Block Disabling Characterization and Improvements in CMPs Operating at Ultra-low Voltages

A. Ferrérón¹, D. Suárez-Gracia², J. Alastruey-Benedé¹, T. Monreal³, V. Viñals¹

¹Universidad de Zaragoza, Spain
²Qualcomm Research Silicon Valley, USA
³Universidad Politécnica de Cataluña, Spain

SBAC-PAD Oct-2014
Operation near the threshold voltage ($V_{th}$)

$V_{dd}$ and $V_{th}$ scaling has stopped

Power density no longer stays constant among technology generations and dark silicon appears

Operation at ultra-low $V_{dd}$

- Reduce the power and energy consumption
- Switch on more cores to exploit parallelism
Operation near the threshold voltage ($V_{th}$): Challenges

Delay increases: lower voltage $\rightarrow$ lower frequency

- Compensate with parallelism: more active cores with the same power budget

Increasing sensitivity to process variation (deviation of device parameters from their nominal values)

- Memory structures especially sensitive to variation
  - Conventional 6T cells: read, write, access, and hold failures
  - Lower voltages $\rightarrow$ stability margins decrease $\rightarrow$ increasing cell failure rate
  - $V_{dd_{min}}$ of memory blocks to guarantee reliable operation
Objective

Lower $V_{dd}$ to near-threshold voltages $\rightarrow$ energy efficient operation

Problem

High sensitivity of SRAM structures to variation at ultra-low $V_{dd}$

Our proposal

Mitigate the impact of SRAM cell failures at ultra-low $V_{dd}$ using low complexity techniques:
Block Disabling with Operational Tags and Block Disabling with Operational Tags and Cache-to-cache Transfers
Outline

Impact of SRAM Failures at Ultra-low Voltages

Block Disabling (BD) Fundamentals and Trade-offs

Improve BD by protecting the cache tags
   BD with operational tags (BDOT)
   BD with operational tags and C2C (BDOT-C2C)

Results

Conclusions
Outline

Impact of SRAM Failures at Ultra-low Voltages

Block Disabling (BD) Fundamentals and Trade-offs

Improve BD by protecting the cache tags
BD with operational tags (BDOT)
BD with operational tags and C2C (BDOT-C2C)

Results

Conclusions
Example of Probability of Failure of SRAM Cells at 22nm
Example of Probability of Failure of SRAM Cells at 22nm

![Graph showing the probability of failure of SRAM cells at 22nm with varying Vdd values. The graph indicates a decreasing probability of failure as Vdd decreases, with a shaded region indicating the objective range.]
Bit Probability of Failure Affects Yield
Bit Probability of Failure Affects Yield
Outline

Impact of SRAM Failures at Ultra-low Voltages

Block Disabling (BD) Fundamentals and Trade-offs

Improve BD by protecting the cache tags
  BD with operational tags (BDOT)
  BD with operational tags and C2C (BDOT-C2C)

Results

Conclusions
Traditional Cache Hierarchy

![Diagram of Traditional Cache Hierarchy]

- L2 bank
- L1i
- L1d
- R
- P
Traditional Cache Hierarchy
Block Disabling Fundamentals

SRAM cell failure detected:
Block Disabling (BD) deactivates entry (tag and data)
Simple implementation and low overhead: 1 bit per cache entry

LLC Bank

<table>
<thead>
<tr>
<th>Tag Array</th>
<th>Data Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>set i</td>
<td>@C @I</td>
</tr>
<tr>
<td>set i+1</td>
<td>@A @H</td>
</tr>
<tr>
<td>set i+2</td>
<td>@D @F</td>
</tr>
<tr>
<td>set i+3</td>
<td>@B @G</td>
</tr>
</tbody>
</table>

@C @I [C] @I
@A @H [A] @H
@D @F [D] @F
@B @G [B] @G

set i
set i+1
set i+2
set i+3
Block Disabling Fundamentals

SRAM cell failure detected:
Block Disabling (BD) deactivates entry (tag and data)
Simple implementation and low overhead: 1 bit per cache entry

<table>
<thead>
<tr>
<th>LLC Bank</th>
<th>Data Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag Array</td>
<td></td>
</tr>
<tr>
<td>set i</td>
<td>@C @I</td>
</tr>
<tr>
<td>set i+1</td>
<td>@A @H</td>
</tr>
<tr>
<td>set i+2</td>
<td>@F [X]</td>
</tr>
<tr>
<td>set i+3</td>
<td>@B @G</td>
</tr>
</tbody>
</table>
Block Disabling at Ultra-low Voltages

At lower voltages capacity and associativity degrade very fast

- Available capacity for 16-way, 1MB cache bank with block disabling (block size is 64 bytes):

<table>
<thead>
<tr>
<th>Vdd</th>
<th>Available capacity (KB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55V</td>
<td>887 KB (86%)</td>
</tr>
<tr>
<td>0.50V</td>
<td>408 KB (40%)</td>
</tr>
<tr>
<td>0.45V</td>
<td>138 KB (13%)</td>
</tr>
</tbody>
</table>
Block Disabling at Ultra-low Voltages

At lower voltages capacity and associativity degrade very fast

- Associativity degradation for 16-way, 1MB cache bank with block disabling (block size is 64 bytes):

![Graphs showing associativity degradation at various voltages](image-url)
Inclusive Hierarchies

LLC Bank (Shared)
Tag Array  Data Array

Private Level

Private Level
Inclusive Hierarchies and Block Disabling Interaction

LLC Bank (Shared)

Tag Array | Data Array
--- | ---
@C | @I
@A | @H
@D | @F
@B | @G

Private Level

A. Ferreron {ferreron@unizar.es} - SBAC-PAD'14
Block Disabling Characterization and Improvements in CMPs Operating at Ultra-low Voltages
Outline

Impact of SRAM Failures at Ultra-low Voltages

Block Disabling (BD) Fundamentals and Trade-offs

Improve BD by protecting the cache tags
  BD with operational tags (BDOT)
  BD with operational tags and C2C (BDOT-C2C)

Results

Conclusions
BD with operational tags: BDOT

Allow blocks to be allocated as just tags: entries with faulty bits can still be used to allocate tag-only blocks in LLC

LLC Bank (Shared)

Tag Array | Data Array
--- | ---
@C | @I
@A | @H
@D | @F
@B | @G

Private Level

@C | @C
@F | @F

...
BD with operational tags (BDOT)

- Protect the tag array
  - Bigger/robust cells: bigger transistors/more transistors per cell (assist circuitry)
  - More complex error correction codes (ECC)
- Why not protect the whole cache structure?
  - Area and power increase when using bigger/robust cells
  - Complex ECC require extra storage and checking hardware: might increase access latency
  - Tag array roughly 10% of the cache area (LLC)
BDOT with cache-to-cache transfers: BDOT-C2C

- Problem: requests to tag-only blocks $\rightarrow$ off-chip transactions
- Observation: shared blocks already on-chip (private levels)

![Diagram of LLC Bank (Shared) and Private Levels]

A. Ferreron {ferreron@unizar.es} - SBAC-PAD'14
BDOT with cache-to-cache transfers: BDOT-C2C

- Problem: requests to tag-only blocks $\rightarrow$ off-chip transactions
- Observation: shared blocks already on-chip (private levels)

![Diagram showing LLC Bank (Shared) and Private Levels with tags and data arrays]
BDOT with cache-to-cache transfers: BDOT-C2C

Provide cache-to-cache transfers of clean blocks: leverage coherence protocol

- The protocol already does cache-to-cache transfers of exclusively owned blocks
- Slight change in the coherence protocol behavior, but no hardware overhead
- Potential gain depends on the applications sharing degree
Outline

Impact of SRAM Failures at Ultra-low Voltages

Block Disabling (BD) Fundamentals and Trade-offs

Improve BD by protecting the cache tags
   BD with operational tags (BDOT)
   BD with operational tags and C2C (BDOT-C2C)

Results

Conclusions
Methodology

- Experimental set-up:
  Simics + GEMS + GARNET + DRAMSim2 + McPAT
- PARSEC benchmark suite
- Random faults + Monte Carlo simulations
On-chip Energy Consumption

Minimum energy on-chip:
volatges values between 0.45-0.5V more active cores → potential higher performance
On-chip Energy Consumption

Energy consumption normalized to 0.8V

- lower Vdd
- Block Disabling (On-Chip)
- BDOT (On-Chip)
- BDOT-C2C (On-Chip)
- the lower the better

- Minimum energy on-chip: voltages values between 0.45-0.5V
- more active cores
On-chip Energy Consumption

Minimum energy on-chip:
volts values between 0.45-0.5V more active cores → potential higher performance
Minimum system energy:
off-chip memory energy consumption main source
higher voltage values (0.55-0.6V)
Outline

Impact of SRAM Failures at Ultra-low Voltages

Block Disabling (BD) Fundamentals and Trade-offs

Improve BD by protecting the cache tags
   BD with operational tags (BDOT)
   BD with operational tags and C2C (BDOT-C2C)

Results

Conclusions
Conclusions

- Operation near $V_{th}$ for energy efficient operation
  - Switch on inactive cores
  - Reduce the overall energy consumption
- SRAM structures fail when lowering $V_{dd}$
  - BD: simple, low overhead, but not effective at ultra-low $V_{dd}$
  - Inclusive hierarchies: BD increases inclusion victims
    - BDOT: allow blocks allocated as tag-only → protect inclusion
    - BDOT-C2C: provide cache-to-cache transfers of shared blocks → reduce off-chip transactions
  - BDOT & BDOT-C2C: substantial reduction on-chip power and energy consumption
Block Disabling Characterization and Improvements in CMPs Operating at Ultra-low Voltages

A. Ferrerón¹, D. Suárez-Gracia², J. Alastruey-Benedé¹, T. Monreal³, V. Viñals¹

¹Universidad de Zaragoza, Spain
²Qualcomm Research Silicon Valley, USA
³Universidad Politécnica de Cataluña, Spain

SBAC-PAD Oct-2014