The Next Era for Chiplet Innovation

Gabriel H. Loh, Raja Swaminathan
Advanced Micro Devices, Inc.
gabriel.loh@amd.com

***Abstract*— Moore's Law is slowing down and the associated costs are simultaneously increasing. These pressures have given rise to new approaches utilizing advanced packaging and integration such as chiplets, interposers, and 3D stacking. We first describe the key technology drivers and constraints that motivate chiplet-based architectures, exploring several product case studies to highlight how different chiplet strategies have been developed to address different design objectives. We detail multiple generations of chiplet-based CPU architectures as well as the recent addition of 3D stacking options to further enhance processor capabilities. Across the industry, we are still collectively in the relatively early days of advanced packaging and 3D integration. As silicon scaling only gets more challenging and expensive while demand for computation continues to soar, we anticipate the transition to a new generation of chiplet architectures that utilize increasing combinations of 2D, 2.5D, and 3D integration and packaging technologies to continue to deliver compelling SoC solutions. However, this next era for chiplet innovation will face a variety of challenges. We will explore many of these technical topics, which in turn provide rich research opportunities for the community to explore and innovate.**

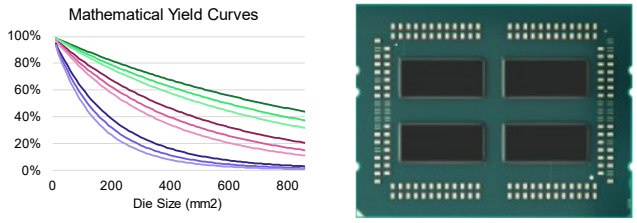
Keywords—chiplets, integration, stacking

I. Introduction

Over the past decade, the semiconductor industry has seen an increasing adoption of advanced packaging and die stacking technologies in a variety of commercial products. Different designs targeting various market requirements have utilized a few of these techniques to solve specific problems or deal with particular constraints. For some products, multi-chip module (MCM) and chiplet designs have been utilized to deal with rising silicon costs and to integrate more logic per package. Other products have utilized die-stacking in different ways to address bandwidth limitations or to increase integration density.

We believe that technology trends such as the slowing of Moore's Law, the increasing costs of silicon, demands on memory bandwidth, solution density and optimizing for total cost of ownership (TCO), and other factors may force future silicon designs to not just adopt these technologies, but aggressively deploy multiple of these technologies at once within the same design. This leads to the potential for exciting new architectures that simultaneously combine different uses of 2D, 2.5D, and 3D technologies.

In this paper, we will start with an overview of several of the advanced packaging and stacking technologies. We will explain what they are, their pros and cons, what applications are most suitable, and provide some examples of commercial use cases for each. We then discuss expected future trends and their implications on the design of semiconductor systems regarding advanced packaging and stacking. In particular, we will walk the reader through some of the interesting and challenging research problems that these future systems will face, which should provide excellent topics for the broader research community to work on.



(a) (b)

Figure 1. (a) Simulated/hypothetical die-yield curves for illustrative purposes, (b) an example four-die MCM.

II. Current Approaches

This section does not provide an exhaustive list of all packaging and stacking options across the entire industry, as there are many options out there especially if one includes technologies used in embedded, industrial, and other customized use cases. We limit the focus here on some of the technologies that have been utilized in high-volume commercial offerings for server and consumer products.

A. 2D MCM

It is well known that semiconductor manufacturing yields decrease in a non-linear fashion as a function of the die size. Intuitively, a larger chip discarded due to a manufacturing defect wastes a larger fraction of the overall silicon wafer, thereby forcing that cost to be amortized over the remaining functional or yielded chip from that wafer. Figure 1(a) shows hypothetical yield curves using the textbook yield equation

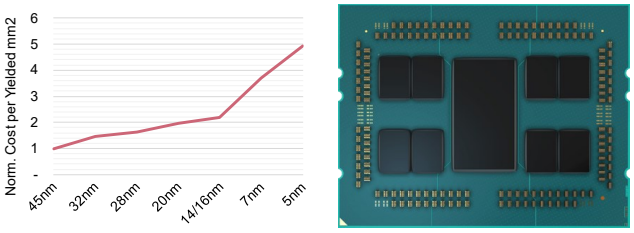
$$\text{Die Yield} = \frac{1}{(1+DefectDensity \times DieArea)^N}$$

using a range of example/arbitrary values for the equation parameters. This is only meant to provide a visualization of the non-linear relationship between die size and yield, and the curves do not relate to any specific process node nor foundry. The example curves also do not account for any yield recovery due to repair and/or redundancy at the design level. The main take away is that, especially for increasingly complex silicon process technology nodes, very large die sizes can become exceedingly expensive as the raw yield rates drop.

The idea and even the commercial usage of multi-chip module (MCM) technology has been around for decades [2][5][10]. MCM take the functionality of what is logically one large die or system on a chip (SoC), and then partitions the design into multiple smaller chips. Due to the non-linear relationship between die size and yield, reintegrating multiple smaller chips can be far more cost effective than constructing a single monolithic SoC.

Figure 1(b) shows a first-generation AMD EPYC™ CPU MCM, where a 32-core CPU has been partitioned into eight smaller 8-core die [1]. Past work has reported that this organization resulted in a ~40% reduction in cost compared to

978-3-9819263-7-8/DATE23/© 2023 EDAA



(a) (b)

Figure 2. (a) Relative cost of a 250mm^2 die across technology nodes, (b) an example chiplet-based processor.

a hypothetical monolithic implementation [8]. One of the design tradeoffs for MCMs is that communication between logical components in different die must now cross a die-to-die communication link across the package substrate. Compared to on-chip metal routing resources, the bandwidth, latency, and power to send data between the chips of an MCM is worse. For SoC that have logical blocks that can be cleanly partitioned in hardware and/or managed by software, the performance impact from the reduced bandwidth of the inter-die links can be kept under control so that the cost benefits of an MCM design can be taken advantage of.

B. 2D Chiplets

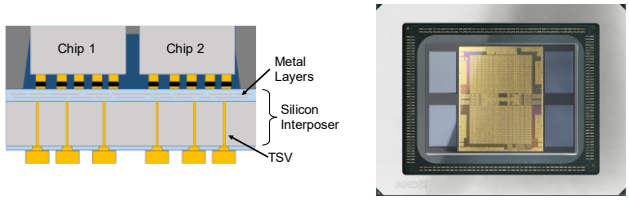
Over the past several years, the cost of silicon has experienced a trend where newer technology nodes are becoming increasingly more expensive even for chips with the same die area. Figure 2(a) shows the cost of a 250mm^2 die normalized to the 45nm node across a range of technologies. While "classic" MCMs partitioned an SoC in multiple smaller and more cost-effective components, AMD's chiplet approach takes things further and implements different die in different process technologies to better match the requirements and/or constraints for each chiplet.

Figure 2(b) shows a second-generation AMD EPYC™ CPU with mixed chiplets. The eight smaller chiplets each implement eight CPU cores in a 7nm technology node. The larger chip in the middle is the "IO Die" that houses memory controllers, IO interfaces, and other system components. Many of these blocks, especially the IO interfaces, do not scale much or at all with improvements in technology nodes. For example, the size of some of these blocks are determined by the area required for the external IO connection. As such, the IOD in this design was implemented in an older and more cost-effective 12nm node [8]. Similar to MCMs, the communications between chiplets may be limited by the substrate-level routing, and so architectural design to effectively partition an SoC into chiplets is an important part of the design process.

C. 2.5 Silicon Interposer

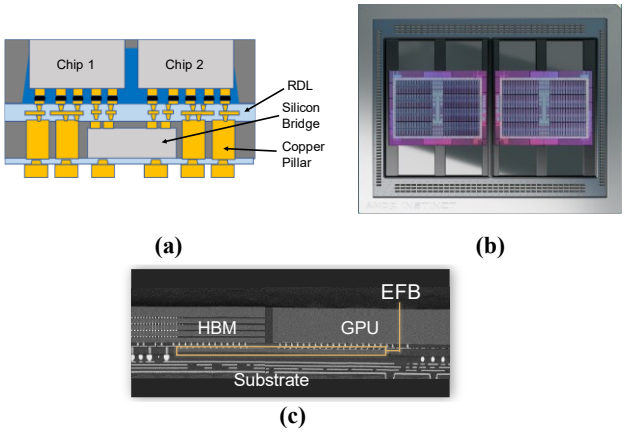
The die-to-die communication links across the package substrate in MCM and chiplet designs is typically limited to around a few 10s of GB/s [1][8]. The primary constraint is the width/density of the metal routing that can be supported in typical organic substrate implementations. For some applications, such as integrating memory directly into the package, 100s of GB/s of bandwidth is required.

A silicon interposer is effectively a "chip" with the purpose of providing interconnections between multiple other chips [7]. Figure 3(a) shows a cross-sectional view of a silicon interposer with two chips stacked on top. This is often referred to as "2.5D" stacking because while the chips are 3D-stacked on top of the interposer, the individual chips are still in a 2D



(a) (b)

Figure 3. (a) Cross-section schematic of two chips stacked on a passive silicon interposer, (b) an example processor die with four memory modules all on a silicon interposer.



(a) (b)

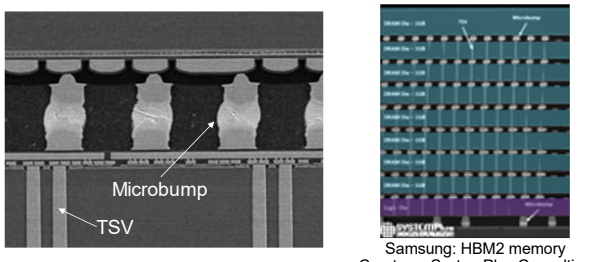
(c)

Figure 4. (a) Cross-section schematic of two chips connected via a silicon bridge, (b) using silicon bridges to connect eight memory modules to two processor dies, and (c) a cross-sectional view of one of the bridge assemblies.

organization relative to each other. The silicon interposer uses conventional back-end-of-the-line processes to construct its metal routing layers, and therefore can provide interconnect densities similar to any other chip. If Chip1 in the figure is a memory device and Chip2 is a compute die, then the metal layers on the silicon interposer can provide thousands of parallel routes in a relatively small area, thereby supporting the 100s of GB/s of bandwidth required by high-performance memory interfaces. The silicon interposer still makes use of through-silicon vias (TSV) to provide IO, power, and ground connections from the individual chips to the outside of the package. Figure 3(b) illustrates an AMD Instinct™ MI100 accelerator, that combines a GPU-based compute die (the larger central chip) with four in-packaged DRAM modules, all stacked on and interconnected with a silicon interposer, supporting a peak theoretical bandwidth of 1.2 TB/s.

D. 2.5 Silicon Bridges

Depending on the specific requirements of a given SoC design, silicon interposers may come with limitations or tradeoffs that make other approaches more desirable. One set of challenges stem from the fact that the size of the interposer must be large enough to accommodate all of the chips that are to be 2.5D stacked on top of it. For a system with a large amount of active silicon to be integrated, this can result in a large interposer. While typical use cases (e.g., memory integration) utilize passive interposers (i.e., without logic devices/transistors), that have very good yield rates, a very large interposer still adds cost to the system. Furthermore, if the total number of chips require an interposer size that exceeds the reticle limit (typically between 800-900mm^2), then additional costs have to be incurred to support reticle stitching techniques to construct larger interposers [4].



(a) **(b)**

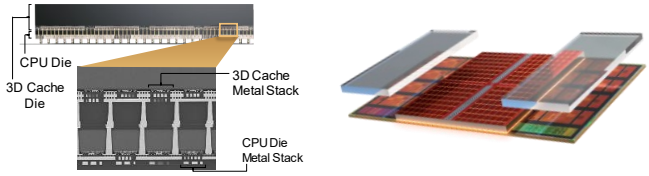
Figure 5. (a) Cross-section micrograph of microbumps connecting two chips, (b) 3D memory stack constructed with microbump 3D stacking.

Silicon bridge technologies have been developed as an alternative packaging solution to provide silicon-levels of wire density while using much smaller pieces of silicon. Figure 4(a) shows a cross-sectional view of AMD's elevated fanout bridge (EFB) technology. The silicon bridge is a small passive chip that presents an electrical interface to the chips above that is very similar to that of the silicon interposer. However, the bridge is much smaller, only needing to be large enough to cover the die-to-die connection interfaces of the two chips that the bridge connects together. Outside of the region occupied by the bridge, conventional copper pillar technology can be used to provide IO, power, and ground signals directly to the chip. The silicon bridges are more cost-effective compared to the silicon interposer due to their smaller size and avoiding the need for additional manufacturing steps to construct TSVs.

Figure 4(b) shows the AMD Instinct™ MI200 accelerator, which consists of two GPU compute dies (the two larger chips along the central lateral axis of the package) and eight in-package memory modules (four at the top, four at the bottom). Each memory module is connected to a GPU compute die via a silicon bridge (EFB), illustrated in the cross-section micrograph shown in Figure 4(c). This also shows how the bridge is above the package substrate (hence the "elevated" nomenclature in EFB).

E. 3D Stacking: Microbumps

The technologies discussed thus far have all been used to connect multiple chiplets together where all of the active components (i.e., not counting the passive interposer or bridges) are placed next to each other. 3D stacking can further increase integration density and die-to-die bandwidth by directly placing one or more active chips on top of each other. Microbumps are effectively very small solder connections (with varying metallurgy depending on the specific implementation). Figure 5(a) shows a cross-sectional micrograph of two chips vertically connected with microbumps. The bottom die also supports TSVs to provide external connections. The microbump stacking process can be repeated to construct stacks with multiple die. Figure 5(b) shows a 3D memory stack with eight layers of DRAM chips all interconnected with TSVs and microbumps. This greatly increases the amount of memory that can be integrated into a given processor package area. Microbump bonding has some challenges, including higher thermal resistance due to the underfill and the additional height from the microbumps and associated metal connection pads. The interconnect density is also limited by the size of the microbumps, which can be difficult to scale to very small sizes and pitches.



(a) **(b)**

Figure 6. (a) Cross-section micrographs of a cache die 3D hybrid bonded to a CPU chiplet, (b) another view of a cache die stacked on a CPU chiplet with additional filler silicon.

F. 3D Stacking: Hybrid Bonding

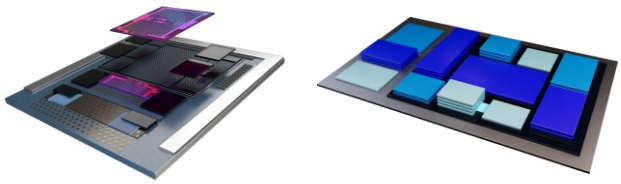
A more recent 3D stacking technology uses a two-phase hybrid bonding process. Rather than have microbumps that connect the metal pads on two chips together, the chips are fused directly together. The first phase consists of forming covalent bonds between the oxide of the two chips' respective surfaces. The second phase consists of a higher-temperature copper-copper bonding process that causes the metal pads on each chip to directly fuse together. By eliminating the microbumps completely, hybrid bonding can support higher interconnect densities (e.g., going from a pitch of 10s of microns for microbumps to single-digit microns for hybrid bonding). Figure 6(a) shows a cross sectional view of a cache die hybrid bonded on top of a CPU die. TSVs on the bottom die connect to the metal bond pads at the hybrid bonding interface. The bond pads on the top die are connected by bond pad vias to the normal metal stack of the top cache die. Figure 6(b) shows a graphical rendering of AMD's V-Cache™ technology that stacks a cache die on top of a CPU chiplet. This provides the ability to triple the capacity of the CPU's L3 cache at full bandwidth. In this implementation, additional passive filler silicon (shown as the floating gray pieces in the figure) is stacked on top of the CPU compute logic to help conduct heat from the processor pipeline to the package's cooling solution (not shown). The direct die-to-die interface without microbumps or underfill provides a thermally superior pathway compared to microbump-based 3D stacking.

III. Future Silicon Cityscapes

The examples discussed so far have all largely used only one or two packaging or stacking technologies at a time to address specific design objectives. However, as system requirements continue to increase, technology scaling slows down, and package sizes stop growing, we believe that the desire to integrate more functionality into a single package will inevitably drive the simultaneous utilization of multiple 2D, 2.5D, and 3D approaches within the same design.

A. Combination Approaches

We are already seeing the beginnings of the move toward SoCs utilizing multiple integration technologies. Figure 7(a) shows a blown up view of the AMD Instinct™ MI200 accelerator that combines both 2.5D silicon bridge technology with 3D microbump-based DRAM stacks all within the same design. Earlier GPUs combined 2.5D passive interposers with 3D DRAM as well. Moving forward, we envision future systems with more complex structures. Figure 7(b) illustrates an example depicting a hypothetical system that simultaneously combined multiple 2D, 2.5D, and 3D technologies in a single solution. From afar, the figure starts to show a resemblance to a small town or city, and so we refer to such designs as "Silicon Cityscapes."



(a) (b)

Figure 7. (a) An exploded view of an AMD Instinct™ MI200 accelerator showcasing multiple packaging technologies, (b) an illustrative example of a Silicon Cityscape combining heterogeneous chiplets with multiple 2D, 2.5D, and 3D advanced packaging and stacking technologies.

B. Design Space – Why Silicon Cityscapes?

From the start of the Exascale Computing era [9] to the emergence of massive machine learning applications, such as large language models [3][12], the demand for more and more computing power does not appear to be slowing. However, the industry and broader community faces several headwinds in our attempt to keep up with computing demand.

One is that Moore's Law is widely acknowledged to have slowed down. This means that instead of the previously reliable cadence of getting twice the transistor density every 18-24 months, we may now have to wait longer for each new silicon process node [8]. As discussed earlier, even when the new nodes arrive, they have been trending toward being increasingly expensive, leading to a slowing of improvement in cost per transistor.

Partitioning an SoC into smaller chiplets and using 2D and 2.5D packaging technologies to reintegrate them together has provided a path to continue scaling. However, placing chiplets and memory side-by-side in 2D/2.5D is quickly running up against the available real estate within the package. The examples previously shown in Figures Figure 3(b), Figure 4(b), and Figure 5(b) all show the silicon components consuming the vast majority of the available package area. Building larger packages is expensive and difficult due to a variety of mechanical engineering challenges. Like a city on an island (e.g., Manhattan, Singapore), when growth is constrained in the 2D plane, the natural option is to go up.

While some problems can be addressed via a "scale out" approach where computations are distributed over increasingly large clusters of processors [6], other computations have communication requirements that desire designs that keep as much of the computational resources co-located within the same package (where 2D/2.5D/3D technologies can provide orders of magnitude higher bandwidths than external package-to-package links). The combination of such "scale up" design requirements, increasing overall computational needs, and the area limits of modern packaging may lead to the eventual adoption of SoC constructed in a Silicon Cityscape fashion.

IV. Silicon Cityscape Research Topics

While the vision of constructing Silicon Cityscapes is exciting as it allows us to continue scaling the integration of more and more heterogeneous components within the same package, there remain many challenging research problems to enable the widespread adoption of this approach.

A. Chiplet Decomposition and Technology Selection

Given an SoC's design requirements, one of the first challenges is in determining how to partition the design into chiplets and how those chiplets should be reintegrated back together. The chiplet decomposition problem can require a deep design-space exploration of different architecture organizations along with cost analyses. A smaller number of larger chiplets reduces the overheads of die-to-die communication, whereas a larger number of smaller chiplet can reduce the silicon cost per chiplet. Some communication paths within the design may be able to tolerate lower bandwidths and/or higher latencies, and therefore may be amenable to using 2D and 2.5D packaging solutions. Other interfaces may require the highest possible bandwidths and lowest latencies, driving such components to be 3D stacked with hybrid bonding. Even for a design that may lean toward utilizing a larger number of smaller chiplets, chiplet reuse is an important factor in keeping the overall design costs under control. Previous work showed the effectiveness of being able to reuse multiple instances of the same chiplet both within a single product and across multiple products [8]. It is already very challenging to try to determine an optimal chiplet partitioning of a single architecture, but the design space increases when one wants to simultaneously optimize the partitioning of multiple architectures while minimizing the number of unique chiplets that have to be taped out. Developing techniques, tools, and methodologies to guide architects and designers in the task of mapping an architecture to a set of chiplets along with effective selection of integration technologies presents a set of interesting and exciting research problems.

B. Interconnect Infrastructure

The different chiplets and chip stacks within a Silicon Cityscape may have different needs based on bandwidth, latency, power, quality of service, priority levels, etc. There are significant research opportunities in exploring how to effectively interconnect everything in a scalable, performant, and energy-efficient manner that also provides easy inter-operability and modularity among the chiplets. This area includes a range of "network on chip" (NoC) topics, but extended across chiplet boundaries, generalizing into the third dimension, and having to negotiate a heterogeneous set of interfaces due to the mix of 2D/2.5D/3D technologies.

Hand-in-hand with the interconnect infrastructure research, effective utilization of Silicon Cityscapes may require additional assistance or explicit co-design with the application layers. Software can potentially allocate and place data in more convenient locations within the package, schedule processing tasks to be co-located near data sources, pre-schedule any necessary data movement to reduce "city-wide" congestion phenomena, and partition workloads across different computational resources (chiplets). Effective hardware-software co-design can potentially relax the requirements of some portions of the Silicon Cityscape's interconnect architecture, perhaps allowing some chiplets to be integrated with simpler or more cost-effective technologies (e.g., 2D/2.5D versus 3D stacking).

Another research topic for Silicon Cityscape interconnects is on designing protocols to support easy interoperability and composability/modularity of many diverse chiplets. Like a physical city, streets and highways are necessary, but for effective utilization we also have rules on their usage (e.g., traffic signals, speed limits, one-way streets). For Silicon Cityscapes, analogous protocols are needed so that different chiplets can communicate efficiently with each other. Some efforts, such as the Universal Chiplet Interconnect Express™ (UCIe™), are already underway in the industry to define some

of these interfaces and protocols [11], but these are only a start. Beyond data communication, further research can be done to devise interconnect solutions and general chiplet infrastructure to provide standard interfaces and approaches for security, power management, memory management, virtualization, boot-up and chiplet discovery, debug, profiling and telemetry, error reporting, and more across the Silicon Cityscape's components.

C. Power Delivery and Thermals

Power delivery in a real-world city can be very challenging due to varying power demands in space (residential vs. industrial) and time (day vs. night, weekday vs. weekend) as well as physical constraints such as where power lines and substations can be placed within a crowded metropolis. Within a densely-integrated Silicon Cityscape, analogous challenges arise, necessitating research into effective power distribution architectures that can span the different 2D/2.5D/3D and silicon-substrate boundaries. Especially for portions of the Silicon Cityscape that utilize 3D die stacking, power delivery to the top levels of the stack can become increasingly difficult due to the accumulated IR effects of traversing many sequential TSVs to reach the top layers. The complex packaging and stacking topologies can also induce an assortment of inductive paths resulting in L·di/dt challenges. Similar to the interconnect discussion, research into co-design with the software layers may be needed, or at least highly beneficial, to schedule work among the chiplets in a way that avoids concentrating too much work in one locale of the Silicon Cityscape, which would overburden the local power distribution capabilities.

Hand-in-hand with power distribution is the challenge of heat removal. Advanced packaging and stacking are used advantageously to increase the amount of compute and memory that can be packed into a given volume, but this can naturally lead to increases in power density as well. Similar to the power distribution for 3D-stacked chiplets, heat removal from a stack is also difficult. Microbump-based 3D stacking can be challenging because the interlayer underfill presents a higher thermal resistance that impedes the flow of heat from the lower layers to the top where the cooling interface typically resides. Hybrid-bonding 3D helps, but fundamental challenges remain from the fact that power density has increased. Either form of 3D stacking can also further aggregate thermal hotspots because die thinning reduces the lateral/planar thermal conductivity of the chiplets, causing heat to get trapped in a smaller area leading to higher temperatures. Research into co-design with software again may be an effective approach to reducing the magnitude of the thermal challenges through intelligent scheduling and work placement across the different compute resources within the overall package.

In a 3D stack of chiplets, power delivery and heat removal provide a dynamic tension that make it challenging to determine the best stack organization and/or placement of work among the layers of the stack. The most compute-intensive and highest power consuming work would favor being placed on the bottom chiplets of a stack to maximize the quality and reliability of the power delivery. However, placement at the bottom of the stack is likely the worst place to be in terms of heat removal, and so optimizing for either power delivery or thermals will tend to be suboptimal for the other. Much research is needed in effective ways to simultaneously manage both within Silicon Cityscapes.

D. Reliability

The Silicon Cityscape approach can potentially enable the integration of a massive amount of silicon within a single package or socket. While the computation promise of such an approach can be attractive for performance and energy efficiency, it may potentially introduce new challenges in terms of reliability, availability, and serviceability (RAS). If a failure occurs within the package, this can result in the potential loss of functionality of one or more components within the package. If the failure is transient, this can represent a potential reduction in availability of the system within the package. However, if the fault is permanent and cannot be repaired, disabled, or worked around, this can result in potentially have to discard the entire package, which may include a significant amount of otherwise functional silicon (e.g., failure of one chiplet in an N-chiplet Silicon Cityscape resulting in losing the remaining N-1 functional chiplets). Advanced packaging and stacking can likewise result in a reduction in potential serviceability, as components that conventionally may have been outside of the package (and hence easier to service or replace) may now be tightly integrated inside the package where they may be no longer accessible for all practical purposes. Architecting and delivering reliability and resilience across all of the components in a Silicon Cityscape, and across the entire lifetime of the part, is a wide-open research subject.

E. Software

Earlier sections already touched on the research opportunities related to hardware-software co-design for Silicon Cityscape systems. Beyond these, a highly chipletized and heterogeneously-integrated system presents additional opportunities for software research. An important open problem is in determining an effective hardware-software interface to expose, describe, program, and manage all of the components and their relationships. The low-level software layers (e.g., operating system, hypervisor) needs to enumerate all of the different types of chiplets and compute resources as well as memory components. To effectively manage performance, job scheduling, and data placement, the software also needs some way to learn about and understand the characteristics of the interconnect interfaces between the components (e.g., how much bandwidth?) and the overall interconnect topology (i.e., how far away are any two components?). Likewise, the Silicon Cityscape may need to provide runtime monitoring capabilities to the software layers, such as reporting performance and utilization statistics, memory behavior, power usage, thermal conditions, suspected security activities, and more. Each of these individually as well as in concert provide rich research problems to tackle.

At the user and programmer level, additional research is needed to develop tools, compilers, programming models, frameworks, and more to effectively map the higher-level problems and algorithms to the underlying hardware capabilities provided by a particular Silicon Cityscape package. Hopefully the continued evolution of higher-level frameworks (e.g., PyTorch, Tensorflow) can abstract much of the potential complexity away from the majority of programmers, but significant research remains to be done on how to provide the necessary runtime, middleware, and other system-software support to those writing and optimizing the frameworks on behalf of the larger programmer communities.

V. Conclusions

This is an exciting time in the semiconductor industry. There are many technological challenges facing us all in the coming years due to the combination of increasing difficulties from running up against physical limits and the accelerating demand for more computational capabilities. However, this also presents many new opportunities to rethink how to organize, architect, and construct new Silicon Cityscapes to deliver the necessary compute for the next decade and beyond. While industry is delivering a range of new products increasingly utilizing more advanced packaging and stacking technologies, there are massive opportunities for the broader research community to innovate and impact where we go from here. The opportunities are immense for transformational research across disciplines and vertically throughout the hardware-software-application stack.

References

[1] Noah Beck, Sean White, Milam Paraschou, Samuel Naffziger, ""Zeppelin": An SoC for Multichip Architectures," IEEE International Solid-State Circuits Conference, February 2018.

[2] Pat Conway, Nathan Kalyanasundharam, Gregg Donley, Kevin Lepak, Bill Hughes, "Blade Computing with the AMD Opteron™ Processor ("Magny-cours")," IEEE Hot Chips 21 Symposium, August 2009.

[3] William Fedus, Barret Zoph, Noam Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," arXiv 2101.03961, January 2021.

[4] Warren Flack, Robert Hsieh, Gareth Kenyon, Manish Ranjan, John Slabbekoorn, Andy Miller, Eric Beyne, Medhat Toukhy, PingHung Lu, Yi Cao, Chunwei Chen, "Large Area Interposer Lithography," IEEE 64th Electronic Components and Technology Conference, May 2014.

[5] R. Kalla, Balaram Sinharoy, J. M. Tendler, "IBM POWER5 Chip: a Dual-core Multithreaded Processor," IEEE Micro, vol. 24, issue 2, March-April 2004.

[6] Pejman Lotfi-Kamran, Boris Grot, Michael Ferdman, Stavros Volos, Onur Kocberber, Javier Picorel, Almutaz Adileh, Djordje Jedvic, Sachin Idgunji, Emre Ozer, Babak Falsafi, "Scale-out Processors," in the International Symposium on Computer Architecture, June 2012.

[7] Joe Macri, "AMD's Next Generation GPU and High Bandwidth Memory Architecture: FURY," IEEE Hot Chips 27 Symposium, August 2015.

[8] Samuel Naffziger, Noah Beck, Thomas Burd, Kevin Lepak, Gabriel H. Loh, Mahesh Subramony, Sean White, "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families," in the International Symposium on Computer Architecture, Industry Track, June 2021.

[9] Oak Ridge National Laboratory, "Frontier Supercomputer Debuts as World's Fastest, Breaking Exascale Barrier," https://www.ornl.gov/news/frontier-supercomputer-debuts-worlds-fastest-breaking-exascale-barrier

[10] David Papworth, "Tuning the Pentium Pro Microarchitecture," IEEE Micro, Vol. 16, April 1996.

[11] Debendra Das Sharma, "Universal Chiplet Interconnect Express (UCIe)®: Building an Open Chiplet Ecosystem," White Paper, http://uciexpress.org, 2022.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention is All You Need," in the Conference on Neural Information Processing Systems, USA.