

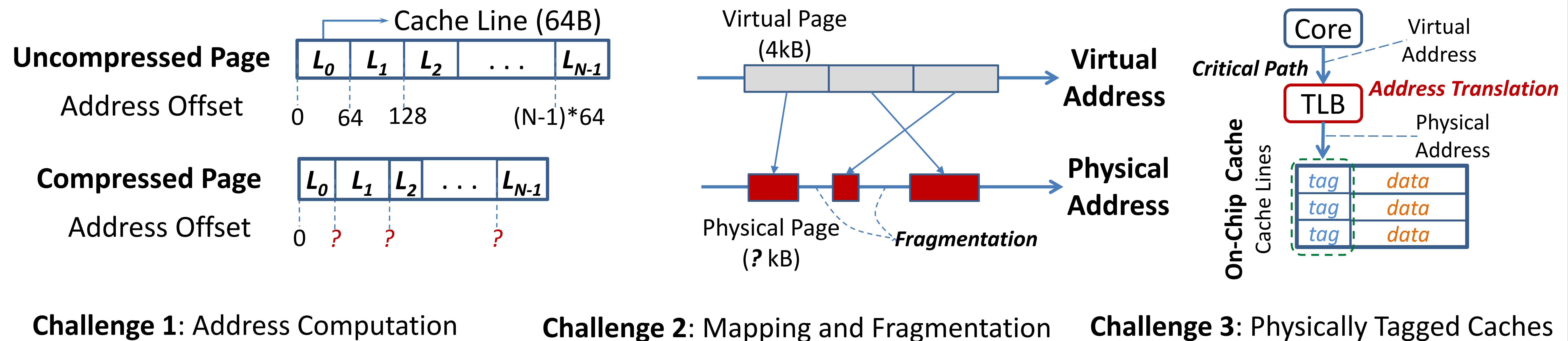
Linearly Compressed Pages: A Main Memory Compression Framework with Low Complexity and Low Latency

Gennady Pekhimenko, Advisers: Todd C. Mowry and Onur Mutlu (Carnegie Mellon University)

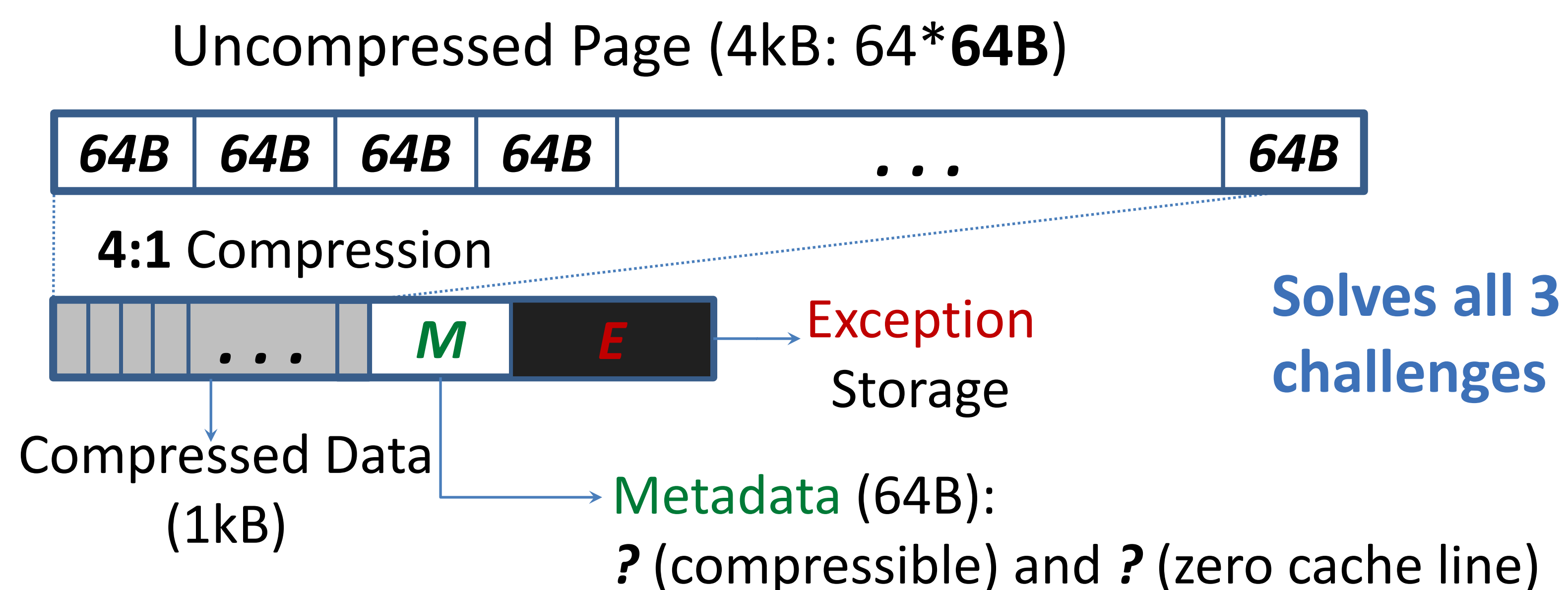
Executive Summary

- Main memory is a limited shared resource
- Observation:** Significant data redundancy
- Idea:** Compress data in main memory
- Problem:** How to avoid latency increase?
- Solution:** Linearly Compressed Pages (LCP): fixed-size cache line granularity compression
 - Increases capacity (**69%** on average)
 - Decreases bandwidth consumption (**46%**)
 - Improves overall performance (**9.5%**)

Challenges in Main Memory Compression



Linearly Compressed Pages (LCP): Key Idea



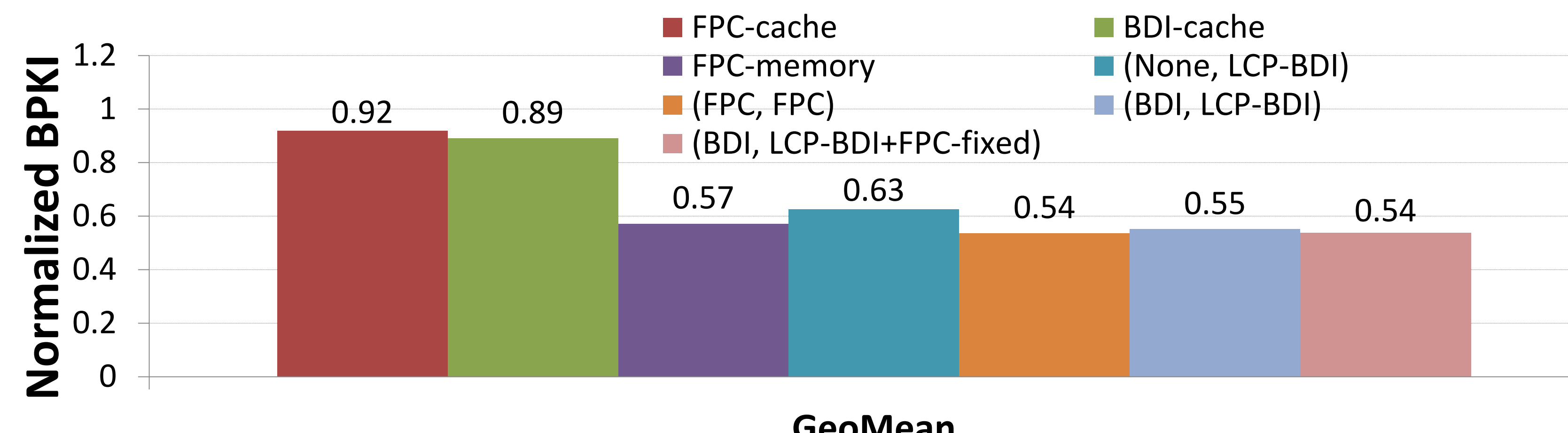
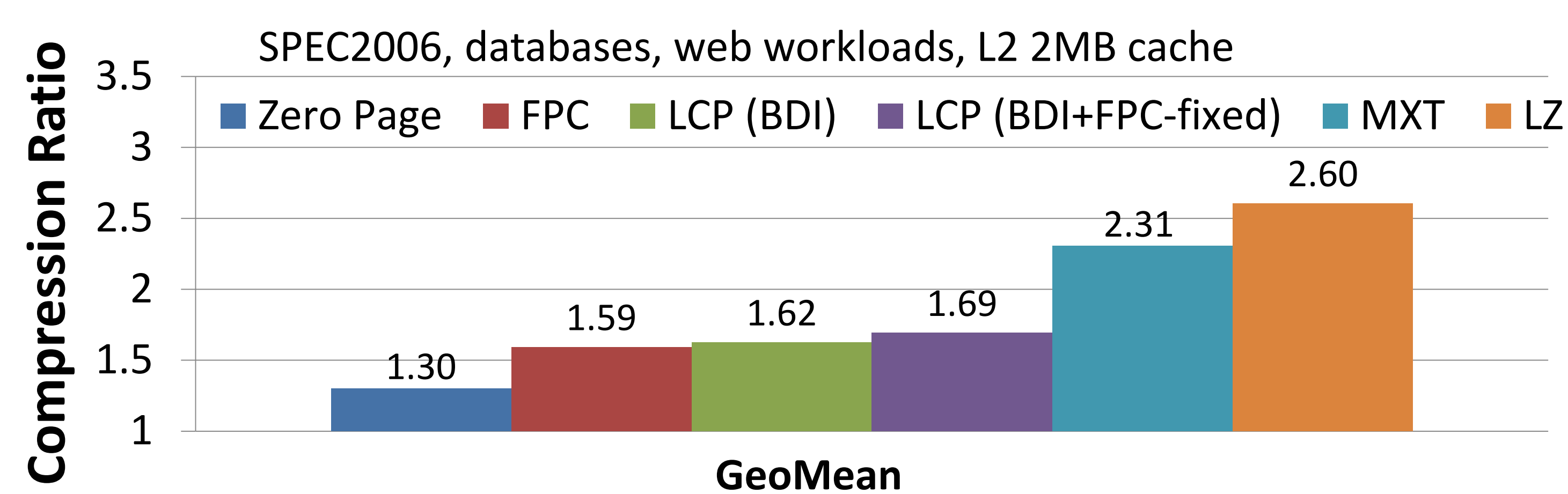
LCP Overview

- 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 for 4 cache lines in 1 transfer
- Zero pages and zero cache lines**: Handled separately in TLB (1-bit) and metadata (1-bit per line)

Key Results: Compression Ratio, Bandwidth, Performance



Average performance improvement:

Cores	LCP-BDI	(BDI, LCP-BDI)	(BDI, LCP-BDI+FPC-fixed)
1	6.1%	9.5%	9.3%
2	13.9%	23.7%	23.6%
4	10.7%	22.6%	22.5%

Evaluated designs

No.	Label: (Cache, Memory)	Description
1	(None, None)	Baseline (no compression)
2	FPC-Cache	LLC compression using FPC [3]
3	BDI-Cache	LLC compression using BDI [2]
4	FPC-Memory	Main memory compression using [1]
5	LCP-BDI	LCP-framework with BDI
6	(FPC, FPC)	Designs 2 and 4 combined
7	(BDI, LCP-BDI)	Designs 3 and 5 combined
8	(BDI, LCP-BDI+FPC-fixed)	Design 3 combined with BDI+FPC-fixed

References

- [1] M. Ekman and P. Stenstrom. A Robust Main Memory Compression Scheme, *ISCA'05*
- [2] G. Pekhimenko et al., Base-Delta-Immediate Compression: Practical Data Compression for On-Chip Caches, *PACT'12*
- [3] A. Alameldeen and D. Wood. Adaptive Cache Compression for High-Performance Processors, *ISCA'04*
- [4] B. Abali et al., Memory expansion technology (MXT): software support and performance. *IBM J.R.D.* '01