Promises and Perils in 3D Architecture

Moein Khazraee
Computer Science and Engineering
University of California at San Diego

***Abstract*—The idea of 3D stacking architecture was initially introduced to postpone the end of Moore's law for CMOS technology. By stacking dies vertically on each other and connecting them with die-to-die vias, overall timing and energy efficiency improve due to better routing and shorter wires. It also introduces the notion of mixed process technology integration, connecting dies with different process technologies with high bandwidth connections. The promise of 3D also brings with it a new set of challenges in managing thermal constraints, power delivery networks, and coming to terms with changes to transistor properties caused by new types of mechanical stress. This paper overviews the primary methods for 3D stacking, the challenges it faces, and studies its opportunities in broad design range: from general-purpose to application-specific systems. In the last part, these implementations are evaluated and some future directions are proposed.**

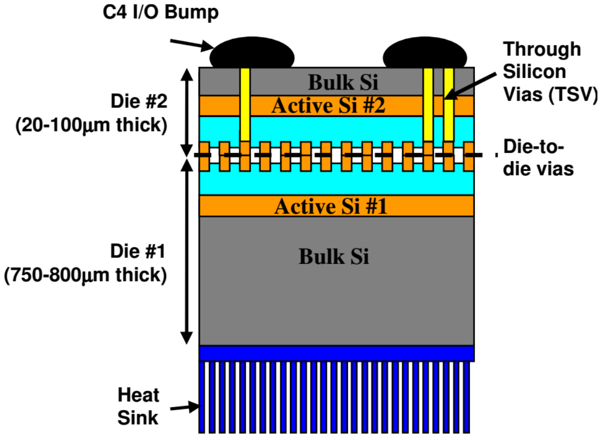

Fig. 1: **Two-die 3D-stacked architecture [5].**

I. Introduction

As CMOS technology moves towards nano-scale transistors, Moore's law has become outdated and unviable. One of the most recent approaches in chip design is 3D stacking, where dies are stacked vertically to increase the number of transistors per unit area and reach Moore's law prediction [1]. In 3D stacking two or more dies are placed upon each other and connected with die-to-die vias, as shown in Fig. 1 [2]. By adding silicon on top, in addition to metal layers, another dimension is added to modules' placement, which makes placement and routing more flexible. This can reduce wire lengths and improve timing. Shorter wires can also improve energy efficiency since, in current designs, a significant portion of total power is wasted in wires. 3D stacking also allows dies with different process technologies to be connected by high-bandwidth connections, which is not an option in 2D designs.

The main challenge for 3D stacking is thermal limitations. Silicon temperature can surpass the allowable limit because two or more hot dies are placed next to each other [3]. On the other hand, due to the dark silicon phenomenon [4], all parts of a chip cannot be powered and running at high frequency at the same time. CAD tools use this opportunity to generate designs in 3D that follow the thermal requirements.

DRAM is a good example for using 3D stacking: they have low power density, several DRAM layers can be stacked to consume less area for the same capacity, and access latency can be reduced [6]. This enables them to be connected to the main processing unit in the same chip using protocols like Wide-IO [7] which increases throughput and energy efficiency substantially. Furthermore, a logic layer can be placed in the DRAM stack as an effective and efficient tool for near memory processing.

3D stacking can be used to design more powerful processors [8], make accelerators for near-data computing such as machine learning applications [9], or tape out application-specific chips such as parallel-data biomedical applications [10]. This paper studies these research efforts, highlights the trade-offs for 3D-stacked designs, and proposes potential future work.

The rest of the paper is organized as follows. Section II and III describe the required background about the 3D stacking process and its challenges. Next, 3D stacking opportunities are discussed in a broad range of designs from general purpose to application specific systems in sections IV, V and VI. Finally, section VII evaluates different aspects of 3D architecture and proposes future directions.

II. Background

There are several methods for 3D stacking with differences in connection pitch, resistance and thermal connectivity of vias, and I/O pad placement. This section will describe two main methods as well as briefly describe other methods based on these two.

The first technique, face-to-face bonding, connects two dies with the metal layer sides facing each other. Fig. 2 shows the process for attaching two dies using the face-to-face technique. Copper via stubs are deposited on top of each die's metal layers similar to vias between metal layers (2). Next, the dies are arranged face to face and are subjected to thermo-compression to bond them. Pressure and temperature will cause the stubs to fuse together. For the purpose of mechanical stability and heat conductivity, the space between the two dies is completely populated by die-to-die (d2d) vias. These vias can also become part of the signal path when routing. Then one of the 3D

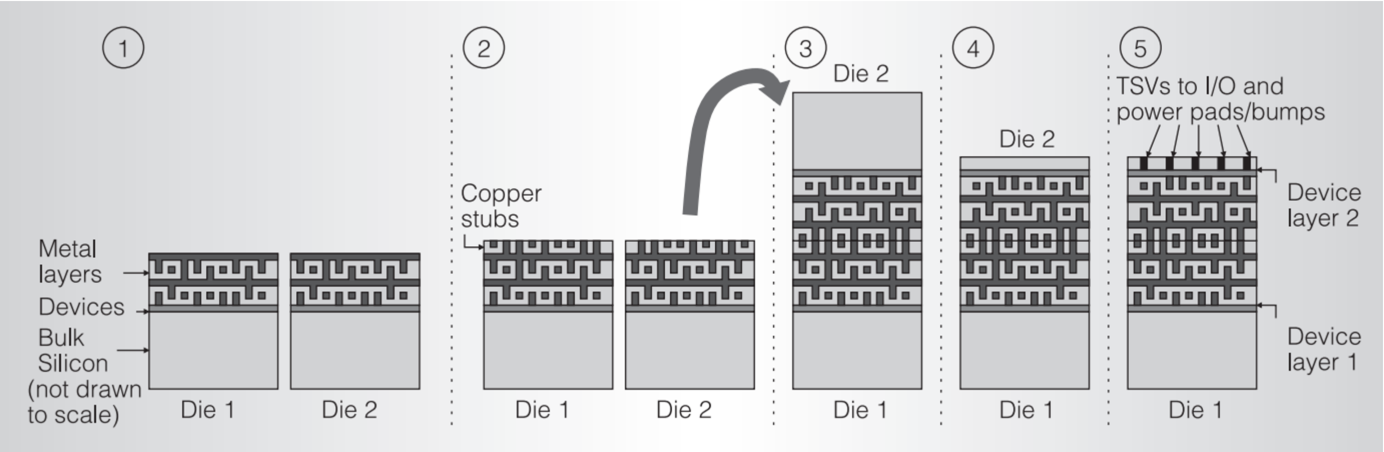

Fig. 2: **Fabrication steps for face-to-face bonding [11].**

stack layers is thinned to only 10 to 20 μm, using chemical-mechanical polishing (CMP) (4). After thinning, we are able to use short through-silicon vias (TSVs) on the top layer for adding pads and bumps for I/O and power with very little power and voltage loss (5). This technique is the most straightforward but it only allows two layers to be connected.

The next technique is called face-to-back bonding where one die is connected from the metal layer side and one die from the silicon side. The process is depicted in Fig. 3. A handle wafer is attached to one of the dies from the metal layer side (1) and this die is thinned to 10 to 20 μm (2). The handle wafer is used to prevent the die from breaking because of its minute height due to thinning. Next, the copper stubs are added to the thicker die similar to the face-to-face method, but for the thinner die they are added to the silicon instead (3). Through thermo-compression the dies are bonded (4) and then the handle is released (5). The thicker die provides the required support for the thinner die not to break. The I/O and power pads can be added, similar to conventional chips, to the available metal layers on top. The benefit of face-to-back bonding is its extendability to multiple layers. However, the procedure is more complicated and the TSVs go all the way through the silicon which makes placement and routing more complicated. In most cases, this increases the pitch for d2d vias. Therefore, to connect more than two layers both face-to-face and face-to-back techniques are usualy used.

Since there are errors in adjusting the two dies for attachment, vias in the bonding interface are thicker and have higher pitch when compared to conventional metal layers. The pitch can range from $10\mu m \times 10\mu m$ to $1\mu m \times 1\mu m$ [11]. The $1\mu m \times 1\mu m$ pitch can be achieved through wafer bonding, by first connecting two wafers together and then cutting out the 3D dies. However, feasibility of this method is based on fabrication circumstances and sometimes half-wafer or even die-to-die attachment is required. Moreover, there are other methods for face-to-face bonding which are easier to implement but have higher pitch, such as bump-to-bump or bond-pad connections instead of d2d vias.

The third technique is called Monolithic 3D IC which is an improved version of face-to-back bonding. In this technique, 3D sequential integration is used which means each layer of silicon is fabricated directly over the previous layer, enabling lower pitch and the use of smaller d2d vias. By replacing TSVs with Monolithic inter-tier vias (MIVs), the contact width would be reduced to about 0.5μm instead of 5μm. MIVs are very similar to inter-metal layer vias and they have very low capacitance ($<< 1fF$) compared to TSVs. Moreover, sequential integration improves alignment accuracy to 0.05μm instead of 1μm[12], but this requires low-temperature fabrication, called process thermal budget constraint, which needs another type of transistors, such as CNFETs that are under development[13].

To summarize, using 3D stacking methods, two or more layers can be placed in a stack and be connected to each other using die-to-die vias. Today's technology can achieve a pitch of 1μm for these vias, meaning a million connections per square millimeter of silicon. This means potentially higher bandwidth as well as faster ad more energy efficient designs. However, physical and thermal constraints introduced in 3D stacking decrease this potential which is the topic of next section.

III. Challenges

There are several physical constraints in chip design such as temperature limit, IR drop, noise control, and process fabrication errors. These must be considered for CAD tools to achieve reasonable die yield. These constraints become more challenging in the 3D regime and thus CAD tools must be updated accordingly. In this section, these additional constraints for 3D designs are described, alongside some high-level rules for CAD tools to address them.

A. Thermal limit

There is a temperature limit for transistors in a silicon layer to work appropriately. In 2D designs, when one side of the die is facing low thermal resistance to the ambient air, it is required to use thermal adhesives and heatsinks. The case becomes more of a challenging for 3D designs, since inner layers are facing hot silicon instead. The larger the number of dies in a stack, the more severe the temperature problem. The most important factor in placement and routing is not to place hot spots of the dies underneath each other. Another technique is voltage down-scaling that is performed by reducing the

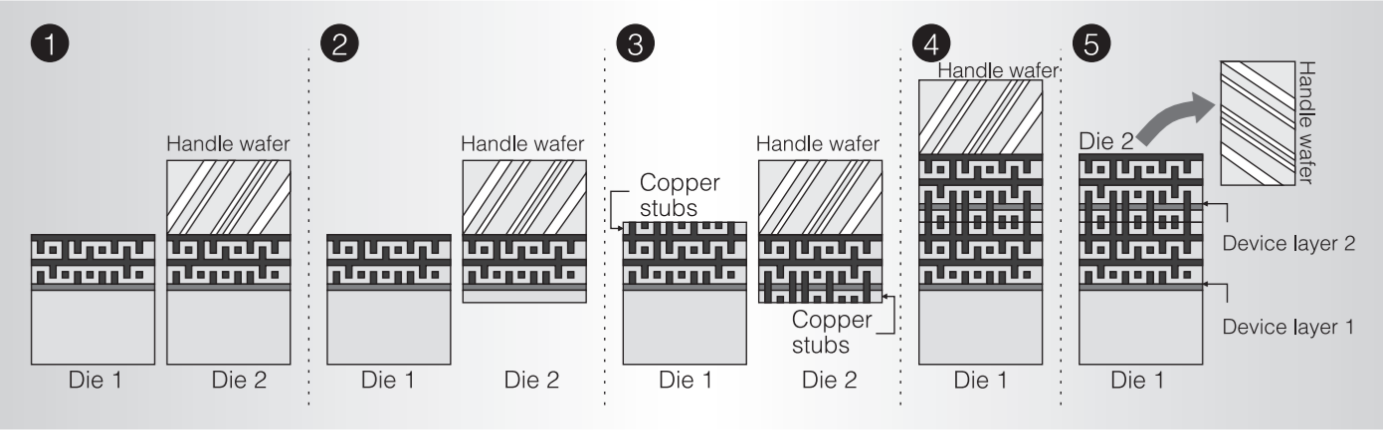

Fig. 3: **Fabrication steps for face-to-back bonding [11].**

operating voltage for power-dense dies. This reduces the power density but degrades the performance and makes 3D design improvements marginal.

B. Power delivery network

IR drop is important to be considered for power delivery network (PDN) routing. The path from input power pads to inner dies becomes longer in 3D, making IR drop considerations more vital. Also PDNs have a significant impact on chip temperature. For example, if a 2D chip is transformed to 3D by simply breaking the die in half and stacking the two smaller dies, the chip area would be halved without any substantial change to the power consumption. The number of available pins and I/O pads are almost halved which requires more current per pin/pad. Moreover, power is delivered to inner dies through die-to-die (d2d) vias, so there is less d2d vias left for signal routing. A rule of thumb is to allocate 30% of d2d vias for power delivery [11]. This degrades the opportunity for shorter signal wire lengths and makes power savings marginal in 3D ICs. It is also more noticeable in monolithic 3D ICs because of its higher integration density. Samal et al. [14] proposed some optimization techniques for PDNs to alleviate these problems:

- If the design is more memory-dominated than interconnect-dominated, leakage will be a significant portion of total power and the PDN's impact on total power increase is negligible and therefore there will be less PDN issues. Yet the memory PDN connections remain important.
- In 3D designs, top metal of the bottom layer(s) is used for d2d vias and cannot be fully utilized for PDNs. The top layer of designs for face-to-face bonding have a similar condition. Hence lower metal layers must be used for PDNs to allow for more flexible placement of d2d vias.
- Having clusters of power/ground supplies allows more continuous space for signal routing and via placement with very little effect on IR drop.
- To reduce the impact of PDN blockages on the signal routing area, the required frequency of PDN wires and the total number of metal layers is determined by overall current demand of a module.

Noise margins are tight in the nanoscale regime and accurate estimations are required for power and signal integrity, especially for chips with multiple power supplies. Reliability of a PDN is dependent on the yield of the d2d vias. Maximum power noise voltage drop has a correlation with d2d reliability: the more reliable the d2d vias, the less the power noise voltage drop. For instance single local TSV failure due to fabrication process or circuit operation can increase maximum voltage variation up to 70% which is significant [15]. Increasing the number of TSVs is helpful but there is a black out region for each TSV where no signals or wires can be routed and no hard macros can be used. There would be less routable area with more TSVs, which makes reducing wire lengths less feasible. Shayan et al. [15] analyzed these trade-offs and provided some insight about the relation between these factors.

Finally, metal layers and power grids must be considered for PDNs in detail. Thier frequency models help identify global and local resonance phenomena, and their time-domain models help identify the worst case supply noise [16]. It has been observed that if the noise response of lower-level tiers is not critical in a certain frequency range, more decoupling capacitance can be used on higher-level tiers. Moreover, resistance between adjacent layers can be reduced using several methods such as increasing the number of d2d vias, connecting TSVs to higher-level metal layers, increasing the wire width, or decreasing wire pitch of metal layers connected to TSVs.

C. Mechanical stress

In 3D stacking, new types of mechanical stress are generated inside the silicon substrate which negatively affect transistor performance. There are two main sources behind this additional mechanical stress which causes the bending shown in Fig. 4. The first is silicon being thinned to less than $50\mu m$ which makes it easier to be bent. The second is local bending stress because of organic adhesive shrinkage. Since the coefficient of thermal expansion (CTE) of an organic adhesive is typically greater than metal micro-bumps, organic adhesive and metal bumps face different volume shrinkage which results in additional mechanical stress. This volume

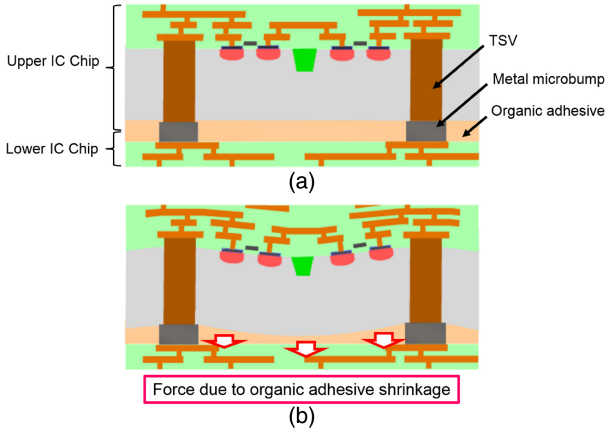

Fig. 4: **Mechanical stress induced by organic adhesive shrinkage in 3D IC (a) before and (b) after curing [17].**

shrinkage is because of cooling down to room temperature after the organic adhesive curing process. Tanikawa et al. [17] proposed a novel local stress evaluation method for 3D DRAM cell arrays. They showed how this stress can change the characteristics of transistors by changing their layout. This change of characteristics should be considered in future designs.

D. Via delay

It is important to consider the added delays from die-to-die (d2d) vias. They are a small resistive-capacitive (RC) component of signal routing. For instance, in 65nm technology, d2d vias increase RC delay by approximately 35% compared to a stack of vias from metal 1 to 9 [11]. A signal that travels between two adjacent dies passes all metal layers of both dies as well as the d2d via. However, this signal delay is still less than a Fo4 delay and significantly less than the delay of a path through one millimeter in a metal layer by about 25 times. So replacing any on-chip interconnect of moderate length can be potentially beneficial. The mentioned numbers are a general estimate since exact latency depends on bonding technology, driving circuits, loading capacitance, and several other factors.

IV. General Purpose Processor Design

Reducing wire lengths and more flexible placement and routing through 3D stacking can improve processors performance and energy efficiency. In this section, after discussing partitioning granularity and its trade-offs, different aspects in which processors can be improved by 3D stacking is studied.

A. Partitioning granularity

Based on the density of d2d vias, a trade-off for 3D-stacked designs is partitioning granularity. For a processor, the coarse-grained use of 3D stacking is merely spreading the major components like L2 cache and cores among the layers, shown in Fig. 5(a). This requires minimal changes to the design and is easy to upgrade but very few d2d vias are used thus the benefits of 3D stacking are not being utilized. The next level of glanularity is distributing smaller components of the design among layers, such as the register file, ALU and first level cache, shown in Fig. 5(b). For example, the register file to ALU wire delay can be reduced enabling the cores to have larger register files. The next level of granularity is inside each module which uses an extensive potential of 3D stacking, shown in Fig. 5(c). The drawback is more complicated design and routing. Therefore, it is generally feasible only for simple repetitive structures like memories. The finest level of 3D partitioning is at the transistor level, shown in Fig. 5(d). For instance, n-channel metal oxide semiconductor (NMOS) transistors are on one die and PMOS transistors are on the other die. However, today's via density is still not sufficient for this and it needs powerful CAD tools which are still being developed. Potential future work could be to gain benefits from this distinguishing factor among transistors to reduce the design cost and complexity.

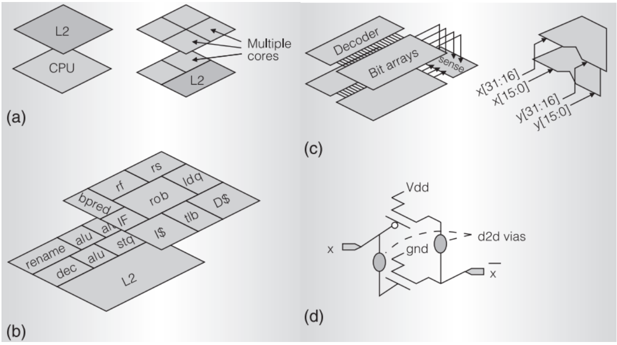

Fig. 5: **Partitioning granularity: (a) entire core, (b) functional unit blocks, (c) logic gates, and (d) transistors [11].**

B. Cache improvements

3D stacking can be used very efficiently for memories due to their homogeneous design as well as their low power density, which makes it easier to apply the new technique with very few limitations. The simplest improvement would be placing the cache on the top layer and the logic part on the bottom to reduce the wire length between logic and cache. However, since the number of SRAM ports is limited in this scenario, it does not use the available throughput and flexibility of vertical routing, which means negligible latency and power improvements.

Next step is to make alterations in the arrangement of banks, as shown in Fig. 6(b) versus (a), which reduces the farthest distance from a SRAM cell to a port by 50%, decreasing the max and average latency. Moreover, the size of the SRAM can be increased with this method; with 2 layers the memory size can be doubled without any latency overhead. Using more than two layers can improve it even further. Bank-stacking still underutilizes the possible connection density in 3D, but it has the benefit of using the same conventional 2D SRAM arrays and getting the savings from global routing.

Array splitting is a more fine-grained technique which reduces the length of wordlines or bitlines, resulting in lower access latency. Fig. 6(c) shows a conventional SRAM cell arrangement. As shown in Fig. 6(e), a wordline can be replaced

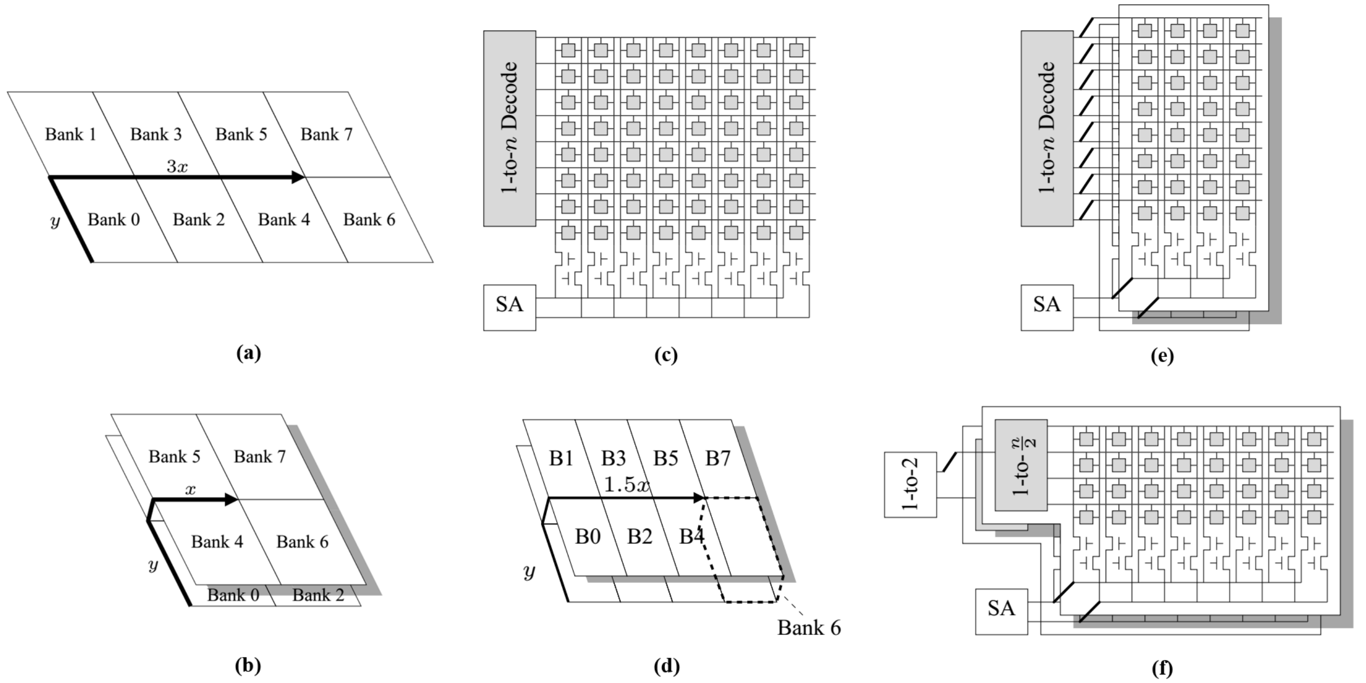

Fig. 6: **Reducing wire length inside SRAM module. (a) conventional SRAM banks. (b) SRAM banks in two layers. (c) conventional SRAM array. (d) SRAM banks after array splitting. (e) wordline splitting. (f) bitline splitting [5].**

by a pair of parallel wordlines so the column select mux would be split. The resistance is reduced by half and the capacitance is not changed. Fig. 6(f) shows the split in row instead of column, causing the bitline length to become half. This requires using two 1-to-n/2 decoders and one 2-to-1 decoder instead of a 1-to-n decoder. To reduce power, one of the larger decoders is not powered, based on the output of the smallest decoder. Both techniques of array splitting would reduce area by half, which further decreases the access latency in bank level, shown in Fig. 6(d). The orientation of split is determined by access latency, whether it is the wordline or bitline delay. For example, array splitting can provide 21.5% latency reduction and 30.9% energy reduction for a 512KB SRAM just within 2 layers [5].

Furthermore, as we will see in the next section, DRAM layers can be 3D-stacked and connected to the main processor via protocols such as Wide-IO. This additional memory can be used as a cheap, large, last-layer cache to improve hit rate.

C. Scheduling units

One of the bottlenecks for a superscalar processor is the instruction scheduler. The delay of the wake-up logic can be expressed as $Delay = T_{tagdrive} + T_{tagmatch} + T_{ormatch}$. Tag-drive is the time taken by buffers to drive the tag bits; tag-match is the time for comparing the tag in CAM structure; and or-match is the required time to OR individual match lines. Issue width and window size affect the tag-drive time which is the dominant part of the total delay. 3D stacking can help significantly in reducing the tag-drive time by reducing the wire length, similar to SRAM, to make the instructor scheduler less of a bottleneck, or enable the processor to use greater issue width and windows size which results in increased instruction level parallelism. Vaidyanathan [8] showed that locating half of each instruction in each layer has significantly better improvement over locating half of each tag line in each layer. They reported 44%, 22% and 16% improvements going from conventional design to 2-layer to 4-layer 3D design, while the other partitioning has only 4% improvement in two-layer 3D design. The power improvements were 23%, 6% and 10% respectively [8].

D. ALU functions

The Kogge-Stone (KS) adder is one of the fastest adders in CMOS technology. Its critical path depends on the number of inputs and wire delay dominates its performance. Fig. 7(a) shows this adder in 2D and Fig. 7(b) show it in a 4-layer 3D design (the top layer is not shown). Due to better organization, the critical path, shown in thick red arrows, is improved by 4 times in terms of cell width. Vaidyanathan et al. [8] reported 20.23%, 23.6% and 32.7% performance improvements and 8%, 15% and 22% power improvements over 2D design for 2-layer to 4-layer 3D designs.

Another example for ALU functionality that can be improved by 3D is logarithmic shifter. Linear dependence of wire length in 2D layout of an 8-bit log shifter and its implementation in 3D within two layers are shown in Fig. 7(c) and (d). Instead of a 10 cell-width wire delay, in 3D design there is only a 4 cell-width wire delay and two d2d via delays. Vaidyanathan et al. [8] reported 13.4% and 28.4% performance improvement and 6.5% and 8% power improvement for Log16 and Log32 implementations within two layers.

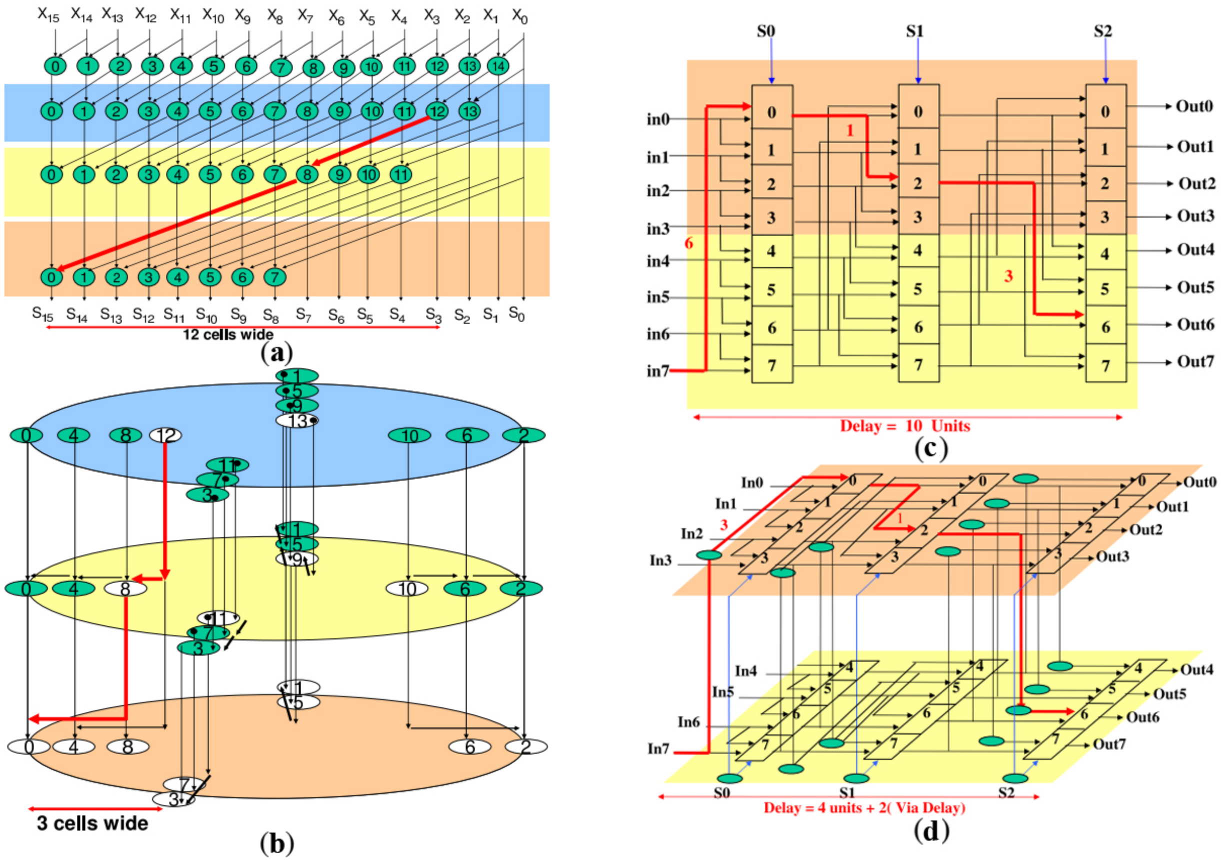

Fig. 7: **ALU functional units: (a) 16-bit KS adder in 2D and (b) 3D, (c) log shifter in 2D and (d) 3D. Critical path is shown in bold line [8].**

E. Register file

Similar to cache cell array optimizations, register files can benefit from 3D stacking. Partitioning registers is equivalent to wordline split where half of the registers are put in another 3D layer. It could be bit-partitioned which is equivalent to bitline splitting in caches. Another method would be port-split where each die contains the bitlines, wordlines and access transistors for half of the ports (either read or write) [18]. All of these methods reduce the area. A combination of these methods can be used to increase the improvements based on the number of layers and the available 3D stacking methods.

Puttaswamy et al. [18] reported up to 16.8% and 24.1% improvements for register files of 128 and 256 registers using 2 layers, and 28.6% and 36.0% using 4 layers. The best energy per access improvements for these register file sizes are 21.5% and 58.5% in 2 layers, and 36.3% and 67.2% in 4 layers. Register partitioning has shown to be the best case for energy improvements. The best method for latency improvement is different among different register file sizes and number of layers, yet bit splitting shows the best results on average.

F. Clock network

The footprint reduction results in shorter distances for the clock distribution network. This decreases both power consumption and timing margins which are required due to clock skew and jitter. Decreasing wire length also allows for some pipeline stages to be removed. For example, Loh et al. [11] showed how they reduced the number of pipeline stages by 25% in Intel Pentium 4. These pipeline modifications increased the performance by 8% and decrease the power by 34%, running at 8% lower operating voltage and thus frequency, while reaching the same maximum temperature. We can also increase the register file or branch predictor size without increasing critical timing paths to improve their performance.

G. Multi-core

3D architecture provides faster on-chip networks because of shorter communication path. Also, larger caches with the same latency or faster caches with the same size are possible. These two are theoretically helpful for multi-core processors. For example, dynamic heterogeneous architecture is only possible in coarse-grained granularity in the 2D regime due to high communication overhead for resource sharing. In heterogeneous architecture, different cores have different capabilities and proper mapping of applications to cores can improve power efficiency in the dark silicon era. However, static scheduling is not efficient in many cases and runtime scheduling requires dynamic heterogeneous architecture. Faster on-chip network and lower-latency memories in 3D designs can make fine-grained dynamic heterogeneous architecture feasible

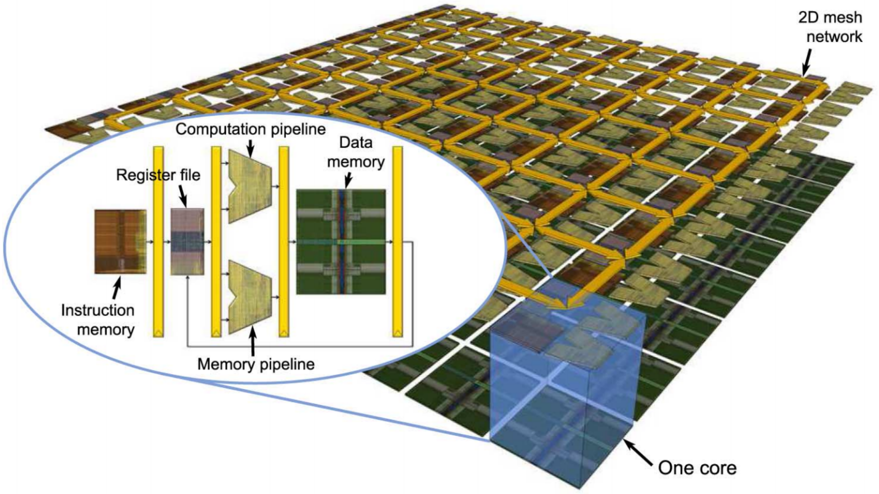

Fig. 8: **3D-MAPS architecture, a massively parallel processor with stacked memory, consisting 64 cores and 64 memory blocks of 4KB SRAM [19].**

[20].

Lastly, shorter wire lengths make larger designs feasible. For example, 3D-MAPS [19] is a massively parallel processor with stacked memory, consisting of 64 cores and 64 memory blocks of 4KB SRAM, shown in Fig. 8. They use face-to-face bond-pads between memories and cores, achieving maximum memory bandwidth of 70.9 GB/s at 277 MHz. Another example is Centip3De [21] which has two stacked dies with 64 ARM M3s operating at near-threshold voltage and 8 DRAM controllers connected to cores with a 128-bit bus providing 2.23 GB/s. 3D enables higher bandwidth both for caches and connection to DRAMs. To address thermal constraints and also fully utilize the bandwidth, cores were run at 10MHz while caches were run at 40MHz. They reached a peak performance efficiency of 3930 DMIPS/W in a 130nm technology (DMIPS stands for Million Instructions per Second).

V. Near-Data Computing Accelerators

Demand for data mining and machine learning applications are increasing and amount of data to be processed is growing exponentially,thus there is a need for improvements in computation speed. Specialized hardware acceleration is used to overcome dark silicon problem, and for data-intensive applications, processing-in-memory (PIM) techniques were introduced[6]. As seen in previous sections, memories gain the most benefit from 3D technology due to their homogeneous structure and low power density. Stacked memories can also be applied to DRAMs which introduce new types of memories called 3D DRAMs. These DRAMs have a much lower area footprint for the same capacity which makes it possible to be placed in the same package as the computing unit. Moreover, due to the possibility of mixing different process technologies, a logic layer in a more advanced process technology can be added to the DRAM stack resulting in substantial opportunity for PIM and near-data computing. In this section, 3D DRAM architecture and some accelerators that use this opportunity are described.

A. 3D DRAM

Bank-level optimization is highly beneficial for DRAMs with little overhead. DRAM dies can be stacked vertically using the conventional DRAM banks, sharing a TSV bus which is different than the common bus inside each layer. Still, the bottleneck is the intra-bank access time and the TSV bus is idle most the of time. To achieve a higher bandwidth, each rank can be broken into several layers [23], but going deeper and changing the DRAM cell arrays is not that beneficial. DRAM bitcells are already optimized for density, and with restricted standards such as DDR3 and DDR4, making fundamental changes is expensive and requires more consideration.

Techniques such as better scheduling [6] and increasing the number of I/O connections to the DRAM module [24] can improve the bandwidth and TSV bus utilization. Due to the large size of DRAM dies, there is no issue with increasing the number of pin connections and their low power density makes stacking practical. Banks from different layers are parallel in stacked DRAMs and the I/O limitation of a conventional DRAM is reduced substantially. Increased throughput and reduced area of DRAMs makes on-chip DRAMs possible for several applications, and new I/O connections such as Wide-IO are developed accordingly. Due to short wires between the 3D DRAM and the processing unit, the connection is highly energy-efficient and can achieve high bandwidth rates.

For memory-intensive applications, Micron introduced Microns Hybrid Memory Cube (HMC), a stacked DRAM with a logic layer underneath. This logic layer can provide efficient near-data computation before sending the data to the main processor. HMC utilizes high-speed SerDes links as I/O and reaches a bandwidth of 80GB/s over 128 pins and has low energy per bit, 10.48 pJ/bit versus DDR3's 70 pJ/ bit [25].

Moreover, CACTI-3DD [26] is an architecture-level modeling framework for 3D off-chip DRAM as main memory. It includes models for power, area, and timing which helps architectures to simulate their designs that have such DRAMs.

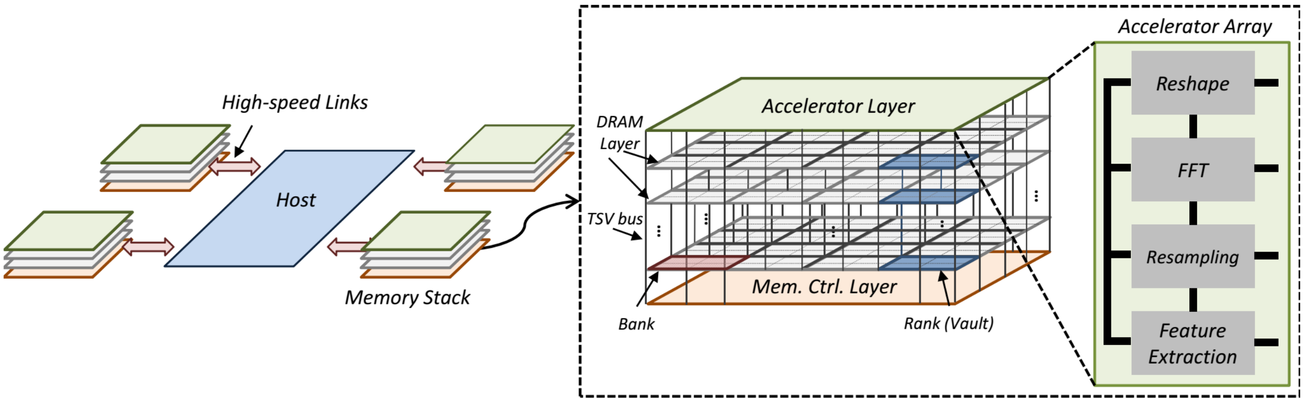

Fig. 9: **Overall architecture of 3D-stacked memory side accelerator system [22]. Each stack of 3D DRAM has an additional accelerator layer implemented in the logic layer of memory stack.**

B. *Improved near-data accelerators*

Guo et al. [22] proposed 3D-Stacked Memory-Side Accelerator (MSA), shown in Fig. 9. Each stack of 3D DRAM has an additional accelerator layer implemented in the logic layer of the HMC. This layer is an array of accelerators, such as FFT, based on the application requirements. The main processor is connected to four of these memory stacks. They implemented a software stack to allow the processor in the decode stage to utilize accelerators for computation. They reported 179x and 96x better energy efficiency than the Intel Haswell processor for the FFT and matrix transposition algorithms.

A similar design introduced 3D-stacked logic-in-memory (LiM) [6] system with only a single stack of DRAM, including a single or multiple LiM layers, to accelerate important data-intensive computation based on the application. For instance, for different memory bandwidth configurations, they reported 1 to 100 Gflops for SpGEMM (Generalized sparse matrix-matrix multiplication). They achieved up to one order of magnitude performance improvements and two orders of magnitude power efficiency improvements over Intel Xeon machines.

For applications with large datasets utilizing MapReduce, it is useful to perform part of the data processing near the data instead of going through the memory hierarchy to the processor. Pugsley et al. [25] used HMC and reported reduced execution time by 23.7% (WordCount) to 92.7% (RangeAgg) and energy by 42% (WordCount) to 93% (RangeAgg) compared to the EECore baseline.

Another category of memory-intensive applications are applications which require data reorganization, such as shuffle, transpose, swap, layout transformations, and pack/unpack. Data reorganization is used in several scientific computing applications such as signal processing, molecular dynamics simulations and linear algebra computations. This relocation is very expensive because of the round-trip to the processor through the memory hierarchy. Akin et al. [23] proposed a mathematical framework to optimize these operations and make them able to run on a DRAM-aware reshape accelerator integrated within the 3D-stacked DRAM. This accelerator performs concise address remapping on the fly to improve parallelism in accessing memory by noting the access pattern. Also, it can perform reorganizations within the main memory instead of going through the memory hierarchy. For example, they achieved an external bandwidth of 320 GB/s in comparison to 25.6 GB/s for CPU and 288GB/s for GPU for a 1GB 3D DRAM that uses only 30W. This DRAM has 8 links, four 3D layers, and 2048 TSVs between them. The Data Reorganization Unit (DRU) accelerator only consumes 0.6% of the 30W. Moreover, this DRU can be integrated into the CPU or GPU memory subsystem. They showed that having the logic layer for the DRU inside the 3D DRAM has a 2.2x performance and a 7x energy improvement when compared to a system with the same 3D DRAM but a CPU-side DMA instead of the logic layer.

Big data processing systems require a very high memory bandwidth to achieve the required response time. Lowe-Power et al. [27] used 3D DRAMs to increase the bandwidth and they found it helps big data processing to have a 256 times better response time as the cost of 50 times more power requirement. Each chip consists of both processor and memory, and increase in memory size requires replication of processor as well, increasing total server energy. Due to servers' energy restriction, 3D DRAMs are not widely used yet, but it could be improved in the near feature by reducing the compute chip power and increasing memory densities of each DRAM layer.

VI. Application-Specific 3D Designs

Mixed process technology in the same stack provides great opportunities for mixed analog and digital signal designs, for example designs with RF/wireless and networking components. Fine-grained dynamic voltage and frequency scaling (DVFS) can become possible by an on-stack DC-DC converter layer [11]. As shown in the previous section, HMC benefits from this feature by having a logic layer in a different process technology than the DRAM layers. Another feature introduced in 3D architecture is more modular chips with potential upgradability because of its layered structure. Some application-specific accelerators and designs could use these features of 3D design and highly benefit from them. This section will review some of these designs.

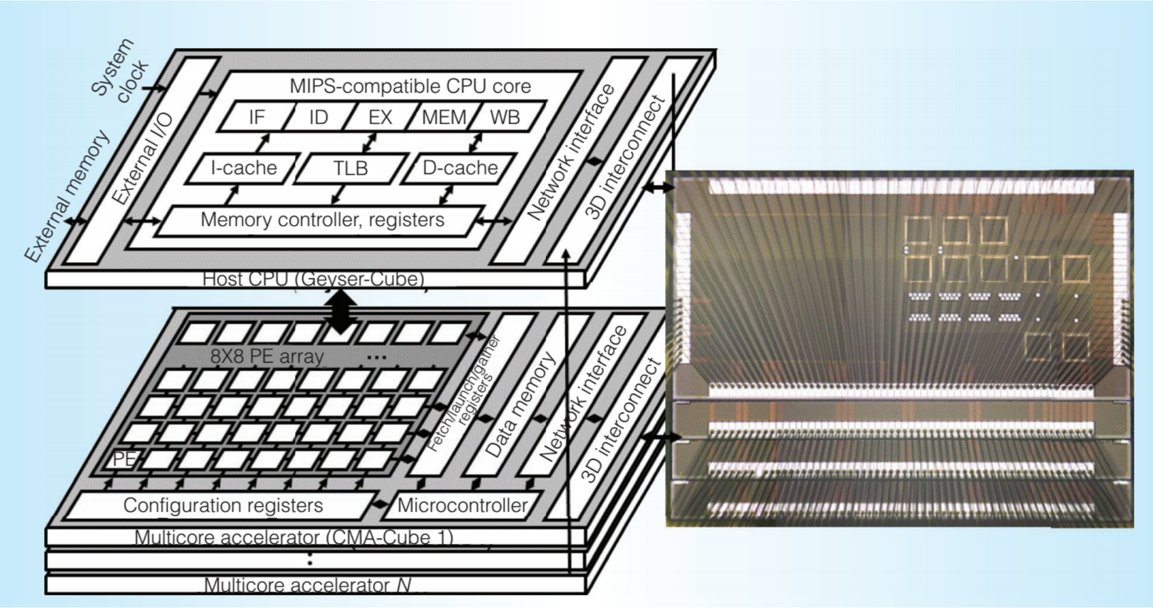

Fig. 10: **Scalable 3D heterogeneous multi-core processor for streaming applications. Geyser-Cube is the base die on top with main processor, and CMA-Cubes are accelerator dies with several processing elements. Network-on-chip on top of the ThruChip Interface (TCI) provides communication among dies [28].**

A. Data-parallel applications

Woo et al. [29] proposed a broad-purpose accelerator architecture for data-parallel applications called parallel-on-demand (POD). POD integrates a specialized SIMD-based die layer atop a CISC superscalar. In the decode stage, the main processor can decide to use the SIMD accelerator. They implemented the required software stack and also modified the ISA to implement the required steps in data transmission between the SIMD accelerators and the processor. For example, an 8×8 POD array in 45nm running at 3GHz would consume 108.7W in the accelerator layer and 148.7W total. 3D enables low-latency (two cycles) communication between adjacent POD accelerators with very low energy overhead. Using four 32 GB/s DRAM controllers, they achieved more than 800 Gflops for dense matrix-matrix multiplication or financial option modeling applications. Due to their low-latency communication, they gained performance improvement by increasing the number of POD accelerators with marginal energy overhead.

B. Streaming applications

Streaming applications commonly consist of a series of tasks on a certain amount of data which can be assigned to an accelerator and be performed in a pipelined manner. Miura et al. [28] proposed the Cube-1 architecture as a low-power processor for streaming applications. They used two new features of 3D stacking: system level modularity for each layer and different operating voltages among layers. By implementing well-defined network and connection interfaces among layers, they achieved extendability. There is a base die running at 1.2v and several similar accelerator dies running at 0.5v to 1.2v using dynamic voltage scaling.

The base die, called Geyser-Cube, has a MIPS-R3000 compatible processor to maintain external I/O. The accelerator layer, called Cube-1, is an ultra-low-energy coarse-grained reconfigurable architecture (CGRA) named Cool Mega Array (CMA). It has several combinational processing elements which are mapped statistically based on applications' data flow, with a simple programmable micro-controller for data management. This layer also has a data memory and network interface, shown in Fig. 10. CMA-Cube achieves 2.72 GOps with only 11.2 mW power consumption without any 3D thermal limit.

Communication is based on a global mapping for the main processor and the accelerators' memories, which is performed through the network interface on top of the ThruChip Interface (TCI). Network interface includes a distributed DMA mechanism to implement this global mapping. TCI uses few d2d vias and is not using the bandwidth provided by 3D stacking, but it enables extendability by having a fixed position and providing a serial communication between dies. In contrast to communication overhead and running the accelerators at lower voltages, they achieved 3.15x speed-up using 3 accelerator layers when compared to just using the base layer for processing 128x96 JPEG images for three tasks of YUV-to-RGB conversion, inverse discrete cosine transform, and inverse quantization. Based on the extendability, they can increase performance by using more accelerator layers.

C. Machine learning applications

Machine learning algorithms such as Neural Networks (NN) have become popular as a modern technique for data processing and are well suited for accelerators. NN accelerators require high cross-layer bandwidth and they have low power density due to large memories and simple computation, which makes it a perfect match for 3D-stacked designs. Belhadj et al. [9] showed that 3D can meet the inter-layer communication requirement using micro-bumps, which was a barrier for 2D designs. They reported 1.64x bandwidth improvement consuming only 48% energy and 66% silicon when compared to a 2D implementation.

D. Biomedical applications

Ultrasound imaging is a non-invasive medical imaging method with high accuracy, safety, and ease of use. Although the imaging itself does not require high power, it is very compute-intensive which requires external processing and cannot be made as a single handheld device. Sampson et al. [10] proposed Sonic Millip3De, a three-dimensional (3D) ultrasonic imaging device that benefits from 3D stacking opportunities such as mixing analog and digital layers. They introduced an accelerator for 3D ultrasound beam-formation, which is the most compute-intensive part, using 3 layers of 3D stack. In simulation, it uses only 16W of energy in 45nm process technology which is 400 times better than a conventional DSP solution. There is a 5W power limit for handheld imaging devices which they expect to be achieved in 11nm process technology.

The first layer is the transducer array which is analog and can be manufactured in an older process technology since these sensor locations are determined by the wavelength of the transmitted signal. The next layer is the power-hungry layer consisting of analog-to-digital converters (ADCs) and an SRAM array to save the read data from sensors. The third layer is the computer layer which includes a main processor and cache, as well as processing elements (PEs) for each SRAM. These PEs perform the beam-formation calculations before sending the data to the main processor.

Every 12 sensors share an ADC and they are sampled consecutively at 40 MHz to satisfy the sampling frequency requirement and their output is stored into the corresponding SRAM. After all 12 data values are ready, they are sent together to the accelerator layer. For a system of 12288 sensors with 12-bit ADC outputs, there are 1024 ADCs and 1024 SRAMs to store 4096 12-bit samples. 24,000 face-to-face bonds were used for routing data and address signals between the ADC/SRAM layer and the processing layer. To achieve one frame per second, load beam-forming constants and read/write image data require 6.2GB/s and 5.5GB/s bandwidths respectively. They achieved a memory bandwidth of 38.4 GB/s from a 192-bit memory channel, which is comprised of 6×16 2Gb LPDDR2-800 memories to provide the necessary capacity of 1.5 GB.

Another advantage that 3D provides for this design is upgradability. If another waveform is desired for sensors, only the sensor layer needs to be replaced, as long as the connection position for each group of sensors is fixed.

VII. ANLYSIS *and* CONCLUTSION

Following Moore's law and doubling the number of transistors per area is becoming harder to achieve and 3D stacking looks to be a solution to this problem. Using two layers doubles the number of transistors per unit area, and it can be increased further by adding more layers. However, in practice, it has stricter power challenges than conventional 2D designs, even more restricted than dark silicon. Placing several hot silicons next to each other would increase the temperature and each die must be run at a lower voltage or frequency to satisfy the temperature limit for CMOS technology. Cooling the inner dies are more challenging because they are facing another hot silicon rather than facing a low thermal resistance to ambient air. Although silicons thermal conductivity is very good, superposition of two heat sources would still increase the temperature. Moreover, the path from power pins or pads to the innermost die is longer than that in the 2D case, passing through die-to-die (d2d) vias which are thinner and hence more resistive. This makes power delivery a challenge and requires dedicating a significant portion of d2d vias for the power delivery network (PDN). These challenges become more and more difficult as the number of stack layers increase. Therefore, although more transistors are placed per unit area, their usability and efficiency is worse than even post-Dennard scaling in 2D designs.

The promising aspect of 3D stacking looks to be more flexible placement and routing since there is another die on top instead of metal layers. Having $1\mu m$ pitch means one million d2d vias per square millimeter, enabling very high bandwidth between layers. However, typically up to 10 metal layers are used in 2D designs and the routing is actually happening in 3D space. Thus 3D stacking is not a ground-breaking opportunity. In addition, due to thermal issues, the placement of blocks are not fully flexible, and because of PDN issues not all of the d2d vias are available. Nonetheless, the ability to have a die in the 3rd dimension of space can make way for designs which were not feasible in 2D. If thermally possible, distant modules in a 2D design can be placed close to each other in 3D and the wires connecting them would become significantly shorter. This would reduce the latency of the system or at least change the timing bottleneck while less energy is wasted in the wires. Moreover, by breaking a large die into smaller dies and stacking them, the clock network hierarchy can be improved and the distance between the farthest point and the main source would be reduced significantly. This translates to less energy used by the clock network and less severe jitter and skew constraints, enabling the use of higher clock frequencies.

For instance, register file access latency forced register file clustering for superscaler processors to meet the timing requirements. By splitting the register file in two 3D layers the length of wires from the cell array to functional units is decreased, enabling larger register files or even higher frequencies in general-purpose processors. From a thermal asect, this is feasible because of the lower power density of the SRAM cells compared to active logic. The same scenario holds for the instruction scheduler which can be the frequency bottleneck. In contrast, the idea of making ALU functional units faster does not seem promising. Generally, they are ran at higher frequencies and are the potential hot spots of the chip, making it infeasible to spread them among 3D layers on top of each other. Also ALU functions are typically not the actual bottleneck of general-purpose systems.

Memory latency and cache hit rate are still performance bottlenecks in current general-purpose systems. Using 3D to reduce cache latency would improve the performance marginally, but the cache misses which force the processor to acquire data from the main memory are still the bottleneck. Nevertheless, the ability to increase cache capacity with the same latency or even using a large 3D DRAM last layer cache on the same chip as the processor can improve the performance. Especially in some embedded systems, the main memory can be fully fitted on the processor chip by using 3D DRAMs, making it possible for the memory to no longer be the bottleneck.

Network-on-chips (NoCs) can greatly benefit from more flexible placement for multi-cores. Previously, caches were placed next to cores and hence the distance between adjacent cores in a mesh grid would not allow for fast communication among cores. However, by putting the memory on another 3D layer, the cores distance shrinks and faster communication is viable among them. This idea is practical because of the low power density of SRAM cells and limited bus width for caches. Similarly 3D stacking opens up new possibilities like dynamic heterogeneous architecture which were not practical in 2D. However, current cores are well-designed for the 2D scenario with high communication delay and they mostly use cache coherency protocols to minimize communication. Therefore using 3D NoCs would have only marginal benefit for current cores, which can be a future direction for computer architects to change the processor design based on this new opportunity.

3D DRAMs fit well with 3D stacking technology, without needing to change the base layers in DDR or similar standards. Reducing the area and improving the bandwidth are promising points for 3D DRAMs, while they are not thermally limited for a reasonable number of layers. The ability to model them using CACTI-3DD and the Microns Hybrid Memory Cube (HMC) which adds a logic layer to the 3D DRAM stack presents further encouragement to design new systems. However, the main latency bottleneck is the DRAM access time inside each layer, not in the communication. Regardless, 3D DRAMs can improve the throughput significantly, and when placing it inside the chip and using protocols like Wide-IO, it can decrease power substantially. For memory-intensive applications, such as machine learning or streaming applications that are not highly sensitive to latency but need very high bandwidth, 3D DRAM is a proper match. Unfortunately 3D DRAMs are still not suitable for big data applications to be run on servers since the required DRAM capacity is very high and cannot be put inside the chip to gain energy benefits. Although they provide a high bandwidth, they are not widely used for servers because of servers power limit. Using more advanced technology nodes to shrink the size of DRAM and redesigning DRAMs for a 3D-stacked scenario are potential opportunities to enable server applications to benefit from 3D DRAM's high bandwidth.

Mixed process integration is an outstanding new feature in 3D which was not available in 2D designs. It makes way for many new mixed signal designs which were not possible before. High bandwidth was only possible for designs on the same die, meaning they needed to use the same process technology. Microns HMC is using this to integrate a logic layer of a more advanced node with the DRAM layers, resulting in promising near-data computing accelerators. Several Systems-on-Chip (SoCs) can substantially benefit from this, such as having the modem and communication layer in another process node which is more suitable. Some parts of the computation can be done using analog accelerators which are implemented in another technology node. For example, Sonic Millip3De design used this opportunity very effectively and implemented a 3D ultrasound imager with an order of magnitude less energy consumption. It greatly utilizes the available bandwidth between different layers and avoids running into thermal issues by having two low-power homogeneous inner layers, and only a single power-dense outer layer which can be cooled similar to conventional designs.

Finally it enables designs to be more systematic and extendable by providing a well-designed interface among layers, such as in Cube-1 design, so a layer can be replaced or more layers can be added to a design. This makes the systems more modular, reusable, and even upgradable. For example, a generic memory layer can be designed and used for many designs by providing the required interface. Another example, which benefits from the ability of using different process technologies as well, is the ability to make cheaper prototypes in an older technology with lower Non-Recurring Engineering (NRE) cost. After achieving the desired results, one can upgrade the system to a better process technology. This can potentially open up several opportunities for more Application-Specific Integrated Circuits (ASICs) to be implemented in the near future.

References

[1] R. Wang, E. F. Young, and C.-K. Cheng, "Representing topological structures for 3-d floorplanning," in *Communications, Circuits and Systems, 2009. ICCCAS 2009. International Conference on*, pp. 1098–1102, IEEE, 2009.

[2] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCauley, P. Morrow, D. W. Nelson, D. Pantuso, *et al.*, "Die stacking (3d) microarchitecture," in *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*, pp. 469–479, IEEE, 2006.

[3] X. Hu, P. Du, and C.-K. Cheng, "Exploring 3d power distribution network physics," in *ASIC (ASICON), 2011 IEEE 9th International Conference on*, pp. 562–565, IEEE, 2011.

[4] M. B. Taylor, "Is dark silicon useful?: harnessing the four horsemen of the coming dark silicon apocalypse," in *Proceedings of the 49th Annual Design Automation Conference*, pp. 1131–1136, ACM, 2012.

[5] K. Puttaswamy and G. H. Loh, "Implementing caches in a 3d technology for high performance processors," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, pp. 525–532, IEEE, 2005.

[6] Q. Zhu, B. Akin, H. E. Sumbul, F. Sadi, J. C. Hoe, L. Pileggi, and F. Franchetti, "A 3d-stacked logic-in-memory accelerator for application-specific data intensive computing," in *3D Systems Integration Conference (3DIC), 2013 IEEE International*, pp. 1–7, IEEE, 2013.

[7] D. W. Kim, R. Vidhya, B. Henderson, U. Ray, S. Gu, W. Zhao, R. Radojcic, and M. Nowak, "Development of 3d through silicon stack (tss) assembly for wide io memory to logic devices integration," in *Electronic Components and Technology Conference (ECTC), 2013 IEEE 63rd*, pp. 77–80, IEEE, 2013.

[8] B. Vaidyanathan, W.-L. Hung, F. Wang, Y. Xie, V. Narayanan, and M. J. Irwin, "Architecting microprocessor components in 3d design space," in *VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on*, pp. 103–108, IEEE, 2007.

[9] B. Belhadj, A. Valentian, P. Vivet, M. Duranton, L. He, and O. Temam, "The improbable but highly appropriate marriage of 3d stacking and neuromorphic accelerators," in *Compilers, Architecture and Synthesis for Embedded Systems (CASES), 2014 International Conference on*, pp. 1–9, IEEE, 2014.

[10] R. Sampson, M. Yang, S. Wei, C. Chakrabarti, and T. F. Wenisch, "Sonic millip3de: A massively parallel 3d-stacked accelerator for 3d ultrasound," in *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, pp. 318–329, IEEE, 2013.

[11] G. H. Loh, Y. Xie, and B. Black, "Processor design in 3d die-stacking technologies," *Ieee Micro*, no. 3, pp. 31–48, 2007.

[12] P. Batude, T. Ernst, J. Arcamone, G. Arndt, P. Coudrain, and P.-E. Gaillardon, “3-d sequential integration: a key enabling technology for heterogeneous co-integration of new function with cmos,” *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, no. 4, pp. 714–722, 2012.

[13] M. M. Shulaker, T. F. Wu, M. M. Sabry, H. Wei, H.-S. P. Wong, and S. Mitra, “Monolithic 3d integration: a path from concept to reality,” in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2015*, pp. 1197–1202, IEEE, 2015.

[14] S. K. Samal, K. Samadi, P. Kamal, Y. Du, and S. K. Lim, “Full chip impact study of power delivery network designs in monolithic 3d ics,” in *Computer-Aided Design (ICCAD), 2014 IEEE/ACM International Conference on*, pp. 565–572, IEEE, 2014.

[15] A. Shayan, X. Hu, H. Peng, C.-K. Cheng, W. Yu, M. Popovich, T. Toms, and X. Chen, “Reliability aware through silicon via planning for 3d stacked ics,” in *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE’09.*, pp. 288–291, IEEE, 2009.

[16] X. Hu, P. Du, J. F. Buckwalter, and C.-K. Cheng, “Modeling and analysis of power distribution networks in 3-d ics,” *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 21, no. 2, pp. 354–366, 2013.

[17] S. Tanikawa, H. Kino, T. Fukushima, M. Koyanagi, and T. Tanaka, “Evaluation of in-plane local stress distribution in stacked ic chip using dynamic random access memory cell array for highly reliable three-dimensional ic,” *Japanese Journal of Applied Physics*, vol. 55, no. 4S, p. 04EC07, 2016.

[18] K. Puttaswamy and G. H. Loh, “Implementing register files for high-performance microprocessors in a die-stacked (3d) technology,” in *Emerging VLSI Technologies and Architectures, 2006. IEEE Computer Society Annual Symposium on*, pp. 6–pp, IEEE, 2006.

[19] D. H. Kim, K. Athikulwongse, M. B. Healy, M. M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y.-J. Lee, D. L. Lewis, T.-W. Lin, *et al.*, “Design and analysis of 3d-maps (3d massively parallel processor with stacked memory),” *Computers, IEEE Transactions on*, vol. 64, no. 1, pp. 112–125, 2015.

[20] V. Kontorinis, M. K. Tavana, M. H. Hajkazemi, D. M. Tullsen, and H. Homayoun, “Enabling dynamic heterogeneity through core-on-core stacking,” in *Proceedings of the 51st Annual Design Automation Conference*, pp. 1–6, ACM, 2014.

[21] R. G. Dreslinski, D. Fick, B. Giridhar, G. Kim, S. Seo, M. Fojtik, S. Satpathy, Y. Lee, D. Kim, N. Liu, *et al.*, “Centip3de: A 64-core, 3d stacked near-threshold system,” *Micro, IEEE*, vol. 33, no. 2, pp. 8–16, 2013.

[22] Q. Guo, N. Alachiotis, B. Akin, F. Sadi, G. Xu, T. M. Low, L. Pileggi, J. C. Hoe, and F. Franchetti, “3d-stacked memory-side acceleration: Accelerator and system design,” in *In the Workshop on Near-Data Processing (WoNDP)(Held in conjunction with MICRO-47.)*, 2014.

[23] B. Akin, F. Franchetti, and J. C. Hoe, “Data reorganization in memory using 3d-stacked dram,” in *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, pp. 131–143, IEEE, 2015.

[24] A. Farmahini-Farahani, J. H. Ahn, K. Morrow, and N. S. Kim, “Nda: Near-dram acceleration architecture leveraging commodity dram devices and standard memory modules,” in *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*, pp. 283–295, IEEE, 2015.

[25] S. H. Pugsley, J. Jestes, R. Balasubramonian, V. Srinivasan, A. Buyuktosunoglu, A. Davis, and F. Li, “Comparing implementations of near-data computing with in-memory mapreduce workloads,” *Micro, IEEE*, vol. 34, no. 4, pp. 44–52, 2014.

[26] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. B. Brockman, and N. P. Jouppi, “Cacti-3dd: Architecture-level modeling for 3d die-stacked dram main memory,” in *Proceedings of the Conference on Design, Automation and Test in Europe*, pp. 33–38, EDA Consortium, 2012.

[27] J. Lowe-Power, M. D. Hill, and D. A. Wood, “When to use 3d die-stacked memory for bandwidth-constrained big data workloads,” 2016.

[28] N. Miura, Y. Koizumi, Y. Take, H. Matsutani, T. Kuroda, H. Amano, R. Sakamoto, M. Namiki, K. Usami, M. Kondo, *et al.*, “A scalable 3d heterogeneous multicore with an inductive thruchip interface,” *Micro, IEEE*, vol. 33, no. 6, pp. 6–15, 2013.

[29] D. H. Woo, H.-H. S. Lee, J. B. Fryman, A. D. Knies, and M. Eng, “Pod: A 3d-integrated broad-purpose acceleration layer,” *IEEE micro*, no. 4, pp. 28–40, 2008.