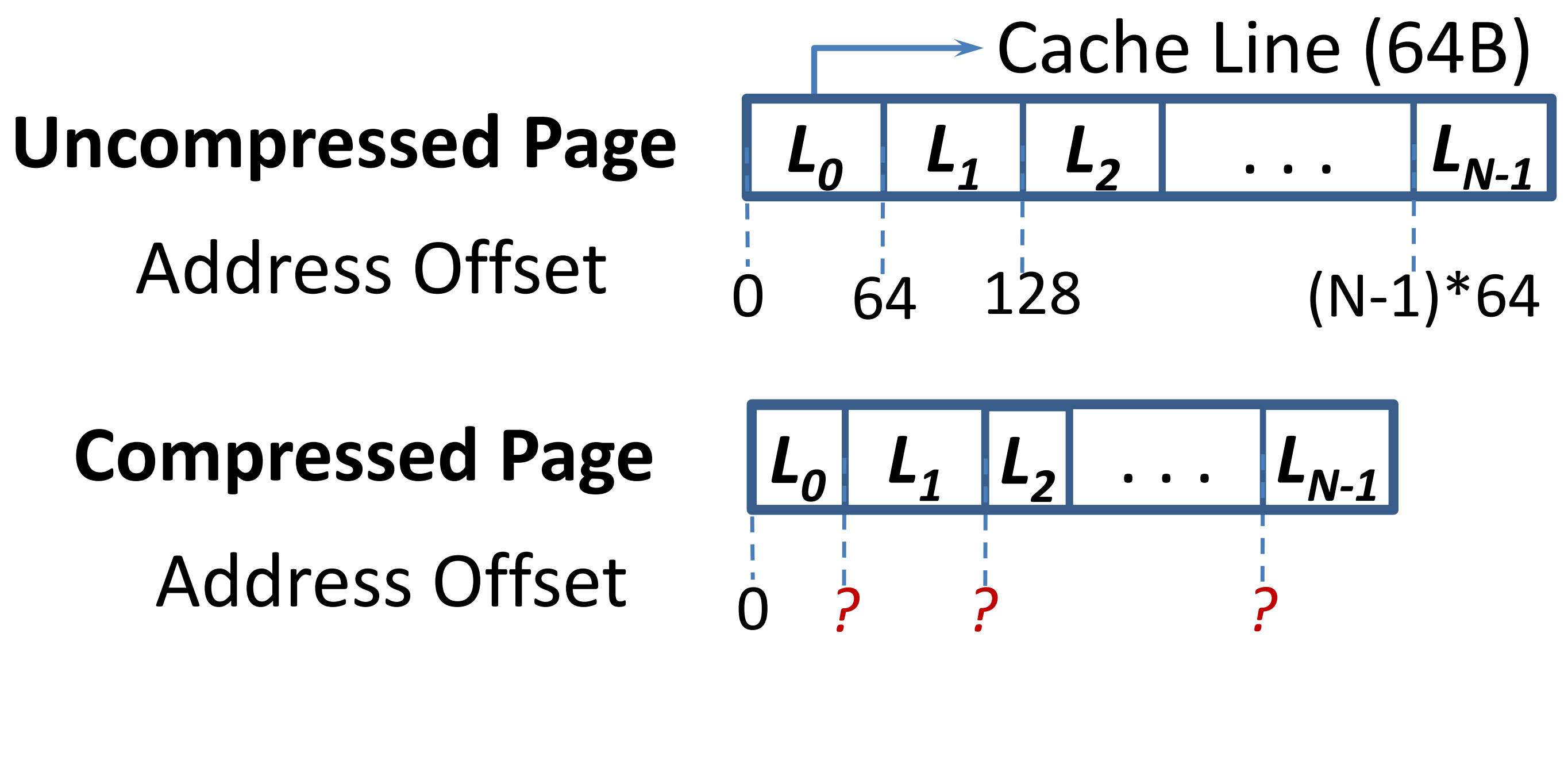
Linearly Compressed Pages: A Low Complexity, Low Latency Main Memory Compression Framework

Gennady Pekhimenko, Vivek Seshadri, Yoongu Kim, Hongyi Xin, Onur Mutlu, Todd C. Mowry (CMU), Phillip B. Gibbons, Michael A. Kozuch (Intel)

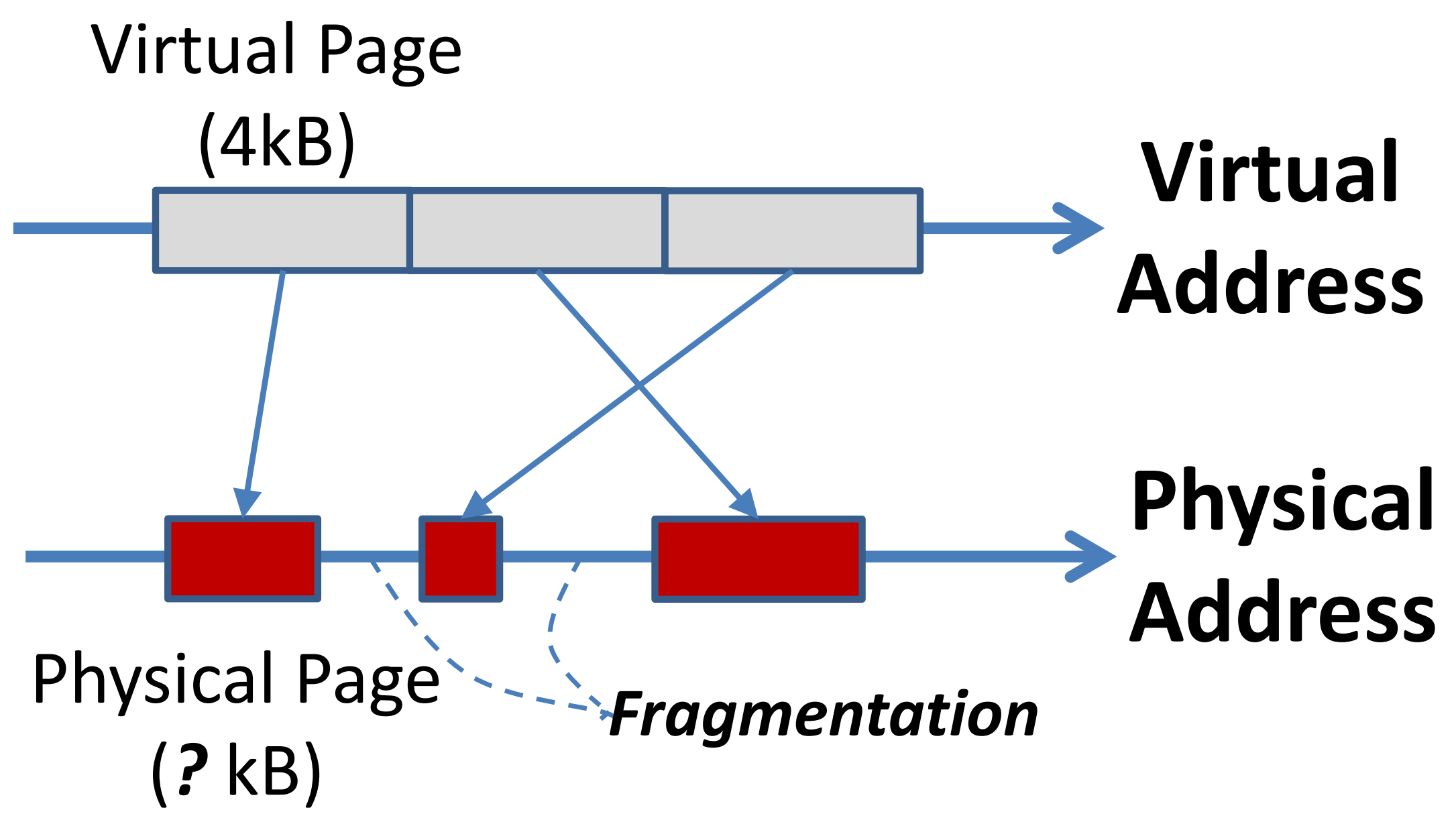
Executive Summary

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

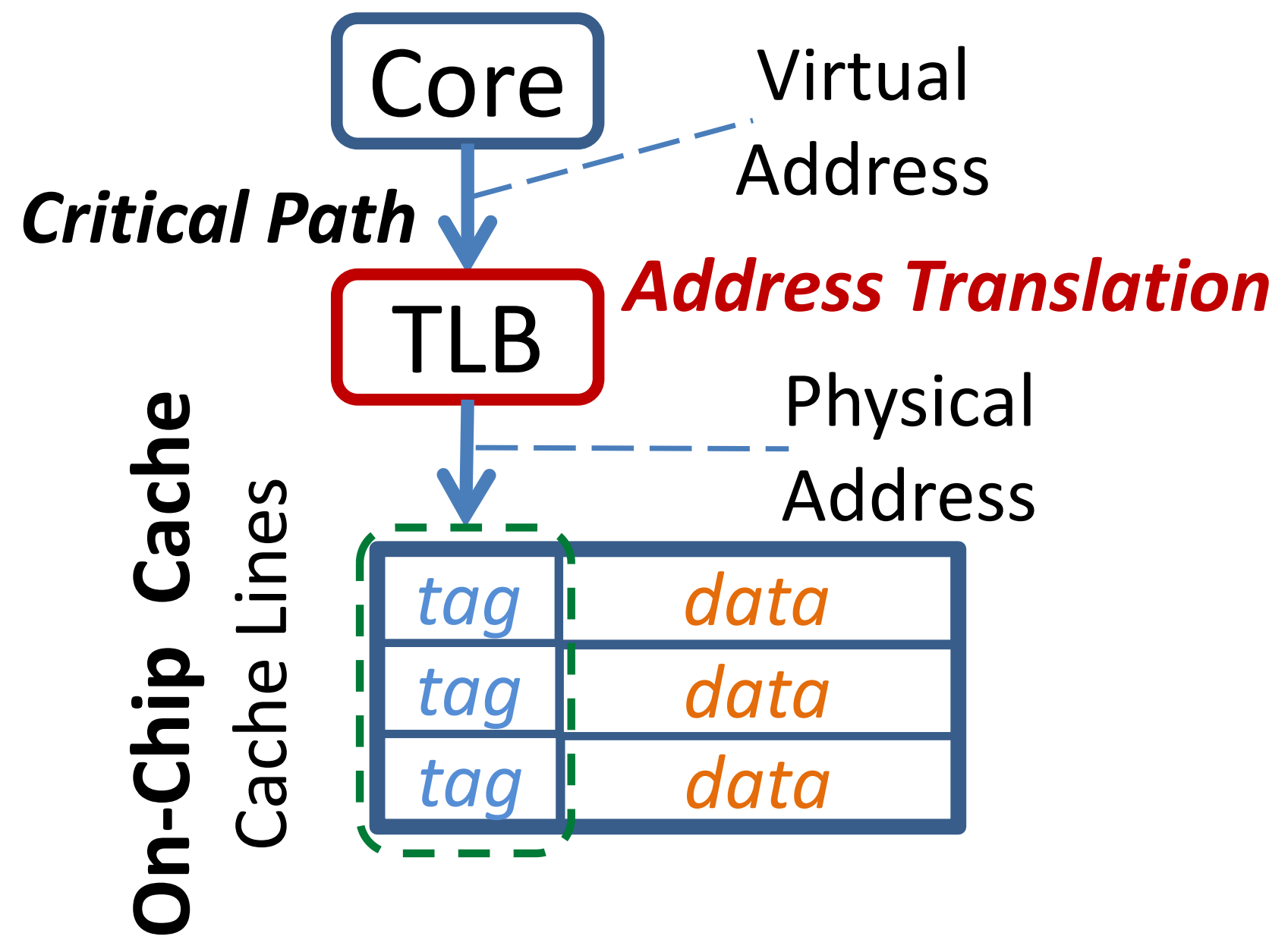
Challenges in Main Memory Compression



Challenge 1: Address Computation

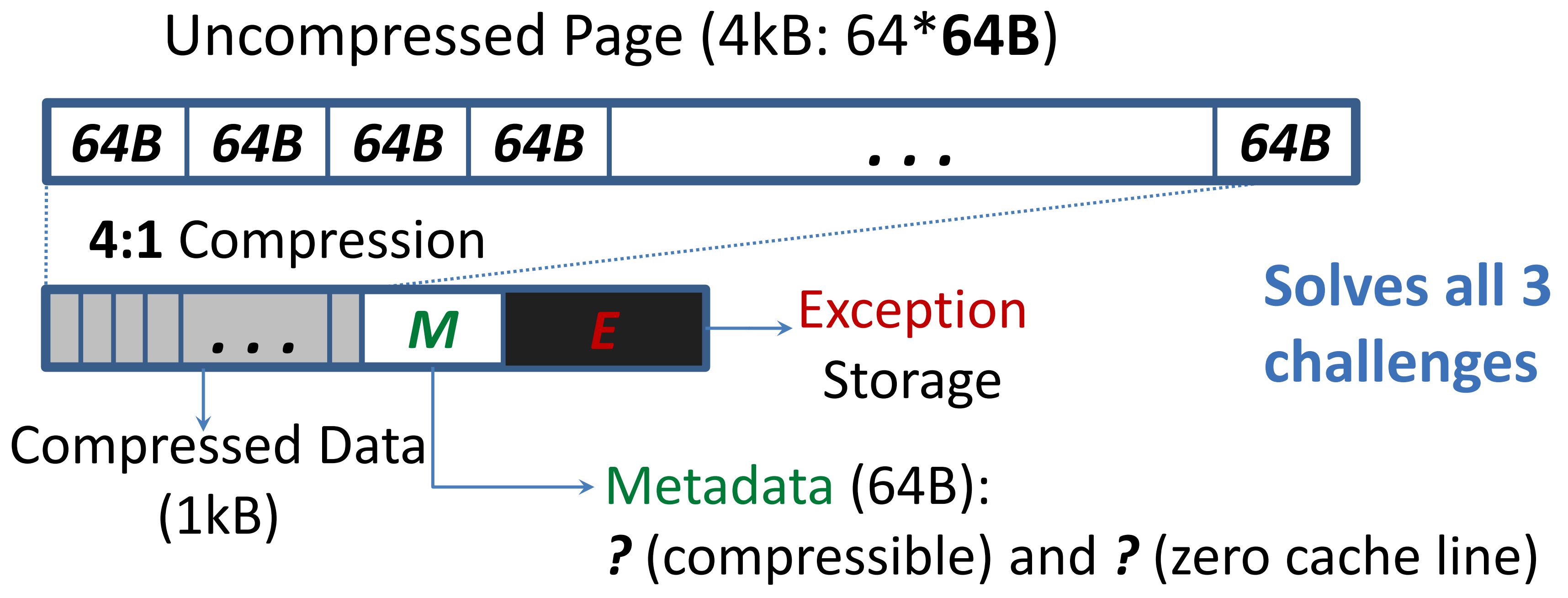


Challenge 2: Mapping and Fragmentation



Challenge 3: Physically Tagged Caches

Linearly Compressed Pages (LCP): Key Idea


Uncompressed Page (4kB: 64*64B)
64B 64B 64B 64B . . . 64B
4:1 Compression
. . .
M
E
Exception
Storage
Compressed Data
(1kB)
Metadata (64B):
? (compressible) and ? (zero cache line)

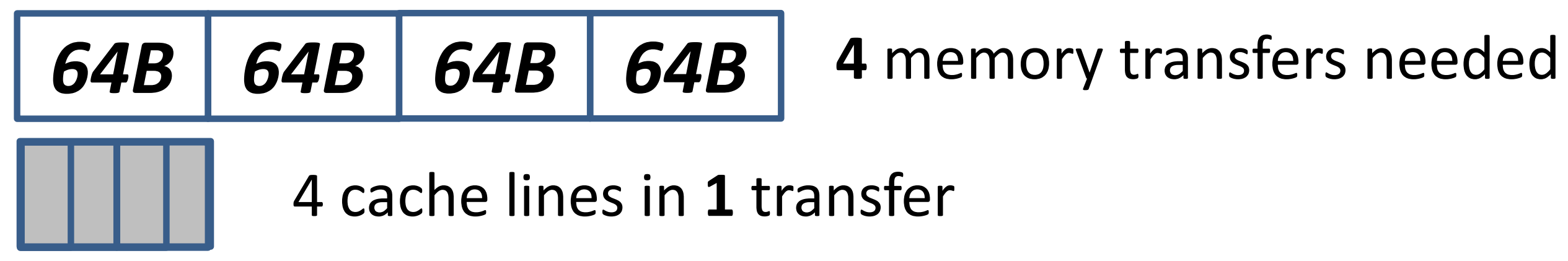
Solves all 3 challenges

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

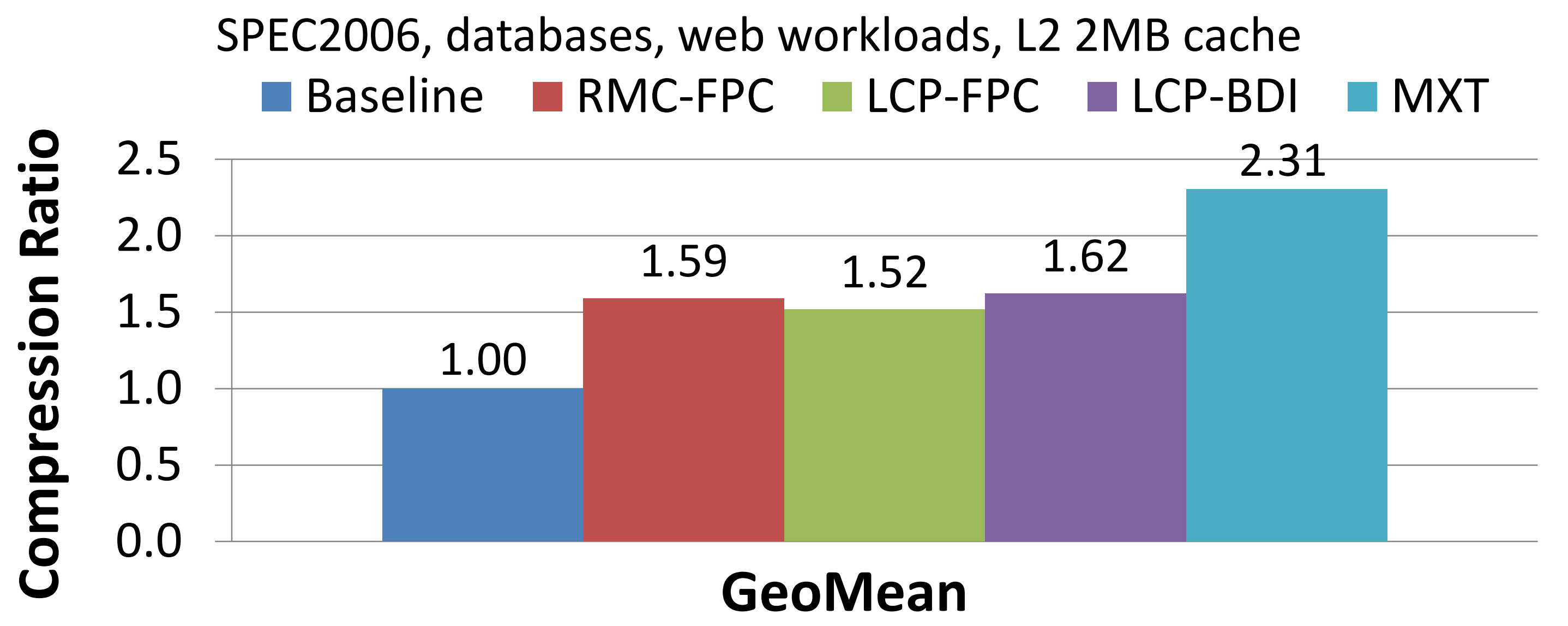
- Metadata cache
  Avoids additional requests to metadata
- Memory bandwidth reduction



- Zero pages and zero cache lines

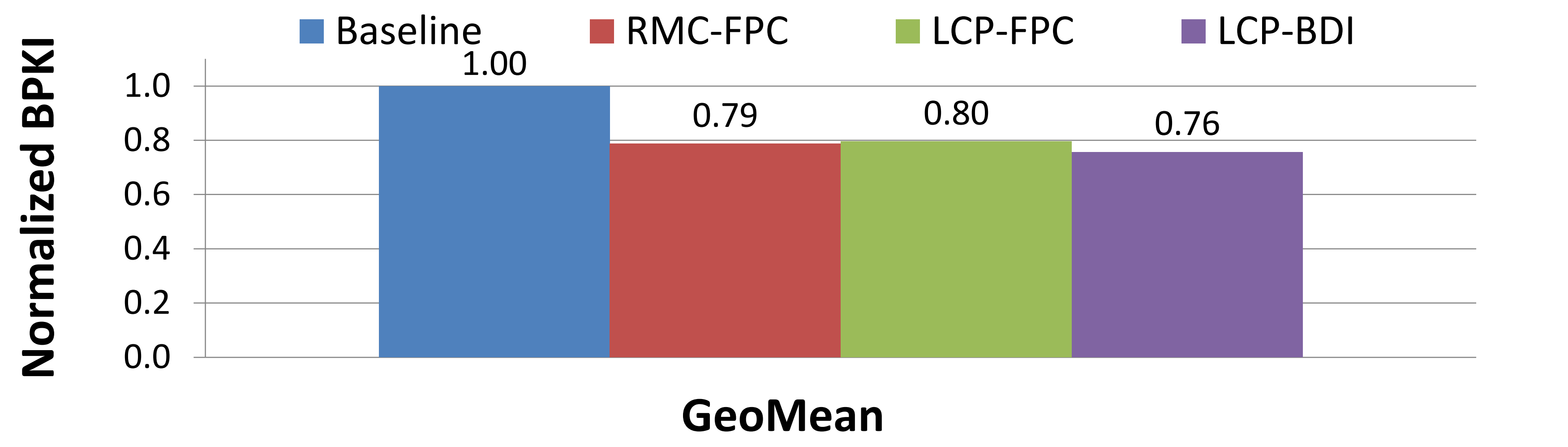
Handled separately in TLB (1-bit) and metadata (1-bit per line)

Key Results: Compression Ratio, Bandwidth, Performance



Average **performance** improvement:

Cores	LCP-FPC	LCP-BDI	(BDI, LCP-BDI)
1	5.0%	6.1%	**9.5%**
2	9.3%	13.9%	**23.7%**
4	7.8%	10.7%	**22.6%**



Evaluated designs

No.	Label	Description
1	Baseline	Baseline (no compression)
2	RMC-FPC	Main memory compression using [1] and FPC [3]
3	LCP-FPC	LCP framework with FPC [3]
4	LCP-BDI	LCP framework with BDI [2]
5	(BDI, LCP-BDI)	Design 4 plus cache compression with BDI [2]
6	MXT	IBM MXT design [4]

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*

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University

Georgia Tech

intel

PRINCETON UNIVERSITY

UC Berkeley

UNIVERSITY of WASHINGTON