Memory Hierarchies in Intelligent Memories: Energy/Performance Design

Wei Huang, Jose Renau, Seung-Moon Yoo and Josep Torrellas

University of Illinois at Urbana-Champaign
Motivation

- Advances in technology:
  - Processor and Memory integration
  - Many processors on a chip
- How to design for high performance
- Energy consumption is a big concern
- Problems in cooling system
Goals of this work

- Evaluate trade-offs in memory hierarchy
  - Energy consumption
  - Performance
  - Area requirements
- Detailed energy consumption analysis
Findings of the Work

- Modest cache size is necessary
- Easy modifications in memory reduce energy consumption
The FlexRAM Architecture

Yi Kang, Wei Huang, Seung-Moon Yoo, Diana Keen, Zhenzhou Ge, Vinh Lam, Pratap Pattanaik and Josep Torrellas - ICCD99
Chip Architecture

- 64 nodes, each one includes:
  - 2-issue processor @800Mhz
  - 1MByte DRAM (12 clk)
  - Row Buffers (6 clk)
  - Cache (1 clk)
How a Memory Bank Works

- 4 Memory sub-banks, each 256KBytes
- 5 Row Buffers, each 1KByte
- 1 Data Buffer 256bits
Small Area Memory Banks

+ More Energy
+ Less Spatial Locality
+ More Localities
Pipelining the requests

Faster memory system without increased energy consumption
Advanced Memory Banks I

- Less Energy and Contention
+ More Area

Active
IS(2,4)
## Advanced Memory Banks II

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
</table>

- Less Energy and Contention

<table>
<thead>
<tr>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
</table>

+ More area

---

Active

IS(2.8)
Terminology for Memory Systems

- Trad(i,j): Traditional
- S(i,j): Segmented
- IS(i,j): Interleaved Segmented
- ISP(i,j): Interleaved Segmented Pipelined

- i: Degree of interleaving
- j: Number of sub-banks per interleaving way
## Energy and Area Issues

<table>
<thead>
<tr>
<th>Access Type</th>
<th>Trad(1,4)</th>
<th>S(1,4)</th>
<th>IS(2,4)</th>
<th>IS(2,8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cache hit (8KB)</td>
<td>191pj</td>
<td>191pj</td>
<td>191pj</td>
<td>191pj</td>
</tr>
<tr>
<td>RB Hit</td>
<td>468pj</td>
<td>468pj</td>
<td>506pj</td>
<td>517pj</td>
</tr>
<tr>
<td>Bank Access</td>
<td>6999pj</td>
<td>3729pj</td>
<td>2287pj</td>
<td>1556pj</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Area (.18µm)</th>
<th>Trad(1,4)</th>
<th>S(1,4)</th>
<th>IS(2,4)</th>
<th>IS(2,8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.25mm²</td>
<td>4.25mm²</td>
<td>4.83mm²</td>
<td>5.23mm²</td>
</tr>
</tbody>
</table>

- More advanced configurations:
  - More Area
  - Less Energy
Evaluation Environment

- **Fixed parameters:**
  - 2-issue processor @800MHz
  - Prefetch
  - Cache, RB, Bank latencies (1,6,12 cycles)

- **Variable parameters:**
  - Cache sizes (256B,1KB,8KB,16KB)
  - Memory Banks:
    - Trad(1,4),S(1,4),SP(1,4)
    - IS(2,4),ISP(2,4),IS(2,8),ISP(2,8)
## Applications

<table>
<thead>
<tr>
<th>Applic</th>
<th>What It Does</th>
<th>Cache Hit Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTree</td>
<td>DM Tree Generation</td>
<td>50.7</td>
</tr>
<tr>
<td>DTree</td>
<td>DM Tree Deployment</td>
<td>98.6</td>
</tr>
<tr>
<td>BSOM</td>
<td>BSOM Neural Network</td>
<td>94.7</td>
</tr>
<tr>
<td>BLAST</td>
<td>Protein Matching</td>
<td>96.9</td>
</tr>
<tr>
<td>Mpeg</td>
<td>Mpeg-2 Motion Estimation</td>
<td>99.9</td>
</tr>
<tr>
<td>FIC</td>
<td>Fractal Image Compressor</td>
<td>97.8</td>
</tr>
</tbody>
</table>
Performance: Memory Banks

- Small performance improvement in advanced configurations with 1KByte cache
Performance: Cache Effect

- Modest cache size is required for performance

![Graph showing IPC for different cache sizes and configurations]
Energy-Delay Product

- Big improvement in energy-delay product with more advanced memory configurations
Energy-Delay Product: Cache

- 8KBytes have the best energy-delay product
Conclusions

- Modest size cache is enough (8KBytes)
  - Improves performance
  - Reduces energy consumption

- Segmentation S(1,4)
  - Reduces energy consumption

- When area is available: use interleaving
  - IS(2,4) increases by 14% the area
Backup Slides
Area-Delay Product: MB

- SP(1,4) best are utilization
Area-Delay Product: Cache

8KBytes is a sweet point for area-delay product
Power Consumption: Cache

- Power is a bad metric, only useful as a constraint
## Memory Access Timing

<table>
<thead>
<tr>
<th>Operation</th>
<th>Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-address buffer</td>
<td>1</td>
</tr>
<tr>
<td>X-address decoder</td>
<td>1</td>
</tr>
<tr>
<td>Wordline enabling</td>
<td>2</td>
</tr>
<tr>
<td>Charge sharing</td>
<td>2</td>
</tr>
<tr>
<td>Bit line sensing</td>
<td>2</td>
</tr>
<tr>
<td>DRAM Data buffer</td>
<td>2</td>
</tr>
<tr>
<td>L1 Cache</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>11</strong></td>
</tr>
</tbody>
</table>

Table showing the timing of various memory access operations.
# Area Requirements

<table>
<thead>
<tr>
<th>Cache size</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>256B</td>
<td>0.07</td>
</tr>
<tr>
<td>1K</td>
<td>0.16</td>
</tr>
<tr>
<td>8K</td>
<td>0.60</td>
</tr>
<tr>
<td>16K</td>
<td>1.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bank size</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,4)</td>
<td>4.25</td>
</tr>
<tr>
<td>(2,4)</td>
<td>4.83</td>
</tr>
<tr>
<td>(2,8)</td>
<td>5.23</td>
</tr>
</tbody>
</table>
Small Area Memory Banks

+Energy
+Spatial Locality
+Localities

0 1 2 3 4

[Diagram showing memory banks with shaded areas indicating active regions and distances marked as 'w']

0 1 2 3 4

[Diagram showing another set of memory banks with shaded areas and distances marked as 'w']
Advanced Memory Banks

Less Energy and Contention
More area

Less Energy and Contention
Even more area