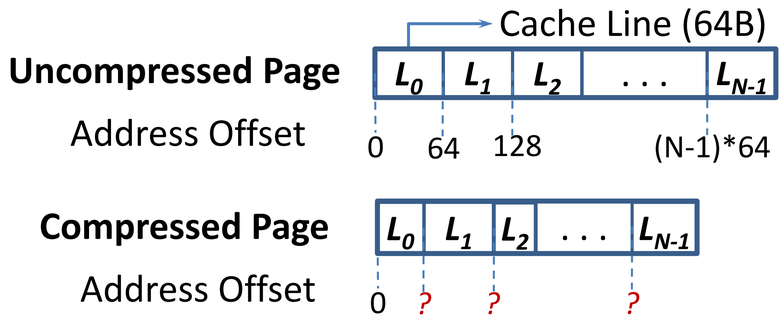
Linearly Compressed Pages: A Low Complexity, Low Latency Main Memory Compression Framework

Summary

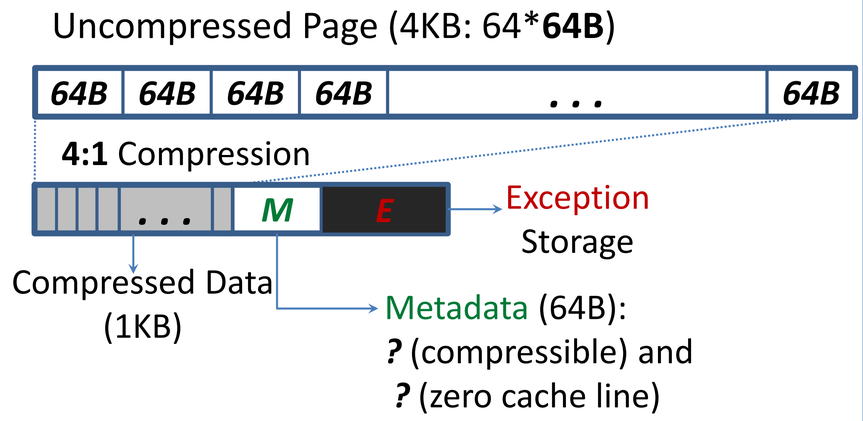
- Main memory is a limited shared resource
- **Observation**: Significant data redundancy
- **Idea**: Compress data in main memory
- **Problem**: How to avoid latency increase?
- **Solution**: Linearly Compressed Pages (LCP): fixed-size cache line granularity compression
  1. Increases capacity (**62%** on average)
  2. Decreases bandwidth consumption (**24%**)
  3. Improves overall performance (**13.9%**)
  4. Decreases memory energy consumption (**9.5%**)

Challenge in Memory Compression



Challenge: Address Computation

Linearly Compressed Pages (LCP)

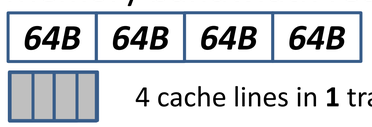


LCP Overview

- 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



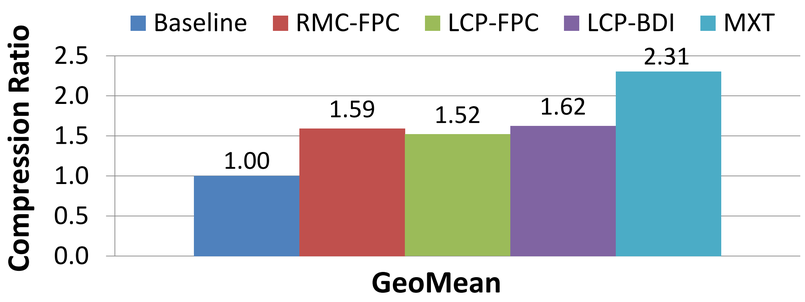
- Zero pages and zero cache lines
  Handled separately in TLB (1-bit) and metadata (1-bit per line)

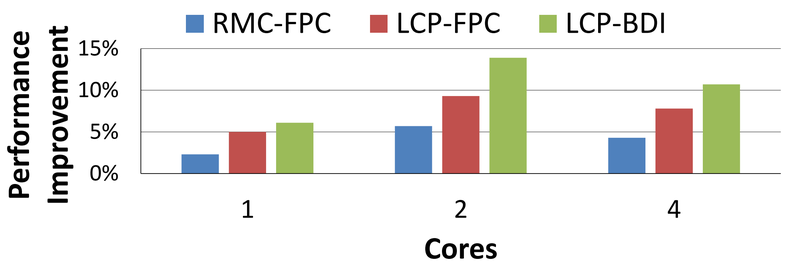
Methodology

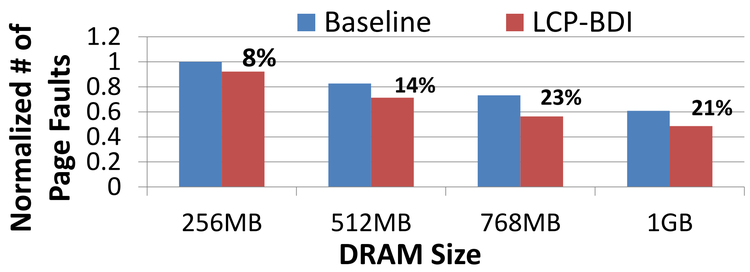
Evaluated designs

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

Key Results: Compression Ratio, Performance, Page Faults







SPEC2006, databases, web workloads, L2 2MB cache

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

Carnegie Mellon University

intel