually lies between 50μm - 200μm. Therefore, l_{min} has to be at least 150μm.

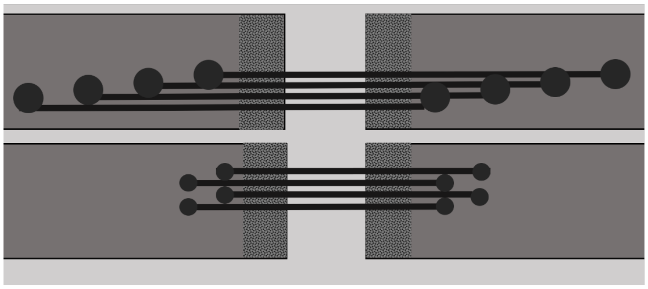


Figure 5: Smaller μbumps help reduce the interconnect length.

Table 1: Parameter values used in our analysis

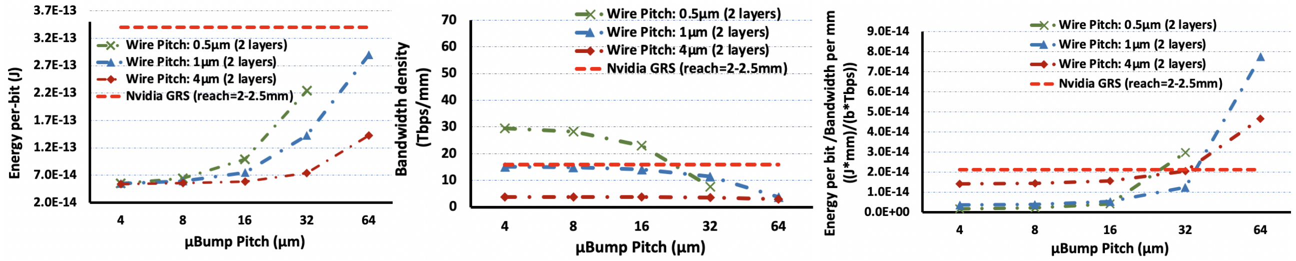
Parameters	Values
$wire_{pitch}$ (width=spacing)	{0.5, 1, 2, 4} μm
IO_{pitch} (width=spacing)	{4, 8, 16, 32, 64} μm
Wire thickness aspect ratio	1.5x (of wire width)
ILD thickness ratio	2x (of wire thickness)
Min. bump-to-bump dist. (l_{min})	{50, 150, 500, 1000, 2500} μm
ESD capacitance (C_{esd})	{0, 20, 50, 100, 200} fF
Metal routing layers (N_{layers})	1, 2, 4
Flip-Flop - $t_{clk-Q+setup}(45nm)$	62ps

Modeling I/O μbump Parasitics: We calculate the capacitance of a μbump (C_{pad}) using the model proposed in [11]. We augment this model by accounting for the coupling capacitance added by the surrounding μbumps and validate the results against the capacitance values presented in [14, 16].

Electrostatic Discharge (ESD) Circuitry Overhead: The individual dies that are placed on the interconnect substrate have to go through multiple post manufacturing processes like die thinning, known good die (KGD) testing, bonding etc. As such, the chiplets are prone to electrostatic discharge related incidents which can potentially damage the I/O circuitry resulting in die yield loss [4]. Therefore, the I/O pads need protection against these catastrophic ESD events, and is provided using large-sized high current carrying diodes. These diodes add significant capacitive load to the interconnect. We model this capacitive load as additional capacitors (C_{esd}) [16] on either end of the interconnect wire as shown in Fig. 3.

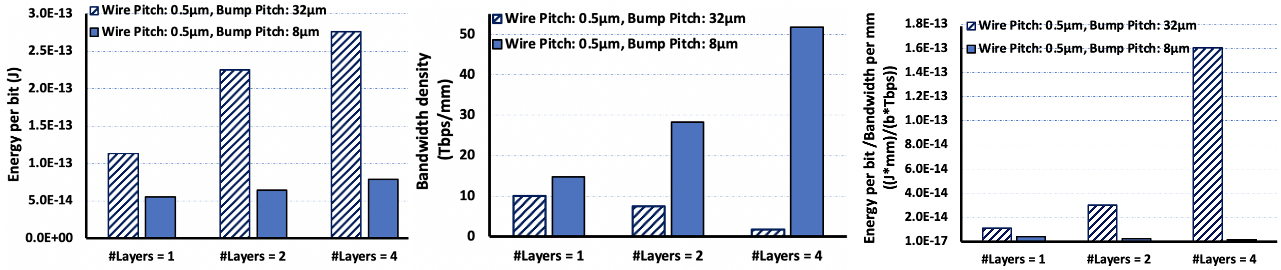
Transceiver Sizing: A cascaded inverter transmitter is used starting with the minimum sized inverter for a given technology Process Design Kit (PDK) and subsequent stages sized to drive a load equivalent to fan-out of four or less. We change the number of stages depending on the amount on load the transmitter is driving. The maximum sized inverter that we use is 128x the minimum size (five stages). The receiver is designed using two minimum sized inverters.

Table 1 shows the parameter exploration space for all the components of the interconnect model. We use HSPICE [3] and 45nm PDK to simulate the model and measure energy-per-bit, propagation delay/latency, rise/fall times. We calculate energy-per-bit by averaging over a pseudo-random binary sequence (PRBS). To calculate the



(a) Energy-per-bit (b) Bandwidth-per-mm (c) Energy-per-bit per bandwidth-per-mm

Figure 6: Scaling of energy-per-bit, bandwidth-per-mm and their ratio with μbump pitch and wire pitch for two metal routing layers (C_{esd}=50fF, l_{min}=150μm).



(a) Energy-per-bit (b) Bandwidth-per-mm (c) Energy-per-bit per bandwidth-per-mm

Figure 7: Scaling of energy-per-bit, bandwidth-per-mm and their ratio with metal routing layers for fixed wire pitch and two different μbump pitch values (C_{esd}=50fF, l_{min}=150μm)

maximum achievable bandwidth, we assume two flip-flops on either side of the interconnect link and consider the clock-to-Q delay and setup time in addition to the link latency. Moreover, we use three different technology nodes to show the trends from technology scaling.

3 HOW SHOULD WE SCALE THE INTERCONNECT SUBSTRATE ?

As mentioned earlier, there has been a recent trend in scaling the μbump size and pitch. Conventional μbumps are made of solder and thermo-compression bonding is used while attaching the die to the substrate. However, solder extrusion issues limit the scalability of the μbumps. Several alternative technologies such as solder-on-copper-pillar [30] and direct copper-to-copper bonding [7] has been proposed to shrink μbump size and pitch to sub-25μm and sub-10μm range, respectively. Though this allows us to pack more μbumps under the die area, these advanced processes are often more complicated and adds to the cost of bonding and assembly. Here next, we evaluate the efficacy of μbump scaling on the characteristics of the inter-die link.

Scaling μbump pitch vs wire pitch: We study the impact of scaling down the μbump pitch on energy-per-bit and perimeter bandwidth density for three different wire pitches and three different metal routing layer schemes. We keep l_{min} and C_{esd} fixed at 150 μm and 50fF, respectively. Scaling down μbump pitch while keeping the wire pitch constant (shown in Fig. 5) decreases the number of staggered I/O columns needed per routing layer to support all the wiring, which, in turn, leads to a reduction in the maximum inter-die link length. As seen in Fig. 6a, for the same wire pitch, the energy-per-bit reduces as μbump pitch is scaled down from 64μm.

However, Figures 6a, 6b and 6c show that scaling down the μbump pitch indefinitely does not improve energy-per-bit or bandwidth. This is because eventually the parasitics coming from l_{min} portion of the wire, C_{esd} and C_{pad} dominate. Due to the maximum link length dependence, the μbump pitch beyond which the benefits saturate increases with increase in wire pitch. This "saturation" happens at μbump pitch of 16μm for a wire pitch of 1μm and at 32μm for a wire pitch of 4 μm (e.g. in EMIB). Therefore, for a fixed wire pitch, incurring additional processing cost to reduce the μbump pitch beyond the saturation knee point might not be beneficial in terms of improving the link characteristics.

Takeaway: Incessant scaling of μbump pitch is not beneficial as the wire load is eventually dominated by ESD capacitance and inter-die separation.

Impact of additional metal routing layers: To support higher bandwidth density, one option is to increase the number of metal routing layers. This results in an increase in the the total amount of wiring per unit die edge. However, the number of I/O columns required to support all the wiring also increases. This, once again, leads to an increase in the worst case link length. In order to offset this effect, it is beneficial to scale down the μbump pitch as seen in Figure 7.

Another interesting observation is that even though increased number of metal layers help increase wiring resources, beyond a certain μbump pitch (>16μm), the added parasitics because of longer wires completely offset the gain from increased wiring and adversely affects the bandwidth/mm. This can be seen in Figures 7a and 7b. For the wire pitch of 0.5μm and μbump pitch of 32μm, the bandwidth/mm with four routing layers is 5x lower than that with a single routing layer. This is especially true at low wire pitch values where the wire resistance, and coupling and area-fringe capacitance values are high. The opposite is true for smaller μbump pitch (<16μm) where increasing the number of metal layers increases the bandwidth almost linearly while having negligible (< 1.5x when increasing from 1 to 4 metal layers) impact on energy-per-bit. Hence, the energy to bandwidth ratio in Fig. 7c decreases with increase in metal routing layers for μbump pitch of 8μm.

Takeaway: Increasing the number of wiring layers *must* be accompanied by a correspondingly smaller μbump pitch to derive bandwidth benefits from available increased wiring.

Comparison with μSERDES scheme: We also compare the link bandwidth and energy-per-bit with a μSERDES scheme. We use Nvidia's on-chip Ground Referenced Signaling (GRS) scheme [34] as the baseline for comparison against parallel interfaces on 2.5D substrates. The GRS links for which the energy-per-bit and bandwidth/mm values are plotted in Figures 6a and 6b have a wire pitch of 2μm, reach of 2-2.5mm and run at 16GHz. As can be seen in Figure 6b, parallel interface with wire pitch of 1μm, can achieve comparable bandwidth as that of the serialized/de-serialized GRS link. However, the parallel interfaces on 2.5D substrates achieve that bandwidth at a much lower energy cost. For example, with two metal routing layers, a parallel interface of 1μm wire pitch and 16μm μbump pitch has almost the same reach and achieves the same perimeter bandwidth density as GRS but at 4.5x lower energy cost. This is because, as mentioned earlier, the parallel links operate at a few GHz and the simple transceiver circuitry has negligible energy overhead. We show this later in Fig. 12 that transceiver circuitry energy is less than 10% of the total link energy. On the other hand, serialized/de-serialized links have much larger transceiver overheads that often dominate the overall link energy. With even smaller wire pitch, the parallel interface bandwidth/mm can be higher than the GRS links. However, the μbump pitch has to be scaled down significantly (16 μm or lower) to achieve any bandwidth benefits (Fig. 6b) and below 32μm for any bandwidth-normalized energy gains (Fig. 6c).

Takeaway: High bandwidth systems which want to move away from complex, energy-hungry serial links should aim for μbump pitches smaller than 16μm, and μbump pitches below 32μm are essential for leveraging parallel link energy efficiency benefits.

Impact of ESD capacitance (C_{esd}): C_{esd} adds a significant amount of load to the interconnect link. In general, about 50fF capacitance is added by the ESD diodes on each side of link [16]. When the maximum link length is small, C_{esd} dominates the link energy and latency. Therefore as mentioned earlier, reducing the link length or reducing the bump pitch (i.e., below the knee points in Fig. 6) doesn't help in improving the overall link characteristics.

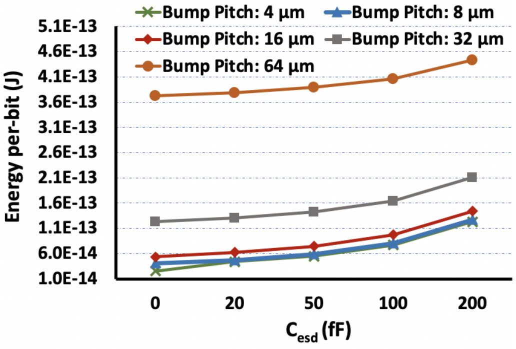


Figure 8: Energy-per-bit scaling with C_{esd} and μbump pitch (l_{min}=150μm, $wire_{pitch}$=1μm, N_{layers}=2)

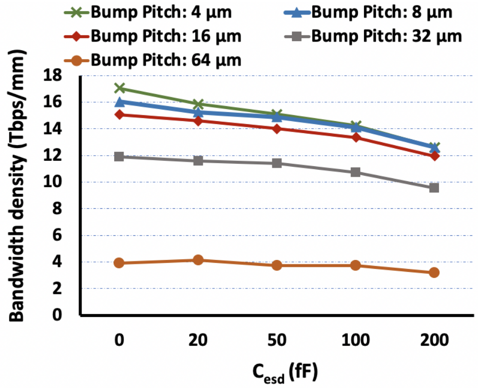


Figure 9: Bandwidth density scaling with C_{esd} and μbump pitch (l_{min}=150μm, $wire_{pitch}$=1μm, N_{layers}=2)

On one hand, with strict ESD control in modern advanced foundries during manufacturing, testing and bonding, the amount of ESD protection required is expected to decrease [5]. On the other hand, the 2.5D ecosystem is expected to accommodate dies from different foundry sources including older non-advanced foundries. As a result, some dies can in fact come with ESD circuitry with even larger amount of capacitance such as up to 200fF to 300fF [5, 26]. Here, we perform sensitivity analysis of ESD diode capacitance overhead.

As expected, in Figures 8 and 9 we see that when C_{esd} increases, even at smaller μbump sizes, the energy-per-bit and bandwidth density are considerably worse. Moreover, energy-per-bit and bandwidth degrades by a smaller fraction when moving to larger μbump sizes than compared to the case when C_{esd} is small. This is due to the smaller amount of load that is added by the additional wire length compared to the fixed overhead of C_{esd} when its value is large. Alternatively, if C_{esd} is small, μbump scaling can provide larger gains in energy efficiency and bandwidth.

Interestingly, reducing C_{esd} from 200fF to 50fF gives energy/bandwidth benefits comparable to reducing bump pitch from 32μm to 16μm. Therefore, it may be worthwhile to control ESD

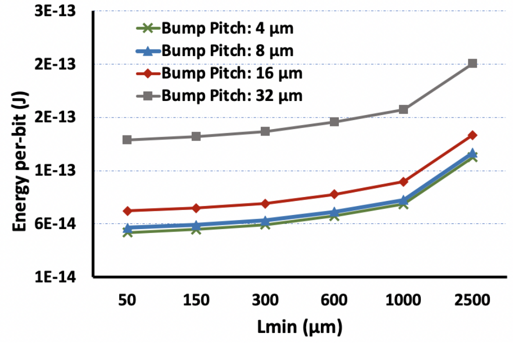


Figure 10: Energy-per-bit scaling with l_{min} and μbump pitch (C_{esd}=50fF, $wire_{pitch}$=1μm, N_{layers} = 2)

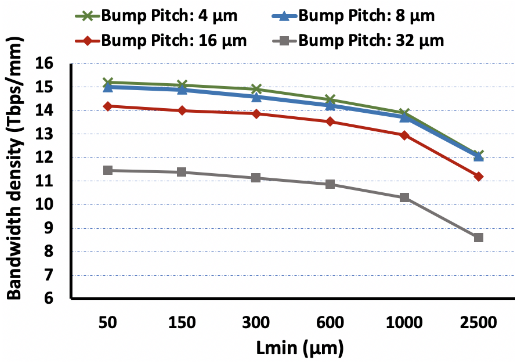


Figure 11: Bandwidth density scaling with l_{min} and μbump pitch (C_{esd}=50fF, $wire_{pitch}$=1μm, N_{layers} = 2)

events in the entire manufacturing and handling process rather than moving to an aggressive (and costly) μbump size.

Takeaway: C_{esd} can be used as a lever to scale both energy-per-bit and bandwidth and can enable us to stay at larger μbump pitches.

Impact of inter-die spacing and dicing overhead: Inter-die spacing and dicing related guard-bands impact the minimum distance between the μbumps on adjacent dies. Usually, mechanical dicing (using dicing saw) is used to singulate the dies on a wafer. This process often creates rough die edges and therefore the width of the die can vary by up to 50μm. Though advanced place and bond tools can achieve inter-die spacing of 50μm or less [7, 12], the dies are usually placed apart at minimum by more than 100μm to avoid die-to-die collision during bonding. Moreover, stress fracture and cracking at the edge of the die is a common occurrence with mechanical dicing [20] and therefore, seal ring and crack stops are added around the perimeter of design of the die [15]; this also affects l_{min}. On the other hand, plasma etch based dicing solutions [18] claim to reduce the die edge roughness as well as have minimal mechanical stress. Therefore, these solutions can reduce the overhead to below 10μm which can potentially reduce l_{min} to about 50μm. Next, we analyze the effect of inter-die spacing on energy-per-bit and bandwidth density of 2.5D substrates to understand if and when advanced processing for die singulation and better inter-die spacing is required.

Similar to the effect of C_{esd}, the capacitive load added by the minimum length wire (l_{min}) affects overall link characteristics. As shown in Figures 10 and 11, as l_{min} decreases, the link characteristics improve. However with C_{esd} of 50fF, reducing the inter-die separation to below 300μm seems to have limited use. 300μm is easily achievable using current generation dicing processes and place and bond tools, indicating that technology investment into better dicing technologies may provide limited benefit. Tighter inter-die spacing would be more useful only if ESD protection requirements can be reduced significantly.

Takeaway: Inter-die separation of 300μm which is achievable by current generation dicing and die placement processes is good enough.

4 INTERCONNECTS IN THE CHIPLET ECO-SYSTEM

As mentioned earlier, 2.5D interconnects can help lower system costs by enabling us to partition larger monolithic dies and re-integrate smaller and high yielding component dies on a 2.5D substrate. On the other hand, the chiplet based eco-system promises to be a platform for heterogeneous integration where chiplets from different technologies and vendors can be assembled to build customized and cost-performance optimal systems. These two use cases have dramatically different effect on the characteristics of the 2.5D interconnects as the design of the transceivers has to be done differently for the two cases.

For the case where all the chiplets come from the same technology, the transceivers on all the chiplets would be homogeneous and can be designed to operate at the core voltage offered by the technology. As the technology node scales down, the voltage of operation usually reduces. This has a quadratic impact on the energy required to switch the wire load. In Fig. 12, we show the energy-per-bit scaling of the interconnects for the different technology nodes of the transceiver circuitry for iso-bandwidth case (the transceivers were sized appropriately).

On the other hand, in a heterogeneous chiplet scenario, the chiplet operating at the highest voltage will determine the peak voltage of operation for the interconnect. Therefore, even though a chiplet can be manufactured in an advanced technology node, the benefits of voltage scaling won't be available. In order to operate transistors in advanced technology node at higher voltage, thick oxide devices may need to be used which could further degrade drive strength. This would require larger transistors resulting in increased transceiver energy although we expect the impact to be small (fraction of the energy spent in the transceiver itself, as shown in Fig. 12, is less than 10%). The minimum operating voltage of the link will be governed by the *oldest* technology the chiplet ecosystem supports. For example, in Fig. 12, a link that supports 45nm-12nm heterogeneous integration will be 70% less energy efficient than 12nm homogeneous integration link.

Takeaway: Link efficiency requirements may need to limit the technologies supported by a chiplet ecosystem.

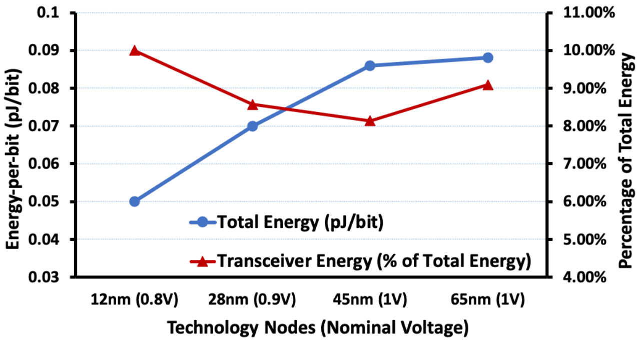


Figure 12: Total link energy and transceiver energy as % of total energy for for dielets (transceiver circuitry) from four different technology nodes. Link length: 1.5mm, W/S:0.5μm, C_{esd}:50fF

5 DISCUSSION AND CONCLUSIONS

Several commercial products today such as Nvidia and AMD GPUs, Xilix and Intel FPGAs, etc. use 2.5D substrates such as CoWos and EMIB to build large systems. As the demand for data-intensive and highly parallel applications grows and technology scales to fit in more compute per die, the amount of resources dedicated to 2.5D interconnect substrates is likely to rise. Therefore, it would be important to scale these interconnects in order to achieve performance and energy scaling proportional to the rest of the system. Nevertheless, it is important to keep a full-system perspective. As an example, let us consider high bandwidth memories (HBM) integrated on silicon interposers that are used in high performance systems today. For HBM2, about 10%-15% (0.4pJ out of 3.9pJ) [24] of memory access energy is used to shuttle data between the memory and compute dies. Memory energy itself would be a fraction of total compute energy (~20% for GPUs [27]) implying system-level energy benefits may be modest from link energy improvements. However, the benefits may be more significant when the non-interconnect part of memory energy is improved or for specialized communication-limited applications such as graph processing or streaming architectures.

Several efforts have been underway (e.g. DARPA CHIPS program [10], Open Compute Project [28]) to design interfaces and protocols to allow multiple chips to communicate. The implementation of these protocols require additional logic which ultimately adds to the inter-chiplet communication overhead. At just 0.1 pJ/bit of link energy or less, large amounts of bandwidth, e.g. 10TBps, can be supported with about 8W of power. However, protocol level logic and synchronization requirements can add 2-5X extra energy overhead [13]. Therefore, alongside optimization of 2.5D interconnect parameters, lightweight and energy efficient protocols need to be designed to enable overall communication energy reduction.

As conventional technology scaling becomes challenging, 2.5D integration provides a viable pathway to compose larger systems using smaller, high yielding dies. Therefore recently, there has been a proliferation of different 2.5D integration technologies. However, the success of this 2.5D approach depends upon optimizing the performance benefits and cost for different use case scenarios. In this work, we develop a pathfinding methodology for 2.5D interconnect technologies and use it to study inter-chiplet interconnect performance and energy as a function of dimensional and technology parameters. We demonstrate that a holistic approach considering features of 2.5D integration technology, chiplet technology and processing techniques. Our analysis indicates that beyond certain point, dimensional scaling (wire and bump pitch) provides marginal benefit in terms of energy-per-bit and bandwidth density; while other factors such as ESD and chip dicing technologies may provide additional levers for further interconnect scaling.

ACKNOWLEDGMENTS

This work was supported in part by the UCLA CHIPS Consortium, C-DEN center, and the Qualcomm Innovation Fellowship. The authors would like to thank Irina Alam, Krutikesh Sahoo, and the anonymous reviewers for their helpful feedback and suggestions.

REFERENCES

[1] 2019. Heterogeneous integration roadmap 2019 edition. https://eps.ieee.org/technology/heterogeneous- integration- roadmap/2019- edition.html.
[2] 2020. 2.5D Interposer(I-Cube™) Development. samsungfoundry.com.
[3] 2020. HSPICE. online. https://www.synopsys.com/verification/ams-verification/hspice.html
[4] A. Amerasekera, W. van den Abeelen, L. van Roozendaal, M. Hannemann, and P. Schofield. 1992. ESD failure modes: characteristics mechanisms, and process influences. *IEEE Transactions on Electron Devices* 39, 2 (1992), 430–436.
[5] ESD Association et al. 2007. ESD Association Standard for the Development of an Electrostatic Discharge Control Program for–Protection of Electrical and Electronic Parts, Assemblies and Equipment (excluding Electrically Initiated Explosive Devices). ANSI/ESD S20: 20-2007. (2007).
[6] A. A. Bajwa, S. Jangam, S. Pal, N. Marathe, T. Bai, T. Fukushima, M. Goorsky, and S. S. Iyer. 2017. Heterogeneous Integration at Fine Pitch (10 µm) Using Thermal Compression Bonding. In *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*. 1276–1284.
[7] A. A. Bajwa, S. Jangam, S. Pal, B. Vaisband, R. Irwin, M. Goorsky, and S. S. Iyer. 2018. Demonstration of a Heterogeneously Integrated System-on-Wafer (SoW) Assembly. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*. 1926–1930.
[8] N. Beck, S. White, M. Paraschou, and S. Naffziger. 2018. 'Zeppelin': An SoC for multichip architectures. In *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*. 40–42.
[9] Y. Chuang, C. Yuan, J. Chen, C. Chen, C. Yang, W. Changchien, C. C. C. Liu, and F. Lee. 2013. Unified methodology for heterogeneous integration with CoWoS technology. In *2013 IEEE 63rd Electronic Components and Technology Conference*. 852–859.
[10] DARPA. 2020. Common Heterogeneous Integration and IP Reuse Strategies (CHIPS).
[11] Pete Ehrett, Vidushi Goyal, Opeoluwa Matthews, Reetuparna Das, Todd Austin, and Valeria Bertacco. [n.d.]. Analysis of microbump overheads for 2.5 d disintegrated design. ([n. d.]).
[12] S. Y. Hou, W. C. Chen, C. Hu, C. Chiu, K. C. Ting, T. S. Lin, W. H. Wei, W. C. Chiou, V. J. C. Lin, V. C. Y. Chang, C. T. Wang, C. H. Wu, and D. Yu. 2017. Wafer-Level Integration of an Advanced Logic-Memory System Through the Second-Generation CoWoS Technology. *IEEE Transactions on Electron Devices* 64, 10 (2017), 4071–4077.
[13] Intel. [n.d.]. Accelerating Innovation Through A Standard Chiplet Interface: The Advanced Interface Bus (AIB).
[14] S. Jangam, A. A. Bajwa, K. K. Thankkappan, P. Kittur, and S. S. Iyer. 2018. Electrical Characterization of High Performance Fine Pitch Interconnects in Silicon-Interconnect Fabric. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*. 1283–1288.
[15] Shin-puu Jeng, Hsien-Wei Chen, Shang-Yun Hou, Hao-Yi Tsai, Anbiarshy NF Wu, and Yu-Wen Liu. 2012. Protective seal ring for preventing die-saw induced stress. US Patent 8,334,582.
[16] M. A. Karim, P. D. Franzon, and A. Kumar. 2013. Power comparison of 2D, 3D and 2.5D interconnect solutions and power optimization of interposer interconnects. In *2013 IEEE 63rd Electronic Components and Technology Conference*. 860–866.

[17] N. Kim, D. Wu, D. Kim, A. Rahman, and P. Wu. 2011. Interposer design optimization for high frequency signal transmission in passive and active interposer using through silicon via (TSV). In *2011 IEEE 61st Electronic Components and Technology Conference (ECTC)*. 1160–1167.

[18] Thierry Lazerand and David Lishan. 2014. Wafer Dicing Using Dry Etching on Standard Tapes and Frames.

[19] John Lee and Mike Kelly. 2018. Amkor's 2.5D Package and HDFO – Advanced Heterogeneous Packaging Solutions. China Integrated Circuits.

[20] Wei-Sheng Lei, Ajay Kumar, and Rao Yalamanchili. 2012. Die singulation technologies for advanced packaging: A critical review. *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 30, 4 (2012), 040801.

[21] C. C. Liu, S. Chen, F. Kuo, H. Chen, E. Yeh, C. Hsieh, L. Huang, M. Chiu, J. Yeh, T. Lin, T. Yeh, S. Hou, J. Hung, J. Lin, C. Jou, C. Wang, S. Jeng, and D. C. H. Yu. 2012. High-performance integrated fan-out wafer level packaging (InFO-WLP): Technology and system integration. In *2012 International Electron Devices Meeting*. 14.1.1–14.1.4.

[22] R. Mahajan, R. Sankman, N. Patel, D. Kim, K. Aygun, Z. Qian, Y. Mekonnen, I. Salama, S. Sharan, D. Iyengar, and D. Mallik. 2016. Embedded Multi-die Interconnect Bridge (EMIB) – A High Density, High Bandwidth Packaging Interconnect. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*. 557–565.

[23] M. Matsuo, N. Hayasaka, K. Okumura, E. Hosomi, and C. Takubo. 2000. Silicon interposer technology for high-density package. In *2000 Proceedings. 50th Electronic Components and Technology Conference (Cat. No.00CH37070)*. 1455–1459.

[24] M. O'Connor, N. Chatterjee, D. Lee, J. Wilson, A. Agrawal, S. W. Keckler, and W. J. Dally. 2017. Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems. In *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. 41–54.

[25] S. Pal, D. Petrisko, R. Kumar, and P. Gupta. 2020. Design Space Exploration for Chiplet-Assembly-Based Processors. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 28, 4 (2020), 1062–1073.

[26] N. Pantano, C. R. Neve, G. Van der Plas, M. Detalle, M. Verhelst, M. Heyns, and E. Beyne. 2016. Technology optimization for high bandwidth density applications on 3D interposer. In *2016 6th Electronic System-Integration Technology Conference (ESTC)*. 1–6.

[27] Next Platform. 2018. Building Bigger, Faster GPU Clusters using NVSwitches.

[28] Open Compute Project. 2020. OCP Open Domain-Specific Architecture Sub-Project. Retrieved Nov, 2020 from https://www.opencompute.org/wiki/Server/ODSA#Project_Leadership

[29] S. Ramalingam. 2016. HBM package integration: Technology trends, challenges and applications. In *2016 IEEE Hot Chips 28 Symposium (HCS)*. 1–17.

[30] Y. Sa, S. Yoo, Y. Shin, M. Han, and C. Lee. 2010. Joint properties of solder capped copper pillars for 3D packaging. In *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)*. 2019–2024.

[31] L. Shan, Y. Kwark, C. Baks, M. Gaynes, T. Chainer, M. Kapfhammer, H. Saiki, A. Kuhara, G. Aguiar, N. Ban, and Y. Nukaya. 2015. Organic Multi-Chip Module for high performance systems. In *2015 IEEE 65th Electronic Components and Technology Conference (ECTC)*. 1725–1729.

[32] Shyh-Chyi Wong, Gwo-Yann Lee, and Dye-Jyun Ma. 2000. Modeling of interconnect capacitance, delay, and crosstalk in VLSI. *IEEE Transactions on Semiconductor Manufacturing* 13, 1 (2000), 108–111.

[33] David Robert Stauffer, Jeanne Trinko Mechler, Michael A Sorna, Kent Dramstad, Clarence Rosser Ogilvie, Amanullah Mohammad, and James Donald Rockrohr. 2008. *High speed serdes devices and applications*. Springer Science & Business Media.

[34] W. J. Turner, J. W. Poulton, J. M. Wilson, X. Chen, S. G. Tell, M. Fojtik, T. H. Greer, B. Zimmer, S. Song, N. Nedovic, S. S. Kudva, S. R. Sudhakaran, R. Bashirullah, W. Zhao, W. J. Dally, and C. T. Gray. 2018. Ground-referenced signaling for intra-chip and short-reach chip-to-chip interconnects. In *2018 IEEE Custom Integrated Circuits Conference (CICC)*. 1–8.