DOI:10.1145/2524713.2524725

Centip3De: A Many-Core Prototype Exploring 3D Integration and Near-Threshold Computing

By Ronald G. Dreslinski, David Fick, Bharan Giridhar, Gyouho Kim, Sangwon Seo, Matthew Fojtik, Sudhir Satpathy, Yoonmyung Lee, Daeyeon Kim, Nurrachman Liu, Michael Wieckowski, Gregory Chen, Dennis Sylvester, David Blaauw, and Trevor Mudge

Abstract

Process scaling has resulted in an exponential increase of the number of transistors available to designers. Meanwhile, global interconnect has not scaled nearly as well, because global wires scale only in one dimension instead of two, resulting in fewer, high-resistance routing tracks. This paper evaluates the use of three-dimensional (3D) integration to reduce global interconnect by adding multiple layers of silicon with vertical connections between them using through-silicon vias (TSVs). Because global interconnect can be millimeters long, and silicon layers tend to be only tens of microns thick in 3D stacked processes, the power and performance gains by using vertical interconnect can be substantial. To address the thermal issues that arise with 3D integration, this paper also evaluates the use of near-threshold computing—operating the system at a supply voltage just above the threshold voltage of the transistors.

Specifically, we will discuss the design and test of Centip3De, a large-scale 3D-stacked near-threshold chip multiprocessor. Centip3De uses Tezzaron's 3D stacking technology in conjunction with Global Foundries' 130 nm process. The Centip3De design comprises 128 ARM Cortex-M3 cores and 256MB of integrated DRAM. Silicon measurements are presented for a 64-core version of the design.

1. INTRODUCTION

High-performance microprocessors now contain billions of devices per chip.[11] Meanwhile, global interconnect has not scaled nearly as well since global wires scale only in one dimension instead of two, resulting in fewer, high-resistance routing tracks. Because of this, and increasing design complexity, the industry is moving toward system-on-chip (SoC) designs or chip-multiprocessors (CMP). In these designs, many components share the same die, as opposed to one giant processor occupying it entirely. These systems typically include processing cores, memory controllers, video decoders, and other ASICs.

Three-dimensional (3D) integration reduces the global interconnect by adding multiple layers of silicon with vertical interconnect between them, typically in the form of through-silicon vias (TSVs). Because global interconnect can be millimeters long, and silicon layers tend to be only tens of microns thick in 3D stacked processes, the power and performance gains by using vertical interconnect can be substantial. Additional benefits include the ability to mix different process technologies (CMOS, bipolar, DRAM, Flash, optoelectronics, etc.) within the same die, and increased yield through "known good die" testing for each layer before integration.[8]

Recently, a DARPA program provided the opportunity to evaluate an emerging 3D stacking technology that allows logic layers to be stacked on top of a DRAM. As part of this project our group at the University of Michigan was given the chance to explore architectures that leverage the high-bandwidth, low-latency interface to DRAM afforded by 3D integration. This paper will discuss the resulting design and test of Centip3De, a large-scale CMP that was recently presented at HotChips'24 and ISSCC[5] (Figure 1). Centip3De uses Tezzaron's FaStack® 3D stacking technology[14] in conjunction with Global Foundries' 130 nm process. The Centip3De design comprises 128 ARM Cortex-M3 cores and 256MB of integrated DRAM. Silicon measurements are presented for a 64-core version of the design in Section 4 which show that Centip3De achieves a robust design that provides both energy-efficient (3930 DMIPS/W) performance for parallel applications and mechanisms to achieve substantial single-thread performance (100 DMIPS) for serial phases of code.

2. BACKGROUND

2.1. 3D integration

The Centip3De prototype uses Tezzaron's FaStack® technology that stacks wafers of silicon (as opposed to individual dies) using copper bonding.[14] A step-by-step process diagram is shown in Figure 2. Before each wafer pair is bonded, a layer of copper is deposited in a regular honeycomb pattern that is then thinned. Due to the regularity

The original version of this paper is entitled "Centip3De: A 3930 DMIPS/W Configurable Near-Threshold 3D Stacked System With 64 ARM Cortex-M3 Cores" and was published in the *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*. San Francisco, CA, February 2012, pp. 190–191.

Figure 1. Floorplan accurate artistic rendering. Centip3De includes seven layers, including two core layers, two cache layers, and three DRAM layers.

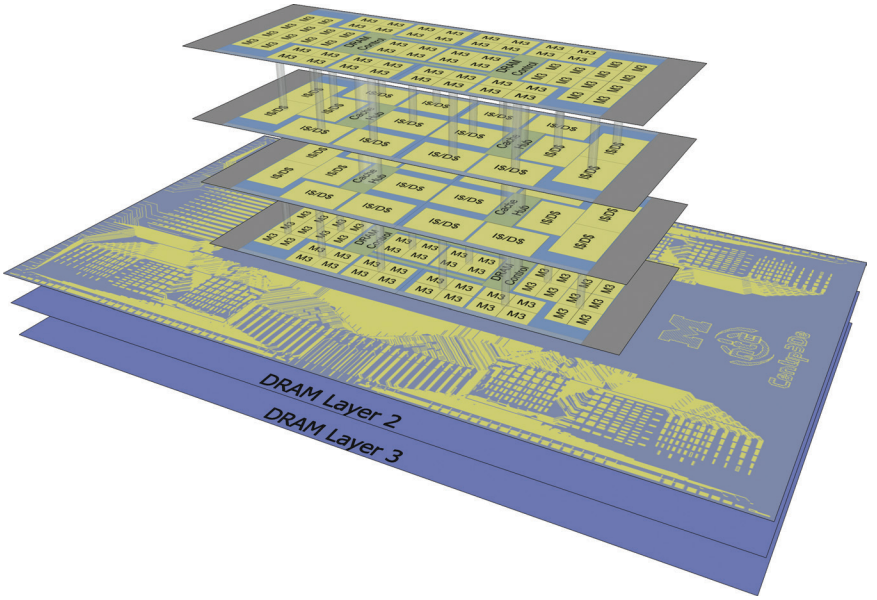
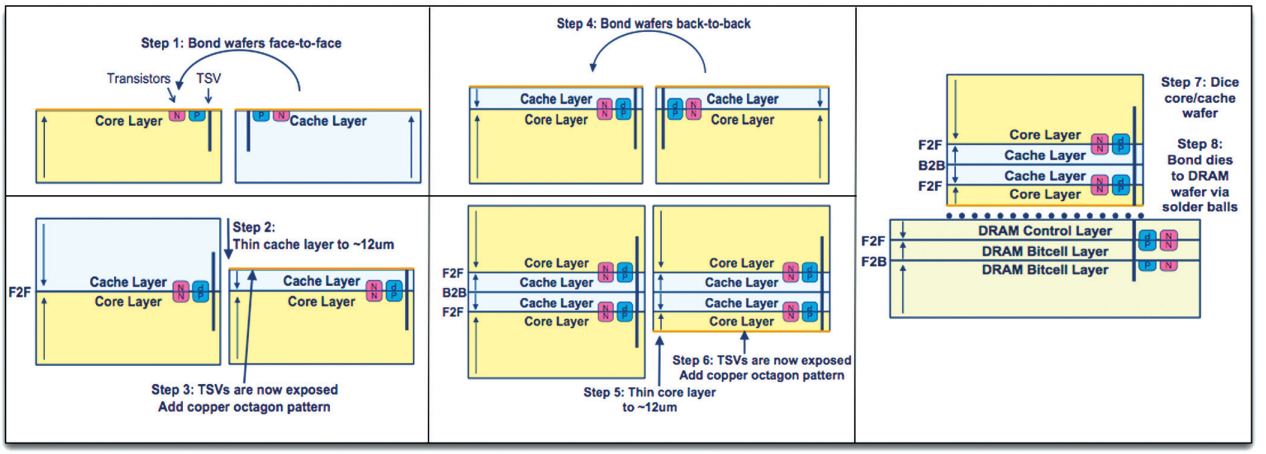


Figure 2. Centip3De's 3D stacking process. The core and cache wafers are bonded with a face-to-face (F2F) connection, then thinned to expose the TSVs. Next two pairs of wafers are bonded back to back (B2B) and thinned. Finally, the logic layers are diced and die-to-wafer bonded to the DRAM.



of the pattern, it is very flat after thinning, and when two wafers with this pattern are pressed together with heat, the copper forms a strong bond. After a bond, one side of the resulting wafer is thinned so that only a few microns of silicon remains, which exposes TSVs for 3D bonding, flip-chip, or wirebonding. The TSVs are made of Tungsten, which has thermal-mechanical properties suitable for creating vias. They are created after the transistors, but before the lowest metal layers, preventing the loss of any metal tracks. Since their diameter is small (1.2 µm), their length is short (around six microns after thinning), and they only couple to the bulk silicon, their parasitic resistance and capacitance is very small (about as much capacitance as a small gate). The finished wafer stack has a silicon pitch of approximately 13 microns with a per-interface TSV density of up to 160,000/mm^2. The combination of thin silicon layers, a high density of Tungsten TSVs, and planes of bonding copper also helps to maximize heat dissipation in this stacking technology.

A similar process is used to stack two dies of different sizes. In this design, the Tezzaron Octopus DRAM[15] is much larger than the core and cache layers. To stack these, the wafer of smaller dies is thinned, coated with copper, and then diced (Figure 2, step 7). The wafer of larger dies is also thinned and coated with copper. The smaller dies are put in a stencil to hold them in place and then are pressed together with the larger wafer to make the bond (step 8). A larger copper pattern than that

used for the wafer-to-wafer bond is needed to support the less-precise alignment of die to wafer.

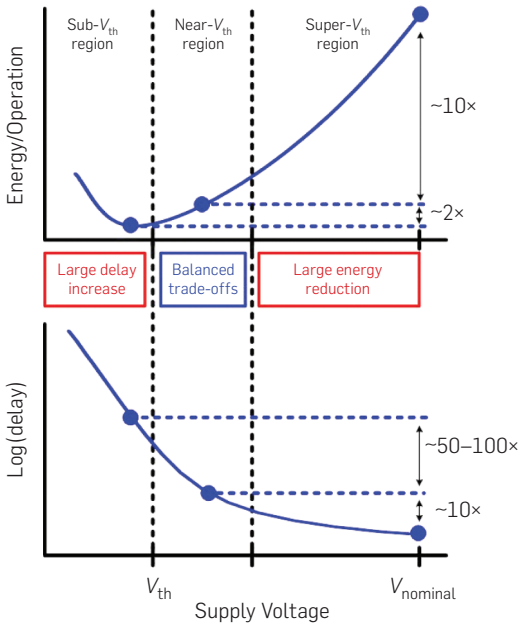
Alternative 3D and 2.5D techniques. The landscape of 3D stacking technologies is quite diverse and provides a wide range of possibilities. For the logic layers of Centip3De, we used the Tezzaron wafer-to-wafer 3D process, which provides unusually small TSVs on an extremely fine pitch (5 µm). This allows complicated designs to be implemented where even standard cells can be placed and routed across multiple layers within a synthesized block. However, this technology relies on wafer-to-wafer bonding for the logic layers which can result in yield issues for large runs. Alternative TSV technologies exist that rely on micro-bump die-to-wafer or die-to-die bonding which resolves some of the yield issues, but have a much larger TSV pitch. The die-to-wafer bonding was used by Centip3De to connect the logic layers to the DRAM layers. These types of 3D stacking are more useful for routing buses and interconnects between synthesized blocks, but are not ideal for 3D connections within synthesis blocks. A more near term integration solution is 2.5D technology that employs silicon interposers. This technique takes the micro-bump a step further by placing dies adjacent to each other and connecting them through an interposer. The 2.5D technology helps mitigate the thermal issues associated with 3D integration, at the expense of longer interconnects and lower TSV densities that are associated with micro-bumps.

2.2. Near-threshold computing (NTC)

Energy consumption in modern CMOS circuits largely results from the charging and discharging of internal node capacitances and can be reduced quadratically by lowering supply voltage (V_{dd}). As such, voltage scaling has become one of the more effective methods to reduce power consumption in commercial parts. It is well known that CMOS circuits function at very low voltages and remain functional even when V_{dd} drops below the threshold voltage (V_{th}). In 1972, Swanson and Meindl derived a theoretical lower limit on V_{dd} for functional operation, which has been approached in very simple test circuits.[6, 13] Since this time, there has been interest in subthreshold operation, initially for analog circuits[16] and more recently for digital processors,[12] that demonstrates operation with a V_{dd} below 200 mV. However, the lower bound on V_{dd} in commercial applications is typically set to 70% of the nominal V_{dd} to guarantee robustness without too significant a performance loss.

Given such wide voltage scaling potential, it is important to determine the V_{dd} at which the energy per operation (or instruction) is optimal. In the superthreshold regime ($V_{dd} > V_{th}$), energy is highly sensitive to V_{dd} due to the quadratic scaling of switching energy with V_{dd}. Hence voltage scaling down to the near-threshold regime ($V_{dd} \sim V_{th}$) yields an energy reduction on the order of 10× at the expense of approximately 10× performance degradation, as seen in Figure 3. However, the dependence of energy on V_{dd} becomes more complex as voltage is scaled below V_{th}. In the subthreshold regime ($V_{dd} < V_{th}$), circuit delay

Figure 3. Energy and delay in different supply voltage operating regions.



increases exponentially with V_{dd}, causing leakage energy (the product of leakage current, V_{dd}, and delay) to increase in a near-exponential fashion. This rise in leakage energy eventually dominates any reduction in switching energy, creating an energy minimum, V_{opt}, seen in Figure 3.

The identification of an energy minimum led to interest in processors that operate at this energy optimal supply voltage[17, 19] (referred to as V_{min} and typically 250–350 mV). However, the energy minimum is relatively shallow. Energy typically reduces by only ~2× when V_{dd} is scaled from the near-threshold regime (400–500 mV) to the subthreshold regime, though delay rises by 50–100× over the same region. While acceptable in ultra-low energy sensor-based systems, this delay penalty is not acceptable for a broader set of applications. Hence, although introduced roughly 30 years ago, ultra-low voltage design remains confined to a small set of markets with little or no impact on mainstream semiconductor products. In contrast, the tradeoffs in the NTC region are much more attractive than those seen in the subthreshold design space and open up a wide variety of new applications for NTC systems.

2.3. NTC and 3D synergy

One of the biggest drawbacks to 3D integration is the thermal density. By stacking dies on top of each other, local hotspots may effect neighboring dies, and hotspots that occur on top of each other complicate cooling.[9] Compounding this problem further is the stagnation of supply-voltage scaling in newer technology nodes leading to a breakdown of Dennard scaling theory. As a result, thermal density in 2D chips is also increasing.[3] Both of these issues are exacerbated by the fact that these hotter logic dies are now stacked on top of DRAM, resulting in the need for more frequent refreshes. On the other hand, NTC operation suffers from

reduced performance because of lower supply voltages.

However, there is an important synergy between these two technologies: (1) by using NTC, the thermal density is reduced allowing more dies to be stacked together as well as on top of DRAM; and (2) by using 3D integration, additional cores can be added to provide more performance for parallel applications, helping to overcome the frequency reduction induced by NTC.

3. CENTIP3DE ARCHITECTURE

Centip3De is a large-scale CMP containing 128 ARM Cortex-M3 cores and 256MB of integrated DRAM. Centip3De was designed as part of a DARPA project to explore the use of 3D integration. We used this opportunity to explore the synergy with NTC as well. The system is organized into 32 computation clusters, each containing four ARM Cortex-M3 cores.[2] Each cluster of cores shares a 4-way 8KB data cache, a 4-way 1KB instruction cache, and local clock generators and synchronizers. The caches are connected through a 128-bit, 8-bus architecture to the 256MB Tezzaron Octopus DRAM,[15] which is divided into 8 banks, each of which has its own bus, memory interface, and arbiters. Figure 4 shows a block diagram for the architecture, with the blocks organized by layer. Centip3De also has an extensive clock architecture to facilitate voltage scaling.

3.1. Designing for NTC: A clustered approach

A key observation in NTC design is the relationship between activity factor and optimal energy point.[4] The energy consumed in a cycle has two main components: leakage energy and dynamic energy. At lower activity factors, the leakage energy plays a larger role in the total energy, resulting in higher V_{opt}. Figure 5 illustrates this effect in a 32 nm simulation, where activity factor was adjusted to mimic different components of a high performance processor. Memories in particular have a much lower activity factor and thus a much higher V_{opt} than core logic. Although NTC does not try to achieve energy-optimal operation, to maintain equivalent energy↔delay tradeoff points, the relative chosen NTC supply points should track with V_{opt}. Thus, V_{NTC_memory} should be chosen to be higher than V_{NTC_core} for them to maintain the same energy↔delay

Figure 4. High-level system architecture. Centip3De is organized into four-core clusters, which connect to eight DRAM buses. The diagram is organized by layer.

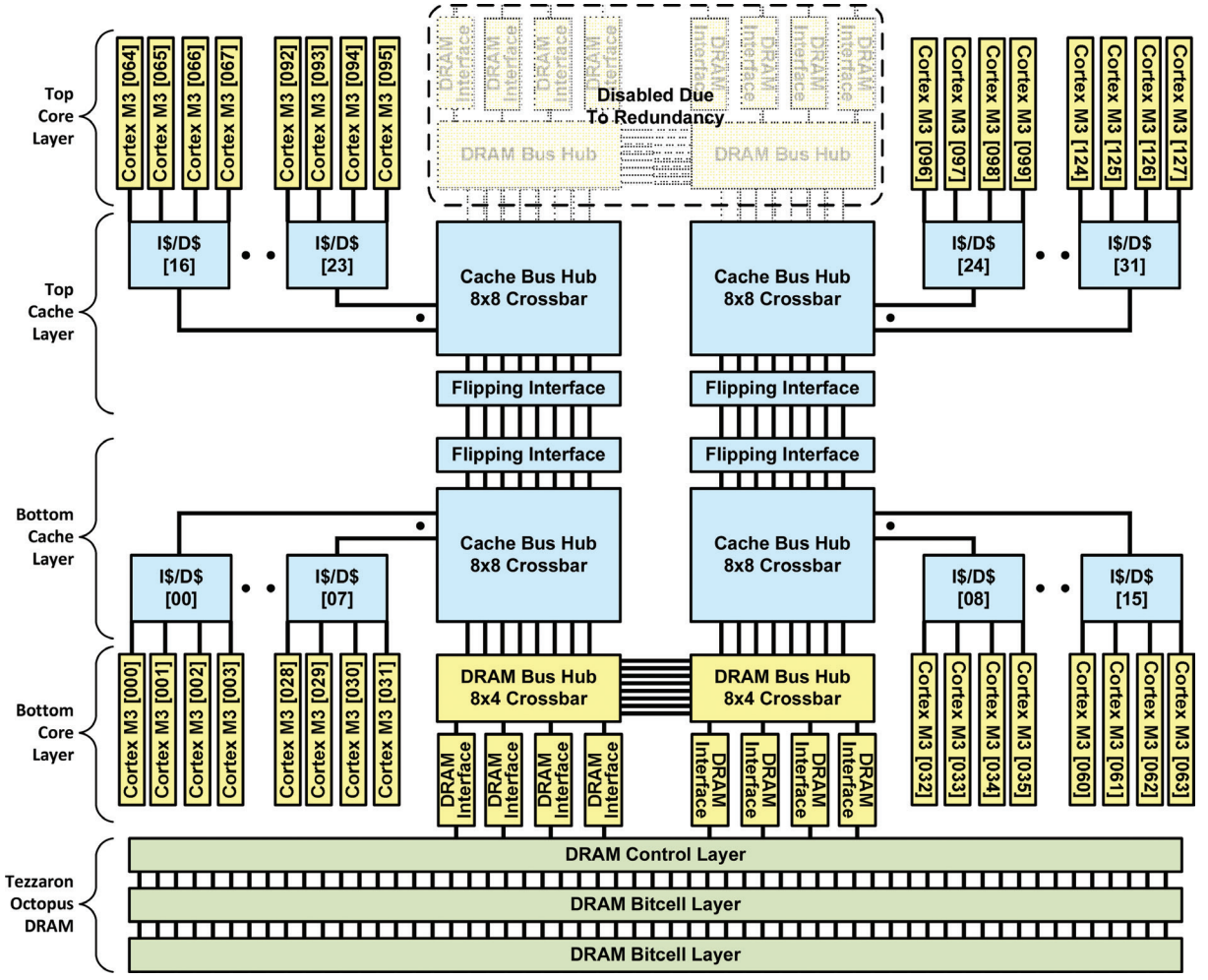


Figure 5. Activity factor versus minimum operating energy. As activity factor decreases, the leakage component of energy/operation increases thereby making the energy optimal operating voltage increase. To account for this, Centip3De operates caches at a higher voltage and frequency than the cores. An 8T SRAM design was used to ensure functionality and yield at V_{opt}.

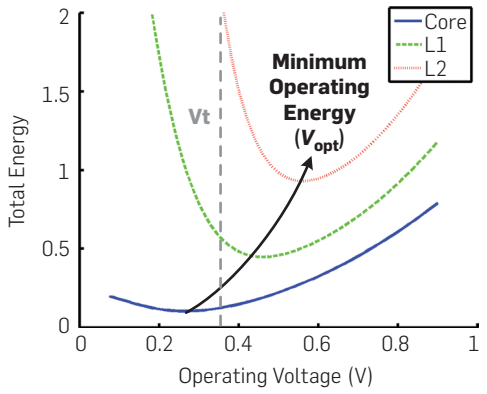
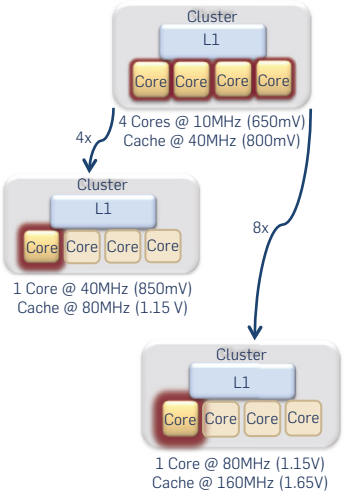


Figure 6. Boosting architecture. Diagram of how a cluster can be boosted to 4× and 8× single-thread performance by disabling some cores and increasing the voltage of the rest.



trade-off point. In addition, because V_{opt} is lower than the minimum functional voltage[a] for typical SRAM designs, a low-voltage tolerant memory design is necessary. This can be achieved by either specially designed 6T SRAM[18] or an 8T SRAM design.[10]

In Centip3De, we have used this observation to reorganize our CMP architecture. Instead of having many cores each with an independent cache, the cores have been organized into 4-core clusters and their aggregate cache space combined into a single, 4×-larger cache. This larger cache is then operated at a higher voltage and frequency to service all four cores simultaneously. To do this, the cores are operated out-of-phase and are serviced in a round robin fashion. Each core still sees a single-cycle interface and has access to a much larger cache space when necessary.

The NTC cluster architecture also provides mechanisms for addressing a key limiter in parallelization: intrinsically serial program sections. To accelerate these portions of the program, Centip3De uses per-cluster DVFS along with architecturally based boosting modes. With two cores of the cluster disabled, the cluster cache access pipeline can be reconfigured to access tag arrays and data arrays in parallel instead of sequentially and change from a 4× core↔cache frequency multiplier to a 2× multiplier. The remaining core(s) are voltage boosted and operate at 2× their original frequency, roughly doubling their single-threaded performance. A diagram of boosting-based approaches with their operating voltage and frequencies in the 130 nm Centip3De design is shown in Figure 6. The performance of the boosted cores is further improved by the fact that they access a larger cache and hence have a lower miss rate. To offset the related increase in cluster power and heat, other clusters can be disabled or their performance reduced.

[a] The operating voltage at which enough SRAM cells, in the presence of variation, function to meet yield constraints.

The Centip3De NTC cluster architecture has manufacturability benefits as well. By using lower voltages, many of the components have improved lifetime and reliability. Redundancy in the cluster-based architecture allows faulty clusters to be disabled and can be coupled with known good die techniques to further increase yield. Centip3De's boosting modes in combination with "silicon odometer" techniques improve performance while maintaining lifetime.[7]

Additional benefits of the NTC cluster architecture include reduced coherence traffic and simplified global routing. Coherence between the cores is intrinsically resolved in the cache while the top level memory architecture has 4× fewer leaves. Drawbacks include infrequent conflicts between processes causing data evictions and a much larger floorplan for the cache. In architectural simulation, however, we found that the data conflicts were not significant to performance for analyzed SPLASH2 benchmarks, and we effectively addressed floorplanning issues with 3D design. In particular, the core↔cache interface is tightly coupled and on the critical path of most processors and by splitting this interface to share several cores in a 2D design results in a slow interface.

To address the floorplan issues, we leverage the 3D interface to co-locate the cores directly above the caches and minimize the interconnect. The floorplan of the core↔cache 3D cluster is shown in Figure 7. The face-to-face connections from the cores appear in the center of the floorplan. This is particularly beneficial since the SRAMs use all five metal layers of this process, thereby causing significant routing congestion. By using 3D stacking, we estimate that cache routing resource requirements were reduced by approximately 30%. Because the parasitic capacitance and resistance of the F2F connections are small, the routing gains are not offset by significant loading (Figure 8).

Figure 7. Cluster floorplan with F2F connections represented as dots. There are 1591 F2F connections per cluster saving approximately 30% of routing and enabling the clustered architecture.

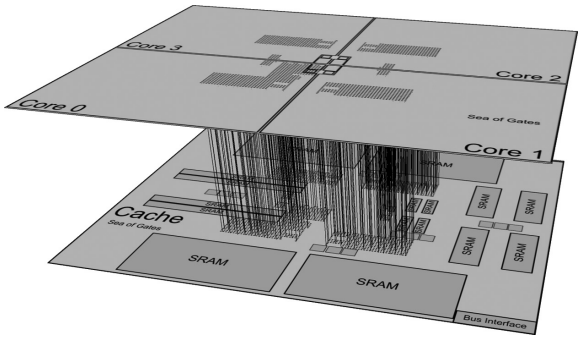
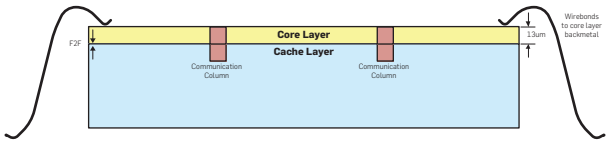


Figure 8. Side view of the measured system. The two-layer stack is formed by a face-to-face wafer bond and then the core layer is thinned to expose TSVs. Wirebonds are used to connect the TSVs to a test package. Both the face-to-face and TSV technologies are tested with this prototype.



4. MEASURED RESULTS

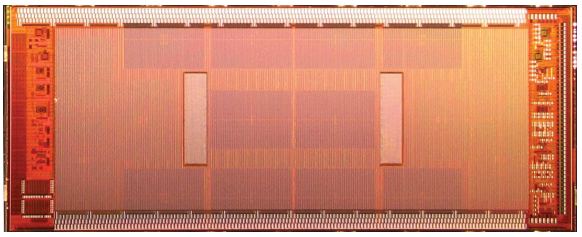
Silicon for Centip3De is coming back in three stages. Already received and tested is a two-layer system that has a core layer and a cache layer bonded face to face. This system is thinned on the core side to expose the TSVs, and an aluminum layer is added to create wirebonding pads that connect to them. A die micrograph is shown in Figure 9.

The next system will have four layers including two core layers and two cache layers. It will be formed by bonding two core–cache pairs face to face, thinning both pairs on the cache side to expose the TSVs, adding a copper layer, and then bonding back to back. One core layer is thinned to expose TSVs and same aluminum wirebonding pads are added.

The final system will include all seven layers, including two core layers, two cache layers, and three DRAM layers. The wirebonding sites will be on the DRAM, which has much larger layers than the core and cache layers.

Figure 10 shows silicon measurements and interpolated data for the fabricated 2-layer system for different cluster configurations running a Dhrystone test. The default NTC cluster (slow/slow) configuration operates with four cores at 10 MHz and caches at 40 MHz, achieving 3,930 DMIPS/W. Based on FO4 delay scaling, 10 MHz is projected to translate to 45 MHz in 45 nm SOI CMOS. Latency critical threads can operate in boosted modes at 8× higher frequency. One-core and two-core boosted modes provide the same throughput, 100 DMIPS/cluster (estimated as 450 in 45 nm), while enabling a tradeoff between latency and power (Figure 10). The system

Figure 9. Die micrograph of two-layer system. Two layers are bonded face to face. The clusters can be seen through the backside of the core layer silicon. Wirebonding pads line the top and bottom edges. The system consists of 16 clusters of 4 cores (64 total cores).



bus operates at 160–320 MHz, which supplies an ample 2.23–4.46GB/s memory bandwidth. The latency of the bus ranges from 1 core cycle latency for 10 MHz cores to 6 cycles when cores are boosted to 80 MHz. As a point of comparison, the ARM Cortex-A9 in a 40 nm process is able to achieve 8000 DMIPS/W.[1] At peak system efficiency, Centip3De achieves 3930 DMIPS/W in a much older 130 nm process, and when scaled to 45 nm Centip3De achieves 18,500 DMIPS/W (Table 1).

The results in Figure 10 present many operating points to select based on workload or workload phase characteristics. When efficiency is desired, choosing the slow core and slow memory results in the most efficient design. For computationally intensive workloads, additional throughput can be obtained, at the expense of power, by using the fast core and slow memory configuration. For memory bound workloads, the core can remain slow and the bus speed can be increased to provide more memory bandwidth. For workloads or phases that require higher single-thread performance (to address serial portions of code), the number of cores in a cluster can be reduced and the core speeds increased.

When boosting clusters within a fixed TDP environment, it may require disabling or down-boosting other clusters to compensate for that cluster's increase in power consumption. A package with a 250 mW TDP can support all 16 clusters in four-core mode (configuration 16/0/0/0, with the number of clusters in each mode designated as 4C/3C/2C/1C). Up to five clusters can be boosted to three-core mode (11/5/0/0) while remaining within the budget. To boost a cluster to one-core mode within the TDP, however, would require disabling other clusters, resulting in system configuration 9/0/0/1. By boosting clusters, Centip3De is able to efficiently adapt to processing requirements. Figure 11 shows a range of system configurations under a fixed TDP of 250 mW. On the left are high-efficiency configurations, with more aggressively boosted configurations on the right, which provide single-threaded performance when needed. Overall, the Centip3De design offers programmers with a wide range of power/throughput/single-thread performance points at which to operate.

5. DISCUSSION

In designing Centip3De we learned several lessons. First, supporting 3D LVS and DRC checking was a challenge

Figure 10. Power analysis of the 64-core system. Power breakdowns for 4-, 3-, 2-, and 1-core modes. Each mode has a slow (NTC) core option or a fast (boosted) option where the frequency/voltage is increased. Each option provides a trade-off in the efficiency/throughput/single-thread performance space. Overall, Centip3De achieves a best energy efficiency of 3930 DMIPS/W.

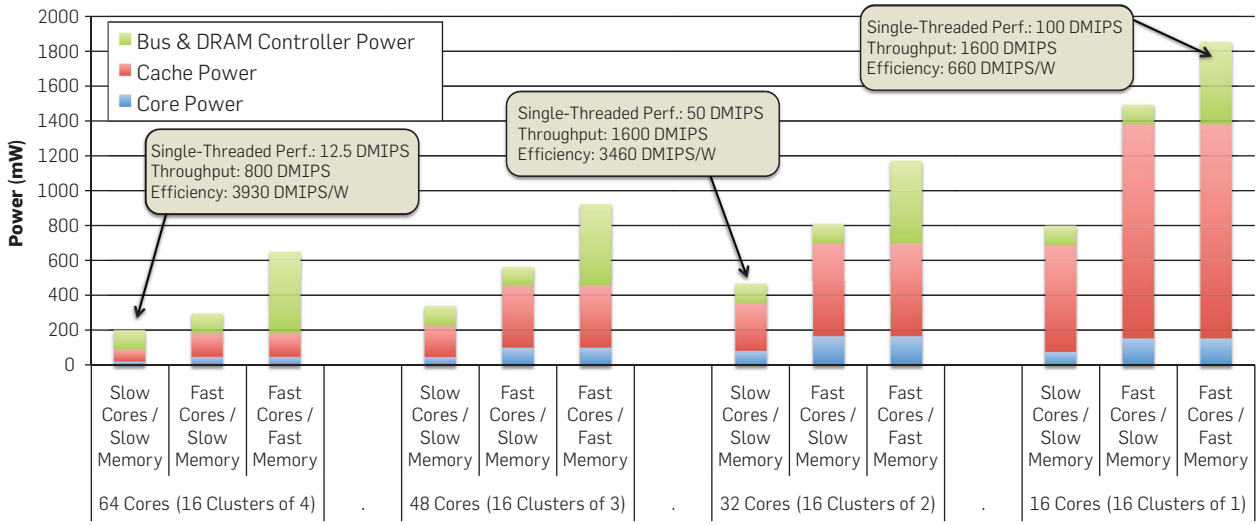
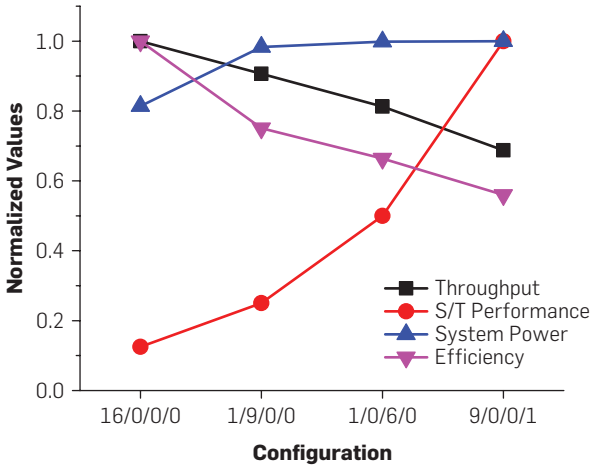


Table 1. Design and technology data. Connection numbers include only signals and do not include minimum density numbers for TSV or power/gnd connections. F2F connections are for a single F2F bonding interface.

Logic layer dimensions	2.66 × 5 mm
Technology	130 nm
Metal layers	5
Core layer devices	28.4 M
Cache layer devices	18.0 M
Core layer thickness	12 µm
F2F connection pitch	5 µm
F2F connections/cluster	1,591
Bus F2F connections	2,992
Total F2F connections	28,485
B2B connection pitch	5 µm
Total B2B connections (TSV)	3,024
DRAM connection pitch	25 µm
DRAM connections (TSV)	3,624

Figure 11. Range of system configurations under a 250 mW fixed TDP. The number of clusters in each mode is listed as 4-core/3-core/2-core/1-core. Each configuration emphasizes a different cluster mode, with the most energy-efficient configurations on the left, and the highest single-threaded performance on the right.



that required a large amount of design time to develop scripts for these operations. The good news is that in the two years since we first started Centip3De, EDA tools have made significant progress in supporting 3D integration. Second, when we designed Centip3De we were concerned with the amount of clock skew that would be introduced by TSVs, so we designed highly tunable clock delay generators as insurance against timing mismatch. However, measurements show that the skew through the TSVs was quite small. Spice simulations indicate that a significant amount of power (~30–60%) is being used in these delay generators, particularly in NTC mode (~60%). Unfortunately, we were unable to subtract the unnecessary power these delay generators consume, because they were not on their own supply rail. If we were able to reduce that power using less tunable delay generators, we expect the efficiency would be far better at NTC than we observed, achieving ~9900 DMIPS/W in 130 nm and ~46,000 DMIPS/W when scaled to 45 nm.

6. CONCLUSION

This paper presented the design of the Centip3De system. Centip3De utilizes the synergy that exists between 3D integration and near-threshold computing to create a reconfigurable system that provides both energy-efficient operation and techniques to address single-thread performance bottlenecks. The original Centip3De design is a 7-layer 3D stacked design with 128-cores and 256MB of DRAM. Silicon results were presented showing a 2-layer, 64-core system in 130 nm technology which achieved an

energy efficiency of 3930 DMIPS/W. At 46.4 M devices, Centip3De is one of the largest academic projects to date.

Acknowledgments

This work was funded and organized with the help of DARPA, Tezzaron, ARM, and the National Science Foundation.

References

1. ARM Cortex-A9. http://www.arm.com/products/processors/cortex-a/cortex-a9.php.
2. ARM Cortex-M3. http://www.arm.com/products/CPUs/ARM_Cortex-M3.html.
3. Dennard, R., Gaensslen, F., Rideout, V., Bassous, E., LeBlanc, A. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE J. Solid State Circ. 9*, 5 (Oct. 1974), 256–268,.
4. Dreslinkski, R.G., Zhai, B., Mudge, T., Blaauw, D., Sylvester, D. An energy efficient parallel architecture using near threshold operation. In *Proceedings of the Parallel Architecture and Compilation Techniques* (2007).
5. Fick, D., Dreslinski, R., Giridhar, B., Kim, G., Seo, S., Fojtik, M., Satpathy, S., Lee, Y., Kim, D., Liu, N., et al. Centip3de: A 3930 dmips/w configurable near-threshold 3d stacked system with 64 arm cortex-m3 cores. In *ISSCC 2012* (2012).
6. Hanson, S., et al. Ultralow-voltage, minimum-energy CMOS. *IBM J. Res. Dev. 50*, 4/5 (Jul/Sep 2006), 469–490.
7. Kim, T.H., Persaud, R., Kim, C. Silicon odometer: an on-chip reliability monitor for measuring frequency degradation of digital circuits. *IEEE J. Solid State Circ. 43*, 4 (Apr. 2008), 874–880.
8. Knickerbocker, J.U., et al. Three-dimensional silicon integration. *IBM J. Res. Dev. 52*, 6 (Nov. 2008), 553–569.
9. Loh, G. 3d-stacked memory architectures for multi-core processors. In *ACM SIGARCH Computer Architecture News* (2008), volume 36. IEEE Computer Society, 453–464.
10. Qazi, M., Stawiasz, K., Chang, L., Chandrakasan, A. A 512 kb 8t sram macro operating down to 0.57 μv with an ac-coupled sense amplifier and embedded data-retention-voltage sensor in 45 nm soi cmos. *IEEE J. Solid State Circ. 46*, 1 (Jan. 2011), 85–96.
11. Rusu, S., et al. A 45 nm 8-Core Enterprise Xeon Processor. In *Proceedings of ISSCC* (Feb. 2009).
12. Soeleman, H., Roy, K. Ultra-low power digital subthreshold logic circuits. In *Proceedings of the 1999 International Symposium on Low Power Electronics and Design* (1999), 94–96.
13. Swanson, R., Meindl, J. Ion-implanted complementary mos transistors in low-voltage circuits. *IEEE J. Solid State Circ. 7*, 2 (Apr. 1972), 146–153.
14. Tezzaron Semiconductor FaStack® Technology. http://tezzaron.com/technology/FaStack.htm.
15 Tezzaron Semiconductor Octopus DRAM. http://www.tezzaron.com/memory/Octopus.html.
16. Vittoz, E., Fellrath, J. Cmos analog integrated circuits based on weak inversion operations. *IEEE J. Solid State Circ. 12*, 3 (Jun. 1977), 224–231.
17. Wang, A., Chandrakasan, A. A 180 mv fft processor using subthreshold circuit techniques. In *IEEE International Solid-State Circuits Conference 2004, ISSCC 2004. Digest of Technical Papers* (15–19 Feb. 2004), volume 1, 292–529,.
18. Zhai, B., Blaauw, D., Sylvester, D., Hanson, S. A sub-200 mv 6t sram in 0.13 μm cmos. In *IEEE International Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers* (Feb. 2007), 332–606.
19. Zhai, B., Nazhandali, L., Olson, J., Reeves, A., Minuth, M., Helfand, R., Pant, S., Blaauw, D., Austin, T. A 2.60pj/inst subthreshold sensor processor for optimal energy efficiency. In *IEEE VLSI Technology and Circuits* (2006).

Ronald G. Dreslinski, David Fick, Bharan Giridhar, Gyouho Kim, Sangwon Seo, Matthew Fojtik, Sudhir Satpathy, Yoonmyung Lee, Daeyeon Kim, Nurrachman Liu, Michael Wieckowski, Gregory Chen, Dennis Sylvester, David Blaauw, and Trevor Mudge

(rdreslin, dfick, bharan, coolkgh, swseo, mfojtik, sudhirks, sori, daeyeonk, rachliu, wieckows, grgkchen, dennis, blaauw, tnm}@eecs.umich.edu), University of Michigan, Ann Arbor, MI.