An Energy-Oriented Evaluation of Buffer Cache Algorithms Using Parallel I/O Workloads

Jianhui Yue, Yifeng Zhu, Zhao Cai,

The authors are with the Department of Electrical and Computer Engineering, University of Maine, Barrows Hall 101, Orono, ME, 04469. E-mail: {jyue, zhu, zcai}@eece.maine.edu.
Outline

• Introduction
• Background
• Energy evaluation
• Conclusions
Introduction (1/2)

• Power management becomes an increasingly important concern.

• The energy breakdown measured on a real server shows that the memory energy consumption is 41% of the total and is 50% more than the processors.

in “Energy Management for Commercial Servers, 2003”
Introduction (2/2)

• It develops a detailed trace-driven memory simulator and uses five parallel I/O workloads to compare the relative energy efficiency of eight replacement algorithms.
  – LRU, Belady, LIRS, ARC, 2Q, MQ, LRFU, and LRU2.
Background

• Cache replacement polices
• DMA Overlapping
Cache replacement polices (1/5)

- **Belady**
  - It is the optimal replacement algorithm.
- **LRU**
- **LRU-2**
  
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last: 12/24</td>
<td>Last: 12/20</td>
<td>Last: 12/22</td>
</tr>
</tbody>
</table>
- **2Q**
  - $A_{in}^{1}$: FIFO queue, $A_{m}$: LRU queue
  - $A_{out}^{1}$: a ghost LRU queue
Cache replacement policies (2/5)

- **LRFU**
  
  \[ C(x) = \begin{cases} 
  1 + 2^{-\lambda} \cdot C(x), & \text{if } x \text{ is referenced at time } t; \\
  2^{-\lambda} \cdot C(x), & \text{otherwise}
  \end{cases} \]

  where \( \lambda, 0 \leq \lambda \leq 1 \)

  when \( \lambda = 1 \) : LRU

  \[ C(x) = \begin{cases} 
  1 + \frac{1}{2} \cdot C(x), & \text{if } x \text{ is referenced at time } t; \\
  \frac{1}{2} \cdot C(x), & \text{otherwise}
  \end{cases} \]

  when \( \lambda = 0 \) : LFU

  \[ C(x) = \begin{cases} 
  1 + C(x), & \text{if } x \text{ is referenced at time } t; \\
  C(x), & \text{otherwise}
  \end{cases} \]
Cache replacement polices (3/5)

• MQ
  – It uses m+1 LRU queues. (typically, m = 8)
  – Q_0, Q_1, Q_2, Q_3, Q_4, Q_5, Q_6, Q_7, Q_{out}
  – Q_i contains blocks that have been referenced at least 2^i times but no more than 2^{i+1} times recently
  – It associates each block with a timer that is set to currentTime + lifeTime.
Cache replacement policies (4/5)

- Buffer
  - LIR
    - LIR block
  - HIR
    - HIR block
    - HIR block metadata

<table>
<thead>
<tr>
<th>Blocks / Virtual time</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>
<th>Recency (stack position)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 (Time7)</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 (Time7)</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 (Time7)</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>0 (Time7)</td>
</tr>
</tbody>
</table>

Stack S size = 3

Stack Q size = 2
Cache replacement polices (5/5)

- **ARC**
  - It uses two LRU lists: $L_1$ and $L_2$ (size: $c$)
  - $L_1$: blocks have been referenced only once.
  - $L_2$: blocks have been accessed at least twice.
Energy Evaluation

• I/O Traces
  – Five traces

• Simulation Framework
# I/O Traces

## Summary of I/O Traces

<table>
<thead>
<tr>
<th>Trace</th>
<th>Data Size (GB)</th>
<th>Working Set (GB)</th>
<th>Run Time (Sec)</th>
<th>Num. of Req.</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>334.96</td>
<td>80</td>
<td>265</td>
<td>5,488,106</td>
</tr>
<tr>
<td>m1</td>
<td>454.30</td>
<td>8.6</td>
<td>265</td>
<td>456,005</td>
</tr>
<tr>
<td>ior2</td>
<td>32.28</td>
<td>16</td>
<td>24</td>
<td>528921</td>
</tr>
<tr>
<td>mpiBlast</td>
<td>1.12</td>
<td>0.9</td>
<td>7.5</td>
<td>110,327</td>
</tr>
<tr>
<td>DB2</td>
<td>229.28</td>
<td>5.2</td>
<td>2,873</td>
<td>3,756,636</td>
</tr>
</tbody>
</table>
Simulation Framework (1/2)

• The simulation framework is composed of three major components:
  – cache simulator+ memory simulator :
    They decide the chip address.
  – disk array simulator :
    Disksim, a well validated disk array simulator
• Our memory simulator only calculates the energy of memory chip used as buffer cache.
Simulation Framework (2/2)

• The simulator adapts the RDRAM memory chips
• Each chip capacity is 32 Mbytes and can support up to 16 concurrent DMA operations (3.2 Gbytes/s).
• Using sequential first-touch policy.
In workload DB2 (1/2)

**DB2 trace**

**Hit rate**

**Runtime**

**Energy consumption**
In workload DB2 (2/2)

Percentage of energy saving by DMA overlap

Formula: $y / x$

$X$: disable DMA   $Y$: saved energy

ex: $X = 250$ mw,

Using DMA, it consumes 100 mw

$=> Y = 250 - 100 = 150$

Energy saving % = $(150/250) * 100\% = 60\%$

In workload ior2 (1/2)

ior2 trace

Hit rate

Runtime

Energy consumption
In workload ior2 (2/2)

Percentage of energy saving by DMA overlap

CDF of requests among banks (cache size: 4 Gbytes).
In workload f1 (1/2)

- **f1 trace**
- **Hit rate**
- **Runtime**
- **Energy consumption**

**Graphs:**
- [f1 trace graph](image-url)
- [Hit rate graph](image-url)
- [Runtime graph](image-url)
- [Energy consumption graph](image-url)
In workload f1 (2/2)

Percentage of energy saving by DMA overlap

CDF of requests among banks (cache size: 128 Mbytes).
In workload m1 (1/2)

- m1 trace
- Hit rate
- Runtime
- Energy consumption

Graphs showing performance metrics for different cache sizes and algorithms, including LIRS.
In workload m1 (2/2)

Percentage of energy saving by DMA overlap

CDF of requests among banks (cache size: 30 Gbytes).
In workload mpiBLAST (1/2)

mpiBLAST trace

Hit rate

Runtime

Energy consumption
In workload mpiBLAST (2/2)

Percentage of energy saving by DMA overlap

CDF of requests among banks (cache size: 0.75 Gbytes).

LIRS

LIRS
Comparison of Clustering Capabilities

- Power consumption rate of memory chips under workload f1 (cache size = 128Mbytes)
Seq. Placement versus Random Placement

- Average energy consumption ratios of seq. placement to random placement under the same configuration.
Conclusions

• Under the same workload, the interplay among the following three important factors appears to be the most important:
  – the cache performance in terms of **hit rates**,  
  – the cache’s ability to **temporally** align memory accesses to the same chips,  
  – and the cache populating schemes to **allocate** buffers.