**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:

```
0x00000000 0x00000000 0x00000000 0x00000000 ...
```

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)

**Base+Delta (B+Δ) Encoding: Key Idea**

4 bytes 32-byte Uncompressed Cache Line

```
0xC04039C0 0xC04039C8 0xC04039D0 ...
```

Baseline

```
0xC04039C0 0x00 0x08 0x10 ... 0x38
```

12-byte Compressed Cache Line

20 bytes space saved

**Base-Delta-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: Performance**

BΔI [4] Performance over Other Mechanisms:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.1%</td>
<td>4.1%</td>
<td>2.1%</td>
<td>1.0%</td>
</tr>
<tr>
<td>2</td>
<td>9.5%</td>
<td>5.7%</td>
<td>3.1%</td>
<td>1.2%</td>
</tr>
<tr>
<td>4</td>
<td>11.2%</td>
<td>5.6%</td>
<td>3.2%</td>
<td>1.3%</td>
</tr>
</tbody>
</table>

**Key Data Patterns**

- **Zero Values**: initialization, sparse matrices
- **Repeated Values**: common initial values
- **Narrow Values**: small values in a big data type

**Key Results: Compression Ratio**

- Use multiple bases to increase compression coverage
- **Pros**: 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
  - 1. First base – first element in the cache lines (base+Δ)
  - 2. Second base – implicit base of 0 (immediate)

![Image of compression ratio graph]


![Image of practical data compression chart]