Executive Summary

- Main memory is a limited shared resource
- Observation: Significant data redundancy
- Idea: Compress data in main memory
- Problem: How to avoid latency increase?
- Solution: Linearily Compressed Pages (LCP): fixed-size cache line granularity compression

1. Increases capacity (62% on average)
2. Decreases bandwidth consumption (24%)
3. Improves overall performance (9.5%)

Challenges in Main Memory Compression

LCP Overview

- Page Table entry extension
  compression type, size, and extended physical base address
- Operating System management support
  4 memory pools (512B, 1kB, 2kB, 4kB)
- Changes to cache tagging logic
  physical page base address + cache line index (within a page)
- Handling page overflows
- Compression algorithms: BDI [2], FPC [3]

LCP Optimizations

- Metadata cache
  Avoids additional requests to metadata
- Memory bandwidth reduction
  4 memory transfers needed
- Zero pages and zero cache lines
  Handled separately in TLB (1-bit) and metadata (1-bit per line)

Key Results: Compression Ratio, Bandwidth, Performance

<table>
<thead>
<tr>
<th>SPEC2006, databases, web workloads, L2 2MB cache</th>
<th>Baseline</th>
<th>RMC-FPC</th>
<th>LCP-FPC</th>
<th>LCP-BDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression Ratio GeoMean</td>
<td>1.00</td>
<td>0.79</td>
<td>0.80</td>
<td>0.76</td>
</tr>
<tr>
<td>Normalized BPKI GeoMean</td>
<td>1.00</td>
<td>0.79</td>
<td>0.80</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Average performance improvement:

<table>
<thead>
<tr>
<th>Cores</th>
<th>LCP-FPC</th>
<th>LCP-BDI</th>
<th>(BDI, LCP-BDI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.0%</td>
<td>6.1%</td>
<td>9.5%</td>
</tr>
<tr>
<td>2</td>
<td>9.3%</td>
<td>13.9%</td>
<td>23.7%</td>
</tr>
<tr>
<td>4</td>
<td>7.8%</td>
<td>10.7%</td>
<td>22.6%</td>
</tr>
</tbody>
</table>

References

[1] M. Ekman and P. Stenstrom. A Robust Main Memory Compression Scheme, ISCA'05