

Software-Controlled Memory Compression Using Informing Memory Operations

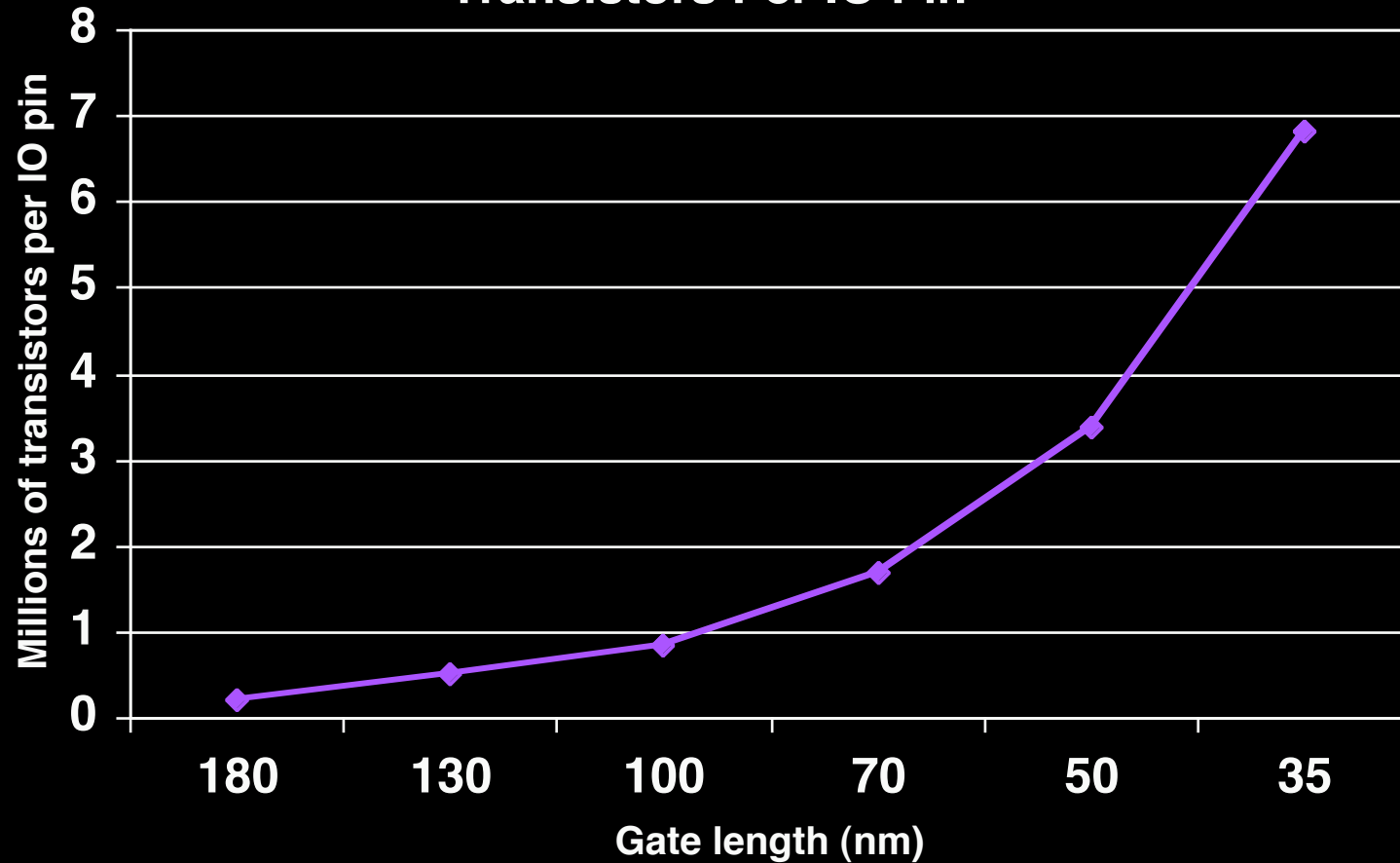
Bert Maher and Katie Coons

Outline

- Motivation - A case for compression in CMPs
- Data-specific Compression
- Informing Memory Operations
- Conclusions and Future Work

Motivation

Transistors Per IO Pin



“In a 35nm technology there will be 45x more transistors per IO pin than in a 180nm technology.”

Emerging CMP Challenges

- Transistor speed increasing faster than DRAM latency or pin bandwidth

Rate of increase per year (ITRS Roadmap 2004)	
Transistor speed	21%
Pin bandwidth	11%
DRAM latency	10%

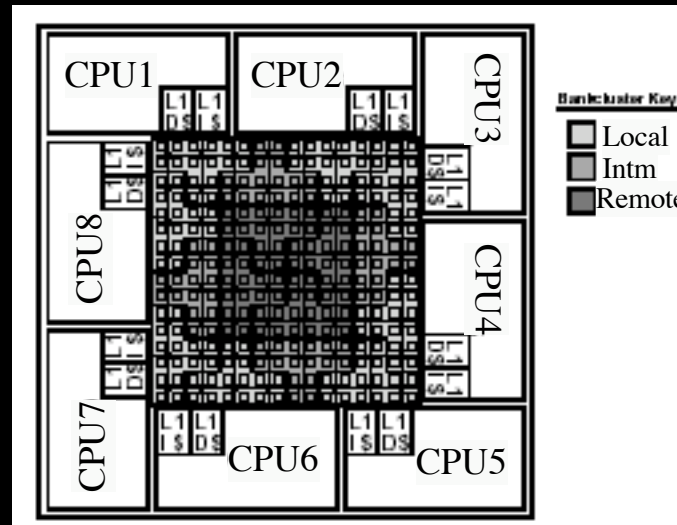
- Use more cores to tolerate DRAM latency with TLP? ← Increases off-chip bandwidth demand
- Use caches to avoid DRAM accesses and conserve off-chip bandwidth?
↑ Cache capacity becomes a bottleneck

Emerging CMP Challenges

- Increasing wire delay

- NUCA?
- D-NUCA?

Power hungry



From "Managing Wire Delay in Large Chip-Multiprocessor Caches"
By Beckmann and Wood

- Transmission Line Caches (TLC)?

- Hardware Prefetching?

Restricted bandwidth



Increases off-chip bandwidth demand significantly

Why use compression?

- Increase capacity...
 - More RAM, L2 cache, or off-chip pin bandwidth
- ... without increasing costs
 - \$\$\$, chip area, power

Basic tradeoff: Increase capacity at the price of increased access latency

Where is compression useful?

- L1 cache - Power savings (embedded systems)
- L2 cache - Fewer off-chip misses, more cache capacity in less die area
- Off-chip link - Bandwidth
- Main memory - Increase memory size for free

Which make the most sense in a CMP?

Not as simple as it sounds...

Choice of compression algorithm
significantly impacts benefits

Outline

- Motivation - A case for compression in CMPs
- Data-specific Compression
- Informing Memory Operations
- Conclusions and Future Work

Data-specific Compression

- Compression algorithms rely on expected regularities in data
- Integer, floating-point, pointer, character data all compress differently
- Application-specific data compression may be even more effective

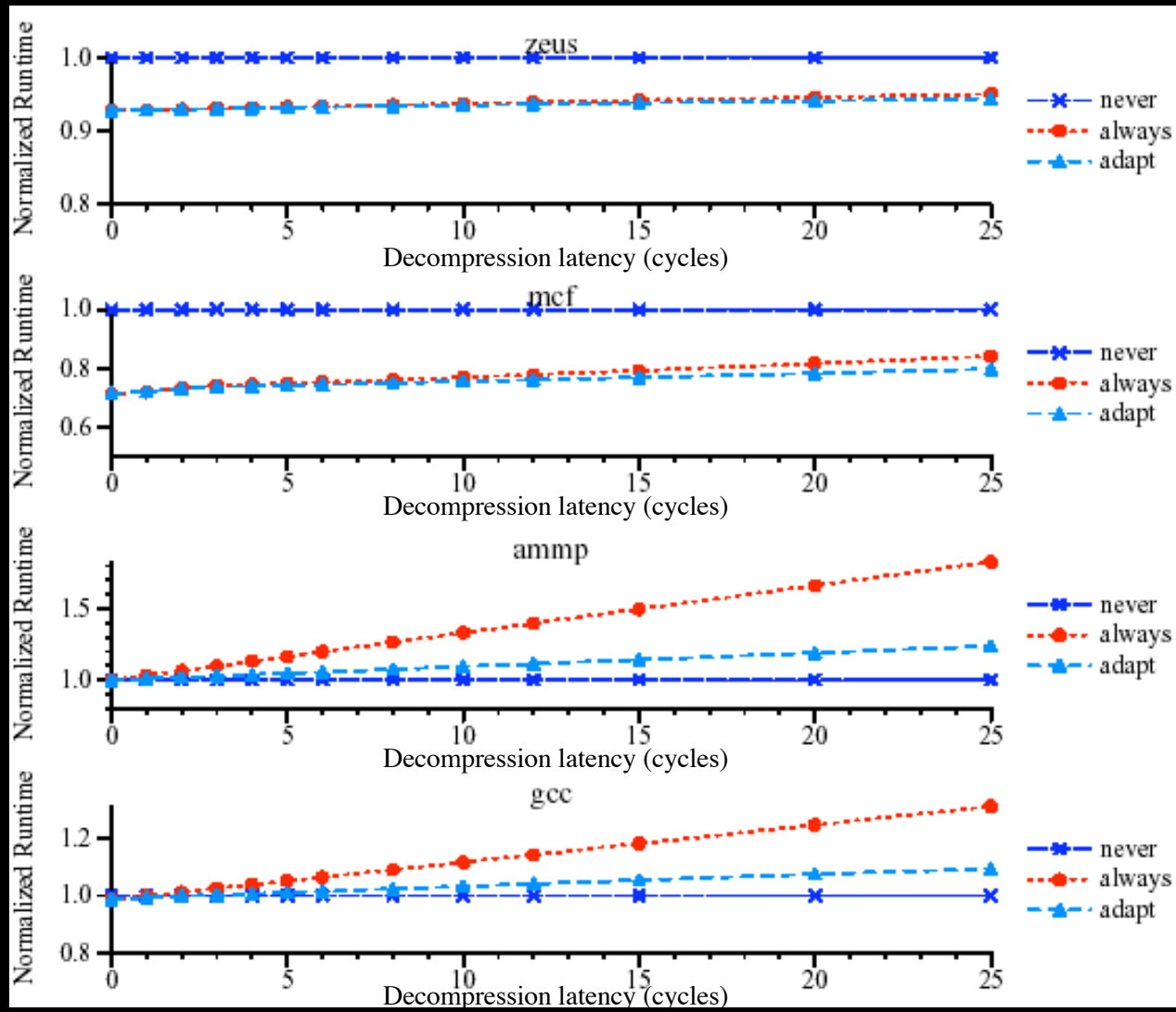
“The main key to good compression is having the right kinds of expectations for the data at hand.”

-- Wilson, Kaplan, Smaragdakis

Data-specific Compression

- Benefit: Significantly higher compression ratios
- Cost: Potentially higher access latencies
- Hybrid solutions:
 - Provide multiple hardware schemes (run-length, dictionary based) as well as a software trap
 - Allow dictionary to be specified as input
 - Provide ISA support for fast compression

Decompression time sensitivity



From Alaa Alameldeen's PhD thesis - "Using Compression to Improve Chip Multiprocessor Performance"

Data-specific Compression

- What we need:
 - Well-balanced compression ratio, compression latency
 - Exploit data-specific knowledge
- What we mostly see today:
 - Data-specific with very high compression ratio, very high latency
 - General-purpose with very fast compression, low compression ratio

Data-specific compression

- Floating point geometry
 - Compress structural data with predictive methods
 - 12MB/s, 33% compression rate
- Inverted indices
 - Byte-oriented Huffman
 - 500 KB/s
- Clearly, these are designed for max compression ratio, not speed

Outline

- Motivation - A case for compression in CMPs
- Data-specific Compression
- Informing Memory Operations
- Conclusions and Future Work

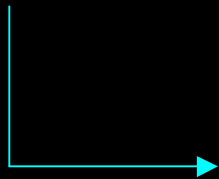
Architectural support

- Informing memory operations
 - Perform a software trap on L2 miss
- Compress data ahead of time
 - No mechanism to compress on evict
- Decompress line on cache miss
 - Line granularity limits compression

TLB-Based Informing Memory Ops

<VA>

Tag	Page	Comp. Type
		LZW
		X-RL



<DATA>



Compressor



<COMPRESSED>

Outline

- Motivation - A case for compression in CMPs
- Data-specific Compression
- Informing Memory Operations
- Conclusions and Future Work

Conclusions

- Compression is not a long-term solution
 - But it does have short-term advantages
- Compression does provide relatively free extra capacity
 - Why not increase your memory size for free?
- Data-specific compression could increase these benefits, but...
 - Appropriate algorithms?