Computing in 3D

Paul Franzon, Eric Rotenberg, James Tuck, W. Rhett Davis, Huiyang Zhou,
Joshua Schabel, Zhenquian Zhang,
J.Brandon Dwiel, Elliott Forbes, JoonmooHuh, Steve Lipa

Dept. of Electrical and Computer Engineering
North Carolina State University, Raleigh NC 27695
Contact Author Email: paulf@ncsu.edu

***Abstract* -- 3D technologies offer significant potential to improve total performance and performance per unit of power. After exploiting TSV technologies for cost reduction and increasing memory bandwidth, the next frontier is to create sophisticated logic on logic solutions that promise further increases in performance/power beyond those attributable to memory interfaces alone. These include heterogeneous integration for computing and exploitation of the high amounts of 3D interconnect available to reduce total interconnect power. Challenges include access for prototype quantities and the design of sophisticated static and dynamic thermal management methods and technologies, as well as test.**

Index Terms --- 3DIC, TSVs, CPUs, Processors

I. INTRODUCTION

3D technologies provide the potential to provide much higher chip-to-chip bandwidths, at lower power levels, than can be achieved with conventional packaging. Sometimes they offer potential for cost reduction through heterogeneous integration. Interposers offer a first step to 3D integration that simplifies physical planning and thermal management. However, intelligently designed 3D stacks offer significant further potential to achieve large improvements in performance per unit of performance beyond that of exploiting low-power memory interfaces. Other potential advantages include a low complexity method to enabling trusted design. This paper explores these potentials, starting with a brief review of the 3D technology set.

II. 3DIC TECHNOLOGY SET

3DIC technology incorporates a number of sub-technologies (Figure 1). To make a 3D chip stack, solder microbumps or copper-copper connections are used to join two chips and/or wafers face-to-face. Solder microbumps are typically limited to a 25 µm pitch today. Copper-copper joins can be made on a 6-8 µm pitch, though tighter is possible. Another key technology are the Through Silicon Vias (TSVs) that are placed within the (thinned) chips to provide backside connections or to enable further stacking. TSVs are fabricated with almost vertical sidewalls so that the area impact of the vias are minimized. The achievable pitch is limited by the chip thickness and sidewall verticality. For example, with chips thinned to 25 µm, a 25 µm TSV pitch is typical [1]. The TSVs are usually copper filled, though Tungsten is also a common choice.

Silicon interposers are built either using thin film metal deposition techniques or by exploiting Back End of the Line (BEOL) processes. The achievable interconnect pitch is in the 5-10 µm range for the former, and can be as tight as 1 µm with a BEOL technology. Typically four metal layers are built, but more are possible. Either potentially provides a capability for running a large number of wires between chips assembled side by side on the interposer. Large TSVs are placed through the interposer to support IO. Using interposers to connect chips horizontally is often referred to as "2.5D technology", and can be thought of as a high density silicon "circuit board").

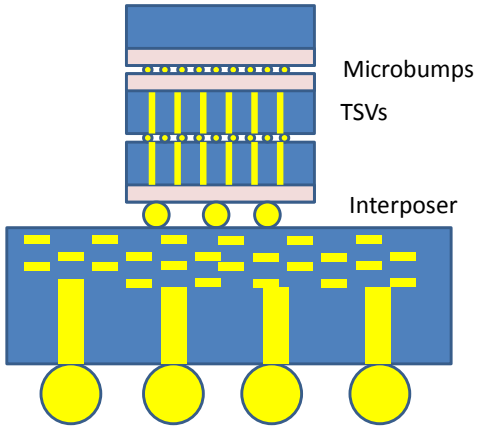


Figure 1. 3D technology elements. Microbumps are often built at 25 µm pitch with solder and 6-8 µm pitch with copper. Chip TSVs are typically built at 25 µm pitch but 2 µm pitch has been demonstrated. Interposers incorporate BEOL or thin film wiring (typically 4 layers but more have been demonstrated) and 100 µm pitch TSVs. Interposers are typically used to connect chips placed side by side on them (not shown).

A further differentiator in 3DIC technology is the overall integration flow. 2.5D technologies permit tight integration without having to deal with the added complexities of chip or wafer stacking. For 3DIC stacks, to date, stacking wafers has been the most common approach as it is the lowest cost and can use the high density copper-copper join processes. Wafer stacking requires all chips in the stack be the same size, and has yield implications. Though less common, chip-to-wafer stacking, including thinned chip to wafer stacking, does exist. To date, chip to wafer stacking can only be done with solder bump technologies. While higher cost, chip to wafer stacking does support pre-test and sort of chips before integration.

III. APPLICATIONS OF 3DIC

Cost Reduction. To date, the primary commercial exploitation of 3DIC has been for cost reduction. Examples include the following:

- Heterogeneous stacks of a photoreceptor layer and silicon circuit layer to cost reduce image sensors (e.g. Sony);
- Partitioning of large chips onto an interposer to cost reduce low-yielding components (e.g. Xilinx); and
- Integration a logic part, built in an advanced node, with an analog or high speed digital part built in a legacy node (e.g. Xilinx).

Memory Bandwidth and Power. The next "obvious" exploitation of 3D technologies is for the provision of high bandwidth, low-interface-power DRAM Assemblies. This solution directly addresses the power and bandwidth issues associated with the "memory wall". Examples include the Hybrid Memory Cube (HMC), WideIO and WideIO2, High Bandwidth Memory (HBM), and the Tezzaron DiRAM4. HMC is a memory chip stack together with one logic chip, the latter being used for control and IO functions. The composite HMC stack is then conventionally packaged, and IO is delivered via high speed SERDES channels. HBM is a DRAM chip stack intended for integration via an interposer or direct 3D stacking. The HBM interface is a wider and slower interface than in the case of HMC. WideIO and WideIO2 are a direct wide memory interface aimed at the mobile markets and is intended for direct 3D integration. The Tezzaron DiRAM4 is a high performance, low latency and low power memories also primarily intended for direct 3D integration via a large number of parallel channels.

Power Efficient Computing and Logic. Table I 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, 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] and [11]. The SIMD core was one designed at NCSU in 65 nm CMOS, and optimized for low-power operation. Some more detail on this core can be found in [6]. These conservative factors capture the slow down 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 figures were taken from [12] and are for DRAMs 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 a potential improvement of about 4x is possible through this approach.

For the interconnect, some of these figures are taken from the modeling and simulation study presented in [2], and again extrapolated to the 7 nm node. The interposer power was based on an extrapolation of the results presented in [13], with an assumption that 2/3rd 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 consumes less power than computation. Thus now it makes sense to move data if an overall advantage can be gained.

Table I. 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 x 64-bit RF	0.5 pJ/word
128 KB SRAM	0.9 pJ/word
L1 Dcache (16 KB)	62 pJ/16 B
L2 Dcache (2 MB)	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

Several examples of situations where an advantage can be gained will now be illustrated.

FFT Processor: Logic on Memory. This system consists of three stacked tiers with eight processing elements, one controller, thirty two SRAMs, and eight ROMs [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. The floorplan is designed so that all communications is vertical – there is no horizontal communications between PEs. The chip was implemented in the Lincoln Labs SOI 3D process. The die photo (Figure 2) 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 2 also shows the stacked chip floorplans. By breaking a large memory into 32 smaller memories memory power was reduced by 58%. (A similar tradeoff exists for DRAMs.)

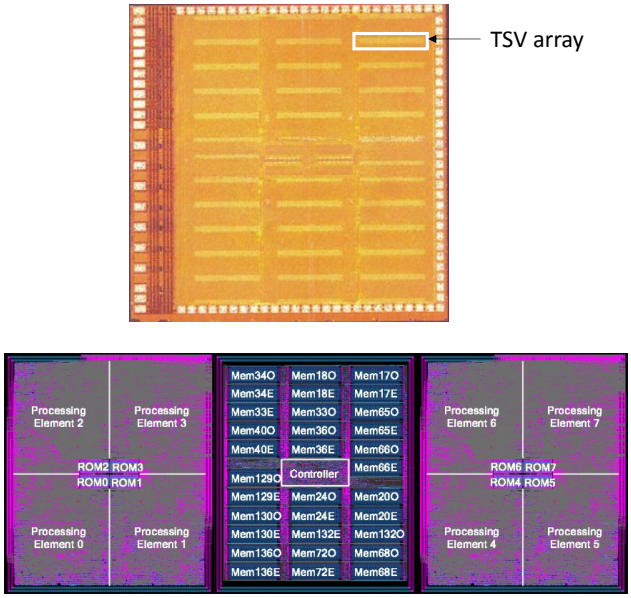


Figure 2. 3D FFT Engine Die Photo and floorplans of the three chips in the stack.

3D Heterogeneous Processor. A stack of two different CPUs are 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 [4]. 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 [5]. 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 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. In contrast 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. I.e. A 25% improvement in power per unit of performance.

This processor stack was taped out in a 3D 130 nm process in summer 2015. 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.

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 1 Gb 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 4x.

A 3D rendering of the overall floorplan is shown in Figure 3 showing the two processor stack integrated with the DRAM acting as the combined L2/L3 cache.

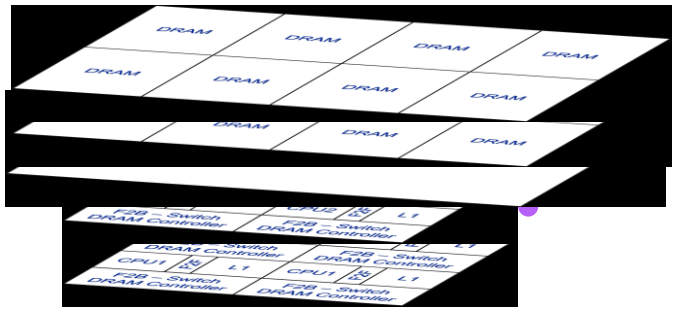


Figure 3. 3D Heterogeneous processor floorplan.

3D Multiprocessing. A 3D processor was designed with customized interconnect switch fabric incorporated into the chip stack, with the interconnect being customized to specific application sets. This design was benchmarked at the 7 nm node to achieve the results presented in Figure 4.

At the core of the design is a very power efficient SIMD core. This core was designed with a control overhead of less than 10% and is simulated to have power efficiency of 32 GFLOPS/W in the 65 nm node. Key aspects that were used to achieve such a power efficiency include shallow arithmetic piplines, a configurable Register File and a software controlled scratchpad SRAM that bypasses the power hungry memory hierarchy. Some more detail can be found in [6].

A layout of a 4x4 array (per chip) of 16-lane SIMD engines is shown in Figure 5. Multiprocessor communications is managed through a software managed light Message Passing Interconnect (MPI) interface. The actual traffic is communicated through a software managed configurable switch fabric. The common centroid 3D layout of the interconnect fabric serves to help minimize the power consumption and area overhead of managing parallelism.

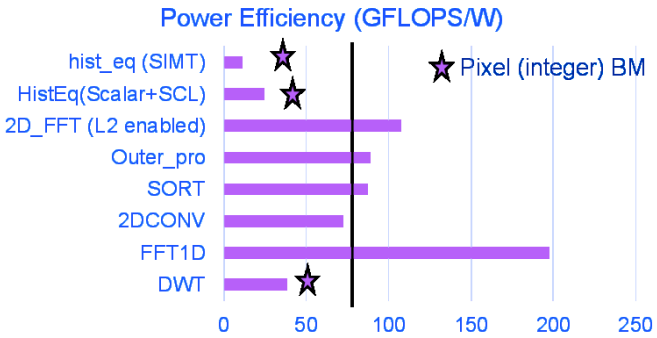


Figure 4. Results achieved in 3DECC our highly power efficient 3D computing platform.

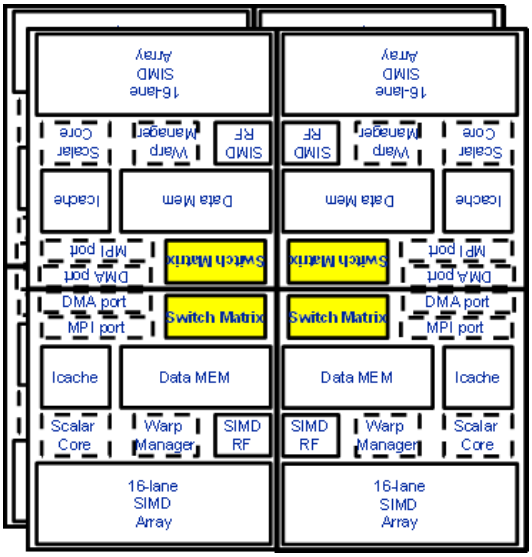


Figure 5. Stacked SIMD cores with common centroid switch matrix. This switch matrix is used to support circuit switching during operation.

Both of the chips above were prepared for a 3D tapeout that went out over summer 2015. The 3D process will be a two-chip stack connected to the DiRAM4 memory. The two-chip stack will itself be interconnected using an 8 µm pitch copper-copper process. Instead of narrow barrel TSVs, a simple wet etch will be used on the thinned die to back-expose the IO. This reduces prototyping cost and time at a loss of area efficiency.

*Logic on Logic with Auto-Partitioning***.** A modified CAD flow was applied to three different designs – a radar Processing Element (PE), an AES encryption engine, and a MIMO multipath radio processing engine. This partitioning approach leverages the high density and bandwidth of the microbump interface when two die are stacked Face to Face with each other. In this case, it was a copper-copper join process at a 6 µm pitch. All flip-flops are kept in one tier so that 3D clock distribution is not required. The radar PE was implemented in the Tezzaron bulk CMOS 3D 130 nm process [7] (Figure 6). This tool flow was run on a number of designs and the results summarized in Table II. On average, performance per unit of power was increased by 22% over a 2D baseline due to the decreases in wire length achieved through this partitioning approach.

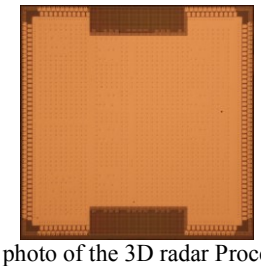


Figure 6. Die photo of the 3D radar Processing Element.

Table II. Results of running an auto-partitioned 3D result on different design. Comparison is made with a 2D chip baseline. "PE" refers to a radar processing elements. The others are AES encryption and MIMO communications design results.

	Total Wire Length (% Change)	Max Frequency (% Change)	Parasitic Power (% Change)	Total Power) (% Change)
PE 3D Seq.	-17.1%	+7.1%	-15.5%	-4.7%
PE 3D Sim.	-17.7%	+16.2%	-27.9%	-7.7%
PE 3D True	-21.0%	+22.6%	-45.2%	-12.9%
AES 3D Seq.	-8.0%	+15.3%	-19.6%	-2.6%
MIMO 3D Seq.	+216.1%	+17.1%	-34.9%	-5.1%

IV. 3DIC CHALLENGES

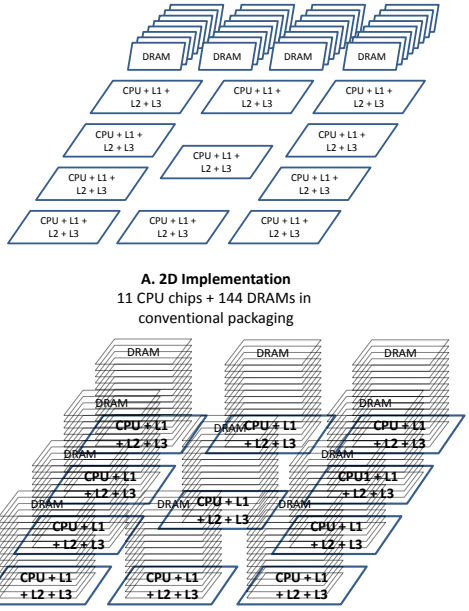
Frankly the main challenge today is getting fabrication access in low volumes suited to prototype exploration. Fortunately, as volume applications like imagers and memories force process standardization and process maturity, ready access to fabrication with reasonable turnaround is likely to become more assured.

Thermal management is a significant challenge for 3DIC systems. When DRAM is mixed with logic, it is desirable to keep the DRAM temperature below 85 C in order to meet JEDEC refresh standards. Meanwhile, the logic typically runs at 105 C or higher creating the need for a thermal barrier. For logic on logic 3D systems, the challenge is compounded that the logic heat flux is higher than the DRAM heat flux, and thus the overall heat flux is greater. In addition, features designed to reduce the total power requirement, such as short reach 2.5D and 3D interconnect, can result in greater proximity and thus

increased heat flux. Table III (adapted from [8]) shows a summary of heat flux, vs. compute efficiency for different 2D and 3D multiprocessor scenarios. The scenarios that were analyzed are illustrated in Figures 7 and 8. The key result is that there is a tradeoff between power density and power efficiency. Within these examples a range of a 37% reduction in power efficiency comes at the cost of an order of magnitude increase in heat flux. This points at the need for improved thermal technologies. None the heat flux issues can be ameliorated through efficient layout [9].

Table III. Heat Flux vs. Power Efficiency for different design scenarios.

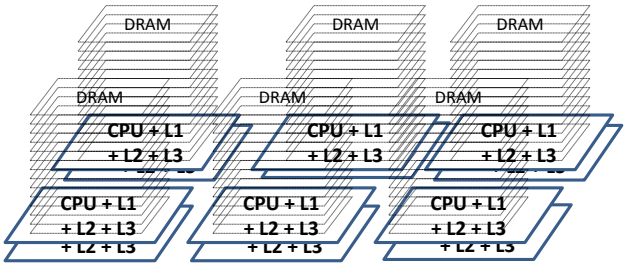
Scenario	Peak Efficiency (mW/GFLOPS)	Av. Power Flux (W/mm^2)
A. 2D	56	0.29
B. Memory on Logic	47	0.26
C. Stacked Memory on two logic layers	35	3.7



A. 2D Implementation
11 CPU chips + 144 DRAMs in conventional packaging

B. Basic Memory On Logic
11 CPU chips + 144 stacked DRAMs

Figure 7. Two scenarios used for thermal analysis of a multiprocessor.



C. Memory On Logic On Logic
6 CPU chips + 144 stacked DRAMs

Figure 8. Third scenario used for thermal analysis of a multiprocessor.

Another commonly cited challenge is test and Design for Test. For 2.5D integration, conventional die sort techniques can be used, though the IO might remain largely untestable until final integration. Partial coverage of IO faults can be done using loop back techniques. There are techniques to probe fine pitch IO but cost issues might preclude their use in production.

For 3D chip stacks, different approaches are needed depending on whether the 3D chips are stacked in wafer form or chip form. If stacked in chip form, then conventional test can be used to sort the die before bonding, again with some reduction in IO test coverage unless a fine pitch probing technology can be used. Then post-assembly at least an integration test is needed to ensure 3D connectivity. There are several ways to do this including 3D extensions of SOC and JTAG test methodologies. In the case of the Heterogeneous Computer described above, the integration bus has self-test support [5] that is intended to be run post-integration.

If the chips are stacked in wafer form, there is an issue that the stacked chips can't be sorted before mating, unless entire wafers are rejected. Thus yield would degrade exponentially with the number of stacked parts. If this becomes a cost issue, then some level for exercising post integration redundancy might be valuable. Also full post integration test will be needed.

V. CONCLUSION

3DIC offers opportunities to gain improvements in performance per unit of power equal to that of a node of Moore's Law. The complexity of access is going down and thermal issues require close attention.

ACKNOWLEDGEMENT

This work was funded in part by the DARPA 3DIC, DAHI and PERFECT programs, and by Sematech, Intel and Qualcomm.

REFERENCES

[1] Y. Civale, et.al., “Through Silicon Via Technology for Three-Dimensional Integrated Circuit Manufacturing,” in Proc. 35th International Electronic Manufacturing Technology Conference, 2012.

[2] A. Karim, P.D. Franzon, A. Kumar, “Power Comparison of 2D, 3D, and 2.5D Interconnect Solutions and Power Optimization of Interposer Interconnect,” in Proc. IEEE ECTC 2013.

[3] W. Davis, E. Oh, A. Sule, T. Thorolfsson, and P.D. Franzon, “Application Exploration for 3-D Integrated Circuits: TCAM, FIFO and FFT Case Studies,” in IEEE Trans. On VLSI, Vol. 17, No. 4, April 2009, pp. 496-506

[4] Rotenberg, E.; Dwiel, B.H.; Forbes, E.; Zhenqian Zhang; Widialaksono, R.; Basu Roy Chowdhury, R.; Tshibangu, N.; Lipa, S.; Davis, W.R.; Franzon, P.D., "Rationale for a 3D heterogeneous multi-core processor," Computer Design (ICCD), 2013 IEEE 31st International Conference on , vol., no., pp.154,168, 6-9 Oct. 2013

[5] Z. Zhang, B. Noia, K. Chakraparthy, P. Franzon, “Face to Face Bus Design With Built-in Self-Test in 3DICs,” in Proc. IEEE 3DIC.

[6] P. Franzon et.al. “3D-Enabled Customizable Embedded Computer (3DECC)” in Proc. 3DIC 2014.

[7] T. Thorolfsson, S. Lipa, and P.D. Franzon, “A 10.35 mW/GFLOP Stacked SAR DSP Unit using Fine-Grain Partitioned 3D Integration,” in Proc. CICC 2012.

[8] P. Franzon, A. Bar-Cohen, “Thermal Requirements in Future 3D Processors,” in Proc. 3DIC 2013.

[9] M Saeidi, K. Samadi, A. Mittal, R. Mittalk, “Thermal Implications of Mobile 3D-ICs,” in Proc. 3DIC 2014.

[10] S. Borkar, “The exascale challenge,” in VLSI Design Automation and Test (VLSI-DAT), 2010 International Symposium on, 2010, pp. 2–3.

[11] H. Esmaeilzadeh, E. Blem, R. St.Amant, K. Sankaralingam, and D. Burger, “Dark silicon and the end of multicore scaling,” in Computer Architecture (ISCA), 2011 38th Annual International Symposium on, 2011, pp. 365–376.

[12] T. Vogelsang, “Understanding the Energy Consumption of Dynamic Random Access Memories,” in Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture, Washington, DC, USA, 2010, pp. 363–374.

[13] J. Poulton, W. Dally, X. Chen, J. Eyles, T. Greer, S. Tell, and C. Gray, “A 0.54pj/b 20gb/s ground-referenced single-ended short-haul serial link in 28nm CMOS for advanced packaging applications,” in Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International, 2013, pp. 404–405.