Linearly Compressed Pages: A Low Complexity, Low Latency Main Memory Compression Framework

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: Linearly Compressed Pages (LCP): fixed-size cache line granularity compression
  1. Increases capacity (62% on average)
  2. Decreases bandwidth consumption (24%)
  3. Improves overall performance (13.9%)
  4. Decreases memory energy consumption (9.5%)

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)
- Handling page overflows
- Hardware support
- Compression algorithms:
  Base-Delta-Immediate (BDI) and Frequent Pattern Compression (FPC)

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

Key Results: Compression Ratio, Performance, Page Faults

Challenge in Memory Compression

Uncompressed Page
Address Offset
0 64 128 (N-1)*64

Compressed Page
Address Offset
0 ? ? ?

Challenge: Address Computation

Linearly Compressed Pages (LCP)

Uncompressed Page (4KB: 64*64B)

4:1 Compression

Exception Storage

Compressed Data (1KB)

Metadata (64B):
? (compressible) and ? (zero cache line)

Methodology

Evaluated designs

No. | Label       | Description
---|-------------|-------------
1  | Baseline    | Baseline (no compression)
2  | RMC-FPC     | Main memory compression using RMC and FPC
3  | LCP-FPC     | LCP framework with FPC
4  | LCP-BDI     | LCP framework with BDI
5  | MXT         | IBM MXT design

Challenges in Memory Compression

Uncompressed Page

\[
L_0 \quad L_1 \quad L_2 \quad \ldots \quad L_{N-1}
\]

Address Offset

0 64 128 (N-1)*64

Compressed Page

\[
L_0 \quad L_1 \quad L_2 \quad \ldots \quad L_{N-1}
\]

Address Offset

0 ? ? ?

Linearly Compressed Pages: A Low Complexity, Low Latency Main Memory Compression Framework

G. Pekhimenko, V. Seshadri, Y. Kim, H. Xin, O. Mutlu, T. C. Mowry (CMU), P. B. Gibbons, M. A. Kozuch (Intel Labs)  
MICRO 2013, Davis, CA