CENTIP3DE: A 64-CORE, 3D STACKED NEAR-THRESHOLD SYSTEM

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
Trevor Mudge
University of Michigan

CENTIP3DE USES THE SYNERGY 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 SEVEN-LAYER 3D STACKED DESIGN WITH 128 CORES AND 256 MBYTES OF DRAM. SILICON RESULTS SHOW A TWO-LAYER, 64-CORE SYSTEM IN 130-NM TECHNOLOGY, WHICH ACHIEVED AN ENERGY EFFICIENCY OF 3,930 DMIPS/W.

High-performance microprocessors contain billions of devices per chip.[1] Meanwhile, global interconnect hasn't kept pace, because global wires scale in only one dimension instead of two, resulting in fewer, high-resistance routing tracks. Because of this problem and increasing design complexity, the industry is moving toward system-on-chip (SoC) designs or chip multiprocessors (CMPs). In these designs, multiple 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 application-specific integrated circuits (ASICs).

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). Whereas global interconnect can be millimeters long, silicon layers tend to be only tens of microns thick in 3D stacked processes, which enables substantial power and performance gains. Additional benefits include the ability to mix different process technologies (for example, CMOS, bipolar, DRAM, Flash, and optoelectronics) within the same die, and increased yield through "known good die" testing for each layer before integration.[2]

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, the University of Michigan had the chance to explore architectures that leverage the high-bandwidth, low-latency interface to DRAM afforded by 3D integration. This article discusses the resulting design and test of Centip3De (pronounced "centipede"), a large-scale CMP that we recently presented at HotChips 24 and the 2012 International Solid-State Circuits Conference (see Figure 1).[3] 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 256 Mbytes of integrated DRAM. We present silicon measurements for a 64-core version of the design, which show that Centip3De

Published by the IEEE Computer Society

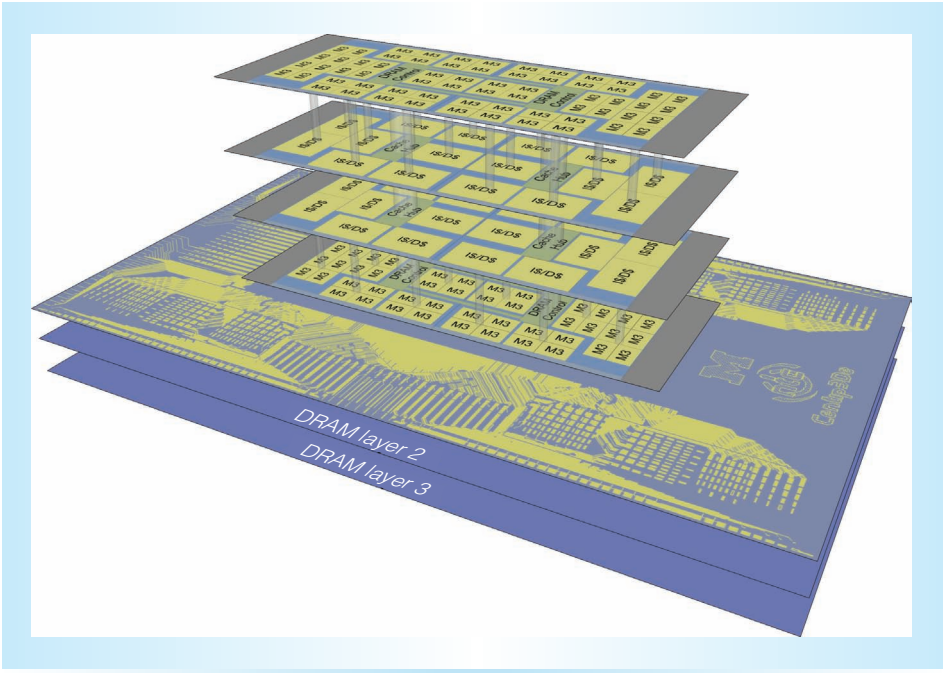

Figure 1. Floorplan-accurate artistic rendering. Centip3De has seven layers, including two core layers, two cache layers, and three DRAM layers.

achieves a robust design that provides energy-efficient performance (3,930 DMIPS/W) for parallel applications while still providing mechanisms to achieve substantial single-thread performance (100 DMIPS) for serial phases of code.

The synergy of 3D integration and near-threshold computing

One of the biggest drawbacks to 3D integration is the thermal density problem. By stacking dies on top of each other, local hotspots could affect neighboring dies, and hotspots that occur in the same place on multiple dies increase the demand on the heat sink.[4] Compounding this problem is the fact that supply-voltage scaling has stagnated at newer technology nodes, meaning that Dennard scaling theory breaks down and thermal density in 2D chips is increasing.[5] Exacerbating these issues is the fact that these hotter logic dies are now stacked on top of DRAM, which will require more-frequent refreshes.

Recent research into the area of near-threshold computing (NTC) has shown that running systems at supply voltages close to the threshold voltage can significantly improve energy efficiency and avoid thermal-density complications.[6] However, the reduced supply voltage results in slower performance.

Thankfully, these two technologies—NTC and 3D—have mutual benefits. By using NTC, the thermal density is much smaller, allowing more dies to be stacked together as well as on top of DRAM. By using 3D integration, additional cores can be added to provide more performance for parallel applications, helping to overcome the frequency degradation experienced with NTC.

Centip3De architecture

Centip3De is a large-scale CMP containing 128 ARM Cortex-M3 cores and 256 Mbytes of integrated DRAM. The system is organized into 32 computation clusters, each containing four ARM Cortex-M3 (http://www.arm.com/products/processors/cortex-m/cortex-m3.php) cores. Each cluster shares a four-way, 8-Kbyte data cache; a four-way, 1-Kbyte instruction cache; and local clock generators and synchronizers. The caches connect through a 128-bit, eight-bus architecture to the 256-Mbyte

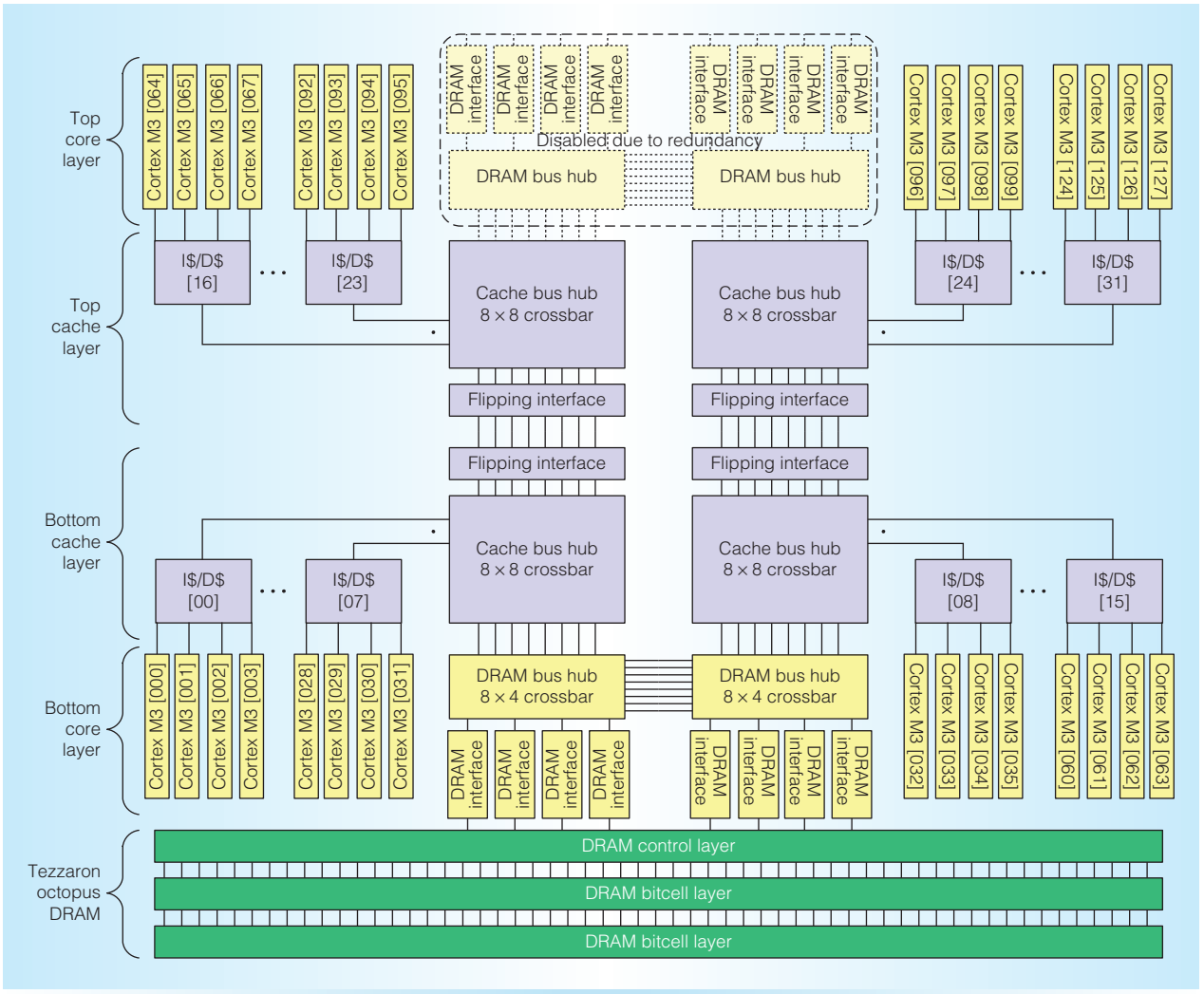

Figure 2. High-level system architecture. Centip3De is organized into four-core clusters that connect to eight DRAM buses. The diagram is organized by layer.

Tezzaron Octopus DRAM (http://www.tezzaron.com/memory/Octopus.html), which is divided into eight banks, each with its own bus, memory interface, and arbiters. Figure 2 shows a block diagram for the architecture, with the blocks organized by layer. Centip3De also has an extensive clock architecture to facilitate voltage scaling.

Designing for near-threshold: A clustered approach

The Centip3De design seeks to leverage some early findings from NTC operation to provide better single-thread performance through voltage boosting—raising the core's voltage and frequency. Owing to the differing activity factors for SRAM and logic, the optimal operating points for caches and cores are different. Zhai et al. showed that SRAM performs better at slightly higher voltages and speeds.[7] This leads to an architecture where the cache is run at a faster rate than the cores, and multiple cores share an L1 cache. In Centip3De, the cache is run at 4× the cores' frequency. The cache is then time-multiplexed to provide each core with a response in a single cycle. This allows cores to communicate shared data within the L1 cache without having to snoop on the bus, thus reducing energy consumption. In most cases, this reduction overcomes

any increase caused by data thrashing by cores sharing the same L1 cache.

The clustered cache also allows the system to be boosted with less impact than a traditional design. Therefore, we implemented a clustered NTC boosting architecture on the basis of the following key observations. First, the cache doesn't need to change frequency, only the cores. The system simply disables three cores and speeds up the remaining core to match the cache's frequency. Because the entire cache space is still visible, miss rates for the remaining core are reduced further than with the traditional approach. Second, the larger cache space also helps hide the increased latency, in relative cycles of the faster core, to main memory. This new architecture will allow for better cache performance. If greater performance is needed, both the core and cache can be boosted to an even higher frequency.

Detailing the clustered cache

On the surface, the cache design seems simple. However, we required a more sophisticated design to meet the ARM Cortex-M3's strict timing budgets. Figure 3 shows the state diagram for the cache and the internal cache pipeline. In the final Centip3De design, the cache operates in two different ways, depending on the number of cores involved. When the cache is in three- or four-core mode, it is pipelined into four stages. In the first stage, the tags are read. In the second stage, a tag comparison is done. Finally (only if a hit is detected) the data is read out of the correct way. This improves energy efficiency by accessing only one way of the data cache. The cores are operated with clocks that are 90 degrees out of phase to allow each core to get a result in each core clock cycle. Figure 4 shows the pipeline for the four- or three-core mode.

When the cluster operates in one- or two-core mode, the cache is repipelined to return the data in a single cycle. In order to achieve single-cycle latency, the tags and data arrays must be read in parallel, and on the second cycle a comparison returns the correct data. This results in more energy consumption, as all data ways must be accessed. Through synthesis, we determined that cycle-stealing from the core is possible, as shown in the pipeline diagram in Figure 3. The cache first captures the access from each core three-quarters of the way through the core cycle and returns the result halfway through the next core cycle.

Measured results

Silicon for Centip3De is coming back in multiple stages. We have already received and tested a two-layer system with 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. Figure 5 shows a die micrograph, and Table 1 lists the values of key design components.

For testing equipment, we use LabVIEW (Laboratory Virtual Instrumentation Engineering Workbench) with a National Instruments PCIe-6535 board and PCI-6723 board. The PCIe-6535 provides 32 digital I/O pins, which we use to control the Joint Test Action Group and scan chain ports on the chip. The PCI-6723 provides analog reference voltages used to tune the phase-locked loop (PLL). Other lab bench equipment includes power supplies, multimeters, and a waveform generator to generate a reference clock for the PLL.

Figure 6 shows silicon measurements and interpolated data for different cluster configurations from running a Dhrystone test on the fabricated two-layer system. The default NTC cluster (slow/slow) configuration operates with 4 cores at 10 MHz and caches at 40 MHz, achieving 3,930 DMIPS/W. Based on fan-out of 4 (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. The system bus operates at 160 to 320 MHz, which supplies an ample 2.23 to 4.46 Gbytes per second (GBps) memory bandwidth. The latency of the bus ranges from one core cycle latency for 10 MHz cores to six cycles

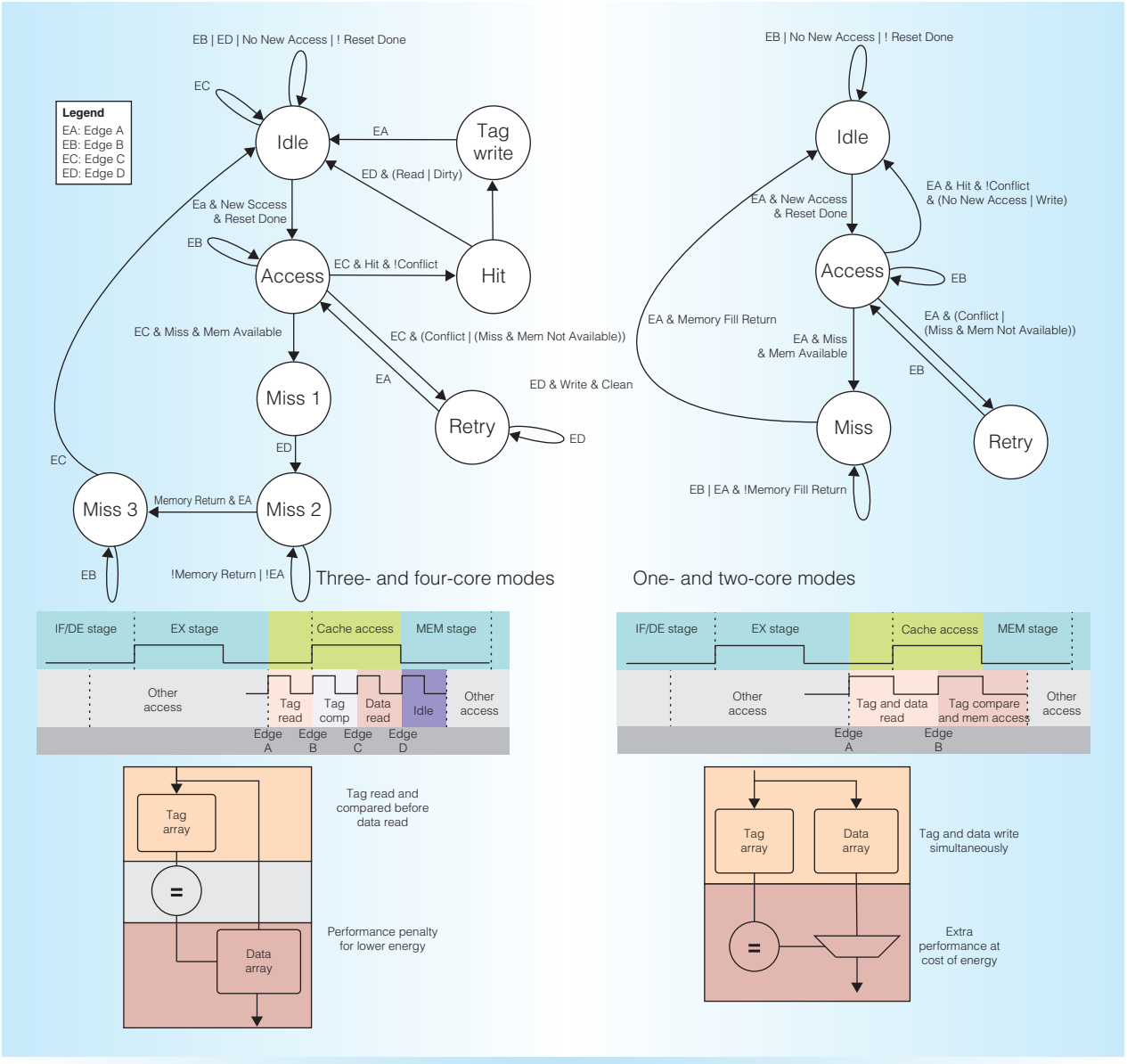

Figure 3. Cache state machines and pipeline diagrams. Four cache modes are supported. In three- and four-core modes, the high latency of the processors is used to improve efficiency by accessing tag and data arrays sequentially. Knowing the tag check results reduces the number of accessed data arrays.

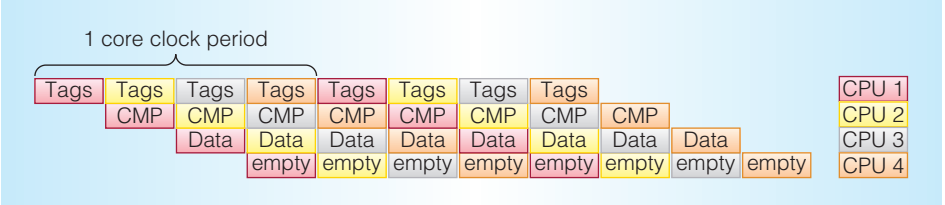

Figure 4. Cache pipelining for four-core mode. Each core is clocked 90 degrees out of phase, so in each cycle one core is reading the tags, one core is doing a comparison on the previous tags, and one core is reading the data array. Because the core operates at one-quarter the cache speed, an entire access finishes in one core cycle.

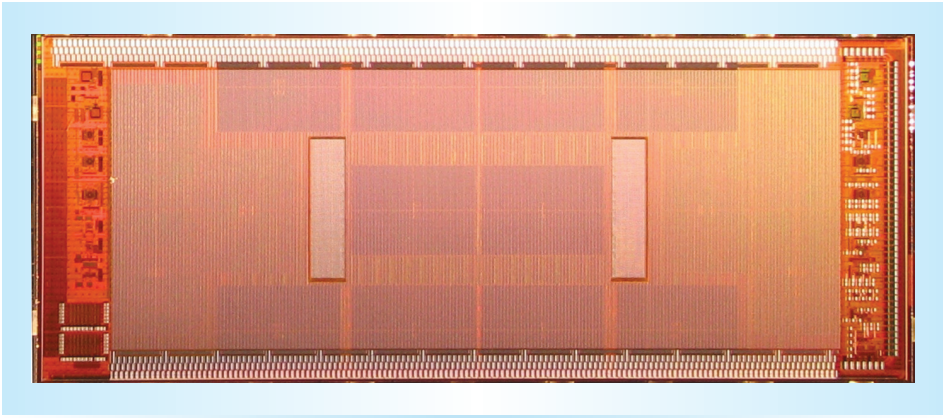

Figure 5. Die micrograph of the two-layer system. The two layers are bonded face-to-face. The clusters can be seen through the backside of the core-layer silicon. Wirebonding pads line the top and bottom edges.

when cores are boosted to 80 MHz. As a point of comparison, the ARM Cortex-A9 (http://www.arm.com/products/processors/cortex-a/cortex-a9.php) in a 40-nm process can achieve 8,000 DMIPS/W. At peak system efficiency, Centip3De achieves 3,930 DMIPS/W in a much older 130-nm process.

The results in Figure 6 present many operating points to choose from on the basis of workload or workload-phase characteristics. Selecting 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. Overall, the Centip3De design offers programmers a wide range of power, throughput, and single-thread performance points at which to operate.

The next system we expect has four layers—two core layers and two cache layers. This system is created 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 the cache sides back-to-back. One core layer is then thinned to expose TSVs, and aluminum wirebonding pads are added. The final system will comprise seven layers, including two core layers, two cache layers, and three DRAM layers. The wirebonding sites for this system will be on the DRAM, which has much larger layers than the core and cache layers.

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

Component	Value
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 and 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

*B2B: back-to-back; F2F: face-to-face; TSV: through-silicon via.

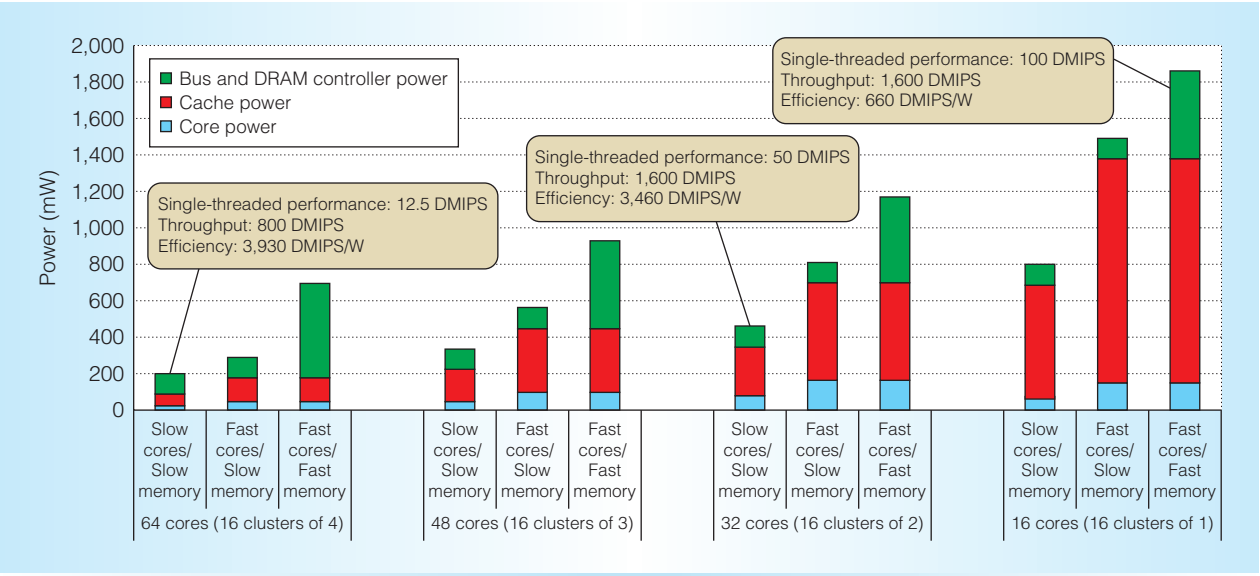

Figure 6. Power analysis of the 64-core system. Power breakdowns for four-, three-, two-, and one-core modes. Each mode has a slow (near-threshold computing) core option as well as a fast (boosted) option with increased frequency and voltage. Each option provides a tradeoff in the efficiency, throughput, and single-thread performance space. Overall, Centip3De achieves the best energy efficiency, at 3,930 DMIPS/W.

The landscape of 3D stacking technologies is diverse, providing a wide range of possibilities. Centip3De used the Tezzaron 3D process, which is an aggressive technology node providing small TSVs at 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, which can result in yield issues for large runs. Alternative TSV technologies rely on microbump die-to-die bonding, which resolves some of the yield issues but has a much larger TSV pitch. 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. At the far end of the 3D spectrum is 2.5D technology, which relies on silicon interposers. This technique takes the microbump 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 microbump TSV densities.

In designing Centip3De, we learned several lessons. First, supporting 3D Layout Versus Schematic and Design Rule Check checking was a challenge that required a large amount of script development time. The good news is that in the two years since we first started work on 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 small. Spice simulations indicate that a significant amount of power is being used in these delay generators, particularly in NTC mode. Unfortunately, we were unable to subtract the unnecessary power these delay-generators consume, because they weren't 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. MICRO

References

1. S. Rusu et al., ''A 45 nm 8-Core Enterprise Xeon Processor,'' *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE, 2009, pp. 9-12.
2. J.U. Knickerbocker et al., ''Three-Dimensional Silicon Integration,'' *IBM J. Research and Development,* Nov. 2008, pp. 553-569.
3. D. Fick et al., ''Centip3De: A 3930 DMIPS/W Configurable Near-Threshold 3D Stacked System with 64 ARM Cortex-M3 Cores,'' *Proc. IEEE Int'l Solid-State Circuits Conf.,* IEEE, 2012, pp. 190-192.
4. G. Loh, ''3D-Stacked Memory Architectures for Multi-Core Processors,'' *ACM SIGARCH Computer Arch. News,* vol. 36, 2008, pp. 453-464.
5. R. Dennard et al., ''Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions,'' *IEEE J. Solid-State Circuits,* Oct. 1974, pp. 256-268.
6. R. Dreslinski et al., ''Near-Threshold Computing: Reclaiming Moore's Law Through Energy-Efficient Integrated Circuits,'' *Proc. IEEE,* Feb. 2010, pp. 253-266.
7. B. Zhai et al., ''Energy-Efficient Near-Threshold Chip Multi-Processing,'' *Proc. ACM/IEEE Int'l Symp. Low Power Electronics and Design* (ISLPED), ACM, 2007, pp. 32-37.

Ronald G. Dreslinski is a research scientist at the University of Michigan. His research focuses on architectures that enable emerging low-power circuit techniques. Dreslinski has a PhD in computer science and engineering from the University of Michigan.

David Fick is a research associate at the Michigan Integrated Circuits Lab at the University of Michigan. His research interests include fault tolerance, adaptive circuits and systems, and 3D integrated circuits. Fick has a PhD in computer science and engineering from the University of Michigan.

Bharan Giridhar is a PhD candidate in the Department of Electrical Engineering at the University of Michigan. His research interests include adaptive computing, robust synchronizer design, and circuit techniques for high performance and low power. Giridhar has an MS in electrical engineering from the University of Michigan.

Gyouho Kim is a PhD candidate in the Department of Electrical Engineering at the University of Michigan. His research interests include ultra low-power and energy-efficient VLSI design. Kim has an MS in electrical engineering from the University of Michigan.

Sangwon Seo is a senior hardware engineer at Qualcomm. His research interests include low-power microarchitecture, parallel computing, and wireless signal processing. Seo has a PhD in electrical engineering and computer science from the University of Michigan, where he performed the work for this article.

Matthew Fojtik is a PhD candidate in the Department of Electrical Engineering at the University of Michigan. His research interests include architecture-independent timing-speculation techniques and the design of several ultra low-power VLSI systems. Fojtik has an MS in electrical engineering from the University of Michigan.

Sudhir Satpathy is a research scientist at Intel's Circuits Research Labs, where he is developing novel high-performance and low-power circuit techniques for next-generation microprocessors. His research interests include on-die interconnect fabrics and hardware security. Satpathy has a PhD in electrical engineering from the University of Michigan, where he performed the work for this article.

Yoonmyung Lee is a research fellow at the University of Michigan, where he researches energy-efficient ultra low-power integrated circuits for low-power, high-performance VLSI systems and millimeter-scale wireless sensor systems. Lee has a PhD in electrical engineering from the University of Michigan.

Daeyeon Kim is a research scientist in Intel's Advanced Design Group, where he works on low-power, high-performance SRAM design, statistical yield analysis, and process variation

mitigation techniques. Kim has a PhD in electrical engineering from the University of Michigan, where he performed the work for this article.

Nurrachman Liu was most recently at Qualcomm Research. His research interests include high-performance digital and custom circuit design, ultra low-power sub- and near-threshold circuit design, VLSI design automation, and EDA for circuit designers and digital design. Liu has a PhD in electrical engineering from the University of Michigan, where he performed the work for this article.

Michael Wieckowski is a senior research engineer at Mala Geoscience, where he designs high-speed mixed-signal circuits for applications in ground-penetrating radar. Wieckowski has a PhD in electrical and computer engineering from the University of Rochester. He performed the work for this article as a research fellow at the University of Michigan

Gregory Chen is a member of Intel's High-Performance Circuits research group. His research interests include networks-on-chip, security, and power management. Chen has a PhD in electrical engineering from the University of Michigan, where he performed the work for this article.

Dennis Sylvester is a professor in the Electrical Engineering and Computer Science Departments at the University of Michigan and is the director of the Michigan Integrated Circuits Laboratory. His research interests include the design of millimeter-scale computing systems and energy-efficient near-threshold computing for a range of applications. Sylvester has a PhD in electrical engineering from the University of California, Berkeley.

David Blaauw is a professor in the Electrical Engineering and Computer Science Departments at the University of Michigan. His research focuses on VLSI design, particularly on ultra low-power and high-performance design. Blaauw has a PhD in computer science from the University of Illinois at Urbana-Champaign. He is an IEEE Fellow.

Trevor Mudge is the Bredt Professor of Engineering at the University of Michigan. His research interests include computer architecture, programming languages, VLSI design, and computer vision. Mudge has a PhD in computer science from the University of Illinois. He is a fellow of IEEE and a member of the ACM, the IET, and the British Computer Society.

Direct questions and comments about this article to Ronald Dreslinski, University of Michigan, 2260 Hayword Ave., Ann Arbor, MI 48109; rdreslin@umich.edu.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.