CACHE INTERCONNECTS

Mahdi Nazm Bojnordi

Assistant Professor

School of Computing

University of Utah

THE UNIVERSITY OF UTAH

CS/ECE 7810: Advanced Computer Architecture

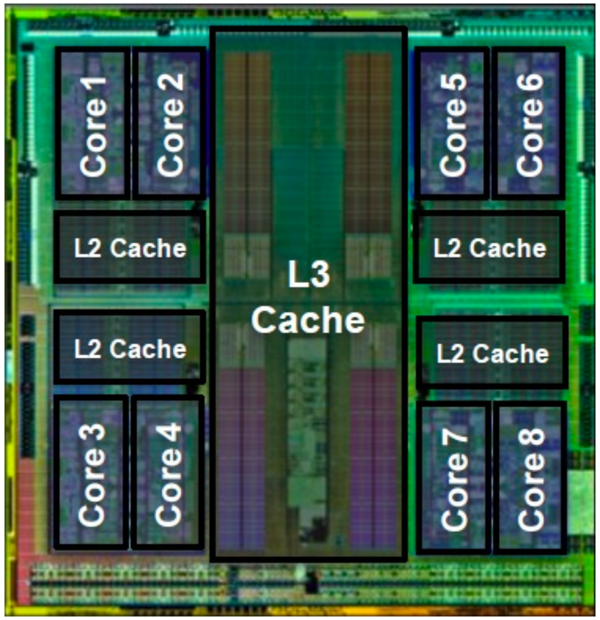
Overview

- Upcoming deadline
  - Feb.1st: project group formation
  - Note: email me once you form a group
- This lecture
  - Content aware optimizations
  - Cache interconnect optimizations
  - Encoding based optimizations

Recall: Cache Power Optimization

- Caches are power and performance critical components
- Performance
  - Bridging the CPU-Mem gap
- Static power
  - Large number of leaky cells
- Dynamic power
  - Access through long interconnects

Example: FX Processors

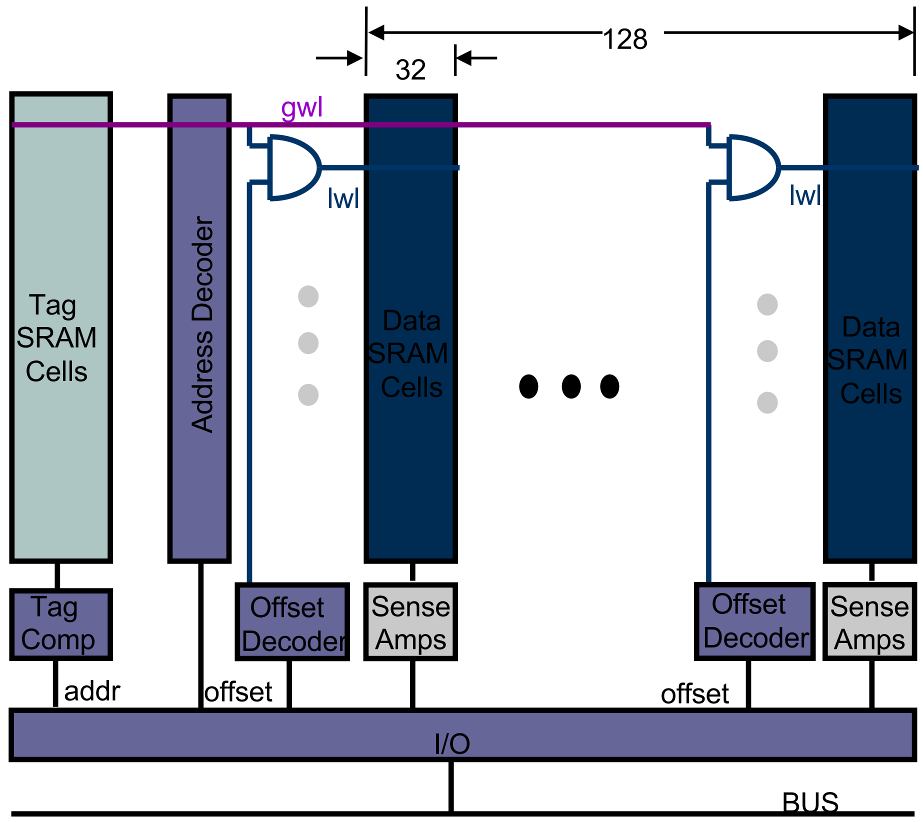


[source: AMD]

Content Aware Optimizations

Dynamic Zero Compression

- More than 70% of the bits in data cache accesses are 0s



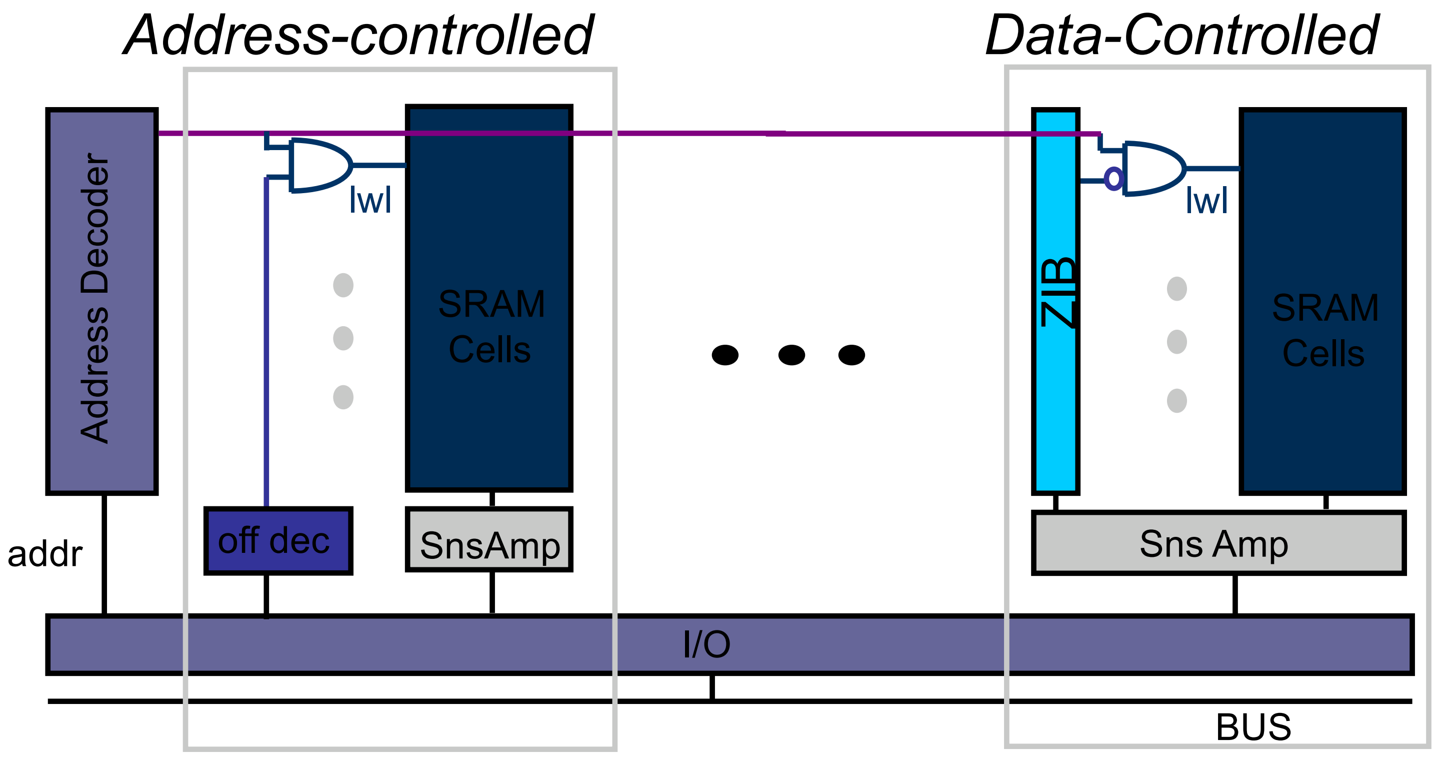
Example of a small cache

	Read (pJ)	Read (%)	Write (pJ)	Write (%)
Total	44.4	100.0	99.1	100.0
Decoder	5.5	12.4	5.5	5.5
word lines	1.1	2.5	1.1	1.1
Tag bitlines and sense-amp	3.0	6.2	3.0	3.0
Data bitlines and sense-amp	14.5	32.7	69.2	69.9
I/O buses	12.1	27.3	12.1	12.2
Other	8.4	18.9	8.4	8.5

[Villa'00]

Dynamic Zero Compression

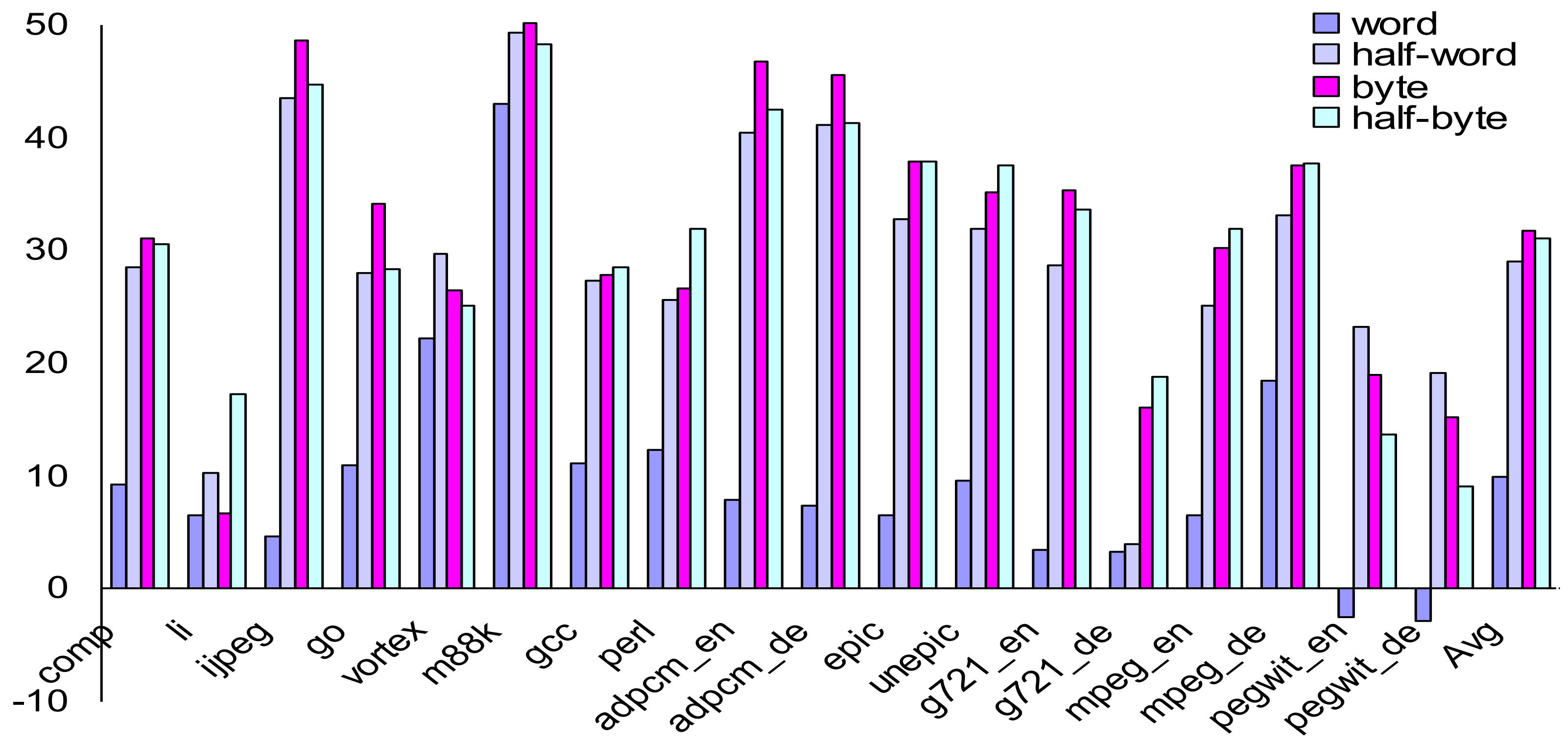
- Zero Indicator Bit; one bit per grouping of bits; set if bits are zeros; controls wordline gating



[Villa'00]

Dynamic Zero Compression

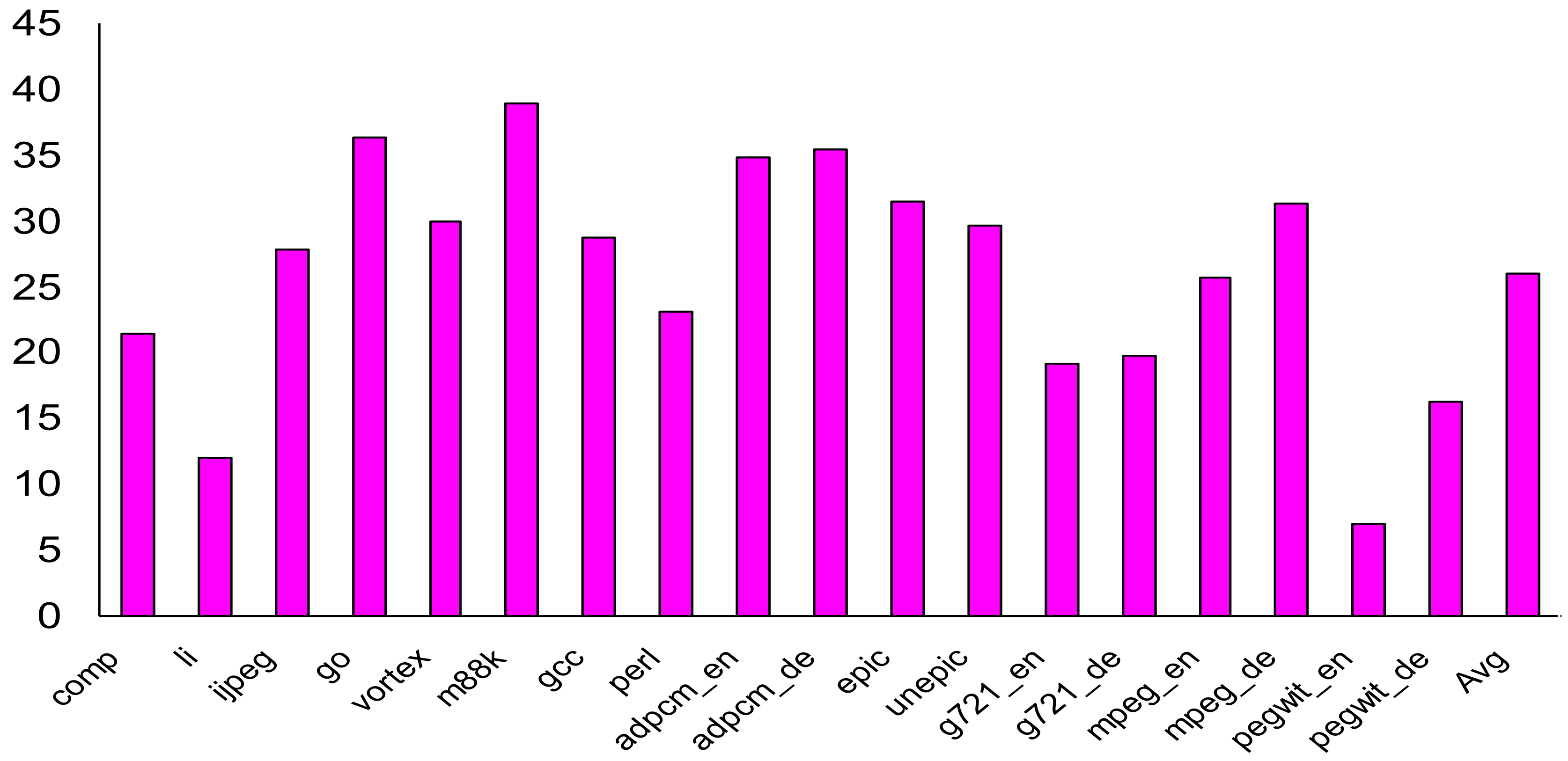
- Data cache bitline swing reduction



[Villa'00]

Dynamic Zero Compression

- Data cache energy savings

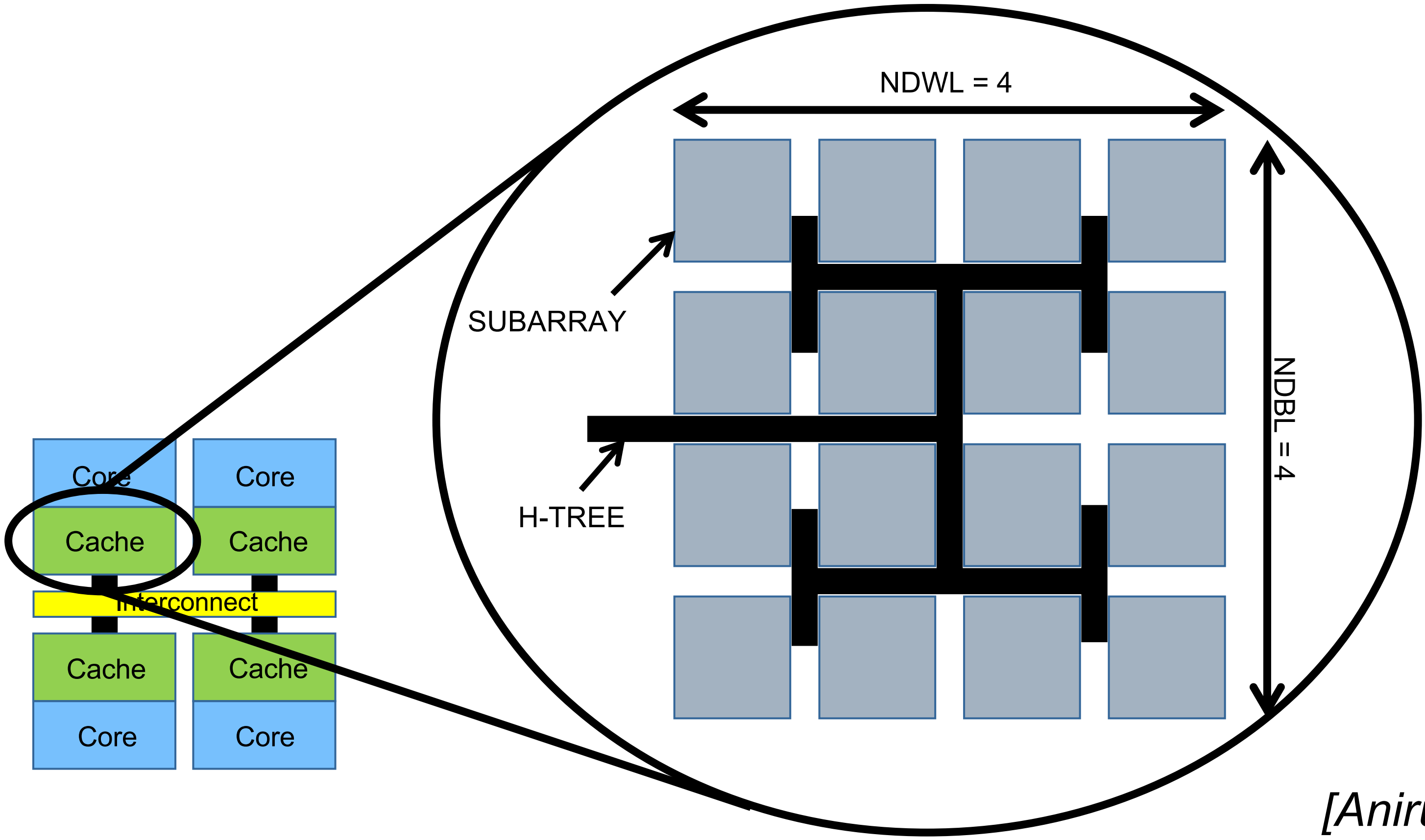


[Villa'00]

Cache Interconnect Optimizations

Large Cache Organization

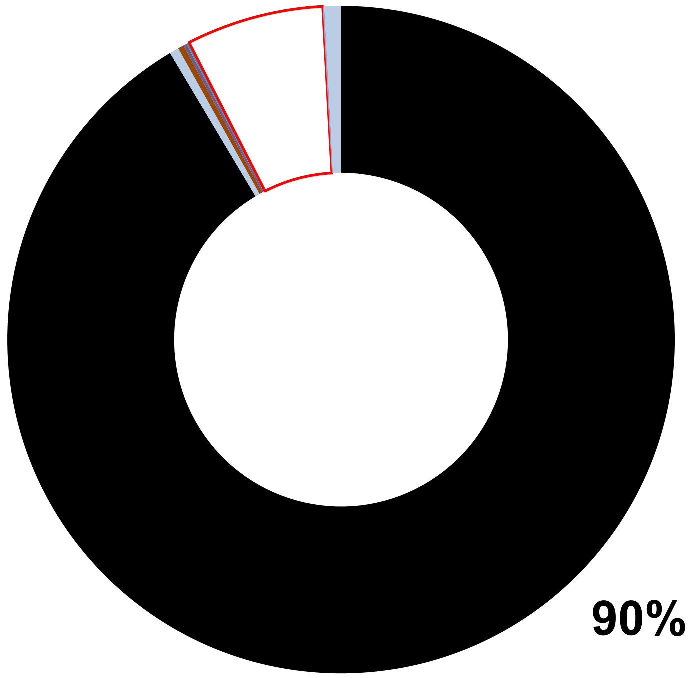
- Fewer subarrays gives increased area efficiency, but larger delay due to longer wordlines/bitlines



[Aniruddha'09]

Large Cache Energy Consumption

- H-tree is clearly the dominant component of energy consumption

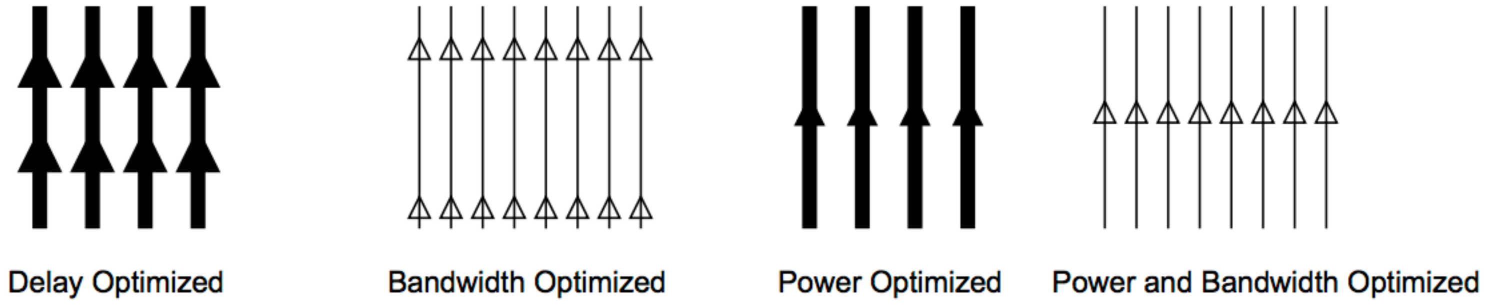


- ■ H-tree
- ■ Decoder
- ■ Wordlines
- ■ Bitline mux & drivers
- ■ Senseamp mux & drivers
- □ Bitlines
- ■ Sense amplifier
- ■ Sub-array output drivers

[Aniruddha'09]

Heterogeneous Interconnects

- A global wire management at the microarchitecture level
- A heterogeneous interconnect that is comprised of wires with varying latency, bandwidth, and energy characteristics



[Balasubramonian'05]

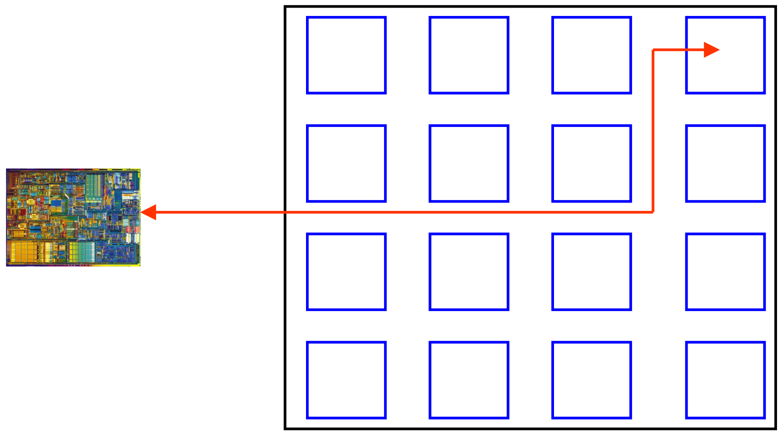
Heterogeneous Interconnects

- Better energy-efficiency for a dynamically scheduled partitioned architecture
  - ED^2 is reduced by 11%
- A low-latency low-bandwidth network can be effectively used to hide wire latencies and improve performance
- A high-bandwidth low-energy network and an instruction assignment heuristic are effective at reducing contention cycles and total processor energy.

[Balasubramonian'05]

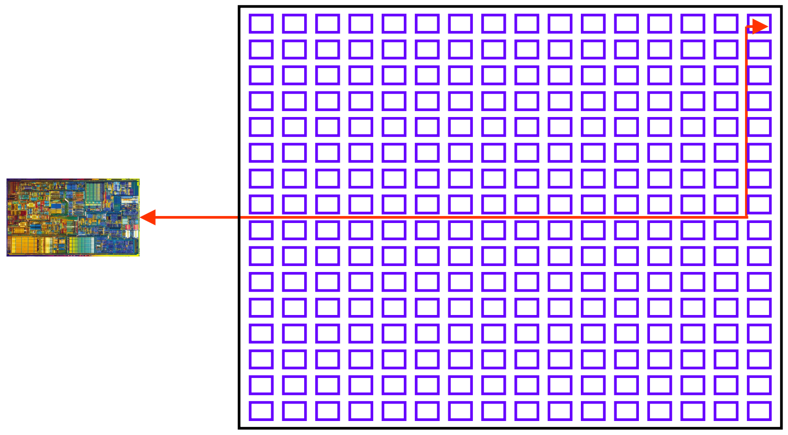
Non-Uniform Cache Architecture

- NUCA optimizes energy and time based on the proximity of the cache blocks to the cache controller.



2MB @ 130nm

Bank Access time = 3 cycles
Interconnect delay = 8 cycles



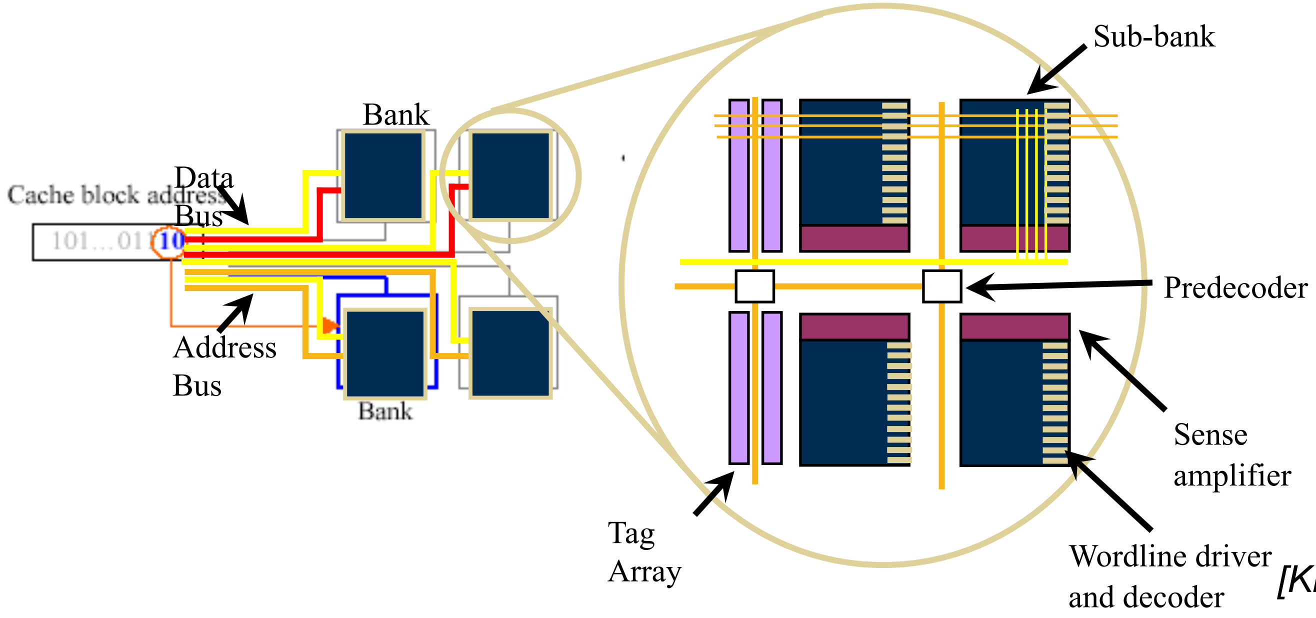
16MB @ 50nm

Bank Access time = 3 cycles
Interconnect delay = 44 cycles

[Kim'04]

Non-Uniform Cache Architecture

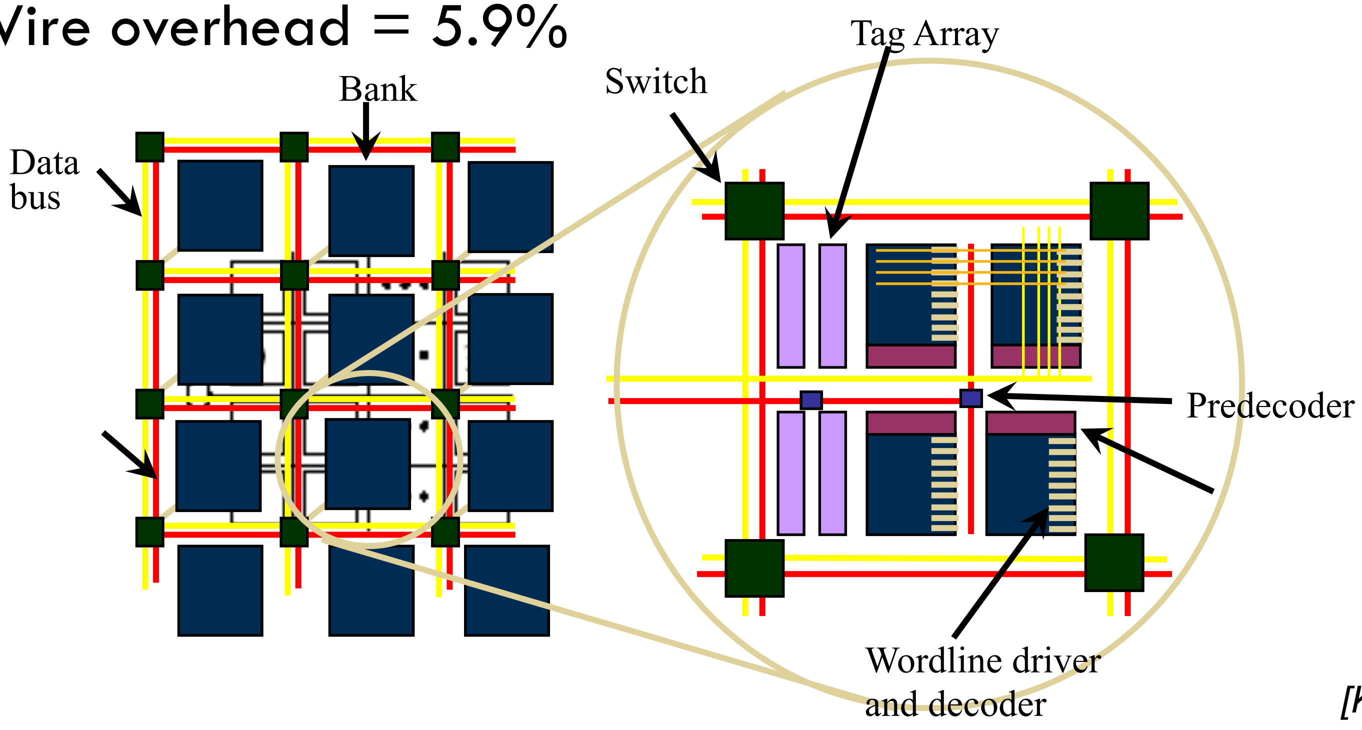
- S-NUCA-1
  - Use private per-bank channel
  - Each bank has its distinct access latency
  - Statically decide data location for its given address
  - Average access latency =34.2 cycles
  - Wire overhead = 20.9% → an issue



[Kim'04]

Non-Uniform Cache Architecture

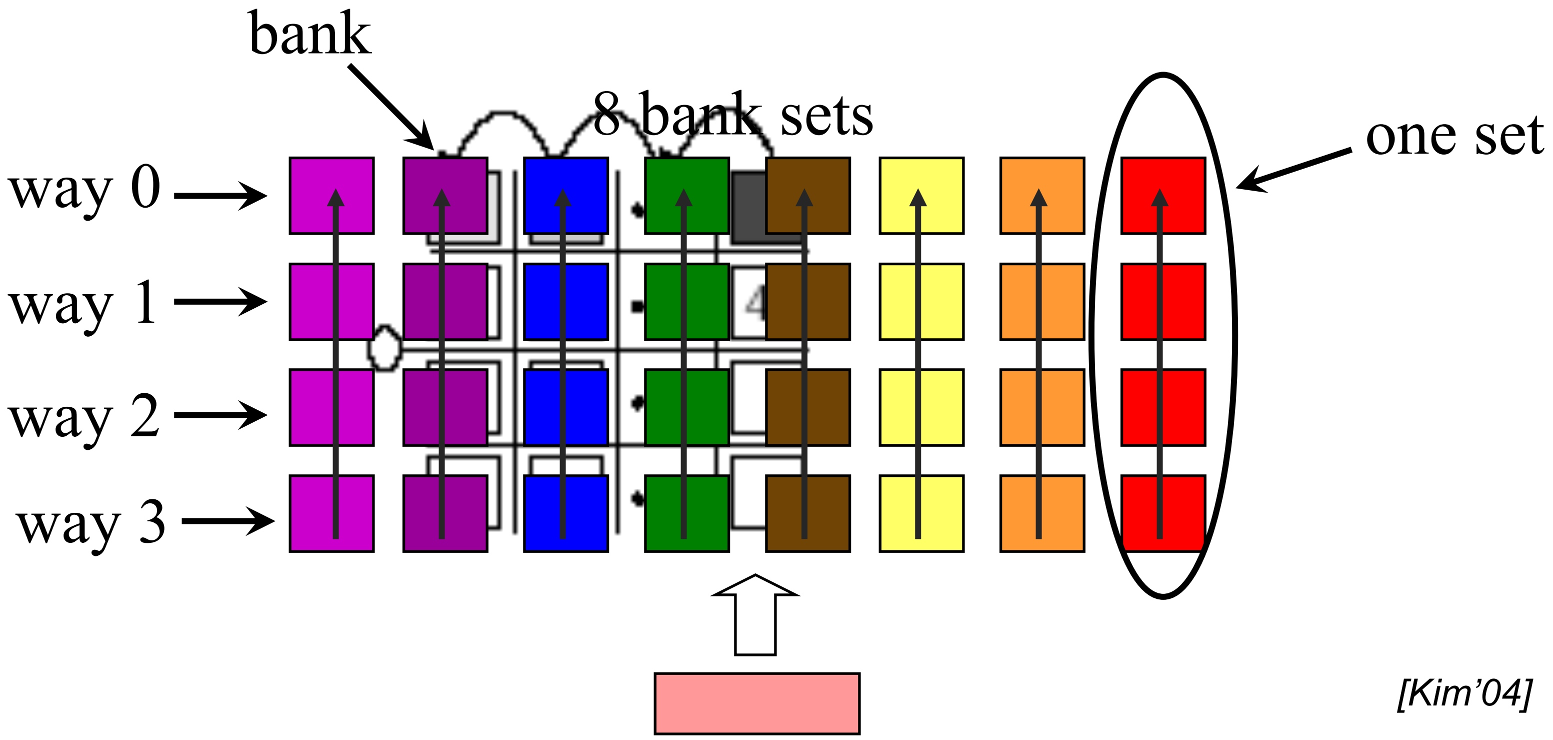
- S-NUCA-2
  - Use a 2D switched network to alleviate wire area overhead
  - Average access latency =24.2 cycles
  - Wire overhead = 5.9%



[Kim'04]

Non-Uniform Cache Architecture

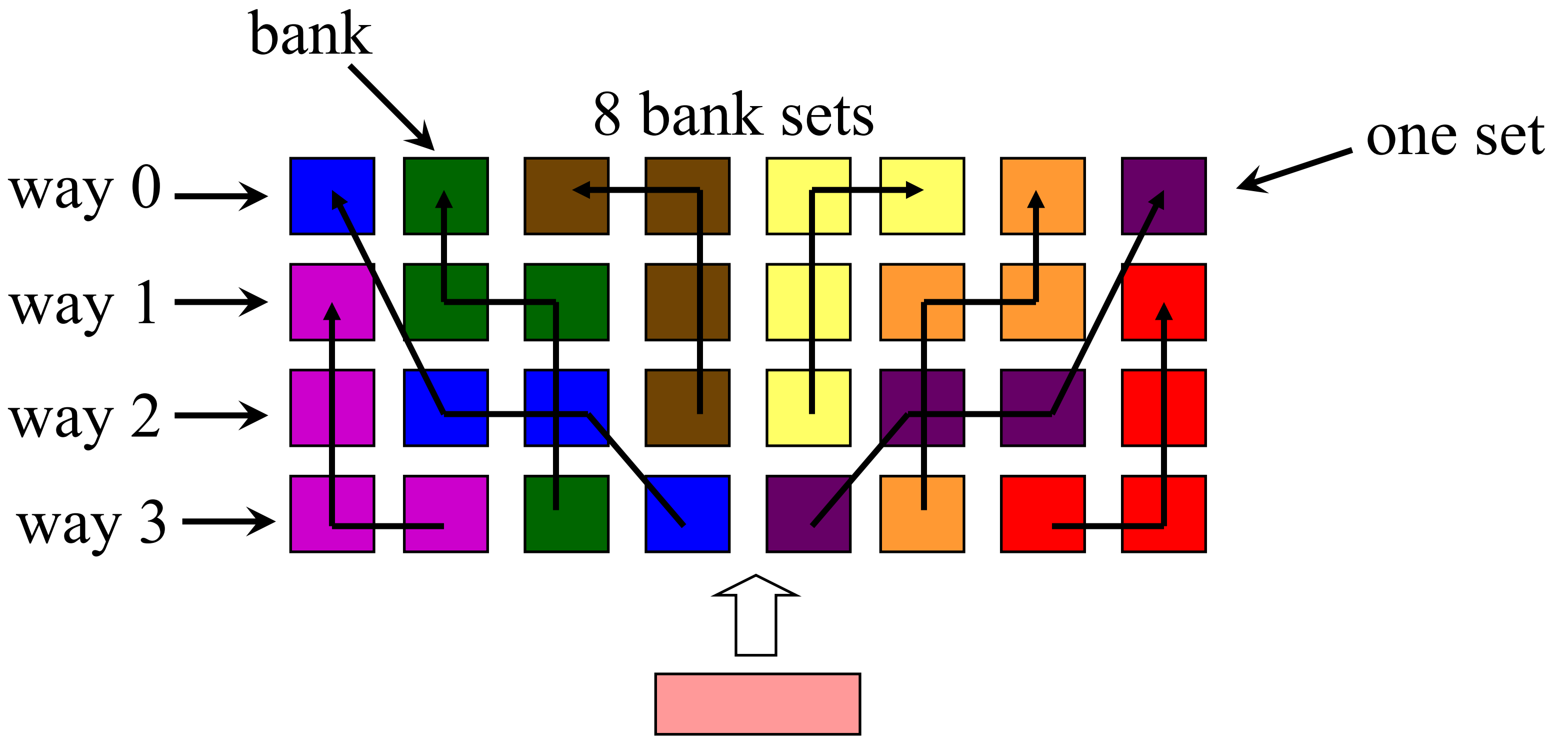
- Dynamic NUCA
  - Data can dynamically migrate
  - Move frequently used cache lines closer to CPU



[Kim'04]

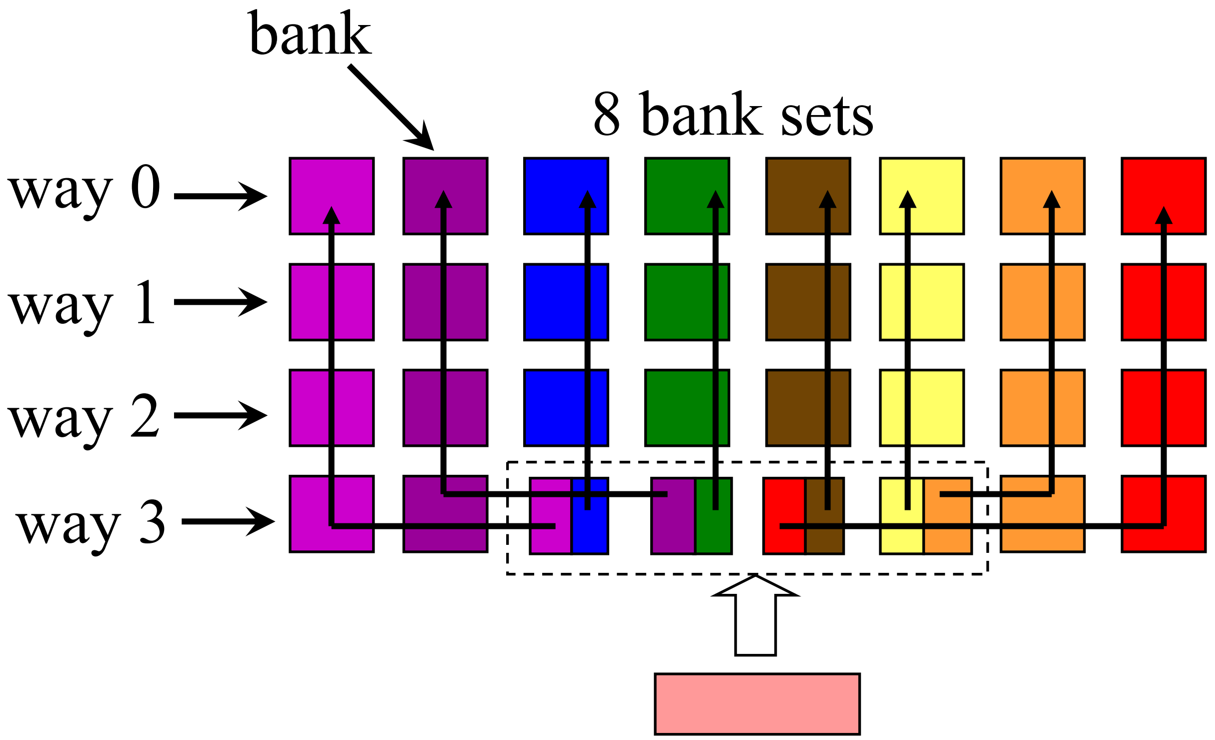
Non-Uniform Cache Architecture

- Fair mapping
  - Average access time across all bank sets are equal



Non-Uniform Cache Architecture

- Shared mapping
  - Sharing the closet banks for farther banks

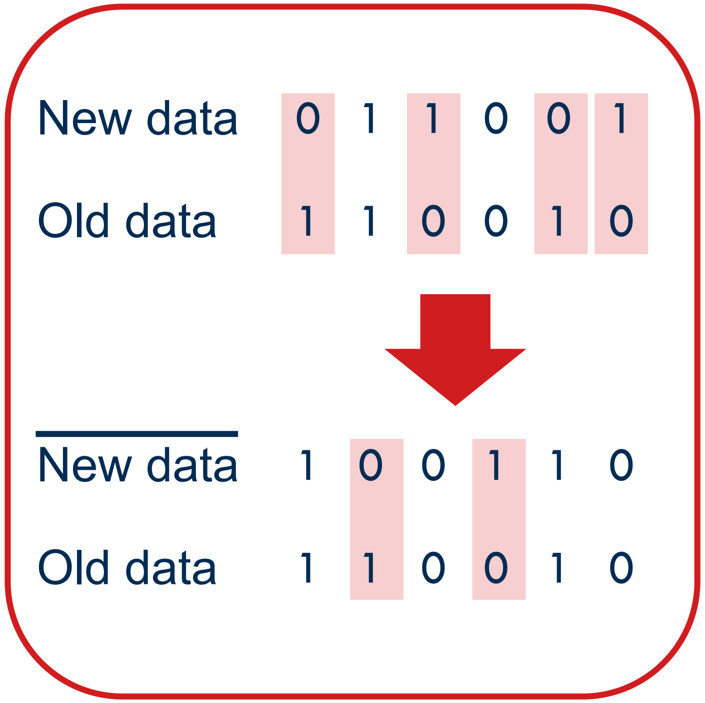


Encoding Based Optimizations

Cache Interconnect Optimizations

- Bus invert coding transfers either the data or its complement to minimize the number of bit flips on the bus.

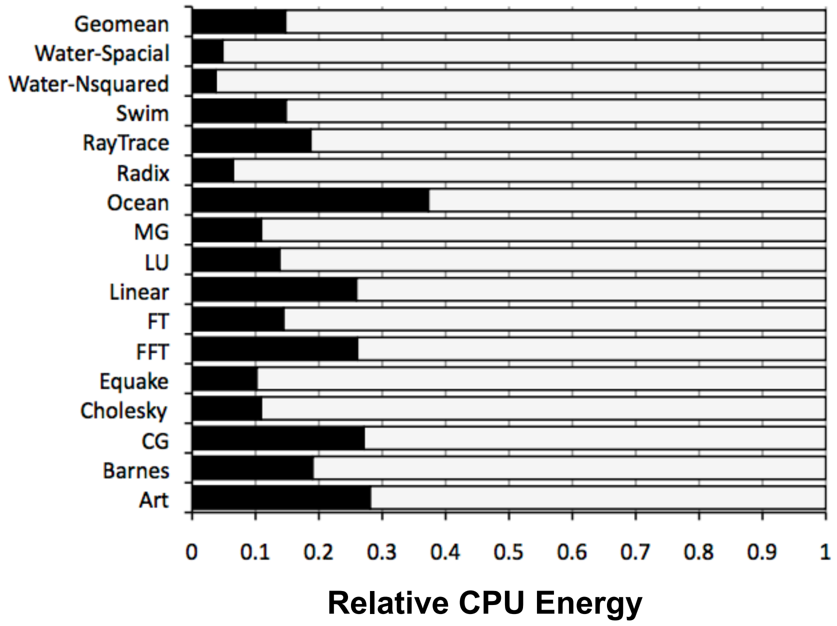
$$P_{\text{switching}} = \alpha C V_{DD}^{2} f$$



[Stan'95]

Time-Based Data Transfer

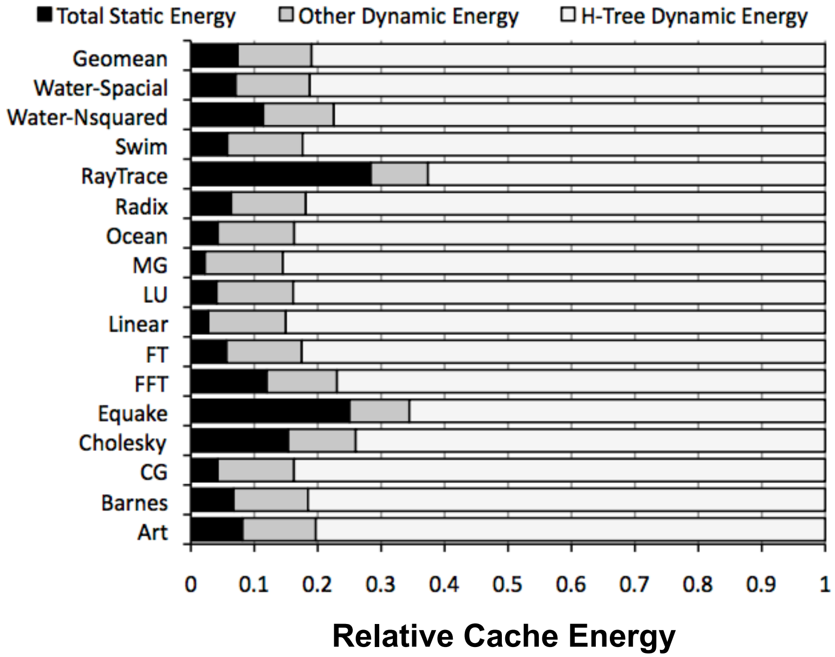
- The percentage of processor energy expended on an 8MB cache when running a set of parallel applications on a Sun Niagara-like multicore processor



[Bojnordi'13]

Time-Based Data Transfer

- Communication over the long, capacitive H-tree interconnect is the dominant source of energy consumption (80% on average) in the L2 cache

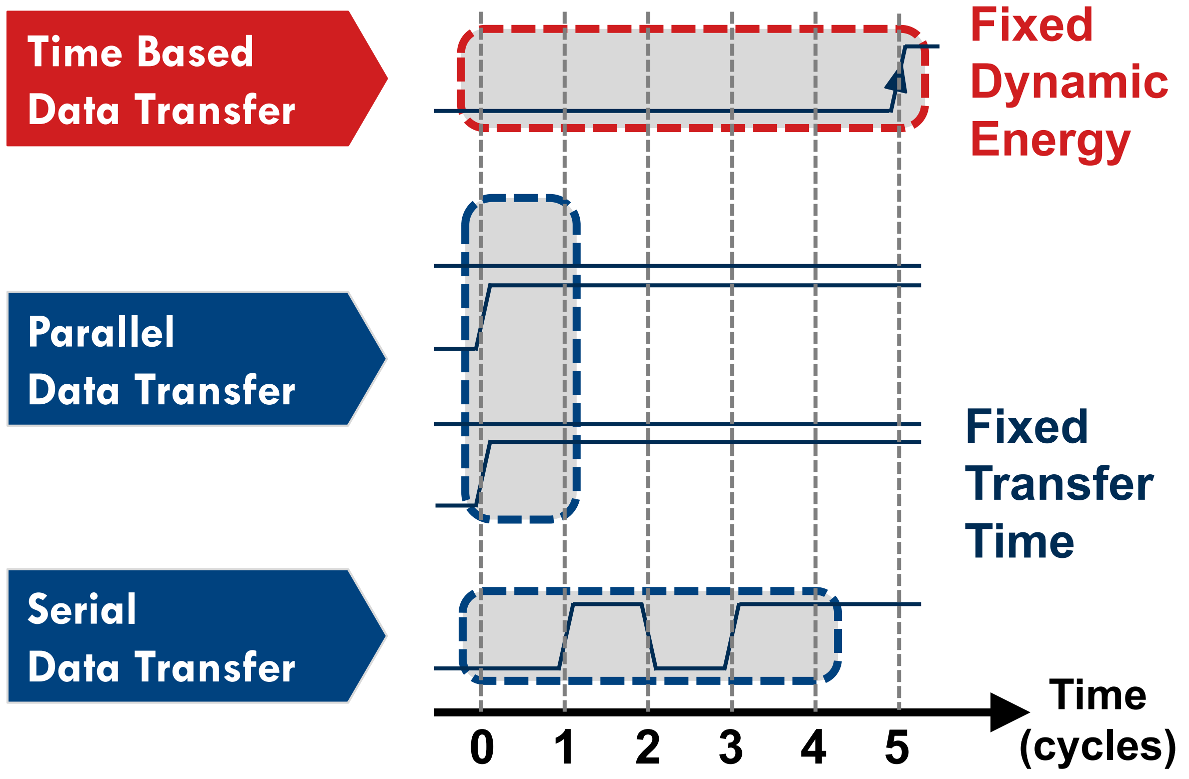


[Bojnordi'13]

Time-Based Data Transfer

Key idea: represent information by the number of clock cycles between two consecutive pulses to reduce interconnect activity factor.

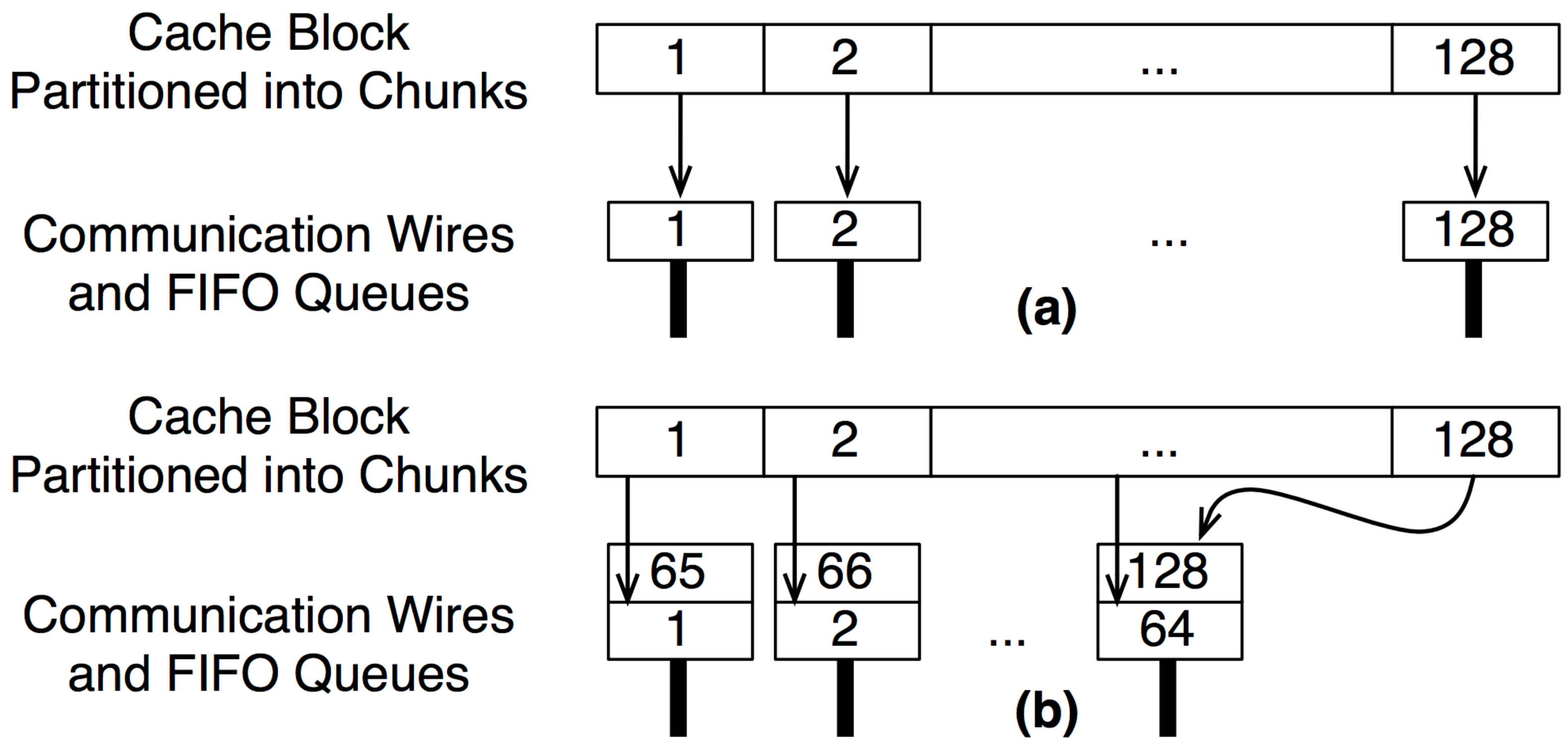
Example: transmitting the value 5



[Bojnordi'13]

Time-Based Data Transfer

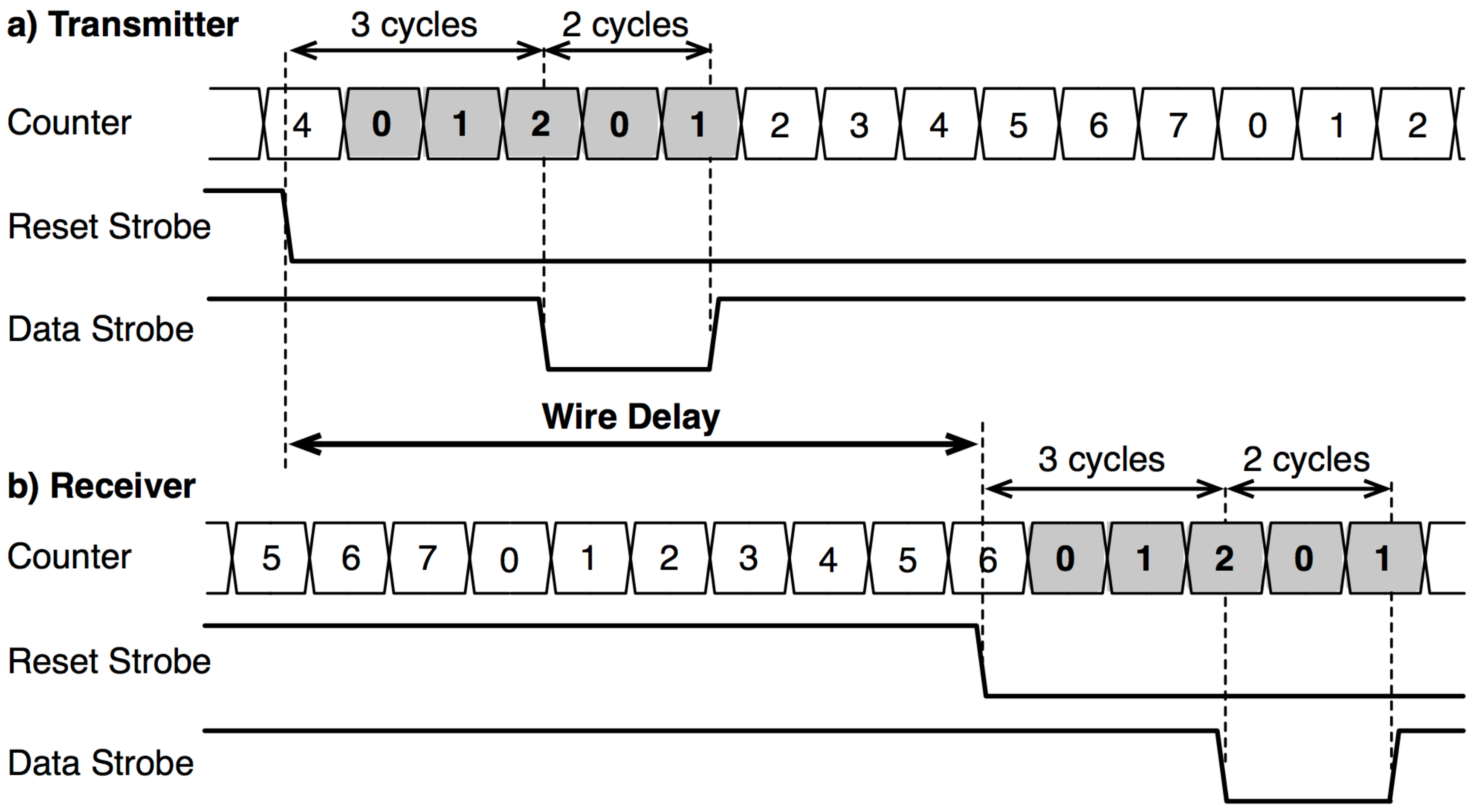
- Cache blocks are partitioned into small, contiguous chunks.



Cache Block Partitioned into Chunks
1
2
...
128
Communication Wires and FIFO Queues
1
2
...
128
(a)
Cache Block Partitioned into Chunks
1
2
...
128
Communication Wires and FIFO Queues
65
1
66
2
...
128
64
(b)

[Bojnordi'13]

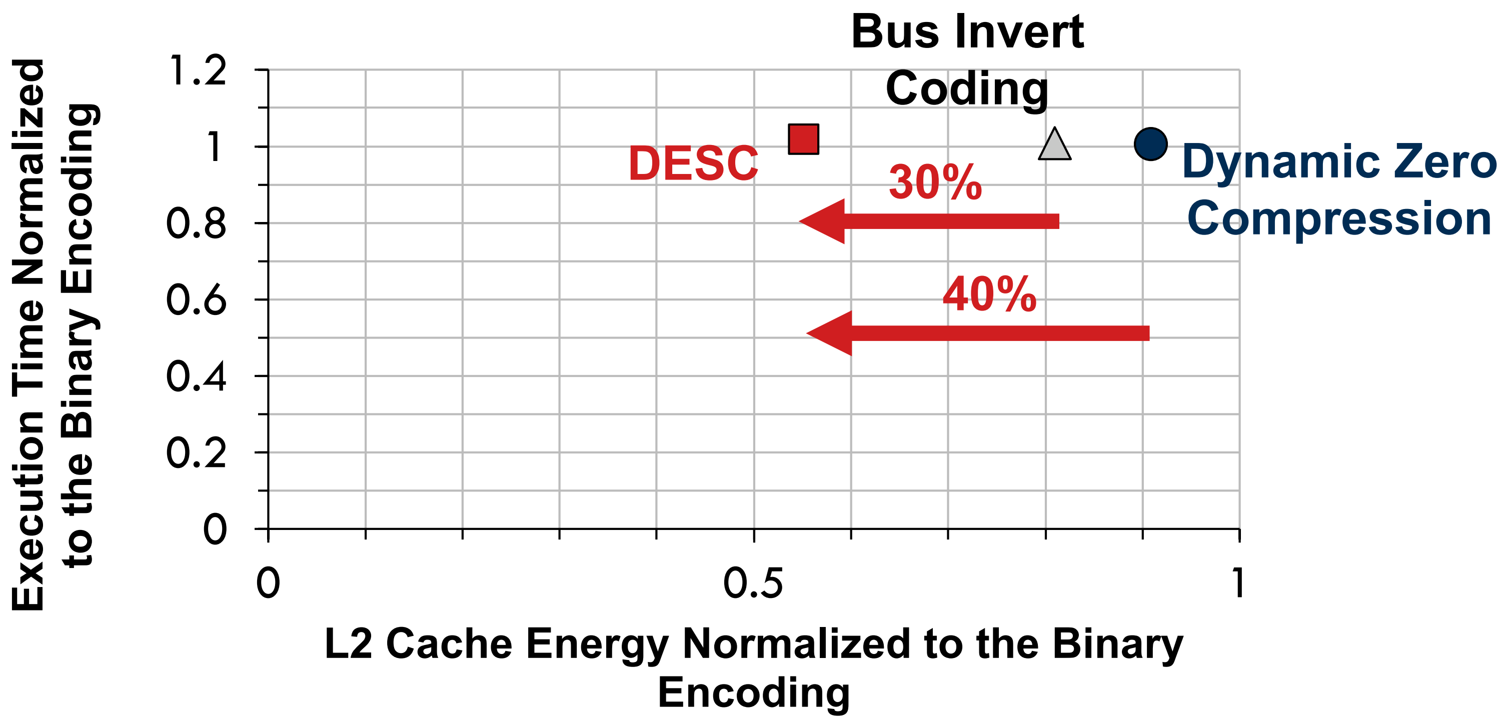
Time-Based Data Transfer



[Bojnordi'13]

Time-Based Data Transfer

- L2 cache energy is reduced by 1.8x at the cost of less than 2% increase in the execution time.



[Bojnordi'13]