Energy-Efficient Data Compression for GPU Memory Systems

Gennady Pekhimenko (Advisors: Todd C. Mowry and Onur Mutlu) – ***Carnegie Mellon University***

High Performance Computing is Everywhere

***Energy efficiency* is key across the board**

Applications today are data-intensive

Memory systems are ***bandwidth constrained***

***Data Compression** is a promising technique to address these challenges*

Potential for HW-Based Data Compression

Multiple simple patterns: zeros, repeated values, narrow values, pointers (**low dynamic range**)

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

Different Compression Algorithms:

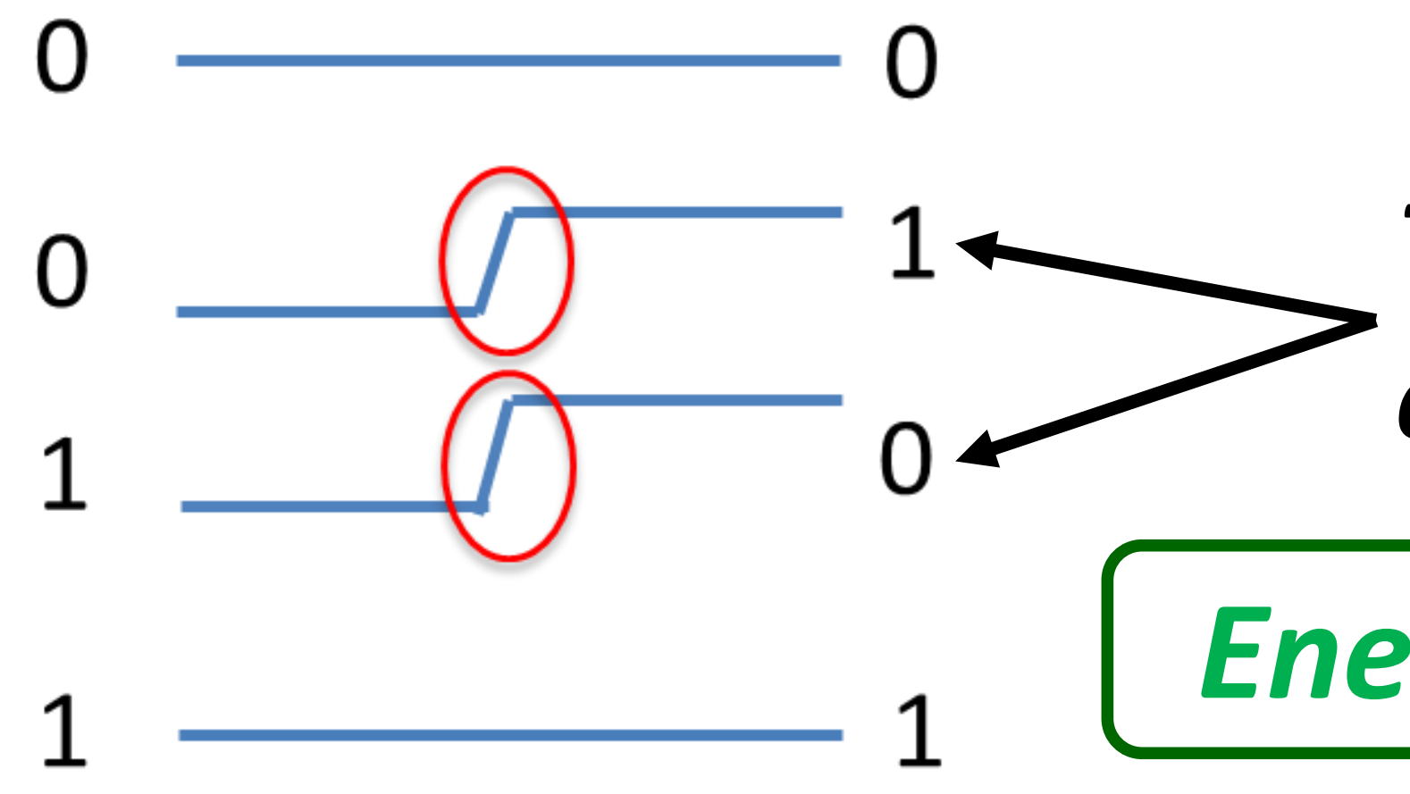
- **BΔI** *[PACT'12]* is based on Base-Delta Encoding
- Frequent Pattern Compression (**FPC**) *[ISCA'04]*
- **C-Pack** *[Trans. on VLSI'12]*
- Statistical Compression (**SC²**) *[ISCA'14]*

> - ***These algorithms improve performance***
> - ***But there are challenges…***

Energy Efficiency: What is a Bit "Toggle"?

How energy is spent in data transfers:

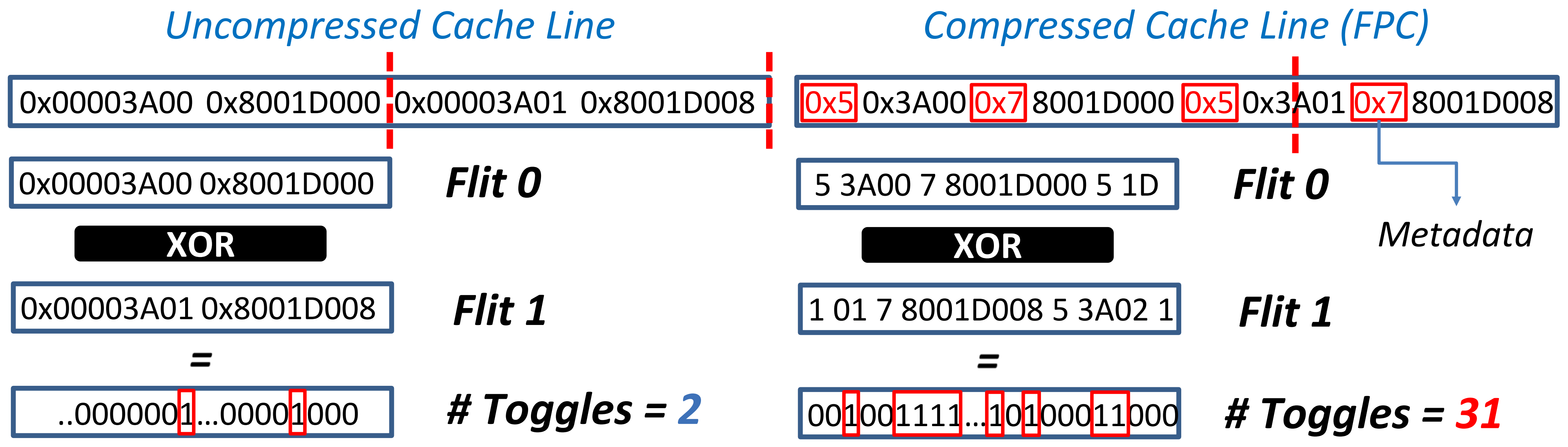
Previous data: 0011 **New data:** 0101



Toggles are expensive

***Energy** = CV² + Fixed*

Excessive Bit Toggles with Data Compression



Uncompressed Cache Line

0x00003A00 0x8001D000 0x00003A01 0x8001D008

0x00003A00 0x8001D000 ***Flit 0***

XOR

0x00003A01 0x8001D008 ***Flit 1***

=

..0000001...00001000 ***# Toggles = 2***

Compressed Cache Line (FPC)

0x5 0x3A00 0x7 8001D000 0x5 0x3A01 0x7 8001D008

Metadata

5 3A00 7 8001D000 5 1D ***Flit 0***

XOR

1 01 7 8001D008 5 3A02 1 ***Flit 1***

=

001001111...101000011000 ***# Toggles = 31***

Toggle-Aware Energy-Efficient Data Compression

Problem:

+ 1.53X effective compression ratio

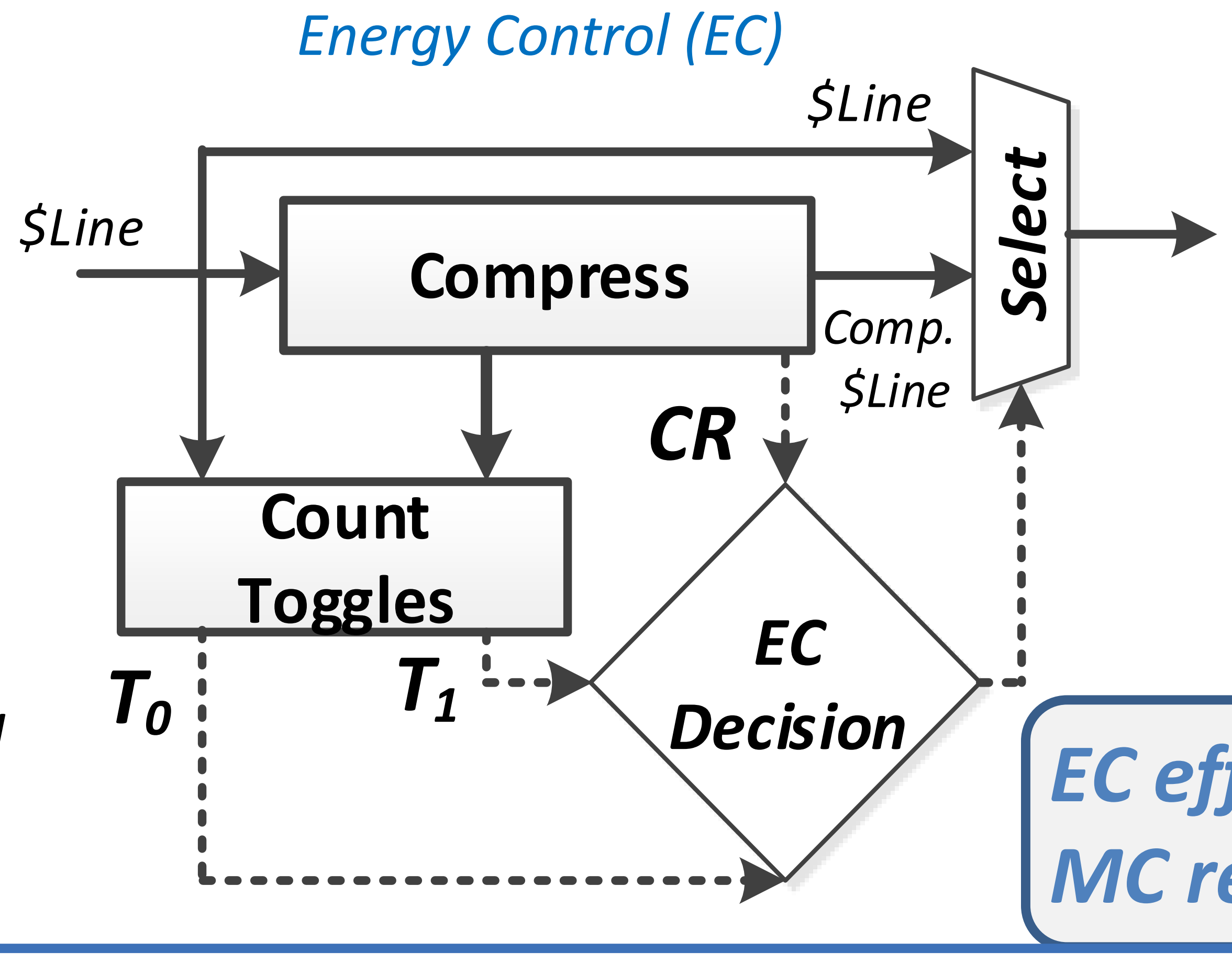
- 2.19X increase in toggle count

Goal:

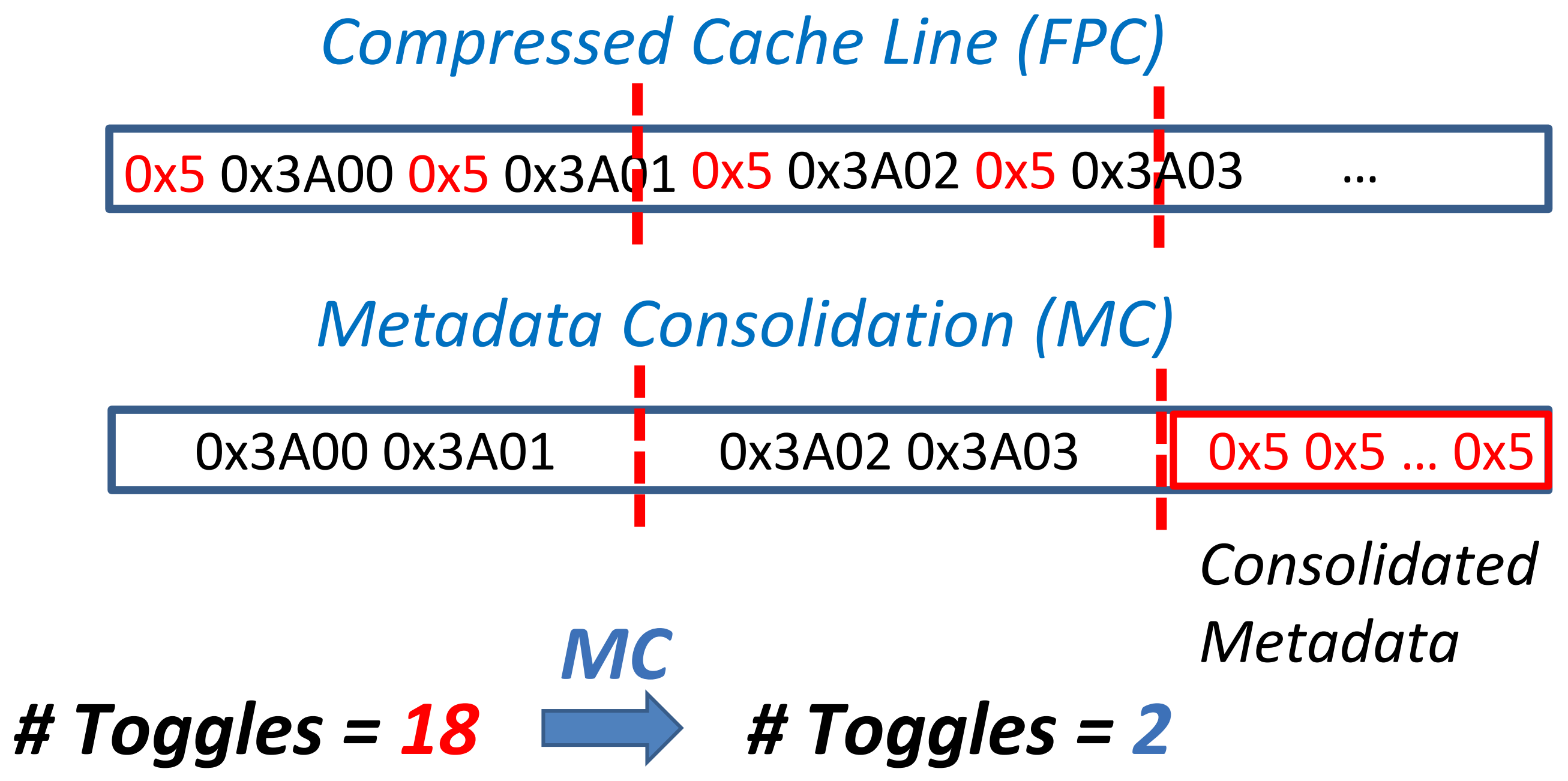
- Find the optimal tradeoff between toggle count and compression ratio

Key Idea – Energy Control (EC):

- Determine toggle count
- Use a heuristic (***Energy X Delay*** *and* ***Energy X Delay²*** metrics)
- Throttle compression when needed



Optimization: Metadata Consolidation (MC)



Compressed Cache Line (FPC)

0x5 0x3A00 0x5 0x3A01 0x5 0x3A02 0x5 0x3A03 …

Metadata Consolidation (MC)

0x3A00 0x3A01 0x3A02 0x3A03 0x5 0x5 … 0x5

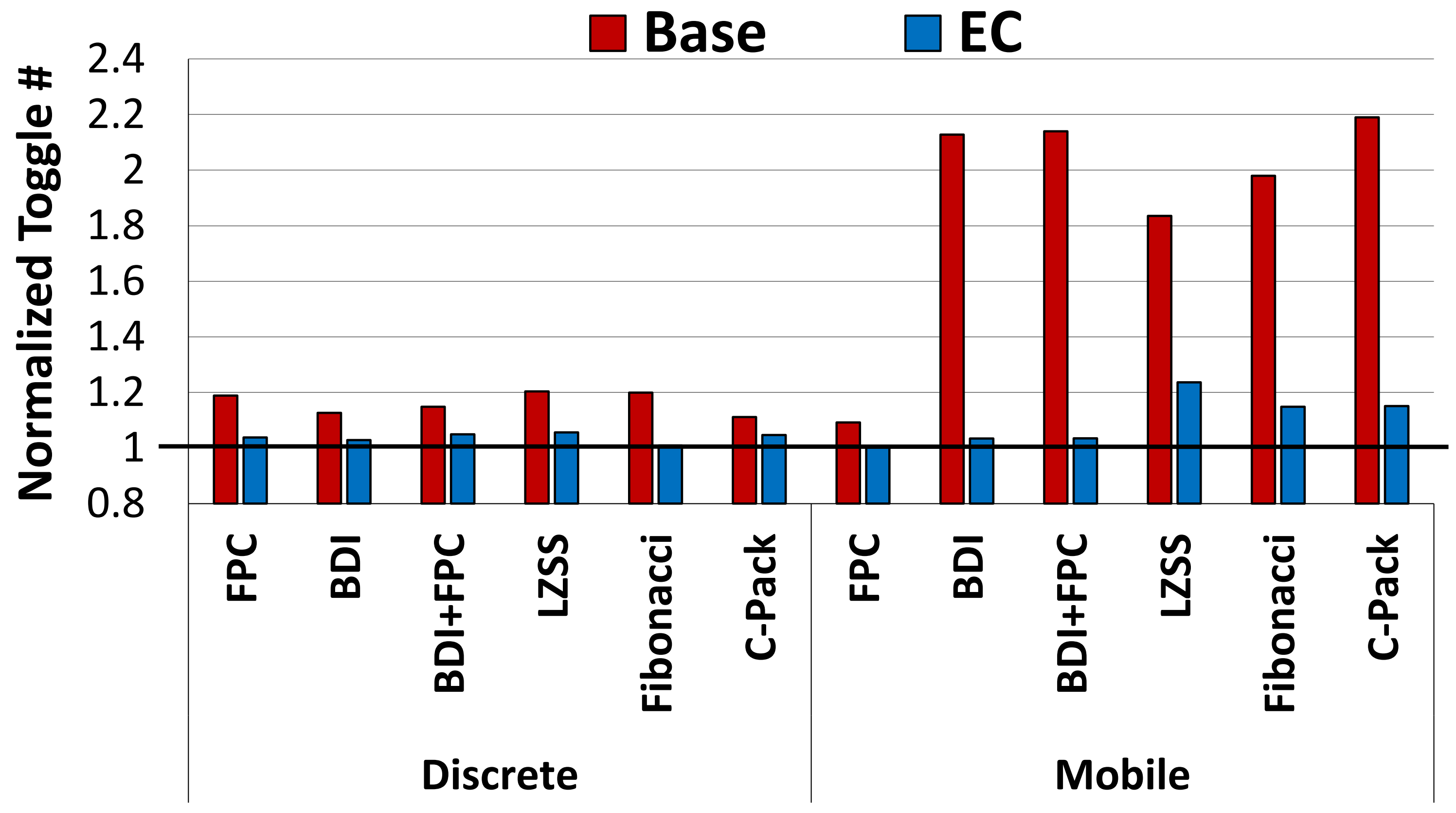
Consolidated Metadata

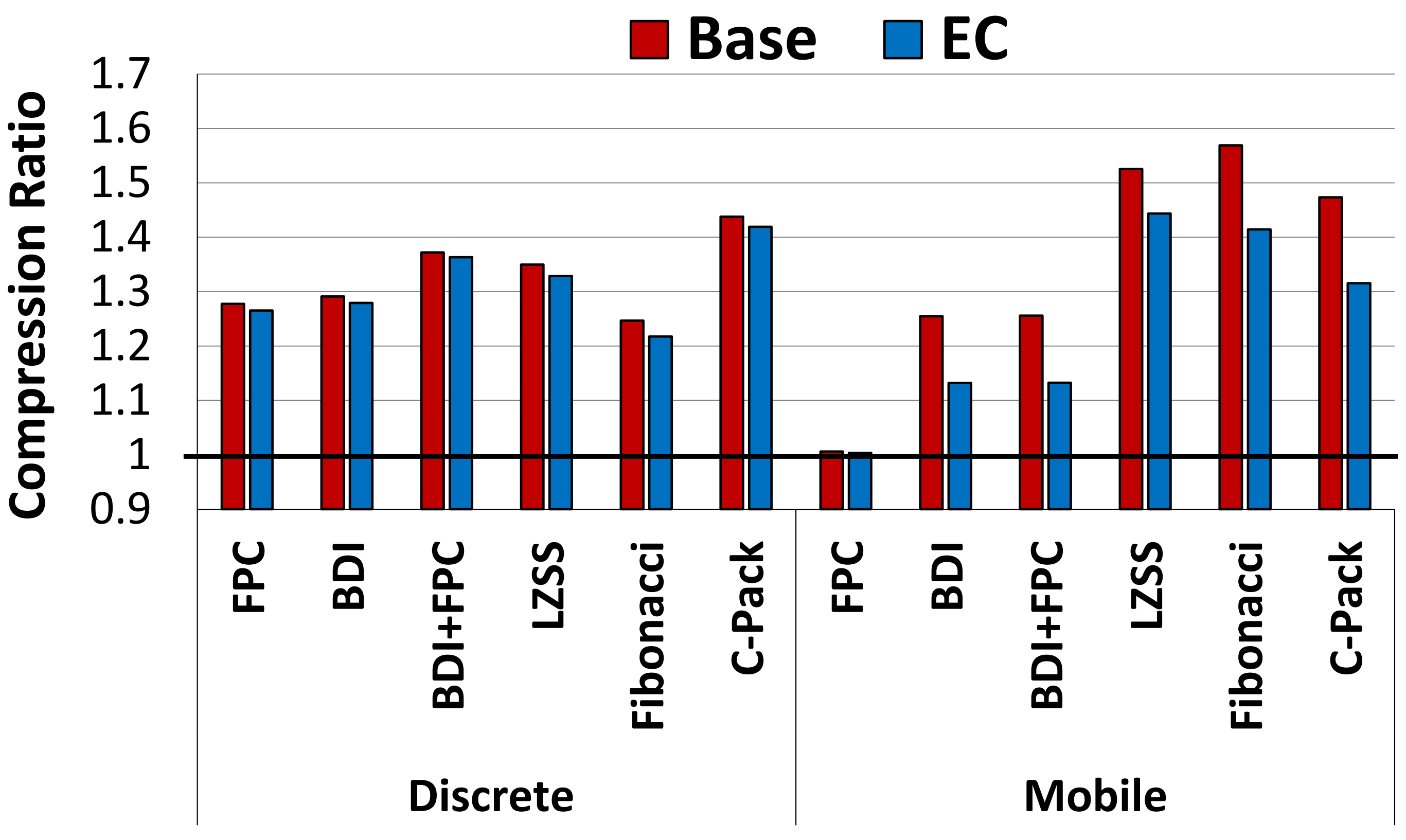
# Toggles = 18 *MC* → ***# Toggles = 2***

> ***EC efficiently trades compressibility with toggles***
> ***MC reduces toggles & preserves compression ratio***

Initial Results: Compression Ratio and Toggle Rate

Applications from NVIDIA: Mobile GPU – 54 in total, Discrete GPU – 167 in total





MC Results:

- ***3.2%/2.9%*** *reduction in toggles for FPC/C-Pack*

Future Work:

- *Detailed Power/Energy model*
- *Effect on different layers in memory hierarchy (DRAM and NoCs)*

> ***EC significantly reduces the number of toggles***

> ***EC preserves most of the compression benefits***