1

3D Design Styles

Paul D. Franzon

North Carolina State University, 2410 Campus Shore Dr., Raleigh, NC 27606, USA

1.1 Introduction

3D-IC and interposer technologies have demonstrated their capability to reduce system size and weight, improve performance, reduce power consumption, and even improve cost as compared with baseline 2D integration approaches. Though not a replacement for Moore's law, 3D technologies can provide significant improvements in performance per unit of power and performance per unit of cost. The main purpose of this chapter is to provide an overview of product and design scenarios that uniquely leverage 3D-IC technologies in 3D specific ways.

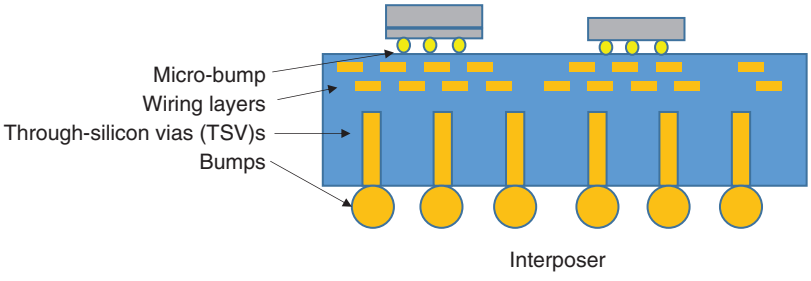
The structure of this chapter is as follows. First, we do a quick review of the 3D technology set. Then we review the main design drivers for using 3D technologies: (i) miniaturization, (ii) provisioning power effective memory bandwidth, (iii) improving performance/power of logic, and (iv) heterogeneous integration for cost reduction to enable unique system capabilities.

1.2 3D-IC Technology Set

There are several technology components that can be mixed and matched in the 3D technology set. The purpose of this section is not to review these in detail, but to introduce them. Other books in this series focus on the technology.

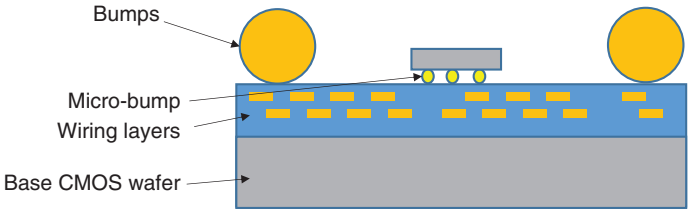
The main 3D-IC technologies of interest are illustrated together in Figures 1.1–1.4. Interposers (Figure 1.1) are so called because they are placed or ***posed*** in between the chip and the main laminate package. Using interposers is often referred to as ***2.5D integration***. A common way to make interposers is to use silicon processing technologies to create a microscale circuit board. Through-silicon vias (TSVs) are fabricated in a silicon wafer, and multiple metal layers are then fabricated on top. These metal layers can be fabricated with thin film processing, typically giving 3–6 metal layers up to a few micrometers

Handbook of 3D Integration: Design, Test, and Thermal Management, First Edition.
Edited by Paul D. Franzon, Erik Jan Marinissen, and Muhannad S. Bakir.



Typical dimensions: Interposer thickness : 100 μm
Micro-bump pitch : 25 μm +
Wiring layers: 4–6 layers; submicron to multi-μm thickness
Typically 1–10 μm width and space
Bump pitch: 150 μm +

Figure 1.1 Interposer or ***2.5D integration***.



Typical dimensions: Interposer thickness : 100 μm
Micro-bump pitch : 25 μm +
Wiring layers: 1–3 layers; submicron to multi-μm thickness
Typically 1–10 μm width and space

Figure 1.2 Redistribution layer.

thick, or can be fabricated with integrated circuit ***back-end-of-line*** (BEOL[1]) techniques, giving 4–6 thinner but planarized metal layers. The latter approach usually reuses a legacy BEOL process, e.g. from the 65 nm technology node. Micron-scale line width and space can be readily achieved. The interposer is usually thinned to 100 μm. Thus 100 μm long TSVs are used to connect the metal layers to the package underneath. The pitch of the TSVs is also typically around 100 μm. Chips are flipped bumped to the top of the interposer, and the interposer connects them to each other and the outside world. The bump pitch between the chip and the interposer can be relatively tight, down to 25 μm, but the interposer package chip must be at conventional scales, typically in the 150+ μm range. The chips on top of the interposer can be single die or multi-chip stacks themselves.

1 The ***front end of line*** refers to transistor fab, which is usually done before metal interconnect fab, which is thus called BEOL.

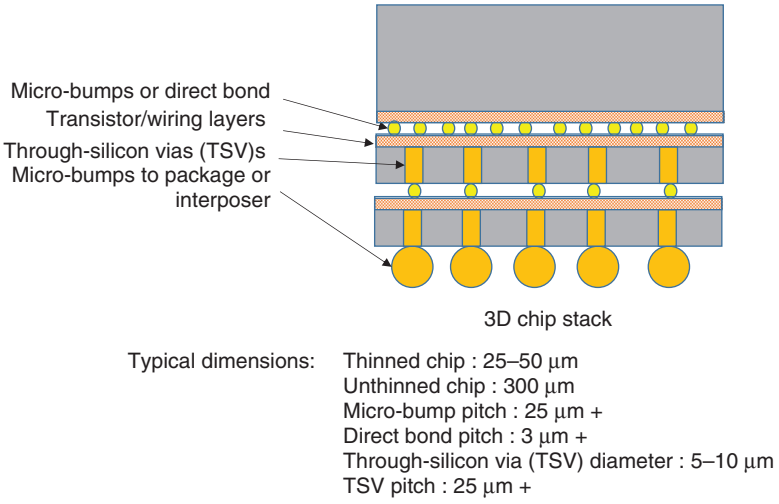


Figure 1.3 3D-IC chip stacking technology set.

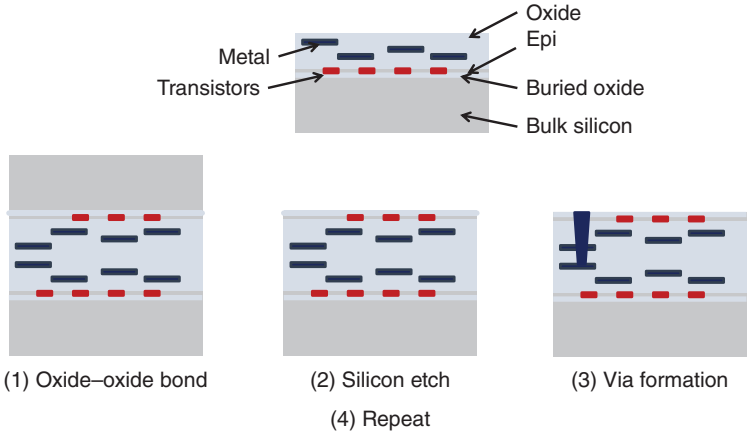


Figure 1.4 3D integration in silicon on insulator technology.

Another interposer technology under active investigation is to use glass as a substrate rather than silicon. Then potentially large panel processing techniques, such as those used to make television screens, can be used, and price reduction achieved.

A related technology is to create interconnect on top of an already finished CMOS wafer and use that to connect to chips and inputs/outputs. This is illustrated in Figure 1.2. Additional thin film wiring layers are processed on top of a completed CMOS wafer to connect the chips in that wafer to chips that are placed on top, together with the chip stack IO. It is referred to as a ***redistribution layer*** (RDL) as the CMOS wafer IOs are ***redistributed***. Not as many wiring layers are possible as with interposers. One application of RDL technology is to

make a chip stack of a larger die, e.g. a memory stack, to a smaller die, e.g. a processor.

An exemplar 3D chip stack, or 3D-IC, is shown in Figure 1.3. This illustrates a three-chip stack, two of which incorporate TSVs. The top two chips illustrated in this stack are mated ***face to face*** (F2F). That is, the transistor and wiring layers are directly mated. This mating can be done with solder bumps or with a thermocompression or direct bonding technology. The latter technologies have been demonstrated down to 3 μm pitch and have potential for 1 μm pitch. An example of a copper direct bond interconnect technology can be found in [1]. This permits a very high interconnect density between the two chips. These F2F connections can be leveraged in multiple ways to enable higher-performance and lower power logic stacks.

TSVs can be used to connect the face of one chip, through the back of another to the transistor/wiring layer, or to connect chip stack IO through a chip backside. Thus they can connect a chip face to back (F2B, shown in Figure 1.3 between the bottom two chips) or even back to back (B2B, not shown). TSVs are made using techniques that create very vertical vias through the bulk silicon substrate. They have a lower density than an F2F connection but are important for creating chip stacks. For example, the TSVs shown in Figure 1.3 connect the primary IO and power grounds at the bottom up through the chip stack. The layers with TSVs have to be thinned. The chip stack often includes one unthinned layer for mechanical stability (though this is not a requirement).

A fourth option that is only possible in a silicon on insulator (SOI) technology is shown in Figure 1.4. In this approach, fabricated wafers are joined F2F using an oxide–oxide bond. Since the transistors are built on top of an oxide layer, a silicon-selective back etch can be used to remove the silicon part of the SOI substrate while not affecting the transistors and interconnect layers. Simple through-oxide vias can then be used to create vertical connections between what were previously separate chips. An example of this process can be found in [2]. If the first two chips in the stack are fabricated without interconnect, then one gets two directly connectable transistor layers in what would be considered a monolithic 3D technology.

1.3 Why 3D

Table 1.1 presents a summary of potential drivers for 3D integration. The desire for thinner smartphone cameras has resulted in the first mainstream high volume use of 3D technologies. However, such miniaturization can also be used for other image sensors and for ***smart dust*** sensors. Provisioning large amounts of power effective memory bandwidth appears to be the next volume application of 3D technologies. In contrast, logic stacking or logic-on-memory stacking has had strong but unrealized potential for improving system performance/power. Finally, 3D offers unique opportunities for heterogeneous integration of different technologies.

Each of these potential design drivers will be explored in detail in the next four sections.

Table 1.1 Issues that are potential drivers for 3D integration.

Driving issue	Case for 3D	Caveats
Miniaturization	Stacked memories ***Smart dust*** sensors Image sensors	For many ***smart dust*** cases, stacking and wire bonding is sufficient
Memory bandwidth	3D memory can dramatically improve memory bandwidth and power consumption	Stacking memory on logic has thermal issues
Interconnect delay, bandwidth, and power	Length of critical paths can be substantially reduced through 3D integration, or benefit can be made of massive vertical bandwidth	Not all cases have a substantial advantage
	In certain cases, a 3D architecture might have substantially lower power or performance/power over a 2D architecture	Thermal issues can be solved with careful floor planning and/or liquid cooling
Mixed technology (heterogeneous) integration	Tightly integrated mixed technology (e.g. III–V on silicon or analog on or next to digital) can bring many system advantages in performance and cost	

1.4 Miniaturization

Obviously, 3D stacking technologies using thinned silicon have direct potential to reduce system volume. An early application of TSVs was for providing the IO connections cell phone camera frontside imaging sensor (http://image-sensors-world.blogspot.com/2008/09/toshiba-tsv-reverse-engineered.html; http://www.semicontaiwan.org/en/sites/semicontaiwan.org/files/docs/4._mkt__jerome__yole.pdf). The goal was not to leverage 3D chip stacks – these were single die – but to reduce the overall sensor height, at least when compared with conventional packaging approaches.

More recently Sony has leveraged a copper–copper direct bonding technology to create an image sensor as a two-chip stack [3], (http://www.sony.net/SonyInfo/News/Press/201201/12-009E/index.html, http://www.3dic.org/3D_stacked_image_sensor). One chip is a backside-illuminated pixel array that does not include interconnect layers or even complete CMOS transistors. The second chip is a complete CMOS chip on which is built the analog-to-digital converters (ADCs) and interconnect for all other functionality required of an image sensor. This approach leverages the high density capability of a direct bonding technology since pixel-scale vertical interconnect is required. Since only one of the two chips goes through a full CMOS fab, there is potential for cost reduction in comparison with a 2D sensor of the same total area having to go through a full CMOS fab. In contrast, here, the sensor-only chip should be substantially smaller per square millimeter. In addition, the volume is reduced substantially through a smaller footprint. This image sensor is probably the first high volume application incorporating a full 3D-IC chip stack.

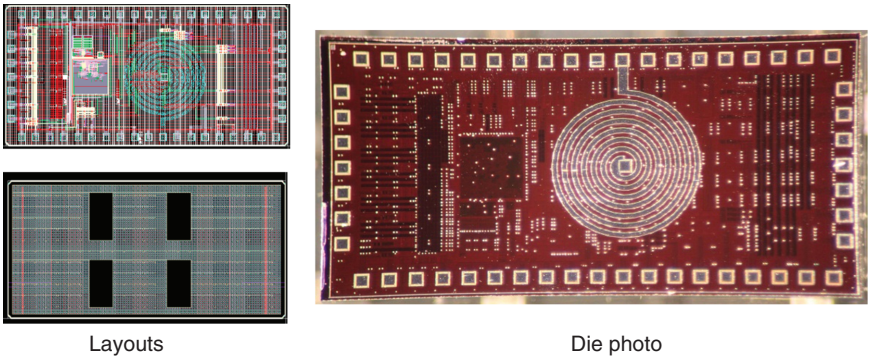


Figure 1.5 Food processing sensor.

In the research domain a number of 3D image sensors have been demonstrated – too many to summarize here. Separating the image and processing layers leads to the potential for improved performance in terms of sensitivity (larger pixels), frame rate (e.g. faster or more ADCs in the CMOS layer), and integrated advanced processing (e.g. edge detection for robotics).

Another interesting use of 3D technologies has been to build non-visible light sensors, sometimes using a non-silicon technology for the sensing layer. Examples include IR imagers, X-ray imagers [4], and other images for high energy physics investigations (http://meroli.web.cern.ch/meroli/DesignMonolitic DetectorIC.html).

3D-IC has been explored for non-image sensors. 3D chip stacking can be used to make such sensors with low integrated volume. Though fabricated using wire bonding, Chen et al. demonstrated an integrated power harvesting data collecting sensor with the photovoltaic power harvesting chip mounted on top of the logic and RF chips [5]. This maximizes the photovoltaic power harvesting area while minimizing the volume. TSVs and bonding technologies would permit further volume reduction. Lentiro [6] describes a two-chip stack aimed at simulating a particle of meat for the purposes of calibrating a new food processing system. One chip is an RFID power harvester and communication chip, and the second is the temperature data logger (Figure 1.5). It is a two-chip stack with F2F connections and TSV-enabled IO. It is integrated with a small battery for data collection purposes only as the RFID cannot be employed in the actual processing pipes. The two-chip stack permits smaller imitation food particles than otherwise would be the case.

1.5 Memory Bandwidth

Memory is positioned as the next large volume application of 3D-IC technologies. To date DRAM has relied on one-signal-per-pin signaling using low cost, low pin count, and single chip plastic packaging. As a result, DRAM has continued to lag logic in terms of bandwidth potential and power efficiency. Furthermore, the IO speed of one-signal-per-pin signaling schemes is unlikely to scale a lot beyond

what can be achieved today in double data rate DDR4 (up to 3.2Gbps per pin) and graphics GDDR6 (8Gbps). Beyond these data rates, two-pins-per-signal differential signaling is needed. Furthermore, the IO power consumption, measured as mW/Gbps, is relatively high, even for the LPDDR standards (intended for mobile applications).

Thus there are multiple 3D-IC enabled memory solutions available in the market, all of which offer improved data bandwidth and power efficiency over conventional memories. These include the hybrid memory cube (HMC), high bandwidth memory (HBM), WideIO, and Tezzaron disintegrated RAM (DiRAM). These are summarized in Figure 1.6 and Table 1.2. Note that in the table, B (byte) and b (bits) are both used. Also note that 1 mW/Gbps is equivalent to 1 pJ/bit.

The HMC is a joint Intel–Micron standard that centers on a 3D stacked part including a logic layer and multiple DRAM layers organized as independent vertical slices. This 3D chip stack is then provided as a packaged part, so the customer does not have to deal with any 3D-IC or 2.5D packaging issues. At the time of writing this chapter, Micron offers 2GB and 4GB parts with a maximum memory

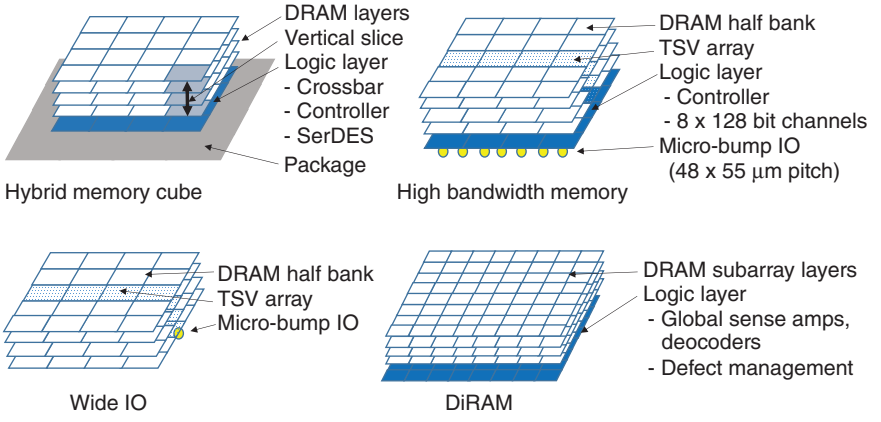


Figure 1.6 3D DRAMs.

Table 1.2 Comparison of 2D and 3D memories.

Technology	Capacity	BW (GBps)	Power (W)	Efficiency (mW/GBps)	IO efficiency (mW/Gbps)	DQ count
DDR4-2667	4GB	21.34	6.6	309	6.5–39	32
LPDDR4	4GB	Up to 42	5.46	130	2.3	32
HMC	4Gb	128GBps	11.08	86.5	10.8	8 Serdes lanes
HBM	16Gb	256GBps		48		1024
WideIO	8–32Gb	51.2GBps[a)]		42[b)]		256
DiRAM4	64Gb	8Tbps				4096

a) WideIO2. WideIO1 was half of this.
b) WideIO1. WideIO2 should be lower.

bandwidth of up to 160GBps (the 128GBps part is used in Table 1.2). The data IO is organized as an eight high-speed serial channels or lanes. HMC is mainly aimed at computing applications.

The HBM is a JEDEC (i.e., industry-wide) standard that is intended for integration via an RDL, 3D-IC stacking, or interposer to logic. It has a lot of data IO (DQ) pins, configured as 8×128-bit wide interfaces with each pin running at up to 2Gbps. Connecting this large number of pins (placed on a 48 μm × 55 μm grid) is the reason why it has to be integrated via an RDL, interposer, or direct 3D stacking. It is fabricated as a stack of multibank memory die, connected to a logic die through a TSV array, the TSV arrays running through the chip centers. Each chip is F2B mounted to the chip beneath it. The eight channels are operated independently. Details for a first-generation HBM (operating at 3.8 pJ/bit power level at 128GBps) can be found in [7]. The use of HBM in graphics module products has been announced by Nvidia and AMD.

WideIO is also a JEDEC-supported standard, aimed largely at low power mobile processors. While intended to be mounted on top of the logic die in a true 3D stack, side-by-side integration on an interposer is also possible. WideIO is a DRAM-only stack – there is no logic layer. Instead the DRAM stack is exposed through a TSV-based interface, and the memory controller is designed separately on the CPU/logic die that is customer designed. An example is the ST/CEA WideIO1 test vehicle [8]. It also supports multiple independent memory channels, operating at up 800Mbs/pin. For example, the Samsung WideIO2 product supports four channels, each 64-bit wide operating at 800Mbs/pin. The standard is currently in its second generation (WideIO2), and a third is being planned. WideIO has yet to enjoy commercial success. To date the thermal challenges of mounting a DRAM on an already hot mobile processor logic die have been insurmountable, especially as it is desired to operate the DRAM at a lower temperature than logic (85 °C for DRAM vs. 105 °C for logic) to control leakage and refresh time. One potential solution is side-by-side integration on an interposer. WideIO has potential for employment in mobile processor-based server solutions as the thermal issues are easier to manage.

The Tezzaron DiRAM4 is a proprietary memory. It has 4096 data IO organized across 64 ports. It is intended only for 3D and interposer integration. It has a unique organization in that the logic layer is not only used for controller and IO functions but also houses the global sense amplifiers and addresses decoders that in other 3D memories are on the DRAM layers. This permits faster operation for these circuits. The DiRAM4 has potential for a very high bandwidth (up to 8Tbps) and fast random cycles (15 ns) [9]. DiRAM4 is being integrated into a number of specialized applications that benefit from its high bandwidth.

1.6 3D Logic

It has always been assumed that the next major employment of 3D-IC, after memories, would be 3D logic, i.e., logic stacks. The argument is simple. On-chip wiring dominates the area, performance, and power consumption of many logic chips. 3D logic stacking would shorten many of those wires, leading to power reduction,

performance enhancement, and area reduction. While these improvements can be achieved, the increased cost and heat flux issues have been challenging. This section describes experiments that have demonstrated these advantages as well as pointing to some solutions to the heat flux question. The main metric of interest in evaluating logic-on-logic stacks is performance per unit of power.

The first two experiments to be described are ones in which a 2D logic chip is partitioned into two 3D stacked chips: first at the module level and second at the circuit level. Before describing those experiments, some discussion on power efficiency in computation is warranted.

1.6.1 Power-Efficient Computing and Logic

Table 1.3 lists the energy per operation for a range of operations, where appropriate, scaled to 0.6 V operation at the 7 nm node (for logic). (Note that 1 pJ/op = 1 mW/Gbps.) This table was constructed by taking simulation or published power results and scaling them using the ***conservative*** scaling factors published by Intel authors in [10, 11]. The single instruction multiple data (SIMD) core was the one designed at NCSU in 65 nm CMOS and optimized for low power operation. Some more detail on this core can be found in [12]. These conservative factors capture the slowdown in performance and power scaling expected after the 22 nm node.

For DRAM, these numbers are for the DRAM core only (not its IO or other overhead) at the 16 nm node, which is the presumed last DRAM node. These

Table 1.3 Energy per operation for a range of operations generally scaled to 0.6 V at the 7 nm node.

Computation	Energy/32-bit word
32-bit multiply–add (SP)	6.02 pJ/op
FPU	1.4 pJ/op
SIMD vector processor (16 lane)	4.6 pJ/FLOP
Data storage	
16 × 64-bit RF	0.5 pJ/word
128KB SRAM	0.9 pJ/word
L1 Dcache (16KB)	62 pJ/16 B
L2 Dcache (2MB)	24 pJ/16 B
16 nm DRAM core	140 pJ/word
Communications	
On-chip	0.23 pJ/word/mm
PCB	54 pJ/word
Interposer	17 pJ/word
TSV	1.1 pJ/word

Source: Adapted from Borkar 2010 [10] and Esmaeilzadeh et al. 2011 [11].

figures were taken from [13] and are for DRAM structures likely for commodity products, with high DRAM cell fill factors. The fill factor is the percentage of total area given over to DRAM cells. Energy/access for a DRAM can be improved by using smaller banks, with lower fill factor. Early studies on this aspect indicate that a potential improvement of about 4× is possible through this approach.

For interconnect, some of these figures are taken from the modeling and simulation study presented in [10] and again extrapolated to the 7 nm node. The interposer power was based on an extrapolation of the results presented in [14], with an assumption that 2/3 of the power is for driving the transmission line and so does not scale.

What is interesting to observe is that for 2D technologies, calculation (computation) is energetically much cheaper than data storage or communications, which creates serious constraints for power-efficient computing. Power efficiency is best achieved by minimizing data motion and by minimizing memory references, especially to DRAM or via the cache hierarchy. In contrast, data motion using 3D technologies takes much less energy than when using 2D technologies. With 3D stacking, vertical data communications using TSVs or a direct bond interface consumes less power than computation. Thus now it makes sense to move data if an overall advantage can be gained. An example of this is given below as a heterogeneous computer.

1.6.2 Modular Partitioning: FFT Processor

This system consists of three stacked tiers with eight processing elements, one controller, thirty-two SRAMs, and eight ROMs [15]. The system performs 32 memory accesses per cycle (16 reads and 16 writes), completing a 1024-point fast fourier transform (FFT) in 653 cycles utilizing five pipeline stages. The floor plan is designed so that all communications are vertical – there is no horizontal communications between PEs. The chip was implemented in the Lincoln Labs SOI 3D process described earlier. The die photo (Figure 1.7) clearly shows the TSV arrays, one of which is specifically pointed out and the locations of which were dictated to be at the SRAM bank interfaces. Figure 1.7 also shows the stacked chip floor plans. This clearly shows the modular nature of the partitioning in that each ***processing element*** (PE) module is preserved as an integrated 2D design. The logic to the interior of the modules is not broken into 3D. Each PE communicates vertically with the memories stacked with it. By breaking a large memory into

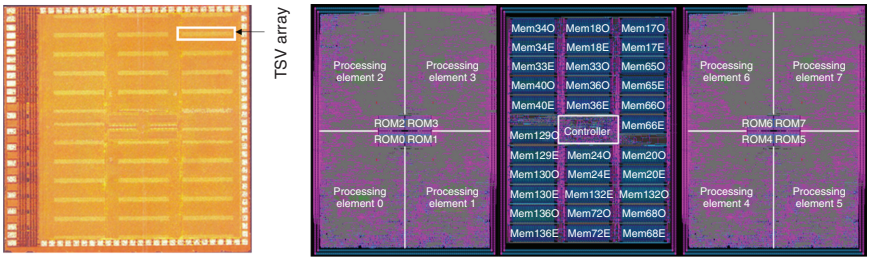


Figure 1.7 3D FFT engine die photo and floor plans of the three chips in the stack.

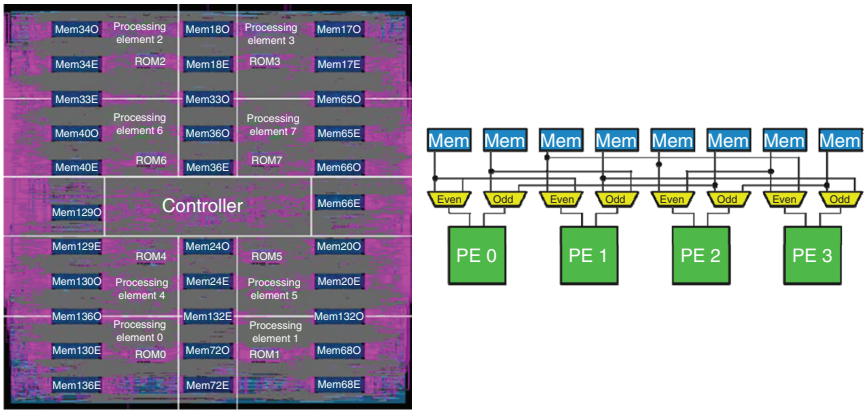


Figure 1.8 2D floor plan and architecture of FFT engine.

32 smaller memories, memory power was reduced by 58%. (A similar trade-off exists for DRAMs.)

This two-chip stack was redesigned as a 2D chip. The floor plan of this chip, together with a module connectivity diagram, is shown in Figure 1.8. A comparison of this with the 3D chip is summarized in Table 1.4. The total area of the 3D chip is 25% less than that of the 2D equivalent. This difference arises due to the need for added area in the 2D chip to route all the additional wiring that was needed. The total length of routed wire went up 57% in the 2D chip. Admittedly, there are around 1800 connections between the PEs and the memories – an amount very affordable in the 3D version but expensive in the 2D version. Due to the reduced wiring load, the 3D version could operate 24.6% faster and with 4.4% less power. Even the logic power is lower in the 3D version due to the reduced capacitive load at the logic outputs. The 3D version of this architecture shows significant advantages due to improvement in routability between the modules.

1.6.3 Circuit Partitioning

A modified CAD flow was applied to three different designs – a radar PE, an AES encryption engine, and a MIMO multipath radio processing engine. The CAD flow was designed to partition a 2D chip into two stacked chips. The partitioning is done at the circuit level – with connected logic gates possibly

Table 1.4 Comparison of 2D and 3D FFT engines.

Metric	2D	3D	Change (%)
Total area (mm^2)	31.36	23.4	−25
Total wire length (m)	19.1	8.23	−57
Maximum speed (MHz)	63.7	79.4	+24.6
Power at 63.7 MHz (mW)	340	325	−4.4
FFT logic energy (μJ)	3.55	3.36	−5.2

Table 1.5 Improvements in 3D design over 2D using logic cell partitioning.

	Total wire length (% change)	Fmax (% change)	Total power (% change)	Power (MHz)
Radar PE	−21.0%	+22.6%	−12.9%	−38%
AES	−8%	+15.3%	−2.6%	−18%
MIMO	+216%	+17.1%	−5.1%	−23%

being on different chips and connected vertically. This partitioning approach leverages the high density and bandwidth of the copper–copper direct bonding interface when two dies are stacked F2F with each other. The minimum bond pitch was 6.3 μm, and the chips were made in a standard 130 nm bulk CMOS process. TSVs were used for backside IO. All flip-flops are kept in one tier so that 3D clock distribution was not required. The radar PE was implemented in the Tezzaron bulk CMOS 3D process [16] (Figure 1.8). The results are summarized in Table 1.5. On average, performance per unit of power was increased by 22% due to the decreases in wire length achieved through this partitioning approach. The radar processor had an improvement in performance per unit of power of 21%. The other designs achieved 18% and 35%.

1.6.4 3D Heterogeneous Processor

This design is very 3D specific. It takes advantage of the vertical dimension and their lower power characteristics in a unique way. A stack of two different CPUs is integrated vertically using a vertical ***thread transfer*** bus that permits fast compute load migration from the high-performance CPU to and from the low power CPU when an energy advantage is found [17]. In this design, the ***high-performance CPU*** can issue two instructions per cycle, while the ***low power CPU*** is a single-issue CPU. The transfer is managed using a low-latency, self-testing multi-synchronous bus [18]. The bus can transfer the state of the CPU in one clock cycle by using a wide interface and exploiting a high density copper–copper direct bond process. The caches are switched at the same time, removing the need for a cold cache restart.

Simulation with Specmark workloads shows a 25% improvement in the power/performance ratio compared with executing the sample workload solely in the high-performance processor. In contrast, if the workload was executed solely in the single-issue (***low power***) CPU, there was a 28% total energy savings, compared with keeping the workload in the high-performance CPU, but at the expense of a 39% reduction in performance. If the workload was allowed to switch every 10 000 cycles, there was a 27% total energy savings but at the expense of only a 7% reduction in performance. That is, a 25% improvement in power per unit of performance is achieved.

This processor stack was taped out in a 3D 130 nm process in fall 2015. A copper–copper direct bond interface, with an 8 μm pitch, is used to build the required vertical connectivity. Key to this design is how the various bus elements are built into the logic tiers so that it can be further stacked with itself or other elements, such as accelerators.

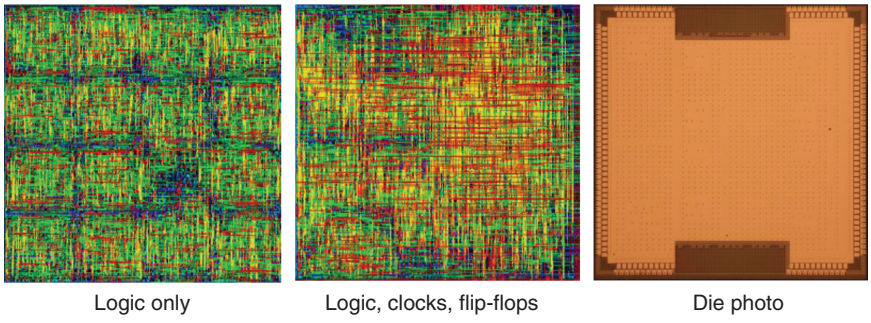


Figure 1.9 Layouts and die photo of the 3D radar processing element.

Another feature of this processor is that it will use the fast multi-port Tezzaron 3D DiRAM4 memory as a combined L2/L3 cache. This DRAM can perform fast RAS–RAS cycles while providing more than 1Gb of total capacity. Compared with an SRAM-based cache hierarchy, it provides a 90% performance improvement while reducing power consumed in these caches by almost 4×.

An illustration of the overall floor plan is shown in Figure 1.9, showing the two-processor stack integrated with each other and their caches. Between the two processors is a 2254-bit wide (1120 data in each direction and 14 control signals) thread transfer bus. This is a very short bus that runs through the copper direct bond pads between the two chips. Each processor is also connected to both caches through a switch. The buses in the switch are again very short. One bus runs horizontally between each cache and the CPU in the same chip; another runs vertically to the different CPUs.

If this architecture was not built as a 3D chip, then at least two of the buses shown here would be long and power hungry and introduce additional delay. Figure 1.10 shows a layout of the two chips in the stack.

1.6.5 Thermal Issues

Based on a simple calculation, the thermal ramifications of 3D-IC are not very positive. In the examples above, the maximum power reduction due to 3D was 13%. Since the footprint area is halved, this means the heat flux is increased 1.7×, which would lead to a significant temperature rise!

However, this simple calculation ignores the fact that temperature rise is very dependent on details of the floor plan. For example, by staggering the high power

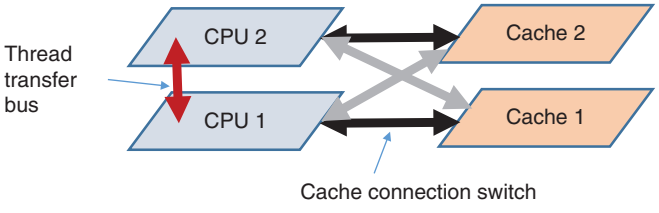


Figure 1.10 3D heterogeneous processor floor plan as a two-chip stack.

density blocks so that they do not overlap, Saeidi et al. [19] showed that a two-chip stack can achieve a junction temperature of only 8 °C more than the 2D equivalent and the same junction temperature if one can achieve a 5% reduction in dynamic power in the 3D version! They then investigated this concept in the framework of a mobile processor design and found through a combination of clever floor planning and/or partitioning, together with the 5–16% power reduction that 3D gives, and the worst hotspot could be less in the 3D design than in the 2D. They achieved this for the CPU through careful modular floor planning and preventing high power density modules from overlapping. They achieved this in the GPU by leveraging the power reduction potential of circuit partitioning across a F2F connection. For servers, another potential solution is to use liquid cooling. Thus with some sophistication, thermal issues do not have to be a barrier to realizing the advantages of 3D design.

1.7 Heterogeneous Integration

Another unique aspect to 3D-IC is that these technologies enable different technologies to be intimately mated with high connectivity. An example that has already been given is that of a CMOS image processor fabricated as a two-chip stack. The two chips are different: one chip just consists of imaging pixels, while the second is a complete CMOS chip. This leads to lower cost than the alternative of two full CMOS chips. DRAM on top of logic also serves as an example of heterogeneous integration.

Three examples of heterogeneous integration will be given in the rest of this section: (i) splitting logic for cost reduction, (ii) mixing different CMOS nodes within one module, and (iii) mixing III–V and silicon technologies within one 3D-IC chip stack.

The first example is only heterogeneous in the sense that it is mixing interposer and CMOS technologies. To a first approximation, the cost of a large CMOS chip goes up with the square of the area. This is because the probability of a defect occurring on the chip and thus ***killing*** the chip goes up with the chip area while the cost of making the chip in the first place also goes up with the area. Thus it is worth considering partitioning a large chip into a set of smaller ones, if the cost of integration and the additional test are less than the savings accrued to increase CMOS yield. Xilinx investigated this concept for large FPGAs and is now selling FPGA modules containing two to four CMOS FPGA chips tightly integrated on an interposer. Details are not available, but they claim an overall cost savings [20].

The second example is that of mixing technology nodes. In general, Moore's law tells us that a digital logic gate costs less to make a more advanced technology due to the reduced area for that gate in that node. However, in contrast, many analog and analog-like functions like ADCs and high-speed serial–deserializer (SerDes) IOs do not benefit in such a fashion. The reason is that the analog behavior of a transistor has higher variation for smaller transistors than for larger ones. Thus, for many analog functions that rely on well-matched behaviors of different transistors in the circuit, no benefit is accrued from building smaller transistors. More simply put, analog circuit blocks do not shrink in dimensions with the use of

more advanced technologies. Thus the cost of these functions in a more advanced process node can actually be higher than in the old node, since the old node costs less to make per unit of area.

While Wu [20] also explores this concept generically, Erdmann et al. [21] have explored this concretely for a mixed ADC/FPGA design. Their design consisted of two 28 nm FPGA logic dies, integrated with two 65 nm ADC array dies on an interposer. Thus two sets of cost benefits are accrued: first the yield-related savings from splitting the logic die in two and second the fabrication cost savings of keeping the ADCs in an older technology.

The third example that will be given is that of mixing III–V and silicon technologies. This is best exemplified by the DARPA diverse accessible heterogeneous integration (DAHI) program in which GaN and InP chips are integrated on top of CMOS chips through micro-bumps and other technologies [22, 23]. More specifically, CMOS can be used for most of the transistors in a circuit, while GaN high electron mobility transistors (HEMTs) can be used for their high power capability and InP HBTs can be used for their very high speed. An example of the latter is an ADC. In an ADC, only a few transistors generally determine the sampling rate. Thus with the DAHI technology, these few transistors can be built in a high-speed but expensive and low-yielding InP chiplet, while the rest of the ADC is built in cheaper and more robust CMOS.

Northrop Grumman is the main fab in the DAHI program [23]. They use gold micro-bumps and through-silicon carbide vias to integrate GaN HEMTs on top of CMOS in a face-to-back process and gold micro-bumps to integrate InP HBT chiplets to CMOS in an F2F process (Figure 1.11). A face-to-back process is used for the GaN parts in order to allow for some heat spreading in the GaN part, as the main conduction path is through the CMOS chip (Figure 1.12).

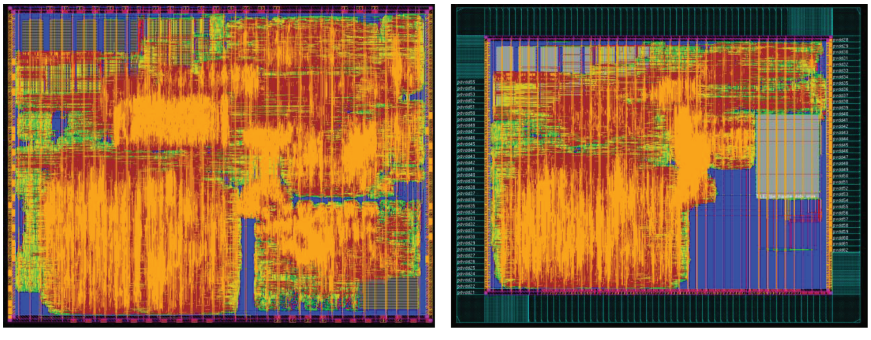


Figure 1.11 Layouts of the two chips in the heterogeneous processor stack.

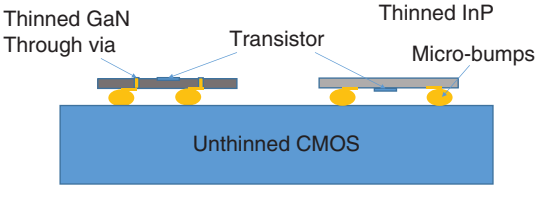


Figure 1.12 DAHI Northrop Grumman heterogeneous integration process.

1.8 Conclusions

3D-IC and 2.5D (interposer) technologies have demonstrated their utility in enabling cost scaling and scaling in performance and power consumption beyond that provided by Moore's law alone. They permit miniaturization of chip assemblies and have had widespread employment in CMOS image sensors for mobile products. Their next impact will be in the form of DRAM stacks, enabling high bandwidth and low power memory integration. By exploiting the high density F2F connection, density of copper direct bonding and logic-on-logic stacks can be designed that improve performance/power by 25% or more. Careful floor planning and placement can be used to solve the thermal challenges that arise. Finally heterogeneous integration, that is, mixing different technologies on an interposer or 3D stack, leads to optimized performance at optimized cost.

References

1 Enquist, P., Fountain, G., Petteway, C. et al. (2009). Low cost of ownership scalable copper direct bond interconnect 3D IC technology for three dimensional integrated circuit applications. In: *IEEE International Conference on 3D System Integration, 2009. 3DIC 2009*, 1–6. San Francisco, CA.

2 Burns, J.A., Aull, B.F., Chen, C.K. et al. (2006). A wafer-scale 3-D circuit integration technology. *IEEE Transactions on Electron Devices* 52 (10): 2507–2516.

3 Enquist, P. (2014). 3D integration applications for low temperature direct bond technology. In: *2014 4th IEEE International Workshop on Low Temperature Bonding for 3D Integration (LTB-3D)*, 8. Tokyo.

4 Deptuch, G.W., Carini, G., Enquist, P. et al. (2016). Fully 3-D integrated pixel detectors for X-rays. *IEEE Transactions on Electron Devices* 63 (1): 205–214.

5 Chen, G., Fojtik, M., Kim, D. et al. (2010). Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells. In: *2010 IEEE International Solid-State Circuits Conference – (ISSCC)*, 288–289. San Francisco, CA.

6 Lentiro, A. (2013). Low-density, ultralow-power and smart radio frequency telemetry sensor. PhD dissertation. NCSU.

7 Lee, D.U., Kim, K.W., Kim, K.W. et al. (2015). A 1.2 V 8 Gb 8-channel 128 GB/s high-bandwidth memory (HBM) stacked DRAM with effective I/O test circuits. *IEEE Journal of Solid-State Circuits* 50 (1): 191–203.

8 Dutoit, D., Bernard, C., Cheramy, S. et al. (2013). A 0.9 pJ/bit, 12.8 GByte/s WideIO memory interface in a 3D-IC NoC-based MPSoC. In: *2013 Symposium on VLSI Circuits*, C22–C23. Kyoto.

9 RTI 3D ASIP (2013). Evolving 2.5D and 3D integration. www.tezzaron.com (accessed 24 August 2018).

10 Borkar, S. (2010). The exascale challenge. In: *2010 International Symposium on VLSI Design Automation and Test (VLSI-DAT)*, 2–3.

11 Esmaeilzadeh, H., Blem, E., Amant, R.S. et al. (2011). Dark silicon and the end of multicore scaling. In: *2011 38th Annual International Symposium on Computer Architecture (ISCA)*, 365–376.

12 Franzon, P.D., Rotenberg, E., Tuck, J. et al. (2014). 3D-enabled customizable embedded computer (3DECC). In: *Proceedings of 3DIC 2014*.

13 Vogelsang, T. (2010). Understanding the energy consumption of dynamic random access memories. In: *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*, 363–374. Washington, DC.

14 Karim, A., Franzon, P.D., and Kumar, A. (2013). Power comparison of 2D, 3D, and 2.5D interconnect solutions and power optimization of interposer interconnect. Proceedings of IEEE ECTC 2013.

15 Davis, W., Oh, E., Sule, A. et al. (2009). Application exploration for 3-D integrated circuis: TCAM, FIFO and FFT case studies. *IEEE Transactions on VLSI* 17 (4): 496–506.

16 Thorolfsson, T., Lipa, S., and Franzon, P.D. (2012). A 10.35 mW/GFLOP stacked SAR DSP unit using fine-grain partitioned 3D integration. Proceedings of CICC 2012.

17 Rotenberg, E., Dwiel, B.H., Forbes, E. et al. (2013). Rationale for a 3D heterogeneous multi-core processor. In: *2013 IEEE 31st International Conference on Computer Design (ICCD)*, 154, 168.

18 Zhang, Z., Noia, B., Chakraparthy, K., and Franzon, P.D. (2013). Face to face bus design with built-in self-test in 3DICs. Proceedings of IEEE 3DIC.

19 Saeidi, M., Samadi, K., Mittal, A., and Mittal, R. (2014). Thermal implications of mobile 3D-ICs. In: *2014 International 3D Systems Integration Conference (3DIC)*, 1–7. Kinsdale.

20 Wu, X. (2015). 3D-IC technologies and 3D FPGA. In: *3D Systems Integration Conference (3DIC), 2015 International*, KN1.1–KN1.4. Sendai.

21 Erdmann, C., Lowney, D., Lynam, A. et al. (2015). A heterogeneous 3D-IC consisting of two 28 nm FPGA Die and 32 reconfigurable high-performance data converters. *IEEE Journal of Solid-State Circuits* 50 (1): 258–269.

22 Raman, S., Dohrman, C.L., and Chang, T.H. (2012). The DARPA diverse accessible heterogeneous integration (DAHI) program: convergence of compound semiconductor devices and silicon-enabled architectures. In: *2012 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT)*, 1–6. Singapore.

23 Green, D.S., Dohrman, C.L., Demmin, J. and Chang, T. (2015). Path to 3D heterogeneous integration. International 3D Systems Integration Conference (3DIC), Sendai, (2015), pp. FS7.1–FS7.3.

24 Gutierrez-Aitken, A., Scott, D., Sato, K. et al. (2014). Diverse accessible heterogeneous integration (DAHI) at Northrop Grumman aerospace systems (NGAS). In: *2014 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, 1–4. La Jolla, CA.