EECS 598: Integrating Emerging Technologies with Computer Architecture

Lecture 8: The Dim Horseman

Instructor: Ron Dreslinski

Winter 2016
Reading Assignment

The Problem of Power

The emerging dilemma:

More and more gates can fit on a die, but cooling constraints are restricting their use.

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

Circuit supply voltages are no longer scaling…

The equation:

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

- \( A = \) gate area \( \rightarrow \) scaling \( 1/s^2 \)
- \( C = \) capacitance \( \rightarrow \) scaling \( < 1/s \)
Today: Super-$V_{th}$, High Performance, Power Constrained

Energy per operation is the key metric for efficiency.

Goal: same performance, low energy per operation

Core 2 Duo
Subthreshold Design

Operating in the sub-threshold gives us huge power gains at the expense of performance → OK for sensors!

Phoenix
30pW @ 200kHz
Near-Threshold Computing (NTC)

Near Threshold Computing (NTC) gives us much lower energy per operation than super-threshold at a manageable performance loss (compared to sub-threshold.) If we can reclaim this 10x slowdown, then our goal will be achieved!
Near-Threshold Computing (NTC):

- 100X power reduction
- 10X energy reduction
- Invest portion of extra transistors from scaling to overcome barriers
Measured Results

Subliminal 2 Processor

- **Energy**
- **Frequency**

\[ V_{dd} = 350 \text{mV}, \ 3.52 \text{pJ/inst}, \ 354 \text{kHz} \]

**Subliminal 2 Design**
- 130nm Design
- 20.5 MHz @ 1.2V
- 33.1 pJ/inst @ 1.2V
- ~11x NTC Performance Loss
- ~7x NTC Energy Reduction

**Phoenix Design**
- 65nm Design
- ~9x NTC Performance Loss
- ~10x NTC Energy Reduction
Performance Loss

- How do we overcome 10x performance loss in NTC?
  - For parallelizable applications, add more cores
  - 100x Power savings per core, so still savings

- Super-linear increase in power with addition of cores
  - Overhead from parallelization
  - Serial portions
- Still better than increasing supply voltage to attain performance

Is there any additional advantages to leverage in NTC?
Increased Variation

- As the supply is lowered, sensitivity to process variation increases dramatically
- Traditionally, this results in margining and waste to guarantee functionality
- These trends become more prevalent at smaller technology nodes

Body bias clustering for post-silicon tuning (TCAD08)

NTC Techniques
- Razor-style self-tuning
- Variation tolerant circuits
- Embedded variation sensors
- Adaptive body biasing
**Increased Functional Failure**

- With reduced supply voltage larger impact from Random Dopant Fluctuation (RDF)
- Increased RDF leads to increased bitcell failure rates
- Traditional solutions not sufficient:
  - Redundancy
  - Increased cell sizes

**NTC Techniques**
- New cell structures
- Decoupled Read/Write
- Wordline boosting
- Alternative cache designs
NTC – Opportunities and Challenges

- **Opportunities:**
  - New architectures
  - Optimized Processes
  - Synergies with 3D Integration – less thermal restrictions, shorter interconnects

- **Challenges:**
  - Performance Loss
    - Many-core designs to improve parallelism
    - Core boosting to improve single thread performance
  - Low Voltage Memory
    - New SRAM designs
    - Robustness analysis at near-threshold
  - Variation
    - Razor [Ernst’03] and other in-situ delay monitoring
    - Adaptive body biasing
Regaining Performance

Full Voltage

NTC

NTC Ideal Parallelism

NTC w/Parallel Overheads & Boosting

Time

Energy/Operation

[IEEEMicro'13]
Where is “Near”? 

- Across a wide range of parallel applications and technology nodes the NTC operating point lies $\sim 100-200 \text{mV}$ above the threshold voltage.
SRAM has a lower activity rate than logic

$V_{DD}$ for minimum energy operation ($V_{MIN}$) is higher

New architectures possible

32nm Simulation

Decreasing activity factor results in increasing $V_{MIN}$
**Architectural Impact of NTC**

- Converting the NTC operating points to frequency:
  - L1 Caches operate ~3x faster than the core @ 32nm
New 3D NTC Architectures

Key Insight:
- SRAM is run at a higher $V_{DD}$ than cores with little energy penalty, allowing caches to operate faster than the core
Clustering Tradeoffs

Tradeoffs
-----------------------
+ Clustered Sharing
- Cluster Conflict
- New Bus (Offset by 3D)
- L1 Speed (Offset by NTC)
L1 Cache Size Tradeoff

- Decreased Miss Rate
- Higher Energy/Access
- Less Cycles
- Less L2 Accesses
Various Scaling Methods

- Baseline
  - Single CPU @ 233MHz

- Simple CMP
  - One core per L1

- Proposed cluster-based CMP
  - Multiple cores per L1
  - Vdd scaling

Normalized Energy/Operation

SPLASH-2 Average: 74% over baseline, 55% over simple CMP

[ISLPED’07]
Results – Energy Optimal L1 Size (Single Core)

- Energy dependency on L1 size
  - Trade-off between L1 and L2 access
Energy Optimal Cluster-based CMP (Fixed Die Size)
Proposed Boosting Approach

**Baseline**
- Cache runs 4x core frequency
- Pipelined cache

**Better Single Thread Performance**
- Turn some cores off, speed up the rest
- Cache de-pipelined
- Core sees larger cache
  - Faster cores need larger caches to hide latency
Cache Timing

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

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

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

<table>
<thead>
<tr>
<th>IF/DE Stage</th>
<th>EX Stage</th>
<th>Cache Access</th>
<th>MEM Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other Access</td>
<td>Tag &amp; Data Read</td>
<td>Tag Compare &amp; Mem Access</td>
<td>Other Access</td>
</tr>
</tbody>
</table>

**Boost Mode (1 or 2 cores)**

- Low latency
- Data and tags read in parallel
- 4 data arrays accessed
Analysis of NTC Across Technology

- Across technologies:
  - 7nm, 10nm, 14nm FinFET technologies
  - 20nm, 28nm, 40nm planar technologies
  - Models provided by ARM, tuned consistently

- Planar trend was NTC becoming less effective
- FinFET is reinvigorating near-threshold
  - 2-3× more scalable
  - better energy to start too
- Fundamental device characteristics (DIBL, SS)