10.7 Centip3De: A 3930DMIPS/W Configurable Near-Threshold 3D Stacked System with 64 ARM Cortex-M3 Cores

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, Dennis Sylvester, David Blaauw

University of Michigan, Ann Arbor, MI

Recent high performance IC design has been dominated by power density constraints. 3D integration increases device density even further, and these devices will not be usable without viable strategies to reduce power consumption. This paper proposes the use of near-threshold computing (NTC) to address this issue in a stacked 3D system. In NTC, cores are operated near the threshold voltage (~200mV above Vth) to optimally balance power and performance [1]. In Centip3De, we operate cores at 650mV, as opposed to the wear-out limited supply voltage of 1.5V. This improves measured energy efficiency by 5.1×. The dramatically lower power consumption of NTC makes it an attractive match for 3D design, which has limited power dissipation capabilities, but also has improved innate power and performance compared to 2D design.

Due to higher leakage current in SRAMs compared to logic, memories reach their optimal energy/delay trade-off at higher voltages than cores: 870mV for SRAM and 670mV for logic in 130nm technology. Hence, SRAMs ideally operate at a higher voltage than cores, improving their speed. Centip3De uses this unique cache/core performance inversion by connecting four cores to each cache, where each cache operates at 4× the core frequency and communicates with the cores in a round-robin fashion. This configuration has the added advantage of automatically resolved coherence within the cluster which reduces coherency traffic and overhead.

To address Amdahl's law, Centip3De allows some cores in a cluster to be boosted by 2, 4 or 8× in frequency by ramping them to a higher voltage while disabling remaining cores in the cluster to offset the higher power consumption. Disabling the non-boosted cores opens up more of the cache to the boosted core, providing it with additional memory performance. In this way, Centip3De can be configured to maximize single-threaded performance, throughput, or a mixture of both, depending on workload.

The fabricated Centip3De system consists of two stacked dies with 64 ARM M3 near-threshold cores that make up 16 four-core clusters, each connected to a 4-way 1kB instruction cache and a 4-way 8kB data cache. The caches communicate over a 3D bus that connects them to DRAM controllers that form the backing store for the caches. Centip3De is designed to be expandable to 4 layers of cores/caches with 2-3 layers of stacked DRAM. This paper provides results for a two-layer system (referred to as the fabricated system), but for completeness we also describe the complete design that will consist of 128 cores, 4-layer logic + 3-layer DRAM (referred to as the expanded system). Final assembly of the expanded system is anticipated at a later date.

Figure 10.7.1 shows the floorplan for a cluster, which separates caches and cores into adjacent layers. The four cores communicate with their adjacent cache through a face-to-face (F2F) 3D interface, which reduces routing resource requirements by providing 331 interface connections in the middle of each core. The F2F interconnects have a pitch of 5μm and a loading equal to a small buffer in this technology. The caches each connect to the closest communication column on their layer. The vertical communication columns reduce required routing resources by approximately 50% compared to a single-layer floorplan; similar gains are obtained in energy and performance, and are not offset by 3D interconnect loading due to its relatively small overhead.

As shown in Fig. 10.7.2, the caches have four modes. In three- and four-core modes, each core operates at 1/4 the frequency of the cache, at 90 degrees out of phase from the other cores. The cache pipeline accesses internally while the cores sees a single-cycle interface. Data tags are read in the first stage of the pipeline and the data is read in the third stage. By determining the correct way in the second stage of the pipeline, the number of data array accesses are reduced, thereby increasing energy efficiency. In one- and two-core modes, each core operates at half the frequency of the cache and is 180 degrees out of phase from the other. Data arrays are read simultaneously with the tag arrays due to the reduced number of pipeline stages, increasing energy consumption per access. In both modes, accesses are monitored for conflicts, which stall later transactions. A custom 8T SRAM provides improved voltage scalability for the cache over a conventional 6T design. A 128b cache line matches the bus system, but each 32b word may be read and written independently.

3D design gives Centip3De greater bandwidth both within a cluster and within its bus system to the DRAM controllers. Eight DRAM controllers are connected to the cores with a 128b bus, providing 2.23GB/s. The eight buses are duplicated in two communication columns (Fig. 10.7.3), each with four DRAM controllers; columns are bridged on the core layer.

The cores, caches, bus interconnect, and DRAM controllers can operate at configurable, scaled frequencies and have independent voltage domains. The cache and core clocks are derived from the globally distributed system bus clock. Due to the low TSV and F2F connection parasitics, additional 3D considerations for the clock tree design were not necessary. However, scaling the core and cache voltages causes leaves of the clock trees to become misaligned. The delay of each clock tree is adjustable via a digitally-controlled, tunable delay buffer at the root of the tree, and clock tree alignment sensors are included *in situ* to assist in adjusting the trees (Fig. 10.7.4). These units and level converters are included between the core and cache interfaces, and the cache-to-bus interfaces (Fig. 10.7.4). The global clock control and alignment is controlled through a scan interface. Cache modes, alignment sensor readings, delay generator settings, and other settings are controlled via memory-mapped I/O in Core 0 of each cluster, accessible via JTAG. The JTAG port for cores 1-3 are enabled via muxes in series with Core 0 by Core 0 MMIO.

The expanded Centip3De design is shown in Fig. 10.7.3. Two pairs of the F2F bonded core/cache dies are thinned to 12μm on the cache side and then bonded back-to-back (B2B) to create a four-layer stack. This four-layer stack is thinned and diced, and the individual dies are bonded to a thinned 2-3 layer DRAM wafer (Fig. 10.7.3). This would provide the system with a maximum of 256MB of shared DRAM connected through eight 128b DDR2 interfaces.

In the expanded system, the core and cache layers have identical twins, and hence, a pull-up circuit is included to identify the position of each layer. The interfacing DRAM provides copper tracing and wirebonding pads for the logic layers. TSVs in the bottom core layer connect to this tracing, while TSVs in the top layer will remain isolated. By attaching a pull-down pad to one of these TSVs and connecting the wirebonding pad to V_{DD}, the top core layer is designed to identify itself. This allows the cache-cache B2B interface to control tri-state buffering on the vertical buses, to disable redundant DRAM controllers, bus bridges, and PLLs on the bottom layer.

Figure 10.7.5 shows silicon measurements for a fabricated 2-layer system for different cluster configurations. The default NTC cluster configuration operates with four cores at 10MHz and caches at 40MHz, achieving 8800 DMIPS/W. Based on silicon measurements of other M3-based projects, 10MHz translates to 45MHz in 45nm 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 45nm), while enabling a trade-off between latency and power (Fig. 10.7.6). The system bus operates at 160-320MHz, which supplies an ample 2.23-to-4.46GB/s memory bandwidth. The latency of the bus ranges from 1 core cycle latency for 10MHz cores to 3 cycles when cores are boosted to 80MHz. An ARM Cortex-A9 in a 40nm process is able to achieve 8000 DMIPS/W [2]. At peak system efficiency Centip3De achieves 3930 DMIPS/W in a 130nm technology.

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

References:

[1] B. Zhai, et al., "Energy efficient near-threshold chip multi-processing," *IEEE International Symp. On Low-Power Electronics Design*, pp. 32-37, 2007.

[2] http://arm.com/products/processors/cortex-a/cortex-a9.php, ARM Ltd, 2011.

978-1-4673-0377-4/12/$31.00 ©2012 IEEE

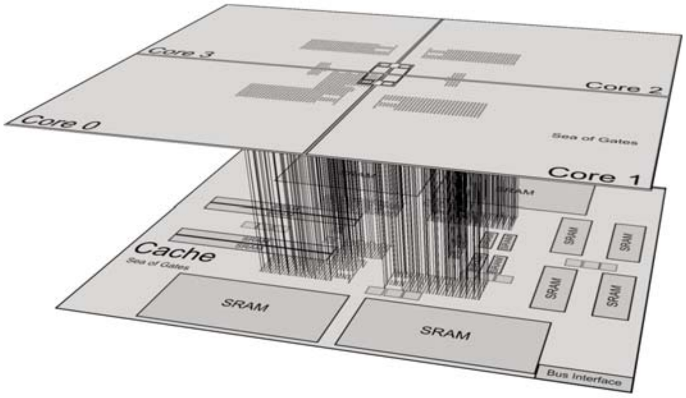


Figure 10.7.1: Cluster floorplan with F2F connections represented as dots with connecting lines between. Each cluster has 1591 F2F connections.

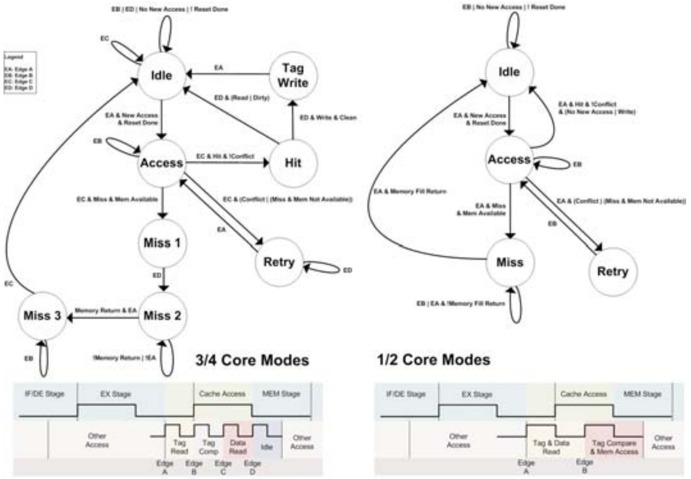


Figure 10.7.2: Cache state machines and pipeline diagrams; four cache modes are supported. Modes 1/2 improve performance, while 3/4 improve efficiency.

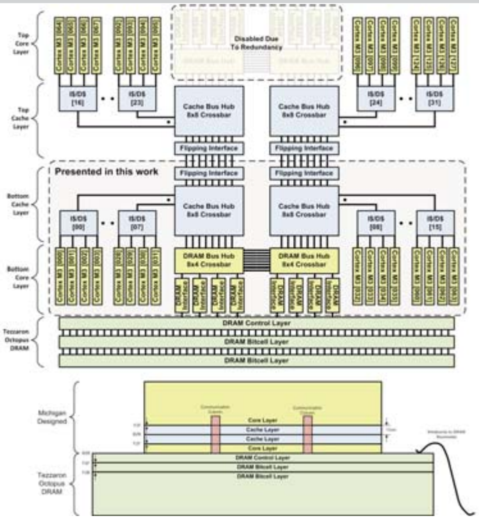


Figure 10.7.3: System block diagram and system side view. The fabricated system has an unthinned cache layer and a thinned core layer, with wirebonds connecting to TSVs on the backside.

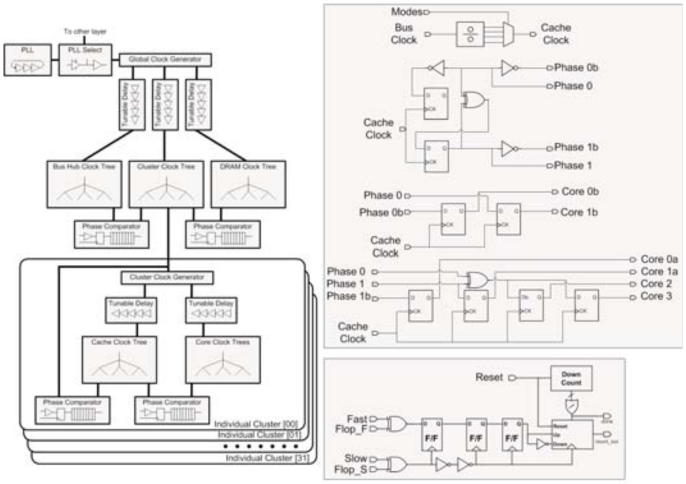
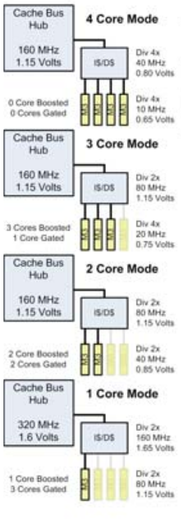


Figure 10.7.4: Left: clock tree with tunable delays and phase detectors. Right top: glitch-free clock generators used in cache. Bottom right: clock tree alignment sensor.



Selected Core Configurations

Mode	Core (MHz)	DMIPS/ Core	DMIPS/ Cluster	Cluster DMIPS/W	Cluster (mW)	Core (V)	Cache (V)	Core (mW)	Cache (mW)
4C	10	12.5	50	8800	5.68	0.65	0.8	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

Selected Bus Configurations

Frequency (MHz)	Voltage (V)	Power (mW)	Modes Supported
160	1.05	107	4C, 3C, 2C
320	1.60	295	4C, 3C, 2C, 1C

Selected System Configurations

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

Figure 10.7.5: Visualization of cluster modes, and measurement results for the cluster modes, bus architecture, and selected system configurations.

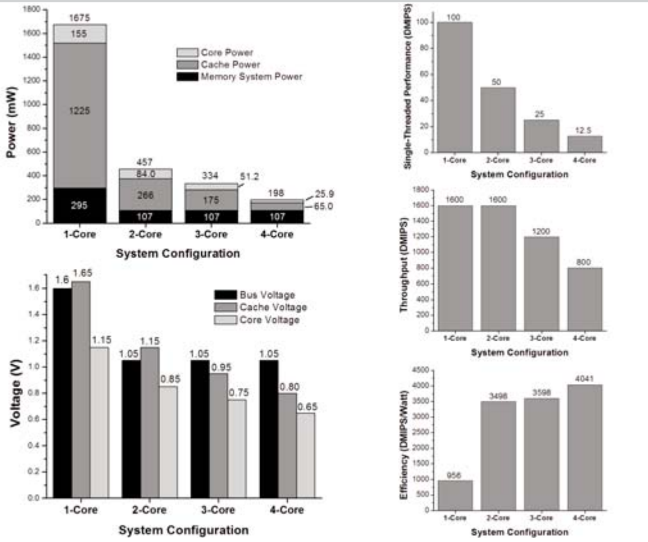
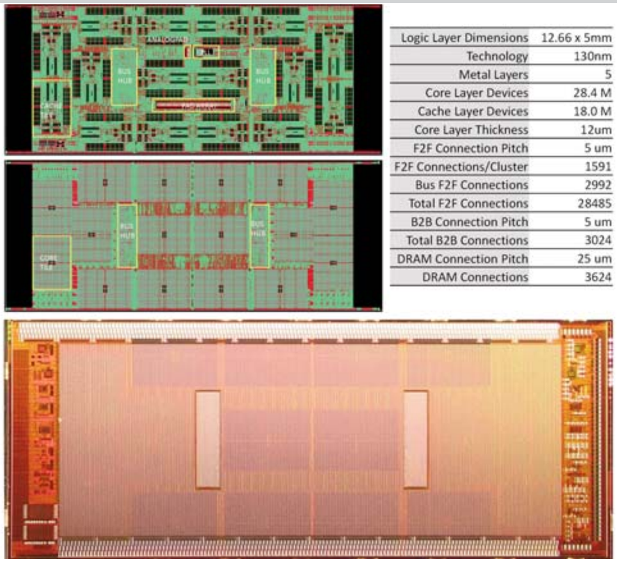


Figure 10.7.6: Measured power and performance breakdowns; each configuration has all clusters in the same mode for this analysis.



Logic Layer Dimensions	12.66 x 5mm
Technology	130nm
Metal Layers	5
Core Layer Devices	28.4 M
Cache Layer Devices	18.0 M
Core Layer Thickness	12um
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

Figure 10.7.7: Die micrograph, floorplan, and technology data.