3D Specific Systems: Design and CAD

Paul D. Franzon, W. Rhett Davis, Thor Thorolfsson, Samson Melamed,
Department of Electrical and Computer Engineering,
North Carolina State University,
Raleigh NC 27695, USA
{paulf, wdavis, trthorol, smelame}@ncsu.edu

***Abstract*—3D stacking and integration can provide significant system advantages. Following a brief technology review, this abstract explores application drivers, design and CAD for 3D ICs. The main 3D exploitation explored in detail is that of logic on memory. This application is explored in a specific DSP example, showing a 25% power advantage when implemented in 3D compared with 2D. Finally critical areas that need better solutions are explored. These include cost management, design planning, test management, and thermal management.**

Keywords-3DIC; 3D IC; three dimensional IC; TSV; stacked memory; memory on logic; FFT

I. INTRODUCTION

This paper discusses approaches to re-architecting systems explicitly to exploit high density Through Silicon Via (TSV) processes to create 3DICs. The core concept presented is that by using high density, fine feature size TSVs, a large number of TSVs can be exploited to permit specific reoptimizations that increase bandwidth and reduce power consumption. A particular focus is on (redesigned) memory on top of logic. This article also serves as a tutorial for the design of 3D specific systems.

The article is organized as follows. First 3D technology is briefly reviewed primarily from a designer's perspective. Then some of the 3D specific optimizations that the designer can explore are expanded and explored. The core of this article is the description of a 3D specific design – a radar DSP application. Finally the outstanding issues in 3D specific design are explored.

II. TECHNOLOGY SELECTION

The basic steps for 3DIC fabrication with TSVs are summarized in Figure 1. This shows what is commonly referred to as a "via middle" process wherein the TSVs are fabricated during wafer fab, just before all the interconnect steps are completed. One other alternative is to use a "via last process" where the TSVs are fabricated after wafer fab using thinned wafers, as a packaging process. For more information about fabrication, the reader is referred to [1].

There are several viable alternatives to assembling a 3D sub-system. The major alternatives are summarized in Figure 2. The first major issue is the scale and pitch of the TSVs. If a large number of TSVs are needed, then large "packaging-scale" TSVs will consume a lot of area. The pitch is determined by the join technology. A lower temperature solder type join is likely to require a 40 – 50 □m pitch, while a high temperature copper-copper join allows a tighter pitch. A high temperature step might be an issue for some types of chips, e.g. pre-tested DRAMs. Obviously larger TSVs incur more capacitive load as well.

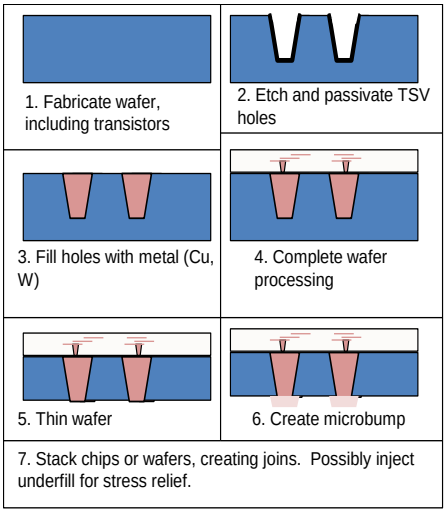

Figure 1. Basic steps in 3DIC fabrication using TSVs.

The next issue is "what is being stacked". If stacking wafers, then it is likely that you are working in a homogeneous technology, designing and fabricating complete wafers intended only for stacking with each other. For example, a memory manufacturer fabricating a memory stack. When stacking wafer on wafer, care must be taken with the cumulative yield loss, assuming there is no method to prevent bad chips being stacked. Thus, the integrated yield will be less than the single tier yield. (An example is given in Table 1.)

Alternatively if stacking die to wafer, then there is more flexibility in the technology choice, technology mix and die sizes. In addition, the die and wafer sites can be tested before integration (creating "Known Good Die" (KGD)), and yield can be maximized.

In many circumstances chips from different vendors will be integrated. It is unlikely that the I/O on one chip (stack) are not physically aligned with the I/O on its soon-to-be mate. In that case some I/O matching is needed. If the only need is re-routing, then a Redistribution Layer (RDL) will often suffice. This consists of one or more layers of thin metal, usually integrated via a spun-on dielectric. If an RDL is not possible (e.g. due to vendor limitations), or power and thermal

management is needed between the chips then an intermediate substrate is needed. I.e. A silicon, ceramic or laminate substrate containing through vias and multiple metal layers. Note if a substrate is used then the interconnectivity through it will be limited by its internal via pitch.

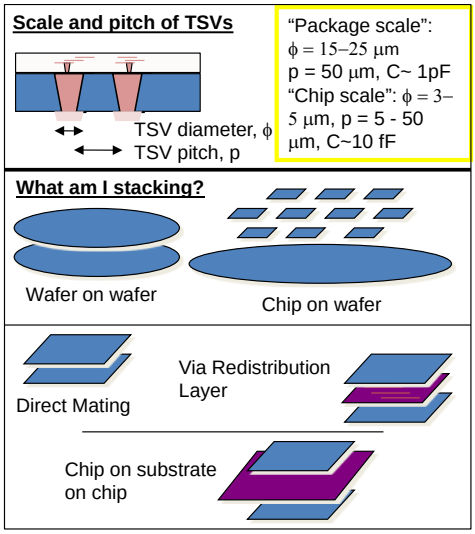

Figure 2. Major choices in 3D assembly from a designer's perspective.

Note that not all 3DICs need TSVs. If only two chips are being stacked, and peripheral bonding of major IO suffices, then face-to-face bonding can be used if the chips are of different sizes.

Table 1. Reduction of integrated yield with stacking using wafer on wafer.

Number of tiers	1	2	3	4
Yield	95%	91%	85%	81%

III. 3D Specific Optimizations

If the chip stack is redesigned to explicitly exploit 3DIC technologies then 3D specific optimizations are possible. Possible optimizations include the following (Table 2):

- Miniaturization, especially in sensors.
- Many studies (e.g. [2, 3]) demonstrate that 3D integration lead to shorter interconnects. Though valuable, the improvement is often incremental and has to be judged against the added cost.
- Use of 3D stacking to increase memory bandwidth. With future multicore CPUs likely to require memory bandwidths of 1 TB/s or above [4], power efficient methods to provisioning this bandwidth are needed. For example, 1024 1Gbps TSV enabled channels are likely to be much more power efficient than 400 20 Gbps channels built in conventional packaging. It is also likely to be more cost-effective.
- Repartitioning the system to decrease power consumption is a unique opportunity for 3D. One potential is to reorganize the memory stack, not just to decrease interconnect power, but also to decrease memory core

Table 2. Potential 3D specific optimizations

Driving Issue	**Case for 3D**	**Caveats**
Miniaturization	Stacked memories. "Smart dust" sensors.	For many cases, stacking and wire-bonding is sufficient
Interconnect Delay	When delay in critical paths can be substantially reduced through 3D integration.	Not all applications will have a substantial advantage
Memory Bandwidth	Logic on memory can dramatically improve memory bandwidth	While memory bandwidth can be improved dramatically, memory size can only be improved linearly
Power Consumption	In certain cases, a 3D architecture might have substantially lower power over a 2D. Memory bandwidth can be provided at a lower bandwidth in 3D.	Limited domain. In many cases, it does not.
Mixed Technology (Heterogeneous) Integration	Mixing an advanced ASIC technology with an older or different analog technology. Incorporating a passives layer.	Though might justify 3D integration, not all examples might justify through-wafer vias.

- Power. This is explored further below.
- Mixed technology (Heterogeneous) integration is a unique opportunity in 3D. Examples include stacking processing with a sensor array; keeping the hard to redesign analog or analog-like (e.g. SerDes) circuits in an older (cheaper) technology node while moving the digital portion to an advanced node; and moving on-chip passives (inductors, capacitors, decoupling) to a low-loss stacked substrate.

IV. Example fo 3D Specific Optimized Design

A 3D optimized synthetic aperture radar (SAR) processor has been designed and is currently in fabrication with Lincoln Labs. The core of this processor is a 1024 point, 32 bit floating point FFT. Power was minimized by using small memories that minimize the energy per access. 3D

interconnect was employed in order to reduce the length of connections to these highly partitioned memories, as shown in Figure 3. The chip layouts are shown below in Figure 4.

The overall system consists of four different components, eight processing elements, one controller, thirty two SRAMs, and eight ROMs and is shown in Figure 3. The system performs 32 memory accesses per cycle (16 reads and 16 writes), completing a 1024-point FFT in 653 cycles utilizing five pipeline stages.

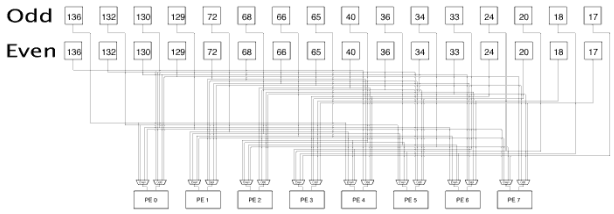

Figure 3. 3D Synthetic Aperture Radar System Block Diagram. Thirty SRAMs (top) are stacked with eight Processing Elements (bottom).

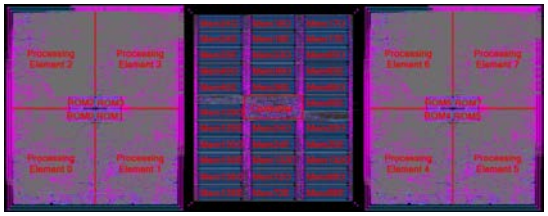

Figure 4. Three-Tier Layout of the 3D Synthetic Aperture Radar Design. Almost all data communications is vertical.

The benefit of splitting the memories into smaller subgroups is that smaller memories are faster and since each memory subgroup can be accessed simultaneously, the system can perform a greater number of reads and writes per cycle. Conversely, a single memory will require less area as only one set of peripheral logic (write driver and sense amp) is required. In this specific implementation the memory was divided into 32 smaller memories (16 even and 16 odd). We use Cacti 4.1[5] to assess the architectural benefit of the partitioned design, by comparing the properties of a single 8 kByte memory to sixteen 512 Byte memories. The results a summarized in Table 3.

Normally in a 2D design, increasing the number of memory to logic interconnect wires from 150 to 2272, leading to an interconnect dominated architecture. Luckily, moving this architecture to 3D ensures that the increase in interconnect power increases does not outweigh the bandwidth and memory access energy consumption gains. This was quantified by redesigning the chip in 2D, with results presented in Table 4. The total silicon area in 2D is 25.3% more than in 3D, due to the extra routing area required for the wires. Even though the PEs ("logic") are essentially 2D in nature (they are not partitioned amongst the tiers), their power and delay are improved in the 3D implementation, due to the reduce wire buffer requirements to the memories. The total power savings of the combined memory and logic structure was 25%.

Table 3. Comparison between the highly partitioned small memories and the unpartitioned big memory.

Metric	Big Memories	Small Memories	%
Wires (#)	150	2272	-1414.7%
Bandwidth (GBps)	13.4	128.4	854.9%
Energy Per Write (pJ)	14.48	6.142	57.6%
Energy Per Read (pJ)	68.205	26.718	60.8%

Table 4. Comparison of the 3D optimized FFT design in both 2D and 3D technologies.

Metric	**2D**	**3D**	**%**
Total Area (mm^2)	31.36	23.40	25.3%
Core Area (mm^2)	29.16	20.16	30.9%
Mean Net Length (μm)	836.0	392.9	53.0%
Total Wire Length (m)	19.107	8.238	56.9%
Max Speed (MHz)	63.7	79.4	24.6%
Critical Path (ns)	15.7	12.6	19.7%
Logic Power (mW)	340.0	324.9	4.4%
FFT Logic Energy (μJ)	3.552	3.366	5.2%

This design was conducted using "chip scale" 5 μm pitch high density TSVs. The 2.6 xx 3 mm design has 17,634 TSVs within it, roughly equally split between signal and power/ground vias. If a coarser "package level" TSV was used, the silicon area impact would be substantially larger (Table 6). The optimizations achieved in this design require high-density, fine-scale vias.

Table 5. Area impact on SAR design of different TSV technologies.

TSV Technology	**Area Impact**
6. μm pitch SOI TSV	0.14 mm^2 (1.7 %)
15 μm pitch (10 μm diameter) intermediate TSV	2 mm^2 (18 %)

More detail about this design can be found in references [3,6].

V. Outstanding Issures in 3D design

Outstanding issues in 3DIC design include cost management, CAD tools, especially for early planning, Test and thermal/power integrity management.

A. *Cost Management*

Since the 3D integration increases the wafer cost by 5% to 15% or more, care must be taken to mitigate the added cost elsewhere in the system. Ideally this cost is compensated not just be a performance and/or power improvement but also by cost reduction elsewhere. Opportunities for cost reduction include using a lower cost functionally optimized technology mix, such as moving passives to an interposer, moving the analog portion to an older technology, or reduced packaging cost, e.g. by reducing pin count, or reduced laminate layer count.

B. *Computer Aideed Design*

With care, 3DICs can be designed and analyzed with the available tools. For example, consider the CAD flow for the radar processor described above. Floorplanning was conducted independently on each tier so that the memories were coarsely aligned with the Processing Elements (PEs). The memory layout was then finalized and the TSV placement determined. A script was written to propagate the TSVs to the other chips in the stack. Normal place and route were then performed on each logic tier. Finally the integrated chip stack was verified using a unified verification stack. By using partitions wherein the local clock distribution is confined to an independent tier and true 3D place and route is needed, detailed 3DIC can be done with modified 2D tools.

The situation is somewhat different for early planning. In particular, floorplanning a design to best use a 3D stacked memory today requires significant hand analysis. Similarly, thermal, power ground and I/O planning presents a difficult chip-package codesign problem that again requires considerable hand analysis. Better tools are needed in this arena.

C. *Thermal Design and Analysis*

The root need in thermal design and analysis is to predict temperature sufficiently accurately so that the resulting unpredictability in signal path delay and clock path delay (and thus skew), and leakage power can both be brought within their budgets. The accuracy needed depends on the budgets allowed. Thermal analysis is complicated in 3D design as the thinned silicon tiers are no longer as good as spreading heat as in an unthinned 2D tier. This leads to local hot spots on the tiers away from the heat sink. For example, the design described above was taken through a full thermal analysis, as discussed in [7] and presented in Figure 5. The relative coolness of the tier (A) closest to the heatsink is apparent, as is the temperature non-uniformity of the tier farthest from the heatsink (tier C). The heatspikes in Tier C are located at the clock buffers, which have a high activity factor.

The main complexity in thermal design is really the complexity in early planning, as described in the previous sub-section.

D. *Test and Design for Test*

When stacking die on wafer, it is highly desirable to know which singulated die are good and which locations are good on the wafer. Testing a die through a TSV array will be difficult especially with “chip-scale” TSVs. It would be very difficult to build a probe card that can reliably and cost-effectively probe thousands of TSVs on a 50 μm pitch. Different solutions are needed, including the use of customized test ports, and internal self-test of the TSV/bond structure before assembly. More information describing possible solutions can be found in [8].

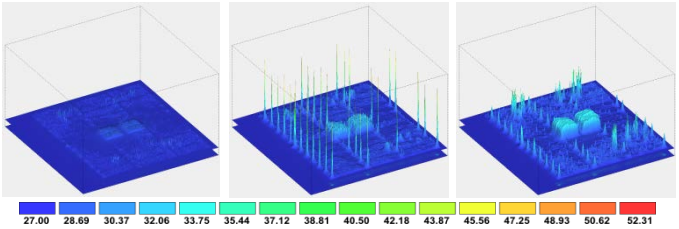

Figure 5. Detailed Thermal Analysis of the three tier FFT design.

VI. CONCLUUSIONS

Designing chips and systems of chips to exploit 3DIC technologies can bring specific advantages in terms of bandwidth and power consumption. A particularly attractive opportunity is redesigning memory for integration with logic functions. An example is given in which power consumption is reduced by 25% compared with the 2D Specific design. The power savings in the memory was 60%. Outstanding issues in 3DIC design include thermal/power codesign, cost, and test.

ACKNOWLEDGMENT

This work was supported, in part, by DARPA under contract FA8650-040C-7127, managed by AFRL, and by SRC as task 1824.

REFERENCES

[1] P. Garrou, C. Bower, and P. Ramm, “Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits,” (Wiley), 2008

[2] K. Banerjee, S. Souri, P. Kapur, K. Saraswat, “3-D ICs: A Novel Chip Design For Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration,” Proc. IEEE, Vol. 89, No. 5, 2001.

[3] T. Thorolfsson, K. Gonsalves, P.Franzon, “Design Automation for a 3DIC processor for synthetic aperture radar: A case study,” in Proc. DAC 2009, July 2009, pp. 51-56.

[4] H.P Hofstee, “Future Microprocessors and off-chip SOP Interconnect,” in IEEE Trans. Advanced Packaging, Vol. 27, No. 2, May 2004, pp. 301-303.

[5] S. Wilton and N. Jouppi. CACTI: an enhanced cache access and cycle time model. Solid-State Circuits, IEEE Journal of, 31(5):677-688, 1996

[6] T. Thorolfsson, S. Melamed, G. Charles, P. Franzon, “Comparative analysis of two 3D integraiton implementations fo a SAR Processor,: in Proc. IIII 3DIC 2009, pp.1-4.

[7] S. Melamed, T. Thorolfsson, A. Srinivasan, E. Cheng, P. Franzon, R. Davis, “Junction-level thermal extraction and simulatin of 3DICs,” in Proc. IEEE 3DIC 2009, pp. 1-7.

[8] E. Marinissen, Y. Zorian, “Testing 3D chips containing through silicon vias,” in Proc. International Test Conference, 2009, pp. 1-11.