Byte-Based Partial-Match Instruction and Data Compression for High-Performance and Low-Power Interconnects

Sujan Kumar Saha and Jiangjiang Liu

Lamar University

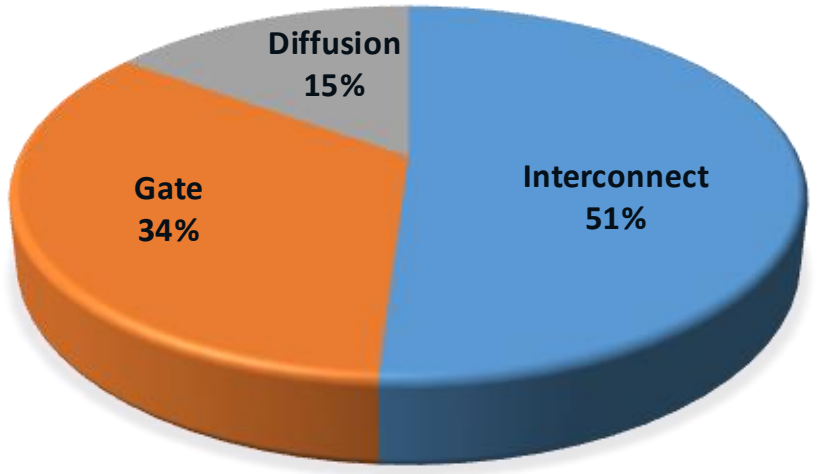
Texas, USA

Presented By

Dr. Jiangjiang Liu
Associate Professor
Department of Computer Science
Lamar University, USA

Challenges of Interconnect in VLSI circuits

- Power consumption
- Delay
- Crosstalk



Dynamic Power Breakdown

Source: Magen, Nir, et al. "Interconnect-power dissipation in a microprocessor." *Proceedings of the 2004 international workshop on System level interconnect prediction*. ACM, 2004.

Reason Behind The Challenges

- Capacitance between adjacent wires
- Inductance between adjacent wires

Device Level Solution to The Challenges

- Using Cu wire and low K dielectric
- Wire spacing
- Buffer insertion
- Track reassignment
- Shielding

Circuit Level Solution

- Staggered repeater
- Charge compensation
- Twisted differential signaling

Architecture Level Solution

- Compression (Reducing number of bits)
- Encoding (Changing bit pattern)

Byte-Based Partial-Match Compression of Instruction and Data transferred on Data Bus

Proposed Memory Hierarchy

Compression has been applied at three memory levels:

- L1 cache to L2 cache
- L2 cache to L3 cache
- L3 cache to Memory

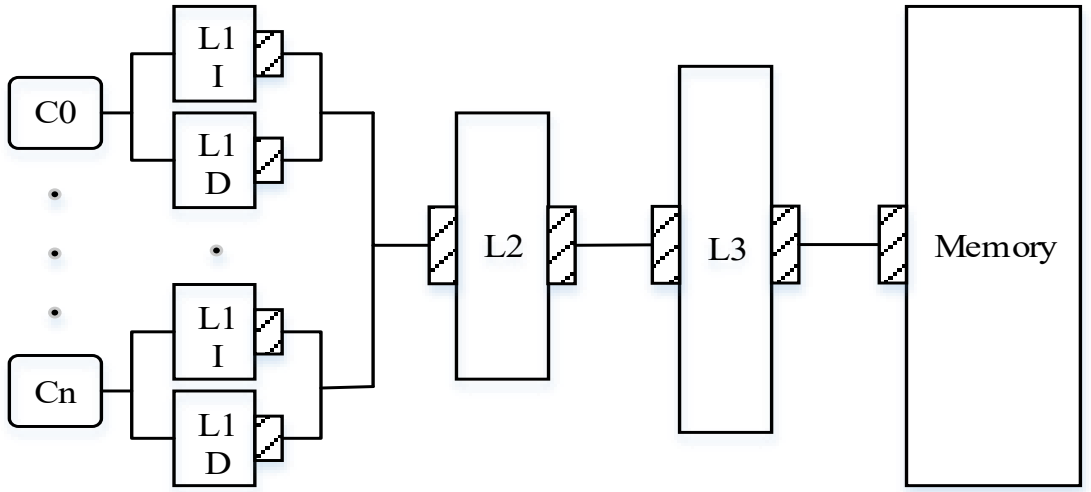


Fig: Multi-core system (Shaded area: Compressor and De-compressor)

Working Principle of Compressor

Number of way = w

Number of set = s

If s = 1, it is called Fully associative

If s > 1, it is called w-way associative

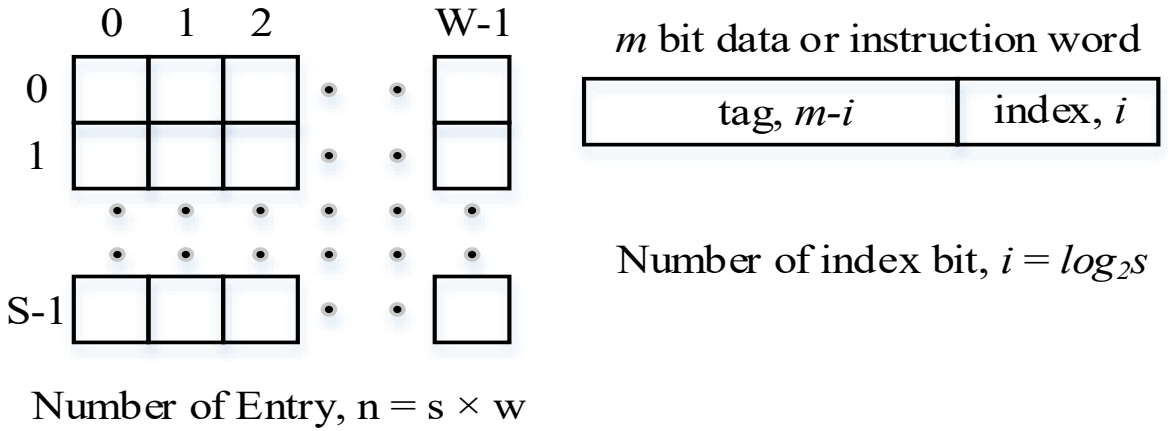


Fig: Memory organization and data indexing

m-bit data or instruction has two parts:

- Index, i bits
- Tag, m-i bits

Working Principle of Compressor

There are three types of match between existing tag and incoming tag:

- Complete Hit (CH): If tag matches completely.
- Partial Hit (PH): If tag matches partially.
- Complete Miss (CM): If no byte is matched.

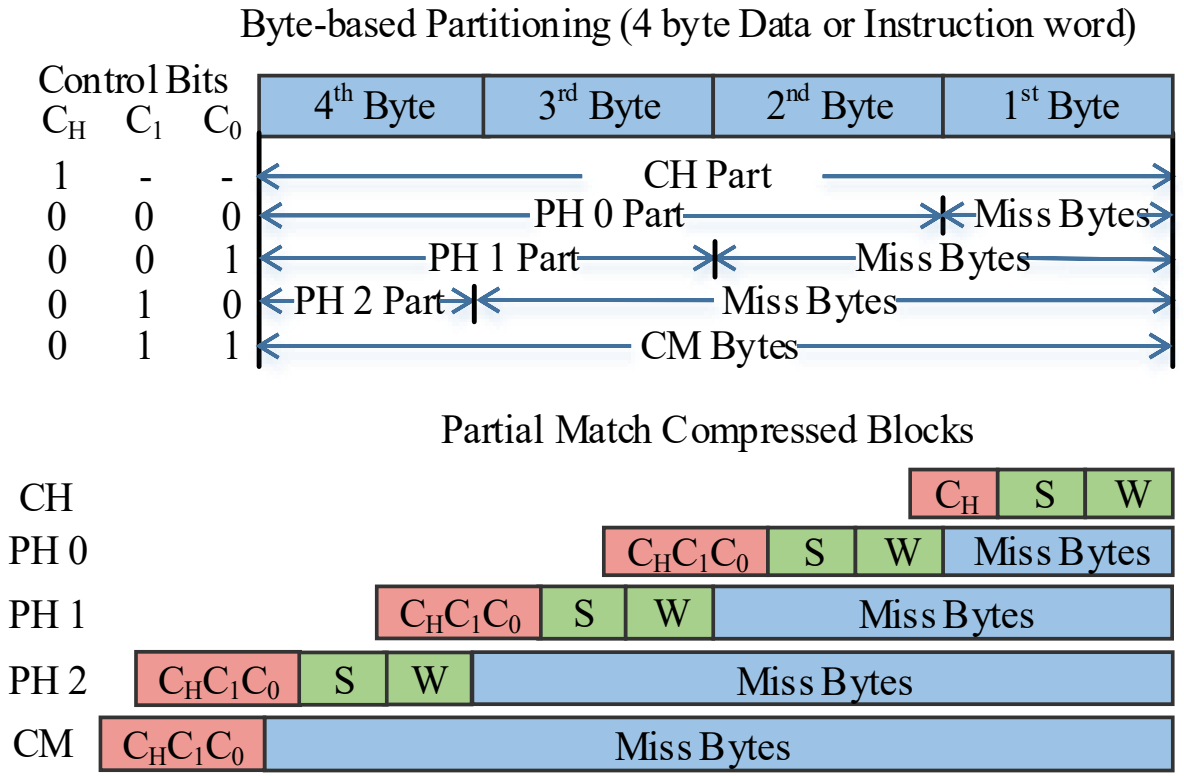


Fig: Byte-Based Partial-Match compression and compressed block formation

Working Principle of Compressor

- Position information (set bit, s and way bit, w) of matched tag are added in the compressed word.
- In case of Partial hit, miss bytes of incoming word are replaced with the miss bytes of existing tag of that position.
- In case of Complete miss, whole word is replaced in Least Recently Used (LRU) position.

Working Principle of Compressor

TABLE I Compressed Word Size in Bits for Hit, Partial Hit, and Miss Cases

Hit/Miss type	**Number of Compressor Entries**					
	8	***16***	***32***	***64***	***128***	***256***
Complete Hit	4	5	6	7	8	9
Partition[0]	14	15	16	17	18	19
Partition[1]	22	23	24	25	26	27
Partition[2]	30	31	32	33	34	35
Complete Miss	35	35	35	35	35	35

Transmission of Compressed word over Bus

Compressed words can be transferred over bus in two ways:

- Using original bus width: system performance increases.
- Using compressed bus width: interconnect power consumption will be reduced.

Evaluation Methodology

- M5 simulator has been used
- SPECCPU2006 benchmarks have been used:

  mcf, bzip2, sjeng, gobmk, libquantum, hmmer, namd, soplex
- Five Multi-core and Multi-thread systems are used:

  2x1, 2x2, 4x1, 4x2, 8x1
- CPU clock rate 2GHz
- L1 cache size 32KB and 2-way set associative
- L2 cache size 1MB and 8-way set associative
- L3 cache size 2MB and 16-way set associative

Evaluation Methodology

- Bus width between L1 and L2 is 8 Byte and latency 2 cycle
- Bus width between L2 and L3 is 8 Byte and latency 2 cycle
- Bus width between L3 and Memory is 18 Byte and latency 4 cycle
- Global interconnect parameters:

Parameters	**Technologies**		
	32nm	***45nm***	***65nm***
Supply Voltage (V)	0.9	1	1.1
Width (µm)	0.22	0.31	0.45
Space (µm)	0.22	0.31	0.45
Thickness (µm)	1.2	1.2	1.2
Height (µm)	0.09	0.14	0.2
Dielectric Constant	2.3	2.5	2.9
Coupling Capacitance, Cc (fF/mm)	118.55	94.65	79.60
Self-Capacitance, Cs (fF/mm)	80.26	87.07	100.81

Simulation Results

A. Compression Ratio (CR)

$$Compression\ Ratio = \frac{Number\ of\ bits\ after\ compression}{Total\ Number\ of\ original\ bits}$$

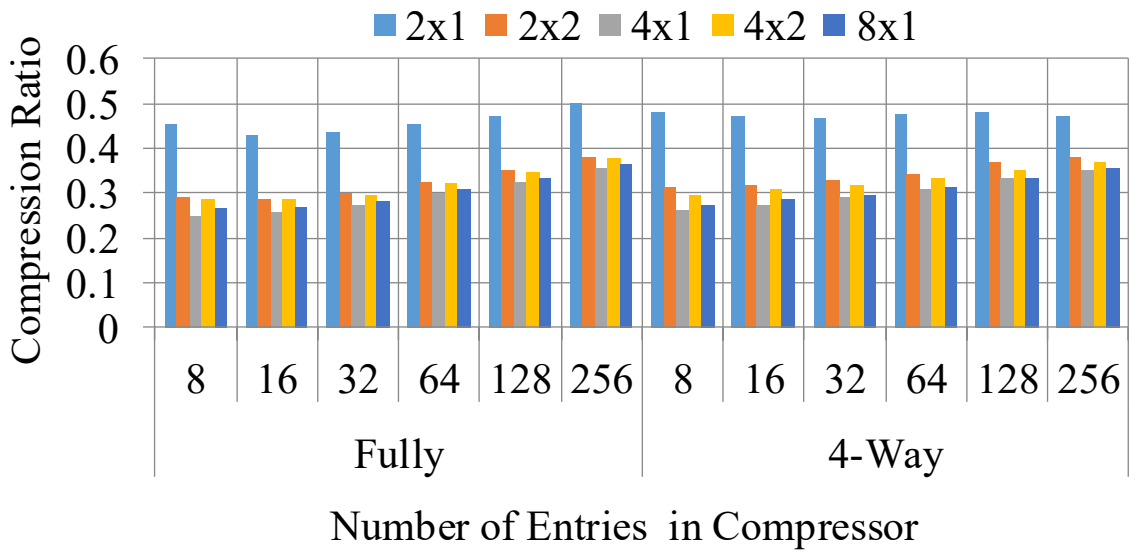


Fig: Average compression ratios of L1-L2, L2-L3 and L3-Memory

Simulation Results

- Compression ratio of 4-way and fully associative compressor are about same
- Compression ratio increases with the increase of compressor entries
- Partial Match works better for 8-entry compressor
- For 2x1 system, CR is higher than other architectures
- Minimum average CR is 28% which is for 4x1 system with 8-entry compressor

Simulation Results

B. Hit Rate, Partial Hit Rate, and Miss Rate

- Hit rate and Partial Hit rate increase with the increase of compressor entries.
- Complete miss rate is higher in 4-way compressor than fully compressor.
- Partial Hit 1 and Partial Hit 2 are higher in fully associative compressor.

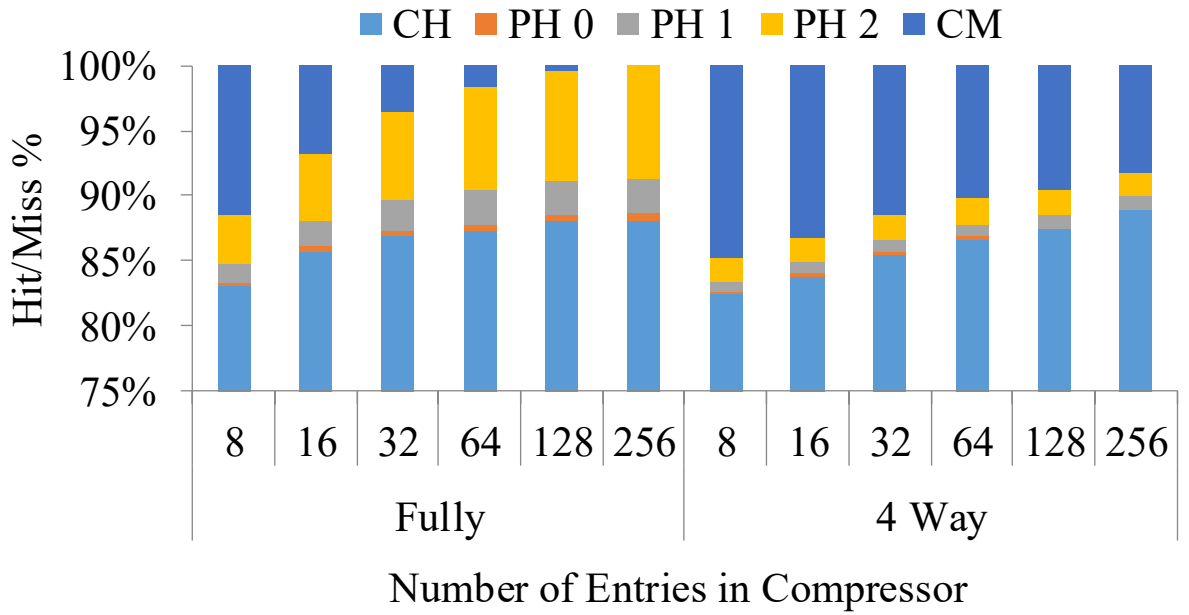


Fig: Average hit and miss rates of compressors between L1-L2, L2-L3, and L3-Memory

Simulation Results

C. Performance Improvement

$$Performance\ Improvement(\%) = \frac{IPC\ after\ compression - Default\ Sytem\ IPC}{Default\ System\ IPC} \times 100$$

- IPC increases with the increase of core number and thread number

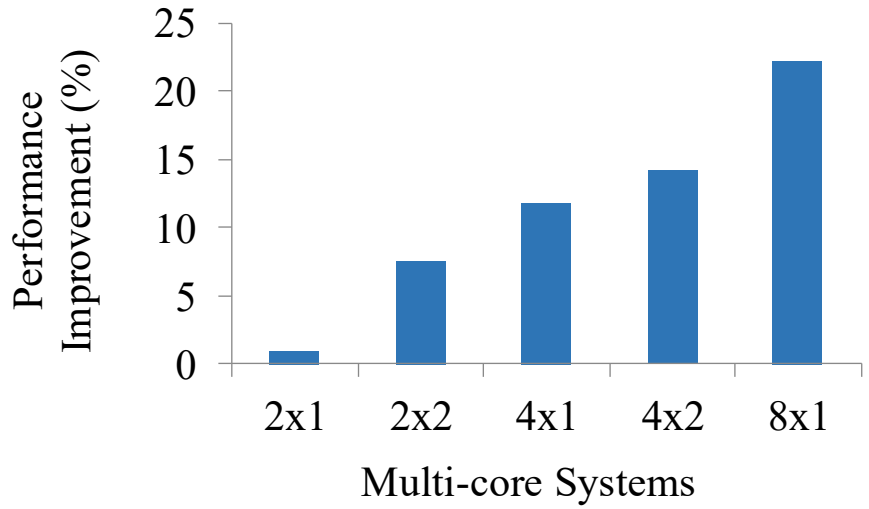


Fig: Performance improvement of Multi-core system

Simulation Results

D. Energy Reduction

The equation to calculate energy of global interconnect:

$$E = \frac{1}{2} \times C_s \times V^2 [N_s + \frac{C_c}{C_s} (N_{c+d} + 4 \times N_t)]$$

Here,

- V is the voltage difference between high logic level and low logic level
- N_s is the number of self-transition
- N_{c+d} is the number of charging and discharging transitions
- N_t is the number of toggle transition
- C_s is self-capacitance if interconnect wire
- C_c is coupling capacitance of adjacent wire

Simulation Results

D. Energy Reduction

$$\text{Energy Ratio} = \frac{\text{Energy after Compression with Compressed Bus}}{\text{Energy of Default System with Original Bus}}$$

- 7% energy reduction in 2x1 and 2x2 system
- About 40% energy is reduced in 4x1, 4x2, and 8x1 system

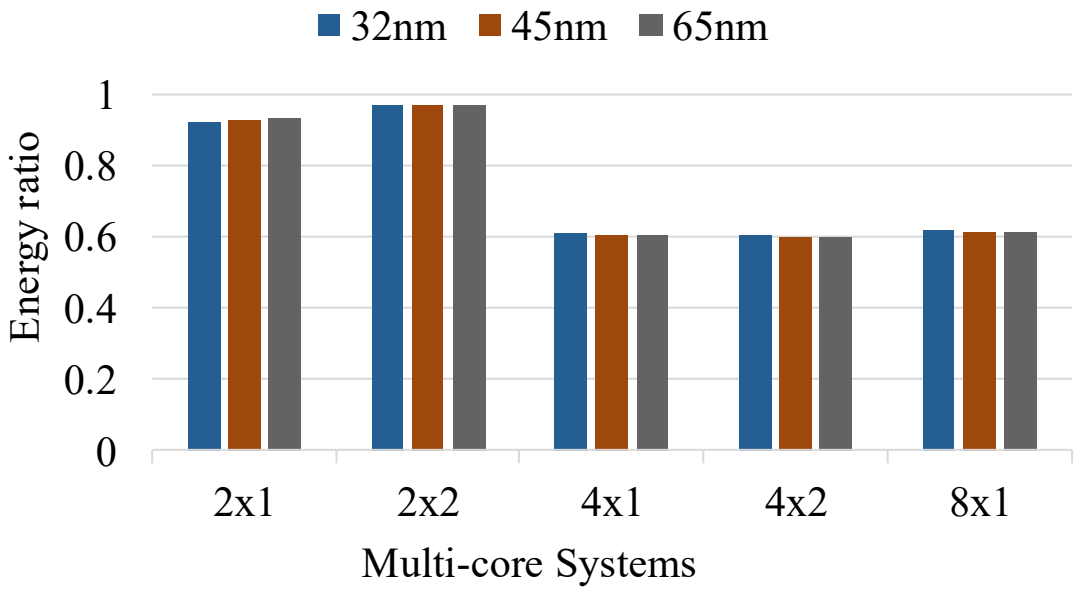


Fig: Energy reduction of Multi-core system with Partial-Match compression

Simulation Results

E. Comparison between Bus-Expander and Partial-Match

- Bus-Expander has only hit and miss
- Partial Match has hit, miss and partial hit
- Using Partial Match, 60.7% more compression is achieved over Bus-Expander

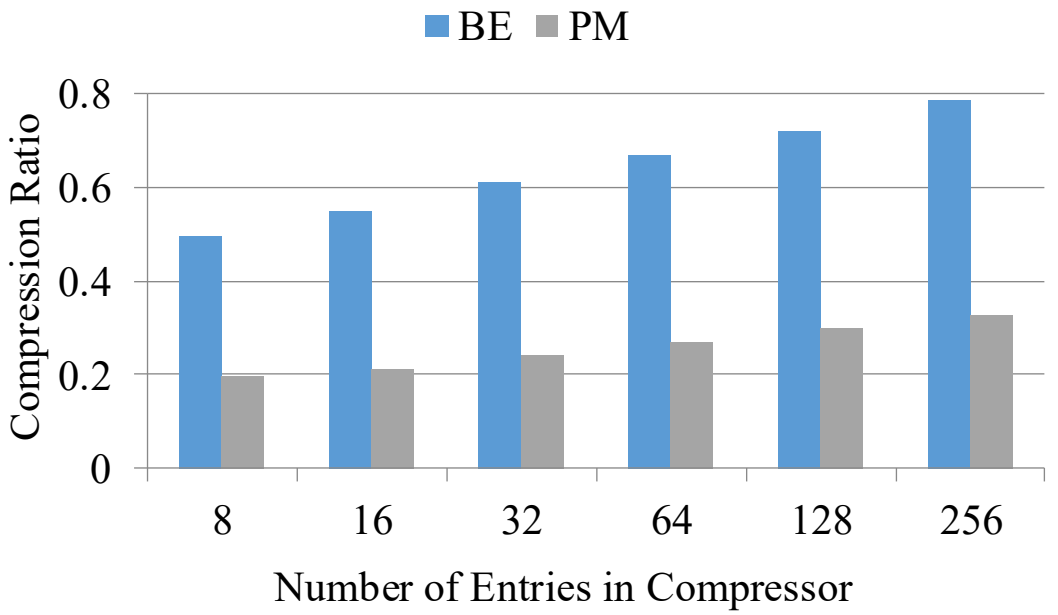


Fig: Comparison of Bus-Expander (BE) and Partial Match (PM)

Conclusion and Future Work

- The goal of the research is to reduce power consumption and increase system performance.
- Byte-Based Partial-Match compression has been proposed.
- Up to 28% average compression ratio is achieved.
- System performance increases up to 22.26 % with default bus width.
- 7~40% reduced power consumption with compressed bus width.
- 60.7% more compression over Bus-Expander.
- In future, area, power and latency of compressor and de-compressor will be measured.

Thank You