Base-Delta-Immediate Compression: Practical Data Compression for On-Chip Caches

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

Motivation & Background

Significant redundancy in data:

0x**00000000**	0x**0000000B**	0x**00000003**	0x**00000004**	…

Cache compression provides effect of a larger cache without making it physically larger

Key requirements:

- **Fast** (low decompression latency)
- **Simple** (avoid complex hardware changes)
- **Effective** (good compression ratio)

Key Data Patterns

Zero Values: initialization, sparse matrices

0x00000000	0x00000000	0x00000000	0x00000000	…

Repeated Values: common initial values

0x000000**FF**	0x000000**FF**	0x000000**FF**	0x000000**FF**	…

Narrow Values: small values in a big data type

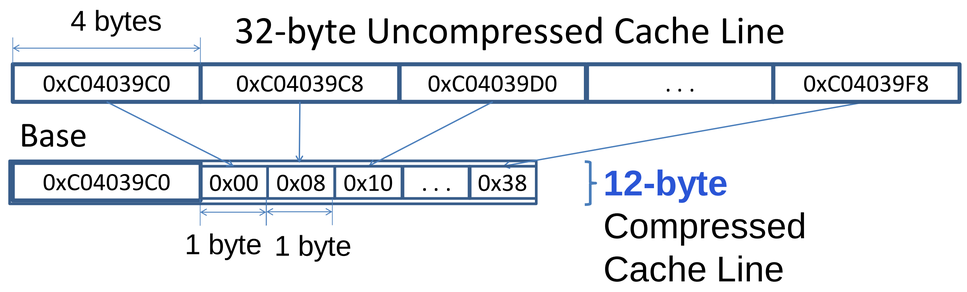
0x00000**00**	0x00000**0B**	0x00000**03**	0x00000**04**	…

Other Patterns: pointers (same memory region)

0xC04039**C0**	0xC04039**C8**	0xC04039**D0**	0xC04039**D8**	…

Low Dynamic Range

Base+Delta (B+Δ) Encoding: Key Idea



20 bytes space saved

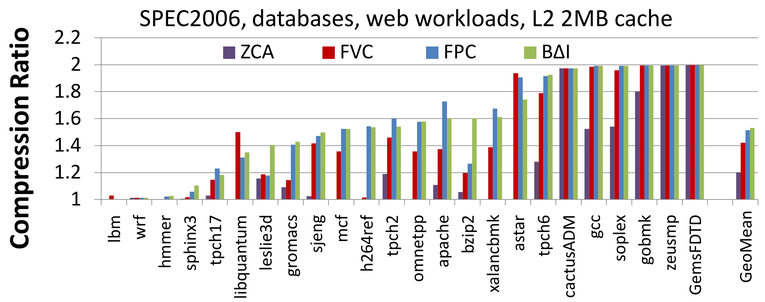
Base-Delta-Immediate Compression

- Use multiple bases to increase compression coverage
- Pro: More cache lines can be compressed
- Cons: 1. Unclear how to find, 2. Higher overhead
- Empirically 2 bases is the best for our set of applications
- Idea:
  1. **First** base – first element in the cache lines (**base+Δ**)
  2. **Second** base – implicit base of **0** (**immediate**)

BΔI Implementation

- Decompressor design: vector addition (fast)
- Compressor design
  - arithmetic (+/-) and comparisons
- BΔI cache organization
  - 2X tags + compr. encoding
  - Data segmenting (e.g., 8-byte)

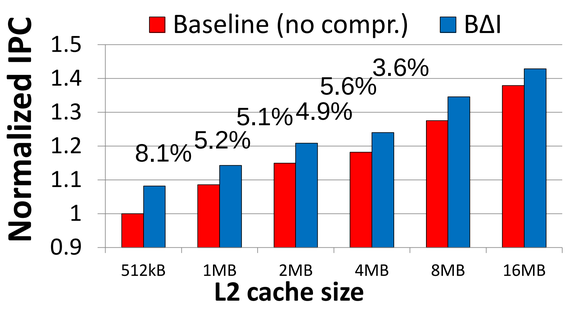
Key Results: Compression Ratio

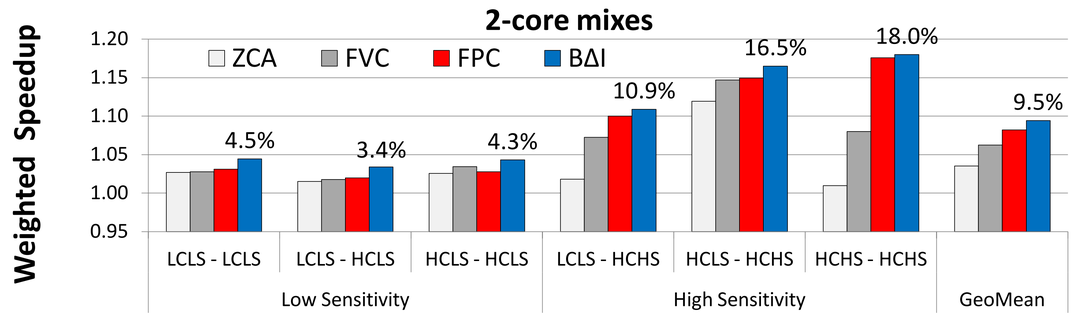


Key Results: Performance

BΔI [4] performance over other mechanisms:

Cores	No Compression	ZCA[3]	FPC [1]	FVC [2]
1	5.1%	4.1%	2.1%	1.0%
2	9.5%	5.7%	3.1%	1.2%
4	11.2%	5.6%	3.2%	1.3%





[1] A. Alameldeen and D. Wood. Adaptive Cache Compression for High-Performance Processors, *ISCA'04*

[2] J. Yang et al., Frequent value compression in data caches. *Micro'00*

[3] J. Dusser et al., Zero-content augmented caches, *ICS'09*

[4] G. Pekhimenko et al., Base-Delta-Immediate Compression: Practical Data Compression for On-Chip Caches, *PACT'12*

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley · University of Washington