HETEROGENEOUS INTEGRATION ROADMAP

2019 Edition

Chapter 2: High Performance Computing and Data Centers

http://eps.ieee.org/hir

The HIR is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

We acknowledge with gratitude the use of material and figures in this Roadmap that are excerpted from original sources.
Figures & tables should be re-used only with the permission of the original source.

Table of Contents

To download additional chapters, please visit
http://eps.ieee.org/hir

Chapter 2: High Performance Computing and Data Centers

1. Introduction: The Need for Heterogeneous Integration

Semiconductor devices targeting the high-performance computing (HPC) and data center markets have always represented the prevalent state-of-the-art in device and process technologies. The needs in these market segments have generally demanded the highest processing rates, highest communication rates (low latencies and high bandwidth, often both of these simultaneously) and highest capacities with extreme requirements for packaging that address the interconnection requirements and higher power dissipations. This is a trend that is likely to continue as a wide variety of applications for HPC systems and data centers have emerged over recent years.

The term **chiplet** has been used to describe a die that is integrated with other such dies (or chiplets) inside a package. An alternative term, **dielet**, is also used synonymously as chiplet. In this chapter, these terms are used interchangeably. As an aside, it is worth noting that the term chiplet strictly means part of a functional chip that is not necessarily stand-alone. In the way this term is used, a chiplet can be a completely functioning die, such as a HBM stack or a multiocre CPU. In its current use, the term chiplet is used to refer to either a part or the whole of a functional chip, in departure from the strict meaning of the term.

This chapter rationalizes the clear need for heterogeneous system integration that realizes systems-in-a-package (SiPs) that target the HPC and data center markets, and that identifies potential solutions and short-term, medium term and longer term challenges that are encountered in realizing these SiPs. Although, as in the past, the processor-memory performance gap remains a key driver for the overall system architecture, new factors that drive the need for heterogeneous integration in the HPC and data center markets have been emerging. These include technology limitations, new and emerging applications, and scaling needs for surmounting power dissipation, power delivery and package IO constraints. These needs and their implications are examined below.

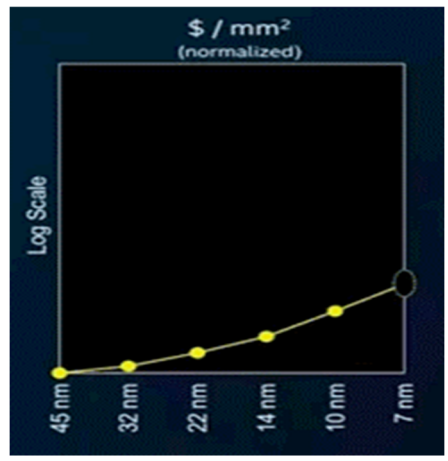


Figure 1. Die cost trends at 45 nm nodes and beyond (from [Bohr 2017])

1.1 Die size limitation

In the past, the technology node (feature size) has been the representative of a specific generation of the mainstream CMOS technology, and the most recent technology was surpassed by a new technology within 18 to 24 months of its introduction. In recent years, as feature size shrank, a node actually encompassed several consecutive technology generations characterized by the shrinking dimension of circuit elements that were realized within the node through process optimizations and circuit redesign. Consequently, a node has begun to last for several years but has actually enabled scaling down of circuit elements to continue through these innovations (dubbed as "hyperscaling" [Bohr 17]) for a relatively fixed feature size. A consensus that has been emerging in recent years is the use of a technology scaling metric that represents the transistors per unit area for some basic circuit elements such as NAND gates or scan flip-flops [Bohr 17] or other cells specific to a vendor [Lu 17]. With hyperscaling in use, the

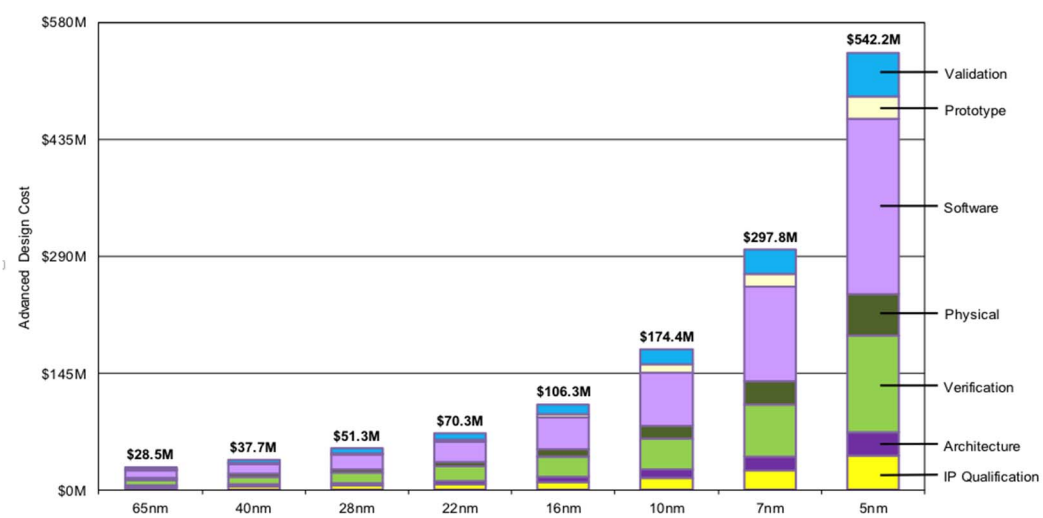
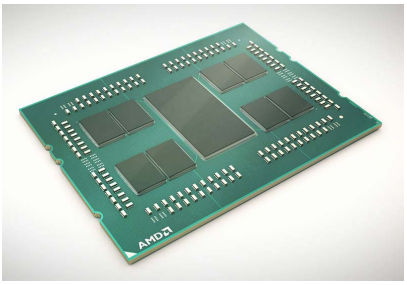


Figure 2. Chip design cost trends at different nodes (from IBS , as reported in [Lap 18])

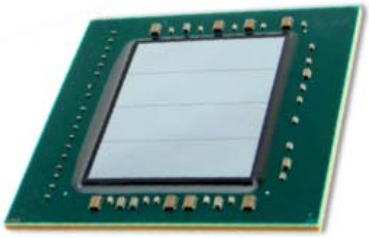
classical generation boundary has to be redefined as the transition between the most refined (that is, best hyperscaled) process technology for a given feature size and the first process implementations for the next technology node.

An implicit assumption behind Moore's Law is that the die size remains unchanged, so that doubling of device or circuit elements doubles performance as a new node is introduced. However, at feature sizes of 10 nm and lower, that assumption is no longer valid. Yield issues, device limitations and escalating design costs are no longer making it economically feasible to maintain the die size with technology scaling [Bohr 17, Naf 18, Ren 19], as seen in Figures 1 and 2. Consequently, even though circuit elements are scaling following Moore's Law, performance scaling is no longer happening following Moore's Law.

Heterogeneous integration certainly offers a solution for performance scaling following the Moore's Law trajectory of the past. For example, instead of fabricating a single large multicore CPU die, smaller dies can be tessellated (tiled) within a package on an interposer with very short connections in-between the dies to realize the same performance offered by a single large die. The smaller dies have higher yields and as long as integration cost is reasonable, the overall SiP solution will scale in performance following Moore's Law. Some high core count CPUs and large FPGAs are actually implemented in this fashion are shown in Figure 3.



AMD EPYC Partitioned Server Processor



Xilinx Virtex Partitioned FPGA

Figure 3. Examples of current products employing die tessellation

1.2 The Processor-Memory Performance Gap

Memory systems and energy-efficiency were listed as a challenge area in a past DOE report prepared for the state of High-Performance Computing (HPC) in the US [Luc 14]. These challenges remain at present, not just for HPC but also for data centers, which consume about 71 billion KWh of electricity annually in the US [She 16]. SiPs that integrate processing and memory chiplets address these challenges, and many SiP product offerings from current and the past several years exist.

DRAM devices are the mainstay of memory systems for virtually all high-end systems, including SiPs that integrate processing/accelerator and memory chiplets. The performance disparity between processing engines and the DRAM memory system have always limited and continue to limit the overall system performance. This performance disparity has been and continues to be mitigated through the use of multiple cache levels, hardware or software prefetching, speculative bypassing of memory updates, and other innovations in the memory interface (including DRAM controllers) and DRAM device and memory system side innovations, as well as software-driven techniques that focus of data layouts and code optimization. These techniques attempt to bridge two aspects of memory system metrics: memory access latency and performance.

Multicore CPU chips and GPUs (and other accelerators) impose a severe demand on the memory system in terms of both latency and particularly bandwidth. Without the availability of low-latency, high-bandwidth connections to memory, the performance potential of these processing engines remain unexploited. Off chip placement of memory from the processing elements, limited by package IO, preclude the realization of low latency, high bandwidth connections. Current high-end heterogeneously integrated product offerings certainly do this, as increases in the core count or processing performance per core demand tight, low-latency, high bandwidth connections to memory within a package, requiring both high lane count and short interconnections to be realized within the package. The memory within the package forms an additional layer within the memory hierarchy and is supplemented with additional memory outside the heterogeneously integrated package.

Two types of dynamic memory devices have emerged as attractive solutions for meeting high-bandwidth, low-latency memory demand:

GDDRx: Graphics double-data-rate DRAM, GDDR, particularly GDDR5x and GDDR6, which effectively uses a wider and modified memory interface to provide higher bandwidth using more traditional DRAM dies, permitting a fast market entry. Aggressive GDDR6 offerings are now available for meeting the data rate demands of new GPUs, and GDDRx technologies are compared to HBM briefly in [Muj 18]. Concurrency in the GDDR6 memory system is limited to 2 channels per chip at present, with 8-bit- and 16-bit-wide connections per chip. A long-term JEDEC standard for GDDRx is not available at this time.

HBM: High bandwidth memory takes a different approach to providing high data rates by stacking memory dies and using wide interconnections that implement 1 or 2 channels to each die, with 128-bit wide connections to each die on the DRAM stack [Jun 15, WikH 19]. The HBM technology thus keeps the real estate needs of the DRAM fixed and grows capacity by increasing the stacking count. A JEDEC standard has been defined and recently updated for five successive generations of HBMs [Hil 19a, JED 19]. As an aside, in the short-term the void left by lack of HBM2 was being met by GDDR5x and GDDR6 in some markets, but a recent HBM2e offering [Ver 19] may change the situation.

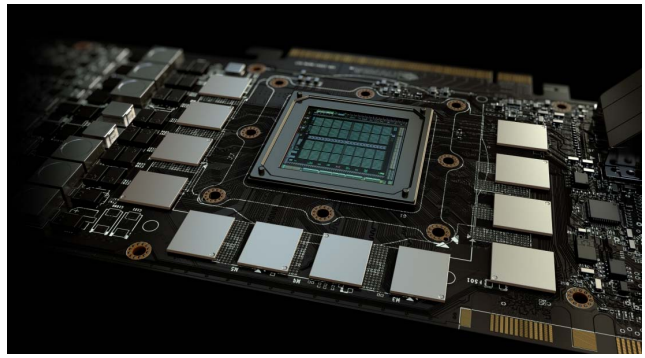
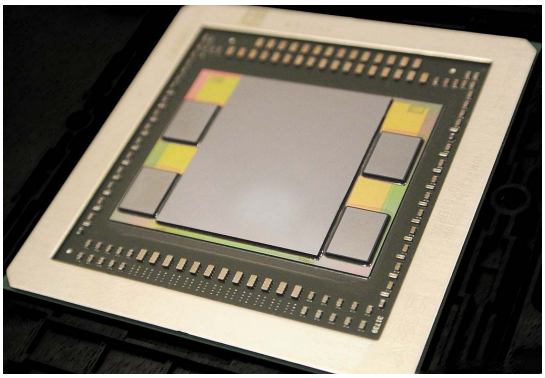
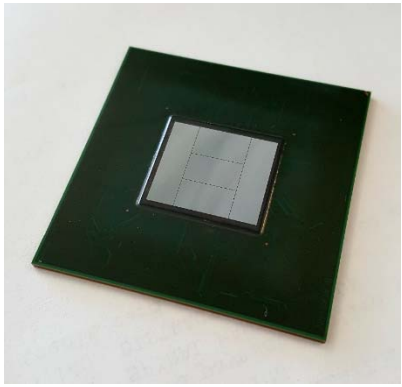


Figure 4. Nvidia Pascal showing GDDR memory surrounding the processing unit. Memory connections are through the PCB

Irrespective of which of these two DRAM technologies dominate, the memory latency needs of emerging processing systems can be met by locating the DRAM memory as close to the processing engine as possible. Heterogeneous integration offers a solution here – the DRAM memory (GDDRx or HBM) dies can be placed adjacent to the processor on an interposer or stacked onto the processor. Adjacent placement appears to be an attractive solution, particularly for HBM that uses wide interconnections, as 3D integration technologies for stacking HBM and processors have to address thermal challenges as well as implement high via counts for the memory channels, and a number of current SiPs actually use this solution today. The short connection between the processor and memory can support high data rates as well as high lane counts. The integration of DRAM and processor dies within a single package not only addresses the bandwidth and latency needs of the memory system but also addresses another potential system-level design constraint – power dissipation. The energy spent in moving data between processor and memory is dramatically reduced compared to what would have been expended when the DRAM was located outside the CPU package. Although the wiring length reduction and subsequent bandwidth and power dissipation improvements are significant for 3D integration, these technologies for stacking HBM and processors have to address thermal challenges as well as implement high via counts for the memory channels. Examples of 2.5D and 3D HBM integration are shown in Figure 5.



AMD Fiji GPU w/HBM integration utilizing 2.5D packaging technology with Si Interposer



GLOBALFOUNDRIES and ASE test chip demonstrating 3D integration of HBM on Logic

Figure 5. Examples of heterogeneously integrated products incorporating HBM in 2.5D and 3D configurations

Stacked SRAM: An emerging player in the memory products market is stacked SRAM, a stack of thinned SRAM dies [McC 18, Ram 18]. The thinned dies have been demonstrated to have retained all of the electrical characteristics of their normal counterparts, offering resiliency against noise and higher temperature as supply voltage scales down. Stacked SRAMs thus appear as a viable alternative for fast, high-bandwidth memory for SiPs targeting both HPC and emerging applications [Arv 18]. For conventional processing solutions, stacked SRAMs may be used as a cache level to the HBMs or external DRAM or as buffers for high-speed transfers in general, both within and across packages. For accelerators for machine learning and graph processing, stacked SRAMs permit high-bandwidth, low-latency access to critical data such as weights, graph node data and the like.

High-Performance Non-Volatile Storage: Semiconductor non-volatile randomly-accessibly storage devices have appeared in the market for a while. The most dominant of these have been in the form of Solid State Drives based on Flash memory technologies that provide bulk storage and faster performance than traditional hard disks. Unfortunately, SSDs do not provide enough performance for many data center applications. Some notable applications in this regard are:

(a) Transaction processing systems that need the state changes made by a completed transaction to non-volatile storage as soon as possible to implement the ACID semantics (Atomicity, Consistency, Isolation, Durability);

(b) Checkpointing in HPC Systems: fast non-volatile storage is also needed to checkpoint the state of a long-running program to avoid complete rollbacks and wasted efforts on errors or crashes; and,

(c) Memory-centric computing, where computing engines are moved closer to the data storage (See Section 1.3).

A significant number of non-volatile, randomly accessible storage devices have emerged in the market in recent years and their characteristics are depicted in Figure 6, compared against NAND-Flash and Intel Optane (formerly, 3d X-point) [Int 19]. These devices certainly appear within the storage hierarchy for a SiP.

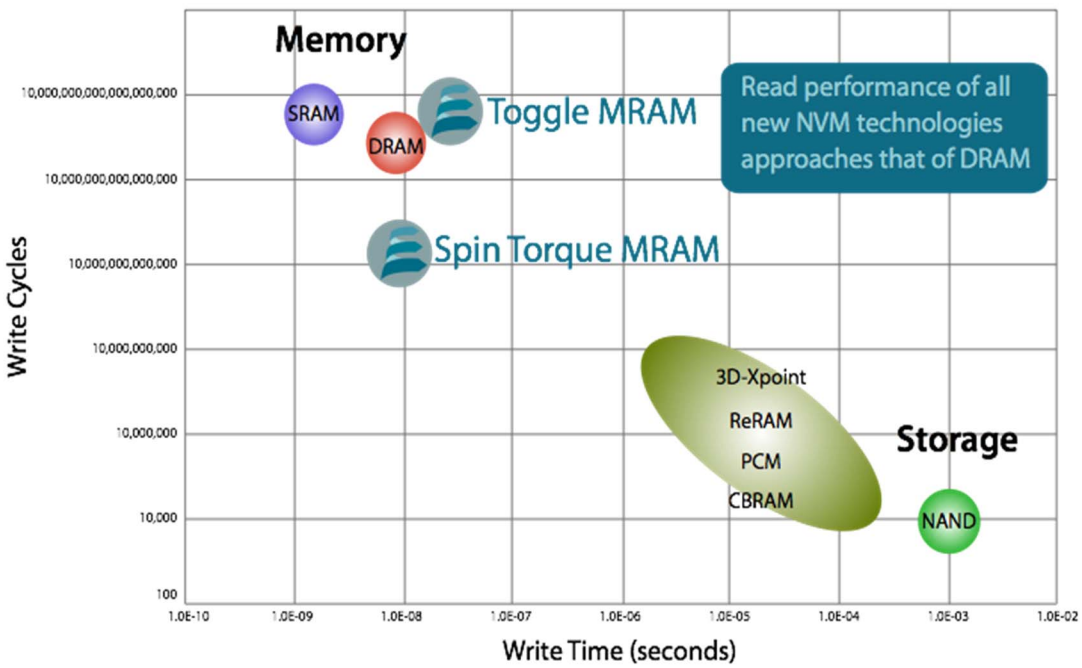


Figure 6. Emerging and existing fast non-volatile storage technologies (from [Hil 18])

The newer non-volatile memory technologies indicated in Figure 6 sharing characteristics close to that of Optane include: (i) Resistive RAM (ReRAM), also known as memristors, where the resistance of a semiconductor material (such as HfO_2, or some forms of tantalum or silicon oxide) can be controlled electrically to represent stored information; (ii) Phase change memory (PCM), where Joule heating can change the state of material phase (from crystalline to amorphous) to indicate what is stored. Two promising non-volatile RAM technologies that come close to RAM speed include Everspin's Toggle MRAM (Magnetic RAM), and Spin Torque Transfer Magnetic RAM (STT-MRAM) [Hil 18]. STT-MRAM can be the basis of forming a fast write buffer for fast transaction commitment before

they are moved to SSD-based storage [Eve 18] and thus will be a prime candidate for inclusion within a SiP that support very high transaction commitment rates offering higher write endurance at close-to-DRAM write speeds.

It is thus inevitable that heterogeneously integrated offerings targeting the data center and HPC market integrate processing engines (and perhaps even accelerators) with a variety of memory chiplets in the form of HBM, HMC, Stacked SRAM and NVRAM. This demands the use of high-lane count, short interconnections within the package between memory and processing elements.

1.3 Exploiting Accelerators for Emerging Applications

In recent years, several important market drivers have emerged, primarily driven by new applications. These include:

- **Internet-of-Things** (IoTs) will have significant applications and processing needs [HP 17] and will demand pre-processing at edge nodes with high IO connectivity to the "things" (sensors/actuators) and final sensor fusion and processing/storage at nodes that would typically be within the cloud. Although the deployment count of IoTs was predicted to be 50 billion devices by 2020 [Eva 11], and to 500 billion devices by 2030 [Cis 16], that prediction has been downgraded to 7 to 28 billion devices by the early 2020s, still representing a very large growth in a new market segment. The IoT product market is predicted to be significant [Hor 18].
- **Data analytics** is a growing need in the mass e-commerce and financial industries which rely heavily on analytics for targeted marketing, inventory optimization, market forecasting, economic forecasting and other aspects of business. Smart healthcare, particularly personalized medicine and pharmaceutical discovery and drug repositioning, have also relied heavily on data analytics. Social networking has also been a driver in this market. In general, all of these analytics require large, potentially unstructured data sets to be processed and analyzed to discern some hidden patterns, requiring the ability to store and process large data sets ("big data") that have volume, dynamics and high data rates. Further, the processing requirements are varied – in addition to numerical processing, symbolic processing is needed, often justifying the need for unconventional processing substrates, starting with customized FPGA-based solutions to the use of GPUs and customized hardware.
- **Intelligence as a service** represents a need to recognize and predict information using machine learning techniques. The advances made on machine learning techniques, notably convolutional neural networks (CNNs) and deep neural networks (DNNs), have seen the deployment of GPUs, FPGAs and special-purpose accelerator chips for implementing CNNs and DNNs. Cloud-based server platforms incorporating these accelerators are already in use.
- **Blockchain processing**: Another application paradigm that will emerge as a driver of heterogeneous integration in the data center market is blockchain processing. Blockchain processing gained its notoriety in implementing bitcoin mining as a way of validating transactions for distributed ledger keeping. The blockchain processing paradigm has significant applications beyond bitcoin mining or financial transaction recording, in domains such as voting, identity establishment, governance and healthcare. Blockchain processing is highly parallelizable and specialized engines other than GPUs may be well-suited for the task, requiring integration with memory and IO chiplets inside a package as a SiP.
- **Special functions and accelerators** have been steadily emerging, starting with GPUs, which have been the vanguard of the accelerator business targeting a wide variety of applications that go beyond graphics and scientific computing. As an example, significant number of products have appeared to accelerate neural networks; examples are described in [Dem 18, Gwe 18, Hal 18]. FPGAs are also emerging as versatile, programmable hardware accelerators/special functions [Whe 18]. Specialized application domains have driven development of new paradigms such as quantum computing, cognitive computing and graph processing [Ozd 17, Gra 17].
- **Memory-centric computing** has emerged as a paradigm for many applications that involve large data sets and requires a high processing rate or low processing time [Sai 16]. Examples of applications using the memory-centric paradigm are in-memory databases, general ACID transactions (implementing Atomicity, Consistency, Isolation, Durability), real-time analytics and others. In this paradigm, high value (that is, often-used) data is kept in high-capacity, high-speed memory (such as HBMs and/or non-volatile semiconductor storage like MRAM or Intel's Optane or ReRAM) and bulk computations are performed on data sets at processing nodes very close to the stored data [Men18]. This avoids the high

cost of moving data back-and-forth across a conventional memory hierarchy composed of secondary storage and traditional DDRs. Ideally, memory-centric computing requires the computing logic to be embedded into the high-speed storage devices and until that becomes a practically viable solution, the processing logic for memory-centric computation can be placed as close to the high speed memory holding the high-value data. For instance, the computing logic performing fairly simple processing on wide memory chunks in parallel can be implemented and connected to HBMs in a 2.5D configuration in an interposer or implemented in a "logic" layer underneath stacked memory dies in a 3D configuration.

These broad market drivers are going to influence the product spectrum and the growth of processing and storage products that fall into the broader HPC and data center markets.

For most of these emerging applications, special-purpose accelerators, including custom ASICs, FPGAs and GPUs, provide an energy-efficient and significantly faster implementation compared to software implementations on general-purpose CPUs. Many of these accelerators also have a need to access significant amounts of data from memory. For instance, neural network implementations require fast (that is, local) memory storage to hold the input data set, weights and activations as the inputs propagate through the network. Additional fast data storage is also needed to support any "lowering" transformation that replaces convolution with GPU-friendly matrix operations. These data sets can easily range from several tens of MBytes (with reduced precision) to hundreds of Mbytes at full hardware-supported precision. Heterogeneous integration provides a solution to meet some of these memory needs by integrating accelerators with Stacked RAM or HBM within a package, akin to what was used to bridge the performance gap between DRAM and general-purpose CPU cores (Section 1.2). Of course, fast local memory implemented on the accelerator itself, as in the Graphcore IPU [Gra 17], can provide the first level of storage, backed up by Stacked SRAM [Arv 18, McC 18], or HBM for potentially faster performance. Similar solutions can be deployed with other types of accelerators.

1.4 Package IO Limitations for future Ethernet Switches and Routers

The Ethernet Alliance has a roadmap that defines the growth of link speed from 100Gbps (Giga bits per second) today to 400Gbps by 2020 and to 6.4 Tbps (Tera bits per second) by 2030, as shown in Figure 7. To implement such a high link speed in the data center will require the network chips to increase their electrical I/O bandwidth to match the link speed of the optical modules connected at the front port of the network switches and routers. The aggregated I/O bandwidth per network chip is calculated by the data rate per I/O lane times the total number of lanes per package. Figure 8 shows the relationship between package size and the total I/O bandwidth in Tbps projected by [Eth 17].

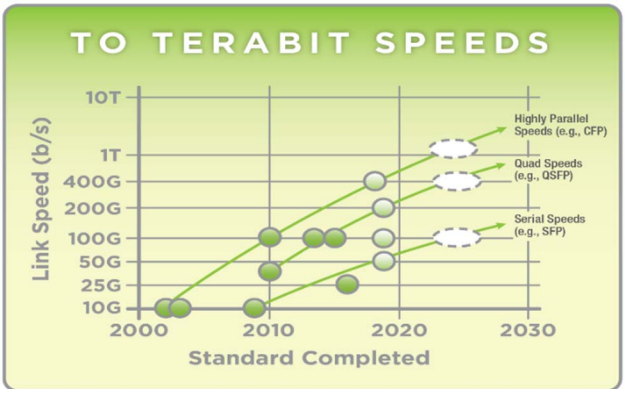


Figure 7. The Ethernet Alliance Roadmap (from [Eth 17])

As of late 2018, the I/O bandwidth of the existing switch devices running at 25Gbps per lane was approaching their limit at 8Tbps as the package body size increased to their 70mm x 70mm limit due to solder joint reliability, warpage, etc. at this large body size. To go beyond 8Tbps with a package size of 70mm x 70mm, we will need to increase the data rate from 25Gbps to 50Gbps as shown in the Figure in order to reach 14Tbps total I/O bandwidth. For short-term solutions that go beyond 14Tbps, the IEEE 802.3 committee is exploring 112Gbps per lane to meet ever-increased Ethernet link speed for Terabit networks. The Nyquist frequency of 112Gbps data rate using NRZ signaling is 56GHz which will have channel loss lower than -35dB after ~6-inch traces on a PCB board. For today's SerDes capability, the receiver will not be able to recover the received signals below -35dB channel loss. Therefore, PAM4 [Zha 17] or higher order PAM signaling is used for data rates >50Gbps to lower the Nyquist frequency to reduce

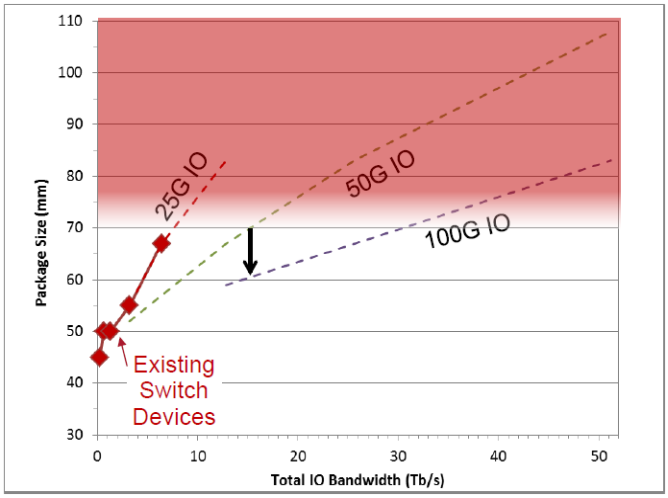
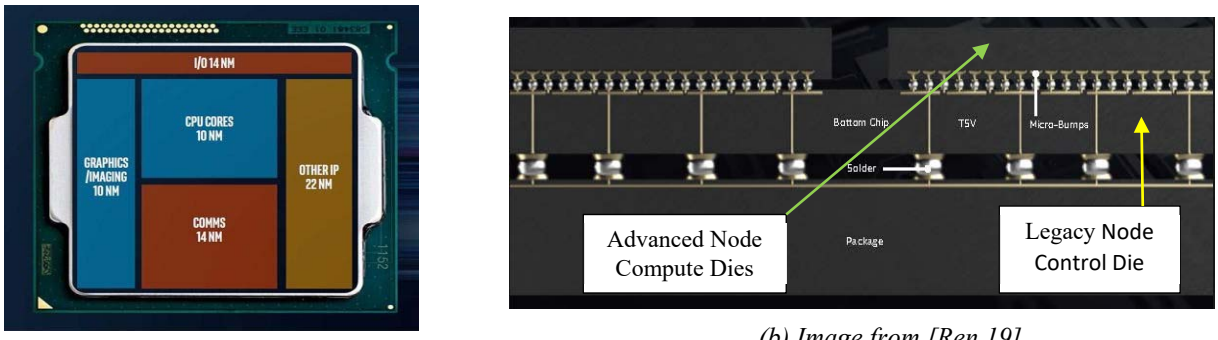


Figure 8. Relationship between package size and total IO bandwidth (from [IEEE 802.3])

the channel loss. However, the price to pay for PAM signals is decreased SNR, increased power, and crosstalk issues. For long-term solutions, we should seek a highly integrated system that can bring devices closer to each other in order to reduce the interconnect distance. Therefore, heterogeneous integration of various electrical and optical devices and connectors used in the data center can pave a way to enable the future Terabit network.

1.5 Integration of dies from diverse nodes and technologies

Fabrication facilities for new nodes cost in excess of several billions of US dollars. Heterogeneous integration offers a way of continuing the use of dies that are not performance-critical with high-performance dies from a newer generation, an example of which is shown in Figure 9(a). An example of this is the Intel Foveros product, shown in Figure 9(b).



(a) Image from [Bohr 17]

(b) Image from [Ren 19]

Figure 9. Examples of heterogeneously integrated dies from different generations in a 2.5D configuration (Figure 9 (a)) and in a 3D configuration (Figure 9 (b))

Another example expected to be quite common in the world of HPC is the integration of HBM, GPU, general-purpose cores and high-speed IO dies (quite likely from different vendors). The AMD Fiji product of Figure 5 is an example of this.

As new processing paradigms emerge, dies with analog and digital capabilities need to be integrated within a SiP. These include the use of analog neural networks (which are often claimed as more energy-efficient than their digital counterparts), or analog dies implementing neuromorphic processors, with dies that use digital storage/processing.

2. Analyzing the Future Demands of SiPs: Approach

In this section, we enumerate the design trends for future SiP products that target the HPC and data center markets and translate their demands into the techniques used for heterogeneous integration.

To begin with, we first identify the main drivers in HPC and data center markets that specify the required features of SiPs in these markets. It is important to note that the market drivers described in this report were identified in more than one independent expert-authored source in publicly available documents or presentations. The markets identified are enabled not only by trending needs but also by new devices that enable new applications. Trends in very specific and narrow domains are not covered in this report.

Cloud-based infrastructures supporting applications as well as platform-as-a-service have ushered in the age of mega data centers that span several acres, and the three market leaders have invested to the tune of $30 billion dollars in growing their mega data centers in the past few years. Although growth in energy consumption of data centers in the US has slowed from its past annual growth rate, the improved energy efficiency of computing and storage platforms have prompted data center operators to pack data centers with additional IT equipment per unit footprint area. Coupled with the introduction of accelerators such as GPUs and specialized ASICs (such as those that are used in very high-speed telecommunication switching/routers), the rack load in the high-end data center and HPC products has gone up dramatically, up at 60 to 80 KWatts per rack in many cases. SiP products offer the promise of reducing the rack power footprints through integration of components that communicate heavily (such as processor/accelerators and memory) as a SiP, dramatically reducing the significant amount of energy that is otherwise spent in moving large amounts of data at a high rate among interacting components.

The following trends appear to have a fairly broad appeal vis-à-vis future HPC/data center SiP product markets.

2.1 HPC System Trends

The demands in the HPC sector for scalable, high-performance and energy-efficient systems have always leapfrogged the capabilities of the current market offerings. Historically, big-science applications have driven the market, and the market trend and system requirements have been described in reports and presentations [Luc 14, Kag 17, Snell 19]. However, HPC systems are seeing increased use in the life sciences domain [Rus 19]. The big-science applications targeting high-end HPC systems have historically looked at particle physics, weather/climate modeling, energy exploration and other fields. These applications have driven an exponential growth in the total number of Tflops delivered by both the highest-performing HPC application (109 Tflops by April 2020) as well as the averages of the top 500 HPC installation (107 to 108 Tflops by April 2020, into the zettascale era) [Sod 16]. Staying on this trajectory requires the heterogeneous integration of processing and storage elements along with communication infrastructures. In the life sciences applications of HPC systems, large data sets acquired from recent instrumentation (such as light sheet microscopes which can generate 25 TByte data sets in a week at 75% utilization), genomics and the longer-term goal of analytics and approaches for precision medicine dominate the application base [Rus 19]. The processing requirements for life sciences are also showing that reliance on both general-purpose processors and GPUs and machine learning aspects can benefit from specialized hardware.

The overall approach for designing PC installations has also changed. The prior strategy of building a HPC system by choosing the highest-performance general-purpose processor and building everything else around it is falling out of favor, as general-purpose processing engines are reaching a per-core performance limit. Current design trends for HPC systems take a more holistic approach and look at the crosscutting issues across the processing nodes, storage, and interconnection fabric [Kag 17]. Interconnection fabrics, in particular, are no longer passive entities that provide connectivity – higher level functions are being steadily implemented within the network adapters and within the switching fabrics. Communication latency reduction will continue to drive the interconnection fabric design for HPC installations. However, this needs to be accompanied by improvements in connection bandwidth to handle increasingly larger data sets. Other implications of the networking infrastructures will parallel the needs seen in the data center segment, as described in Section 3.2.

Storage systems in HPC installations (and data centers) are evolving at all levels. SSD drives are displacing hard-disks at the first level of bulk, non-volatile storage systems. Non-volatile memory devices that operate at DRAM-like speeds, like Intel's Optane and other non-volatile semiconductor memory devices, are being introduced to add a hitherto missing level in the storage hierarchy. Fast bulk storage devices are useful for many long-running HPC applications that require the computation state to be checkpointed. As a solution to address the memory bottleneck imposed by package IO limitations in terms of both bandwidth and access latency, stacked DRAM chips, most notably HBM, are showing up in SiPs that provide the computing infrastructure for HPC systems.

2.2 Data Center Market Trends

In reality, the differences between HPC installations and data centers are blurring rapidly. The use of traditional, general-purpose multicore architectures is likely to continue as the main processing infrastructure for data centers. Aside from incremental improvements to the microarchitecture, the emphasis of accommodating a growing variety of applications with large cache footprint and/or poor cache locality are being addressed by trading of cores for larger caches at the lowest level and the use of non-inclusive cache hierarchies. As in the HPC market, the integration of stacked memory dies, multicore/vector CPU die and high-speed communication infrastructures in a SiP offering is going to cater to the high end data server market. The use of a large number of simpler cores for processing non-numerical workloads in a die integrated with other components or within a single package will also be a preferred way of making the processing substrate energy-efficient while retaining a performance advantage on non-numerical workloads. Last, but not the least, heterogeneously integrated SiPs for the processing substrates are likely to integrate chips from different process generations to permit critical components to be implemented and re-engineered to take advantage of process improvements [Bohr 17].

Acceleration engines – in the form of dedicated hardware in the form of GPUs, FPGAs or dedicated hardware for special functions (such as engines to support deep neural networks) – are also going to appear prominently in the data center market segment as analytics, machine intelligence and other similar applications permeate the market. The current systems/products from Intel [Deo 17], Microsoft [Fow 18] and Google [Jou 17] for deep learning applications are SiPs that integrate FPGA or dedicated hardware devices with stacked memory dies and other components. Future offerings are likely to become mainstream and integrate general-purpose processing engines along with these components to offer enhanced performance and improved energy efficiency.

Newer non-volatile memory devices operating at DRAM-like speeds offer the promise of accelerating transactional systems where results need to be committed to a non-volatile memory device. It is thus conceivable to see such memory devices to be integrated with processing substrates and accelerators in SiP offerings. Acceleration of big data applications are likely to incorporate accelerators in SiPs and incorporate substantial amounts of SRAM in a distributed fashion on the processing dies [Gra 17], along with generous local DRAM memory within the package in the form of stacked memory dies.

Internet-of-things will also drive the design of processing substrates at the edge [Eva 11, Deo 17, HP 17], where a relatively large number of data streams at low data rate need to be pre-processed and conditioned for further processing in the cloud data centers. SiPs integrating memory, a large number of analog and digital processing cores and possibly switching logic for multiplexing across low speed links for pre-processing appear attractive both in terms of performance and their lower power needs.

As in the HPC systems, networks at the data center level will have to evolve to accommodate the full potential of SiPs that provide the processing/storage infrastructures for servers. Although a large class of data center applications are not latency-sensitive, some are, and as in the world of HPC systems, the networking infrastructures will have to address high bandwidth and lower latency needs. At the data center scale, Ethernet will continue to remain the mainstream protocol of choice [Cis 19] and the switching and routing infrastructures themselves need to be implemented as SiPs to enable high-speed processing of higher-level routing functions. At the projected data rates of 100 Gbps and beyond (to 1 Tbps), line-speed signal processing functions need to be incorporated in these SiPs to perform signal recovery and conditioning functions.

3. Demands of Future SiPs and Solutions for the HPC/DC Market

We enumerate the requirements dictated by these SiPs on the heterogeneous integration methodologies and processes, as well as any influence the integration may have on the components being integrated. To do this, we use the conservative scaling assumptions described in Section 2 and see how this influences the following:

1. On-package interconnections
2. Off-package interconnections
3. Signal integrity and distribution needs
4. Power distribution and regulation
5. SiP-level global power management and overview of thermal management
6. Security and reliability issues
7. Design tools
8. Impact on the supply chain

We address Items 1 through 6 of these needs in detail in following sections. Items 7 and 8 are covered in their respective chapters and are not part of this chapter.

3.1 On-package and Off-Package Interconnections

It is useful to note that in high-end SiPs that target the HPC and data center markets, the challenges as well as the solutions used for in-package and off-package communication links are similar. The key differences are centered on the connection widths and drive requirements.

Irrespective of other components that are integrated with memory and processing dies in a SiP, it is critical to provide high bandwidth, low latency connections between processing and memory elements. Specifically, for connecting multicore processor die with stacked memory dies, point-to-point interconnections are needed and the number of memory dies will be proportional to the number of cores on the processor die. Conservatively assuming that core counts scale by a factor of 1.4X per generation, 1.4 times as many memory dies need to be accommodated per generation in the SiP. Simultaneously, if we assume that advances in the stacked memory technologies enable twice as many data bits to be delivered per generation and assuming that the clock rate on the processor-memory link remain unchanged, the number of bit links between the multicore die and the stacked memory dies will have to grow by a factor of 2.8X with each process generation.

As an example, at 14 nm, Intel implements 1024-bit-wide bit links as EMIBs (embedded multi-die interconnection bridge) on a Silicon substrate to each HBM inside the SiP with a core count of 56. When the transition is made to hyperscaled 10 nm, the core count grows to 78 (=56 X 1.4), requiring 2048-bit wide links to each HBM and the ability to connect to 1.4 times as many HBMs. This will require finer interconnection pitches in the EMIB or other enhancements that will require additional metal layers (beyond the 4 to 6 metal layers in use now on the silicon bridge) and additional vias in the EMIBs or alternative on-package interconnection techniques. In general, the on-

chip interconnection problem may be exacerbated when dies integrating general-purpose cores and accelerators are integrated with other components, as off-die connections may be grossly limited by the physical dimensions of the dies, requiring reduced pitches in the links in the bridge to enhance the number of connections in-between adjacent dies.

3D integration can also be a promising solution if thermal, yield and reliability issues are addressed to permit stacking of memory dies and processor dies. This will be more realistic for stacking lower power, energy-efficient integer cores targeted to specific data center applications with DRAM memory/HBM dies.

The possible solutions for addressing these needs are as follows:

Short-term:

- **High-density wiring** (e.g., Si Interposer w/ TSV, EMIB): Currently, high-density integration in applications such as HBM is done with TSVs using silicon interposer technology. Embedded Multi-die Interconnect Bridge (EMIB) is an approach that avoids the use of TSVs and was developed by Intel to interconnect heterogeneous chips inside the package with high connection density. The industry refers to this application as 2.5D package integration. Instead of using a large Si interposer typically found in other 2.5D approaches (like TSMC's CoWoS and Unimicron's embedded interposer carrier), EMIB uses a very small silicon bridge die with multiple routing layers, but without TSVs. This bridge die is embedded as part of Intel's substrate fabrication process. With further improvement and broader applications, EMIBs will continue to play a dominant role in the near future with enhancements in the choice of organic materials, number of metal layers, and improved driver/receiver circuitry for signal integrity enhancements.
  Short-term advances will be focused on cost reduction and potentially increasing the density of interconnections within the silicon bridge. This is necessitated for accommodating wider, higher data rate parallel signal links that are needed in future HBM generations, with additional connections links needed for shielding, ground and other lines needed for maintaining signal integrity. These additional lines increase the total link count considerably at the PHY level.
  Other cost-reduced alternates will continue to develop, including interposers with fine-line laminates, glass carriers, and integration into a package with fine line layers such as iTHOP, or even multi-chip fan out technologies for the right application. The choice of technology will depend on the particular application being pursued. Breaking up a chip into smaller components will have different power demands and different interconnection bandwidth and latency than, say, an accelerator with advanced memory stacks. These high-density interconnects will play a dominant role in the near future with enhancements in the choice of organic materials, number of metal layers, and improved driver/receiver circuitry for reduced power. For the example of 140 Watts available for power, the power allocated for the Tx/Rx circuits needs to be in the single digits.

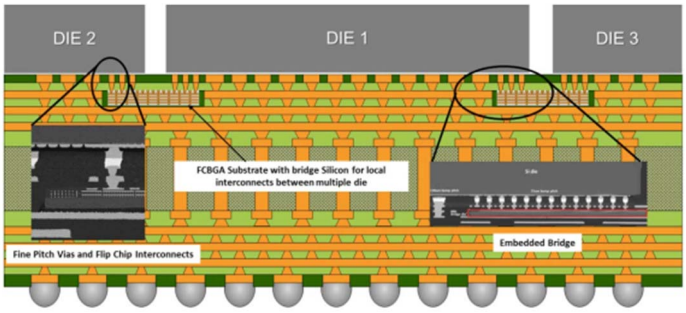


Figure 10. Embedded multi-die interconnect bridge (from [Mahajan 2018])

- **High-density organic substrate**: By combining with thin-film processes, high-density flip-chip packaging is emerging as a potential heterogeneous integration carrier. 8/8µm line/spacing and <50µm via pitch will be commercially available at reasonably low cost by the end of 2018. Various solutions are proposed, and there will be multiple options to choose as a substitution to silicon interposer and/or EMIB-like hybrid. Even though there are still gaps, particularly line width/spacing, compared to silicon technologies, organic substrates are much easier for designs. On the other hand, fine lines may cause

RC delays due to high line resistance, as for silicon chips, and therefore there is an optimal line width number, which is roughly between 2-3µm. Cost must also be considered. Although theoretically an organic substrate should be considerably lower cost than a TSV-based Si interposer, this is currently not the case. Due to advanced techniques required by the substrate suppliers to achieve 2-3µm L/S in the build-up layers, their cost is high enough that the technology is prohibitively expensive today. Further investment in substrate manufacturing capability and tooling will be required to reduce cost and drive adoption of this promising technology.

- **Denser and improved vias**: There are two aspects to via design: high-density and high-frequency. To facilitate high-density wiring, a higher density via is required. Currently high-density wiring carries data signal rates in the 1-2 Gbps range. The capacitance loading is important, but the characteristic impedance of the vias is not as critical as it is for the high-speed signals at 25 Gbps and above. For high-density signal design, a significant focus is on ground rules to maximize the signal wire density such as via diameters and capture pads. For high-frequency signals, the balance of inductance and capacitance (characteristic impedance) and the isolation of signals (crosstalk) become essential to defining the wiring rules. In addition, for high-speed signals, the material properties resulting in high-frequency dielectric and conductor loss become critical. Materials with low loss tangents and conductors with smooth surfaces are critical to properly operating packages.

3D integration is also an upcoming technology that will allow for higher levels of functional integration in a package, as well as bridging the gap between traditional fab node scaling (which is getting longer and longer for each subsequent generation). For the HPC market in particular, the use of 3D integration to bring large amounts of memory close to processing cores allows for significant advantages in compute speeds and power consumption. As development cycles continue over time, the industry is providing solutions to overcome barriers to widespread adoption of 3D integration in advanced logic devices (i.e. reliability, cost/yield, and thermal dissipation).

Memory Integration

For both HPC and data center applications, low access time and high access bandwidth are critical for meeting the demands of the processors. In SiPs that incorporate a diversity of processing dies (such as combinations of multicore CPUs, GPUs, accelerators, etc.), buffers used for facilitating the asynchronous operations of these elements are critical, as a common clock domain may not be preferred. These buffers often have to be implemented as SRAMs in the interest of facilitating fast access rates and data transfer bandwidth. Early SiPs targeting the HPC and data center segments have incorporated HBMs to work under the constraint of limited IO connections on the package and to reduce the power consumption that would otherwise exist had memory been located outside the package. However, HBMs inside the package may not be able to address the speed disparity between the DRAM-based HBM stack and the processing elements located on one or more dies inside the package. This is precisely where a SRAM layer under the HBM can serve as a L4 cache for the processing elements in a 2.5D or 3D configuration. Going beyond this, SRAM stacks can begin to take over some of the role played by the HBMs inside a SiP and can even be stacked over dies with processing elements in a 3D configuration (Figures 11 and 12). The feasibility of stacking SRAM dies, where thinning the SRAM dies has little impact on the minimum supply voltage and read/write times, has been demonstrated [McC 2018], paving the way for using stacked SRAMs in the near future. Compared to stacking

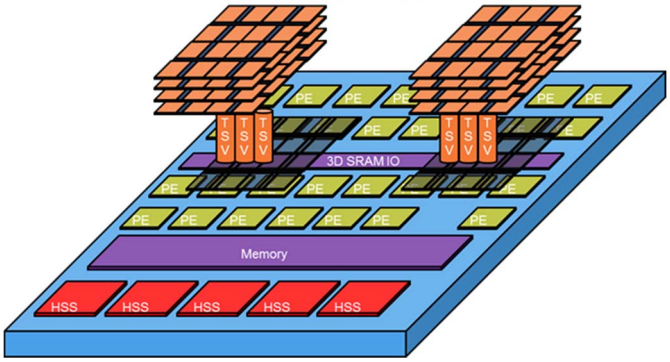


Figure 11. Integration of stacked SRAM

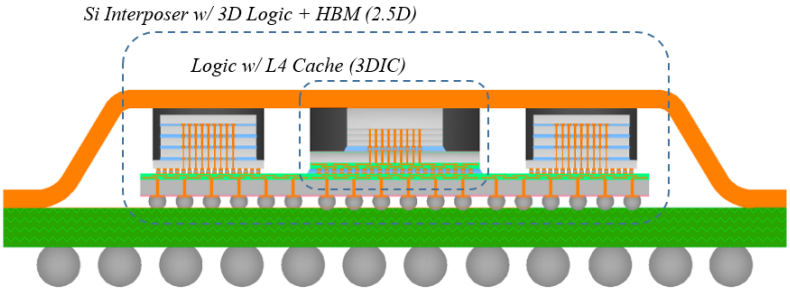


Figure 12. Diagram highlighting how supplemental SRAM cache is likely to be implemented alongside HBM in a 2.5D package. Source: [Eng 17].

DRAM dies on top of a logic die, SRAMs have another advantage – unlike DRAMs, whose data retention properties (operating voltage range, refresh rates, etc.) are affected by high temperature, SRAM dies have the same operating temperature range as the logic die and offer significantly improved retention capabilities. Stacked SRAM elements can be supplemented with HBM dies in higher-end SiPs where the lower capacity of the SRAM dies, relative to DRAM dies, may be insufficient for meeting the storage capacity needs inside the package.

The need for low-latency and high-bandwidth applications is particularly pronounced in systems for processing big data and AI/machine learning applications. For instance, in Graphcore's prototype product, designed specifically for high-performance, energy-efficient graph processing with applications to big data and machine learning, processing elements on the die are traded off for high-speed SRAM to hold subgraph data or node weights [Gra 17, Arv 18]. An SRAM die stacked on a processing element die to permit short, fast and high-bandwidth, low-power links to the SRAM die can permit larger graphs to be handled and thus facilitate scaling. High-end networking chips can also benefit from the use of stacked SRAM to permit line-rate processing demands to be met.

Depending on the end application, the following types of memory integration can be expected in SiPs:

- **HBM-on-Logic**: Several advantages can be gained by placing HBM directly on a logic die. For instance, by simply moving the HBM closer to the logic, a bandwidth gain from 2.0Gb/s to >2.7Gb/s is expected (in the case of HBM2). This is because for 3D stacking, the distance can be considered as ~50µm (the thickness of the die), whereas the wiring length on a Si interposer in a 2.5D configuration is ~2mm from the logic PHY to the HBM I/O block. When taking power consumption into consideration, we can achieve a 97% reduction in pin capacitance with these wiring length differences, which results in a substantial overall power reduction. In addition, cost calculations have shown a ~20% cost reduction by moving from 2.5D to a 3D configuration for HBM integration due to the elimination of the interposer as a cost component [Eng 17b].
- **L4 Cache**: The addition of a standalone memory die (i.e. SRAM or eDRAM) stacked on a logic die in 3D fashion allows us for easier implementation of a L4 cache. The addition of a L4 cache has been previously demonstrated first by IBM starting with their Power6 servers and later by Intel in more mainstream products with their Haswell processors, both using a separate eDRAM die as an L4 cache. These eDRAM dies are packaged in a multi-chip module (MCM) format, with die-to-die wiring occurring through the package substrate. The use of TSV technology and 3D packaging will allow the L4 cache to be placed directly on the backside of a logic die, providing both power and performance improvements when compared to the MCM option.
- **SRAM Partitioning**: As design capability improves and die partitioning becomes more mainstream, we will see SRAM being partitioned from the logic die and stacked on the logic die as a separate die instance. Since many high-performance logic designs are roughly 50% SRAM cache in the layout, it is a relatively easy piece of the design to break out. The effect of a standalone SRAM is L3 cache performance with the ability to stack multiple SRAM dies for an extended capacity. In addition, for ultra-large dies that are near reticle size, a significant fab yield improvement can be had by halving the die area and fabricating the logic and SRAM components on separate wafers to be combined later using 3D stacking.
- **Non-Volatile Memory**: Applications that require non-volatility are expected to utilize MRAM or a similar high-density memory technology. These can be stacked on a logic die independently, or combined with other memory types for a multi-stack option.

As scaling becomes more difficult, the use of the z-axis to create true 3D SoC devices is inevitable. 3D SoC designs are created by combining two separate wafers together using wafer bonding technology (typically in a Face-to-Face configuration, where the BEOL of each wafer is bonded together). First implementations of this technology are expected to utilize design partitioning at the IP block level. This allows for designs to progress with few changes to the fab technology process design kit (PDK) other than the addition of wafer bond layers, and also enables the re-use of existing IP block designs. Power and performance benefits are gained by placing key IP blocks that communicate with each other on top of each other in the overall die layout, enabled by the z-axis wiring and preventing long lateral wiring connections between IP blocks that also typically require multiple repeaters, resulting in significant voltage droop. A side benefit of the power/performance gains by die partitioning are potential lowered cost of the entire system. As the long lateral wiring is replaced with short z-axis wiring, we can eliminate some levels of BEOL wiring that are needed on a standard 2D design. Obviously, the amount of gain here is application and design dependent. In the longer term, monolithic integration for 3D SoC designs is expected. This is defined by the

fabrication of p- and n-type devices on separate wafers, using wafer bonding to create a vertical FEOL system, with a single BEOL stack. This technology is under development now at the consortia level, but will require significant changes to foundry technology PDKs to enable implementation of the technology for real products.

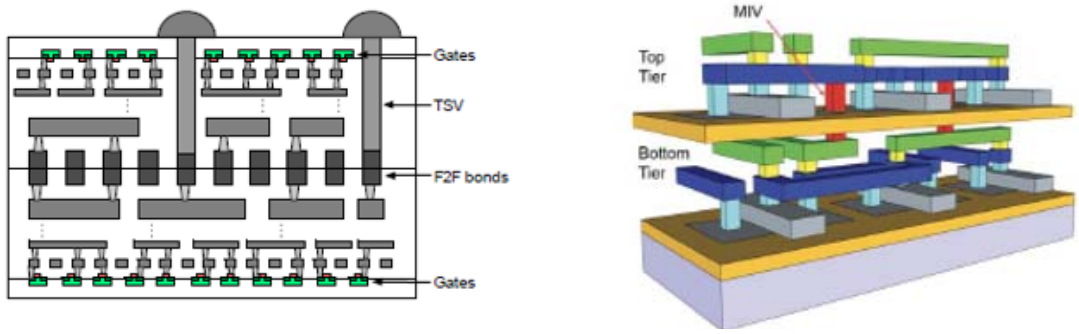


Figure 13. IP block partitioning for 3D integration (left) and monolithic 3D integration (right)

The utilization of process core tiling to create a scalable compute system provides several system level advantages, including the ability for more efficient distribution of power utilization and subsequent thermal loading. By breaking up a large, multi-core processor unit into smaller homogeneous dies (giving significantly higher yield overall), this allows us to utilize one processor design for multiple applications. As an example, a single-core unit could be used by itself for a low-performance device (such as an IoT product), whereas multiple instances of the single core unit could be combined at the package level to provide enhanced capability for high-performance computing needs. This scaling can be done using two methods: 1) 3D packaging, as in Figure 13; or 2) multi-chip modules where the dies are placed side-by-side to scale laterally instead of vertically as in the 3D case, as shown in Figure 14.

Expected interconnect solutions to support these technologies are as follows:

Short-term:

- **Ultra Large Interposers**: Large 2.5D packages containing 4 HBM memory stacks can currently be considered mainstream technology. In many cases, these products have a near-reticle-size logic die. In order to accommodate 4 HBM (2 HBM on two opposing sides of the logic), the Si interposer must be fabricated with reticle stitching on one axis. As the need for additional memory grows for high performance applications, end products will require up to 8 HBM stacks, driving two-axis reticle stitching and ultra-large interposer sizes. In addition, scaled processors can be placed in a tiled configuration across an ultra large interposer with fine pitch wiring connecting the processor tiles, creating a scalable computing system.

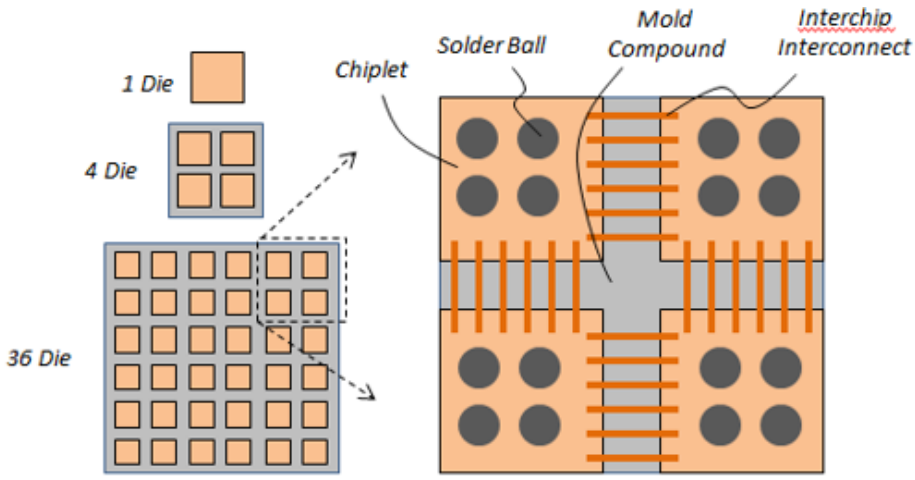


Figure 14. Scalable computing packaging concept using HDFO packaging. The Inter-chiplet interconnect utilizes minimum feature size RDL wiring. [Eng 19]

- **Embedded Die Bridges in Substrate**: Die bridges embedded in substrates will likely play a large role in the near future for high-performance computing applications. Options for embedded die bridges include both Si and glass. Since these embedded bridges enable dual connection of standard Cu Pillar and Micropillar to an organic laminate, the obvious advantage is the elimination of a Si interposer. As die placement accuracy of embedded bridges improves, and site-to-site overlay of Micropillar landing pad sites can be maintained across multiple bridges, this technology is likely to be an alternative to traditional 2.5D packaging, especially for products using 8 HBM stacks where two axis interposer stitching would be required.
- **TSV Scaling**: Si interposers are currently utilizing a TSV size of 10 X 100µm (diameter X depth) in high volume production, with 6x50µm or 5x50µm being utilized for DRAM or logic stacking. As wafer bonding and true 3D packaging is qualified and becomes mainstream, TSV dimensional scaling will be required to facilitate fine-pitch wiring between die levels. In order to maintain good manufacturability of the TSV structures, we will likely see a 10:1 aspect ratio being maintained for scaled TSV sizes. For example, GLOBALFOUNDRIES has highlighted a 2 X 20µm TSV [Eng 17] for use in stacking multiple wafers with hybrid bond technology.
- **Face-to-Face Stacking**: Face-to-face (FTF) integration will also emerge as a viable approach to stacking dies. With FTF integration, the bottom of the top die and the top of the bottom die have microbumps at locations where vertical connections are needed. Solder material is used to connect corresponding microbumps. FTF integration has been used for lower-powered products (Intel Foveros [Ren 19]), but the technique can be used for higher end products targeting the HPC/data center markets.
- **Hybrid Bonding**: Hybrid bonding in W2W or D2W format is emerging as a critical technology to enable true 3D products. Hybrid bonding is defined as an inorganic dielectric-to-dielectric bond of opposing dies or wafers, where Cu pads are planarized with the inorganic dielectric, and make contact upon bonding. Following a thermal anneal, the opposing Cu pads form a single pad across the interface. The use of hybrid bonding for F2F connections is highly advantageous because it allows for sub-micron pad and pitch sizes, providing a true connection at the BEOL wiring scale, enabling the IP block partitioned and monolithic 3D designs as described previously. This technology is already becoming mainstream for high end CMOS image sensors with integrated memory and logic, and is expected to soon be released for high performance computing devices.

Slightly longer term:

- **Ultra-Fine Pitch Substrates** (i.e. 2.1D): Although ultra-fine pitch substrates (1-2µm line/space) have been under development by several suppliers in recent years for replacement of Si interposers, they have not become mainstream yet due to reliability challenges. It is expected that these issues will be worked out over time as new materials become available. Even so, cost will likely be a challenge for widespread adoption. In order to achieve the ultra-fine feature sizes in the upper levels of the substrate, advanced techniques are required that drive the cost above and beyond that of standard substrate wiring levels.
- **Wireless Die-to-Die Communication**: In a multi-high wafer/die stack, the use of inductive coupling between I/O sites on each die has the potential to replace TSVs for die-to-die communication. This has the effect of reducing wiring complexity for TSV insertion, and preventing issues with interconnect yield (i.e. hybrid bond pads). Detailed characterization will be required to understand the impact of various parameters such as substrate resistivity, die thickness, etc. on the inductive coupling and resulting I/O quality at various operating frequencies.
- **Photonics-based solutions**: These center on the use of integrated photonics or photonics chiplets within the package and are briefly discussed in Section 3.2, leaving detailed presentations to the companion chapter on the role of Photonics in heterogeneous integration.
- **3D integration on the horizon**: The development of silicon-level 3D integration for high-performance systems has been slowed down due to thermal and power delivery issues. However, package-level 3D integration is becoming a trend for performance and miniaturization, which is more viable in terms of scalability and cost.

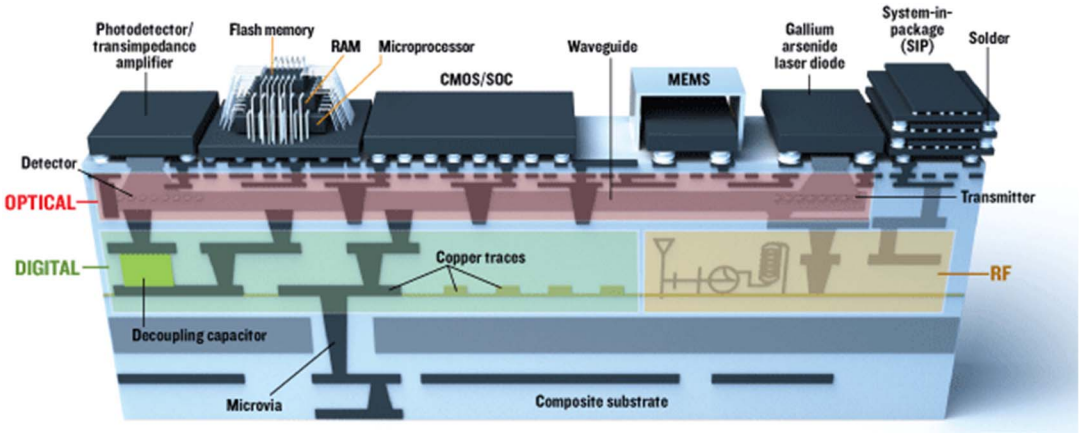


Figure 15. A System-in-a-Package (from [Tum 06)

Very similar scaling rules apply to the point-to-point interconnections between GPU dies and stacked memory dies or between special function FPGA dies and stacked memory dies.

Connections to off-package interfaces and DRAM controllers on the SiP substrate can continue to rely on the PCIe standard, and the evolution path for multi-lane PCIe have been well-defined. The implementation of alternatives to direct links based on point-to-point interconnection technologies will require multiple metal layers in the silicon substrate, and the exact topologies used are specific to the SiP architecture. Signal integrity needs on longer links in the substrate, symbol encoding and clock synchronization issues have to be addressed here. If higher speed serial links are used, the silicon-imposed limits on SERDES have to be observed. Photonics links will be a viable interconnection alternative for implementing high-data-rate, relatively longer links on the substrate, but this will require significant advances to be made for realizing low power emitters whose wavelength drifts are limited with temperature variations, as well as the design of reliable detectors.

3.2 Off-package Interconnections

As additional components are integrated within a single package, the demands on the off-chip bandwidth go up commensurately with the number of processing elements that are integrated. The newer generation of PCIe links can possibly meet these needs, but the ultimate limitation will be imposed by the package pinout. As an example, when 1.4X more cores are accommodated on a multicore die, the off-package bandwidth will need to go up commensurately. With a limit on the pin count, this need can be met by increasing the link data-rate and multiplexing multiple logical links on a single physical link. Photonics links can be an alternative to copper links, as techniques like wavelength division multiplexing can be used to implement several connections concurrently on a single photonic link, especially when extended reach of interconnect is needed because of the distance between components.

Possible Solutions

- Future-generation links (PCIe5, others)
  Package-level system integration tends to blur the line between on-package and off-package I/O. Many I/O standards are commonly used for both I/O scenarios. PCIe is one of the most popular I/O standards, and it takes over four years for each generation evolution (double data-rate). However, as PCIe Gen4 had hardly settled down in 2017, the industry already started searching solutions for PCIe Gen5, which is a clear indication of package-level system integration advancement. PCIe Gen5 is expected to carry 32Gbps per data channel without changing the TX/RX specifications. IBM and Amphenol Corporation jointly developed a new PCIe connector and demonstrated PCIe Gen5 bandwidth in early 2018, which significantly accelerates the availability of the new standard.

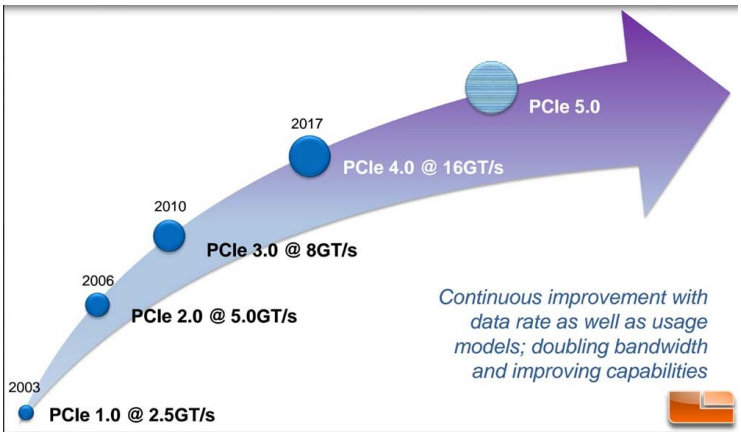


Figure 16. PCIe Roadmap (adapted from [Pir 17b])

Driven by package-level integrations, numerous proprietary I/O standards have been emerging in recent years, such as GenZ, Omni-Path, and NVLink. Most are evolving towards 32Gbps in the next couple of years. The table below shows SERDES I/O speed, distance, and channel topologies. Off-package 56Gbps data-rate is expected by 2020 with PAM4 signaling.

Parameter	USR	XSR	VSR	MR	LR
Data Rate (Gbps)	18 -58	36 -58	36 -58	36 -58	36 -58
BER	1E-15	1E-15	1E-15	1E-15	1E-15
Distance	10 mm (~0.4")	50 mm (~2")	150 mm (~6")	500 mm (~20")	886 mm (~34.9")
Interconnect	MCM	PCB+ 0 connector	PCB + 1 connector	PCB + 1 connector	PCB + 2 connectors
Insertion Loss (dB) (at f_N)	2	4 (PAM4) 8 (NRZ)	10 (PAM4) 23 (NRZ)	20 (PAM4) 36 (NRZ)	30 (PAM4)
Modulation	NRZ	NRZ or PAM4	PAM or NRZ	PAM4	PAM4
FEC	N	N	Y/N	Y	Y

Figure 17. SERDES IO trend (compiled from various sources)

These interconnection standards cover several types of channel topologies, such as backplane, daughter card, and cable-based interconnections. All of these are evolving support for higher bit-rate communication. As a result, there is development to overcome bottlenecks. Technology is being developed to provide higher bandwidth connectors, lower loss laminate materials, and smoother copper. This includes the use of twin-ax cables, optical fiber, and re-driven signals to extend the reach of the transmitters. The receiver circuits are being designed to recover signals with smaller amplitudes through a more aggressive use of equalization circuits and analog-to-digital techniques. Signaling techniques are adopting modulation techniques such as PAM4 and encoding to manage the frequency bandwidth of the signaling. In addition, techniques such as FEC (forward-error correction) become more widely used for appropriate applications. The application needs to determine the signaling based on power limits, latency impact, and interoperability needs of the system.

- Photonics will undeniably play a significant role in both enabling the use of high-end SiPs for HPC and data centers into the rest of the system as package IO solution in the near term, and ultimately to facilitate tighter integration of chiplets in high-end SiPs in the longer term [Zuf 13, Bot 17, Kri 17]. As a package IO alternative for overcoming the traditional package escape bandwidth limitations, discrete photonics chiplets are likely to be integrated with other dies and multiple photonics links, or wavelength

division multiplexing (WDM) can be used to overcome the limitations of SERDES circuitry. In the longer term, to overcome the bandwidth and latency limitations of interconnections within the package, photonics can play a significant role. Interposer waveguides and, ultimately, plasmonic interconnections appear promising in this respect. In any photonics solutions that are employed in the short or long term, cooling of the photonics components becomes a critical need as photonics components can malfunction even with slight changes in temperature. The promise of photonics and spot-cooling challenges and solutions are discussed in detail in the Photonics chapter and the crosscutting Thermal management chapter.

Finally, the limitations of copper as an interconnection material need to be addressed for SiPs targeting th HPC and data center markets, where low interconnection latency is important. The RC delays of copper can no longer be ignored as nodes shrink below 3nm [Lu 17]. Solutions to stretch the use of copper connections include annealing, use of shielding materials, and others. Alternative material solutions to replace copper are also being explored and the use of other metals, such as cobalt and ruthenium or "compounded" copper.

3.3 Signal Integrity Issues

In general, to exploit the capabilities of a SiP without IO bottlenecks, dense parallel connections need to be used on-package, and higher bandwidth off-package connections operating at very high link rates become a necessity. These certainly introduce potential signal integrity problems and they need to be dealt with adequately. Powerful error correction capability going beyond ECC will be necessary for critical on-package connections and alternative symbol encoding, and signal processing necessary for recovering data waveforms for off-chip links may well become the norm in very high-end, high-availability SiPs.

With growing data-rate, both loss and crosstalk increase significantly, and channel signal integrity can be compromised. Therefore, new materials, connectors/sockets, and via transitions are required to achieve link specifications. For dielectric materials, 3-4 times lower dielectric loss (compare with FR4, $\tan\delta=0.22$) will be widely available, combined with smooth copper foil to mitigate skin effects. To support Cu smoothness requirements required, advances in adhesion promotors for build-up dielectric to Cu adhesion will be required (i.e. CZ8401 or FlatBond) to prevent layer delamination within the organic substrate. Meanwhile, a low dielectric constant (<3.2) may help reduce within-layer channel-to-channel cross-talk. For via transitions, via-stub removal by using blind via or backdrilling is critical, and smaller a via diameter may be needed for via impedance control and cross-talk reduction. Further, signal conditioning and equalization will be widely adopted to compensate for excessive loss, ISSI, and cross-talk. For data rates beyond 50Gbps, PAM4 signaling will prevail, for much lower Nyquist frequency.

Addressing signal integrity will require advances in electrical analysis tools. As electrical signaling approaches a fundamental frequency of 25-32 GHz to provide 50+ Gbps NRZ and 100+ Gbps PAM4 data rates, the primary electromagnetic modeling tools will move to full 3D extraction for the full channel, even including traces, to enable simulation of the reflections and crosstalk for these signals. The impedance tolerance specs will need tighter bounds (½ to 2/3 the impedance tolerance of 10 Gbps signals) and the crosstalk amplitude more accurately modeled because of the lower margin on higher speed interconnects (up to 6 dB less margin than 10 Gbps signals). With these tighter tolerances, the power supply noise becomes a larger concern and improved modeling of signal amplitude and power noise will be desired. To achieve the tighter tolerances, the physical dimensions of the traces, planes, vias, and dielectrics will need to be more closely controlled. Machine learning using Bayesian optimization will be applied to more evaluation of channels to speed the analysis and gain insight into the design and operation of these high-speed applications.

3.4 Power Distribution

Power distribution and power quality issues become dominant as more components that operate at lower voltages (0.7 V to 1 V) are integrated. In the case of a 200 Watt package TDP where the components dissipate 70% of the package power (that is 140 Watts), the current draw from the regulated source will be up to 200 Amps. With many components drawing high levels of current that are placed at different positions on the substrate, a larger number of pins need to be devoted to power connections. Worse, inductive noise on the power connections will be significant,

affecting power quality and requiring additional decoupling capacitors. Additionally, Ohmic (that is, resistive losses) may be non-negligible, affecting the overall energy efficiency.

A potential remedy to these issues will be to incorporate local voltage regulators within the package itself as a separate integrated component or part of some high-power chiplet [DiB 10, Tie 15, Sag 18], but adequate cooling needs to be provided. Inductorless integrated switched-capacitor regulator technologies have certainly evolved and can be operated in a distributed configuration to provide point-of-load regulation. This is a strong contender as the best solution, whether it is used intra-die or intra-package [Kose 13]. Complementing these solutions, distributed point-of-load power regulators, implemented in the mainstream CMOS process technologies that enable DVFS control and have a low setting time, such as [And 14, And 17], appear to be an attractive solution at the die level. The microprocessor industry has been using distributed regulators on the die for the past few years, and SiP-level solutions extending these are thus viable for meeting the very short-term needs.

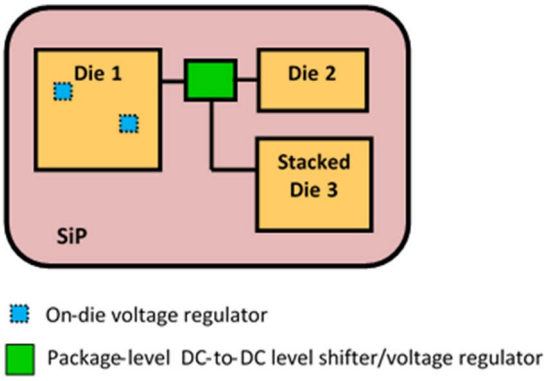


Figure 18. Power distribution and regulation inside package for SiPs [Eng 19]

For 2.5D packaging, we will continue to see adoption of localized decoupling capacitors embedded within the Si interposer or bridge. MIMCAPs integrated between BEOL layers are widely available and in use today. Deep trench (DT) capacitors are also available (Figure 19), but have been slow to adopt due to added cost of fabrication. As power and performance requirements are increasing, and DT fabrication costs are decreasing, adoption is expected to coincide with products utilizing HBM2e.

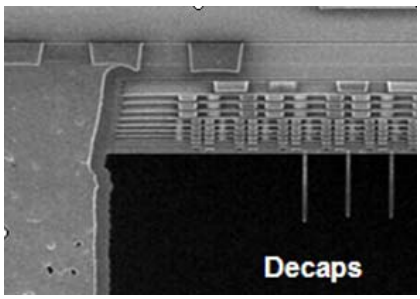


Figure 19. Active Si Interposer from that includes Deep Trench (DT) capacitors formed in the bulk Si wafer [Courtesy of Globalfoundries]

For the medium to long term, high-efficiency, high current point-of-load DC-to-DC converters that provide high conversion efficiency over a wide load range are required. This need can be satisfied with emerging power conversion devices. Converters based on new GaN power devices are likely to permeate high-end SiPs and offer improved efficiency, reliability and availability in power distribution systems for emerging and future SiPs. It is encouraging to note that GaN devices are scaling almost in parallel with CMOS devices with acceptable yield and cost, easing their deployment within SiPs within discrete regulator dies [Lid 15, Lid 16].

The use of a 48 V DC power feed to SiPs seems to be attractive, as there is an already established ecosystem for high-efficiency DC-to-DC converters from a variety of vendors. It is noteworthy that some vendors have already developed GaN-based converters for converting 48 VDC input to chip-level voltages [Lid 16] and these are likely to be the first entries into the high-end SiP market. The use of 12 VDC also seems to be a viable alternative for lower TDP packages, as they have seen some use in the Open Computing platforms, thereby establishing another supply ecosystem.

A final solution that has the potential for scaling well with SiP complexity will be to use distributed regulators within the package that operate at higher input DC voltage and regulate these down in a distributed configuration to the 1 V or less as needed [MJ 18, Wiw 17]. This solution will certainly reduce ohmic losses on the power connections, but their benefit in terms of reducing inductive noise is not clear and may not be commensurate with the reduced current draw on the power lines to the package.

When switching converters for high current loads are moved inside the package, high conversion efficiency is a requirement. Switched-capacitor converters appear attractive in this respect, as they eliminate the need for inductors. Unfortunately, switched capacitor converters have a lower conversion efficiency, so developments are needed for improving their efficiency to be comparable or possibly better than that of switching converters that use inductors. Non-linear control techniques for switching appear to be attractive in this respect, but other solutions need to be devised.

In general, the noise from switching regulators placed close to the point-of-load, as is desired for high-power SiPs, can disrupt the operations of the logic they power. However, recent products, available for downconverting from 48 VDC to chip-level voltages designed for close placement to the package, use two-stage conversion and lower

switching frequency to reduce switching noise [Xin 17, MJ 18]. Similar techniques, as well as filtering circuitry, can be used to reduce switching noise further and enable use of 48 VDC to chip level voltage converters inside a SiP package, close to the die that they power. To enable noise suppression circuitry to fit inside the package, innovations are needed to provide inductors with small area and height (such as thin film magnetic core inductors [Fer 19], thin film inductors on glass substrates using $Ni_{45}Fe_{55}$ and $Co_{80}P_{20}$ magnetic materials [Laf 18]) as well as ultracapacitors with similar properties. Realizing these passive components will probably require significant advances in materials used for these components.

A related and added requirement that has security implications is the need for electromagnetic shielding around the load and in-package converter for avoiding side channels based on EM signal monitoring, further explored in chapter 19 on Security.

The use of high-power converters within a SiP introduces a thermal challenge. For example, when a converter with a 95% efficiency is used to power a 200 Watt load inside the chip, the switching converter will dissipate 10 Watts of power, most of it within the power device with a small footprint, creating a hot spot that needs to be cooled aggressively. One solution for dealing with this may be to distribute the hot spot using one or more converters adjacent to each of the high-power dies inside the SiP to provide power to these dies. The problem of hot spots centered around power devices inside the package exacerbates the thermal challenge for SiPs using stacked dies, requiring aggressive cooling solutions and, where needed, thermal shielding.

3.5 Global Power and Thermal Management

The various components integrated onto a single substrate in a SiP can each have their own power management strategy. A global power management scheme is essential to synergistically manage the power dissipation of all integrated components to not only stay within the package TDP but to also address any inevitable hot spot that may result. There are several ways to implement a global power management scheme and all require the ability to sense temperature and the power dissipated within key blocks of the various dies. A dedicated controller for power management may be needed, similar to the PMU microcontrollers used in many multicore processor chips. Several power management policies are possible that use static of dynamically allocated power budgets. PMUs implementing machine learning-based global power and temperature management is also possible. This is an open area of research and may well dictate the standardization of sensor and actuator interfaces for each integrated component, including voltage regulators, inside the package.

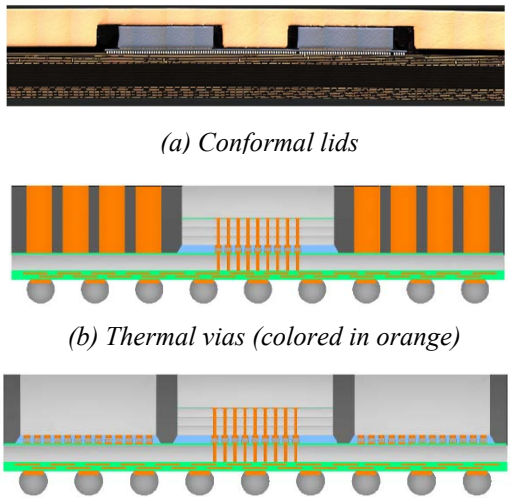
SiPs targeting the HPC/Data Center applications are going to include high-power chiplets such as multicore CPUs, GPUs, domain-specific accelerators and high data rate transceivers. Additionally, high power densities in some of these components are inevitable. Additionally, memory components, specifically stacked DRAM, are likely to be part of the SiP. The SiP cooling challenge comes from many sources:

- High power, high power density chiplets.
- Chiplets that are affected adversely by high temperatures, such as stacked DRAM.
- Chiplets that are susceptible to temperature variations, such as photonic transmitters.
- Chiplets that have different heights when integrated.

To address these challenges, cooling solutions will need to not only provide high cooling capacities, but also deal with multiple and possibly dynamic hot spots, and with heat transfer from one chiplet to another inside the package.

We envisage the short-term cooling solutions to use:

- Conformal package lids whose package side matches the height of various components inside to reduce the thermal resistance from each chiplet to the lid, as shown in Figure 20(a).
- Thermal vias in 3D stacked structures, which are useful for top die thicknesses of less than ~200um, as shown in Figure 20(b). Alternatively, for thicker top die, dummy Si dies stacked with micropillars can be utilized as a direct drop-in to a molded D2W process flow, as depicted in Figure 20(c).



(a) Conformal lids

(b) Thermal vias (colored in orange)

(c) Dummy die with micropillars (colored in orange)

Figure 20. Examples of near-term solutions to transfer heat to the top of 3D-integrated SiPs

- Internal heat shields to protect temperature-sensitive components where needed.
- Conventional forced-air cooling for package TDPs up to 200 Watts and the use of heatpipes at higher power levels.
- Use of coldplates circulating warm water (water at close to ambient temperature) or chilled water circulated through coldplates that replace traditional heatsinks to cool the package to handle package TDP up to 300 to 400 Watts.
- Liquid impingement cooling through 3D printed lids, which contain a liquid delivery pattern that is customized for each individual product.
- Use of thermal vias in 3D stacked structures.

In the longer term, new cooling technologies need to come into play, particularly to handle 3D-integrated SiPs. These include the use of the following:

- Package lids with microchannels to support water cooling or evaporative cooling.
- Advanced thermal interface materials.
- Aggressive single and two-phase liquid cooling solutions to handle TDPs to 800 Watts.
- Heat spreading layers in-between chiplets in a 3D configuration.
- Dense thermal vias or wide pillars to remove heat from stacked chiplets.
- Immersion cooling systems.

These and other solutions are detailed in a companion chapter.

3.6 Security and Reliability Issues

As the industry moves towards the use of heterogeneous integration as a mainstream technology for the HPC and data center markets, it will become necessary to integrate components from a variety of vendors. The integration system will have to proactively address security and reliability issues that may be critical when the sourcing of components to be integrated addresses a broader supplier base. Reverse-engineering of a SiP is also a concern, as the components and interconnections are readily accessible on opening the package. Trojans and other security flaws that are present in a component that was introduced maliciously or resulted from design flaws can potentially jeopardize the operation of a SiP.

The security issue can be addressed in hardware or software, or using a combination of both. From a purely software perspective, test tools need to be augmented to detect Trojans. Hardware solutions, such as a programmable Security Management Unit (SMU), can also be a component that needs to be integrated into the SiPs that address critical application for continuous monitoring and possible isolation from the rest of the system within the package. Tamper-resistant solutions at the package or at the die level (such as die-internal fuses) can be incorporated to act as a barrier against reverse engineering.

Reliability issues parallel the security issues mentioned above in terms of how system availability is affected. Expensive SiPs targeting the HPC and data center segments need to have the ability to degrade gracefully on failure. Thus, facilities that detect the failure of individual components and isolate failed components from others to enable at least partial functionality where possible may become a necessary part of the SiP design process. In fact, this facility can well be integrated into the SMU, which can also use the isolation capability to isolate rogue components.

The problem of side channels based on the EM signals for a high-powered die also needs to be addressed, requiring electromagnetic shielding (see Section 3.4) or active solutions.

Finally, to support both reliable and secure operations, adequate sensing capabilities including at least the ability to monitor the interconnections among components on the SiP substrate is necessary. It is probably useful to standardize such sensors and their interfaces for widespread deployment.

4. Chiplet Standards for Heterogeneous Integration Targeting HPC and Data Centers

The vision of heterogeneous integration is to overcome scaling, performance and cost barriers of single chip solutions, as well as to facilitate rapid product development. Specifically, this calls for the ability to integrate chiplets from different vendors and different technologies. To facilitate this goal, the industry has been converging for a while at devising PHY layer interconnection standards for communicating across chiplets (like different generations of PCIe, the HBM interface). While these standards provide a rapid transition path towards integration, they are inadequate at meeting all of the needs of the higher-level layers for heterogeneously integrated products targeting the HPC and data center markets. Integrators can benefit significantly from IP/standards at the subsystem to facilitate rapid development. These standards apply not just to the communication protocols but should extend to facilities

that provide a consistent view of memory shared by chiplets for general-purpose processing, IO logic and domain-specific accelerators. The Open Domain-Specific Architecture (OSDA) is an industry forum that has been formed to facilitate standardization of die-to-die interfaces. These include communication protocols at multiple higher layers beyond the physical layer, and associated power management mechanisms and support for memory coherence across chiplets. Some of the emerging standards include the following:

- **Compute Express Link** (CXL): this standard, built on top of the PCIe physical and electrical interfaces, represents an interconnection standard among the CPU, memory, IO and accelerator packages as well as chiplets, while providing memory coherence [CXL 19, Das 19].
- **Short Range SerDes** (XSR): this standard accommodates short connections, up to 50 mm.
- **Ultra Short Reach** (USR): Examples of interfaces available include Kandou Glasswing, Aquantia AQlink, and Nvidia NVLink.
- **High Speed Parallel Interface** (HSPI): The usage of a massively wide bus with many wires in parallel to move data between chips without the complexity of an advanced SerDes or USR design is becoming a reality due to new enabling advanced packaging technologies. Due to the inherently more simplified design, this interface type is easier to implement across chiplets from different fab technologies. One example of this type of parallel interface is Intel AIB.

These standards facilitate integration and reduce design and verification costs. It is likely that other standards will emerge in the future. Other emerging standards are likely to include active interposer standards that incorporate coherence support as part of the embedded interconnection fabric.

Another promising approach is to develop interconnection standards, including interfaces that can be embedded in the interposer, effectively resulting in what can be called an "active interposer". An example of a proposal that incorporates support for memory coherence is presented in [Yin 18].

As the industry embraces heterogeneous integration for the HPC and data center markets on a wider scale, standards are also expected to emerge for power distribution, testing and system-level power management. At this time, these standards do not exist.

5. Applicable Tracking Metrics

The formulation of system-level metrics for tracking the emerging needs for SiPs that target the HPC and data center market is challenging for two reasons. First, the applicable metrics depend on the types of dies that are integrated within the SiP. This impacts the nature of the interconnections inside the package, off-package IO needs, overall system power, power conversion and distribution and other attributes. The second factor that complicates the formulation of tracking metrics has to do with the area of the interposer. The bigger the interposer is, the higher is the number of components that can be put inside the package. Unfortunately, the generational limit of the interposer size is almost impossible to ascertain, as it is vendor-specific (and possibly of a proprietary and competitive nature), depends on the interposer material and on the components that are placed on the interposer for integration.

An alternative to the use of system-level metrics will be to use lower-level metrics such as:

- Generational needs of interconnections between specific types of components (such as those between HBM and general-purpose cores, HBM and GPU dies, general-purpose cores and accelerator, etc.)
- Package power limits based on the type of cooling used (passive air, passive water via coldplates, two-phase cooling, etc.)
- Other component-level metrics directly representative of some aspect of performance provided by the SiP.

At this time, the metrics proposed for use are represented in the following tables; some metrics are still not quantified. These metrics will be reformulated based on discussions with other TWGs and domain-specific experts beyond the TWGs.

Broader Issue	Specific Needs, Potential Solutions: 5-Year Horizon	Specific Needs, Potential Solutions: 10-Year Horizon
Logic Integration: Processor/Logic Subsystem/ Accelerator Integration	Tightly-coupled 2D tiled configuration; Wide-lane connection among adjacent dies • Silicon/EMIB bridges to support inter-tile connections up to 4000 lanes at low latency • ECC + Symbol encoding on links • Large interposers realized with reticle stitching • Locally-synchronous, globally-asynchronous clock	Tightly-coupled 2D tiled and 3D configuration (stacking) • Dense vias implementing multiple 1000+ lane links • Up to 8000 low-latency lanes for 2D tiling • Plasmonic interconnections
Logic/DRAM Integration	• Interconnections to support up to 4000 lanes/256GB per sec. per HBM to accommodate HBM2 thru HBM 3 • Silicon/EMIB bridge • Alternative imposers • DRAM stack on logic/SRAM layer implementing memory acceleration artifacts and L4-Cache • Advanced symbol encoding on links + ECC	• Interconnections to support HBM 3 and beyond with 4000-8000 lanes/ >512 GB/sec. per HBM; DRAM stacked over processing elements; SRAM stack on top of processing die; • SRAM, DRAM stacks with processing element die(s) • Dense vias implementing multiple 1000 lanes • Combination of 2.5D and 3D subsystems • Photonic/Plasmonic interconnections
Logic/Memory Integration	• SRAM stacks or STT-MRAM at edge with lane counts similar to DRAM; • SRAM stack serving as cache for DRAM stack, external memory; Stacked SRAM for use by FPGA engines etc. • Silicon/EMIB bridge • Alternative interposers	• SRAM stack on top of processing die; Distributed SRAM for supporting big data/ML/AI applications; Logic close to or embedded with STT_MRAM and SRAM/DRAM stacks to support memory-centric computing; SRAM, DRAM stacks with processing element die(s) • Dense vias implementing 4000+ lanes with low latency • Combination of 2.5D and 3D subsystems • Limited photonic connections/optical vias to SRAM stack
Logic/NVRAM Integration	• Needs and solutions parallel those for SRAM/DRAM depending on type of NVM	• Needs and solutions parallel those for SRAM/DRAM depending on type of NVM
Package IO	• High bandwidth wide-lane IO channels; • Limited use of optical transceivers on high BW IO links using integrated photonics TXRX die(s) • Aggressive signal equalization • Advanced symbol encoding • Integrated photonics component with high thermal immunity • Limited number (2 to 16) of wavelengths on WDM	• Multiple high-BW IO channels • AJ/bit photonic links • Advanced symbol encoding • Dense WDM • Advanced copper IO
SiP-level Power Management	• Fine-grained DVFS control of processing, memory elements, package IO interface • Rich sensing • Distributed power management in package at common voltage islands and within die • Application-specified power budgets for subsystems within SiP	• Advanced power-management of SiP • Power-budget assigned dynamically with ML controller with fully-decentralized power controllers • Tight integration with in-package converters • Potential coordination with active package cooling system • Thermally-aware load shifting/distribution inside package

Broader Issue	Specific Needs, Potential Solutions: 5-Year Horizon	Specific Needs, Potential Solutions: 10-Year Horizon
Power Delivery to Package and Distribution Inside Package	• Support for multiple voltage islands; • Reduction of Ohmic losses in power delivery network – power limited to 200 to 250 W per package; • Noise reduction in power delivery system; • Use of high-voltage (48 VDC) to package with few DC-to-DC converters inside package • Use of noise reduction techniques based mainly on passive components • Advanced inductors and capacitors • Advanced DC-to-DC converter designs: switched capacitor converter with non-linear control and GaN power device(s), 2-stage conversion	• Reduction of Ohmic losses in power delivery network – up to 250 to 800 W per package; • Noise reduction in power delivery system; • Active coordination with power management system • Use of high-voltage (48 VDC) to package with more distributed DC-to-DC converters inside package • Use of active noise reduction techniques • Advanced switched capacitor DC-to-DC converter design using GaN power devices; 2-stage conversion; lower-frequency switching for noise reduction • Advanced solutions for mitigating side channels based on power line noise and other EMF emissions
Security Needs	• IP Protection against reverse engineering/tampering; • Potential information leakage via interconnection probing in opened package; • Tamper-proof package; self-destruction fuses; • Certified supply chain • Limited forms of link-encoding • EM shielding of radiating components inside package to mitigate side channels • Static and run-time testing	• Protection against compromised dies; • Active side channel mitigation techniques • Full-blown security management co-processor monitoring SiP • Isolation of compromised dies (also used for isolating faulty dies) • Active protection against power viruses • Aggressive run-time test/diagnosis/repair
Package Cooling	• Up to 250 W heat removal with air/warm water coldplate cooling; • Ability to deal with hot spots near power-conversion devices/specific logic/optical TXRX etc. • Potential need for thermal shielding • Conformal lids • Thermal vias • Dummy dies with micropillars • Coldplates • Other Potential Solutions: TBD, including possibly very limited use of immersion cooling in niche market segments	• Up to 800 W heat removal with advanced cooling solutions; • Potential need for significant thermal shielding; • Coordination with SiP power management system; • Heatpipes/dense thermal vias for stacked SiPs • Inter-layer cooling • Widespread use of liquid/2-phase cooling • Other Potential Solutions: TBD, including immersion cooling
Others: Alternative processing paradigms (Quantum, neuro-morphic etc.), analog component integration	• WORK IN PROGRESS – will appear in the next version of this chapter	• WORK IN PROGRESS – will appear in the next version of this chapter

High Performance Computing and Data Centers TWG Team and Contributors

Dale Becker	William Bottoms	William Chen
Luke England	Kanad Ghose	Rockwell Hsu
Ravi Mahajan	Lei Shan	

Selected References

[And 14] Andersen, T. M. et al, "A Sub-ns Response On-Chip Switched-Capacitor DC-DC Voltage Regulator Delivering 3.7W/mm2 at 90% Efficiency Using Deep-Trench Capacitors in 32nm SOI CMOS" in Proc. ISSCC 2014.

[And 17] Andersen, T. M. et al, "A 10 W On-Chip Switched Capacitor Voltage Regulator with Feedforward Regulation Capability for Granular Microprocessor Power Delivery", IEEE Transactions on Power Electronics, vol. 32, no. 1, pp. 378-393, Jan. 2017.

[Arv 18] Arsovski, I., "Predictions for the future of Artificial Intelligence", presentation as part of the 2018 ECTC forum, 68th Electronic Components and Technology Conference, June 2018.

[Bohr 17] Bohr, M., "Moore's Law Leadership", presentation at the Intel Technology and Manufacturing Day, 2017.

[Bot 18] Bottoms, B., "Heterogeneous Integration Roadmap & Photonics", presentation at Confab 2018.

[Chu 17] Chung, E. et al, "Accelerating Persistent Neural Networks at Datacenter Scale", in Proc. Hot Chips 2017.

[Cis 16] Cisco Systems Inc., "Internet-of-Things at a Glance", available at: https://www.cisco.com/c/dam/en/us/.../internet-of-things/at-a-glance-c45-731471.pdf

[Cis 19] Cisco Systems Inc., Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper, Feb. 27, 2019. Available at: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

[CXL 19] Compute Express Link Documents available at: https://www.computeexpresslink.org/

[Das 19] Das Sharma, D., "Compute Express Link" whitepaper, available at: https://docs.wixstatic.com/ugd/0c1418_d9878707bbb7427786b70c3c91d5fbd1.pdf

[Deo 17] Deo, M., "Enabling Next-Generation Platforms Using Intel's 3D System-in-Package Technology", White Paper No. WP-01251-1.5, 2017.

[DiB 10] Dibene, J. T. et al, "A 400 Amp Fully Integrated Silicon Voltage Regulator with In-die Magnetically Coupled EMbedded Inductors", presentation at APEC 2010 Meeting, Feb. 2010.

[Dem 18] Demler, M., "Mythic Multiples in a Flash: Analog In-Memory Computing Eliminates DRAM Read/Write Cycles", Microprocessor Report, Aug. 27, 2018.

[Eng 17] England, L., and Arsovski, I., "Advanced Packaging Saves the Day! - How TSV Technology Will Enable Continued Scaling", Proc. IEDM 2017.

[Eng 17b] England, L., Presentation at panel "3D Packaging: A Key Enabler for Further Integration and Performance", SEMI European 3D Summit, 2017.

[Eng 19] England, L. et al, "Advanced Packaging Drivers/Opportunities to Support Emerging Artificial Intelligence Applications," 2019 Electron Devices Technology and Manufacturing Conference (EDTM), Singapore, 2019

[Eth 17] Ethernet Alliance, IEEE 802.3 New Ethernet Applications Adhoc IEEE 802 March 14, 2017 Plenary Meeting presentation.

[Eva 11] Evans, D., "The Internet of Things - How the Next Evolution of the Internet is Changing Everything", Cisco White Paper, April 2011.

[Eve 18] Everspin Technologies Inc., "Accelerating Fintech Applications with Lossless and Ultra-Low Latency Synchronous Logging using nvNITRO", Application Note, Jan. 2018.

[Fer 19] Ferric Inc., Thin film magnetic core inductor overview at: http://www.ferricsemi.com/technology

[Fow 18] Fowers, J., "A Configurable Cloud-Scale DNN Processor for Real-Time AI", in Proc. 45-th. Int'l. Symposium on Computer Architecture, 2018.

[Gra 17] Graphcore Limited, presentation on the Graphcore IPU available at: https://www.graphcore.ai/technology, 2017.

[Gwe 18] Gwennap, L., "Gyrfalcon Shrinks AI Accelerator", Microprocessor Report, Feb. 19, 2018

[Hal 18] Halfhill, T.R., "Tachyum Targets Data Centers", Microprocessor Report, June 11, 2018.

[Hil 18] Hilson, G., "Everspin Targets Niches for MRAM", EE Times, Jan. 22, 2018.

[Hil 19a] Hilson, G., "Updated HBM Standard Geared for HPC, Networking". EE Times, Jan. 18, 2019, available at: https://www.eetimes.com/document.asp?doc_id=1334218&page_number=2

[Hil 19b] Hilson, G., "Samsung Doubles HBM Density with Flashbolt", EE Times, March 27, 2019, available at: https://www.eetimes.com/document.asp?doc_id=1334488

[Hor 18] Horwitz, L., "Technology trends in 2018: AI, IoT and conversational interfaces will redefine customer experience", available at: https://www.cisco.com/c/en/us/solutions/data-center/2018-technology-trends.html.

[HP 17] Hewlett Packard Enterprise, "Capitalizing on the Sustainable Benefits of the IoT", Business white paper, No. a00000273ENW, January 2017.

[Int 18] Intel Corpn., Information on the Foveros technology and summary of the Intel Architecture Day presentation by R.Koduri on Dec. 11, 2018, available at: https://newsroom.intel.com/news/new-intel-architectures-technologies-target-expanded-market-opportunities/#gs.ivshe8

[Int 19] Intel Corpn., "Intel Optane Technology", Jan. 13, 2019, web pages at: https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html

[JED 19] JEDEC Solid State Technology Association, High Bandwidth Memory (HBM) DRAM, updated version (Item 1797.99J.), Document JESD235B, Nov. 2018.

[Jou 17] Jou. et al, "In-Datacenter Performance Analysis of a Tensor Processing Unit", in Proc. 44-th. Int'l. Symposium on Computer Architecture, 2017.

Jun 15] Jun, H., "HBM (High Bandwidth Memory) for 2.5D", presentation at Semicon Taiwan, Sept. 2015.

[Kag 17] Kagan, M., "Networking Trends in High-Performance Computing", CIO Review, available at: https://high-performance-computing.cioreview.com/cxoinsight/networking-trends-in-highperformance-computing--nid-12770-cid-84.html.

[Kos 13] Kose, S. et al, "Active Filter-Based Hybrid On-Chip DC–DC Converter for Point-of-Load Voltage Regulation", IEEE Trans. On VLSI, 21(4), April 2013.

[Kri 17] Krishnamoorthy, A. V. et al, "From Chip to Cloud: Optical Interconnects in Engineered Systems", IEEE Jrnl. Of Lightwave Technology, 35(15), August 2017.

[Laf 18] Lafage, V. et al, "2D Magnetic Inductors for DC-DC converters on Glass Interposer", in Proc. 68-th. Electronic Components and Technology Conference (ECTC), 2018.

[Lap 18] Lapadeus, M., "Big Trouble At 3nm", Semiconductor Engineering, June 21, 2018.

[Lid 15] Lidow, A. et al, "GaN Integration for Higher DC-DC Efficiency and Power Density", EPC Inc. Application Note AN-018, 2015.

[Lid 16] Lidow, A. et al, "Getting from 48 V to Load Voltage: Improving Low Voltage DC-DC Converter Performance with GaN Transistors", APEC Tutorial, March 2016, downloadable from http://epc-co.com.

[Lu 17] Lu, L. C., "Physical Design Challenges and Innovations to Meet Power, Speed, and Area Scaling Trend", presentation at ISD 2017.

[Luc 14] Lucas, R. et al, "DOE Advanced Scientific Computing Advisory Subcommittee (ASCAC) Report: Top Ten Exascale Research Challenges", Feb. 2014, available at: https://www.osti.gov/biblio/1222713.

[Mah 16] Mahajan, R. et al, "Embedded Multi-Die Interconnect Bridge (EMIB) – A High Density, High Bandwidth Packaging Interconnect", Proc. 66-th. Electronic Components and Technology Conference (ECTC), 2016.

[McC 18] McCan, D., "Packaging and Heterogeneous Integration for HPC, AI, and Machine Learning", Presentation at SEMICon West 2018.

[MJ 18] McCauley, S. and Jiang, S.,"Google 48V Update: Flatbed and STC", presentation at OCP Summit, March 2018.

[Mel 18] Mellor, C., "Samsung preps for Z-SSD Smackdown on Intel Optane Drives", The Register, Jan. 30, 2018.

[Men 18] Menon, J., "The Rise Of Memory-Centric Architectures", Forbes Council Post, Nov. 16, 2018, available at: https://www.forbes.com/sites/forbestechcouncil/2018/11/16/the-rise-of-memory-centric-architectures/#287712ac5952

[Muj 18] Mujtaba, H., "Samsung Begins Mass Producing Fastest 18 Gbps GDDR6 Memory For High Performance Graphics Cards – Up To 24 GB of VRAM, 864 GB/s Bandwidth", in WCCFTech, Jan 18, 2018, available at: https://wccftech.com/samsung-gddr6-16gb-18gbps-mass-production-official/

[Naf 18] Naffziger, S., "CPU Performance for the Next Decade", Microprocessor Report, Sept. 24, 2018.

[Ozd 17] M. M. Ozdal *et al*., "Graph Analytics Accelerators for Cognitive Systems," in *IEEE Micro*, 37(1), Jan.-Feb. 2017.

[Pie 17] Pierson, R., "48V-to-1V Conversion – the Rebirth of Direct-to-Chip Power", EPC GaN Talk, May 26, 2017, available at: https://epc-co.com/epc/GaNTalk/Post/14229/48V-to-1V-Conversion-the-Rebirth-of-Direct-to-Chip-Power

[Pir 17] Pirzada, U., "HBM3 Memory Will Double Transfer Rates To 4 GT/s For At least Twice The Memory Bandwidth – DDR5 Design Specs Aiming To Offer Up To 2x Performance", WCCFETCH newsletter, Dec. 6, 2017, available at: https://wccftech.com/hbm3-ddr5-memory-early-specification-double-bandwidth/

[Pir 17b] Pirzada, U., "Hot Chips 2017: PCI Express 4.0 Standard Coming In 2017 But Will Be Short-lived – PCIe 5.0 Landing in 2019, in Wccftech, Aug 29, 2017, available at: https://wccftech.com/pci-express-4-0-standard-coming-in-2017-but-will-be-short-lived-pcie-5-0-landing-in-2019/

[Ram 18] Rambo, S., "GlobalFoundries stacks the chips for machine learning", RCR Wireless News, June 18, 2018, available at: https://www.rcrwireless.com/20180608/5g/globalfoundries-stacks-the-chips-for-machine-learning-tag41

[Ren 19] Renduchintala, M., (Intel) 2019 Investor Meeting presentation, May 8, 2019.

[Rus 19] Russell, J., "HPC in Life Sciences Part 1: CPU Choices, Rise of Data Lakes, Networking Challenges, and More", HPCWire, February 21, 2019.

[Sag 18] Saggini, S. et al, "High current switching capacitor converter for on-package VR," in Proc. 2018 IEEE Applied Power Electronics Conference and Exposition (APEC), 2018.

[Sai 16] Sainio, A., "NVDIMM – Changes are Here - So What's Next", presentation at In-Memory Computing Summit, 2016

[Sam 18] Samsung Corporation, "Supercharge Your Applications with Samsung High Bandwidth Memory", 2018.

[SIA 17] Semiconductor Industry Association and Semiconductor Research Corporation, "Semiconductor Research Opprtunities: An Industry Vision and Guide" March 2017.
[She 16] Shehabi,A. et al, United States Data Center Energy Usage Report, Lawrence Berkeley National Lab Report No. LBNL-1005775, June 2016.
[Sne 19] Snell, A., "The New HPC", video of presentation at the Swiss HPC Conference, available at: https://insidehpc.com/2019/04/addison-snell-presents-the-new-hpc/
[Sod 16] Sodani, A., "Knight's Landing: Second Generation Intel Phi Product", in IEEE Micro Magazine, March/April issue, 2016.
[Tie 15] K. Tien et al., "An 82%-efficient multiphase voltage-regulator 3D interposer with on-chip magnetic inductors," in Proc. 2015 Symposium on VLSI Technology (VLSI Technology), Kyoto, 2015.
[Top 19] Top 500 HPC rankings at: https://www.top500.org/
[Tum 06] Tummala, R., "Moore's Law Meets Its Match", IEEE Spectrum, 43(6), June 2006.
[Vin 19] Vinnakota, B., "ODSA: Technical Introduction", presentation, March 28, 2019, link available at: https://www.opencompute.org/wiki/Server/ODSA
[Ver 19] Verheyede, "Samsung Announces Flashbolt HBM2E: Up to 16GB and 1.64 TBps Per Stack", Tom's Hardware. March 20, 2019. Available at: https://www.tomshardware.com/news/samsung-flashbolt-hbm2e-hbm2-memory,38874.html
[Whe 18] Wheeler, B., "Marvell Doubles PAM4 PHY Density", Microprocessor Report, March 26, 2018.
[Whe 19] Wheeler, B., "Xilinx Delivers Server Acceleration", Microprocessor Report, Feb. 18, 2019.
[WikH 19] High Bandwidth Memory Wikipedia pages at: https://en.wikipedia.org/wiki/High_Bandwidth_Memory
[Wiw 17] Wiwynn Corpn., "48V: An Improved Power Delivery System for Data Centers", White paper, June 2017.
[Xin 17] Xin, L. and Jiang, S., "Google 48V Power Architecture", APEC Conference presentation, March 27th 2017.
[Yan 16] P. Yang *et al*., "Inter/intra-chip optical interconnection network: opportunities, challenges, and implementations," *in Proc. Tenth IEEE/ACM International Symposium on Networks-on-Chip (NOCS), 2016.*
[Yin 18] Yin, J. et al, "Modular Routing Design for Chiplet-based Systems", in Proc. ACM/IEEE 45th Annual International Symposium on Computer Architecture, 2018.
[Zha 17] Zhang, G., PAM4 Tutorial at DesignCon 2017.
[Zuf 13] Zuffada, M., "Vision on Silicon Photonics for Efficient Data Communications", presentation at the Photonics 21 - WG6 Workshop, April 30th, 2013.

Edited by Paul Wesling