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, Trevor Mudge, Dennis Sylvester, David Blaauw

University of Michigan
The Problem of Power

Circuit supply voltages are no longer scaling...

Power does not decrease at the same rate that transistor count increases, resulting in increased energy density

Dynamic dominates

\[ U \approx \frac{CV_{dd}^2}{A} + \frac{I_{\text{leak}}V_{dd}}{Af} \]

A = gate area \(\rightarrow\) scaling \(1/s^2\)
C = capacitance \(\rightarrow\) scaling < \(1/s\)

The emerging dilemma:

More and more gates can fit on a die, but cooling constraints are restricting their use
Today: Super-$V_{th}$, High Performance, Power Constrained

Large gate overdrive favors performance with unsustainable power density

**Must design within fixed TDP**

Goal: maintain performance, improved Energy/Operation
Operating in sub-threshold yields large power gains at the expense of performance.

Applications: sensors, medical
Subthreshold Design

Operating in sub-threshold yields large power gains at the expense of performance.

Applications: sensors, medical
Near-Threshold Computing (NTC):

- >60X power reduction
- 6-8X energy reduction
- Enables 3D integration
Caches have higher $V_{opt}$ and operating frequency

Smaller activity rate when compared to core logic

Leakage larger proportion of total power in caches

New Architectures Possible
Proposed NTC Architecture

- SRAM is run at a higher $V_{DD}$
  - Caches operate faster than core
- Can introduce clustered architecture
  - Multiple cores share L1
  - Cores see private L1
  - L1 still provides single-cycle latency

Advantages:
- Less coherence/snoop traffic
- Larger cache for processes that need it

Drawbacks:
- Core conflicts evicting L1 data
  - Not dominant in simulation
- Longer interconnect
  - 3D addressable
Proposed Boosting Approach

Measured results for 130nm LP design
10MHz becomes ~110MHz in 32nm simulation
140 FO4 delay core

Baseline
- Cache runs 4x core frequency
- Pipelined cache

Better Single Thread Performance
- Turn some cores off, speed up the rest
- Cache de-pipelined
- Faster response time, *same* throughput
- Core sees larger cache
  - Faster cores needs larger caches
Cache Timing

NTC Mode (3/4 Cores)
Low power
Tag arrays read first
0-1 data arrays accessed

Boost Mode (1/2)
Low latency
Data and tags read in parallel
4 data arrays accessed
## Cache Timing

### NTC Mode (3/4 Cores)
- Low power
- Tag arrays read first
- 0-1 data arrays accessed

<table>
<thead>
<tr>
<th>Stage</th>
<th>Label</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF/DE Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EX Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cache Access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEM Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other Access</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Other Access
- Tag Read
- Tag Comp
- Data Read
- Idle

- Other Access

- Edge A
- Edge B
- Edge C
- Edge D
Boost Mode (1/2)
Low latency
Data and tags read in parallel
4 data arrays accessed
Centip3De System Overview

- 7-Layer NTC system
- 2-Layer system completed fabrication with measured results
- Full 7-layer system expected End of 2012
Centip3De System Overview

- Cluster architecture
  - 4 Cores/cluster
  - 1kB I$, 8kB D$
  - Local clock controller operates cores 90° Out-of-phase
  - 1591 F2F connections per cluster

- Organized into layer pairs (cache↔core)
  - Minimizes routing
  - Up to two pairs
  - 16 clusters per pair
  - Cores have only vertical interconnections
Centip3De System Overview

- Bus interconnect architecture
  - Up to 500 MHz
  - 9-11 cycle latency
  - 1-3 core cycles

- 8 lanes, each 128b
  - One per DRAM interface
  - Each cluster connects to all eight
  - 1024b total

- Vertically connected through all four layers
  - Flipping interface enables 128-core system
Centip3De System Overview

- **3D-Stacked DRAM**
  - Tezzaron Octopus

- **1 control layer**
  - 130nm CMOS

- **1 Gb bitcell layers**
  - Up to two layers
  - DRAM process

- **8x 128b DDR2 interfaces**
  - Operated at bus frequency (up to 500 MHz)
Centip3De System Overview

28485 F2F

3024 B2B

28485 F2F

3624 B2B
Centip3De System Overview

130nm process
12.66x5mm per layer
28.4M device core layer
18.0M device cache layer
2-Layer Stacking Process Evaluated

For the measured 2-layer system, aluminum wirebond pads were used instead.
Cache 3D Connections

Cache

Sea of Gates

SRAM

Bus Interface

SRAM

Sea of Gates
Core 3D Connections

Core 0  Core 1
Core 2  Core 3

Sea of Gates
Cluster 3D Connections

1591 F2F Connections
Each saved ~600-1000um in routing
Prevented wiring congestion around SRAMS
Silicon Results

- **Top Core Layer**
  - Cortex M3 [064]
  - Cortex M3 [065]
  - Cortex M3 [066]
  - Cortex M3 [067]

- **Top Cache Layer**
  - IS/D$[16]$
  - IS/D$[23]$

- **Bottom Core Layer**
  - Cortex M3 [000]
  - Cortex M3 [001]
  - Cortex M3 [002]
  - Cortex M3 [003]
  - Cortex M3 [004]
  - Cortex M3 [005]
  - Cortex M3 [006]
  - Cortex M3 [007]

- **Bottom Cache Layer**
  - IS/D$[00]$
  - IS/D$[07]$

- **Tezzaron Octopus DRAM**
  - Disabled Due To Redundancy

- **DRAM Control Layer**
- **DRAM Bitcell Layer**
Die Shot

130nm process
12.66x5mm per layer
28.4M device core layer
18.0M device cache layer
System Configurations

**Cache Bus Hub**

**4 Core Mode**
- 160 MHz
- 1.15 Volts
- Div 4x 40 MHz 0.80 Volts
- Div 4x 10 MHz 0.65 Volts
- 0 Core Boosted
- 0 Cores Gated

**2 Core Mode**
- 160 MHz
- 1.15 Volts
- Div 2x 80 MHz 1.15 Volts
- Div 2x 40 MHz 0.85 Volts

**3 Core Mode**
- 160 MHz
- 1.15 Volts
- Div 2x 80 MHz 1.15 Volts
- Div 4x 20 MHz 0.75 Volts
- 3 Cores Boosted
- 1 Core Gated

**1 Core Mode**
- 320 MHz
- 1.6 Volts
- Div 2x 160 MHz 1.65 Volts
- Div 2x 80 MHz 1.15 Volts
- 1 Core Boosted
- 3 Cores Gated
Measured Results

Boosting a single cluster to 1-core mode requires disabling, or down-boosting other clusters.

1-core cluster:
- = 15x 4-core clusters
- = 6x 3-core clusters
- = 4.5x 2-core clusters

Baseline configuration depends on TDP and processing needs.
Measured Results

Single-Threaded Performance (DMIPS)

System Configuration

4-Core  3-Core  2-Core  1-Core

Core Power
Cache Power
Memory System Power

Power (mW)
Measured Results:
Centip3De – 3,930 (130nm)

Industry Comparison:
ARM A9 – 8,000 (40nm) [1]

Estimated Results:
Centip3De – 18,500 (45nm)

Conclusion

- Near threshold computing (NTC)
  - Need low power solutions to maintain TDP
  - Achieves 10x energy efficiency => 10x more computation to give TDP
  - Offers optimum balance between performance and energy
  - Allows boosting for single threaded performance (Amdahl's law)

- Large scale 3D CMP demonstrated
  - 64 cores currently
  - 128 cores + DRAM in the future
  - 3D design shown to be feasible

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