System-Technology Co-Optimization for Advanced Integration: A Perspective in the Computing Context

Saptadeep Pal[1], Arindam Mallik[2], Puneet Gupta[3]
[1] Etched.ai, U.S.A. [2] IMEC, Belgium [3] ECE Department, UCLA, U.S.A.
[3]puneetg@ucla.edu

April 17, 2025

Abstract

Advanced integration and packaging will drive the scaling of computing systems in the next decade. Diversity in performance, cost, and scale of the emerging systems implies that system-technology co-optimization (STCO) would be essential to develop these integration technologies for future systems. Such STCO would need to comprehend not only integration technology, circuits, architectures, and software but also their interactions with the power delivery, cooling, and system costs. We present a perspective on what would be needed from these STCO approaches with exemplar case studies covering the current state of the art and the future outlook.

1 Introduction

Traditionally, dimensional scaling has been the primary driver for dramatic improvements in the power, performance, form factor, and cost of electronic integrated systems. Aggressive scaling of CMOS silicon and wiring minimum features of over 1000× for over four decades (enabled by advancements in patterning technologies) coupled with performance boosters such as the adoption of copper wiring, strained silicon, and FinFETs have delivered on the promise of Moore's law. Unfortunately, this scaling has come at exponentially increasing cost [1–3]. This is becoming increasingly untenable as we approach physical limits. This is forcing the semiconductor industry to take a careful look at the "system on chip" trend of the past few decades.

A chip is rarely the whole system. It is packaged and bonded to a printed circuit board (PCB), with a "fan-out" at each level (see Fig. 2). Although the dimensions within the chip have been scaled by more than three orders of magnitude in the last five decades, the dimensions of package/PCB input/output or I/O (Ball Grid Array or BGA) bumps have scaled barely 5X [4]. As a result,

multi-package systems on a PCB suffer in all aspects: power, performance, area, and cost, which drove the industry toward systems on chip. With chip scaling becoming more difficult, there is a new focus on advancing packaging to scale inter-chip connectivity. This approach has the potential to reduce the cost of large systems, making communication overheads much better and enabling new types of systems with intimately connected heterogeneous components. Advanced integration will be a system-scaling driver in the coming decade.

The semiconductor industry has long relied on separating concerns between design and manufacturing. Several abstraction aids, such as design rules and compact device models, have been developed to preserve the clean abstraction of technology available to circuit designers. This has made design and technology development largely independent of each other. Unfortunately, difficulty in scaling has blurred these boundaries and made the co-optimization of design approaches and technology development essential. This has resulted in a strong interest in design-technology co-optimization (DTCO), especially in the development of device technology [5–9] and lithographic patterning [10–12]. The eventual choice of patterning scheme at any technology node has as much been dictated by design considerations (e.g., ease of design, availability of design automation tools, block-level power/performance/area metrics) as by the difficulty of the technology itself. Over time, DTCO approaches have become increasingly sophisticated, ranging from the earlier manual design of small benchmarks [12] to elaborate stitched electronic design automation (EDA) tool flows [12, 13] to principled and fast frameworks [10, 14].

Integration and packaging are going through a renaissance and will see significant innovation in the coming years. Limiting DTCO to a single die is no longer sufficient, and evaluating an entire system that can consist of several chips integrated together using packaging technology would be essential [15, 16]. This system-technology co-optimization (STCO) is needed to guide innovation in the right direction. STCO approaches are still in their infancy owing to the lack of automated frameworks. Eventually, STCO would need to account for multiple facets, such as within-chip technologies (device, patterning, interconnect), heterogeneous system component technologies (e.g., memory types), ways of connecting chips (2.5D or 3D integration), power delivery and cooling infrastructure, and architecture and software applications running on hardware. As shown in (Figure 1), the future of system scaling is highly dependent on cross-layer optimization of different abstraction layers of computing systems. Traditionally, the semiconductor industry has scaled logic, memory, and interconnects separately and largely independently of the systems being constructed using them. The future trend would be to optimize system functions or modules using the process technology best suited to it. In practice, that means building each on its chiplet. Then an advanced packaging scheme, such as advanced 3D stacking, would bind those together using technology so that all the functions act as if they were on the same piece of silicon.

This paper motivates the need for STCO in the context of packaging for computing systems. Section 2 provides recent industry examples of different choices of packaging approaches driven by the system context. Section 3 details

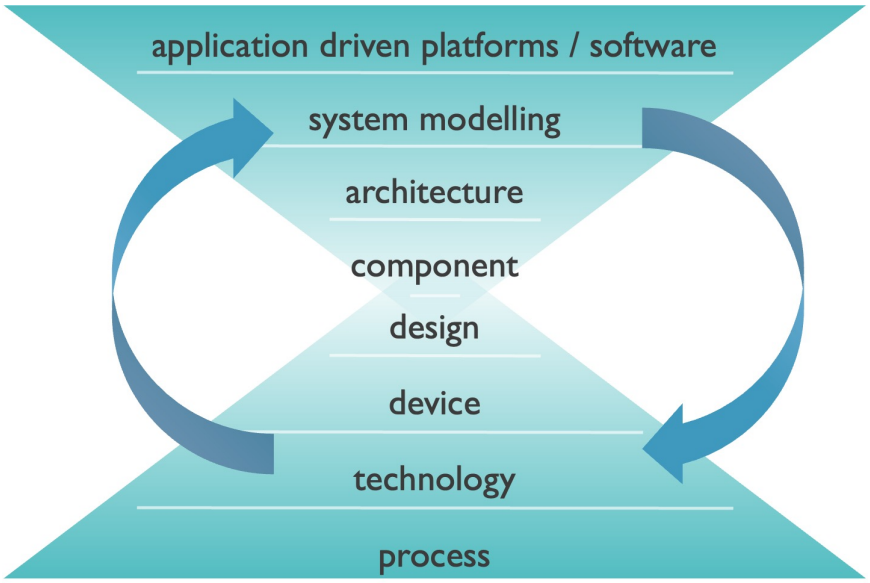


Figure 1: Cross-stack STCO optimization. Cross-layer optimization paves the way for system scaling [17]. Section 2 gives examples of such cross-layer optimizations.

the system drivers and the enablers for advanced integration. Different points on the multidimensional Pareto frontier of these drivers/enablers would dictate viable integration technologies. Section 4 discusses some of the emerging approaches for DTCO/STCO for advanced packaging and integration. Finally, we conclude the paper in section 5.

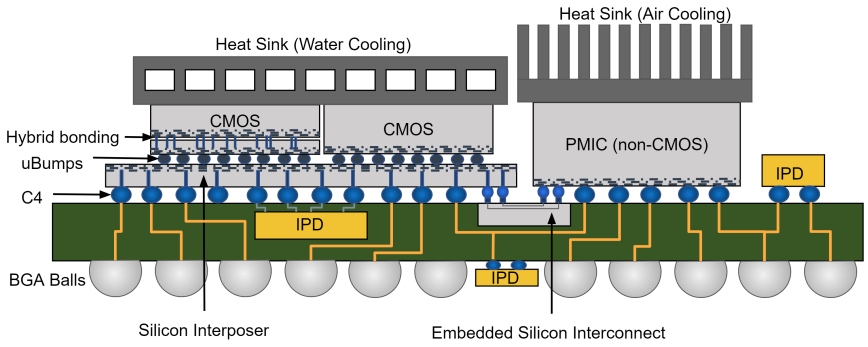


Figure 2: A cross-section view of a multi-chiplet packaged system is shown. A diversity of chiplet integration technologies alongside power delivery and thermal management components are tightly integrated to realize the full potential of such a system.

2 Design-Dependent Choice of Advanced Packaging: A Comparative Case Study

With the proliferation of applications demanding high performance (e.g., artificial intelligence), the requirement for larger silicon systems has grown exponentially as shown in Figure 4. Over the past decade, advanced packaging technologies have improved the scale, performance, and energy efficiency of silicon systems. It allows us to build large chips by dense integration of multiple silicon dies inside a package. These technologies have different flavors, with trade-offs between integration density, scale, and cost. For example, organic interposers are cheaper but allow for a lower interconnect density between adjacent dies compared to silicon interposers. Therefore, depending on the target application and market, the right advanced packaging technology needs to be chosen, and the architecture needs to be co-optimized.

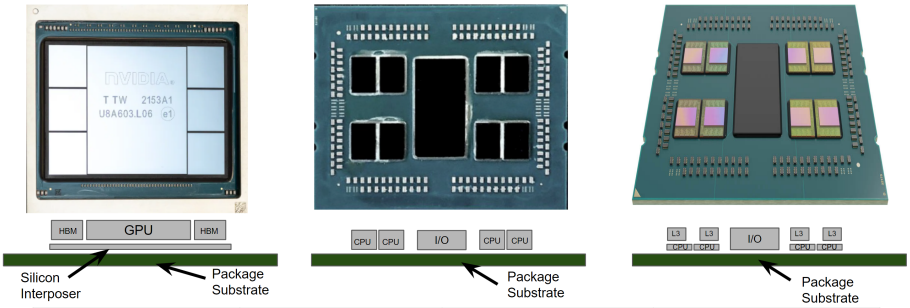
Recent developments have showcased a trend toward the design-dependent co-optimization of system architecture and advanced packaging across various products. In the following section, we will delve into a comparative analysis of several notable examples of this trend.

The field-programmable gate array (FPGA) industry is one of the earliest adopters of silicon interposers [18]. In the late 2000s, because of their easier reconfigurability and quick turnaround time, FPGAs gained popularity, and larger systems based on FPGAs began to develop. FPGAs have 20-40x lower compute density. Therefore, FPGA silicon started becoming as large as a full reticle [18] and systems were regularly built using multiple FPGAs on a board. Reticle-sized silicon is yield-limited and, therefore, costly. In addition, multi-FPGA solutions often exhibit poor performance. To alleviate these issues, Xilinx used silicon interposers to build large FPGAs. Silicon interposers allow the integration of multiple known-good-dies at a high interconnect density, allowing for lower-cost FPGA products. Moreover, it allows FPGAs to be built with integrated high-bandwidth memory (HBM) thus making them viable alternatives to building application-specific integrated circuits (ASICs). For instance, Microsoft adopted FPGAs as the de facto platform to build custom accelerators [19].

Similarly, manufacturing yield concerns for building large core-count monolithic central processing units (CPUs) pushed AMD to adopt a chiplet-based architecture. Disintegrating a large monolithic processor into smaller chiplets allowed AMD to build processors with known-good dies and save on cost, often as much as 2.1x [20]. Moreover, AMD leveraged the cost benefits of heterogeneous integration by integrating external I/O circuitry into an I/O chiplet on a lower-cost 12 nm node, as opposed to the core chiplets fabricated on an expensive 7 nm node. Cost constraints forced AMD to use organic substrates for chiplet integration rather than the expensive silicon interposers used by FPGAs and graphics processing units (GPUs). This was enabled by the co-design of the architecture with the packaging substrate characteristics, and the fact that the inter-chiplet bandwidth required was only a few 100s of GB/s. Additionally, a chiplet-based methodology provides flexibility for building multiple product

lines by altering the number of chiplets. AMD and Xilinx leveraged this flexibility to save non-recurring engineering (NRE) costs and improve the time to market for different product lines.

The demand for high-performance computing (HPC) and artificial intelligence (AI) applications is driving the adoption of very high-bandwidth in-package integration technologies such as silicon interposers and silicon bridges. These applications are highly parallel and primarily run on accelerators such as general-purpose graphics processing units (GPGPUs) and Google tensor processing units (TPUs). These accelerators are highly parallel (e.g. 14,592 FP32 cores in NVIDIA H100) with a large amount of computing throughput, often more than one PFLOPs of compute per die. Large computing throughput requires higher memory bandwidth [21]. Consequently, accelerator architectures rely on on-package DRAMs to provide the required bandwidth (e.g., 3 TB/s on an NVIDIA H100 GPU [22]). Multiple HBM devices are integrated with the accelerator compute die within the package [23]. HBMs use wide memory interfaces (e.g. 16x DDR channels per device), and each pin supports a data rate of <10 Gbps to maintain low I/O energy and area overhead. Integration technologies using silicon for inter-die links can accommodate a 10x higher density of signal pins and traces. Consequently, accelerators such as GPGPUs and TPUs use technologies such as CoWoS-S [24], CoWoS-L [25], and EMIB [26, 27] instead of organic substrates for inter-chiplet connectivity.



	NVIDIA GPU	AMD CPU	AMD CPU w/ 3D V-Cache
Interconnect Substrate	Silicon Interposer	Organic Substrate	Organic Substrate
Bump Pitch	45 um (Die-to-Interposer)	130 - 150 um (Die-to-Substrate)	130-150 um (Die-to-Substrate) 9 um (3D Die-to-Die)
Inter-Chiplet I/O Type and Width	Parallel (1024 Data + ~80 Control)	SerDes (72 Data + 8 Control)	SerDes (72 Data + 8 Control) Parallel (10,000+ for 3D Die-to-Die)
Total Inter-Die Data Bandwidth	512 GB/s (GPU-HBM)	~1 TB/s (Core Complex Die - I/O Die)	~1.8 TB/s (Core Complex Die - I/O Die) 2.5 TB/s (3D L3 - Core Complex Die)
Cost	High	Low	Medium

Figure 3: Various integration schemes provide different interconnect characteristics and integration density. The sources of the images and data of NVIDIA GH100, AMD EPYC 2nd Generation, and AMD EPYC with V-Cache are [24, 28], [20, 29] and [30, 31] respectively.

Beyond 2.5D integration using chiplets, 3D integration of two active dies

on top of each other, is gaining steam. Certain HPC, gaming, and multimedia workloads benefit from larger caches [31]. However, static random access memory (SRAM) cost and area scaling have been underperforming compared to logic scaling over the past few technological nodes [32–34]. AMD introduced 3D integration of a cache die on top of a CPU die in their V-Cache technology. This is a clever and elegant co-design of architecture and packaging. 3D integration using hybrid bonding can provide 25x I/O density [31, 35] and a shorter interconnect distance and energy than 2.5D integration. Therefore, it can provide the on-chip-like bandwidth needed by the cache subsystem with minimal energy overhead. In one incarnation, the bottom CPU die is built in an expensive 5 nm node, whereas the cache die is built in a relatively cheaper 7 nm node optimized for SRAM, thus improving the overall cost of the system.

These case studies show how careful co-design of the chiplet-based system architecture and integration scheme can lead to optimized product solutions. Figure 3 shows recent commercial products such as NVIDIA GH100 [24, 28], 2nd generation AMD EPYC [20, 29] and 3rd generation AMD EPYC with V-Cache [30, 31] and show the characteristics of the respective integration schemes used. We argue that system-technology co-optimization is critical to the success of next-generation products when the cost benefits of moving to newer technology nodes are dwindling. In addition, with the recent surge in demand for AI [36] and other HPC workloads [37], custom ASICs are becoming mainstream. STCO frameworks are required to guide the choice of both architecture and technology selection to extract the most value from these systems.

3 Advanced Integration: Key Drivers and their Design Interactions

In this section, we discuss system metrics that drive multi-chip integration as well as system enablers for advanced packaging of future computing systems.

3.1 System Drivers

Let's begin by asking: what are the primary drivers behind the surging demand for advanced packaging technologies? The need for high-performance and energy-efficient connectivity between components inside a package is growing in scale. Additionally, the need for cost optimization and form factor minimization is driving the development of advanced packaging. We will now explore these factors in detail.

3.1.1 Connectivity

Connectivity is the primary driver behind the development of advanced packaging solutions. Poor scaling of off-package links becomes a barrier to system performance and power scaling when integrating multiple packaged chips on a PCB. Integrating chiplets inside a package is driven by the increased inter-die

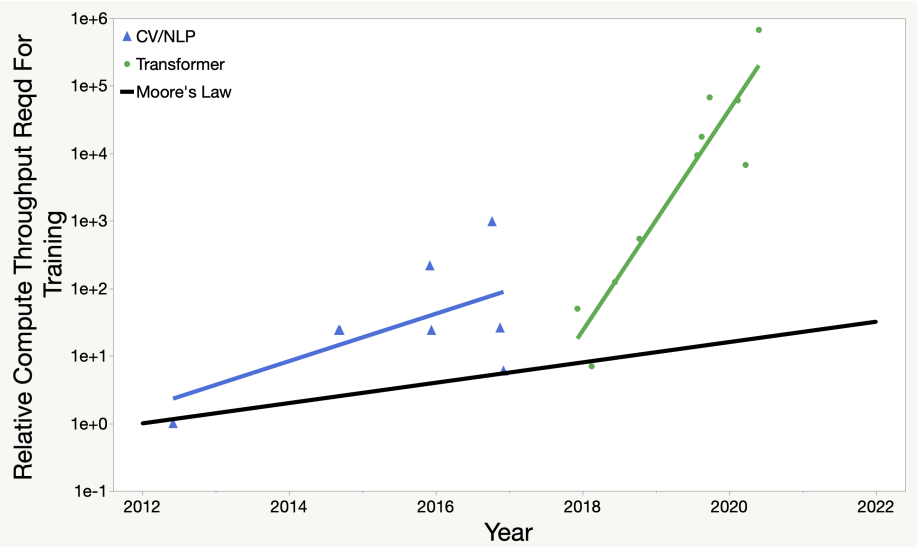


Figure 4: Computing demand for AI and HPC workloads is orders of magnitude higher than what Moore's Law can provide [38].

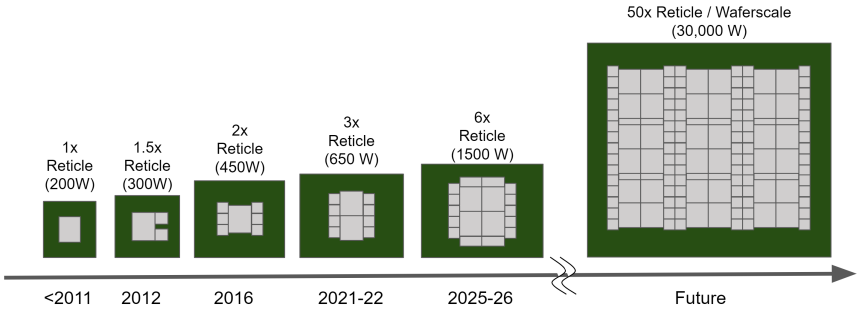


Figure 5: Driven by the extreme growth of computing demand in the HPC and AI workloads, advanced packaging technologies are evolving to integrate large amounts of silicon in a single package. This alone is insufficient; co-optimization is necessary to extract maximum performance from the silicon area.

connectivity that can be achieved inside a package. Today's HPC and AI workloads demand multiple TB/s of bandwidth between different compute and memory chiplets. Therefore, the development of advanced packaging technologies is geared towards enabling high inter-chiplet bandwidths at low energy overheads. This is accomplished by reducing I/O pitch ($<20\mu m$ vs $>200\mu m$ for off-package I/Os) [4, 39], interconnect wiring pitch ($<5\mu m$ vs $>50\mu m$) and length ($<1mm$ vs >10cm) [39], which enables efficient highly parallel interfaces. Furthermore, this reduces the need for power-hungry high-speed serializer-deserializer (SerDes) circuitry that is needed to drive high data rates over individual interconnects in I/O-constrained designs. Today, >10x bandwidth at equivalent interconnect power can be achieved between chiplets integrated on a package compared to chips interconnected over a PCB. (e.g., up to 6TB/s of memory bandwidth can be achieved using six HBM3 [40] modules at ∼160W of inter-chiplet interconnect power). This is an order of magnitude higher bandwidth than that can be achieved using off-package memories over double data rate (DDR) interface [41, 42] at iso-power. Similarly, 3D integration enables another step function improvement in I/O density (>15x) and energy efficiency (>3x) [43].

3.1.2 Scale

Improved connectivity facilitates system scaling within a package. As new workloads and data processing techniques demand increasingly parallel hardware, this scaling becomes essential. Compute requirements for machine learning workloads alone have far outpaced gains from Moore's Law (see Figure 4). As evident from several recent trends [25, 44], silicon area per chip is growing fast to meet this seemingly insatiable demand (see Figure 5). This is driving enormous research and development efforts for future advanced packaging technologies. As discussed before newer advanced packaging technologies, such as CoWoS-L are being developed to integrate up to 5000 mm^2, i.e., six reticles worth of silicon [25] in a single package. At the extreme, waferscale integration technologies are being developed commercially [45, 46] and in academia [44, 47] to build systems that are large as an entire 300mm wafer. For some classes of applications, these technologies would enable systems that can provide an order of magnitude performance gain over systems built using conventional packages [47, 48].

3.1.3 Cost

Though advanced packaging provides us with newer platforms for more connected and scaled systems, the primary driver behind the acceptance of a new technology is cost (often cost per performance). Given the manufacturing complexities of advanced packaging, can it offer economic advantages for the next generation of electronic systems? The traditional path for improving the cost of digital systems through silicon CMOS scaling is becoming increasingly difficult [1–3]. Chiplets are best thought of as an alternative design methodology to monolithic chips in a world where Moore's Law has largely stopped being an economic benefit. It can help improve yield and reduce costs by allowing man-

ufacturers to use smaller, more specialized chiplets rather than a single, monolithic chip for certain tasks [49, 50]. AMD has demonstrated the economics of the chiplet approach to building its Ryzen client processors. A 16-core Ryzen chip, such as the Ryzen 9 5950X, built on a monolithic 7 nm die, would have cost AMD 2.1 times more in comparison to its chiplet-based approach of using two 8-core 80 mm^2 core complex dies paired with a cheaper 12 nm I/O die. [20]. By modularizing the system based on chiplets, it can be customized for each market segment by simply adding or removing more chiplets. This not only saves cost but enables faster design and time-to-market. The overall benefit can be seen in the total cost of ownership (TCO) of the hardware [51]. Hence, chiplets are fueling a new era of innovation within the semiconductor industry based on a flexible and cost-effective economic model.

3.1.4 Form-factor

Consumer electronics devices (laptops, mobile phones, smartwatches, etc.) have aggressively driven several packaging and integration technologies over the past couple of decades to maximize miniaturization and energy efficiency. Packaging technologies such as integrated fan-out wafer level packaging (InFO) [52], package-on-package (PoP) [53], wire-bonded chip scale packages (WB-CSP), flip-chip system-in-package (SiP) allow systems to be built with minimal area and volumetric footprint. For example, smartwatches and mobile phones integrate power management IC and memory chips with the system-on-chip (SoC) using PoP and SiP techniques. Similarly, Apple's new M-series processors integrate LPDDR memory packages with processor SoC die on the same package substrate. These technologies improve the form factor of these devices by as much as 50% [54, 55]. These examples show that advanced packaging plays a key role in enabling different use cases which would not have been possible with traditional single-chip packaging technologies.

3.2 System Enablers

Figure 2 captures an outlook of a future system platform where heterogeneity of technology nodes, better connectivity, and co-integration of specialized components help to provide a step function improvement in the factors mentioned in Section 3. In this section, we will discuss the system-level metrics that enable the future of advanced packaging schemes.

3.2.1 Technology Heterogeneity

Chipletization opens a major avenue for improved functional integration: intimate connection of disparate process technologies. In the past, the trend in the semiconductor industry has been toward a "siliconification" of all functions due to cost, form factor, and short-hop connectivity to the silicon CMOS compute fabric (i.e., the SoC trend). Advanced integration (both 2.5D and 3D) allows system designers to buck this trend with possible gains in power and

performance. Some examples of such technological heterogeneity include the following.

- *Intimately Connected Memories.* High-bandwidth memories (HBMs) which use a DRAM process are now connected at very short distances (<5mm) to the compute substrate with very high bandwidth [56–58]. This has improved performance, especially for memory-bottlenecked machine learning workloads. One can envision similar tight integration with other types of memory and storage technologies such as Flash.
- *Intimately Connected Off-Package Interconnect.* High-bandwidth, low-energy, low-latency photonic interconnect [59] has been another representative example of leveraging chiplet heterogeneity, which would otherwise have required much worse pluggable optics or electrical links.
- *Intimately Connected Power Delivery Infrastructure.* Efficient integrated voltage regulators (e.g., using Gallium Nitride technology (GaN) transistors [60, 61]) and within-package or within-interposer passives (capacitors and inductors) can dramatically improve power delivery efficiencies for large high-power systems [62].

Although multi-chip modules [63, 64] and systems in package [65, 66] of the past allowed heterogeneous integration as well, the proximity of the different chiplets was over 1-2 orders of magnitude worse ($\sim$1cm vs. $\sim$100μm).

3.2.2 Power Delivery

Advanced packaging enables systems with higher power density in a package. As a result, power integrity challenges in these systems need to be addressed by holistically looking at the integration technology. Novel techniques (architecture, design) and technologies (materials, in-substrate capacitors) are being developed and more are needed to provide power reliably. Recently, TSMC has started embedding deep-trench capacitors in the silicon interposer. Similarly, newer versions of CoWoS, (CoWoS-R [67] and CoWoS-L [25]) are being developed with integrated passive devices for better power integrity [68]. GraphCore [69] used 3D integration (based on wafer-to-wafer bonding) to integrate a deep-trench capacitor die alongside the compute die resulting in approximately 40% higher performance. To build a system-in-package solution with CPU, GPU, accelerator, and memory dies on an interposer, the platform Voltage Regulator (VR) needs to be integrated on the interposer close to the logic dies. This could be enabled using high-voltage complementary GaN (CGaN) devices with inductors embedded in the package using high-frequency high permeability materials. [60, 62]

Stable power supply to the microprocessor is important ensure optimal performance. As technology nodes shrink, power density and IR drop increase, challenging designers to maintain the 10 percent margin that is allowed for the power loss between the voltage regulator and the transistors. The development of a high-efficiency, dense Integrated Voltage Regulator (IVR) will be critical

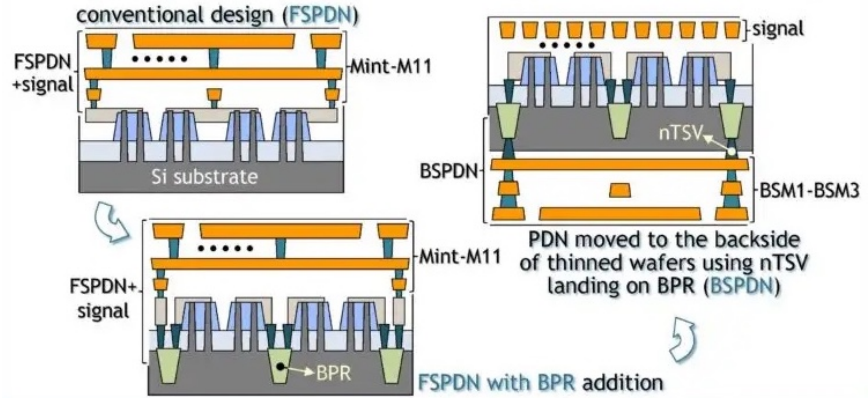


Figure 6: Backside power delivery network (BS-PDN) enablement for power delivery [70, 71]

to meet the requirements of future high-performance microprocessors [62]. Alternatively, a backside power delivery network (BSPDN) decouples the power delivery network from the signal network by moving the entire power distribution network to the backside of the silicon wafer. Figure 6 shows the BSPDN enablement at the process technology level. This approach promises to benefit the IR drop, improve the power delivery performance, reduce routing congestion in the back-end-of-line (BEOL), and allow standard cell height scaling [72] [73, 74]. A backside PDN looks promising for the performance improvement of 3D systems-on-chip (3D SOCs) [70]. For both 2D and 3D designs, the concept of exploiting the wafer's free backside can potentially be expanded by adding specific devices in the backside, such as I/Os or ESD (electrostatic discharge) devices [75].

3.2.3 Thermal Management

The rise of hyperscaled data centers and artificial intelligence (AI) computing has already increased the rack power density from 10-20 kW per rack to more than 30 kW per rack. In the near future, this number is expected to double. Increased power density exacerbates the thermal problem in a system. This necessitates advanced cooling technologies such as liquid cooling, phase-change cooling, and even techniques such as immersion cooling [76, 77].

With heterogeneous packaging, there is a power density disparity across the total area of the package. It corresponds to a higher temperature gradient across the whole package, which can be addressed by novel heat spreader methodologies [78]. At the same time, the challenge of dissimilar heights of individual chiplets (e.g. a logic die chiplet versus a high bandwidth memory (HBM) module) needs varying cavity depth to use an integrated heat spreader [79]. On the positive side, chipletization benefits thermal performance because heat-generating components are spread apart, thus reducing their thermal crosstalk

[80]. Additionally, it helps the reliability of thermally sensitive components in the package, as well as overall system-level reliability [81–83].

The novel utilization of features specific to 2.5D or 3-D integration, such as through-silicon vias (TSVs) for heat dissipation and management is an interesting aspect. Thermal-aware floorplanning can manage heat loads by optimizing the distribution of circuit components and TSVs, effectively reducing junction temperatures across the die [84–86]. Multiple pieces of research have been carried out to co-optimize thermal and electrical design challenges [87]. TSVs have been used as a heat-removal mechanism [88]. In addition, Codesign approaches that couple TSVs with microfluidic cooling [89], silicon micropin fin [90], or air gaps [91] have been reported recently.

Overall, thermal management challenges in advanced packaging are closely related to electrical performance and manufacturing. These coupled phenomena often present critical trade-offs and constraints that must be correctly recognized and accounted for, through system technology co-optimization.

4 STCO Methodologies and Frameworks

Advanced packaging innovation will enrich system-in-package (SiP) technology helping the semiconductor community to continue the benefits of Moore's Law but at a system-scale. Moore's Law has allowed for the production of less expensive semiconductors, that dissipate less power and have higher performance. This has led to a large demand for semiconductor systems with a wide range of integrated functionalities on a single die. On the other hand, Moore's Law scaling is slowing down and Dennard's Law has been near-dead for over a decade now. As a result, building high-performance, low-power, and cost-effective silicon systems is no longer just about realizing a design in one semiconductor manufacturing process. The monolithic SoC way of designing electronic systems is losing its viability as a cost-efficient, functional option for system integration. SiP, however, opens the door to the design of a nearly limitless variety of complex systems. SiP provides opportunities as well as new challenges across the entire stack that encompasses technology development, design, manufacturing, testing, and system software.

Recent examples of such co-optimization have been emerging both in industry and academia. Let us look into a few such examples of STCO.

- Cerebras [45]attempted to solve the problem of accelerating large AI workloads run across multiple chips in a compute cluster. Instead of cutting up a wafer into dies to make traditional chips, they carve out a larger square within the round 300-millimeter wafer. That's a total of 84 dies, with 850,000 cores, all on a single chip. Cerebras architecture enabled running large ML models on a single chip without portioning, scaling becomes easy and natural. This required them to rethink system architecture. New packaging technology, power delivery techniques, and cooling systems were co-developed with the waferscale architecture to realize a truly unique cluster-in-a-box system.

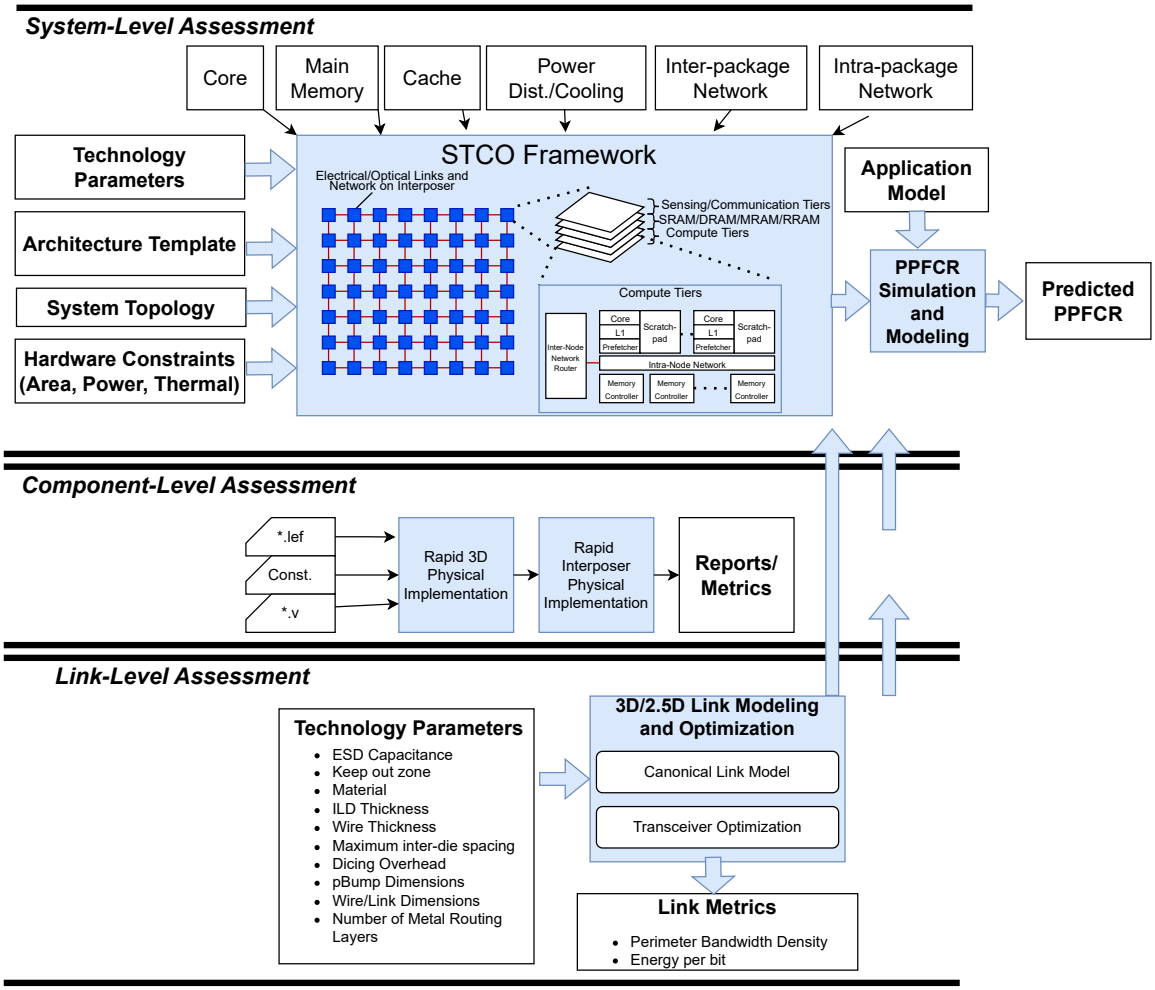


Figure 7: An overview of a STCO framework to predict system-level Power, Performance, Form-Factor, Cost, and Reliability (PPFCR).

- GraphCore was faced with a problem of dynamic voltage droop in the package causing performance loss. They used a wafer-on-wafer hybrid bonding technology to 3D stack the accelerator die on top of a power delivery die [92]. This allowed them to improve AI workload performance by 40%. CMOS process technology scaling alone has stopped providing such leaps in performance, whereas the use of clever design, integration, and manufacturing techniques can help realize the true performance potential of a system.

- A recent work [50] attempted to understand what the minimum size of a chiplet should be such that the overall cost is minimized. The authors showed that the cost of high-performance 2.5D substrates, inter-chiplet I/O overheads, assembly yield issues, and cost of the die-to-substrate bonding can out-strip the yield and system composability benefits that chipletization of large silicon systems offer. The results show that for micro-processor class chiplets, the minimum size of chiplets would be around $40mm^2$, and that increases to $200mm^2$ for random logic. These results mean that bring-your-own-hardened intellectual property (IP) business models may not be feasible as $40mm^2$ is very large real estate and would require multiple IPs in a chiplet. Selection of the right IPs requires an understanding of the diverse set of applications such chiplets would be targeted towards.

- NVIDIA showed how careful optimization of architecture, design, and packaging technology can be leveraged to target GPUs for different markets such as HPC, AI, etc. They propose a Composable On-PAckage GPU (COPA-GPU) architecture [93] to provide domain-specialization. In one incarnation, an additional cache layer can be realized by either 3D integrating a cache die beneath the GPU die or 2.5D integrating multiple cache dies between the GPU and the HBM devices on the package. Each of these options offers different performance, power, and physical size trade-offs and just by leveraging packaging constructs with architectural optimizations, the paper showed that the same training performance can be achieved with a 50% lesser number of GPU instances.

These examples show that STCO can unleash the true potential of SiPs. To enable this, we need frameworks, methodologies, and tools for STCO. Though industrial organizations have internal methodologies and frameworks, neither the tools nor the methodologies are public. That has changed in recent years. We categorize the set of STCO frameworks into three categories: link-level, component-level, and cross-stack system-level. Figure 7 shows an overview of these frameworks. All three levels of STCO can provide useful information about power, performance, cost, and form-factor metrics but at different levels of abstraction and detail. Further, link and component-level modeling can feed into the true cross-stack STCO. Below, we discuss the present state-of-the-art and their shortcomings. Subsequently, we outline emerging directions in full-stack STCO.

4.1 STCO: Link-Level

Advanced packaging technologies bridge the large gap between on-chip and off-package interconnects. Several frameworks have been built in the recent past to model and optimize the inter-die link characteristics. Recent works such as [39, 94, 95] analyze how different parameters of silicon substrates such as interconnect length, inter-layer dielectric (ILD) material, μbump pitch, inter-die spacing, ESD capacitance, etc., affect inter-chiplet bandwidth and energy efficiency. Using these tools, one can figure out which of these parameters should be improved upon or invested in. For example, authors in [39] showed that scaling μbump pitches below $20\mu m$ wouldn't provide meaningful bandwidth or energy efficiency gains unless the I/O ESD requirement is reduced. Signaling figures of merit have been developed as well [96]. These frameworks rely on simple I/O circuits to do the design space exploration, which is suitable for integration technologies where the links are very short. On the other hand, [20, 26] have shown that organic substrates are suitable for cost reasons and, therefore, they co-optimized the I/O circuit-design for organic substrates with comparatively longer links (5-10x) and less density than that of silicon interposers.

Link-level STCO however doesn't cover the system-level implications of link characteristics. For example, a 2x reduction in link energy efficiency may affect total power by a couple of percentage points while improving reliability and cost significantly. Though past works have laid a foundation, more comprehensive exploration tools are needed to explore the design space of different integration schemes and their impact on the overall chip. First, characteristics of substrate technologies such as their material properties (which affect cost and reliability), and interconnect (wiring and bump) characteristics are available for a subset of these technologies. Process design kits (PDKs) and models in standardized formats need to be available to chip designers for simulation even during early phases of technology development (similar to early PDKs made available for advanced CMOS nodes). Second, details and requirements for the ESD protection circuitry are rarely available, and often designers over-design ESD circuitry and rely on post-silicon statistics to understand the impact of ESD events. Therefore, standardized ESD requirements based on the manufacturing environment should be available to the designers. Third, I/O circuits are often over-designed and made available as IPs. These IPs are usually used as is or with minor tweaks inside a chiplet, thus leading to sub-optimal system-level power-performance-area characteristics. Therefore, I/O circuit generators alongside compact analytical models should be developed such that end-to-end interconnect characteristics, including the receiver and transmitters, can be evaluated and their impact on the overall area, power, and performance of the chip architecture can be analyzed early on. This would enable better co-optimization of the I/Os on the chiplets alongside the parameters of the integration technology. Future link STCO research and development should address these shortcomings and develop tools and models using standard EDA and design tools for us to fully leverage today's integration technologies and drive the next generation of integration technologies. Link-level STCO tools should generate abstract final

models of the links and I/Os which can then be used in higher-level tools such as component-level STCO tools. This would enable us to evaluate the interconnect technology's true impact at the system level.

4.2 STCO: Component-Level

The second class of STCO approaches is a natural extension of DTCO methodologies and leverages commercial and academic physical implementation EDA tools. These approaches take one or more benchmark designs (usually modestly sized) and go through an entire chip and system realization flow (placement, routing, power distribution, etc) for multiple chiplets integrated into a system (2.5D or 3D). Such component-level STCO approaches have been used to compare interposer types [97–99], assess back-side power delivery [100–103], evaluate benefits of monolithic [9, 104, 105] or other 3D integration [106], etc. The primary advantage of such approaches is the accuracy of the analyses performed. They exposes the design enablement challenges of new integration technologies. Unfortunately, there are several limitations, especially in the context of STCO for advanced integration. First, these approaches are not scalable to real systems which can have gate counts exceeding 100M spanning several chiplets. Such component implementation approaches worked reasonably well for DTCO in the context of patterning [10, 14, 107–109] where results from small design blocks could be generalized to larger systems-on-chip. However, generalization is difficult for large multi-chiplet packages. For example, inter-chiplet signaling overhead can look much worse for small chiplets [50] while thermal and power delivery problems can look easier. Second, these approaches require the underlying integration approaches to be evaluated to be mature enough to have tool-usable models (e.g. PDKs, Assembly Design Kits (ADKs), etc) which for early technology exploration are rarely available. Third, most assessments require new EDA capability development (e.g., a pseudo-3D design implementation flow using 2D tools such as [110, 111] to do any 3D integration STCO). Although this has the benefit of simultaneous design-enablement of the technology, it severely limits the pathfinding space in STCO. Finally, such component-level approaches ignore the system tradeoffs which are only visible when the system architecture and the application workload running on it are accounted for.

Some of these shortcomings can be addressed by future research. Scalability issues can be partly dealt with by abstracting the physical implementation to block-level rather than gate-level which should give 1-2 orders of magnitude speedup at the cost of hiding some detail (e.g., see [112]). Further, block-level flows could be fed by automated system-level benchmark generators (which don't currently exist) built on top of architecture design space exploration tools such as [113–115]. To better connect the component-level STCO with applications and architectures, analytical performance/power macro models can be developed to be used in conjunction with physical estimates (for example to constrain the physical implementation appropriately).

4.3 Emerging STCO Approaches: Cross-stack

Advanced chiplet integration technologies are platforms for building large systems inside a package. However, traditional DTCO approaches and piecemeal STCO approaches (link and component level) are not suitable for understanding the system-level impact of these technologies. This is because such traditional approaches often do not model the entire system and do not allow us to understand the impact of decisions at the lower levels on application performance and power. Therefore, newer cross-stack frameworks are needed where the interplay of technology, design hardware, and software architecture can be explored by evaluating the impact of the choices made at each layer of the stack at the application level. Recently, a set of efforts to build cross-stack STCO tools have emerged. On the 3D integration front, several hardware/software co-synthesis frameworks have been proposed [112, 116–118] to explore the 3D SoC design space. For interposer-based designs, Floorplet [119] is a framework that can optimally partition a fixed SoC design into chiplets based on yield and reliability, generate the chiplet design, optimize the interposer floorplan and perform cycle-accurate performance simulation to optimize the entire system. DeepFlow [113] on the other hand allows one to co-optimize the chip and the scale-out distributed system architecture alongside the software parallelization strategy for a given machine learning workload. Low-level technology parameters such as area, power, and performance of building blocks (ALUs, SRAMs, DRAMs, interconnects) and physical constraints (power, thermal, pin density, etc.) are provided as inputs. The framework can automatically search the architecture design space, model the performance, and using gradient-descent search approaches search the vast space of hardware-software co-design. These approaches are critical to understanding the end-to-end impact of advanced technology development on application-level performance. Unfortunately, these tools suffer from shortcomings. Since rich cross-stack frameworks need to comprehend and search over an impossibly large parameter space (technology, design, architecture, software, etc.), these tools are built around simplified abstractions and assumptions that render them useful for limited subsets of the design space. For example, Floorplet works with a given design/architecture and therefore isn't able to show what the changes in technology parameters can lead to if one had to re-architect the SoC. DeepFlow targets deep learning workloads and targets exploration of a vast design space. The architecture generation and performance simulation portions of the tool are designed to be workload-specific for runtime efficiency reasons and therefore lack generality.

These tools however provide a solid foundation for future research. As evident, system-level performance modeling is critical for cross-stack STCO tools but these models need to be fast, relatively accurate, and scalable for it to be useful when doing large design space exploration. This requires building composable analytical models for different types of architectural blocks (CPU cores, SIMD cores, accelerators, register, and memory blocks). Such analytical models then need to be coupled with abstract workload modeling where the characteristics of the key kernels can be abstracted and application dataflow graphs can be

input to the simulation model. On the architecture generation front, link-level and component-level tools can help guide the generation of feasible and realizable architectures and SoC designs by providing abstract power, performance, and area models of the different hardware components. On top of this, several novel search techniques (e.g., genetic algorithms, particle swarm optimization, data-driven ML-assisted techniques) can be employed to assess the design space and co-optimize across the stack of technology, chip and system architecture, and software strategies.

4.4 Future STCO Contexts and Directions

Although this paper has looked at advanced integration only from a conventional computing lens, packaging is enabling completely new sensing/computing paradigms. Flexible computing systems [120, 121], biocompatible electronics [122, 123], and heterogeneously co-integrated sensor and compute [124–126] are few examples of such emerging areas of research. STCO in these contexts will take a different flavor than conventional computing but is equally important.

Lastly, we want to emphasize a few important system metrics that we have not discussed but are becoming increasingly important [127], especially in use contexts like automotive [128]. The choice of materials in integration can have a substantial impact on thermo-mechanical stresses [129–131] but this needs to be balanced against cost and performance considerations. Advanced packaging can both help with supply-chain security [132–134] and expose more challenges in both system security [135–138]. Environmentally sustainable manufacturing and reducing the lifecycle carbon and waste footprint of electronics has become critical [139–144]. Packaging is a big part of this footprint and system design using chiplets can open up novel ways of looking at the sustainability problem as well as potentially additional carbon footprint. For example, any tradeoff between the recyclability of packaging materials and their performance implication is an STCO task.

5 Conclusions

Advanced packaging is seeing a renaissance as it is seen as a way to enable "more than Moore" scaling. Including integration/packaging as part of the performance, energy, and total cost of ownership optimization requires expanding the scope of design-technology co-optimization to include the system. Such System-Technology Co-Optimization touches not only aspects of logic/memory chip design/manufacturing but also heterogeneously integrated power delivery, integrated cooling approaches, and off-package interconnect. In this paper, we have discussed some of the existing approaches to STCO. Furthermore, to truly derive full value from technology, we argue that one needs to expand the scope of STCO to be cross-stack and account for micro-architecture and software/algorithms as well. Such materials-to-software frameworks and methodologies that would allow for true STCO are still in their infancy and are likely to be domain-specific

to bound the problem to be tractable.

STCO for advanced integration is going to be a vibrant, high-impact area of research and development in the coming decade and we encourage researchers to take a cross-disciplinary software-hardware-technology cross-stack approach to it.

References

1. Mallik, A., Ryckaert, J., Mercha, A., Verkest, D., Ronse, K. & Thean, A. V.-Y. Maintaining Moore's law - Enabling cost-friendly dimensional scaling. **9422** (Apr. 2015).
2. Doug O'Laughlin, S. *The Rising Tide of Semiconductor Cost* `https://semiwiki.com/semiconductor-services/308018-the-rising-tide-of-semiconductor-cost/`. 2022.
3. Mallik, A., Vandooren, A., Witters, L., Walke, A., Franco, J., Sherazi, S. M. Y., Weckx, P., Yakimets, D., Bardon, M., Parvais, B., Debacker, P., Ku, B., Lim, S., Mocuta, A., Mocuta, D., Ryckaert, J., Collaert, N. & Raghavan, P. *The impact of sequential-3D integration on semiconductor scaling roadmap* in (Dec. 2017), 32.1.1–31.1.4.
4. Iyer, S. S. Heterogeneous Integration for Performance and Scaling. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **6,** 973–982 (2016).
5. Zhang, J., Patil, N., Philip Wong, H.-S. & Mitra, S. *Overcoming carbon nanotube variations through co-optimized technology and circuit design* in *2011 International Electron Devices Meeting* (ieeexplore.ieee.org, Dec. 2011), 4.6.1–4.6.4.
6. Gupta, S. K. & Roy, K. Device-Circuit Co-Optimization for Robust Design of FinFET-Based SRAMs. *IEEE Des. Test Comput.* **30,** 29–39 (Dec. 2013).
7. Zhang, Z., Wang, R., Chen, C., Huang, Q., Wang, Y., Hu, C., Wu, D., Wang, J. & Huang, R. *New-Generation Design-Technology Co-Optimization (DTCO): Machine-Learning Assisted Modeling Framework* in *2019 Silicon Nanoelectronics Workshop (SNW)* (ieeexplore.ieee.org, June 2019), 1–2.
8. Wang, S., Pan, A., Chui, C. O. & Gupta, P. PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices. *IEEE Transactions on Very Large Scale Integration Systems* (2015).
9. Wang, W.-C. & Gupta, P. Efficient layout generation and design evaluation of vertical channel devices. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **35,** 1449–1460 (2015).

10. Ghaida, R. S. & Gupta, P. DRE: A framework for early co-evaluation of design rules, technology choices, and layout methodologies. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **31,** 1379–1392 (2012).

11. Ryckaert, J. *et al. Design Technology co-optimization for N10* in *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference* (ieeexplore.ieee.org, Sept. 2014), 1–8.

12. Yeric, G., Cline, B., Sinha, S., Pietromonaco, D., Chandra, V. & Aitken, R. *The past present and future of design-technology co-optimization* in *Proceedings of the IEEE 2013 Custom Integrated Circuits Conference* (ieeexplore.ieee.org, Sept. 2013), 1–8.

13. Capodieci, L., Gupta, P., Kahng, A. B., Sylvester, D. & Yang, J. *Toward a methodology for manufacturability-driven design rule exploration* in *Proceedings of the 41st annual Design Automation Conference* (2004), 311–316.

14. Kahng, A., Kahng, A. B., Lee, H. & Li, J. PROBE: A placement, routing, back-end-of-line measurement utility. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **37,** 1459–1472 (2017).

15. Collaert, N. *1.3 Future Scaling: Where Systems and Technology Meet* in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)* (ieeexplore.ieee.org, Feb. 2020), 25–29.

16. Samavedam, S. B., Ryckaert, J., Beyne, E., Ronse, K., Horiguchi, N., Tokei, Z., Radu, I., Bardon, M. G., Na, M. H., Spessot, A. & Biesemans, S. *Future Logic Scaling: Towards Atomic Channels and Deconstructed Chips* in *2020 IEEE International Electron Devices Meeting (IEDM)* (ieeexplore.ieee.org, Dec. 2020), 1.1.1–1.1.10.

17. imec. *Getting the most out of your system* `https://www.imec-int.com/en/articles/getting-most-out-your-system`. 2021.

18. Lenihan, T. G., Matthew, L. & Vardaman, E. J. *Developments in 2.5D: The role of silicon interposers* in *2013 IEEE 15th Electronics Packaging Technology Conference (EPTC 2013)* (2013), 53–55.

19. Chiou, D. *The microsoft catapult project* in *2017 IEEE International Symposium on Workload Characterization (IISWC)* (2017), 124–124.

20. Naffziger, S., Beck, N., Burd, T., Lepak, K., Loh, G. H., Subramony, M. & White, S. *Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product* in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)* (2021), 57–70.

21. Zhu, M., Zhuo, Y., Wang, C., Chen, W. & Xie, Y. *Performance evaluation and optimization of HBM-Enabled GPU for data-intensive applications* in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017* (2017), 1245–1248.

22. Elster, A. C. & Haugdahl, T. A. Nvidia Hopper GPU and Grace CPU Highlights. *Computing in Science Engineering* **24,** 95–100 (2022).

23. Lee, C.-C., Hung, C., Cheung, C., Yang, P.-F., Kao, C.-L., Chen, D.-L., Shih, M.-K., Chien, C.-L. C., Hsiao, Y.-H., Chen, L.-C., Su, M., Alfano, M., Siegel, J., Din, J. & Black, B. *An Overview of the Development of a GPU with Integrated HBM on Silicon Interposer* in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)* (2016), 1439–1444.

24. Huang, P. K., Lu, C. Y., Wei, W. H., Chiu, C., Ting, K. C., Hu, C., Tsai, C., Hou, S. Y., Chiou, W. C., Wang, C. T. & Yu, D. *Wafer Level System Integration of the Fifth Generation CoWoS®-S with High Performance Si Interposer at 2500 mm2* in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)* (2021), 101–104.

25. Hu, Y.-C., Liang, Y.-M., Hu, H.-P., Tan, C.-Y., Shen, C.-T., Lee, C.-H. & Hou, S. Y. *CoWoS Architecture Evolution for Next Generation HPC on 2.5D System in Package* in *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)* (2023), 1022–1026.

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

27. Duan, G., Kanaoka, Y., McRee, R., Nie, B. & Manepalli, R. *Die Embedding Challenges for EMIB Advanced Packaging Technology* in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)* (2021), 1–7.

28. Techpowerup.com. *NVIDIA H100 SXM5 96 GB* `https://www.techpowerup.com/gpu-specs/h100-sxm5-96-gb.c3974`. Online; accessed 18 February 2024.

29. Michigan State University. *2nd Generation AMD EPYC CPU* `https://docs.icer.msu.edu/Cluster_amd20_with_AMD_CPUs/`. Online; accessed 18 February 2024.

30. Server The Home. *3rd Gen EPYC With AMD 3D V Cache Delidded$_{T}$op* `https://www.servethehome.com/amd-milan-x-delivers-amd-epyc-caches-to-the-gb-era/3rd-gen-epyc-with-amd-3d-v-cache-delidded_top/`. Online; accessed 18 February 2024.

31. Wuu, J., Agarwal, R., Ciraula, M., Dietz, C., Johnson, B., Johnson, D., Schreiber, R., Swaminathan, R., Walker, W. & Naffziger, S. *3D V-Cache: the Implementation of a Hybrid-Bonded 64MB Stacked Cache for a 7nm x86-64 CPU* in *2022 IEEE International Solid- State Circuits Conference (ISSCC)* **65** (2022), 428–429.

32. Chang, J., Chen, Y.-H., Chan, G., Lin, K.-C., Wang, P.-S., Lin, Y., Chen, S., Lin, P., Wu, C.-W., Lin, C.-Y., Nien, Y.-H., Fujiwara, H., Katoch, A., Lee, R., Liao, H.-J., Liaw, J.-J., Wu, S.-Y. M. & Li, Q. *A 3nm 256Mb SRAM in FinFET Technology with New Array Banking Architecture and Write-Assist Circuitry Scheme for High-Density and Low-VMIN Applications* in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (2023), 1–2.

33. Liu, J., Mukhopadhyay, S., Kundu, A., Chen, S., Wang, H., Huang, D., Lee, J., Wang, M., Lu, R., Lin, S., Chen, Y., Shang, H., Wang, P., Lin, H., Yeap, G. & He, J. *A Reliability Enhanced 5nm CMOS Technology Featuring 5th Generation FinFET with Fully-Developed EUV and High Mobility Channel for Mobile SoC and High Performance Computing Application* in *2020 IEEE International Electron Devices Meeting (IEDM)* (2020), 9.2.1–9.2.4.

34. Semiconductor Engineering. *5nm Vs. 3nm* `https://semiengineering.com/5nm-vs-3nm/`. Online; accessed 14 September 2023.

35. Chia, H.-J., Tai, S.-P., Cui, J. J., Wang, C.-T., Tung, C.-H., Yee, K.-C. & Yu, D. C. *Ultra High Density Low Temperature SoIC with Sub-0.5 m Bond Pitch* in *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)* (2023), 1–4.

36. Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M. & Villalobos, P. *Compute Trends Across Three Eras of Machine Learning* in *2022 International Joint Conference on Neural Networks (IJCNN)* (2022), 1–8.

37. Shaw, D. E. *et al. Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch* in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (Association for Computing Machinery, St. Louis, Missouri, 2021). ISBN: 9781450384421. `https://doi.org/10.1145/3458817.3487397`.

38. AI and Memory Wall. *Medium.com.* `https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8` (2023).

39. Pal, S. & Gupta, P. *Pathfinding for 2.5D Interconnect Technologies* in *Proceedings of the Workshop on System-Level Interconnect: Problems and Pathfinding Workshop* (Association for Computing Machinery, San Diego, California, 2020). ISBN: 9781450381062. `https://doi.org/10.1145/3414622.3431906`.

40. Park, M.-J. *et al.* A 192-Gb 12-High 896-GB/s HBM3 DRAM With a TSV Auto-Calibration Scheme and Machine-Learning-Based Layout Optimization. *IEEE Journal of Solid-State Circuits* **58,** 256–269 (2023).

41. Park, S. J. *et al. Industry's First 7.2 Gbps 512GB DDR5 Module* in *2021 IEEE Hot Chips 33 Symposium (HCS)* (2021), 1–11.

42. Park, S. & Huddar, V. A. *Design and Analysis of Power Integrity of DDR5 Dual In-Line Memory Modules* in *2022 IEEE Electrical Design of Advanced Packaging and Systems (EDAPS)* (2022), 1–3.

43. AMD. *3D V-Cache™ Technology* https://www.amd.com/en/technologies/3d-v-cache. Online; accessed 28 August 2023.

44. Pal, S., Liu, J., Alam, I., Cebry, N., Suhail, H., Bu, S., Iyer, S. S., Pamarti, S., Kumar, R. & Gupta, P. *Designing a 2048-Chiplet, 14336-Core Waferscale Processor* in *2021 58th ACM/IEEE Design Automation Conference (DAC)* (2021), 1183–1188.

45. Lie, S. *Cerebras Architecture Deep Dive: First Look Inside the HW/SW Co-Design for Deep Learning : Cerebras Systems* in *2022 IEEE Hot Chips 34 Symposium (HCS)* (2022), 1–34.

46. Talpes, E., Williams, D. & Sarma, D. D. *DOJO: The Microarchitecture of Tesla's Exa-Scale Computer* in *2022 IEEE Hot Chips 34 Symposium (HCS)* (2022), 1–28.

47. Pal, S., Petrisko, D., Tomei, M., Gupta, P., Iyer, S. S. & Kumar, R. *Architecting Waferscale Processors - A GPU Case Study* in *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (2019), 250–263.

48. Rocki, K., Essendelft, D. V., Sharapov, I., Schreiber, R., Morrison, M., Kibardin, V., Portnoy, A., Dietiker, J. F., Syamlal, M. & James, M. *Fast Stencil-Code Computation on a Wafer-Scale Processor* in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (2020), 1–14.

49. Feng, Y. & Ma, K. Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration. *arXiv e-prints,* arXiv:2203.12268. arXiv: 2203.12268 (Mar. 2022).

50. Graening, A., Pal, S. & Gupta, P. *Chiplets: How Small is too Small?* in *Proc. ACM/IEEE Design Automation Conference (DAC)* (July 2023), 6.

51. Peng, H., Davidson, S., Shi, R., Song, S. L. & Taylor, M. *Chiplet Cloud: Building AI Supercomputers for Serving Large Generative Language Models* 2023. arXiv: 2307.02666 [cs.AR].

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

53. Lujan, A. P. *Comparison of Package-on-Package Technologies Utilizing Flip Chip and Fan-Out Wafer Level Packaging* in *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)* (2018), 2089–2094.

54. Shah, M., Kumar, R., Kim, C.-K., Aldrete, M., Syed, A., Gupta, P. & Lane, R. *Module/SiP Packaging Trends* in *2019 Electron Devices Technology and Manufacturing Conference (EDTM)* (2019), 82–84.

55. Octavo Systems. *SiP Technology* https://octavosystems.com/sip-technology/. Online; accessed 28 August 2023.

56. Choquette, J. & Gandhi, W. *NVIDIA A100 GPU: Performance amp; Innovation for GPU Computing* in *2020 IEEE Hot Chips 32 Symposium (HCS)* (IEEE Computer Society, Los Alamitos, CA, USA, Aug. 2020), 1–43. https://doi.ieeecomputersociety.org/10.1109/HCS49909.2020.9220622.

57. Macri, J. *AMD's next generation GPU and high bandwidth memory architecture: FURY* in *2015 IEEE Hot Chips 27 Symposium (HCS)* (2015), 1–26.

58. Sodani, A., Gramunt, R., Corbal, J., Kim, H.-S., Vinod, K., Chinthamani, S., Hutsell, S., Agarwal, R. & Liu, Y.-C. Knights Landing: Second-Generation Intel Xeon Phi Product. *IEEE Micro* **36,** 34–46 (2016).

59. Wade, M., Anderson, E., Ardalan, S., Bhargava, P., Buchbinder, S., Davenport, M. L., Fini, J., Lu, H., Li, C., Meade, R., *et al.* TeraPHY: a chiplet technology for low-power, high-bandwidth in-package optical I/O. *IEEE Micro* **40,** 63–71 (2020).

60. Ren, H., Sahoo, K., Xiang, T., Ouyang, G. & Iyer, S. S. *Demonstration of a Power-efficient and Cost-effective Power Delivery Architecture for Heterogeneously Integrated Wafer-scale Systems* in *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)* (ieeexplore.ieee.org, May 2023), 1614–1621.

61. Desai, N., Then, H. W., Yu, J., Krishnamurthy, H. K., Lambert, W. J., Butzen, N., Weng, S., Schaef, C., Radhakrishnan, K., Ravichandran, K., Tschanz, J. W. & De, V. A 32-A, 5-V-Input, 94.2% Peak Efficiency High-Frequency Power Converter Module Featuring Package-Integrated Low-Voltage GaN nMOS Power Transistors. *IEEE J. Solid-State Circuits* **57,** 1090–1099 (Apr. 2022).

62. Radhakrishnan, K., Swaminathan, M. & Bhattacharyya, B. K. Power Delivery for High-Performance Microprocessors—Challenges, Solutions, and Future Trends. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **11,** 655–671 (2021).

63. Hagge, J. State-of-the-art multichip modules for avionics. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology* **15,** 29–42 (1992).

64. Rinne, R. & Barbour, D. Multi-Chip Module Technology. *ElectroComponent Science and Technology* **10,** 31–49 (1982).

65. Sun, P., Leung, V., Yang, D., Lou, R., Shi, D. & Chung, T. *Development of a new Package-on-Package (PoP) structure for next-generation portable electronics* in *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)* (2010), 1957–1963.

66. Fontanelli, A. *System-in-Package Technology: Opportunities and Challenges* in *9th International Symposium on Quality Electronic Design (isqed 2008)* (2008), 589–593.

67. Jeng, S.-P. & Liu, M. *Heterogeneous and Chiplet Integration Using Organic Interposer (CoWoS-R)* in *2022 International Electron Devices Meeting (IEDM)* (2022), 3.2.1–3.2.4.

68. Roth, A., Zhou, C., Wong, M., Soenen, E., Huang, T.-C., Ranucci, P., Hsu, Y.-C., Lin, H.-C., Kuo, C., Wang, M.-J., Yang, S.-Y., Chu, J. R., Yeh, T.-Y., Ting, K. C., Loke, A. L. S., Rusu, S., Chen, M., Lee, F. Y. H., Zhang, K. & Kalnitsky, A. *Heterogeneous Power Delivery for 7nm High-Performance Chiplet-Based Processors using Integrated Passive Device and In-Package Voltage Regulator* in *2020 IEEE Symposium on VLSI Technology* (ieeexplore.ieee.org, June 2020), 1–2.

69. Moore, S. K. 3 paths to 3D processors. *IEEE Spectrum* **59,** 24–29 (June 2022).

70. Sisto, G., Chehab, B., Genneret, B., Baert, R., Chen, R., Weckx, P., Ryckaert, J., Chou, R., van Der Plas, G., Beyne, E. & Milojevic, D. *IR-Drop Analysis of Hybrid Bonded 3D-ICs with Backside Power Delivery and -n- TSVs* in *2021 IEEE International Interconnect Technology Conference (IITC)* (2021), 1–3.

71. Veloso, A. *et al.* Scaled FinFETs Connected by Using Both Wafer Sides for Routing via Buried Power Rails. *IEEE Transactions on Electron Devices* **69,** 7173–7179 (2022).

72. Cline, B., Prasad, D., Beyne, E. & Zografos, O. Power from Below: Buried Interconnects Will Help Save Moore's Law. *IEEE Spectrum* **58,** 46–51 (2021).

73. Kobrinsky, M. *et al. Novel Cell Architectures with Back-side Transistor Contacts for Scaling and Performance* in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (2023), 1–2.

74. Hafez, W. *et al. Intel PowerVia Technology: Backside Power Delivery for High Density and High-Performance Computing* in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (2023), 1–2.

75. Chen, R., Sisto, G., Stucchi, M., Jourdain, A., Miyaguchi, K., Schuddinck, P., Woeltgens, P., Lin, H., Kakarla, N., Veloso, A., Milojevic, D., Zografos, O., Weckx, P., Hellings, G., Van Der Plas, G., Ryckaert, J. & Beyne, E. *Backside PDN and 2.5D MIMCAP to Double Boost 2D and 3D ICs IR-Drop beyond 2nm Node* in *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* (2022), 429–430.

76. Siricharoenpanich, A., Wiriyasart, S., Srichat, A. & Naphon, P. Thermal management system of CPU cooling with a novel short heat pipe cooling system. *Case Studies in Thermal Engineering* **15,** 100545. ISSN: 2214-157X. https://www.sciencedirect.com/science/article/pii/S2214157X19303703 (2019).

77. Pambudi, N. A., Sarifudin, A., Firdaus, R. A., Ulfa, D. K., Gandidi, I. M. & Romadhon, R. The immersion cooling technology: Current and future development in energy saving. *Alexandria Engineering Journal* **61,** 9509–9527. ISSN: 1110-0168. https://www.sciencedirect.com/science/article/pii/S1110016822001557 (2022).

78. Elliott, J., Lebon, M. & Robinson, A. Optimising integrated heat spreaders with distributed heat transfer coefficients: A case study for CPU cooling. *Case Studies in Thermal Engineering* **38,** 102354. ISSN: 2214-157X. https://www.sciencedirect.com/science/article/pii/S2214157X22005949 (2022).

79. SemiEngineering. *Thermal Management Implications For Heterogeneous Integrated Packaging* https://semiengineering.com/thermal-management-implications-for-heterogeneous-integrated-packaging/. 2022.

80. Zhou, M., Li, L., Hou, F., He, G. & Fan, J. Thermal Modeling of a Chiplet-Based Packaging With a 2.5-D Through-Silicon Via Interposer. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **12,** 956–963 (2022).

81. Lin, S.-C. & Banerjee, K. Cool Chips: Opportunities and Implications for Power and Thermal Management. *IEEE Transactions on Electron Devices* **55,** 245–255 (2008).

82. Eris, F., Joshi, A., Kahng, A. B., Ma, Y., Mojumder, S. & Zhang, T. *Leveraging thermally-aware chiplet organization in 2.5D systems to reclaim dark silicon* in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)* (2018), 1441–1446.

83. Chen, X., Fu, Y., Feng, J., Zhang, J., Chen, S. & Xu, J. Improving the thermal reliability of photonic chiplets on multicore processors. *Integration* **86,** 9–21. ISSN: 0167-9260. https://www.sciencedirect.com/science/article/pii/S0167926022000372 (2022).

84. Luo, G., Shi, Y. & Cong, J. An Analytical Placement Framework for 3-D ICs and Its Extension on Thermal Awareness. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **32,** 510–523 (2013).

85. Goplen, B. & Sapatnekar, S. *Efficient thermal placement of standard cells in 3D ICs using a force directed approach* in *ICCAD-2003. International Conference on Computer Aided Design (IEEE Cat. No.03CH37486)* (2003), 86–89.

86. Ma, Y., Delshadtehrani, L., Demirkiran, C., Abellan, J. L. & Joshi, A. *TAP-2.5D: A Thermally-Aware Chiplet Placement Methodology for 2.5D Systems* in *2021 Design, Automation Test in Europe Conference Exhibition (DATE)* (2021), 1246–1251.

87. Zhang, Y., King, C. R., Zaveri, J., Kim, Y. J., Sahu, V., Joshi, Y. & Bakir, M. S. *Coupled electrical and thermal 3D IC centric microfluidic heat sink design and technology* in *2011 IEEE 61st Electronic Components and Technology Conference (ECTC)* (2011), 2037–2044.

88. Lau, J. H. & Yue, T. G. *Thermal management of 3D IC integration with TSV (through silicon via)* in *2009 59th Electronic Components and Technology Conference* (2009), 635–640.

89. Shi, B., Srivastava, A. & Bar-Cohen, A. *Hybrid 3D-IC Cooling System Using Micro-fluidic Cooling and Thermal TSVs* in *2012 IEEE Computer Society Annual Symposium on VLSI* (2012), 33–38.

90. Zhang, Y., Dembla, A. & Bakir, M. S. Silicon Micropin-Fin Heat Sink With Integrated TSVs for 3-D ICs: Tradeoff Analysis and Experimental Testing. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **3,** 1842–1850 (2013).

91. Zhang, Y., Zhang, Y. & Bakir, M. S. Thermal Design and Constraints for Heterogeneous Integrated Chip Stacks and Isolation Technology Using Air Gap and Thermal Bridge. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **4,** 1914–1924 (2014).

92. Graphcore Uses TSMC 3D Chip Tech to Speed AI by 40%. *IEEE Spectrum.* `https://spectrum.ieee.org/graphcore-ai-processor` (2022).

93. Fu, Y., Bolotin, E., Chatterjee, N., Nellans, D. & Keckler, S. GPU Domain Specialization via Composable On-Package Architecture. *ACM Transactions on Architecture and Code Optimization* **19,** 1–23 (Mar. 2022).

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

95. Jangam, S., Pal, S., Bajwa, A., Pamarti, S., Gupta, P. & Iyer, S. S. *Latency, bandwidth and power benefits of the superchips integration scheme* in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)* (2017), 86–94.

96. Jangam, S. & Iyer, S. S. A signaling figure of merit (s-FoM) for advanced packaging. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **10,** 1758–1761 (2020).

97. Stow, D., Xie, Y., Siddiqua, T. & Loh, G. H. *Cost-effective design of scalable high-performance systems using active and passive interposers* in *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (2017), 728–735.

98. Kim, J., Murali, G., Park, H., Qin, E., Kwon, H., Chekuri, V. C. K., Rahman, N. M., Dasari, N., Singh, A., Lee, M., *et al.* Architecture, chip, and package codesign flow for interposer-based 2.5-D chiplet integration enabling heterogeneous IP reuse. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **28,** 2424–2437 (2020).

99. Stow, D., Akgun, I. & Xie, Y. *Investigation of cost-optimal network-on-chip for passive and active interposer systems* in *2019 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)* (2019), 1–8.

100. Zhu, L., Jo, C. & Lim, S. K. Power Delivery Solutions and PPA Impacts in Micro-Bump and Hybrid-Bonding 3D ICs. *IEEE Trans. Compon. Packaging Manuf. Technol.* **12,** 1969–1982 (Dec. 2022).

101. Lanzillo, N. A., Chu, A. M., Perez, N., Vega, R., Clevenger, L. & Dechene, D. Benchmarking Power Delivery Network Designs at the 5-nm Technology Node. *IEEE Trans. Electron Devices* **69,** 7135–7140 (Dec. 2022).

102. Choi, S., Jung, J., Kahng, A. B., Kim, M., Park, C.-H., Pramanik, B. & Yoon, D. PROBE3.0: A Systematic Framework for Design-Technology Pathfinding with Improved Design Enablement. arXiv: `2304.13215 [cs.AR]` (Apr. 2023).

103. Sisto, G., Chen, R., Milojevic, D., Zografos, O., Weckx, P., Hellings, G. & Ryckaert, J. *System-level evaluation of 3D power delivery network at 2nm node* en. in *DTCO and Computational Patterning II* **12495** (SPIE, Apr. 2023), 207–217.

104. Abdi, D. B., Salahuddin, S. M., Boemmels, J., Giacomin, E., Weckx, P., Ryckaert, J., Hellings, G. & Catthoor, F. 3D SRAM Macro Design in 3D Nanofabric Process Technology. *IEEE Trans. Circuits Syst. I Regul. Pap.* **70,** 2858–2867 (July 2023).

105. Liebmann, L., Smith, J., Chanemougame, D. & Gutwin, P. *CFET Design Options, Challenges, and Opportunities for 3D Integration* in *2021 IEEE International Electron Devices Meeting (IEDM)* (ieeexplore.ieee.org, Dec. 2021), 3.1.1–3.1.4.

106. Agnesina, A., Brunion, M., Kim, J., Garcia-Ortiz, A., Milojevic, D., Catthoor, F., Mirabelli, G., Komalan, M. & Lim, S. K. Power, Performance, Area, and Cost Analysis of Face-to-Face-Bonded 3-D ICs. *IEEE Trans. Compon. Packaging Manuf. Technol.* **13,** 300–314 (Mar. 2023).

107. Liebmann, L., Zeng, J., Zhu, X., Yuan, L., Bouche, G. & Kye, J. *Overcoming scaling barriers through design technology cooptimization* in *2016 IEEE Symposium on VLSI Technology* (2016), 1–2.

108. Ryckaert, J., Raghavan, P., Schuddinck, P., Trong, H. B., Mallik, A., Sakhare, S. S., Chava, B., Sherazi, Y., Leray, P., Mercha, A., *et al. DTCO at N7 and beyond: patterning and electrical compromises and opportunities* in *Design-Process-Technology Co-optimization for Manufacturability IX* **9427** (2015), 101–108.

109. Kagalwalla, A. A., Lam, M., Adam, K. & Gupta, P. *EUV-CDA: Pattern shift aware critical density analysis for EUV mask layouts* in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)* (2014), 155–160.

110. Ku, B. W., Chang, K. & Lim, S. K. *Compact-2D: A physical design methodology to build commercial-quality face-to-face-bonded 3D ICs* in *Proceedings of the 2018 International Symposium on Physical Design* (2018), 90–97.

111. Agnesina, A., Brunion, M., Kim, J., Garcia-Ortiz, A., Milojevic, D., Catthoor, F., Mirabelli, G., Komalan, M. & Lim, S. K. Power, Performance, Area, and Cost Analysis of Face-to-Face-Bonded 3-D ICs. *IEEE Transactions on Components, Packaging and Manufacturing Technology* **13,** 300–314 (2023).

112. Priyadarshi, S., Hu, J., Choi, W. H., Melamed, S., Chen, X., Davis, W. R. & Franzon, P. D. *Pathfinder 3D: A flow for system-level design space exploration* in *2011 IEEE International 3D Systems Integration Conference (3DIC), 2011 IEEE International* (2012), 1–8.

113. Ardalani, N., Pal, S. & Gupta, P. DeepFlow: A Cross-Stack Pathfinding Framework for Distributed AI Systems. *ACM Trans. Des. Autom. Electron. Syst.* ISSN: 1084-4309. `https://doi.org/10.1145/3635867` (Dec. 2023).

114. Rashidi, S., Sridharan, S., Srinivasan, S. & Krishna, T. *ASTRA-SIM: Enabling SW/HW Co-Design Exploration for Distributed DL Training Platforms* in *2020 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (Aug. 2020), 81–92.

115. Parashar, A., Raina, P., Shao, Y. S., Chen, Y.-H., Ying, V. A., Mukkara, A., Venkatesan, R., Khailany, B., Keckler, S. W. & Emer, J. *Timeloop: A Systematic Approach to DNN Accelerator Evaluation* in *2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (Mar. 2019), 304–315.

116. Zou, Q., Chen, Y., Xie, Y. & Su, A. *System-level design space exploration for three-dimensional (3D) SoCs* in *2011 Proceedings of the Ninth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)* (2011), 385–388.

117. Agrawal, P., Milojevic, D., Raghavan, P., Catthoor, F., Van der Perre, L., Beyne, E. & Varadarajan, R. System Level Comparison of 3D Integration Technologies for Future Mobile MPSoC Platform. *IEEE Embedded Systems Letters* **6,** 85–88 (2014).

118. Siozios, K., Papanikolaou, A. & Soudris, D. A method and tool for early design/technology search-space exploration for 3D ICs (Jan. 2008).

119. Chen, S., Li, S., Zhuang, Z., Zheng, S., Liang, Z., Ho, T.-Y., Yu, B. & Sangiovanni-Vincentelli, A. L. *Floorplet: Performance-aware Floorplan Framework for Chiplet Integration* 2023. arXiv: `2308.01672 [cs.AR]`.

120. Biggs, J., Myers, J., Kufel, J., Ozer, E., Craske, S., Sou, A., Ramsdale, C., Williamson, K., Price, R. & White, S. A natively flexible 32-bit Arm microprocessor. *Nature* **595,** 532–536 (2021).

121. Bleier, N., Lee, C., Rodriguez, F., Sou, A., White, S. & Kumar, R. *FlexiCores: low footprint, high yield, field reprogrammable flexible microprocessors* in *Proceedings of the 49th Annual International Symposium on Computer Architecture* (2022), 831–846.

122. Jahanshahi, A., Salvo, P. & Vanfleteren, J. *Stretchable biocompatible electronics by embedding electrical circuitry in biocompatible elastomers* in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2012), 6007–6010.

123. Wang, M., Luo, Y., Wang, T., Wan, C., Pan, L., Pan, S., He, K., Neo, A. & Chen, X. Artificial skin perception. *Advanced Materials* **33,** 2003014 (2021).

124. Lindsay, M., Bishop, K., Sengupta, S., Co, M., Cumbie, M., Chen, C.-H. & Johnston, M. L. Heterogeneous integration of CMOS sensors and fluidic networks using wafer-level molding. *IEEE transactions on biomedical circuits and systems* **12,** 1046–1055 (2018).

125. Xie, L., Yang, G., Mantysalo, M., Xu, L.-L., Jonsson, F. & Zheng, L.-R. Heterogeneous integration of bio-sensing system-on-chip and printed electronics. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* **2,** 672–682 (2012).

126. Haruta, T., Nakajima, T., Hashizume, J., Umebayashi, T., Takahashi, H., Taniguchi, K., Kuroda, M., Sumihiro, H., Enoki, K., Yamasaki, T., *et al. 4.6 A 1/2.3 inch 20Mpixel 3-layer stacked CMOS Image Sensor with DRAM* in *2017 IEEE International Solid-State Circuits Conference (ISSCC)* (2017), 76–77.

127. Liu, C., Ren, P., Sun, Y., Gao, D., Luo, W., Chen, Z. & Xia, Y. *Reliability Challenges in Advanced Technology Node: from Transistor to Circuit (invited)* in *2020 IEEE 15th International Conference on Solid-State Integrated Circuit Technology (ICSICT)* (2020), 1–4.

128. Sham, M.-L., Gao, Z., Leung, L. L.-W., Chen, Y.-C. & Chung, T. *Advanced Packaging Technologies for Automotive Electronics* in *2007 8th International Conference on Electronic Packaging Technology* (ieeexplore.ieee.org, Aug. 2007), 1–5.

129. Iyer, S. S. & Bajwa, A. A. *Reliability challenges in advance packaging* in *2018 IEEE International Reliability Physics Symposium (IRPS)* (ieeexplore.ieee.org, Mar. 2018), 4D.5–1–4D.5–4.

130. Chase, N. S., Irwin, R., Yang, Y. T., Ren, H. & Iyer, S. S. *Reliability Considerations for Wafer Scale Systems* in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)* (ieeexplore.ieee.org, June 2021), 84–89.

131. Yip, L., Lin, R., Lai, C. & Peng, C. *Reliability Challenges of High-Density Fan-out Packaging for High-Performance Computing Applications* in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)* (ieeexplore.ieee.org, May 2022), 1454–1458.

132. Xie, Y., Bao, C. & Srivastava, A. 3D/2.5 D IC-based obfuscation. *Hardware protection through obfuscation,* 291–314 (2017).

133. Gu, P., Li, S., Stow, D., Barnes, R., Liu, L., Xie, Y. & Kursun, E. *Leveraging 3D technologies for hardware security: Opportunities and challenges* in *Proceedings of the 26th edition on Great Lakes Symposium on VLSI* (2016), 347–352.

134. Imeson, F., Emtenan, A., Garg, S. & Tripunitara, M. *Securing Computer Hardware Using 3D Integrated Circuit ({{{{{IC}}}}}) Technology and Split Manufacturing for Obfuscation* in *22nd USENIX Security Symposium (USENIX Security 13)* (2013), 495–510.

135. Nabeel, M., Ashraf, M., Patnaik, S., Soteriou, V., Sinanoglu, O. & Knechtel, J. 2.5 D root of trust: Secure system-level integration of untrusted chiplets. *IEEE Transactions on Computers* **69,** 1611–1625 (2020).

136. Xie, Y., Bao, C., Serafy, C., Lu, T., Srivastava, A. & Tehranipoor, M. Security and vulnerability implications of 3D ICs. *IEEE Transactions on Multi-Scale Computing Systems* **2,** 108–122 (2016).

137. Knechtel, J. & Sinanoglu, O. *On mitigation of side-channel attacks in 3D ICs: Decorrelating thermal patterns from power and activity* in *Proceedings of the 54th Annual Design Automation Conference 2017* (2017), 1–6.

138. Dofe, J., Gu, P., Stow, D., Yu, Q., Kursun, E. & Xie, Y. *Security threats and countermeasures in three-dimensional integrated circuits* in *Proceedings of the on Great Lakes Symposium on VLSI 2017* (2017), 321–326.

139. Wang, Y.-C., Chen, T.-C. T. & Wang, L.-C. Simulating a semiconductor packaging facility: Sustainable strategies and short-time evidences. *Procedia Manufacturing* **11,** 787–795 (2017).

140. Harland, J., Reichelt, T. & Yao, M. *Environmental sustainability in the semiconductor industry* in *2008 IEEE International Symposium on Electronics and the Environment* (2008), 1–6.

141. Kerbusch, J. *Why the electronics industry must address sustainability* in *EASS 2022; 11th GMM-Symposium* (2022), 1–3.

142. Gandhi, A., Ghose, K., Gopalan, K., Hussain, S., Lee, D., Liu, Y., Liu, Z., McDaniel, P., Mu, S. & Zadok, E. *Metrics for sustainability in data centers* in *Proceedings of the 1st Workshop on Sustainable Computer Systems Design and Implementation (HotCarbon'22)* (2022).

143. EECKHOUT, L. Towards sustainable computer architecture: A holistic approach. *HiPEAC Vision 2023,* 216 (2023).

144. Bardon, M. G., Wuytens, P., Ragnarsson, L.-Å., Mirabelli, G., Jang, D., Willems, G., Mallik, A., Spessot, A., Ryckaert, J. & Parvais, B. *DTCO including sustainability: Power-performance-area-cost-environmental score (PPACE) analysis for logic technologies* in *2020 IEEE International Electron Devices Meeting (IEDM)* (2020), 41–4.

6 Highlighted References

1. Iyer, S. S. Heterogeneous Integration for Performance and Scaling. IEEE Transactions on Components, Packaging and Manufacturing Technology 6, 973–982 (2016).

   This paper makes a case that packaging dimensions have scaled much more poorly so far than within-chip dimensions but heterogeneous integration will be the backbone of sustaining Moore's law in the years ahead.

2. Naffziger, S., Beck, N., Burd, T., Lepak, K., Loh, G. H., Subramony, M. White, S. Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families : Industrial Product in 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA) (2021), 57–70.

   This paper details the technology challenges that motivated AMD to use chiplets in their product families.

3. Jangam, S. Iyer, S. S. A signaling figure of merit (s-FoM) for advanced packaging. IEEE Transactions on Components, Packaging and Manufacturing Technology 10, 1758–1761 (2020).

   This paper proposes a simple signaling figure of merit for inter-chiplet interconnect and compares various link+packaging technologies using it: an example of link-level STCO.

4. Zhu, L., Jo, C. Lim, S. K. Power Delivery Solutions and PPA Impacts in Micro-Bump and Hybrid-Bonding 3D ICs. IEEE Trans. Compon. Packaging Manuf. Technol. 12, 1969–1982 (Dec. 2022).

   Agnesina, A., Brunion, M., Kim, J., Garcia-Ortiz, A., Milojevic, D., Catthoor, F., Mirabelli, G., Komalan, M. Lim, S. K. Power, Performance, Area, and Cost Analysis of Face-to-Face-Bonded 3-D ICs. IEEE Trans. Compon. Packaging Manuf. Technol. 13, 300–314 (Mar. 2023).

Both these papers are examples of component-level STCO approaches which evaluate technologies using physical implementation of benchmark designs.

5. Ardalani, N., Pal, S. Gupta, P. DeepFlow: A Cross-Stack Pathfinding Framework for Distributed AI Systems. ACM Trans. Des. Autom. Electron. Syst. issn: 1084-4309. https://doi.org/10.1145/3635867 (Dec. 2023).

   This is one of the earliest attempts at developing an algorithms to technology cross-stack STCO framework in context of distributed training of large neural networks.

7 Acknowledgements

Puneet Gupta discloses support for the research of this work from DARPA/SRC CHIMES JUMP 2.0 Center and National Science Foundation.

8 Competing Interests

The authors have no competing interests.

9 Key Points

- Connectivity, scale, cost, and form-factor are the main drivers for use of advanced integration techniques in emerging computing systems.
- Support for technology heterogeneity, advanced power delivery, and cooling within the advanced packaging are key system enablers.
- System-technology co-optimization approaches can be classified as link-level, component-level or cross-stack system-level. Automated, fast cross-stack STCO frameworks are still in their infancy but are essential to guide high-value technology development.

10 Short Summary

Advanced packaging is emerging as the way to provide system scaling for next-generation computing. System-technology co-optimization across the technology-hardware-software stack is key to guiding expensive research and development efforts and enabling future heterogeneously integrated computing systems.