Energy-Efficient Data Compression for Modern Memory Systems

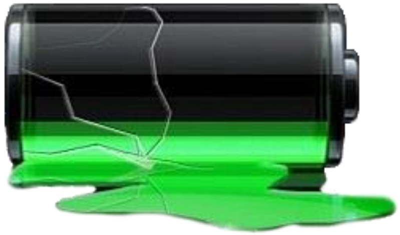
Gennady Pekhimenko

ACM Student Research Competition

March, 2015

Carnegie Mellon University

High Performance Computing Is Everywhere



Energy efficiency **is key across the board**

Applications today are data-intensive

Modern memory systems are ***bandwidth constrained***

Data Compression is a promising technique to address these challenges

Potential of Data Compression

- **Multiple simple patterns**: zeros, repeated values, narrow values, pointers

0x*C04039***C0**	0x*C04039***C8**	0x*C04039***D0**	0x*C04039***D8**	…

- **Different Algorithms:**

Low Dynamic Range:

Differences between values are significantly smaller than the values themselves

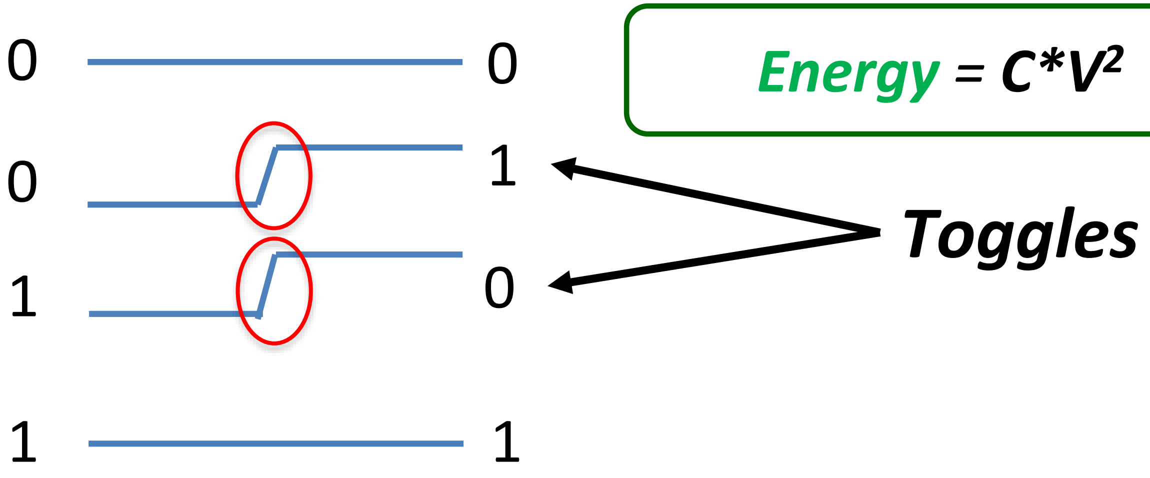
- ***These algorithms improve performance***
- ***But there are challenges…***

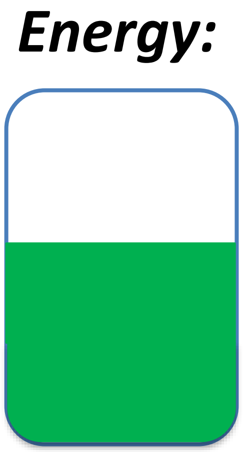
Energy Efficiency: Bit Toggles

How energy is spent in data transfers:

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

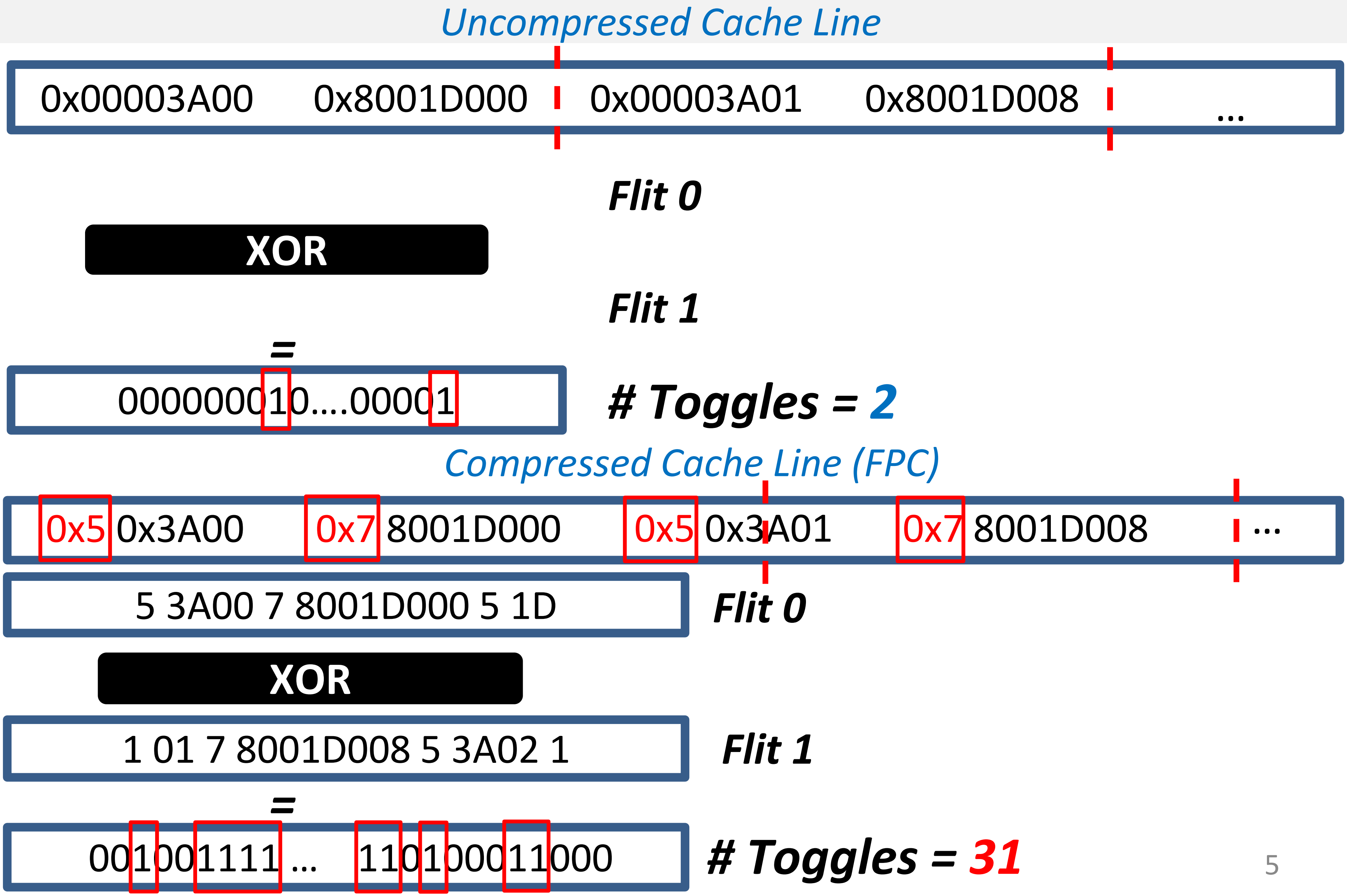
Energy = $C*V^2$





Energy of data transfers (e.g., NoC, DRAM) is proportional to the number of toggles

Excessive Number of Bit Toggles



Uncompressed Cache Line

0x00003A00 0x8001D000 0x00003A01 0x8001D008 ...

Flit 0

XOR

Flit 1

=

000000010....00001

# Toggles = 2

Compressed Cache Line (FPC)

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

5 3A00 7 8001D000 5 1D

Flit 0

XOR

1 01 7 8001D008 5 3A02 1

Flit 1

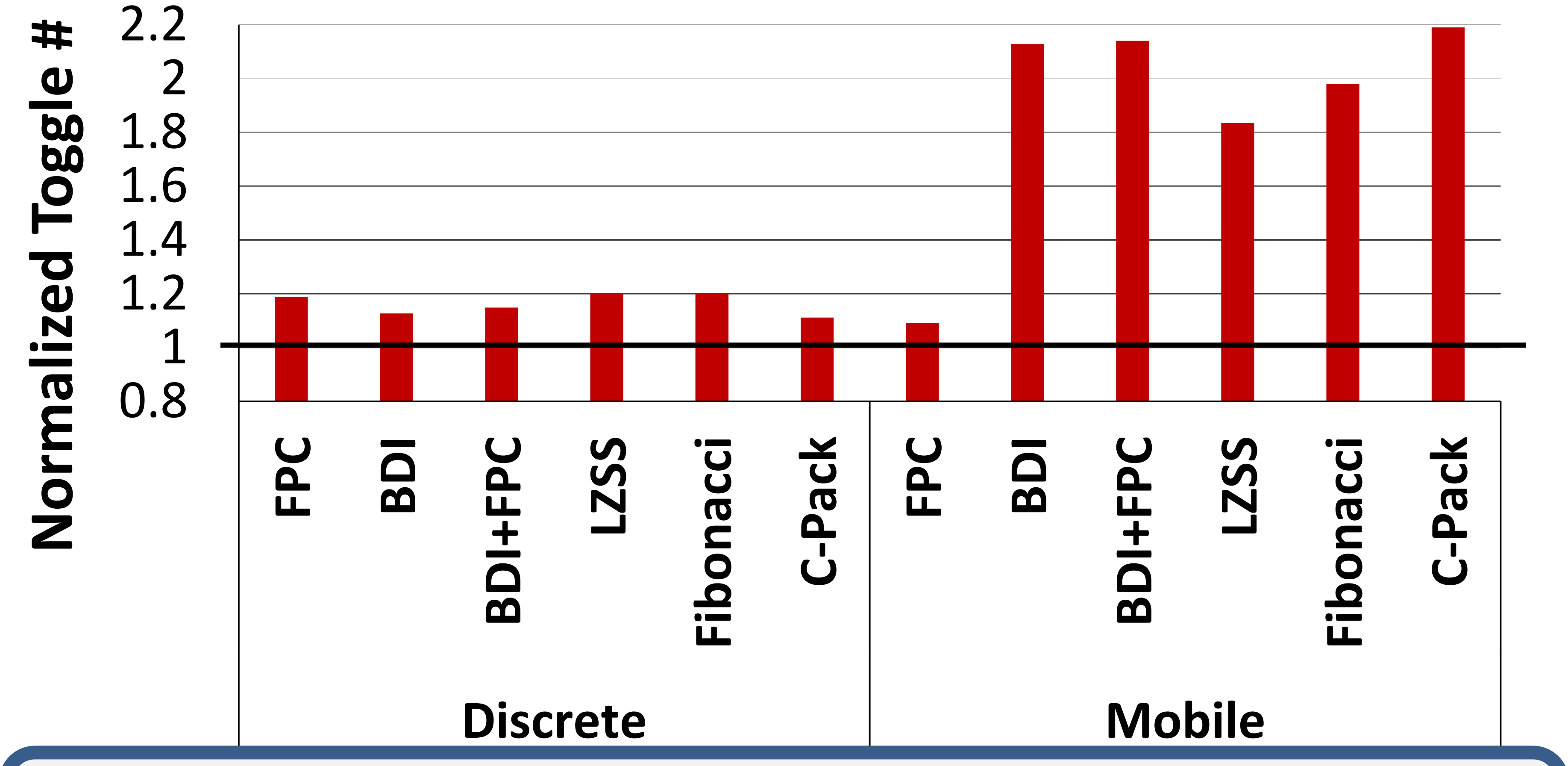
=

001001111 ... 1101000011000

# Toggles = 31

Effect of Compression on Bit Toggles

NVIDIA Apps: *Mobile GPU* – *54* in total, *Discrete GPU* – *167* in total



Significant increase in the number of toggles, hence potentially increase in consumed energy

Toggle-Aware Data Compression

Problem:

- *1.53X* effective compression ratio
- *2.19X* increase in toggle count

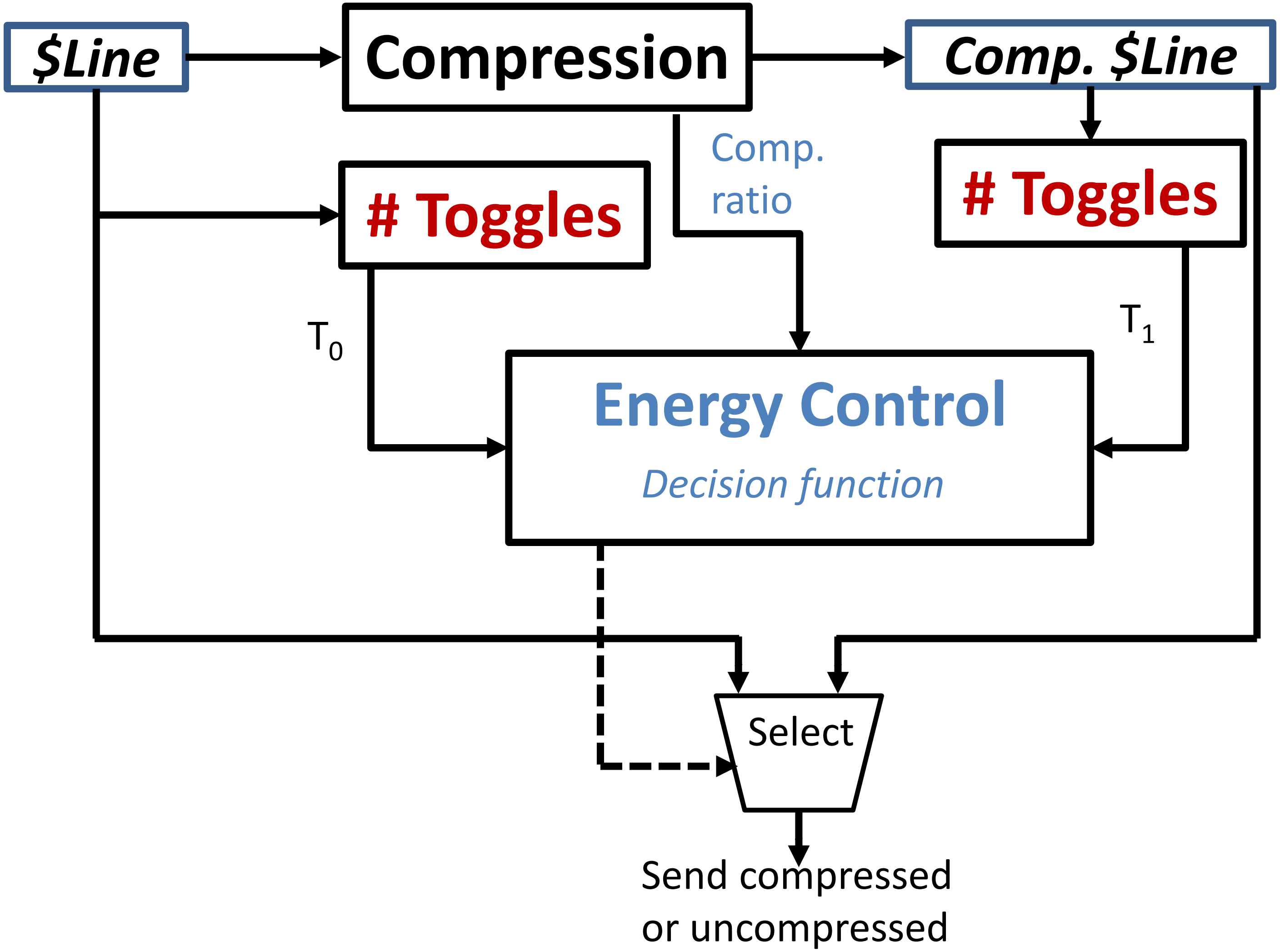
Goal:

- Find the optimal tradeoff between toggle count and compression ratio

Key Idea:

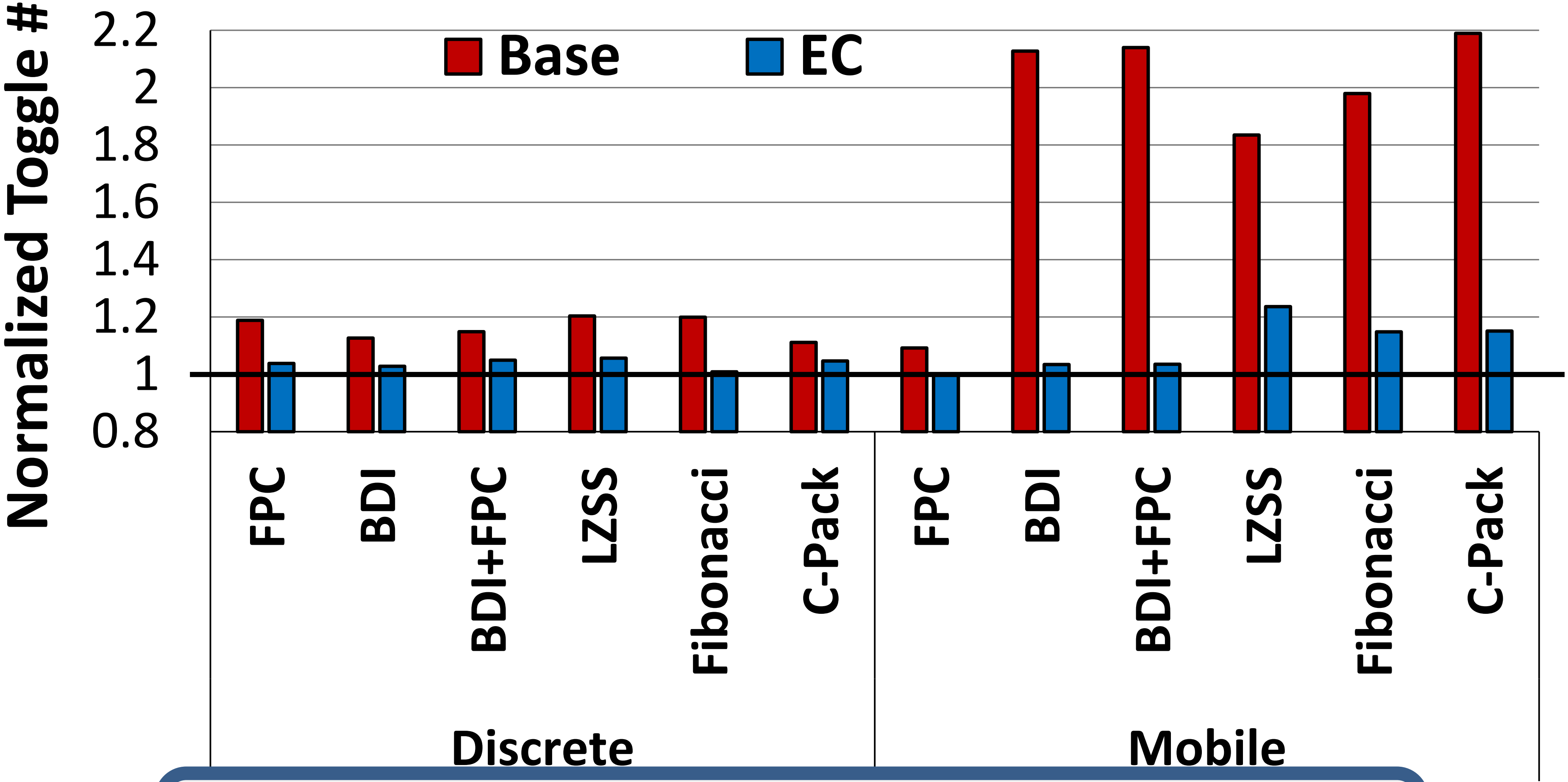
- Determine toggle rate for compressed vs. uncompressed data
- Use a heuristic (*Energy X Delay* or *Energy X Delay*^{2} metric) to estimate the tradeoff
- Throttle compression to reach estimated tradeoff

Energy Control (EC) Flow



Energy Control: Effect on Bit Toggles

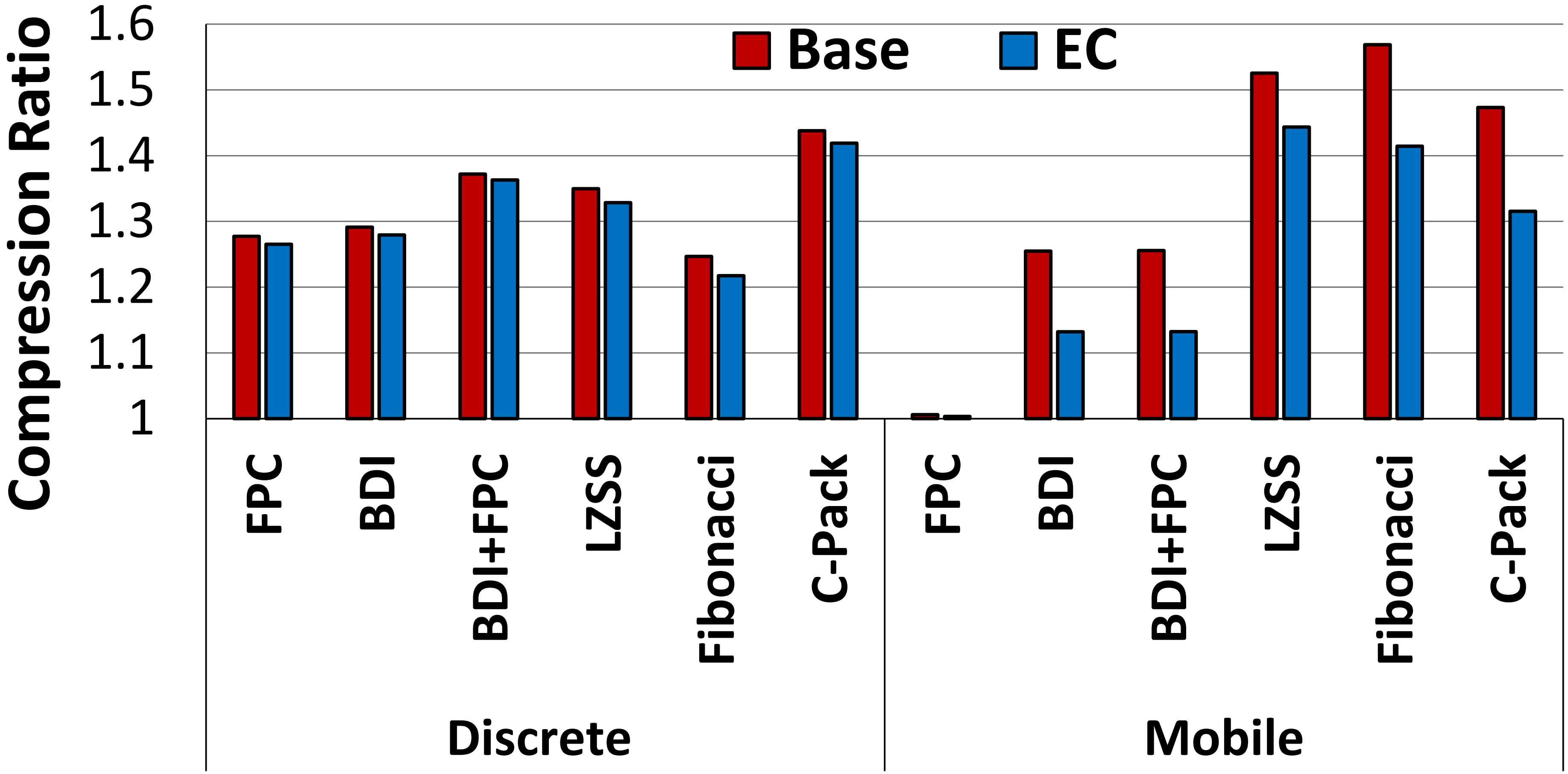
NVIDIA Apps: *Mobile GPU* – *54* in total, *Discrete GPU* – *167* in total



Significant **decrease** in the number of toggles

Energy Control: Effect on Compression Ratio

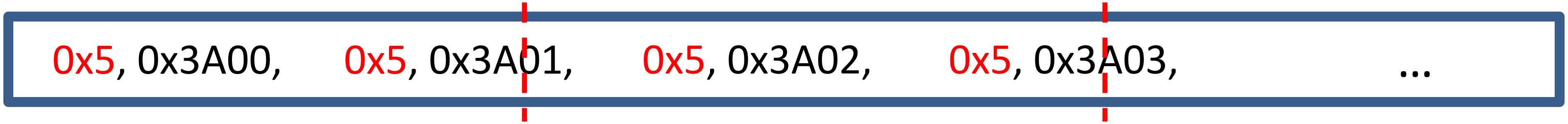
NVIDIA Apps: Mobile GPU – 54 in total, *Discrete GPU – 167* in total



Modest **decrease** in compression ratio

Optimization: Metadata Consolidation

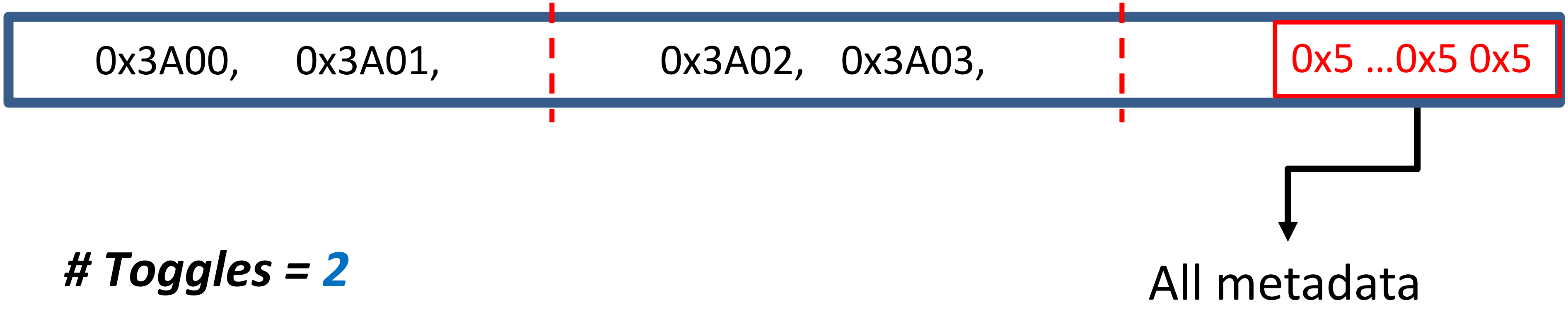
Compressed Cache Line with FPC, 4-byte flits



0x5, 0x3A00, 0x5, 0x3A01, 0x5, 0x3A02, 0x5, 0x3A03, ...

# Toggles = 18

Toggle-aware FPC: all metadata **consolidated**



0x3A00, 0x3A01, 0x3A02, 0x3A03, 0x5 ...0x5 0x5

All metadata

# Toggles = 2

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

Summary

- Bandwidth and energy efficiency are the first order concerns in modern systems
- **Data compression** is an attractive way to get higher effective bandwidth efficiently
- ***Problem:*** Excessive toggles ('0'⇔'1') waste power/energy
- ***Key Idea:***
  - Estimate the tradeoff between compression ratio and energy efficiency (*Energy X Delay* or *Energy X Delay^2*)
  - Throttle compression when the overall energy increases

Energy-Efficient Data Compression for Modern Memory Systems

Gennady Pekhimenko

ACM Student Research Competition

March, 2015

Carnegie Mellon University

SAFARI