HETEROGENEOUS INTEGRATION ROADMAP

2021 Edition

Chapter 2: High Performance Computing and Data Centers

For updates, visit 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 and tables should be re-used only with the permission of the original source.

IEEE ELECTRONICS PACKAGING SOCIETY · ELECTRON DEVICES SOCIETY · IEEE Photonics Society · ASME SETTING THE STANDARD · semi

Table of Contents

Chapter 2: High Performance Computing and Data Centers

1. Introduction: The Need for Heterogeneous Integration

Send corrections, comments and suggested updates to the TWG chair, using our HIR SmartSheet:

https://rebrand.ly/HIR-feedback

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.

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. Heterogeneous system integration realizes a SiP using multiple dies and their interconnections. 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 multicore 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.

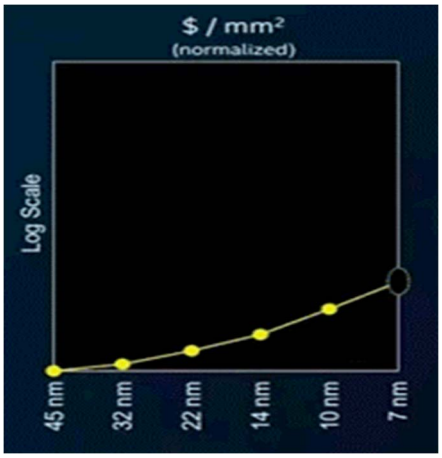


Figure 1. Die cost trends at 45 nm nodes and beyond (from [Bohr 2017])

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.

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 classical generation boundary has to be redefined as the transition between the most

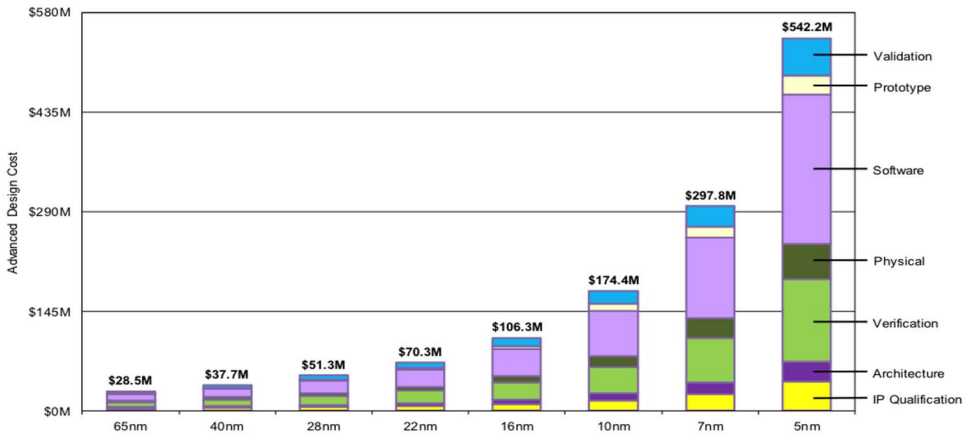
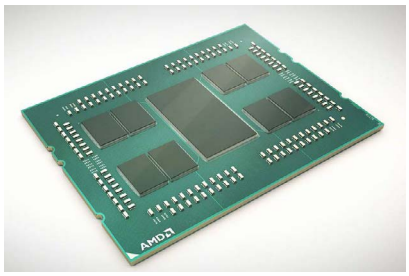


Figure 2. Chip design cost trends at different nodes (from IBS, reported in [Lap 18])

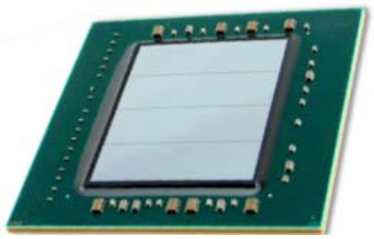
refined (that is, best hyperscaled) process technology for a given feature size and the first process implementations for the next technology node.

An implicit assumption behind Moore's Law is that the die size remains unchanged. 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. It is to be noted that the data in Figure 1 on die cost per unit area is vendor-specific and data for recent technology nodes is not readily available. Nevertheless, it is included here to illustrate a general trend that is applicable to all chip vendors. The relative cost increase at smaller nodes is affected not only by yield issues but also by other factors seen separated in Figure 2. Further, circuit elements do not scale commensurately with technology. Consequently, the performance scaling expected from Moore's Law is no longer happening as technology nodes advance.

Heterogeneous integration certainly offers a solution for performance scaling that was seen in the past as device counts increased following Moore's Law. 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 chiplets (dies) have higher yields and as long as integration cost is reasonable, the overall SiP solution will scale in performance as expected from Moore's Law. Some high-core-count CPUs and large FPGAs are actually implemented in this fashion, as 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 a higher number of parallel connections and shorter interconnections 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 beyond GDDR6 is not available at this time. Recent advances in GDDR technologies, particularly GDDR6 and GDDR6x, make them very attractive for HPC SiPs, as they provide a high IO data rate comparable to HBM2 and are cheaper compared to HBMs. GDDR6x is still not a JEDEC standard. GRRD6x replaces the NRZ encoding of IO data with PAM4 to not only get a higher data rate (2 bits per clock cycle vs. one bit per cycle with NRZ) but also improve the end-to-end transport energy per bit by as much as 15% (to about 7.2 pJ/bit) [Micron 20a, Micron 20b]. This is very close to that for HBM2, as seen in Figure 5.

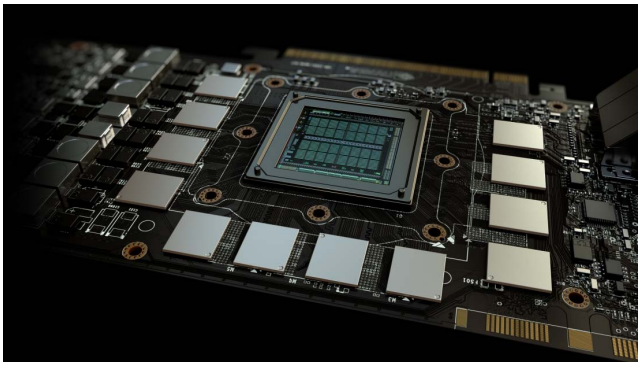


Figure 4. Nvidia Pascal showing GDDR memory surrounding the processing unit. Memory connections are through the PCB

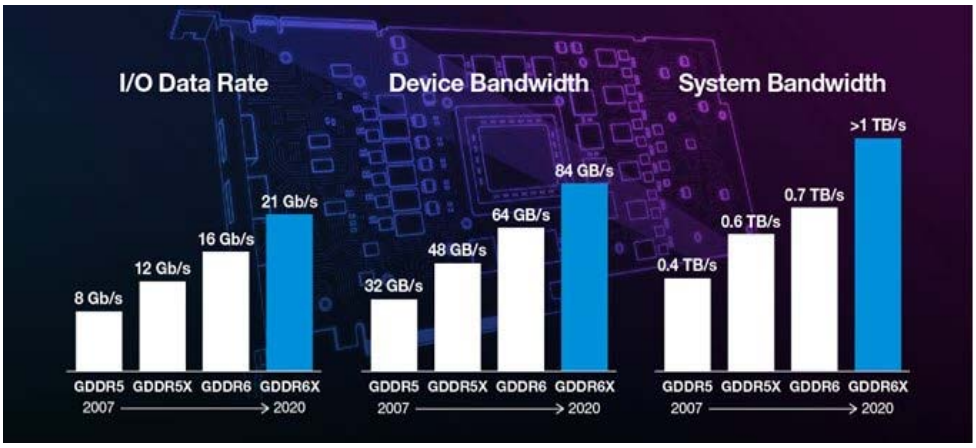
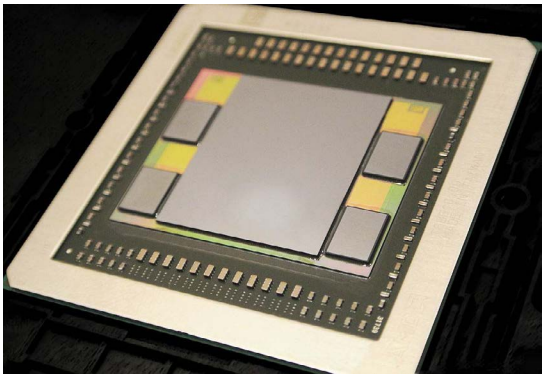


Figure 5. GDDR Specifications, highlighting the GDDR6x generation (from [Micron 20a])

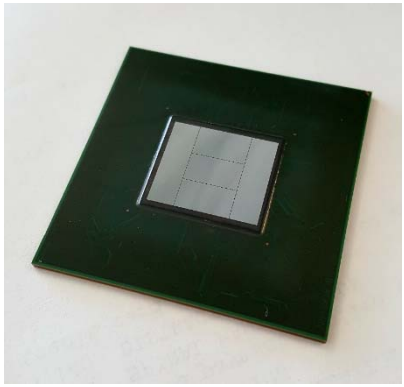
HBM: High bandwidth memory takes a different approach to providing high data rates by stacking memory dies and using wide interconnections that implement 1 or 2 channels to each die, with 128-bit wide connections to each die on the DRAM stack [Jun 15, WikH 19]. The HBM technology thus keeps the real 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, Smi 21]. As an aside, in the short-term the void left by lack of HBM2 was being met by GDDR5x and GDDR6 in some markets. HBM2e products are available from a number of vendors and JEDEC has also recently published a standard for HBM 3. Hynix reports implementations with 1024-wide connections per stack and IO rates of 6.4 Gbits/sec. per pin, with stacks of 8 to 12 chiplets in forthcoming offerings.

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 6.



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 6. 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. A 2-layer stacked SRAM cache has been announced recently by AMD.

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 7, compared against NAND-Flash and Intel Optane (formerly, 3d X-point) [Int 19]. These devices certainly appear within the storage hierarchy for a SiP.

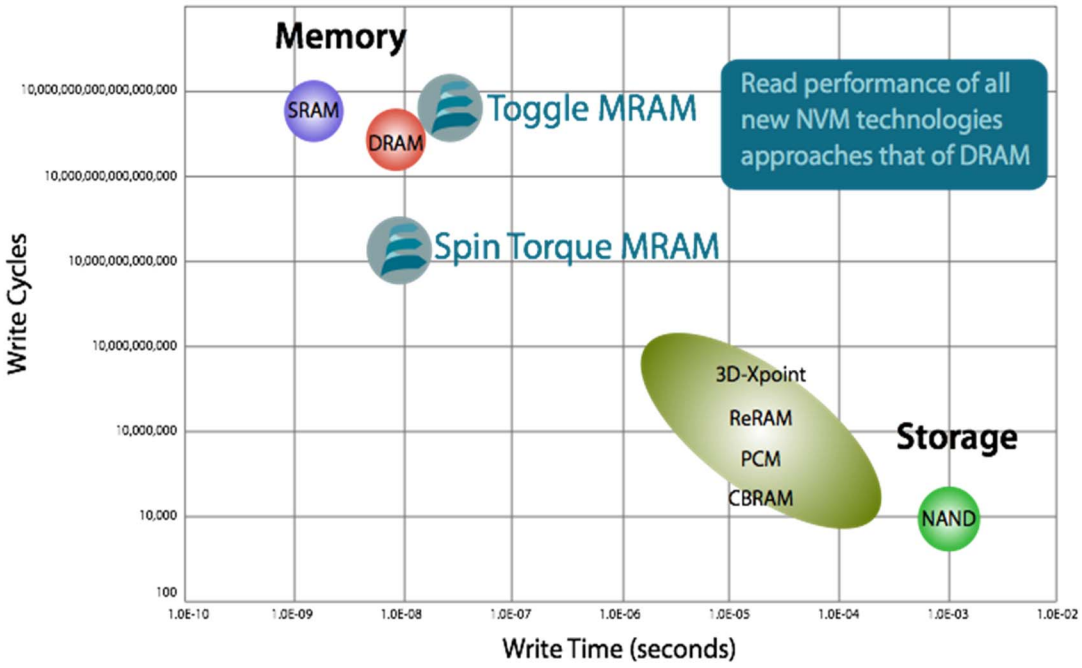


Figure 7. Emerging and existing fast non-volatile storage technologies (from [Hil 18])

The newer non-volatile memory technologies indicated in Figure 7 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.

STT-MRAMs are also attractive as a SRAM replacement, particularly for implementing Last-Level-Cache (LLC) for high-performance CPUs. SRAM scaling is slowing down and the leakage current is increasing despite the fact that the requirement for SRAM is increasing for AI and big data applications. STT-MRAM is the leading candidate for SRAM replacement because it offers significantly higher density than SRAM (Figure 8), higher endurance, zero array leakage in standby, with low static power during operations.

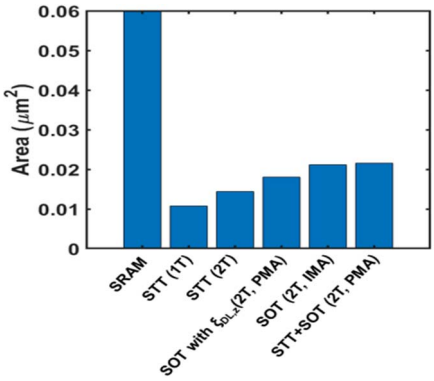


Figure 8. Area comparison of various MRAM technologies to SRAM [Kum 20]

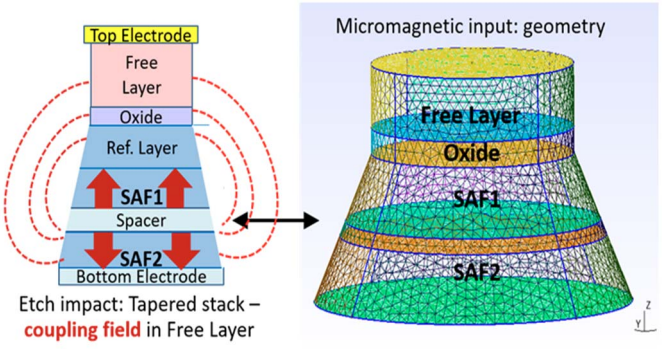


Figure 9. Tapered MTJ bit cell [Dix 20]

Developments are on-going in Magnetic Tunnel Junction (MTJ) stack design, particularly in materials optimization and etch and integration, as shown in Figure 9 (from [Dix 20]) for realizing fast switching and enhanced endurance [Dix 20]. Different MTJ stack profiles give different R-H and R-V curves. A 40 Mb macro [Lee 20] has

been demonstrated with 10ns read and write times over -40°C to 125°C with data retention of one month at 125°C. For MTJs, tight switching voltage distribution enables improved Write Error Rate (WER) with reduced back-hopping. A write time of 3 ns has been reported for a 4kb test array, where the write pulse duration is proportional to the Rmin distribution [Edw 20].

SOT-MRAM (Spin-Orbit-Torque MRAM) has several advantages over STT-MRAM for HPC but has lower density and requires higher write current. The SOT-MRAM operates in a fundamentally different way than STT-MRAM. To begin with, it has separate write and read paths, so it is a three-terminal device (Figure 10), requiring two selectors, which is the main reason for its larger area. The higher speed and improved endurance allow it to replace SRAM for L1-L3 applications, in contrast to STT-MRAM which is targeted at LLC [Gup 20]. Challenges include one order larger write current than STT-MRAM – in the range of 200-500 µA – and sensitivity to bit-line resistance. In addition, write current needs to be assisted by an in-plane magnetic field, although recent work aims to obviate this. The write current can be reduced by increasing the spin efficiency θ_{SHE}, which is the conversion ratio between charge and spin currents. The MTJ area here is 32nm. Write energy and delay vs. θ_{SHE} are shown in Figure 11 (from [Gup 20]), showing that for the technologically enhanced spin efficiency θ_{SHE} of 1.1, significant improvements in energy and delay are possible along with reductions in bit line resistance, making these devices ideal for implementing caches.

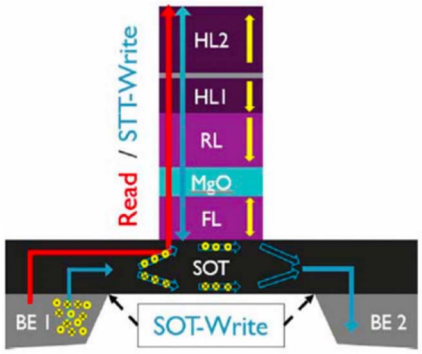


Figure 10. Separate read and write paths in a SOT-MRAM

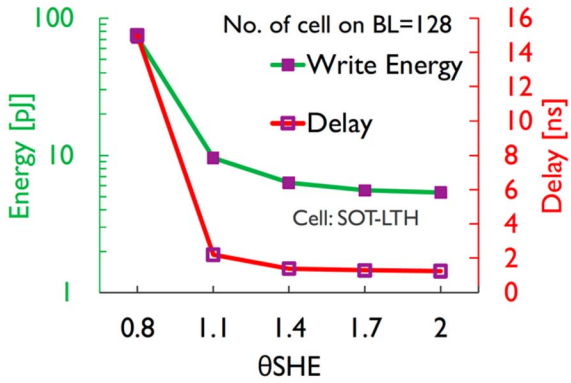


Figure 11. Energy and delay of SOT-MRAM as a function of spin efficiency, θ_{SHE}

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, with 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]. Additional examples of accelerators include those for data-intensive computing, such as bit-serial, data-parallel processors, and AI accelerators incorporating analog processing components.
- **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) in 2010 to 400Gbps by 2020 and to 6.4 Tbps (Tera bits per second) by 2030, as shown in Figure 12 (from [Eth 17]). To implement such a highspeed link 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 13 shows the relationship between package size and the total I/O bandwidth in Tbps projected by the IEEE 802.3 Technical Committee [IEEE 802.3].

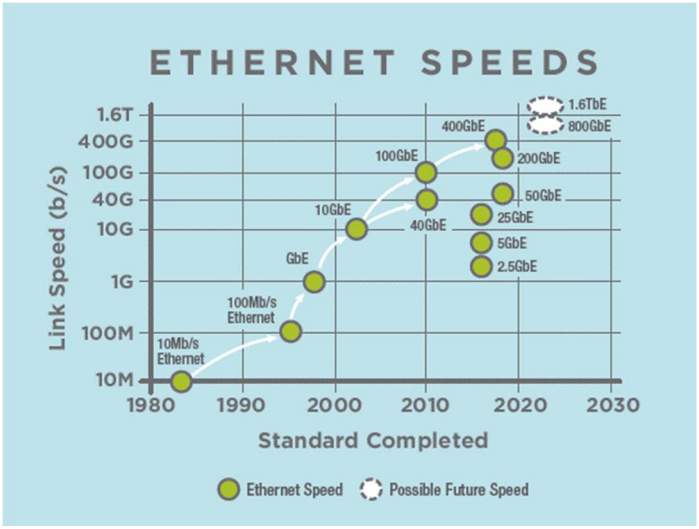


Figure 12. The Ethernet Alliance Roadmap (from [Eth 17])

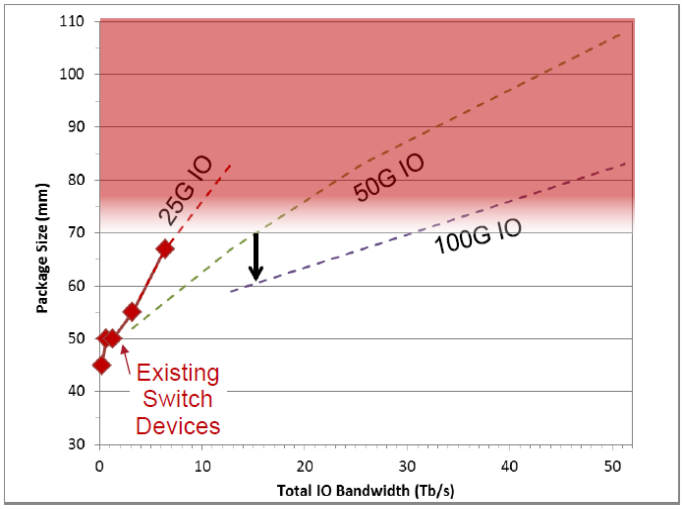
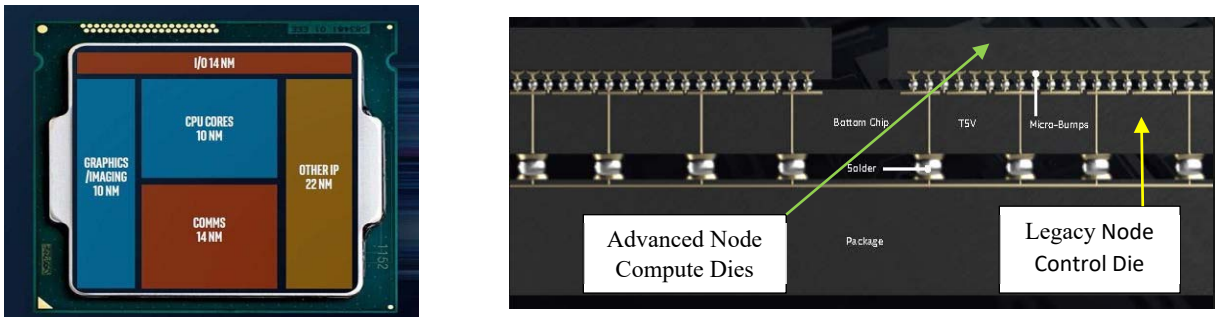


Figure 13. Relationship between package size and total IO bandwidth (from [IEEE 802.3])

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 its 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 Optical Internetworking Forum (OIF) CEI-112G committee is exploring 112Gbps per lane for an optical channel, and the IEEE 802.3ck draft 1.4 was released on December 10, 2020 for 106.25Gbps per lane signaling specifications. The final IEEE 802.3ck will be released in 2021 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 -36dB 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 -36dB channel loss. Therefore, PAM4 [Zha 17] or higher order PAM signaling is used for data rates >50Gbps to lower the Nyquist frequency to reduce 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 the way to enable the future Terabit network. For example, in 2020, using chiplets to integrate SerDes IP and optical devices to core logic die is gaining more attention in the industry, and replacing PCB traces with coaxial cables is under exploration for 100G signaling in the system design.

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 14 (a). An example of this is the Intel Lakefield processor that uses the Intel Foveros technology, shown in Figure 14 (b).



(a) Image from [Bohr 17]

(b) Image from [Ren 19]

Figure 14. 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

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 (10^9 Tflops by April 2020) as well as the averages of the top 500 HPC installations (10^7 to 10^8 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 HPC 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 tradeoffs in terms of cost, performance and power 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 off 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. Recent advances in photonics IO (Sec. 3.3) are going to enable the use of alternative data center interconnection technologies and tighter coupling at tens of Tbps in the next few years. Continuing advances in photonics will pave the way to disaggregated architectures where basic resources for storage and computing will be shared effectively at high data rates for a large class of data-intensive applications.

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

As the industry moves to smaller process nodes, costs for yielding large dies continues to increase. Compared to 250 mm² die on the 45 nm process, the 16 nm process more than doubles the cost/mm² and the 7 nm process nearly double that to 4x the cost per yielded mm². Moving to the 5 nm and even 3 nm nodes, the cost is expected to continue to increase. Fabricating large monolithic dies will become increasingly less economical. One solution to easing the economics of manufacturing chips with a large amount of transistors, the industry has started shifting to chiplet-based design whereby a single chip is broken down into multiple smaller chiplets that are "re-assembled" on package-level, which demands significant interconnect bandwidth. In addition, other heterogeneous components, such as HBM, GPU, and FPGA, are to be integrated in package simultaneously. The scale and complexity of SiPs requires greater carrier dimensions as well as higher interconnect density, which in turn drives the development of innovative packaging solutions.

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 with 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) [Mah 18], Figure 15, 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.

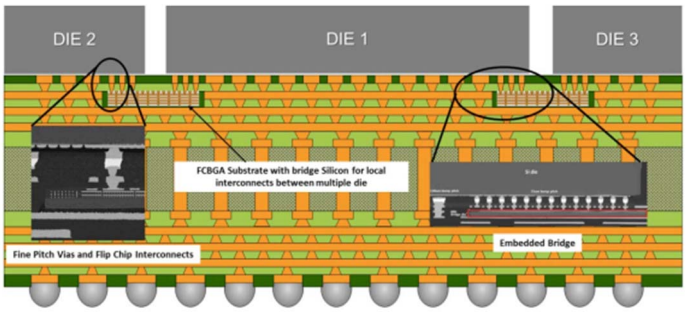


Figure 15. Embedded multi-die interconnect bridge (from [Mah 2018])

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 less than 10% of the chip total power budget.

- **High-density organic substrate**: By combining with thin-film processes, high-density flip-chip packaging is emerging as a potential heterogeneous integration carrier. 6/6µm line/spacing and <50µm via pitch will be commercially available at reasonably low cost by 2022. 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.
- **Ceramic-Based Heterogeneous Carrier (CBHC):** Ceramic substrates had been widely used to integrate multi-chip modules for decades until gradually replaced by organic laminates attributing to continuous advancement of semiconductor technology. Now that system integration comes back to package level, leading ceramic companies, such as NTK technologies, are developing Ceramic-Based Heterogeneous Carrier (CBHC) by taking advantages of both ceramic and organic materials. Such low-cost, large dimension, low-CTE, bond&assembly friendly, reliable, and reworkable heterogeneous substrate technology is expected to become commercially available in the near future.

InFO-RDL

Embedding interconnections at the wafer level is an attractive alternative to bridges. The INFO-RDL technology does this with wafer-level fan-out packaging. Wafer-level packaging refers to assembling dies on a wafer before dicing. Wafer-level fan-out packaging embeds the IO interconnections within the wafer, allowing them to scale with the die shrinks. The IO interconnection and the dielectric layers are fabricated on the wafer using a "chip-first" process [TSMC 2021]. The embedded IO connections make up what is called the redistribution layer (RDL) and the packaging technique is referred to as InFO-RDL (Integrated Fan-Out packaging with RDL). In recent years, InFO-RDL has become very attractive as yet another high-performance interconnection solution, with connection pitches currently down to 2µm and with support for up to 5 redistribution layers for enhanced 2D (e.g., 2.5D) integration [TSMC 21a]. Variations of InFO are also available to support 2-layer 3D and enhanced 2D integration as shown in Figure 16 [TSMC 21a]. InFO-RDL is expected to continue as an available and proven integration technology for high-end HPC applications.

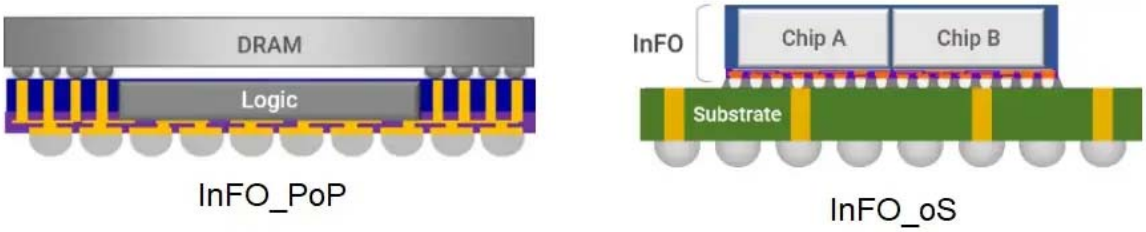


Figure 16. Variations of InFO from TSMC (from [TSMC 2021)]

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 (Figure 17). 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 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.

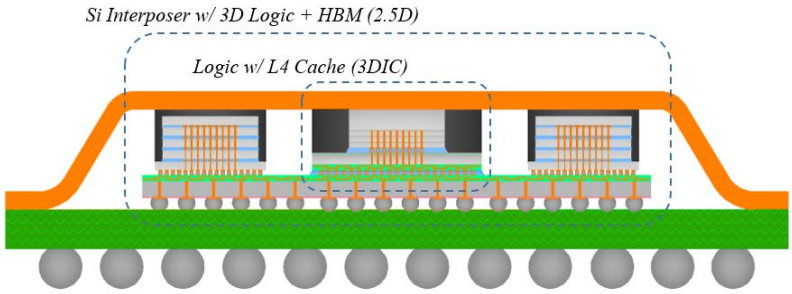


Figure 17. Diagram highlighting how supplemental SRAM cache is likely to be implemented alongside HBM in a 2.5D package. Source: [Eng 17].

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 significant bandwidth gain is expected (in the case of HBM3, in particular). This is because for 3D stacking, the distance can be considered as ~50um (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, 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 L3 or 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. A multicore SiP with a stacked SRAM cache has been recently introduced by AMD.
- **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.

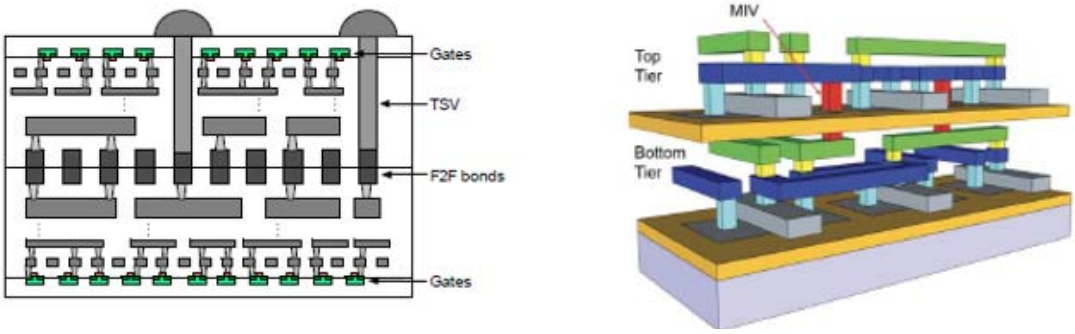


Figure 18. IP block partitioning for 3D integration (left) and monolithic 3D integration (right)

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 18; 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 19.

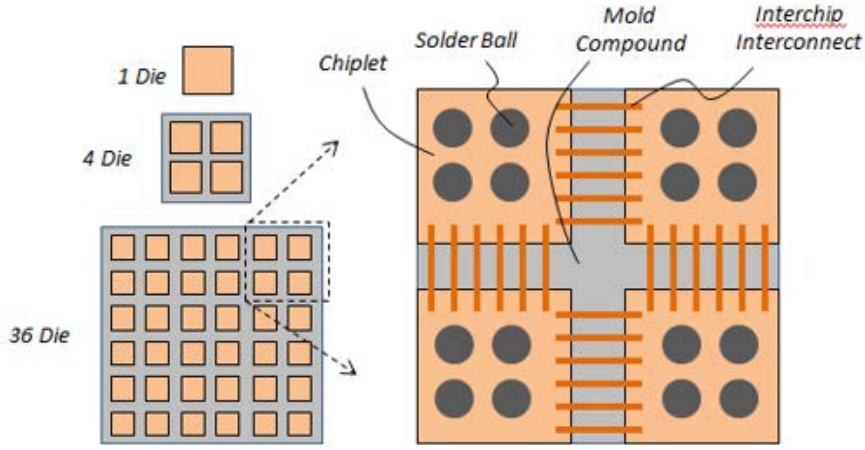


Figure 19. Scalable computing packaging concept using HDFO packaging. The Inter-chiplet interconnect utilizes minimum feature size RDL wiring. [Eng 19]

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 (eg, reliability, cost/yield, and thermal dissipation).

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.
- **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. One µm diameter TSVs have also been demonstrated [Dudo 20].
- **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. Production-level microbump pitches of 10 nm (or, equivalently, a maximum bump density of 10,000 per square mm) and lower have been announced by Intel for the Foveros-Direct technology in 2021, featuring low-resistance direct copper-to-copper bonding. The pitch is to be reduced further in the upcoming generations [Intel 21a]. Similar technologies include TSMC's SoIC (System on Integrated Chips) offering [TSMC 21b].
- **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.

In general, bumpless stacking solutions have the potential of lowering the (vertical) thermal resistance of the stack for reducing the junction temperature increases in the stacked chiplets. Bumpless connections also reduce the interconnection latencies (shorter lengths and lower resistance, combined with lower capacitance, due to the absence of underfills between stacked chiplets). This makes them particularly suited for logic-logic and logic-memory connections required in HPC/Data Center solutions. Thinning dies to be stacked also helps similarly, by reducing the vertical thermal resistance. Die thinning has been used and matured for memory devices, including DRAM and SRAM stacks and imaging products. Many of these interconnection solutions are useful for wafer-on-wafer, die-on-die, and die-on-wafer solutions (including CoWoS). For high-cost dies, die-on-die stacking using thinned dies and bumpless connections implemented with a Damascene process is attractive, as it can test the dies before stacking.

SoIC is TSMCs 3D integration technology and represents a 3rd party solution for 3D integration. SoIC will be available in two flavors, SoIC WoW and CoW [TSMC 21b]. In SoIC, identical-sized dies on two wafers are

connected using face-to-face bonding – a bad chiplet can be bonded with a good chiplet and the entire 2-chiplet stack will be defective. However, once alignment is done at the wafer-level, bonding can proceed. The tradeoff in SoIC-WoW is thus between the time-to-market and the yield. TSMCs SoIC-WoW process is offered for wafers with identical process technologies – for 7 nm wafers at the end of 2021 and for 5 nm wafers in 2022, with 3nm wafer solutions projected for 2023. SoIC-CoW permits chiplets of different dimensions to be stacked, currently with bump pitches about 10 microns, similar to Intel's initial Foveros-Direct offering.

Finally, in larger systems, 2.5D and 3D interconnection technologies can be combined. For example, Intel EMIB and Foveros technologies are combined in the Ponte-Vecchio offering. TSMC's SoIC can be combined with the relatively mature CoWoS technology or the InFO-PoP technology. A detailed evolution and roadmap of the interconnections available for heterogeneous integration is presented in the companion chapter on Interconnections.

For applications in the HPC/Data Center domain that involve significant amounts of data (as in big-data applications, machine learning, graph processing and others), vertical stacking of memory and logic devices are attractive, as it reduces the connection length for data accesses and simultaneously reduces the drive power. Additionally, clock-forwarded data transfers (where the sending chiplet's clock is forwarded along with the data to the receiver), are possible, precluding the use of complex clock recovery circuitry that would otherwise be needed at the receiving chiplet. To permit performance scaling for stacked logic and memory SiPs, the areal density of vertical connections has to be increased. The use of bump connections (of the order of 100 um and upwards in diameter) is not feasible in this scenario. Microbumps, which are a few 10s of um or less in diameter, offer a promising alternative in the near term. For dense vertical connections, bumpless face-to-face or face-to-back bonding offers low latency and low thermal resistance. With process technology improvements, bumpless contacts, used in imaging products, will permeate into HPC/Data Center SiPs in the coming years as viable choices for signal and clock lines. Ultra-short-reach connection PHY protocols, discussed for chiplets (Sec. 4), are geared to handle the vertical connection needs for stacked chiplets, including dense vertical connections.

Recent developments in the industry and research labs have all focused on reducing the bump diameter, from 20 um pitch microbumps to 1 um pitch microbumps realized with hybrid bonding, and ultimately down to 300 nm pitch with the use the Coolcube integration technology [Dut 20, Thu 20]. TSVs have been similarly reduced, with via diameters shrunk down to 1 um [Dut 20]. In the next few years, it is expected that the higher connections densities will be transitioned from the research labs to production scales.

Examples of 3D stacking technologies and stacked-chiplet systems demonstrated recently include:

- The Intel Foveros technology, which is used in the Lakefield line of multicore CPUs. The Lakefield family used two stacked chiplets: a lower-power 22 nm chip that contains a security engine, PCIe controllers, USB interfaces serving as an active substrate; and a 10 nm chiplet on top that contains multicore CPUs, graphics and DRAM controllers. The vertical connections between the two chiplets use microbumps with 36 um pitch to realize short reach, low latency and energy-efficient connections. The cross section of the stack is shown on the left of Figure 19 [Intel 21a].
- Intel's ODI (Omni Directional Interconnect) technology which uses wider vias for connections requiring low resistance. The use of fat vias for power (or signal) that can be supplied at the edge of the top die in this manner reduces the density of thinner vias that would otherwise be needed for the same purpose in the lower die and require its area to grow to accommodate the power vias. Thinner, finer pitched and short-length vias with face-to-face connections are used only for vertical connections that do not require a low resistance path, as shown on the right of Figure 20.

Figure 20. The Foveros technology used in the Intel Lakefield line (left) and use of fat vias for power connections and face-to-face microbumps in the Intel ODI technology (right)

- A demonstration vehicle by CEA-LETI that stacks six 16-core chiplets using microbumps on an active interposer [Viv 21]. The active interposer contains voltage regulators that are more efficient (replacing less efficient dropout regulators that are normally part of the multicore chiplet), NoC interconnections that connect the 16-core chiplets to implement coherent caches, IO and memory interfaces, and other peripheral logic. Vertical connections between each of the multicore chiplets and the active interposer use microbumps with 20 um pitch.
- Solder-based microbumps with 7 um pitch have also been demonstrated for die-to-die or wafer-to-wafer bonding in the past year by IMEC [Der 19]. Development continues into realizing even finer microbump pitches to support 3D chiplet stacking with microbump pitches at a few hundreds of nm at LETI [Thu 20].
- TSMCs 3DFabric suite, which includes three different 3D stacking technologies for a spectrum of products, including HPC stacking solutions, currently for 7 nm and 5 nm process technologies.

In the coming years, increasing the number of chiplets in the stack beyond two must address heat dissipation challenges and power distribution challenges aggressively. Smart physical organization within the stacked components can also partly mitigate the thermal issues, but additional research work is needed here.

Finally, it is to be noted that 3D chiplet stacking solutions are orthogonal to 2.5D interconnection technologies for SiPs, and both technologies can be deployed within a SiP. An example of this is shown in Figure 21.

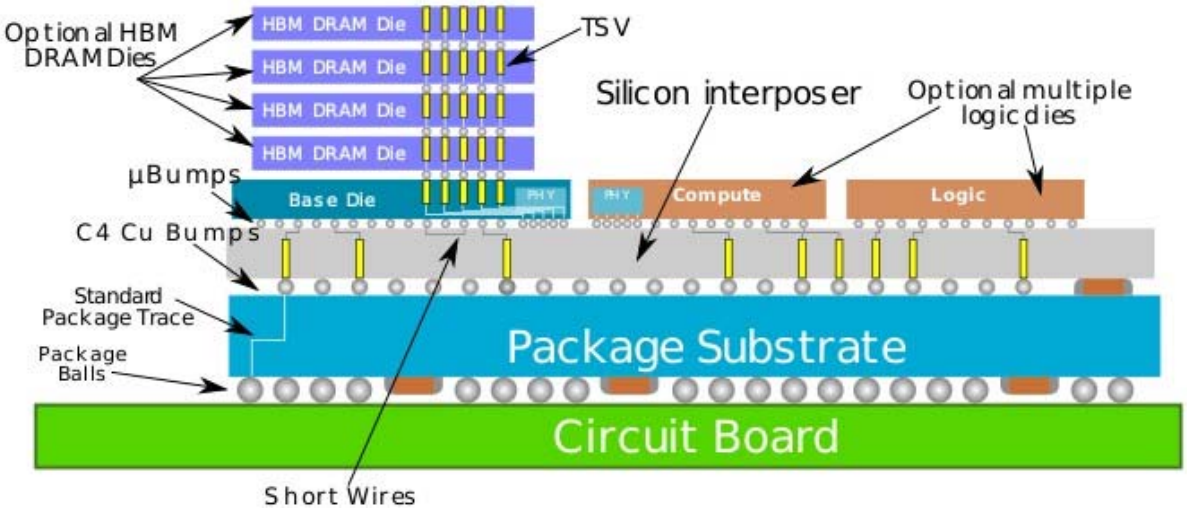


Figure 21. An example showing the use of 2D and 3D interconnections (courtesy TSMC)

A companion chapter (Interconnects for 2D and 3D Architectures) gives some projected scaling trends for the interconnections used for die stacking.

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. Commercial offerings like the Intel Lakefield family of CPUs and other demonstrations at a smaller scale are examples that show that 3D integration is permeating the HPC/Data Center market for SiPs. Reduction in dimensions of the 3D interconnections remain a top priority and is driven by the need to have higher interconnection densities. These advances will encourage new 3D SiP architectures for the HPC and Data Center markets, as 3D integration offers a significant reduction in the latencies among stacked chiplets in many cases. Many of the demonstrations mentioned under the short-term trends are likely to be transitioned into production.

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

- 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 has historically taken more than four years for each generation evolution (double data-rate). However, this evolution cycle is accelerating since the 4th generation. As PCIe Gen 4 was beginning to be offered, PCIe Gen5 was in development and, likewise, the industry has already started searching for solutions for PCIe Gen6 as Gen5 is starting to be offered. This is a clear indication of package-level system integration

advancement. The upcoming PCIe Gen6 will further double the data rate of PCIe Gen5 to 64Gbps by adopting the PAM4 signaling scheme. In the meantime, driven by package-level integration, numerous proprietary I/O standards have been emerging in recent years, such as GenZ, Omni-Path, and NVLink. Most are evolving towards 64Gbps in the next couple of years. The Table 1 shows SERDES I/O speed, distance, and channel topologies. Off-package 112Gbps data-rate is expected by 2022 with PAM4 signaling.

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.

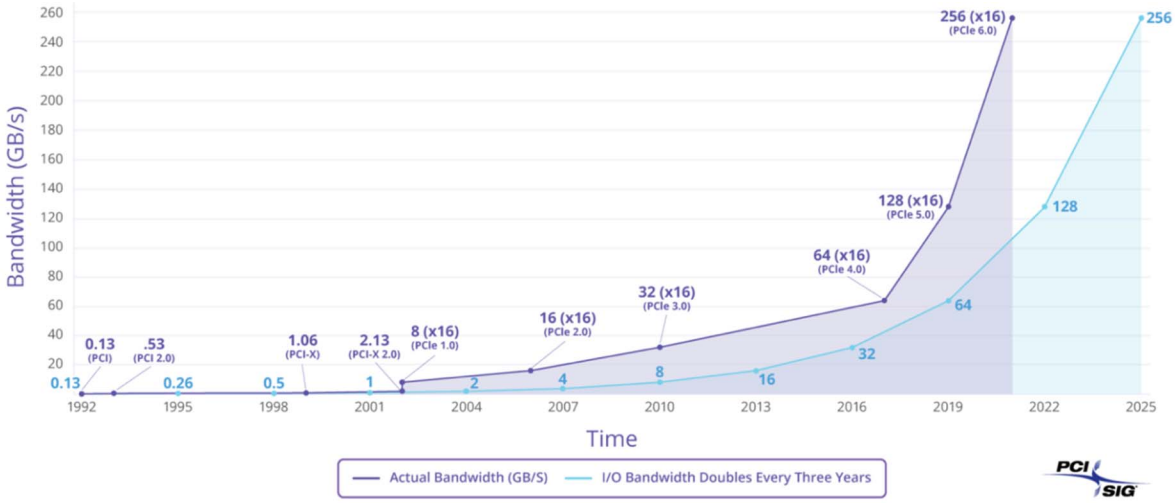


Figure 22. PCIe Roadmap (adapted from [Pir 17b])

Table 1. SERDES IO trend (compiled from various sources)

Parameter	Ultra Short Range	Very Short Range	Short Range	Mid Range	Long Range
Data Rate, Gbps	32 – 112	32 – 112	32 – 112	32 – 112	32 – 112
Bit Error Rate	1E-15	1E-15	1E-15	1E-12	1E-9
Distance, cm	1	5	15	50	100
Interconnect	MCM	PCB + 0 connector	PCB + 1 connector	PCB + 1 connector	PCB + 2 connector
Insertion Loss, dB @ f_N	3	6 (PAM4), 14 (NRZ)	15 (PAM4), 35 (NRZ)	30 (PAM4)	27 (PAM8), 45 (PAM4)
Modulation	NRZ	NRZ or PAM4	NRZ or PAM4	PAM4	PAM4 or PAM8
Forward Error Correction	N	N	Y	Y	Y

- 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 the 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, as is the use of other metals, such as cobalt and ruthenium or "compounded" copper.

3.3 Recent Advances in Package-Level Photonics IO and Switching

Recent advances are paving the way for in-package optical IO and photonics communication for the HPC/Data Center segments. The notable advances in 2019 and 2020 included the following:

- **Terabits/second Integrated Photonics**: CMOS SoC-implemented integrated photonics TeraPHY transceivers from Ayar Labs that overcome the line rate of 112 Gbps due to SERDES limitations for long-range connections, with a PHY-level connection latency of 10 ns (vs. ~100 ns for copper lines) at end points [Mea 19, Wade 20]. (The PHY end-to-end connection latency, of course, equals the media latencies plus the end-point latencies.) TeraPHY uses WDM implemented using an external multi-wavelength laser source to the SoC package that incorporates micro-resonator arrays and all peripheral logic. TeraPHY relies on direct fiber-attach techniques to couple the signal-encoded continuous light beam via a vertical grating to a fiber connector assembly as part of its packaging solution (Figure 23, top). Electronics innovations include the use of bit-statistical tuners which permits corrections to be made for self-heating over a large tuning range, for detecting zeros and ones. A demonstration vehicle implemented by Ayar Labs and Intel that connects a Stratix FPGA via EMIB to a TeraPHY SoC realizes a 1 Tb/s (16 X 25 Gbps) connection with 16 physical lanes and with a bit power budget of 0.8 pJ/bit, with an edge bandwidth of 1 Tbps/mm. The low-latency, Tb/sec communication bandwidth of TeraPHY not only supports HPC and data center scale applications but also allows resources such as RAM, accelerators, general-purpose computers, SSDs and HDDs to be disaggregated as in OCP for efficient resource sharing at the rack scale with very low connection latencies.
- **Ultra-miniature Optical Frequency Comb**: Miniaturized combed laser sources providing stable multi-color light streams are needed for dense WDM photonic connections. A low-power (1 Watt), miniature (measuring 1 cm^3) wafer-scale combed source using a nonlinear Kerr comb and hi-Q Silicon-nitride microresonators has been demonstrated [Raja 19, PV 19] and shown in the middle row of Figure 23. Optical feedback is used to provide stable wavelenths, eliminating active electronic or other tuning mechanisms on the resonator chip. The combed source provides equispaced wavelengths and relies on low-power InP laser diodes for the optical power source.
- **Co-Packaged Photonic Interface and Switch Fabric**: Intel demonstrated a CoWoS-implemented 12.9 Tbps programmable protocol-independent switch that accommodates 16 co-packaged optical interface modules, with each module rated at 1.6 Tbps [AgKi 20], Figure 23, bottom row. The extreme switching bandwidth of the switch allows for dense, high power connectivity at the rack scale or across racks.

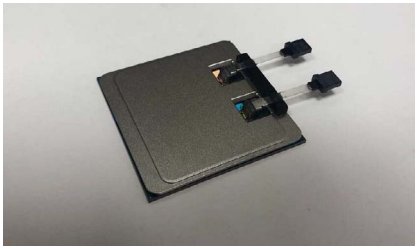
These technology advances enable the realization of package-level photonic IO with Terabits of IO with low power requirements and relatively high-stability, as well as fast switching of such links, paving the way for densely integrated SiPs for disaggregated OCP-style HPC architectures, new data center-scale switched interconnection technology that are geared towards large scale compute, machine learning and data processing applications. Photonics interconnection technologies, their implications and roadmap are detailed in a companion chapter.

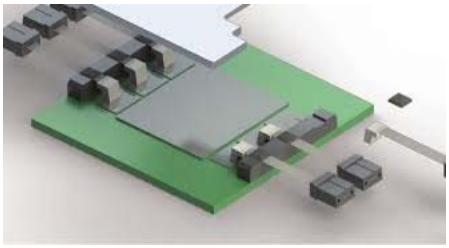
3.4 Signal Integrity Issues

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

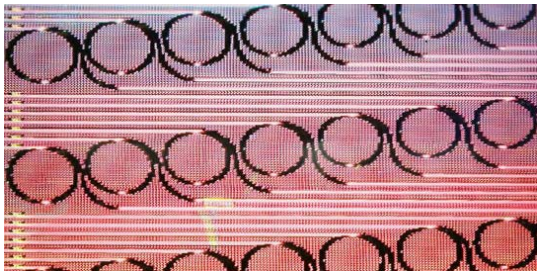
error correction capability going beyond ECC will be necessary for critical on-package connections and alternative symbol encoding, and signal processing necessary for recovering data waveforms for off-chip links may well become the norm in very high-end, high-availability SiPs.

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, tanδ=0.22) will be widely available, combined with smooth copper foil to mitigate skin effects. To support Cu smoothness requirement, 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, ISI, and cross-talk. For data rates beyond 50Gbps, PAM4 signaling will prevail for much lower Nyquist frequency.

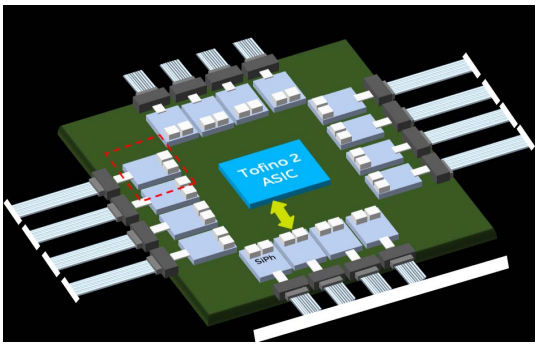




TeraPHY integrated photonics transceivers with packed integrated media connectors



Microresonators within miniature, low-power optical frequency comb (from PV [19])



Tofino 2 switch fabric and interfaces (left), Tofino 2-based 12.9 Tbps programmable switch module (right)

Figure 23. Examples of recent innovations that enable the use of practical package-level Tbps IO and electronically-switched low-latency, Tbps photonic interconnections. (Picture credits: top and bottom row pictures: Intel, middle row: EPFL)

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 and the crosstalk amplitude more accurately modeled because of the lower margin on higher speed interconnects. 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 [Hak 18].

3.5 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. The use of active interposers with stacked CPU dies permits the use of higher-efficiency converters within the interposer that can replace the relatively lower efficiency regulators that are implemented in the CPU die. A prototype system using this approach has been demonstrated recently [Viv 19].

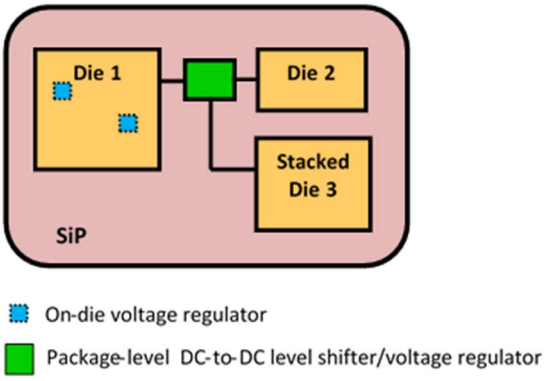


Figure 24. Power distribution and regulation inside package for SiPs

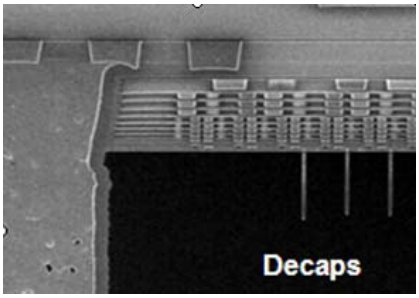


Figure 25. Active Si Interposer from that includes Deep Trench (DT) capacitors formed in the bulk Si wafer [Courtesy of Globalfoundries]

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 25), 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.

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.

3.6 Power Delivery

Power delivery is a challenge in general with SiPs for the HPC/Data Center segment. For 2.5D solutions, large power vias can be used. Additionally, power routing networks may have to be implemented within the interposer that can interfere and impose constraints on the routing of signal interconnections. New techniques have emerged to alleviate this problem. Delivering power from the backside of a wafer or die directly to the transistor layers avoids the relatively longer connections for power delivery and greatly reduces the losses in power delivery. It also opens up area on the die around the transistors for other connections, including those used for data delivery, avoiding complexities involved in routing power and closely spaced signal connections on the same layer. Intel's PowerVIA [Intel 21c] and IMEC/ARM's backside power delivery technology [Pra 19] are two examples. As transistors shrink with node advances, backside power delivery will be increasingly used to avoid the inefficiencies of using longer power routing networks on the die, to reduce noise and power droops. Intel's PowerVIA technology is shown in Figure 26. For 3D integration, the power delivery network has to rely on low resistance wide power vias or place power vias at the periphery of stacked packages to deliver power directly to the chiplets in the 3D stack. An example of this is seen in Intel's Foveros technology (Figure 26).

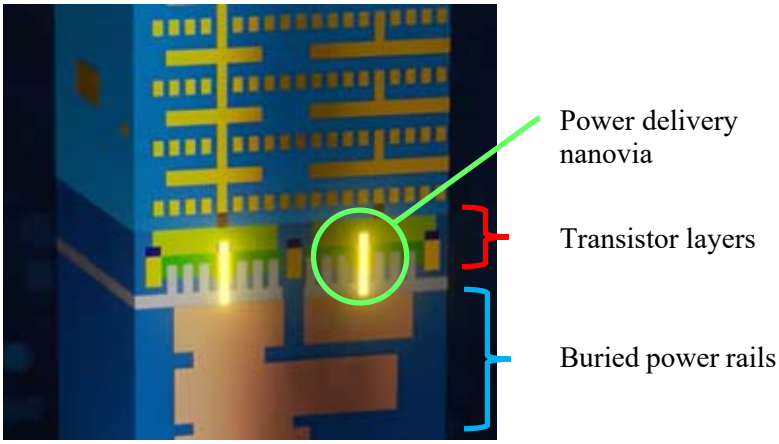


Image of cross-sectional view: courtesy of Intel

Figure 26. The Intel PowerVIA technology for backside power delivery

3.7 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.

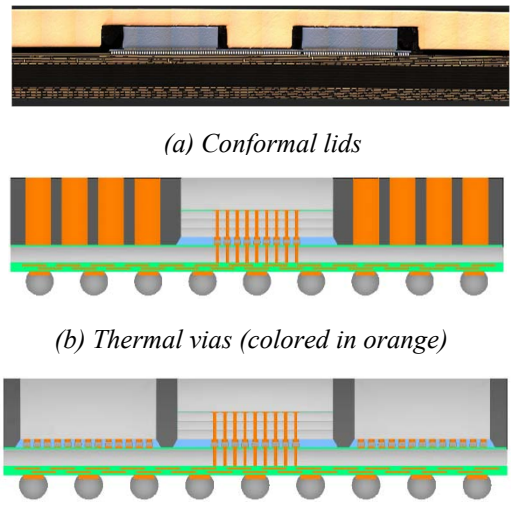
We envisage the short-term cooling solutions to use:

- 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 27(a).
- Thermal vias in 3D stacked structures, which are useful for top die thicknesses of less than ~200um, as shown in Figure 27(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 27(c).
- 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 on Thermal Management.



(a) Conformal lids

(b) Thermal vias (colored in orange)

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

Figure 27. Examples of near-term solutions to transfer heat to the top of 3D-integrated SiPs

3.7 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. Chiplets are quickly developing as the way to pursue this vision and leverage advanced packaging to implement the quickly evolving capability of chiplets. Initially, it was recognized that die disaggregation, such as separating out the high-speed SERDES transceivers from FPGA cores, could speed development of the silicon dies. To then offer this product, the individual dies need to be reassembled with a low-latency, high-bandwidth, and low-power interconnect which can approach the metrics achievable in a monolithic die. [Keh 20a] Ultimately, a diverse set of dies can be assembled in a heterogeneously integrated package that not only reduces development effort and product cost but, using device technology tuned to each application, can provide function that would be difficult to replicate in a monolithic die.

Specifically, this calls for the ability to integrate chiplets from different vendors and different technologies. To facilitate this goal, the industry has been devising PHY layer interconnection standards for communicating across

chiplets. These include communication protocols at multiple higher layers beyond the physical layer, and associated power management mechanisms and support for memory coherence across chiplets.

DARPA-funded research in the CHiPS program has produced a royalty-free die-to-die interface called the Advanced Interface Bus (AIB) to standardize the physical interface design which various protocols can support. The Open Domain-Specific Architecture (ODSA) is an industry forum that has been formed to facilitate standardization of die-to-die interfaces is developing Bunch of Wires (BoW) and OpenHBI standards. Other standards also exist; one example is the Optical Internetworking Forum's CEI-112G-XSR [Lap 20]. As the industry embraces heterogeneous integration for the HPC and data center markets on a wider scale, standards are also expected to develop to include power distribution, testing and system-level power management. These details will be needed for reliable interoperability of the chiplets. Figure 28 shows an overview of the die-to-die interface being pursued by ODSA (from Vin [20]).

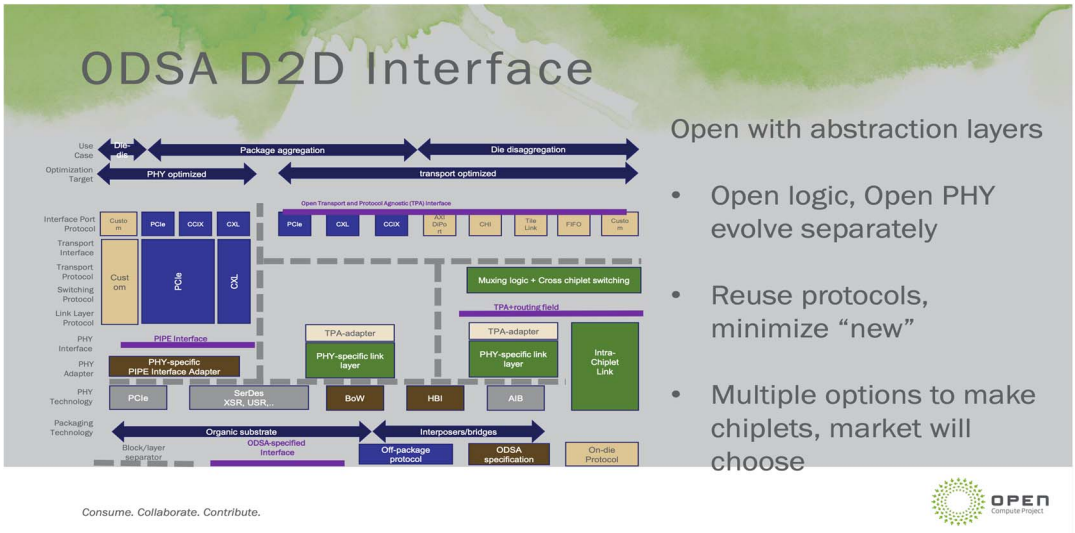


Figure 28. The die-to-die interface standards as defined by an industry group, ODSA. [Vin 20]

Using chiplets does not depend on a specific packaging technology, and the interface standards are set for flexibility in using 2D, 2DO, 2DS, and even 3D technology [Kan 20]. OEMs and the supply chain will need to develop the infrastructure needed to support and exploit the capability that chiplets will provide.

As these standards develop and chiplets become available, the function available to HPC and Data Centers will extend to include accelerator integration into the packaged components using chiplets. This gives a modular design and deployment opportunity

The developing standards will facilitate integration and reduce design and verification costs. The standards will address the challenges of power and signal integration and focus on energy efficiency of the die-to-die interfaces which will be a key focus item as wide-spread adoption is attained.

The design tools are key to successful deployment of chiplet-based designs. The tools need to comprehend the developing standards, partition the function, route power and signal on package and chips to minimize latency, minimize power, while also minimizing wiring resources needed [Yin 18]. Electrical tools must analyze the end-to-end interfaces for signal and power requirements by efficiently integrating multiple levels of packaging and chips into the analysis tools. This will necessitate standardization of interface files between tools and interoperability between design teams to build upon the standards that are being developed.

As stated above, there are several developing communication channel standards and protocols. Some examples are given below:

Advanced Interface Bus (AIB) [Keh 20b]: DARPA-supported wide scalable bus developed by Intel. AIB is a PHY interconnection specification emphasizing use of a large number of concurrent links and uses a forwarded clock from the sending side to the receiving side. As listed in the reference, at a data rate of 2 Gb/s increasing to 4 Gb/s with improved designs, the power of 0.85 pJ/bit has been demonstrated with an expected path to get below 0.5 pJ/bit. A total delay of less than 4 ns for the interface is demonstrated and is available for a ball pitch of 55µm. The interface lines and connections are as follows:

- TX from the sending chiplet to the RX connection of the receiving chiplet (and a similar pair of connections in the opposite direction)
- Clock lines (both directions, double-ended (clock-complement of clock pair)). Depending on the connection type, a single chiplet can provide the clock for the other chiplet it is connected to (as in a DRAM memory system that relies on the processor's clock).
- Control lines.

AIB has an advanced implantation called AIB Plus which adds double data rate (DDR) capability to AIB as well as improvements to handle the higher signaling speed. DDR launches data on both edges of the clock. Clock/data synchronization is implemented using DLL circuitries to maintain phase and duty cycles. Calibration circuitries are used to set up reliable PHY link operation. Specifically, for DDR, a duty cycle of 50% is maintained (with a 3% tolerance limit specified in the standard), which requires a complex duty cycle correction logic. Similarly, a delay-lock loop-based phase correction is employed to correct for clock skews on the way from the sender to the receiver to permit high data rates. AIB Plus also specifies the use of (optional) retiming circuitry prior to the IO block that samples the links based on the forwarded clock to determine if the data sent was a one or a zero. From an implementation perspective, link lengths are kept equal by using staggered bumps. Further redundant microbumps and links are needed to deal with yield issues.

Extra Short Reach SerDes (XSR): this standard accommodates short connections, up to 50 mm. [Ram 20, Ngu 19]. The Optical Internetworking Forum (OIF) has specified the Common Electrical I/O (CEI) 112G XSR (eXtra Short Reach) PHY interface. This specification details the interconnection between chiplets, as well as between chiplets and optical engines. The 112G XSR interface is designed for high performance, low-complexity implementation and low end-to-end energy per transmitted bit, as seen from the data available at [Syn 20]. A 16-lane configuration is used in the XSR interface. The DLL-based, clock-forwarded links employ PAM-4 (or NRZ) signaling and realize data rates from 72 to 116 Gbps for reaches up to 50 mm. Link timing recovery is facilitated using low jitter PLLs.

ODSA Bunch of Wires (BoW) [OCP 19, ODSA 19, Vin 20]: BoW is a specific die-to-die parallel PHY interconnection standard developed by ODSA that uses bidirectional transfers with NRZ symbol encoding. A single-polarity power source is used, and reaches up to 50 mm are supported. Specifically, energy-hungry forward-error correction used with aggressive symbol encoding techniques like PAM-4 are avoided in the BoW standard. Source data are synchronized to differential clock signals. The three classes of BoW are as follows:

- BoW-Base, relying on simple transceiver designs and unterminated lanes, bidirectional links and are limited to reaches of 10 mm and capable of realizing data transfer rates of 4 Gb/sec per wire.
- BoW-Plus, relying on terminated lanes, pushes the data rate to 16 Gb/sec on each wire and provides a reach up to 50 mm.
- BoW-Turbo, relies on bidirectional wires terminated at both ends to double the data rate to 32 Gbit/sec when both wires in a bidirectional wire are sending data in the same direction.

Backward compatibility is automatically provided in the BoW standard. As an example, to connect a BoW-Base interface to a BoW-Plus interface, the BoW-Plus side only needs to disconnect its terminations. The BoW standard also specifies bump patterns (called a "bump map") with specific functions assigned to each bump on the die to provide the analog of a pin-compatible interconnection solution with multiple parallel wires. Each bump map accommodates 16 data wires. and bump maps can be arrayed or stacked to increase the bandwidth at the edge of a chiplet. A prototype BoW implementation in Globalfoundries' 14 nm process for use with organic substrates realizes bump pitches of 130 um with < 0.7 pJ per bit transport energy efficiency with a 0.7 to 0.9 Volt power supply. Interconnection latencies range from 3 ns to 10 ns, depending on the desired link error rate (BER). [ODSA BOW]

ODSA OpenHBI [ODSA 19]: OpenHBI is a new chiplet interconnection standard that builds on the relatively mature HBM standard for DRAMs. The standard specifies 128 channels (with dual simplex links) for the interface, split into groups of 32 channels. Each 32-channel group has 32 data wires, a parity bit wire, 4 DBI wires (which are used to minimize toggling of the links to save power), 4 wires to specify protocol-specific framing information, and a lane-repair indicator. The OpenHBI standard is still in the specification phase and prototype demonstrations of very similar specs are available from Xilinx.

Low voltage In Package INterCONnection (LIPINCON) [Dill 20, WC 20]: TSMCs' LIPINCON (Low-voltage-In-Package-INterCONnect) is a standard designed for CoWoS packaging, and short, thin vias for vertical connections are used with delay compensation circuits to avoid the use of PLL/DLL-based circuitries (and their associated die

real estate needs) to provide low-latency interconnections with a high areal interconnection density. Competing and improved technologies are emerging from other vendors.

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]. Currently, PCIe interfaces are exclusively off-package interfaces including the CXL standard. As the accelerator components move from off-socket to on-socket, these interfaces will need to also develop to meet the power and latency requirements for on-package interconnection or the chiplets themselves optimized for the afore-mentioned interfaces.

Note that some of these standards may be initiated to be more suited to silicon interposers (e.g., AIB) and others more suited for organic-based advanced packaging (e.g., BoW), but as the adoption advances, standards develop, applications are designed, and packaging advances, this will be a dynamic area and will need to be closely monitored.

Chiplets and chiplet standards represent a powerful solution to the challenges of silicon technology development and performance demands of developing systems. The capability of heterogeneous integration using package technology gives a unique opportunity to advance HPC and data center systems using these technologies.

5. Heterogeneous Integration and its Role in Quantum Computing

Quantum computing techniques have matured in recent years and have received significant attention from industries as a viable approach for solving problems that require a long time to run on traditional computing platforms. To understand the role of heterogeneous integration in quantum computing, a fairly detailed introduction to quantum computing is provided before examining the potential impact heterogeneous integration techniques can have in this area.

5.1. Overview of Quantum Computing

Current and emergent quantum computers are evolving into three distinct classes:

- Quantum annealers
- Quantum computers with analog processing
- Quantum computers with digital processing implemented using digital quantum gates

All three classes rely on the notion of two-level quantum systems storing information in **qubits** and exploit superposition of states and quantum entangling (Sec. 5.1.1). Quantum annealers are good for optimizing complex systems, while the two other classes represent more general and powerful classes of quantum computers. All three classes of quantum computers can ultimately benefit from the use of heterogeneous integration technologies. D-Wave's quantum annealer was one of the first quantum processing systems to be deployed to serve customer needs in specific application areas [DW 21].

5.1.1. Qubits

In quantum computing, the qubit is the basic unit of information storage used in quantum computers that rely on two-level quantum mechanical systems. Unlike traditional bits, the state of a qubit is quantified probabilistically. A classical single bit can be either in the zero or one state. However, a qubit can be in other states in-between a one and zero due to quantum state **superposition**. Because of the superposition, a qubit can represent multiple values. The superposed state of a single qubit actually corresponds to a vector, instead of a single value as in a traditional bit and can, in theory, potentially store an infinite amount of information. The level of precision with which a qubit state can be established via external sources ultimately determines how many different values one can store in a qubit for practical purposes. Operations on qubits thus correspond to operations on vectors (instead of on a single bit in a traditional system); this is exactly why quantum computers have a significant computing power compared to a traditional computer.

Current dominant variants for quantum computing [NA 19, Ne Ch 21] implement a qubit as:

- An electron in an atom, with its extreme spin direction corresponding to a zero (down or negative spin) and one (up or positive spin). These qubits are called spin qubits.
- A resonant inductor-capacitor tank, with the inductor implemented as a superconducting Josephson junction. These qubits are often called superconductor qubits.
- An ion-trap with the nuclear spin state reflecting the qubit content. These are called ion-trap qubits.
- An isolated photon, with the qubit content corresponding to the polarization direction.

Other qubit implementations are also possible and are described in Sec. 5.2.1.

Quantum computing algorithms are based on the stipulation that qubits are in specific states with an associated probability. The readout ("measurement") of the state of a qubit is destructive (as it destroys the superposition) and

the qubit value after the readout is always a zero or one, depending on the superimposed qubit state just before the measurement.

For simplicity, the superposed state of a qubit, which is a vector, can be viewed as a distinct point on the surface of a sphere ("Bloch sphere", Figure 29) and can be represented by a "longitude" and "latitude" value, with a one and zero at the poles. The destructive measurement process simply moves the point representing the state to one of the poles. Alternatively stated, the sign of the latitude (equivalently, North or South latitude) effectively determines what contents of the qubit will be after its actual superposed state is measured. The North pole on the Bloch sphere has been traditionally equated to the classical zero state as it is the lower energy state for spin qubits under the application of a magnetic field along the North-South axis, directed to the North (or +Z) direction.

Over time, the state represented by a qubit can also be altered by noise, both thermal noise and environmental noise; the resulting information loss is called **decoherence** (Sec. 5.1.2). It is thus critical to have a robust way of maintaining the state of a qubit, relying first on the quantum phenomenon of **entanglement** and then on constructing a robust **logical qubit** (Sec. 5.1.2).

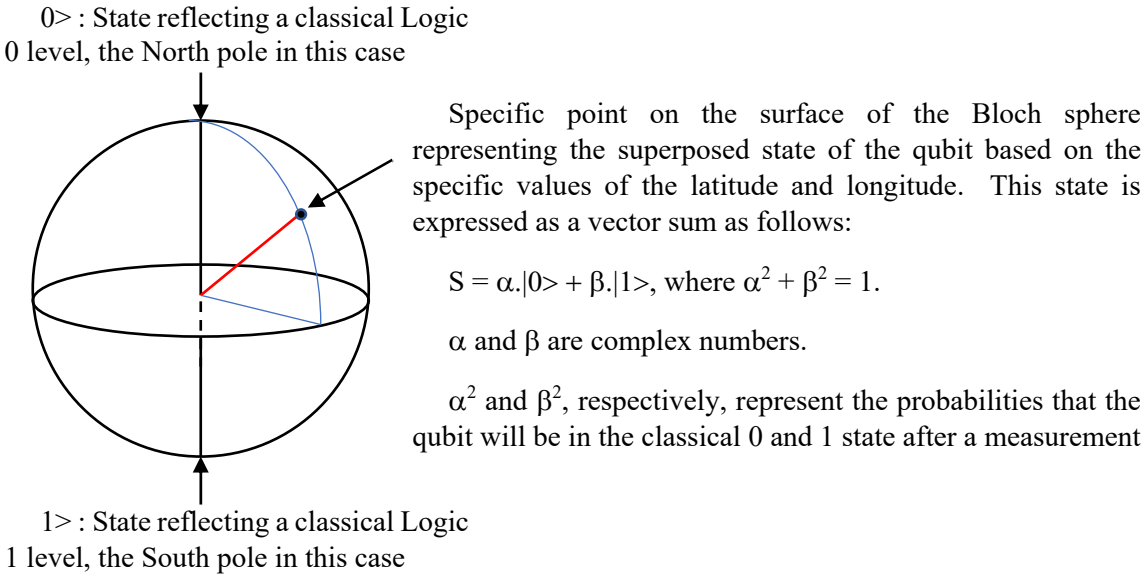


Figure 29. The Bloch sphere representation of pure superposed state of a qubit

Quantum entanglement is an unintuitive (and hitherto unexplained) phenomenon that essentially couples the state of two (or more) qubits, even at higher distances among the entangled qubits, so that they share a common state [NA 19]. With entangled qubits, the state of one qubit can be inferred from the state of the other qubit(s). To entangle qubits, a variety of techniques, depending on the specific qubit implementation, can be used. One such technique for chip-level qubit implementation relies on the existence of connections among the qubits and on some directed energy source (such as a microwave pulse) for altering the state of the individual qubits in a manner that entangles them – that is, makes them share a common desired state. Quantum entangling essentially permits a robust logical qubit to be constructed using multiple entangled qubits to provide resiliency against decoherence (Sec. 5.1.2).

5.1.2. Qubit Decoherence

If a qubit is left undisturbed, under ideal isolation conditions, it maintains its state. However, unintended interactions that violate the isolation requirements corrupt the qubit state – a phenomenon referred to as decoherence. The time in-between setting the state of a qubit and the time at which its state changes due to inevitable and unintended interactions with the environment is called the **decoherence time**. Keeping the qubit at near zero degrees Kelvin increases decoherence period, as thermal noise is minimized. The specific implementation of qubits also affects decoherence time. For example, spin-electron qubits have a lower decoherence time (a few tens of microseconds) compared to ion trap qubits using nuclear spin (several tens of seconds). Currently, one of the practical limitations of current computing platforms is the cumulative error introduced by decoherence. Techniques are thus needed to mitigate decoherence.

To detect unintended state changes to a single qubit, it is not possible to copy the state through a measurement, as measurement destroys the superimposed state in the source qubit. Thus, qubits cannot be replicated for redundancy to correct errors, requiring a different approach to provide resilience.

In a minimalist manner, a single and robust **logical qubit** can be implemented using three entangled qubits, say QA, QB and QC and two ancillary qubits, say, Q1 and Q2. Three distinct entanglements are used to realize a single, robust logical qubit. These are: a 3-way entanglement among QA, QB and QC, a second one among QA, QB and Q1 and a third one among QB, QC and Q2. Now consider the state of an individual qubit in the triad (QA, QB, QC), and say QA is altered due to noise. When the state of the ancillary qubits Q1 and Q2 are measured, the two measurements will now disagree and the altered qubit in the triad can easily be identified and set back to its proper value through an externally-applied signal – for example via a microwave pulse. It is to be noted that a single physical qubit cannot replace the triad, since measurements on ancillary entangled qubits will both agree when the sole qubit reflecting the stored value is corrupted. The example used implies that to read the state of a single qubit, measurements on the ancillary qubits have to precede the reading. The degree of robustness (that is, resiliency to noise) can be enhanced further using more qubits (often hundreds to thousands of qubits) to implement a single logical qubit.

The exact number of qubits needed to set up a logical qubit is also a function of the precision with which the state of a qubit can be set, with a higher precision calling for commensurately higher number of qubits to make up a logical qubit. Of course, for a given precision for the setting of a qubit state, fewer qubits will be needed to make up a logical qubit if the noise-resiliency of an individual qubit is inherently higher. Consequently, significant attention has been devoted to implement qubits in a manner that improves their inherent noise resiliency [NA 19].

5.1.3 Quantum Gates

Quantum gates represent the information processing element of a quantum computer. Unlike traditional logic gates, quantum gates are reversible – that is, no information is lost as they are processed. This requires quantum gates to have the same number of inputs and outputs. Entangled qubit contents also propagate through quantum gates as they are processed without any information loss. Compositions of quantum gates are also reversible. Quantum gates not only implement information processing but gates also exist to entangle qubits and swap qubit contents. Often, some of these gates can be embedded into the qubit array structures.

A variety of primitive quantum gates are available and these can be used to construct more elaborate functions. Different sets of universal quantum gates exist; gates from any particular set can be used to implement a complete set of arbitrary functions that operate on qubits. Some reversible quantum gates, such as the Deutsch gate (3-bit generalization of Toffoli gates) can perform classical Boolean operations [deWo 21, Ne Ch 21]. Although quantum gate designs have been proposed extensively and demonstrated in laboratory and controlled settings or as simpler qubit-embedded gates, an extensive set of robust quantum gates deployable into a practical general-purpose quantum computer is currently not available. Claims to the contrary are controversial.

The physical implementations of quantum gates are just as varied as the physical implementations of qubits and range from gated/coupled interconnections etched on the same die as qubits to large, external artifacts. Interestingly, the quantum gate speeds for the processing logic for qubits is lower for qubits with lower coherence time and higher for qubits with higher decoherence time, indicating the tradeoffs needed in a quantum computer between decoherence time and processing speed. Quantum gate control has to also address noise in the control lines for the gates (Sec. 5.1.5 and 5.2.2).

5.1.4 Computing with Qubits

The superposed state of a qubit can be set to a specific value through excitation signals such as microwave pulses (see Sec. 5.2.1). The vector components reflecting the superposed qubit state can be set to the desired values to store information in the qubit. This essentially amounts to changing the location of the current qubit content on the surface of the Bloch sphere via changing of the longitude and latitude values. Changing the location of the point corresponding to the qubit state is equivalent, in an abstract sense, to applying controlled rotation in the X, Y, Z dimension, through appropriate excitation, in the 3-dimensional space represented by the Bloch sphere. To entangle qubits prior to the computations, the qubits to be entangled are coupled using couplers or gates, and their states are set as desired.

Quantum computing requires the use of gates. In the state-of-the-art, this is done by setting qubits representing the initial computing state to their desired values and then connecting them appropriately using gates. The gates are themselves controlled using traditional electronic circuitry. A typical quantum computing sequence is as follows:

1. **Pre-initialization:** In this step, the qubits involved in the computation are set to superposed state where the coefficients of the possible superposed states have equal values. Doing this makes it easy to change the coefficients to their desired values before the computations begin.
2. **Initialization:** The qubits are entangled as needed and the qubit states are set to the desired values prior to any computations.
3. **Problem Encoding and Computation:** The initialized qubits are then gated and connected as needed by the computation algorithm. Qubit interactions through interference, the natural equilibrium processes in the quantum sense, now cause the qubits to come to an agreement and converge to values reflecting the final solution.
4. **Measurement:** A final measurement step reads out the result stored in the qubits.

A series of such computing sequences are used to implement a specific algorithm. Section 5.2.2 presents how these steps are implemented in response to the software that implements the algorithm.

5.1.5. Sources of Errors in Quantum Computers

There are three major sources of error in quantum computers. The first of these has to do with the precision with which the vector components representing the qubit state can be set. This precision, in turn, determines the error introduced in subsequent computations. The second source of error is due to changes in the vector components representing the state during the decoherence process, which alters the superposed qubit state (that is, location on the surface of the Bloch sphere), affecting the accuracy of the result stored in the qubit. The third source of error is actually a set of errors that have to do with the connections to the qubit from the control and measurement circuitry, including crosstalk among lines that affect states of qubits that were not addressed by the control, errors due to mechanical misalignments, and others. Practical quantum computers must address all sources of errors. High-precision control and measurement circuitry as well as the use of resilient logical qubits (Sec. 5.1.2) built using multiple qubits come to the rescue here but wiring for control/measurement lines also have to be designed and packaged physically to reduce crosstalk and other signal artifacts that introduce errors [NA 19].

5.2 Quantum Computer Building Blocks

The qubit and its associated control and measurement circuitry are the most critical components of a quantum computer. This section presents some existing qubit implementations and then presents the subsystems within a typical quantum computer.

5.2.1. Qubit Implementations

Qubits can be implemented in different ways and some prevailing ones are presented here, with the names of companies using the implementations in parentheses.

Superconducting Qubits (Google, IBM, Rigetti, D-Wave, Alibaba, Quantum Circuits, Oxford Quantum Circuits and prior Intel designs) [Aru 19, DW 21, IBM 21, Mar 20, StDi 20]: Each qubit is essentially an oscillator implemented as a resonant tank with an inductor implemented as a Josephson junction, coupled in parallel with a capacitor. The typical qubit area is a few square microns. The Josephson junction, formed by a thin sandwiched

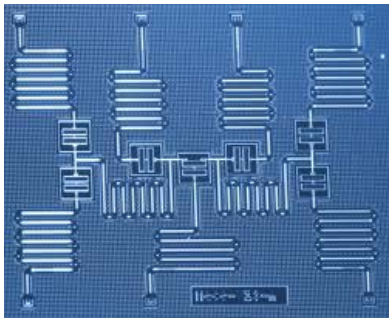


Figure 30. IBM's 9-qubit group

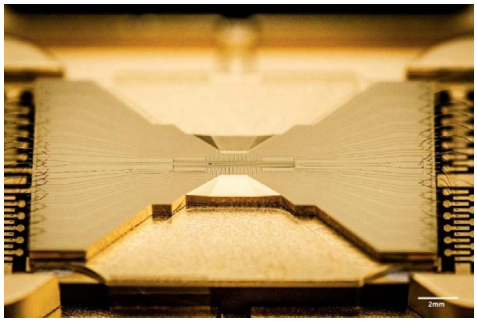


Figure 31. IonQ's 32-qubit assembly. The ions are trapped in the center (constricted part) of the assembly

insulating layer between two aluminum plates, starts to conduct as the qubit is cooled down to mK temperatures (20 mK is typical). Exciting the tank with microwave signals near or at the resonant frequency of the tank (typically, a few GHz) causes the tank to oscillate. The non-linear response characteristics of the Josephson junction and the tank

makes it easy to discern the vector components representing the qubit state for readout. The transmon qubit, as employed in IBM's design, is a special variant of a superconducting qubit that uses transmission-line technologies (two Josephson inductors shunted by relatively large capacitors) to reduce plasma oscillations. The shunt capacitors increase the energy stored in the superconducting inductor (Josephson energy) in proportion to charging energy for the capacitor, improving the resiliency of the stored state to charge noise. This improved noise resiliency increases the decoherence period. However, the shunt capacitor decreases the resonant tank's anharmonicity – that is, the ability of the tank to oscillate at frequencies other than what is expected from a harmonic oscillator is reduced. To set the qubits state to specific values, anharmonicity is needed and states are set with specially-shaped microwave pulses. The name transmon qubit is derived from the two artifacts characterizing these qubits (transmission line techniques and plasma oscillations). At the chip level, several qubits are implemented on a single die, with appropriate interconnections, such as grid interconnection patterns, where qubits are connected to the grid network via an electronically-controlled coupler. Separate circuitry allows the microwave signals to be directed to individual qubits and to direct electronic signals to enable or disable the coupler. Quantum gates are also formed similarly using couplers.

Ion Trap Qubit (IonQ, Honeywell, Alpine Quantum Technologies) [QA 20]: These qubits are implemented as ions trapped in an electric field, and the state of the electron in the outermost orbit of the trapped atom represents the qubit. These qubits have a long decoherence time and do not need to be cooled down to near-zero degrees Kelvin temperatures but cooling is still needed to increase decoherence time in the face of thermal noise and other potential sources of perturbation. These qubits are relatively slower, and require significant ancillary components (lasers, enhanced vacuum, microwave sources etc.) and have relatively larger area.

FinFET-Implemented Spin Qubit (Intel/QuTech, Silicon Quantum Computing) [Cla 20, Intel 21, Pil 18, Pil 19, va Di 19]: These qubits are implemented as isolated electrons in the transistor's channel with their spin reflecting the quantum state, and are implemented within a modified FinFET device. The modified FinFET device has multiple gates and some of these are used to admit and trap an isolated electron from a pool of electrons via electromagnetic control of the admission gates. Careful gate voltage control is the key to isolating electrons, and process variation and geometry variations have to be taken into account, possibly through a calibration process. Once isolated, the spin state is controlled/set with the application of microwave pulses from an ESR (Electron Spin Resonance) transmission line sitting on top of the FinFET. The ESR line generates a magnetic field that affects the spin state. With multiple FinFET gates on a single channel (7 in the current Intel prototype in a 22 nm process), two isolated electrons can be trapped in the FinFET channel and can be entangled. Operations on two or more qubits and measurements are performed by varying the potentials of the barrier and plunger gates. Qubit measurements essentially rely on spin to charge conversion followed by the use of a charge sensing circuitry. FinFET qubits have several advantages:

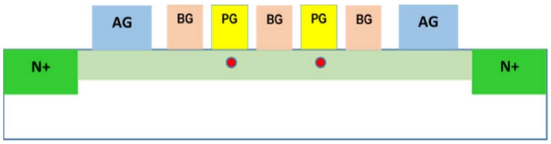


Figure 32. FinFET Qubit, with two isolated electrons (red dots) trapped under Plunger Gates (PG) whose potential controls the spin state. The solitary electrons are isolated and held in place by Barrier Gates (BG). The Admission Gates (AG) enable electrons to be supplied into the channel (light green) for isolation and use

a) The qubit area is reduced to a few square nm.
b) These qubits can be easily implemented in a slightly extended CMOS process.
c) These qubits operate at 100 mK temperature and do not need to be cooled down to the 20 mK temperatures for operations like other qubits. Operations at temperatures exceeding 1 degree K have also been demonstrated by Intel. This advantage is a direct consequence of eliminating superconducting elements within the qubit.
d) Quantum gate logic can be easily built in existing CMOS technology around these qubits.

It has also been claimed that FinFET-implemented spin qubits have a long coherence time and this effectively permits it to operate at temperatures close or above one degree K.

Photonic Qubits (PsiQuantum; Xanadu) [Arr 21, Choi 20, Pool 20, Xanadu 21]: Photonic qubits can be implemented as an isolated photon or as a beam of superposed squeezed photons. The state of a single photon qubit is characterized by two parameters: amplitude and phase, based on the wave analog of a photon. For a normal photon, based on Heisenberg's Uncertainty Principle, the uncertainty of measuring the amplitude or the phase are distributed

identically. In a squeezed state, the uncertainty between the amplitude and phase of the wave equivalent of a photon is skewed either towards the phase or towards the amplitude, unlike normal photons where the phase and amplitude distributions are identical – that is, not favoring the amplitude over the phase or vice-versa. Squeezing, in essence, reduces the uncertainty in measurements of either the amplitude or the phase of a photonic qubit, allowing the state to be represented by either the phase or the amplitude, while the other (and the larger uncertainties in the associated measurements) can be ignored. Squeezing also permit a larger number of photons in squeezed states to be entangled.

Quantum computers using isolated photons as qubits, which have to be made robust to survive measurements (Sec. 5.1.2), require complex setups for processing. Quantum computers based on photon beams fall into the class of Continuous Wave Quantum Computers (CWQC). Representing a significant advance, the key components of a programmable CWQC using photon beams with squeezed states have recently been implemented on a single chip in the X-series of photonic quantum computers from Xanadu (Figure 33) [Xanadu21, Arrazola 21]. The 4 mm X 10 mm chip includes squeezed photon generators and all processing logic for processing eight squeezed-photon beams. No complicated physical alignment for the optics,nor any extensive refrigeration are needed, as the chip integrates the photon squeezing and electronically-controlled processing logic to support operations at room temperature. The use of high-energy photon beams in Xanadu's system greatly reduces disturbances due to low-energy "noise" photons of thermal origin even at room temperature operations of the chip.

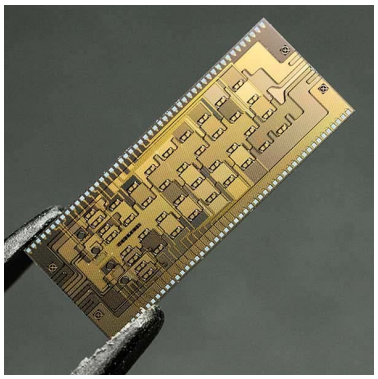


Figure 33. Xanadu's Silicon Nitride quantum computer chip based on photonic qubits with squeezed states

Xanadu's quantum computer generates beams of the squeezed photons from an external pulse laser source that is fed through a flexible fiber cable into micro-ring resonators at one end of the quantum computing chip depicted in Figure 33. The squeezed photons exiting the micro-resonator are filtered out from other photons and then directed via on-chip waveguides to a series of on-chip beam splitters and phase shifters which are controlled electronically based on the desired gating configuration that meets the processing needs. The beam splitters and phase shifters collectively form an interferometer and serve as a series of computing gates. The interferometer performs linear optical transformations and also entangles the squeezed photons, contributing to the final results based on the dynamically-formed gating configuration. The entangled photons then exit the chip and are counted using individual photon counters at each output of the chip. The integer array of photon counts of the chip outputs represents an encoding of the solution. Although limited to solving restricted classes of problems, in part due to the limited connectivity among qubits on the chip, Xanadu's chip represents a practical way of implementing photonic quantum computing at room temperature using the squeezed qubits, which are intrinsically robust. At the present time, Xanadu's quantum computing system requires a small tabletop refrigeration unit for the transition edge-sensing photon counters (maintained below 100 mK) to perform consistent measurements with high accuracy on the entangled beam of photons exiting the chip. In the future, advances in compact photon counter developments will eliminate the use of near zero Kelvin cooling, implementing a quantum system that operates completely at room temperature.

Cold Atom Qubits (ColdQuanta) [[Burt 21, CQ 21, CQ21a]: Cold atom qubits use entire atoms as qubits. ColdQuanta's cold qubit-based system traps isolated Cesium atoms in a two-dimensional array of laser beams within a small space in vacuum. The entrapment cools down the isolated atom to microKelvin temperatures (< 5 microKelvins), where noise signal levels that can perturb the qubit state are practically absent. In the resulting (almost) noise-free environment, the quantum properties of the cold atoms are adjusted appropriately and positioned relative to configure the quantum gates for the computations. Repositioning and state manipulations are done by modulating the laser beams and microwave sources. The microKelvin temperature also permits qubit contents to be retained for a relatively longer period; the increased time-to-decoherence period permits relatively more computations to be done in sequence before measurements are done to "copy" the quantum state for the next series of computations. The smaller dimensions of the isolated atoms, compared to semiconductor-implemented superconductor qubits, also enable dense packing of cold atom qubits in a compact space. This enables the system to scale up very easily to accommodate a significantly higher qubit count ("a million qubits on a thumbnail"). ColdQuanta also claims a connectivity (the number of qubits that can interact) of 4 for its qubits, double the connectivity of most other qubits. Finally, Cesium atoms are all identical and do not suffer from manufacturing tolerance variations seen with other types of superconducting qubits implemented in Silicon. The Hilbert ColdQuanta platform targets to incorporate over 100 qubits by early 2022.

Topological Qubits (Microsoft) [Cas 21, Rei 21]: Topological qubits draw their name from elastic shapes that retain their topological characteristics (= contents) irrespective of distortions in the shape (= perturbance introduced by noise). Microsoft's implementation of topological qubits uses nanowires with a thin indium antimonide core surrounded by a sheath of aluminum. At ultra-low temperatures and on the application of a magnetic field, the aluminum becomes a superconductor. Applications of signals via gates at the center and extremes of the wire separate out halves of a single electron (quasiparticles called Majorna Fermions) drive them to either ends of the wire and keep them separated, where one subparticle can be manipulated independent of the other. The typical qubit dimensions of Microsoft's topological qubit are about 1 um square. However, the claim about finding and isolating Majorna particles in Microsoft's qubit design have been recently withdrawn, raising serious questions about this specific implementation of qubits and the viability of the approach in general, making the future of implementing topological qubits implemented at the chip scale somewhat uncertain, although Microsoft appears to continue with its topological qubit efforts [Cas 21].

Other qubit implementations (notably, neutral atom qubits [CQ 21]) are also being used, as noted in Figure 34 (from QA[20]).

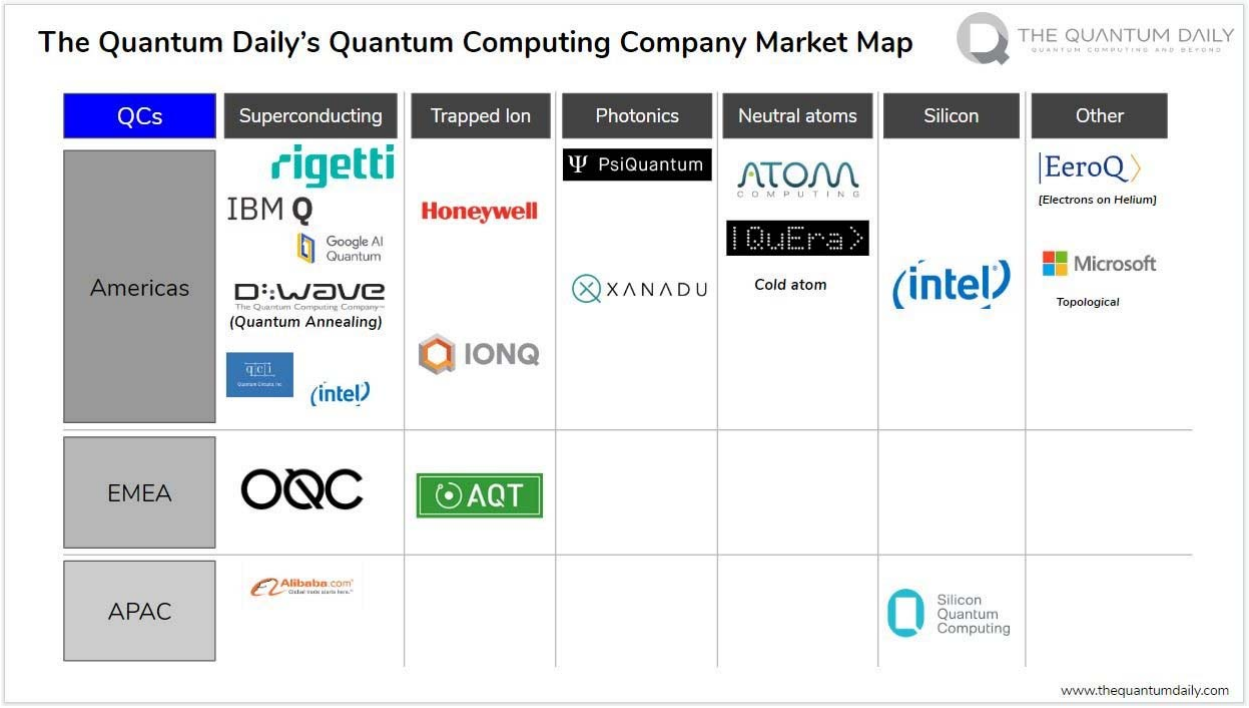


Figure 34. Existing qubit implementations [QD 20]

5.2.2 Subsystems within a Quantum Computer

At the system level, the major building blocks for a quantum computer, shown in Figure 35, are as follows:

Quantum data plane: This plane is represented by the chip/artifacts that implement the qubits and facilities for reading out their contents and for setting qubit contents. For gate-based architectures, appropriate programmable (and typically, somewhat limited in terms of connectivity) interconnection networks are part of the data plane and form the gates. The limited qubit connectivity in the data plane also constrains the variety of computation that can be performed and how they have to be performed under the limited connectivity. The data plane is kept in a very controlled low-temperature environment.

Control/Measurement plane: The control plane acts as the interface between the binary digital data used in the world outside the quantum computer for both input and output. It also passes on the commands from the control processor to orchestrate the computation sequences. The control plane is used to send the analog control information using wired connections or using optical or microwave signals via free space or waveguides from the control plane to the qubit for trapped-ion qubits. The transmission of the control information has to ensure that other qubits in the data plane are unaffected; this becomes difficult with an increase in the qubit number and density in the data plane.

In the process of reading out the analog qubit state, inevitable errors in the readout can accumulate and lead to larger errors in the final output. Crosstalk across the channels used between the data and control plane can also

inadvertently send commands to qubits that are not addressed, again introducing errors. Mechanical misalignment and improper shielding introduce crosstalk. Control pulse shaping and periodic recalibration can reduce these errors.

The control plane logic implemented in current CMOS technology can accommodate the data plane speed and is not a performance limiter.

Control and Host Processor: The control processor triggers the desired sequence of quantum gate operations via the control plane. The trigger sequence is generated in response to the algorithm specified in the host processor, so the control processor can be viewed as the command interpreter for the host's program. The host processor is a traditional processor with OS, libraries and the program that is run.

The control processor also runs error correction steps when needed, and such steps are implemented with traditional binary processing engines and can be complex. This complexity demands the use of highly concurrent and partly- or fully-customized hardware in the control processor, which should correct errors overlapped with quantum processing and measurement steps.

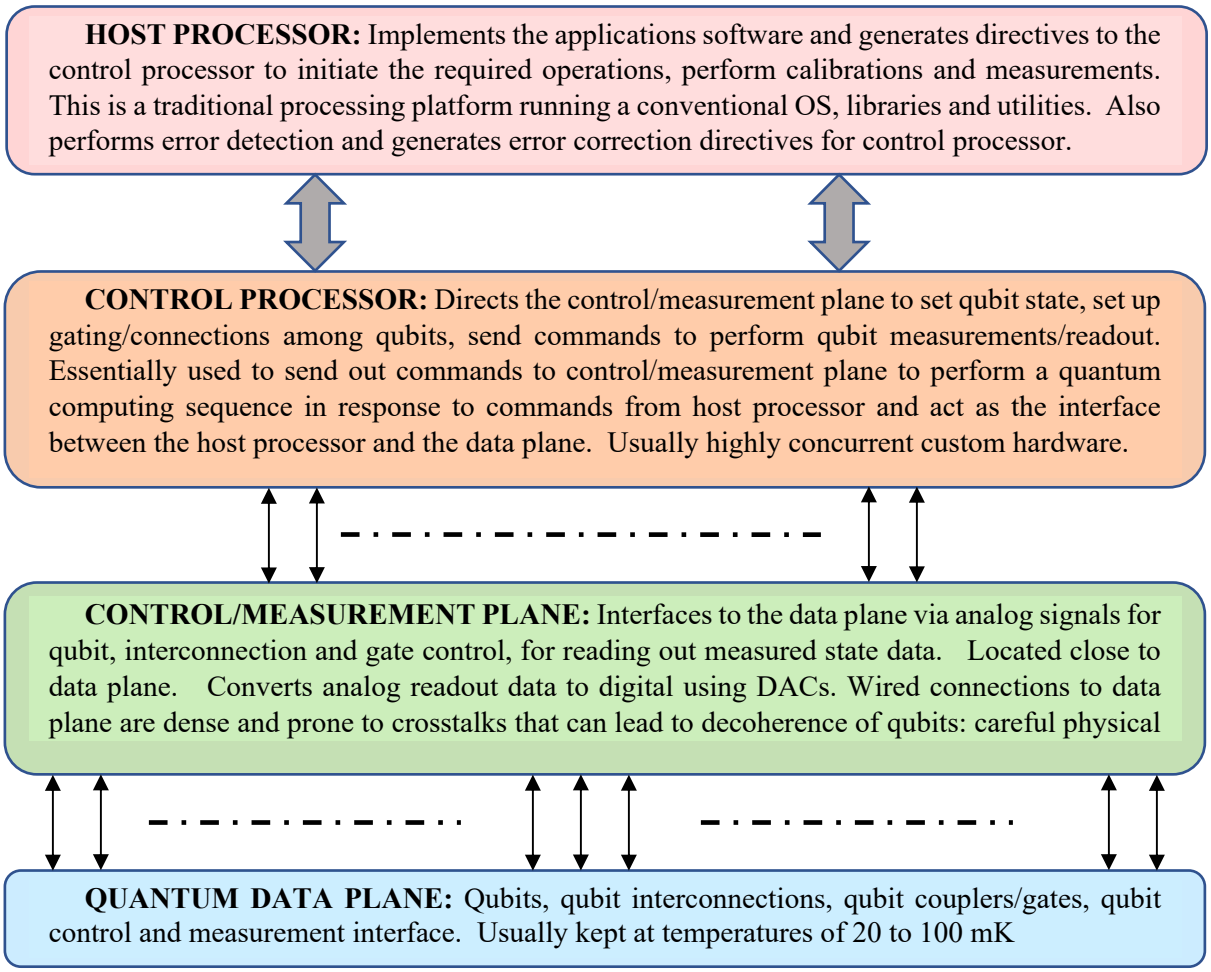


Figure 35. Key subsystem components of a generic quantum computer

Figure 35 depicts how the various subsystems come together within a typical quantum computer.

5.3. Considerations for Heterogeneous Integration

This section presents potential heterogeneous integration solutions that can benefit quantum computers and also overviews the cooling needs of exiting quantum computers.

5.3.1. Candidates for Heterogeneous Integration

The control/measurement plane is a good candidate to benefit from heterogeneous integration as it has analog and digital components. The digital components can themselves benefit from a chiplet implementation to capitalize on the use of different process technologies. However, it is to be noted that the control measurement plane can also be

implemented as a SoC when the qubit count is relatively low. As the qubit count goes up, a SiP implementing the control plane will begin to look attractive.

The most serious candidates for heterogeneous integration will, however, be the realization of the control/measurement plane and the quantum data plane as a SiP (or several SiPs operating in parallel when both planes are partitioned into groups). The SiP realization offers some immediate advantages: wiring complexity between the control/measurement plane and the quantum data plane can then be inside the package and perhaps be more manageable, reducing errors due to crosstalk and mechanical misalignments. FinFET qubits, which can operate at relatively higher temperatures compared to superconducting qubits, will encourage the integration, although the entire SiP needs to be cooled down. To bear the microwave signals, traditional interconnections can be replaced by waveguides in the interposer. The SiP integrating the control/measurement plane and the data plane will also allow the system to scale up effectively, as error sources are minimized. The development of ultra-low power chiplets for the delicate qubit control/measurements and control plane logic in general is critical for this integration.

5.3.2. Cooling and Packaging Needs

The vast majority of quantum computers emerging today have one common need: critical elements of the system need to be cooled to near zero degrees Kelvin temperature. The cooling solutions seen today are quite varied but most cooling systems used are based on a mixture of isotopes of helium that are pumped across the stages of a multi-stage dilution refrigerator.

Most superconducting qubit-based systems requires large refrigeration units to house the critical components [Mos 20]. A typical example of the cooling needs is shown in Figure 36. This translates to higher cooling power requirement in general. Of course, the inevitable tradeoff between decoherence time and operational reliability have to be traded off against the cooling power needs. The one downside to large refrigeration units is that they may get in the way of building quantum computers with a large number of qubits.

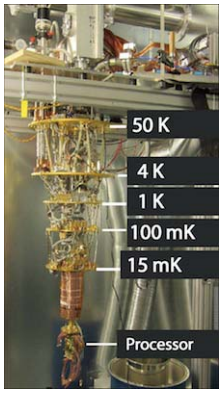


Figure 36. D-Wave's quantum annealing assembly showing cooling needs for different parts of the system. Most quantum computers have similar cooling needs

Figure 37. An example of the type of small tabletop refrigeration unit (from [Xanadu 21])

In Intel/QuTech's system using FinFET qubits, the dilution refrigeration system provides cooling in stages to cool the qubits in the data plane down to tens of milli Kelvins. In the coldest stage, only a few mWs of cooling capability exists, while for the higher stages it's a few Watts. This implies that the control plane power needs to be grossly limited, as it needs to be in close proximity of the data plane. (Intel's Horse Ridge II chip, which is essentially the data plane's control processor operating at 4 degrees Kelvin, is an example of a fairly integrated 100^{+} million transistor control plane SoC in a 22 nm CMOS.) Intel's CMOS FinFET qubit data plane chip can run at around one degree Kelvin, instead of requiring few tens of mKelvin temperature and lower cooling power.

In IonQ's system, directed laser Doppler "cooling" keeps the ion representing a qubit stationary in a low-energy field below the threshold for movements [Mos 20]. Several directed laser beams together implement this field. As the ion moves, the laser wavelength is blue-shifted to let the ion absorb energy and slow down after emitting a photon in a random direction. The photon emission takes away extra energy that led to movement and increases decoherence time. However, this may not be enough to take away heat from control plane, so refrigeration is still needed, albeit at a lower level.

Electrical cooling systems for quantum computers have also been proposed, essentially transferring heat via electrons across a special junction [VTT 20]. This technology has the potential for reducing the size and power requirements of the large cooling units seen in current quantum computers.

The extreme cooling needed in quantum computers also presents a different challenge that is not seen in today's system. To prevent thermal contraction and expansion from damaging the circuit boards and packaged chips used in the system, rapid temperatures decreases and increases, in the process of bringing an off-lined system to operational state and vice versa, have to be avoided. Dilution refrigerators have to be raised to room temperature over a period of days to avoid issues induced by rapid temperature changes over a very wide range. Cryogenic cycling needs also contribute to down time. Packaging solutions that can withstand rapid temperature cycling over a very large temperature ranges need to be developed to improve the overall reliability of the quantum computing system. The solutions developed should also be aggressive enough to handle the high pinout demands of any heterogeneously-integrated SiP developed for quantum computing.

Photonic quantum systems, such as the system from Xanadu, do not require extreme low temperatures for operation and all but the photon counters operate at room temperatures (Section 5.2.1). This cooling is currently handled by a small tabletop refrigeration unit (Figure 37). As with bigger refrigerators, Xanadu's cooling system has a downtime of 2 hours for every 24 hours of operation to accommodate cryogenic cycling needs. Emerging photon counter designs operating at higher temperatures will eliminate this cooling need.

5.4 Projections

Quantum computing is a paradigm that is constantly evolving at this time, with new technologies being introduced on a routine basis. This makes it difficult to predict the technology trends specifically, although some general trends can be noted:

Short term:

- Efforts will continue on reducing errors through technology, architectural and packaging innovations.
- Development of room-temperature alternatives will accelerate.
- Additional gate types and increased gate depths will be supported.
- Integration of control/measurement plane logic at low power with a qubit/gates chip as a solution to avoiding complex wiring; reducing errors due to crosstalk will be emphasized.
- Incremental scaling solutions will be demonstrated.
- Verifiable quantum supremacy demonstrations on systems with limited programmability will happen.

Longer term:

- Availability of products with versatile programming capabilities and scalability that target a wider application base will be offered.
- Significant reductions in form factor will take place, with heterogeneous integration playing an important role.
- Room-temperature quantum computers will become mainstream.
- Quantum supremacy will be demonstrated on fully programmable quantum systems.
- Potentially, several alternative implementations and architectures with new technologies will appear, which are difficult from those predicted at this point.

6. 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; • Silicon/EMIB bridges to support inter-tile with up to 4000 single-ended bit connections at low latency • ECC + Symbol encoding on links • Large interposers realized with reticle stitching • Locally-synchronous, globally-asynchronous clock • Initial offerings of 3D stacked products including both proprietary and accepted chiplet interconnection standards	• Tightly-coupled 2D tiled and 3D configuration (stacking) using chiplet interconnection standards • Dense vias implementing multiple 1000+ single-bit connections • Up to 8000 low-latency single-ended bit connections for 2D tiling • Aggressive use of 3D high-density 3D integration of chiplets for HPC/data center markets
Logic/DRAM Integration	• Interconnections to support up to 4000 bit connections /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 single-ended bit connections/ >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 • Broader integration of new NVM types
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 • Availability of integrated photonic IO with tens of Tbps links in switched configuration	• Multiple high-BW IO channels • AttoJ/bit photonic links • Advanced symbol encoding • Dense WDM • Emergence of fine-grained disaggregated systems for the HPC/Data Center markets based on tens to hundreds of Tbps links • Advanced copper IO

(continued …)

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 300 to 500 W per package; Specialized larger packages can demand 10+ KW per package. • Noise reduction in power delivery system; • Use of high-voltage (12VDC or 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 and more in larger packages; • 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 • Product-level deployment of active interposers with embedded, distributed high-efficiency converters at point-of-load • 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
Quantum Computing	• Initial demonstrations and appearance of few specialized products, initially via the Cloud • Development of truly general-purpose quantum systems, including the use of a broader set of quantum gates • Maturing of Silicon-implemented quantum computing platforms with reduced or no extraordinary cooling needs	• Appearance of complete, general-purpose and scalable quantum computing solutions • Increasing deployment of room-temperature quantum computing systems • Quantum computing systems, with increasing generality that operate at room temperatures • New, unforeseen paradigm and technology shifts, but likely based on the use of semiconductor technologies

Others: Alternative processing paradigms (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: Key Contributors

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

High Performance Computing and Data Centers: full TWG Team

Tawfik Arabi	Ivor Barber	Dale Becker
Bill Bottoms	Tahir Cader	Don Draper
William Chen	Luke England	Eric Eisenbraun
Kanad Ghose	Ali Heydari	Rockwell Hsu
Madhu Iyengar	Sam Karikalan	Michael Liehr
Ravi Mahajan	Gamal Refai-Ahmed	Tom Salmon
Lei Shan,	Bahgat Sammakia	John Shalf

Selected References

[AgKi 20] A. Agrawal and C. Kim, *Intel Tofino2 –A 12.9Tbps P4-Programmable Ethernet Switch*, presentation at HotChips 2020.

[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.

[Arr 21] J. M. Arrazola et al, "Quantum circuits with many photons on a programmable nanophotonic chip", *Nature* Vol. 591**,** pp. 54–60, March 2021.

[Aru 19] F. Arute et al, "Quantum supremacy using a programmable superconducting processor", *Nature*, Vol. 574, pp. 505-511, Oct. 24, 2019.

[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.

[Burt 21] J. Burt, "ColdQuanta Uses Cold Atoms to Build a Quantum Computing (sic)", The Next Platform, July 16, 2021.

[Cas 21] D. Castelvecchi, Evidence of elusive Majorana particle dies — but computing hope lives on, *Nature,* News page, March 10, 2021, available at: https://www.nature.com/articles/d41586-021-00612-z.

[Choi 20] C. Q. Choi, "First Photonic Quantum Computer on the Cloud", IEEE Spectrum Techtalk, Sept. 9, 2020. Available at: https://spectrum.ieee.org/tech-talk/computing/hardware/photonic-quantum.

[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

[Cla 20] J. S. Clarke, *Towards A Large Scale Quantum Computer Using Silicon Spin Qubits*, presentation at HotChips 2020 (HC 32) conference, August 2020.

[CQ 21] ColdQuanta web pages at: https://coldquanta.com/ and https://www.businesswire.com/news/home/20201111005101/en/ColdQuanta-Previews-its-Cold-Atom-Quantum-Computer-Technology.

[CQ 21a] ColdQuanta, "The Cold Atom Method", presentation at www.coldquanta.com.

[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

[Dem 18] Demler, M., "Mythic Multiples in a Flash: Analog In-Memory Computing Eliminates DRAM Read/Write Cycles", Microprocessor Report, Aug. 27, 2018.
[Deo 17] Deo, M., "Enabling Next-Generation Platforms Using Intel's 3D System-in-Package Technology", White Paper No. WP-01251-1.5, 2017.
[Der 19] J. Derakhshandeh, "Novell embedded microbump approach for die-to-die and wafer-to-wafer interconnects with variable microbump diameters and down to 5 um interconnect pitch scaling", in Proc. 22nd European Microelectronics and Packaging Conference & Exhibition (EMPC), Sept. 2019, pp.
[deWo 21] R. de Wolf, Quantum Computing Lectures Notes, available at: https://homepages.cwi.nl/~rdewolf/qcnotes.pdf, 2021.
[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.
[Dill 20] T. Dillinger, "Optimizing Chiplet-to-Chiplet Communications", article at SemiWiki.com, June 20, 2020. Available at: https://semiwiki.com/semiconductor-manufacturers/tsmc/287582-optimizing-chiplet-to-chiplet-communications/.
[Dix 20] H. Dixit et al, "TCAD Device Technology Co-Optimization Workflow for Manufacturable MRAM Technology", in Proc. IEDM 20 pp. 278.
[Dut 20] P. Dutoit et al, "How 3D integration technologies enable advanced compute node for Exascale-level High Performance Computing?", Proc. IEDM 2020.
[DW 21] D-Wave Systems, Quantum Computing pages at: https://www.dwavesys.com/quantum-computing.
[Edw 20] E. R. J. Edwards et al, "Demonstration of narrow switching distributions in STT MRAM arrays for LLC applications at 1x nm node", in Proc. IEDM 20 pp. 512.
[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, https://ethernetalliance.org
[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.
[Gup 20] M. Gupta et al, "High-density SOT-MRAM technology and design specifications for the embedded domain at 5nm node", in Proc. IEDM 20 pp. 516.
[Gwe 18] Gwennap, L., "Gyrfalcon Shrinks AI Accelerator", Microprocessor Report, Feb. 19, 2018
[Hak 18] Hakki M., "A Bayesian Framework for Optimizing Interconnects in High-Speed Channels", 2018 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO)
[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.
[IBM 21] IBM Quantum Computing System web pages at: https://www.ibm.com/quantum-computing/quantum-computing-at-ibm/
[IEEE 802.3] New Ethernet Applications Adhoc IEEE 802 March 14, 2017 Plenary Meeting presentation.
[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
[Intel 21] Intel Quantum Computing information at: https://newsroom.intel.com/press-kits/quantum-computing/
[Intel 21 a] A. Kelleher, presentation on PowerVIA at Intel Accelerated, July 2021. Webcast available at:

https://www.intel.com/content/www/us/en/newsroom/news/intel-accelerated-webcast-livestream-replay.html#gs.ezui39
[Intel 21b] B. Safi, presentation on Foveros at Intel Accelerated, July 2021. Webcast available at: https://www.intel.com/content/www/us/en/newsroom/news/intel-accelerated-webcast-livestream-replay.html#gs.ezui39
[Intel 21c] S. Natarajan, presentation on Intel PowerVIA as part of the Intel Accelerate Event, July 2021. Webcast available at: https://www.intel.com/content/www/us/en/newsroom/news/intel-accelerated-webcast-livestream-replay.html#gs.ezui39
[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.
[Kan 20] David Kanter, "Intel Extends 3D Packaging With ODI," Microprocessor Report, Jan. 20, 2020.
[Keh 20a] D. C. Kehlet, "Chiplets on the rise," presentation at the 2020 Heterogeneous Integration Roadmap Symposium, Feb. 21, 2020.
[Keh 20b] D. C. Kehlet, "Accelerating Innovation through a Standard Chiplet Interface: The Advanced Interface Bus (AIB)", Intel whitepaper, available at: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/accelerating-innovation-through-aib-whitepaper.pdf.
[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.
[Kum 20] Y.-C. Liao et al, "Spin-Orbit-Torque Material Exploration for Maximum Array-Level Read/Write Performance", in Proc. IEDM 20, pp. 282
[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.
[Lap 20] M. Lapedus, "Chiplet Momentum Rising," Semiconductor Engineering, Feb. 26, 2020.
[Lee 20] T. Y. Lee et al, "Advanced MTJ Stack Engineering of STT-MRAM to Realize High Speed Applications", in Proc. IEDM 20 pp. 234.
[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.
[Mar 20] J. Martinis, *Quantum Supremacy Using a Programmable Superconducting Processor*, presentation at HotChips 2020 (HC 32) conference, August 2020.
[McC 18] McCan, D., "Packaging and Heterogeneous Integration for HPC, AI, and Machine Learning", Presentation at SEMICon West 2018.
[Mea 19] R. Meade et al, "TeraPHY: A High-density Electronic-Photonic Chiplet for Optical I/O from a Multi-Chip Module," in Optical Fiber Communication Conference (OFC) 2019, OSA Technical Digest (Optical Society of America, 2019), paper M4D.7.
[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
[MJ 18] McCauley, S. and Jiang, S.,"Google 48V Update: Flatbed and STC", presentation at OCP Summit, March 2018.
[Micron 20a] Micron infographic, September 2020, available at: https://media-www.micron.com/-/media/client/global/images/infographics/gddr6x_infographic.jpg?la=en&rev=2a0569c5bd5c4fe5bb086168680b1f3b
[Micron 20b] Micron Technical Brief, Doubling I/O Performance with PAM4 – Micron Innovates GDDR6X to Accelerate Graphics Memory, 2020.
[Mos 20] S. Moss, "Cooling quantum computers", Data Center Dynamics, Cooling Supplement, Dec. 2020, pp. 13-15.
[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/

[NA 19] *Quantum Computing Progress and Prospects*, Consensus Report, National Academies Press, 2019.
[Naf 18] Naffziger, S., "CPU Performance for the Next Decade", Microprocessor Report, Sept. 24, 2018.
[NeCh 21] *Quantum Computation and Quantum Information, 10th Anniversary Ed.*, by M. Neilsen and I. L. Chuang, Cambridge Univ. Press, 2021.
[Ngu 19] N. Nguyen, "Accelerating Chiplets With 112G XSR SerDes PHYs", Semiconductor Engineering, Oct. 31, 2019. Available at: https://semiengineering.com/accelerating-chiplets-with-112g-xsr-serdes-phys/, online
[OCP 19] R. Farjarrad, B. Vinnakota, M. Kuemerle and B. Bahdori, "Bunch of Wires (BoW) Interface Standard for Chiplets", presentation at the OCP Summit, Aug. 2019.
[ODSA 19workshop] K.Ma, "ODSA OpenHBI Workstream proposal: Open High Bandwidth Interconnect for chiplets", ODSA Workshop, Dec. 2019.
[ODSA BOW] R. Farjadrad, M. Kuemerle and B. Vinnakota, "A Bunch-of-Wires (BoW) Interface for Interchiplet Communication," in IEEE Micro, vol. 40, no. 1, pp. 15-24, 1 Jan.-Feb. 2020, doi: 10.1109/MM.2019.2950352.
[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
[Pil 18] R. Pillarisetty et al, "Qubit Device Integration Using Advanced Semiconductor Manufacturing Process Technology", Proc. IEDM 2018.
[Pil 19] R. Pillarisetty, Spin Qubit Device Integration, Intel presentation, March 19, 2019.
[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/
[Pool 20] R. Pool, "A new kind of quantum", SPIE Technology focus pages, Nov. 2020. Available at: https://spie.org/news/photonics-focus/novdec-2020/a-new-kind-of-quantum?SSO=1.
[Pra 19] D. Prasad et al, "Buried Power Rails and Back-side Power Grids: Arm® CPU Power Delivery Network Design Beyond 5nm", Proc. IEDM 2019.
[PV 19] *The Smallest Optical Frequency Comb*, Photonics View, Feb. 22, 2019. Available at: https://www.photonicsviews.com/the-smallest-optical-frequency-comb/
[PW 20] "Cool technology enables quantum computing", Physics World article, March 16, 2020, Available at: https://physicsworld.com/a/cool-technology-enables-quantum-computing/.
[QD 20] *A Detailed Review of Qubit Implementations for Quantum Computing,* The Quantum Daily, May 21, 2020. Available at: https://thequantumdaily.com/2020/05/21/tqd-exclusive-a-detailed-review-of-qubit-implementations-for-quantum-computing/
[Raja 19] A. S. Raja et al.: Electrically pumped photonic integrated soliton microcomb, Nat. Commun. 10, 680 (2019). DOI: 10.1038/s41467-019-08498-2.
[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
[Ram 20] Rambus Inc., 112G XSR Multi-protocol SerDes PHY interface webpage at: https://www.rambus.com/interface-ip/serdes/112g-xsr-phy/, online
[Rei 20] D. Reilly, *If Only We Could Control Them: Challenges and Solutions in Scaling the Control Interface of a Quantum Computer*, presentation at HotChips 2020 (HC 32) conference, August 2020.
[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.
[She 16] Shehabi,A. et al, United States Data Center Energy Usage Report, Lawrence Berkeley National Lab Report No. LBNL-1005775, June 2016.
[SIA 17] Semiconductor Industry Association and Semiconductor Research Corporation, "Semiconductor Research Opprtunities: An Industry Vision and Guide" March 2017.
[Smi 21] R. Smith, "SK Hynix Announces Its First HBM3 Memory: 24GB Stacks, Clocked at up to 6.4Gbps", Anandtech article, October 20, 2021. Available at: https://www.anandtech.com/show/17022/sk-hynix-announces-its-first-hbm3-memory-24gb-stacks-at-up-to-64gbps
[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.
[StDi 20] M. Steffen and O.Dial. *Underneath the Hood of a Superconducting Qubit Quantum Computer*, presentation at HotChips 2020 (HC 32) conference, August 2020.
[Syn 20] Synopsis Inc., DesignWare USR/XSR PHY IP web pages at: https://www.synopsys.com/dw/ipdir.php?ds=dwc_usr-xsr_phy.
[Thu 20] S. Thuries et al, "M3D-ADTCO: Monolithic 3D Architecture, Design and Technology Co-Optimization for High Energy Efficient 3D IC", Proc. DATE 2020, pp. 1740-1745.
[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/
[TSMC 21a] T. Dillinger, "Highlights of the TSMC Technology Symposium 2021 – Packaging", June 2021, available at: https://semiwiki.com/semiconductor-manufacturers/tsmc/299955-highlights-of-the-tsmc-technology-symposium-2021-packaging/
[TSMC 21b] P. McLellan, "Advancing 3D Integration", June 2021, available at: https://semiengineering.com/advancing-3d-integration/
[Tum 06] Tummala, R., "Moore's Law Meets Its Match", IEEE Spectrum, 43(6), June 2006.
[va Di 19] J. P. G. van Dijk et al, "Impact of Classical Control Electronics on Qubit Fidelity", Physical Review Applied, **12**, 044054-1 to 044054-20, October 24, 2019.
[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
[Vin 19] Vinnakota, B., "ODSA: Technical Introduction", presentation, March 28, 2019, link available at: https://www.opencompute.org/wiki/Server/ODSA
[Vin 20] B. Vinnakota, "Open Domain Specific Architecture," presentation at the 2020 Heterogeneous Integration Roadmap Symposium, Feb. 21, 2020.
[Viv 21] P. Vivet et al, "IntAct: A 96-Core Processor with Six Chiplets 3D-Stacked on an Active Interposer With Distributed Interconnects and Integrated Power Management". IEEE J. Solid State Circuits, Vol. 56, No. 1, pp. 79-97, 2021.
[VTT 20] VTT Technical Research Centre of Finland, "New electronic cooling technology to enable miniaturization of quantum computers" Phy.org Quantum Physics pages, April 14, 2020. Available at: https://phys.org/news/2020-04-electronic-cooling-technology-enable-miniaturization.html.
[Wade 20] Wade, M. et al, "TeraPHY: A Chiplet Technology for Low-Power, High-Bandwidth In-Package Optical I/O", IEEE Micro Magazine, March-April 2020, pp. 63-71, vol. 40. DOI Bookmark: 10.1109/MM.2020.2976067
[WC 20] WikiChips artcle: Low-voltage-In-Package-INterCONnect (LIPINCON), available at: https://en.wikichip.org/wiki/tsmc/lipincon.
[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.
[Xanadu 21] Xanadu Quantum Technologis, Inc., "On the Road to Room Temperature Quantum Computation", June 2, 2020. Available at: https://medium.com/xanaduai/on-the-road-to-room-temperature-quantum-computation-d1bd356dcf57.
[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