Centip3De: A Cluster-Based NTC Architecture with 64 ARM Cortex-M3 Cores in 3D Stacked 130nm CMOS

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

Author Contact: David Fick, 1301 Beal Ave, Ann Arbor, MI 48104, (734) 915-4091, dfick@umich.edu

Abstract

We present Centip3De, a large-scale 3D CMP with a cluster-based near-threshold computing (NTC) architecture. Centip3De uses a 3D stacking technology in conjunction with 130nm CMOS. Measured results for a two-layer, 64-core system are discussed, with the system achieving 3930 DMIPS/W energy efficiency, which is >3x improvement over traditional operation at full supply voltage. This project demonstrates the feasibility of large-scale 3D design, a synergy between 3D and NTC architectures, a unique cluster-based NTC cache design, and how to maximize performance in a thermally-constrained design.

I. Introduction

Process scaling has resulted in exponential growth of the number of transistors available to designers, with high-performance designs now containing billions of devices per chip [1]. However, with the stagnation of supply voltage scaling in advanced technology nodes, power dissipation has become a limiting factor in high-performance processor design. As a result, designers have moved away from a single, high-complexity, super-scalar processor and instead have opted for multiple, simpler, higher energy efficiency cores. In these system-on-chip (SoC) or chip-multiprocessor (CMP) designs, many components share the same chip. These systems typically include processing cores, memory controllers, video decoders, and other ASICs [2], [3].

Due to the large number of individual components in an SoC or CMP, the interconnect network between components has become critical to these systems. However, global interconnect has not scaled nearly as well as transistor count since global wires scale in only one dimension instead of two, resulting in fewer, high resistance routing tracks.

Three-dimensional (3D) integration seeks to address the global interconnect scaling issue by adding multiple layers of stacked silicon with vertical interconnect between them, typically in the form of through-silicon vias (TSVs). Since global interconnect can be millimeters long, and silicon layers tend to be only tens of microns thick in 3D stacked processes, the power and delay reductions by using vertical interconnect can be substantial, often 30-50% [4].

Additional benefits of using 3D integration 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" techniques, where each layer is tested before integration [5]. Integrated DRAM in particular has shown significant performance improvements [6], [7]. Recently, several industrial and academic 3D systems have been demonstrated [8]–[12].

Heat dissipation is a salient issue with 3D integration. High performance designs reached

a maximum practical thermal design power (TDP) years ago. Since then, power density has been increasing *further* due to non-ideal process scaling [13], which is exacerbated by having multiple layers of silicon. In this work, we propose using near-threshold computing (NTC) in 3D design to address these issues [14], [15]. We show how NTC has a unique synergy with 3D design and propose a new clustered cache architecture that exploits the unique properties of NTC design. We demonstrate the proposed approaches in a 64-core 3D CMP design and present silicon measurements.

In previous work, subthreshold computing has been widely used for maximum energy efficiency. In this realm, the supply voltage is reduced below the threshold voltage of the devices down to V_{opt}, the optimal voltage that minimizes energy/cycle. By operating at V_{opt}, the leakage and dynamic power components become nearly balanced, maximizing energy efficiency (~12-16x greater than nominal operation). The cost of this improvement is that performance is reduced by ~1000x. This trade-off, however, is particularly suitable for environmental sensing applications [16], [17] and medical sensor applications [18], where low activity rates are needed.

NTC focuses on high performance applications. To accomplish this, the supply voltage is instead reduced from the wear-out limited nominal supply voltage to just *above* the threshold voltage of the technology (V_{NTC}), resulting in a ~60-80x power reduction [14], [15]. This reduction in power facilitates heavily 3D stacked designs. An associated ~10x performance loss is also incurred, resulting in ~6-8x total energy savings. This can be shown to be true across a range of processes, with measured data from a 180nm process and simulated data from a 32nm process shown in Figure 1. This loss of performance is more manageable, and can be recovered through parallelism and the performance benefits of 3D-design, resulting in both improved energy/operation and increased overall performance for a fixed TDP [19].

A key observation in NTC design is the relationship between activity factor and optimal energy point [19]. 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 2 illustrates this effect in a 32nm 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 trade-off points, the relative chosen NTC supply points should track with V_{opt}. Thus, V_{NTC_memory} should be chosen to be relatively higher than V_{NTC_core} for them to maintain the same energy↔delay trade-off point. This strategy does not conflict with other energy saving techniques such as "drowsy caches" [20]. But, since V_{opt} is lower than V_{min} (minimum functional voltage) for typical SRAM designs, a low-voltage memory design will have be used, such as an 8T design [21] or a specially designed 6T [22]. For this reason, 8T SRAMs have become popular in industry for lower level caches and register files [23], [24].

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, 4x-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 (described in detail in Section II-C) 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 can reconfigure its pipeline to access tag arrays and data arrays in parallel instead of serially and change from a 4x core↔cache frequency multiplier to a 2x multiplier. The remaining core(s) are voltage boosted and operate at 2x their

original frequency, roughly doubling their single-threaded performance. 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.

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 which known good die techniques can be coupled with to further increase yield. Centip3De's boosting modes in combination with "silicon odometer" techniques improve performance while maintaining lifetime [25], [26].

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 4x 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.

To demonstrate the proposed design concepts, we describe Centip3De [12], a large-scale 3D CMP with a cluster-based NTC architecture. Centip3De uses Tezzaron's 3D stacking technology in conjunction with Chartered 130nm process. Measured results for a two-layer, 16-cluster system are discussed. Section II starts with an overview of the architecture, then describes each of the components of the system in detail. Section III describes Tezzaron's 3D integration technologies and Section IV discusses how 3D integration factored into design reuse. Silicon results are presented in Section V, future expansions of the 3D stack are described in Section VI, and the paper concludes in Section VII.

II. System Architecture

Centip3De is a large-scale, 3D CMP containing clusters of ARM Cortex-M3 cores [27] designed with the NTC principles described in Section I. A system containing two layers bonded face-to-face (F2F) has been measured, where 16 clusters localize the cores to a single layer and the caches to another. Upper metal layers are used to create F2F connections, as described in Section III, while TSVs are used for off-chip communication. TSVs are also present for the future back-to-back (B2B) and face-to-back (F2B) stack expansions described in Section VI.

To determine the system architecture, analysis was performed using SPLASH2 benchmarks (Cholesky, FFT, FMM, LU, Radix, and Raytrace) on the gem5 simulator as performed in [19]. A 128b, eight bus architecture was chosen based on cache miss bandwidth requirements, floorplanning constraints, and to match data widths of future connect memories (Section VI). A cluster size of four cores per cluster, a 1kB instruction cache, and a 8kB data cache were chosen to maximize energy efficiency while working within our area constraints. In this study, four-core cluster systems were found to be 27% more energy efficient while providing 55% more throughput than one-core cluster systems for Centip3De's design constraints (technology parameters, available area, type of core, SRAM performance, and voltage scaling targets).

For multi-program support, each of the 64 cores runs with an independent stack pointer in a non-virtualized memory space with access to a core ID and other core-specific registers programmed by JTAG. Each cluster contains a cluster-level memory-mapped semaphore to support fine-grained data locking. With these tools, Centip3De can support completely separate programs, or a single program with many threads.

In addition to cores and caches, each cluster contains local clock generators and synchronizers. Figure 3 shows a block diagram for the architecture with the blocks organized by layer. Centip3De also has an extensive clock architecture to facilitate voltage scaling, which is discussed in detail in Section II-E.

A. Floorplanning

Floorplanning in 3D design addresses the main two classes of interconnect: signal routing and power delivery. Signal routing between layers can be through fine grained (transistor level) or coarser grained (bus level) vertical interconnect. Centip3De uses bus level connections for design simplicity. However, due to the low parasitics of vertical interconnect in Tezzaron's process (Section III), Centip3De places high-performance, routing-dense buses vertically, which in this system are between cores and caches, and within the bus architecture.

To improve power routability, the cores and caches were separated into different layers since they use separate power supplies. This partitioning eliminated one power class from each layer which reduced resistive droop by approximately 15-25% for the same resources. Localizing the caches to a single layer also simplified the bus design. Careful planning of bus hub ports eliminated cache↔bus routing congestion on that layer. The twin bus hub columns are bridged with eight buses on the core layer, which has little additional global routing.

By building the bus architecture vertically, required routing resources reduced by approximately 50% compared to a single-layer floorplan. Similar gains are obtained in energy and performance, and were not offset by 3D interconnect loading due to its relatively small overhead.

B. Processing Core

Centip3De contains 64 ARM Cortex-M3 cores, which have a 3-stage, in-order, single issue pipeline. In this 130nm design, the core operates between 10MHz at 650mV and 80MHz at 1.15V.

The core has a JTAG interface that allows access to status registers within the core, reset and halt capabilities, and access to the memory system. JTAG also provides access to the memory system, which includes memory mapped IO (MMIO). The JTAG data signals are daisy-chained between the four cores of each cluster, then between multiple clusters, as described in Section II-F.

MMIO registers control the stack pointer location, general purpose registers, a hard reset, and core clock adjustment controls. The clock adjustment controls allow the measurement and control of core↔cache clock skew that is further described in Section II-E. MMIO also allows access to a hardware based semaphore contained in the cache, which organizes memory operations within the cluster. Core 0 has additional MMIO registers (they exist vestigially in cores 1-3) to access cache mode, cache clock skew measurement and control, and cache clock multiplier controls. The core 0 MMIO has additional registers to control clock gating for the other three cores and multiplexers to add or remove those cores from the JTAG daisy chain.

The cluster floorplan for the core layer is shown in Figure 4. Each of the four cores is identical - the same layout is mirrored four times to achieve the cluster layout. In the center of the cluster is the clock delay generator and the skew measurement unit. Core 0 has 598 signals, while cores 1-3 each have 331 signals, for a total of 1591 connections to the cache. This results in 25,456 core↔cache vertical interconnects for the 64-core system. The vertical interconnections are visualized in Figure 4. Outputs from the core are differential to facilitate level conversion in the cache.

All core signals connect F2F to the cache on the adjacent layer, with no direct connections to other modules on the same layer (Figure 3). The core also contains a square grid of dummy TSVs on a 50um pitch to meet TSV density rules, and are visible in Figure 5. Since the cluster is rotated to many orientations (as seen in Figure 5), both grids are square (x and y dimension matched) and can rotate together on the relevant copper honeycomb interface patterns discussed in Section III.

C. Cluster Cache

The cluster cache contains a 1kB instruction cache, an 8kB data cache, clock generators, and hardware semaphores. The cache operates between 40 MHz at 800mV and 160 MHz at 1.65V. The cache can operate in high-efficiency four- and three-core modes or high-performance two-

and one-core modes. In three-/four-core modes, the cache operates 4x the frequency of the cores, and the enabled cores operate with separate clocks that are 90° out-of-phase. In one-/two-core modes, the relationships are 2x and 180°, respectively. The cache is pipelined such that in all modes the cores see a typical, single-cycle interface to the cache.

For robust low-voltage operation, the cache uses a custom 8T bitcell design with tunable pulse generators. A word length of 32b and a line length of 128b were chosen to match the core and bus datapath widths, respectively. The cache supports 4-way set association nominally and a direct mapped mode for debugging.

The three/four-core mode supports higher efficiency than the one-/two-core mode by first reading and checking the tag arrays, then accessing only the necessary data array(s). By doing this, at most one data array is read per hit, or all four data arrays for a miss with eviction. In contrast, each of the four data arrays are read for every access in one/two-core mode, in case there is a hit. When a miss occurs, the other cores are able to continue operation until a conflict occurs.

The cluster clock generator is shown in Figure 6. These two modes require different frequency and phase relationships with the core clocks. These clocks are generated locally based on configuration bits from core 0 and synchronized with the bus clock with a tunable delay buffer. Before generating the core clocks, the cache clock is first divided by a factor between one and eight. The clocks transition glitch-free during mode changes to prevent core logic corruption, which is particularly important for core 0. The individual core clocks can also be gated by configuration bits from core 0 to reduce energy consumption in unneeded cores.

To assist in multi-core programming, hardware-based semaphores are included in the cache to provide a read-modify-write operation to a select number of addresses. Inputs from the core are differential and are level converted upon entry into the cache. Similarly, outputs to the bus are also differential.

The floorplan in the cache is shown in Figure 4. The F2F 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%. These benefits were not offset by the loading requirements of the 3D interface, since the parasitic capacitance and resistance of the F2F connections are small.

A square grid of dummy TSVs exists on a 50um pitch and is aligned with the TSV grid of the cores such that that the cluster design may be rotated. Within the SRAM arrays, some dummy TSVs are safely forgone while others are aligned with pre-existing power routing. The dummy TSV impact on array utilization is less than 1.5%.

D. Bus Architecture

The bus architecture includes eight independent, 128b buses that operate between 160 MHz at 1.05V and 320 MHz at 1.6V (all at the same frequency). The eight buses represent independent address spaces and service cache misses from all sixteen clusters. A round-robin arbiter determines the order of priority of the requests (Figure 3). The buses are each physically split into two columns that span the cache and core layers of the chip.

Each bus can service a single request at a time, which takes six to fifteen bus cycles. Crossing from one column to the other induces a single-cycle penalty each way. For cores operating in higher efficiency modes, a cache miss and resulting memory request incur as little as a single-cycle penalty for the core. In the highest performance core modes a cache miss becomes a four-cycle minimum penalty.

The bus arbiter contains configuration bits to mask out faulty clusters, although this proved to be unnecessary in our testing. Configuration bits also control which three bits of the memory address select the bus, with options being either the top three or bottom three. These settings allow Centip3De to either distribute traffic to all of the memory buses, or localize traffic to

particular buses.

The two communication columns are bridged on the core layer with eight buses, while the routing to the clusters is on the cache layer. Routing lanes exist between the clusters on both layers to facilitate this routing. The main portion of the logic exists in two bus hubs on each layer. Vertical interconnect in the bus hub modules alleviates routing congestion within the module, reducing the footprint and making additional perimeter space accessible for global routing.

E. Clock architecture

Centip3De is designed so that cores, caches, and the main bus operate on different power domains with level converters included between these power domains. Since each power domain can be scaled individually depending on the desired power and performance target, inter-clock-domain skew becomes an issue. For example, if a cluster that runs at half the frequency of the main bus is re-tuned to run at a quarter of the frequency, then the delay through its clock tree will change, resulting in timing violations when crossing clock domain boundaries.

To address inter-clock-domain skew caused by voltage scaling, each clock domain has a delay generator and each relevant pair of clock domains has a skew detector connected at the leaves. This method also simplified clock tree matching between separately designed modules. By tuning the clock tree once, the settings can be saved and restored, even for different dies, although re-tuning a particular die is possible if needed.

As shown in Figure 7, an initial clock is generated by a phase-lock loop (PLL), which locks to an off-chip clock reference, or can be controlled directly (no-feedback) via an off-chip reference voltage. This clock is used to generate three global clocks that can be aligned using two phase comparators. Each computation cluster has its own clock generator for cache and core clocks. The core clocks are phase compared to the cache clock, which is also phase compared to the bus clock.

The phase comparator design is similar to simple PLL phase comparators. A flip flop is used

to compare the phases of two clock trees by connecting the slower and faster clocks to the flip flop clock and data pins, respectively. The phase result progresses through two additional flip flops to protect against metastability. In lieu of an RC filter, an up/down counter counts for a preset number of cycles (2048). The phase comparators are controlled via scan chain at the system level or JTAG at the cluster level.

F. Offchip I/O and Power

Aluminum wire bonding pads were added to the core backside for packaging. These structures are visible in the periphery of the die in Figure 5.

Due to floorplanning constraints, the digital and analog pads were moved to the center of the floorplan on the cache layer (Figure 5). On the core layer, these areas are used to route bridging buses between the two communication columns. Routing from these pads to the chip periphery used extra-wide wiring for robustness.

In a trade-off between testing time and number of IO pads, the JTAG data signals were daisy-chained through the four cores of a cluster and two clusters on the same cache layer. In this way, there are eight JTAG data interfaces, each connected to eight cores, which facilitates the parallel loading of data. A scan chain is included to control system-level configuration settings, such as clock delay chain settings and phase generator controls [28].

III. TEZZARON'S 3D TECHNOLOGY

Tezzaron's FaStack® technology stacks wafers of silicon (as opposed to individual dies) using copper bonding [29]. Before each wafer pair is bonded, a layer of copper is deposited in a regular honeycomb pattern that is then thinned. Due to the homogeneity of the pattern, it is very flat after thinning, and when two wafers with this pattern are pressed together with heat, the copper bonds strongly. 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 patterns. The

TSVs are made of Tungsten, due to preferable thermal-mechanical properties, and are created after the transistors, but before the lowest metal layers, preventing the loss of any metal tracks. Since their diameter is small (1.2um), their length is short (around six microns), and they're only coupling 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 together 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 [30] 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. The wafer of larger dies is also thinned and coated with copper. The smaller dies are put in a tray to hold them in place and the tray of smaller dies and the larger wafer are pressed together to finish the bond. A larger copper pattern is used to support less-precise alignment, using 27 TSVs for each connection. This process would not be needed if the two designs had the same physical dimensions.

IV. Design Reuse

Design reuse in 3D ICs has challenges at both module and layer levels. At the module level, we reduced design effort for the core and cache designs by tiling instead of recreating each core and cache instance. This simplified verification, DRC, and LVS, however placing vertical interconnect in a design that may be rotated or flipped proved challenging. TSVs and F2F interconnects both connect to a pre-set honeycomb pattern. These designs could only use TSV and F2F interconnect locations that are rotationally symmetric around the origin of the module, relative to the honeycomb pattern. For the F2F interfaces, this created a 5um square grid of interconnect locations that could be used. For the B2B interface, however, this was a 50um grid. This restriction only applied to modules that needed to be flipped and rotated, which included

the clusters only. The B2B interface in the cluster was needed for placing filler TSVs only.

At layer level a modification was created to expand the system from two logic layers to four. Since the two-layer logic stack is symmetric across the Y-axis, it is possible to combine two completed two-layer stacks such that their bus interfaces align. The core-side of the two-layer stack contained the DRAM interface, so the cache-side was used to expand the logic stack with B2B TSVs. This design strategy requires only one additional mask for fabrication, the copper pattern for the B2B interface. To complete the expansion, a flipping interface and side-detector were added, as discussed in Section VI-A.

V. Measured Results

A two-layer system was fabricated with results provided. Technology data and system information are shown in Table I. A die micrograph is shown in Figure 5. In this system a core layer and a cache layer were bonded face to face. The backside of the core layer was then ground down, exposing TSVs for the DRAM interface and for offchip I/O. A layer of aluminum was then applied to create wirebonding pads for the TSVs and to cover exposed DRAM and dummy TSVs. The chip was then wirebonded in a 256-pin, ceramic PGA package, and tested using a custom-designed PCB and LabVIEW.

A measured frequency and energy profile of the core is included in Figure 9. Although Centip3De can support a wide variety of voltages, frequencies, and modes for each cluster, four configurations in particular were used in this analysis, as detailed in Table II. The modes were chosen to illustrate a wide range of operations, with maximum energy efficiency at low voltage in four-core mode, and maximum boosting performance at high voltage in one-core mode.

The test program was written in C, compiled into the ARM Thumb ISA, and loaded into the cluster via JTAG. It contains a number of arithmetic operations based on an example power virus program provided by ARM and also uses the hardware semaphores to ensure that all four

cores start the main portion of the program simultaneously. It is designed to stress the caches, but also fit within them since the DRAM is currently unavailable for use. The direct-mapped cache mode facilitates operation without DRAM.

For each performance point the core voltage is adjusted first. Each core voltage requires a different clock tree alignment setting, which was applied using JTAG controls. After a particular operating voltage for the core is set, the minimum necessary cache voltage is determined. After the cache voltage is adjusted, the cache clock is first re-aligned to the bus clock and then the core clock is re-aligned to the cache clock. During clock phase alignment, the phase comparator counter typically returned one of the two extreme values, meaning that the clock jitter was significantly smaller than the delay generators FO1 delay increment, and a three or four bit counter would work as well as eleven bits. The phase generator and phase comparators were always able to align the clock trees across a variety of voltage scaling scenarios. In a commercial application, the phase information would be stored in a look-up table off-chip and would be managed by the system BIOS. The operating system would make advanced configuration and power interface (ACPI) requests for particular power states, and the BIOS would provide the voltage and configuration definitions of these states.

The highest efficiency four-core mode operates with cores at 10MHz and caches at 40MHz, achieving 8800 DMIPS/W (Table II). Latency critical threads can operate in the boosted mode at up-to 8x higher frequency. The boosted one-core mode operates the core at 80MHz and the cache at 160MHz. The difference in energy efficiency and single-threaded performance between four-core and one-core modes is 7.6x and 8x, respectively.

Boosting clusters within a fixed TDP environment may require disabling or down-boosting other clusters to compensate for that cluster's increase in power consumption. A package with a 250mW TDP can support all sixteen 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, 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 250mW. On the left are high-efficiency configurations, with more aggressively boosted configurations on the right, which provide single-threaded performance when needed.

A wider system-level analysis is performed in Table III. A variable number of clusters are boosted from four-core mode to other modes, providing a variety of trade-offs between single-threaded performance and energy efficiency. Four system-level modes are analyzed in detail in Figure 10, with power breakdowns, voltages, single threaded performance, energy efficiency, and throughput visualized. An ARM Cortex-A9 in a 40nm process is able to achieve 8000 DMIPS/W [31]. At peak system efficiency Centip3de achieves 3930 DMIPS/W.

VI. Future Directions

This paper describes a measured two-layer, 64-core system, while the final system will include up to seven layers, 128-cores, and 256MB of DRAM, via already existing cross-layer interfaces. These include a flipping interface to allow the addition of another core and cache layer pair, and a DRAM interface, which allow the addition of three layers of DRAM as shown in Figure 8. These additional layers use back-to-back (B2B), and face-to-back (F2B) bonds in addition to the F2F bonds used in the measured system.

A. Flipping Interface

The seven-layer system includes duplicated core and cache layers. These layers are designed to be reused, *i.e.*, they are identical copies. To accomplish this, a flipping interface is designed into the bus hub on the cache layer, as seen in Figure 3. The two bus hubs are placed equidistant

from the center of the chip so that when two cache layers are aligned B2B, the bus hubs also align. The flipping interface in the bus hubs includes pairs of tri-state buses.

The direction of each of these buses is determined by detecting which side of the B2B interface is connected to the DRAM and/or wirebonding pads. A single internally pulled-down IO pad is used to make this determination. The side-detector connected to the outside world will be wirebonded to V_{DD}, whereas the other will automatically pull-down to ground instead. By doing this, we can safely negotiate the directions of the tri-state buses. The side-detector also disables redundant units such as the unconnected DRAM interfaces and unnecessary bus hubs, labeled at the top of Figure 3.

B. DRAM Interface

In the seven-layer system, there will be 256MB of Octopus DRAM organized into eight banks, each with its own DRAM interface [30]. The DRAM interfaces are similar to DDR2 and use double edge clocking for data transfer to the DRAM. They operate between 160 MHz at 1.05V and 320 MHz at 1.6V, on the same frequency and voltage domain as the bus architecture. They also use a double frequency clock at 320/640MHz to facilitate DDR operation. With the bus architecture, the DRAM interfaces provide 2.23-4.46 GB/s of memory bandwidth to the clusters.

Similar to typical DDR2 interfaces, a page must first be opened before it can be accessed. The controller keeps track of which DRAM page is currently open, and opens new pages when necessary. Memory operations are performed in 4x128b bursts that are also cached. Depending if the page is closed, the page is open, or the needed data is already cached, a DRAM operation can take between one and eight bus cycles. DRAM interface settings include typical RAS and CAS settings, among others.

The DRAM controllers are synthesized with the bus and share the bus hub module area on the core layer. The bus hubs were placed directly on top of the DRAM interface. Large clusters of TSVs were used to connect to the DRAM through the die-to-wafer stacking process. These

TSVs provide sufficient density to meet TSV density requirements.

The DRAM has one control layer and 1-4 gigabit bitcell layers; the seven-layer Centip3De system will include two bitcell layers. The control and sense amplifier layer was designed by Tezzaron in the same Chartered 130nm process used for the core and cache layers. The bitcell layers were provided by an outside vendor to Tezzaron, who then stacked all 2-5 wafers together for the final DRAM wafer stack. To configure the DRAM internal timing and program the number of bitcell layers, the DRAM contains a "message box" control interface that is accessible by the scan chain.

VII. Conclusion

In this work we proposed a cluster-based NTC architecture as a solution for maximizing performance in a TDP-constrained environment. Centip3De was implemented in Tezzaron's 3D stacking process, demonstrating the feasibility of 3D design, particularly when coupled with energy-efficient computing. Issues such as design reuse, floorplanning, and voltage scaling in a 3D environment were discussed. A two-layer system was fabricated and measured, achieving 3930 DMIPS/Watt energy efficiency in a 130nm process. At 46.4M 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] S. Rusu, S. Tam, H. Muljono, J. Stinson, D. Ayers, J. Chang, R. Varada, M. Ratta, and S. Kottapalli, "A 45nm 8-core enterprise Xeon® processor," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, Feb. 2009, pp. 56–57.

[2] S. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, "An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 1, pp. 29–41, Jan. 2008.

[3] S. Bell, B. Edwards, J. Amann, R. Conlin, K. Joyce, V. Leung, J. MacKay, M. Reif, L. Bao, J. Brown, M. Mattina, C.-C. Miao, C. Ramey, D. Wentzlaff, W. Anderson, E. Berger, N. Fairbanks, D. Khan, F. Montenegro, J. Stickney, and J. Zook, "Tile64 - processor: A 64-core soc with mesh interconnect," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, feb. 2008, pp. 88 –598.

[4] R. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214 –1224, june 2006.

[5] J. U. Knickerbocker, P. S. Andry, B. Dang, R. R. Horton, M. J. Interrante, C. S. Patel, R. J. Polastre, K. Sakuma, R. Sirdeshmukh, E. J. Sprogis, S. M. Sri-Jayantha, A. M. Stephens, A. W. Topol, C. K. Tsang, B. C. Webb, and S. L. Wright, "Three-dimensional silicon integration," *IBM Journal of Research and Development*, vol. 52, no. 6, pp. 553 –569, Nov. 2008.

[6] G. H. Loh, "3D-Stacked Memory Architectures for Multi-core Processors," *SIGARCH Comput. Archit. News*, vol. 36, no. 3, pp. 453–464, 2008.

[7] T. Kgil, A. Saidi, N. Binkert, S. Reinhardt, K. Flautner, and T. Mudge, "PicoServer: Using 3D stacking technology to build energy efficient servers," *J. Emerg. Technol. Comput. Syst.*, vol. 4, no. 4, 2008.

[8] T. Karnik, D. Somasekhar, and S. Borkar, "Microprocessor system applications and challenges for through-silicon-via-based three-dimensional integration," *Computers Digital Techniques, IET*, vol. 5, no. 3, pp. 205 –212, May 2011.

[9] M. Wordeman, J. Silberman, G. Maier, and M. Scheuermann, "A 3d system prototype of an edram cache stacked over processor-like logic using through-silicon vias," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, feb. 2012, pp. 186 –187.

[10] S. Borkar, "3d integration for energy efficient system design," in *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, June 2011, pp. 214–219.

[11] D. H. Kim, K. Athikulwongse, M. Healy, M. Hossain, M. Jung, I. Khorosh, G. Kumar, Y.-J. Lee, D. Lewis, T.-W. Lin, C. Liu, S. Panth, M. Pathak, M. Ren, G. Shen, T. Song, D. H. Woo, X. Zhao, J. Kim, H. Choi, G. Loh, H.-H. Lee, and S. K. Lim, "3d-maps: 3d massively parallel processor with stacked memory," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, Feb. 2012, pp. 188–190.

[12] D. Fick, R. G. Dreslinski, B. Giridhar, G. Kim, S. Seo, M. Fojtik, S. Satpathy, Y. Lee, D. Kim, N. Liu, M. Wieckowski,

G. Chen, T. Mudge, D. Sylvester, and D. Blaauw, "Centip3de: A 3930dmips/w configurable near-threshold 3d stacked system with 64 arm cortex-m3 cores," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, Feb. 2012, pp. 190–192.

[13] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein, "Scaling, power, and the future of CMOS," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, Dec. 2005, pp. 7–15.

[14] B. Zhai, R. Dreslinski, D. Blaauw, T. Mudge, and D. Sylvester, "Energy efficient near-threshold chip multi-processing," in *Low Power Electronics and Design (ISLPED), 2007 ACM/IEEE International Symposium on*, Aug. 2007, pp. 32–37.

[15] R. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, feb. 2010.

[16] G. Chen, M. Fojtik, D. Kim, D. Fick, J. Park, M. Seok, M.-T. Chen, Z. Foo, D. Sylvester, and D. Blaauw, "Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, Feb. 2010, pp. 288–289.

[17] J. Kwong, Y. Ramadass, N. Verma, and A. Chandrakasan, "A 65 nm sub-v_t microcontroller with integrated sram and switched capacitor dc-dc converter," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 1, pp. 115–126, Jan. 2009.

[18] G. Chen, H. Ghaed, R. Haque, M. Wieckowski, Y. Kim, G. Kim, D. Fick, D. Kim, M. Seok, K. Wise, D. Blaauw, and D. Sylvester, "A cubic-millimeter energy-autonomous wireless intraocular pressure monitor," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, Feb. 2011, pp. 310–312.

[19] R. G. Dreslinkski, B. Zhai, T. Mudge, D. Blaauw, and D. Sylvester, "An Energy Efficient Parallel Architecture Using Near Threshold Operation," *Proceedings Parallel Architecture and Compilation Techniques*, pp. 175–188, 2007.

[20] K. Flautner, N. S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power," in *Proceedings of the 29th annual international symposium on Computer architecture*, 2002, pp. 148–157.

[21] M. Qazi, K. Stawiasz, L. Chang, and A. Chandrakasan, "A 512kb 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," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 1, pp. 85 –96, jan. 2011.

[22] B. Zhai, D. Blaauw, D. Sylvester, and S. Hanson, "A sub-200mv 6t sram in 0.13 μm cmos," in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, feb. 2007, pp. 332 –606.

[23] J. Kulkarni, B. Geuskens, T. Karnik, M. Khellah, J. Tschanz, and V. De, "Capacitive-coupling wordline boosting with self-induced vcc collapse for write vmin reduction in 22-nm 8t sram," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, feb. 2012, pp. 234 –236.

[24] L. Chang, R. Montoye, Y. Nakamura, K. Batson, R. Eickemeyer, R. Dennard, W. Haensch, and D. Jamsek, "An 8t-sram for variability tolerance and low-voltage operation in high-performance caches," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 4, pp. 956 –963, april 2008.

[25] E. Karl, D. Blaauw, D. Sylvester, and T. Mudge, "Reliability modeling and management in dynamic microprocessor-based

systems," pp. 1057–1060, 2006.

[26] T.-H. Kim, R. Persaud, and C. Kim, "Silicon odometer: An on-chip reliability monitor for measuring frequency degradation of digital circuits," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 4, pp. 874 –880, april 2008.

[27] "ARM Cortex-M3," http://www.arm.com/products/CPUs/ARM_Cortex-M3.html.

[28] "OpenCores: Scan Based Serial Communication," http://opencores.org/project,scan_based_serial_communication.

[29] "Tezzaron Semiconductor FaStack®Technology," http://tezzaron.com/technology/FaStack.htm.

[30] "Tezzaron Semiconductor Octopus DRAM," http://www.tezzaron.com/memory/Octopus.html.

[31] "ARM Cortex-A9," http://www.arm.com/products/processors/cortex-a/cortex-a9.php.

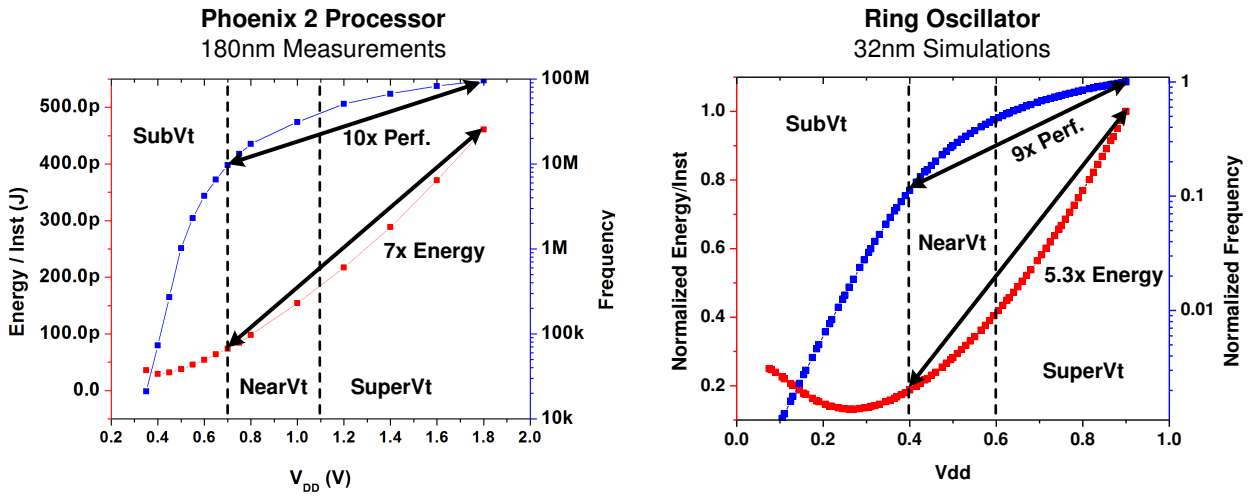


Fig. 1: Trade-offs of NTC operation in a range of process technologies. Data from 180nm is measured, while the 32nm data is simulated. NTC operation is shown to be effective in both older and newer process technologies. Results from Centip3De are from a 130nm process.

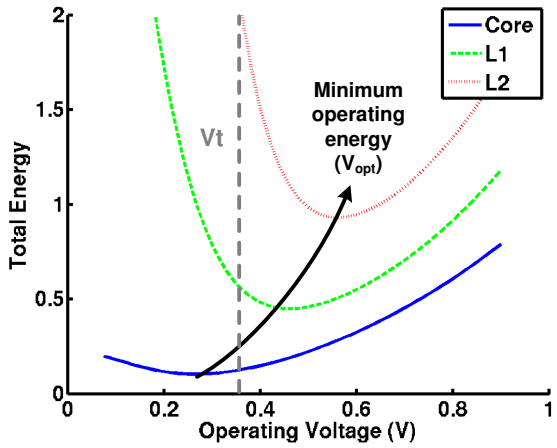


Fig. 2: Activity factor versus minimum operating energy. As activity factor decreases, the leakage component of energy/operation increases thereby making the energy optimal operating voltage increase. To account for this, Centip3De operates caches at a higher voltage and frequency that the cores. An 8T SRAM design was used to ensure that V_{min} is below V_{opt}.

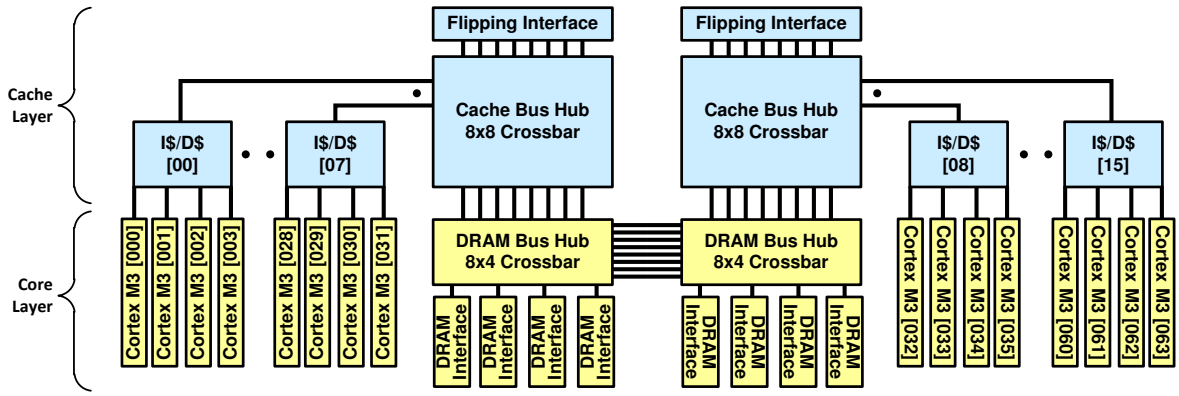


Fig. 3: System block diagram. The block diagram is organized and shaded by layer, with F2F connections shown as lines crossing between these layers. The 8 buses each connect to all 16 clusters as well as a round-robin arbiter.

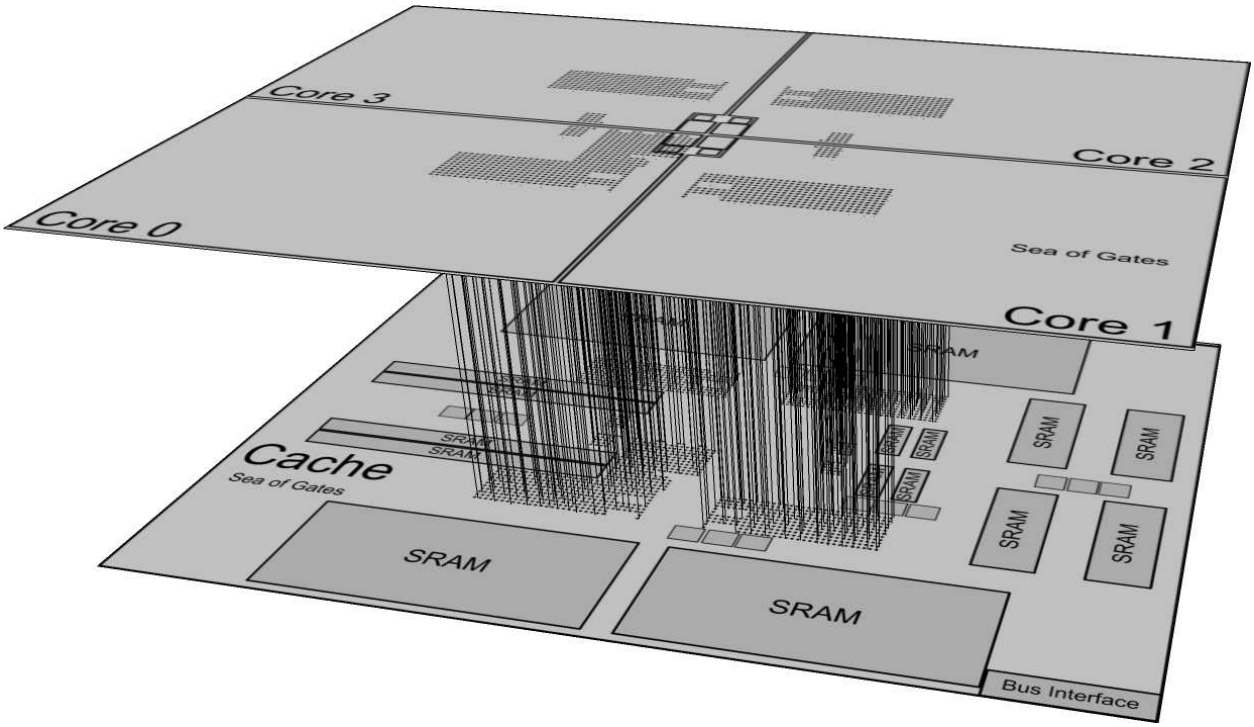


Fig. 4: Artistic rendering of a cluster. Relative module size and placement are accurate. F2F connections are represented as dots with lines between.

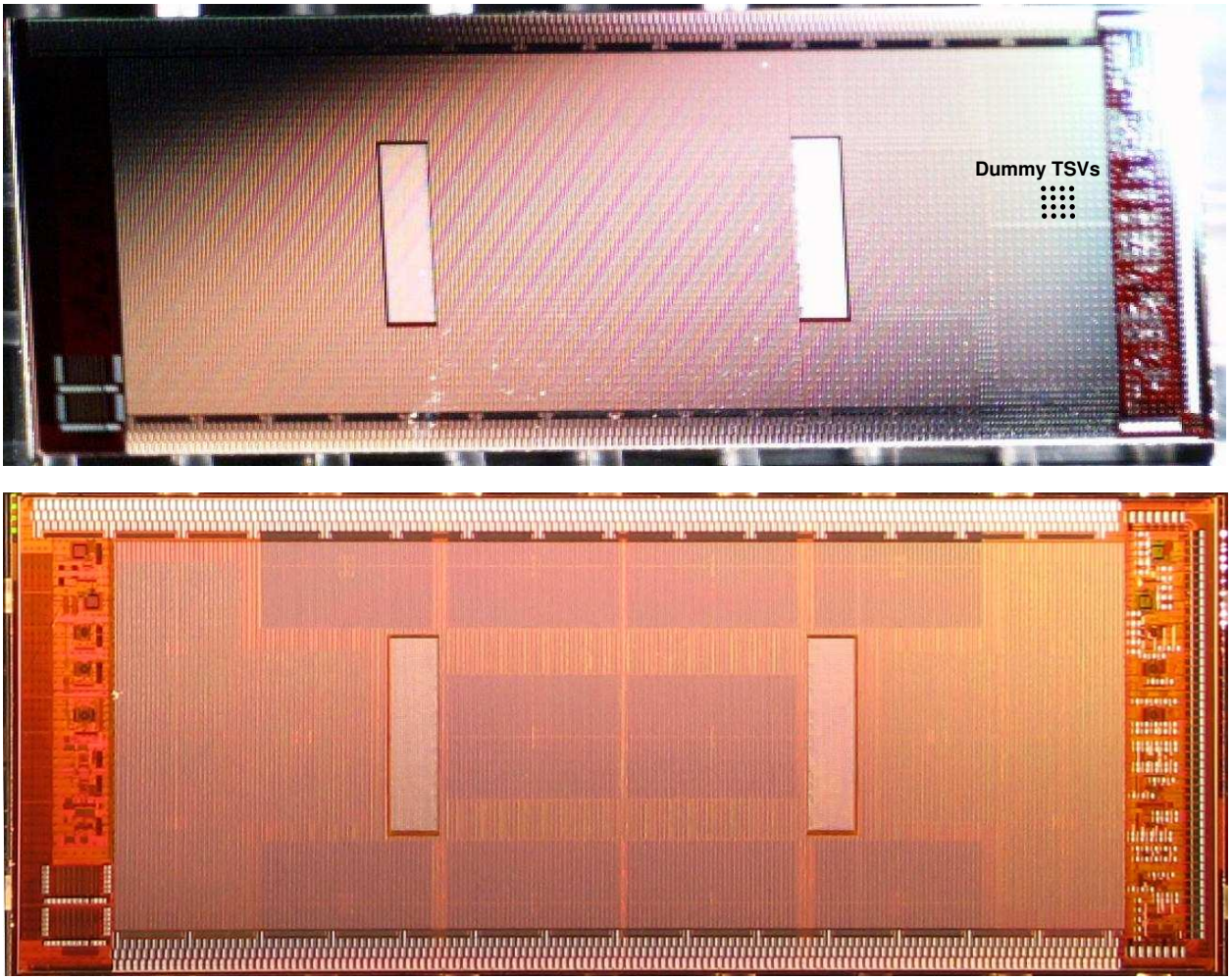


Fig. 5: Die micrographs of two-layer system. Two layers are bonded face-to-face. The clusters can be seen through the backside of the core layer silicon. Wirebonding pads line the top and bottom edges. Above, the clusters can be seen by the pattern of dummy TSVs; the clusters have a square grid of TSVs (for rotational symmetry), whereas the space between cores has more densely packed rows of TSVs.

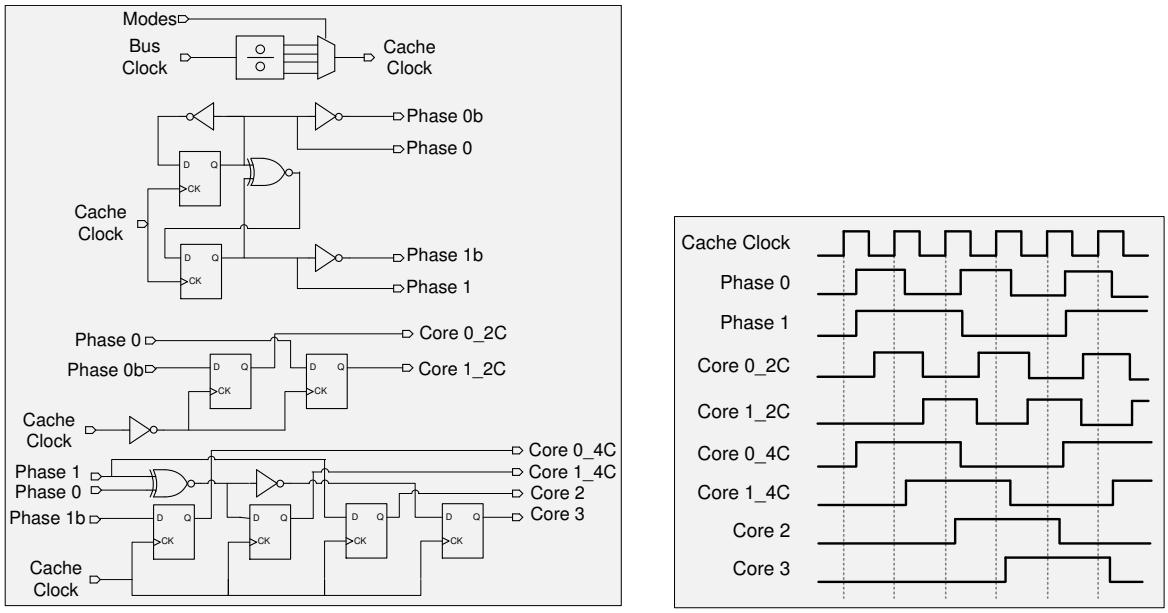


Fig. 6: Glitch-free clock generator and associated waveform. The bus clock is divided to generate the cache clock. The core clocks are then generated from the cache clock, depending on the mode. Since Core 0 sets the clock mode but also uses one of the generated clocks, it is important that its clock never glitches. For safety, this unit is designed such that no clock glitches. Additional components include clocked multiplexers for selecting between modes and gating-off partial cycles post-reset.

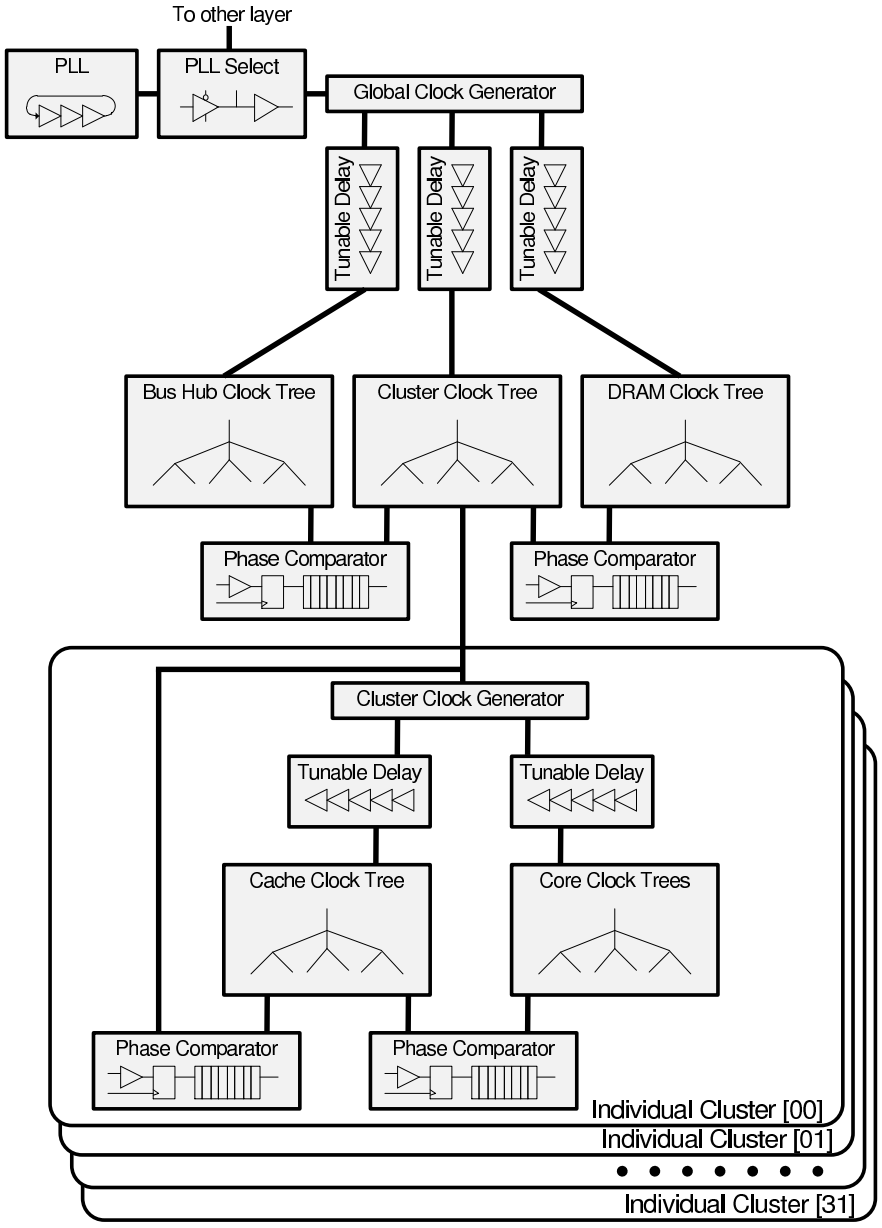


Fig. 7: Clock architecture. To align clock phases after phase changes due to voltage scaling, the clock architecture includes clock phase comparators and digitally controlled tunable delay buffers.

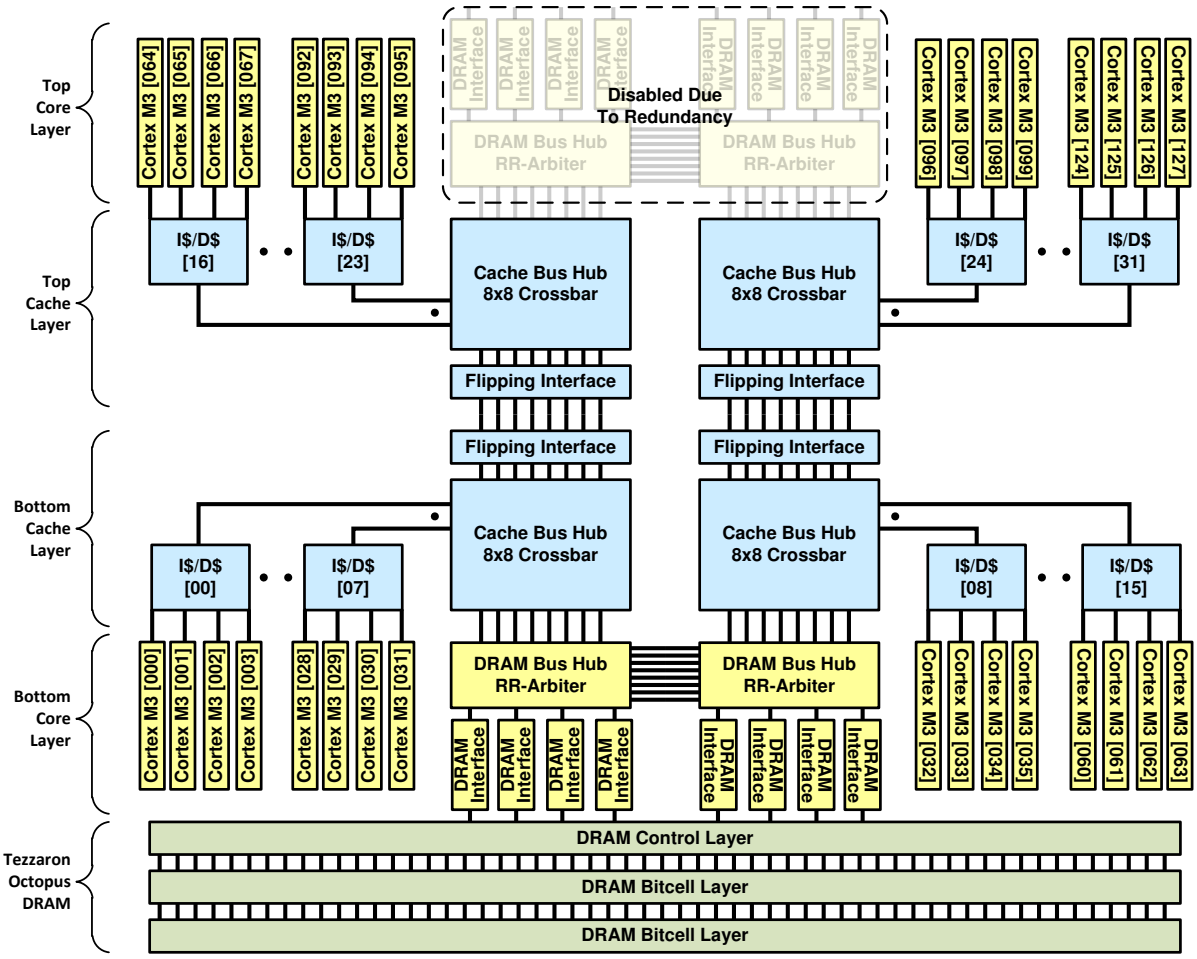


Fig. 8: Seven-layer system block diagram. Centip3De includes up to seven layers in future versions, including two core layers, two cache layers, and three DRAM layers.

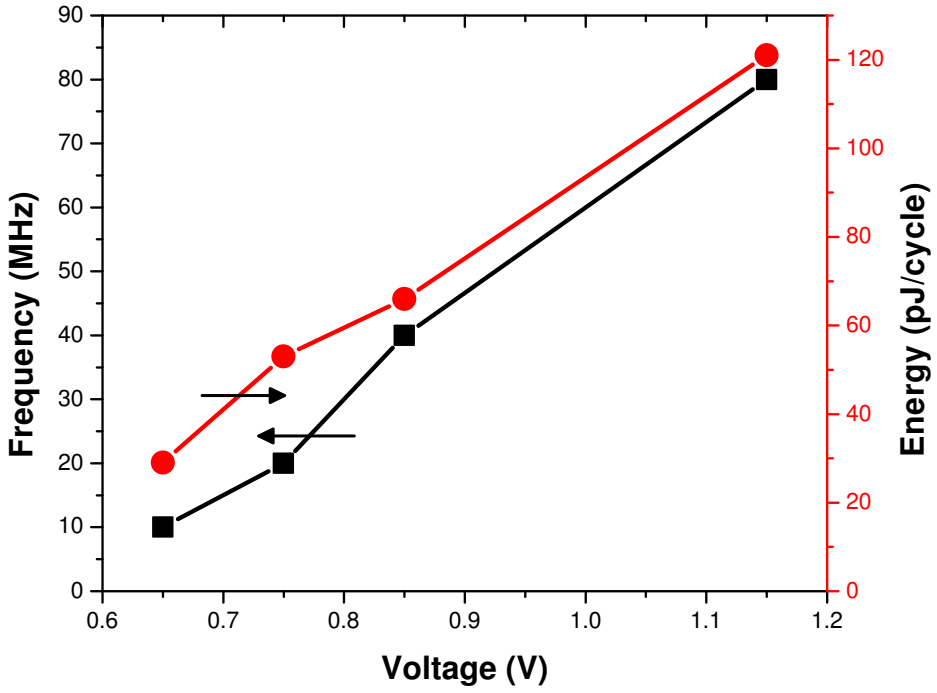
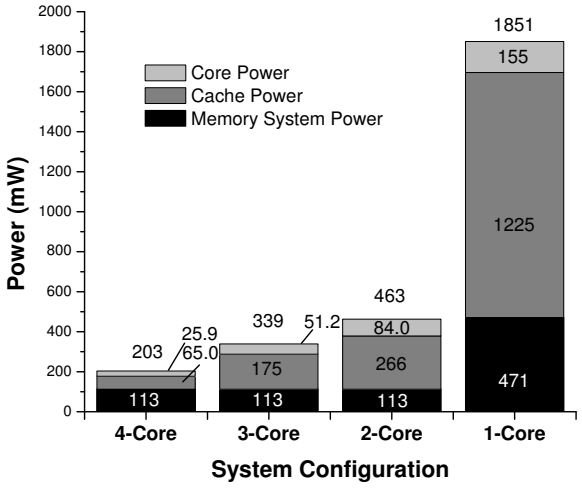
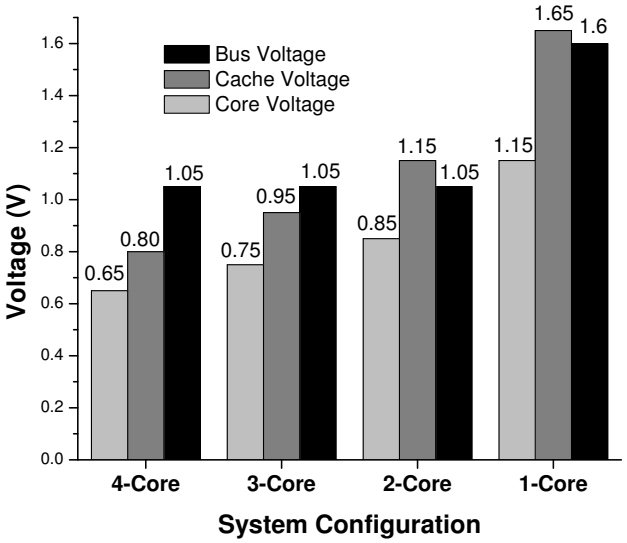


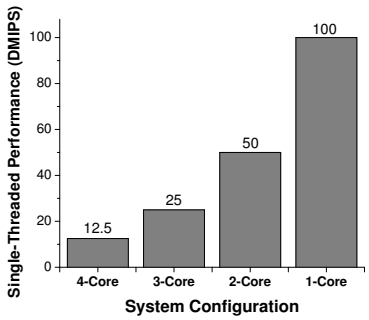
Fig. 9: Measured core frequency and energy profile.



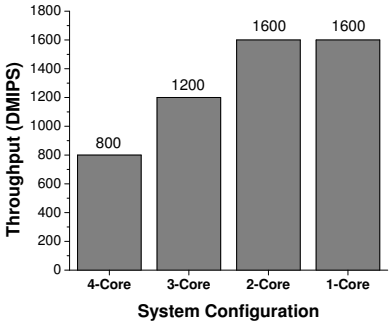
(a) System Power Breakdown



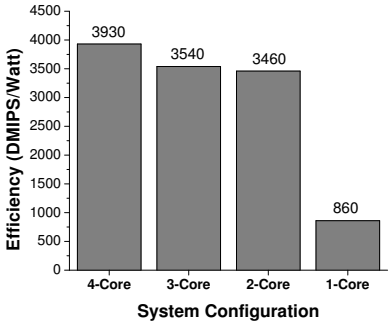
(b) System Voltage Configurations



(c) Single Threaded Performance



(d) System Throughput



(e) System Energy Efficiency

Fig. 10: Power and performance results of four system modes from Table II.

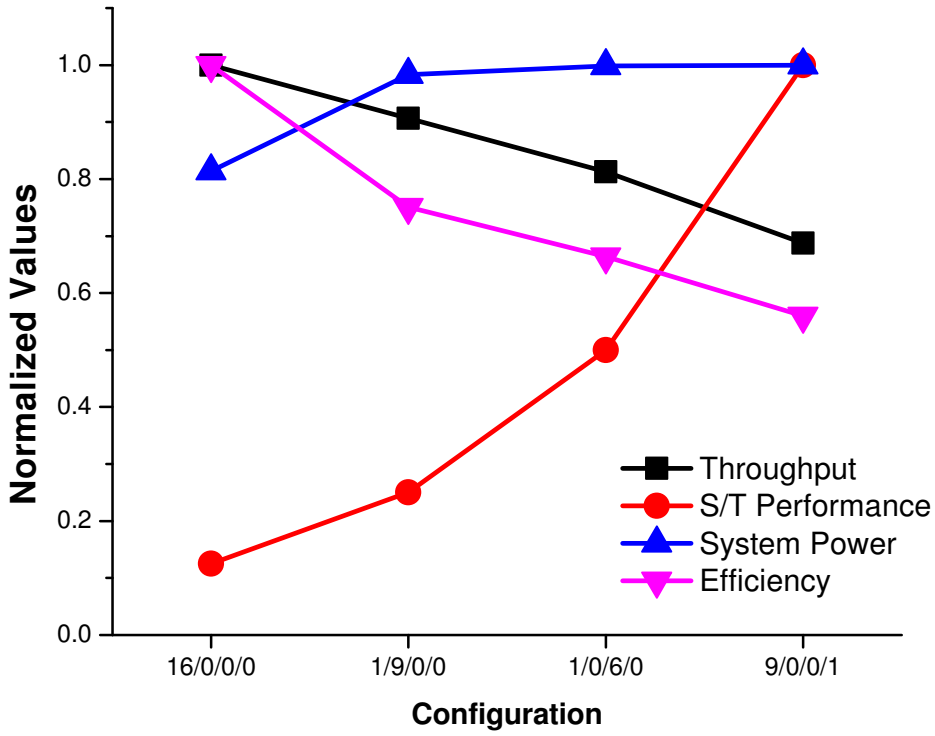


Fig. 11: Range of system configurations under a 250mW fixed TDP. The number of clusters in each mode is listed as 4-core/3-core/2-core/1-core. Each configuration emphasizes a different cluster mode, with the most energy efficient configurations on the left, and the highest single-threaded performance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

TABLE I: Design and technology data. Connection numbers include only signals. F2F connections are for a single F2F bonding interface.

Logic Layer Dimensions	2.66 x 5mm
Technology	130 nm
Metal Layers	5
Core Layer Devices	28.4 M
Cache Layer Devices	18.0 M
Core Layer Thickness	12 um
F2F Connection Pitch	5 um
F2F Connections/Cluster	1591
Bus F2F Connections	2992
Total F2F Connections	28485
B2B Connection Pitch	5 um
Total B2B Connections	3024
DRAM Connection Pitch	25 um
DRAM Connections	3624

TABLE II: Cluster configurations used in power and performance analysis.

Mode	Core (MHz)	DMIPS /Core	DMIPS /Cluster	DMIPS /W	Cluster (mW)	Core (V)	Cache (V)	Core (mW)	Cache (mW)
4C	10	12.5	50	8800	5.68	0.65	0.80	1.16	4.06
3C	20	25	75	5300	14.2	0.75	0.95	3.20	11.0
2C	40	50	100	4560	21.9	0.85	1.15	5.25	16.6
1C	80	100	100	1160	86.3	1.15	1.65	9.71	76.5

TABLE III: Power and performance analysis of selected system configurations. Total number of clusters is 16 for each system configuration.

Bus Freq	#4C	#3C	#2C	#1C	DMIPS	mW	DMIPS/W	MHz/mW
160	16				800	203	3930	3.15
160	15	1			825	211	3890	3.12
160	12	4			900	237	3790	3.03
160	15		1		850	219	3870	3.10
160	14		2		900	235	3820	3.05
160		16			1200	339	3540	2.83
160			16		1600	462	3460	2.77
320	14		2		900	594	1510	1.21
320	15			1	850	642	1320	1.06
320			16		1600	821	1950	1.56
320				16	1600	1851	860	0.69

LIST OF FIGURES

LIST OF TABLES