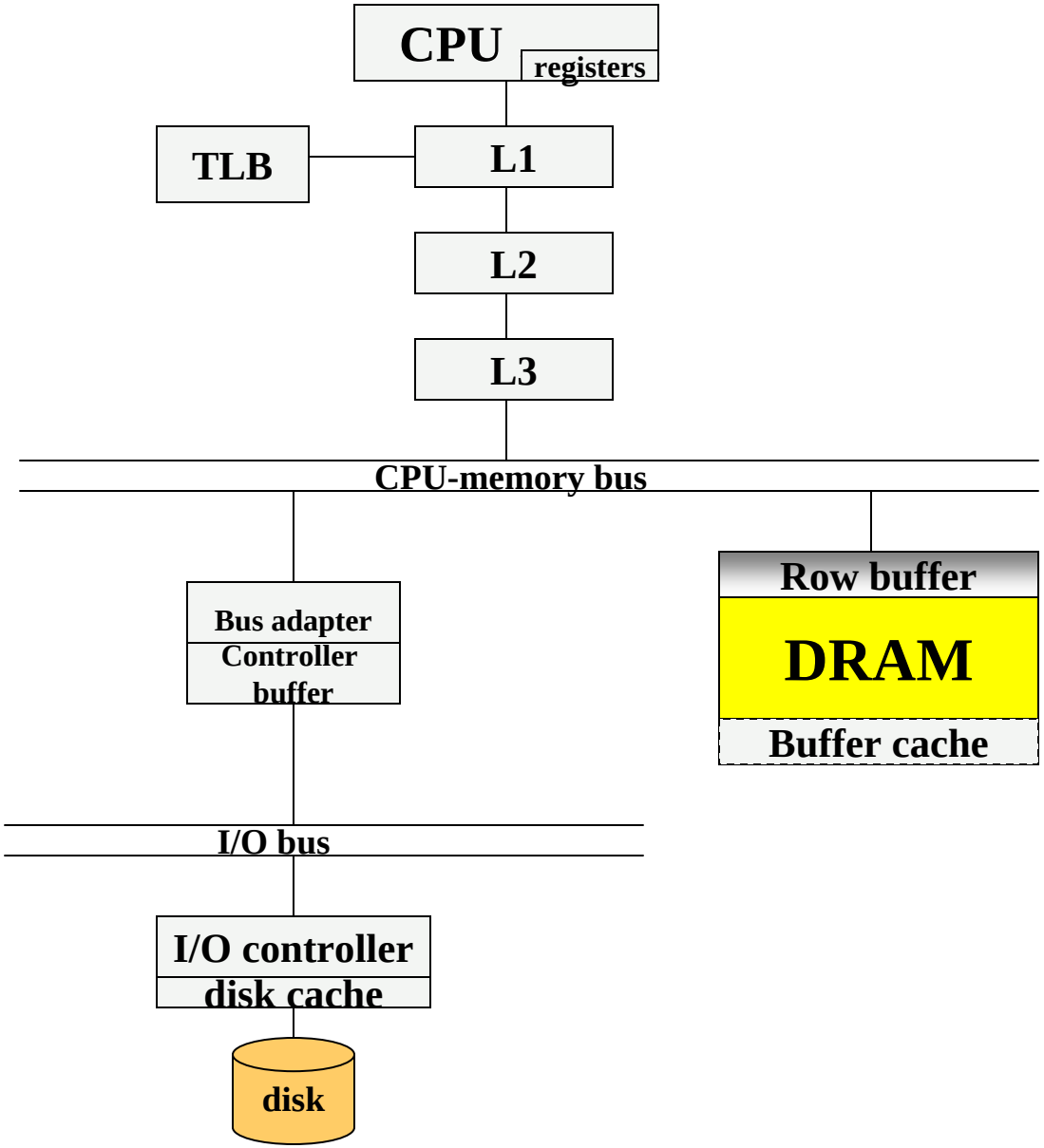
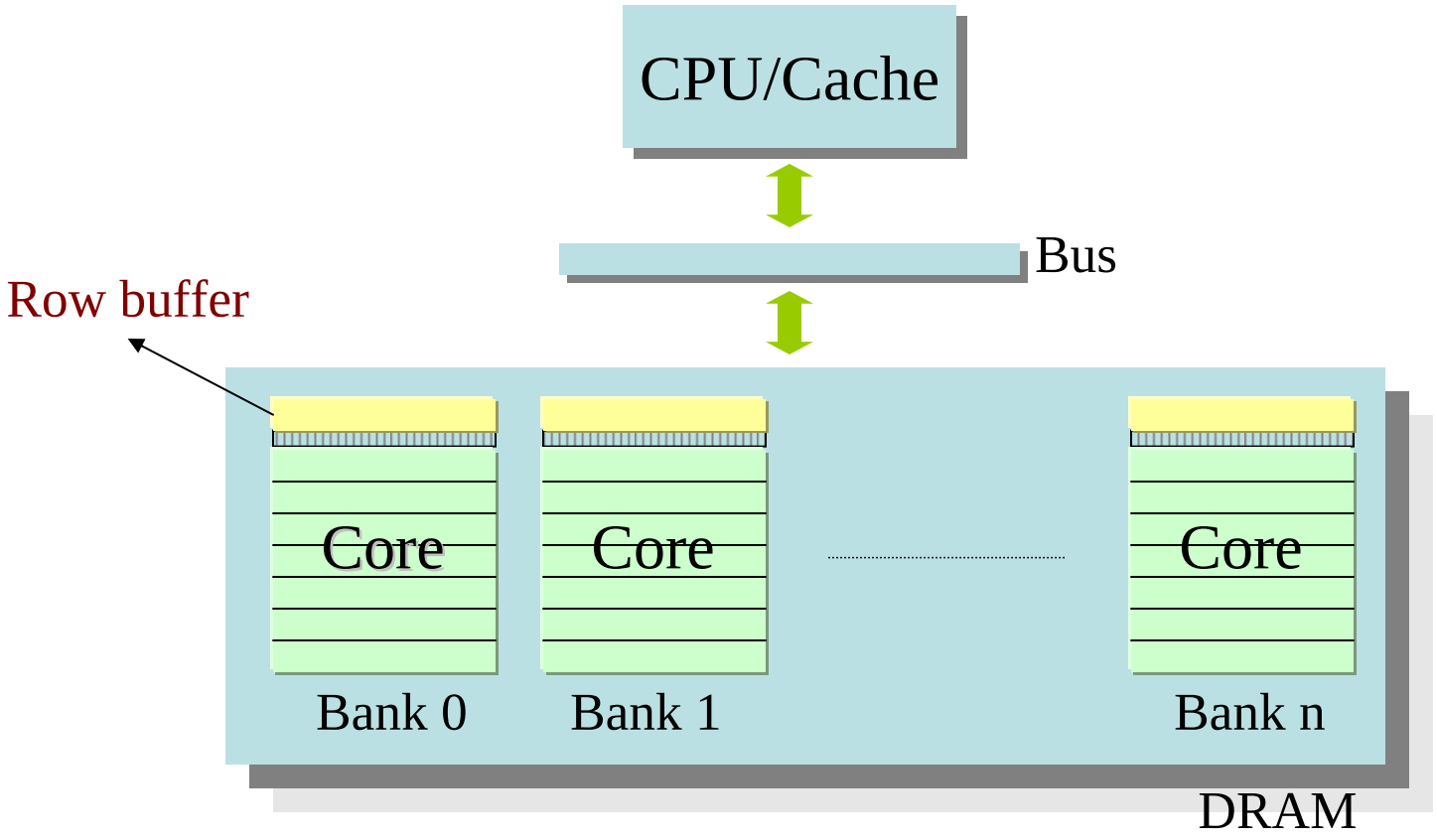
Lecture 7: Caching in Row-Buffer of DRAM

Adapted from “A Permutation-based Page Interleaving Scheme: To Reduce Row-buffer Conflicts and Exploit Data Locality” by x. Zhang et. al.

A Bigger Picture



DRAM Architecture

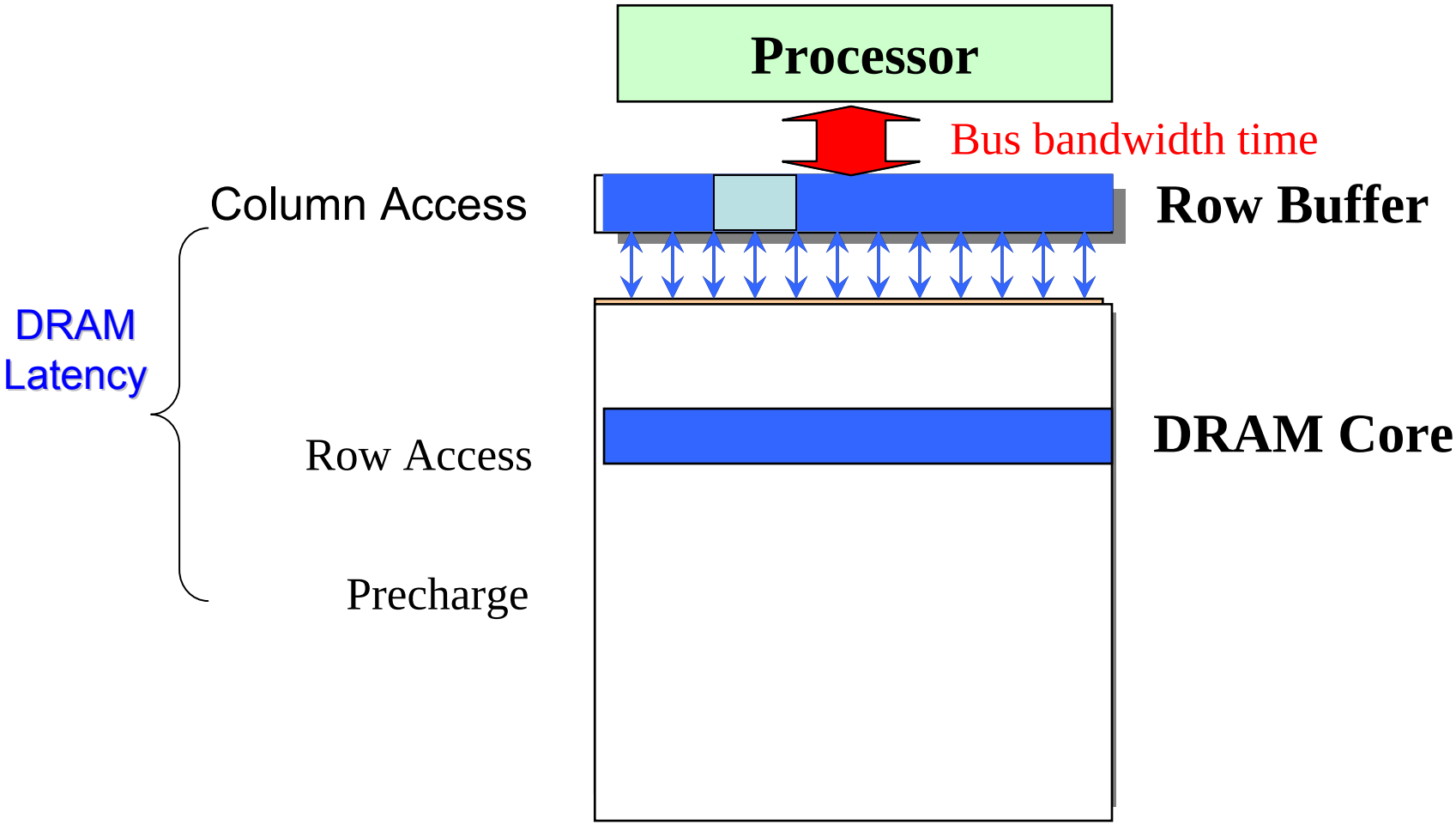


Caching in DRAM

- DRAM is the center of memory hierarchy:
  - High density and high capacity
  - Low cost but slow access (compared to SRAM)
- A cache miss has been considered as a constant delay for long time. This is wrong.
  - Non-uniform access latencies exist within DRAM
- Row-buffer serves as a fast cache in DRAM
  - Its access patterns here have been paid little attention.
  - Reusing buffer data minimizes the DRAM latency.

DRAM Access

- Precharge: charge a DRAM bank before a row access
- Row access: activate a row (page) of a DRAM bank
- Column access: select and return a block of data in an activated row
- Refresh: periodically read and write DRAM to keep data



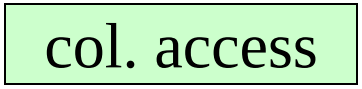
Row buffer misses come from a sequence of accesses to different pages in the same bank.

When to Precharge --- Open Page vs. Close Page

- Determine when to do precharge.
- Close page: starts precharge after every access
  - May reduce latency for row buffer misses
  - Increase latency for row buffer hits
- Open page: delays precharge until a miss
  - Minimize latency for row buffer hits
  - Increase latency for row buffer misses
- Which is good? depends on row buffer miss rate.

Non-uniform DRAM Access Latency

- Case 1: Row buffer hit (20+ ns)

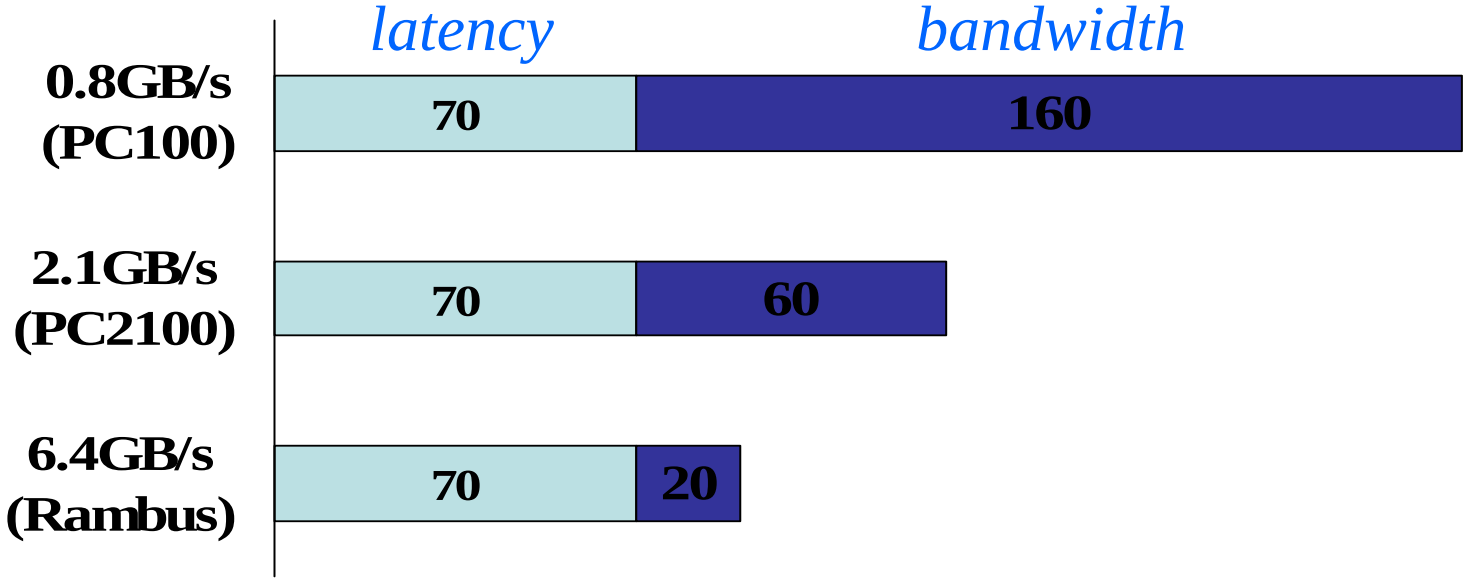


- Case 2: Row buffer miss (core is precharged, 40+ ns)

- Case 3: Row buffer miss (not precharged, ≈ 70 ns)

Amdahl's Law applies in DRAM

- Time (ns) to fetch a 128-byte cache block:



- As the bandwidth improves, DRAM latency will decide cache miss penalty.

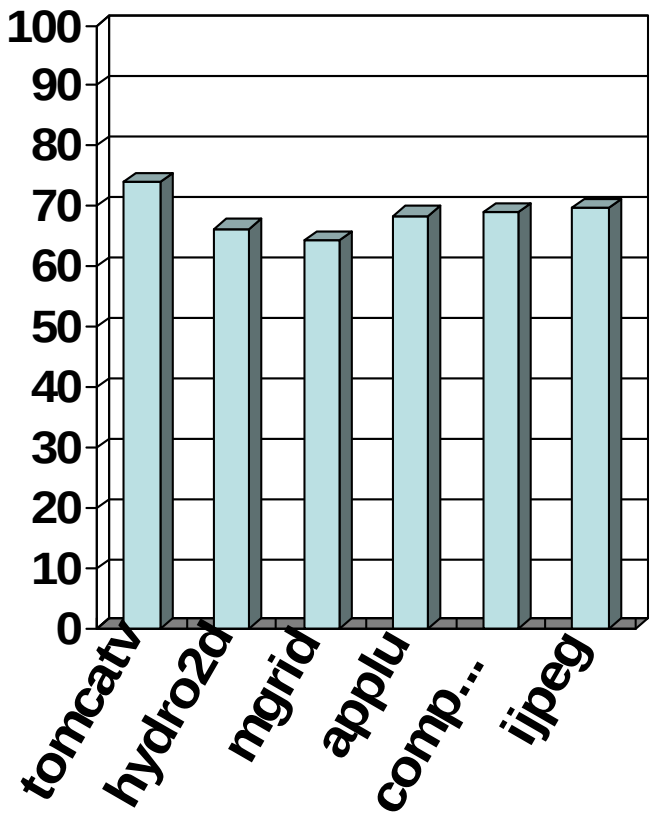
Row Buffer Locality Benefit

$$Latency_{\text{row buffer hit}} < Latency_{\text{row buffer miss}}$$

Reduce latency by up to 67%.

Objective: serve memory requests without accessing the DRAM core as much as possible.

SPEC95: Miss Rate to Row Buffer

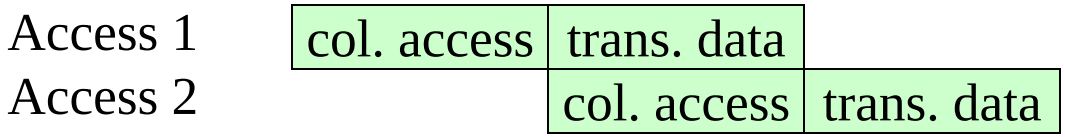


- Specfp95 applications
- Conventional page interleaving scheme
- 32 DRAM banks, 2KB page size

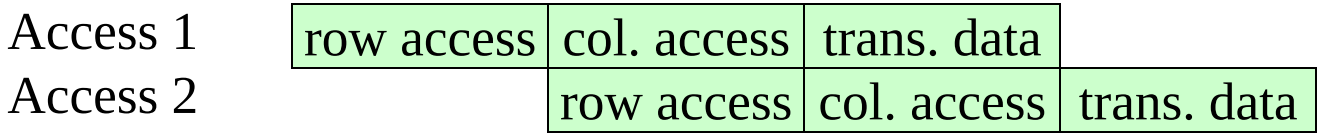
- **Why is it so high?**
- **Can we reduce it?**

Effective DRAM Bandwidth

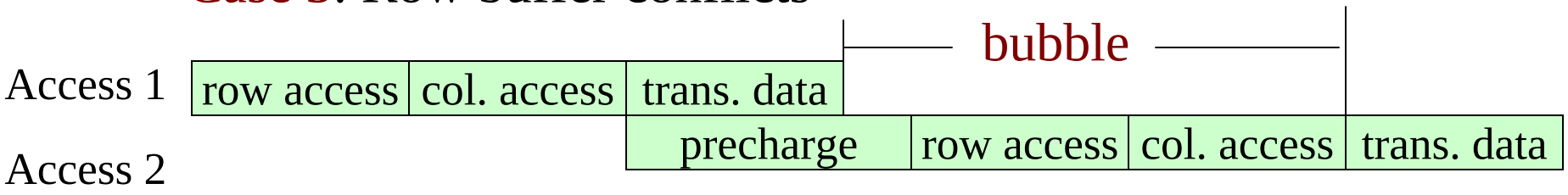
- Case 1: Row buffer hits



- Case 2: Row buffer misses to different banks



- Case 3: Row buffer conflicts

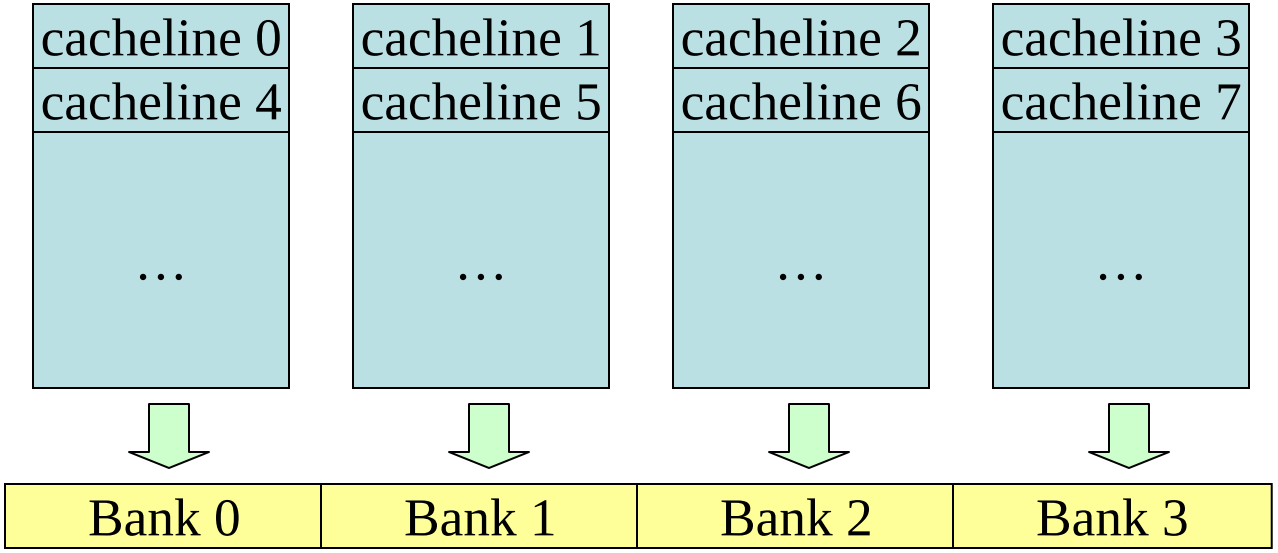


Parameters in a Memory System

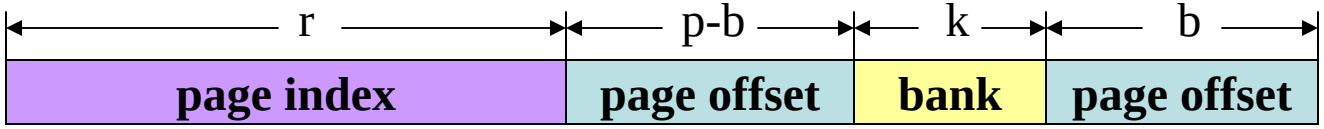
Parameter	Parameter descriptions
m	the length of the memory address in bits.
Cache-related	**Parameter descriptions**
C	the cache size in bytes.
S	the number of sets in the cache.
N	the number of blocks in a set.
B	the block size in bytes.
s	the length of the cache set index in bits. $s = \log S = \log C/(BN)$.
b	the length of the cache block offset in bits. $b = \log B$.
t	the length of the cache tag in bits. $t = m - (s + b)$.
Memory-related	**Parameter descriptions**
K	the number of memory banks.
P	the page size in bytes, which is also the size of the row_buffer.
R	the number of pages (rows) in a memory bank.
k	the length of the memory bank index in bits. $k = \log K$.
p	the length of the page offset in bits. $p = \log P$.
r	the length of the page index in bits. $r = \log R = m - (k + p)$.

Table 1: Parameters of a memory system.

Conventional Data Layout in DRAM
---- Cacheline Interleaving

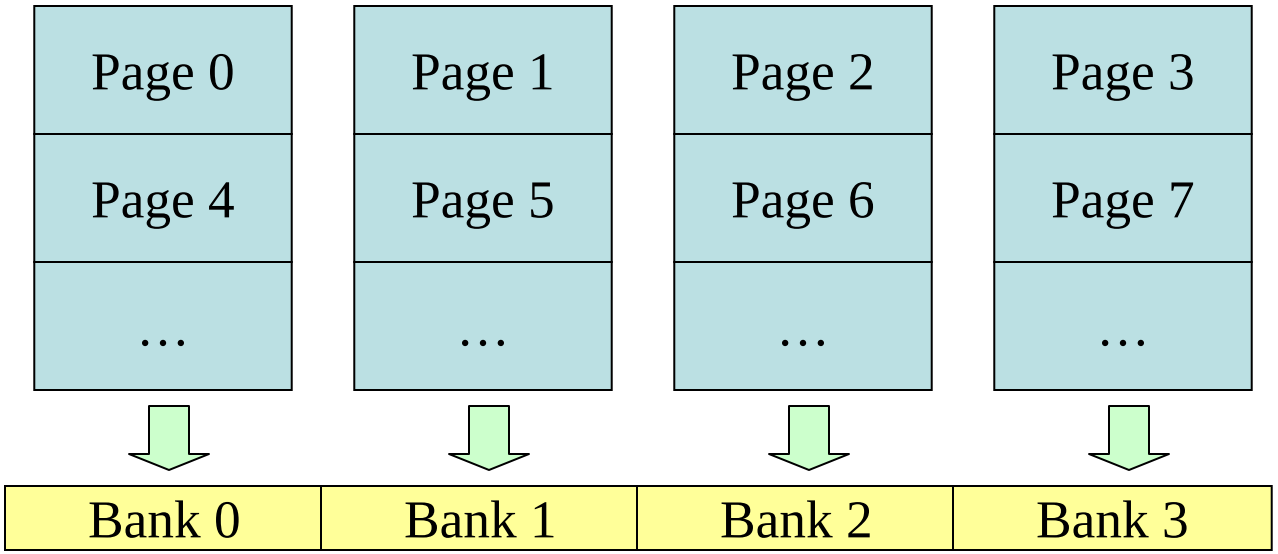


Address format

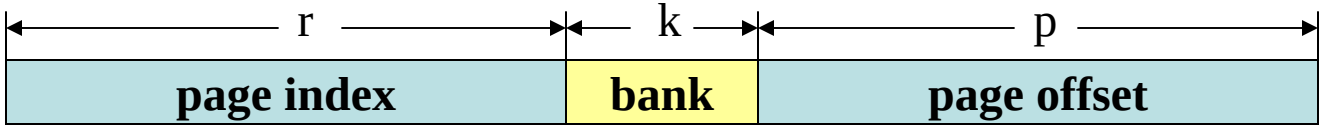


Spatial locality is not well preserved!

Conventional Data Layout in DRAM
---- Page Interleaving

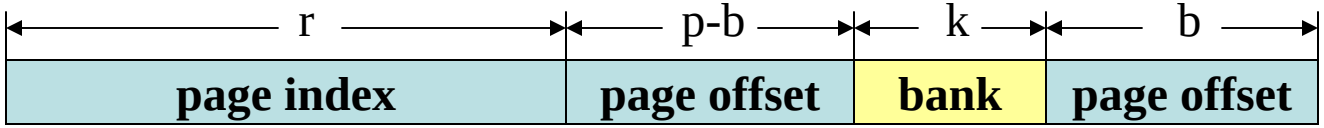


Address format

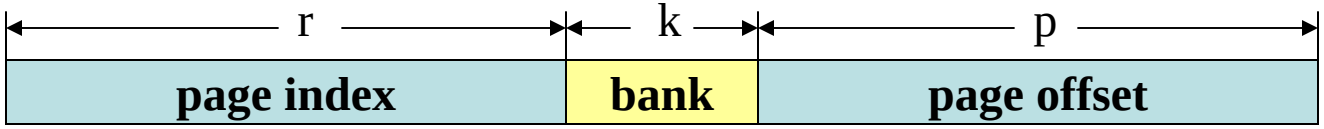


Compare with Cache Mapping

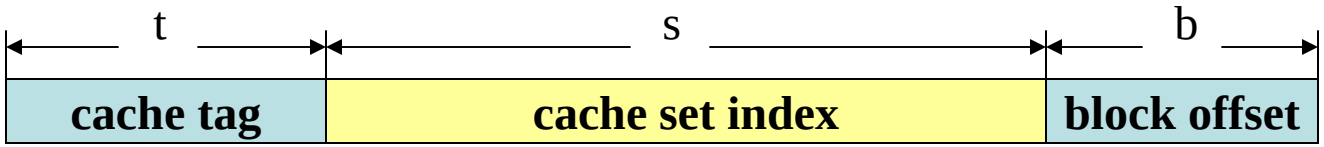
Cache line interleaving



Page interleaving



Cache-related representation



1. Observation: bank index $\subseteq$ cache set index
2. Inference: $\forall x \forall y$, x and y conflict on cache $\Rightarrow$ x and y conflict on row buffer

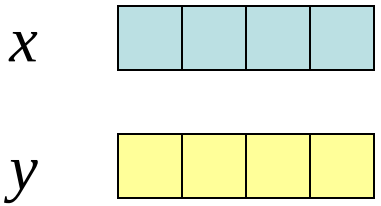
Sources of Row-Buffer Conflicts --- L2 Conflict Misses

- L2 conflict misses may result in severe row buffer conflicts.

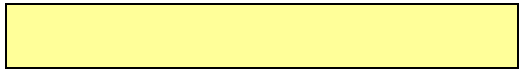
Example: assume x and y conflicts on a direct mapped cache (address distance of X[0] and y[0] is a multiple of the cache size)

```
sum = 0;
for (i = 0; i < 4; i ++)
    sum += x[i] + y[i];
```

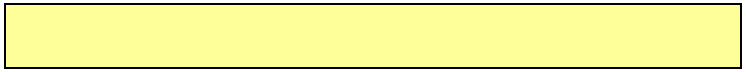
Sources of Row-Buffer Conflicts --- L2 Conflict Misses (Cont'd)



Cache line that x,y resides



Row buffer that x,y resides



Cache misses 8

Row buffer misses 8

Thrashing at both cache and row buffer!

Sources of Row-Buffer Conflicts --- L2 Writebacks

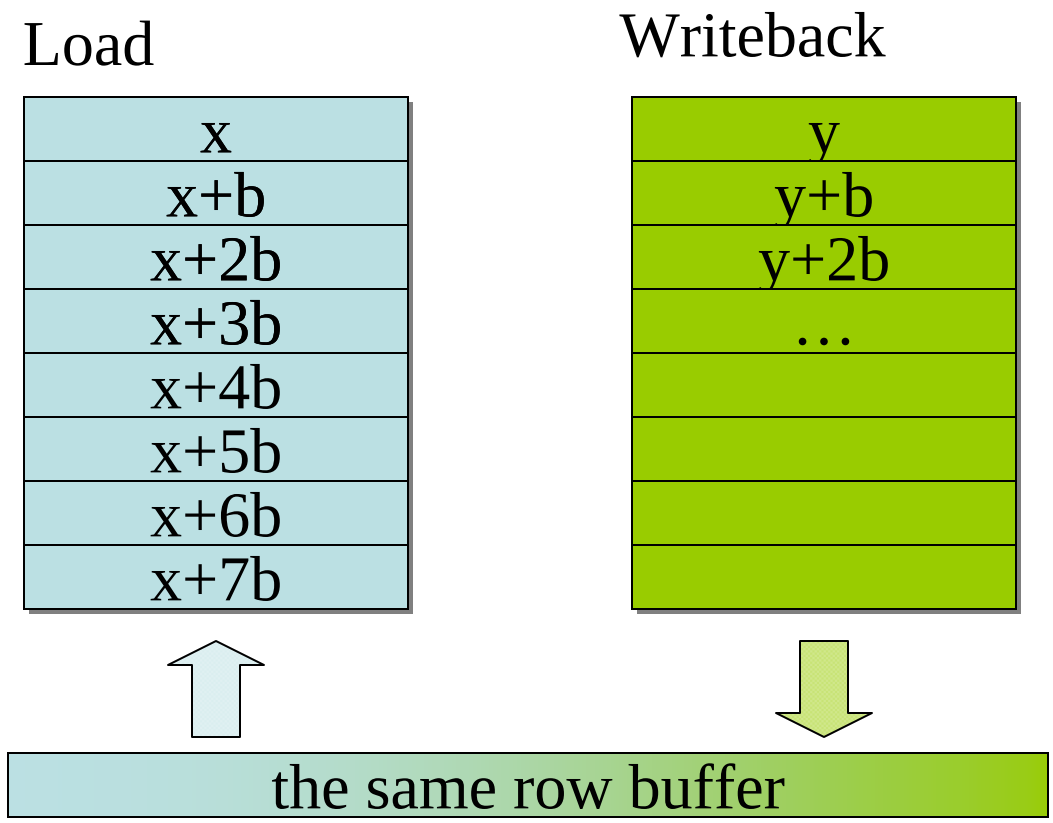
- Writebacks interfere reads on row buffer
  - Writeback addresses are L2 conflicting with read addresses

Example: assume writeback is used

(address distance of X[0] and y[0] is a multiple of the cache size)

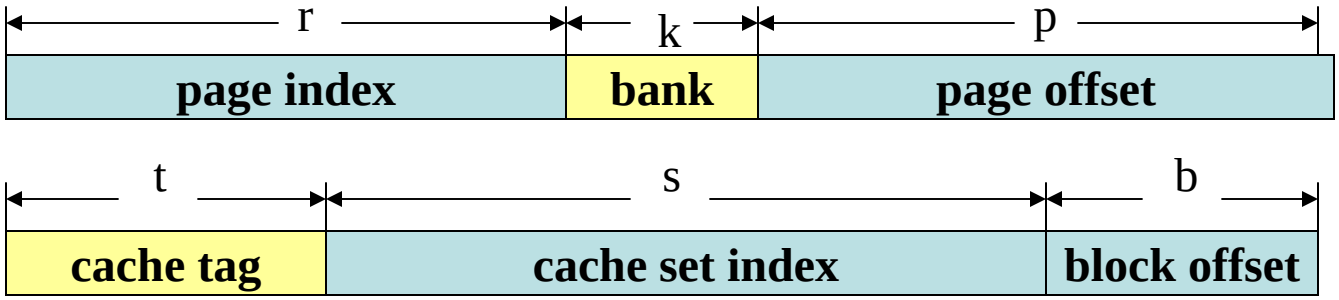
```
for (i = 0; i < N; i ++)
    y[i] = x[i];
```

Sources of Row-Buffer Conflicts --- L2 Writebacks (Cont'd)

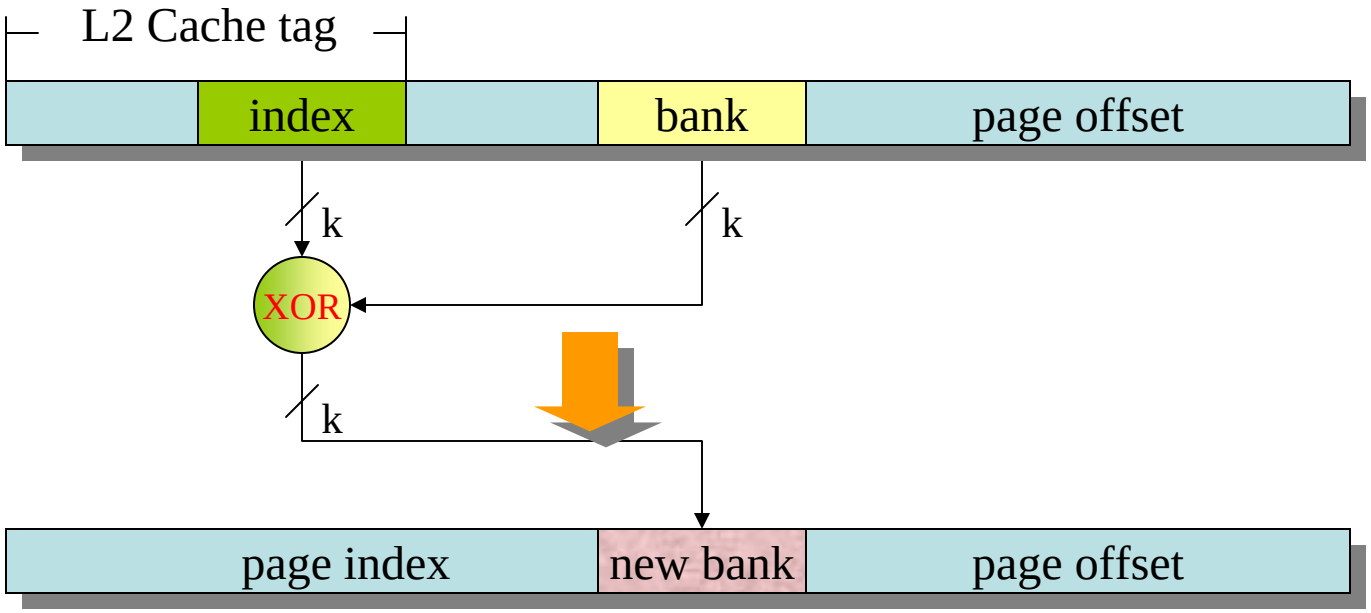


Key Issues

- To exploit spatial locality, we should use maximal interleaving granularity (or row-buffer size).
- To reduce row buffer conflicts, we cannot use only those bits in cache set index for “bank bits”.

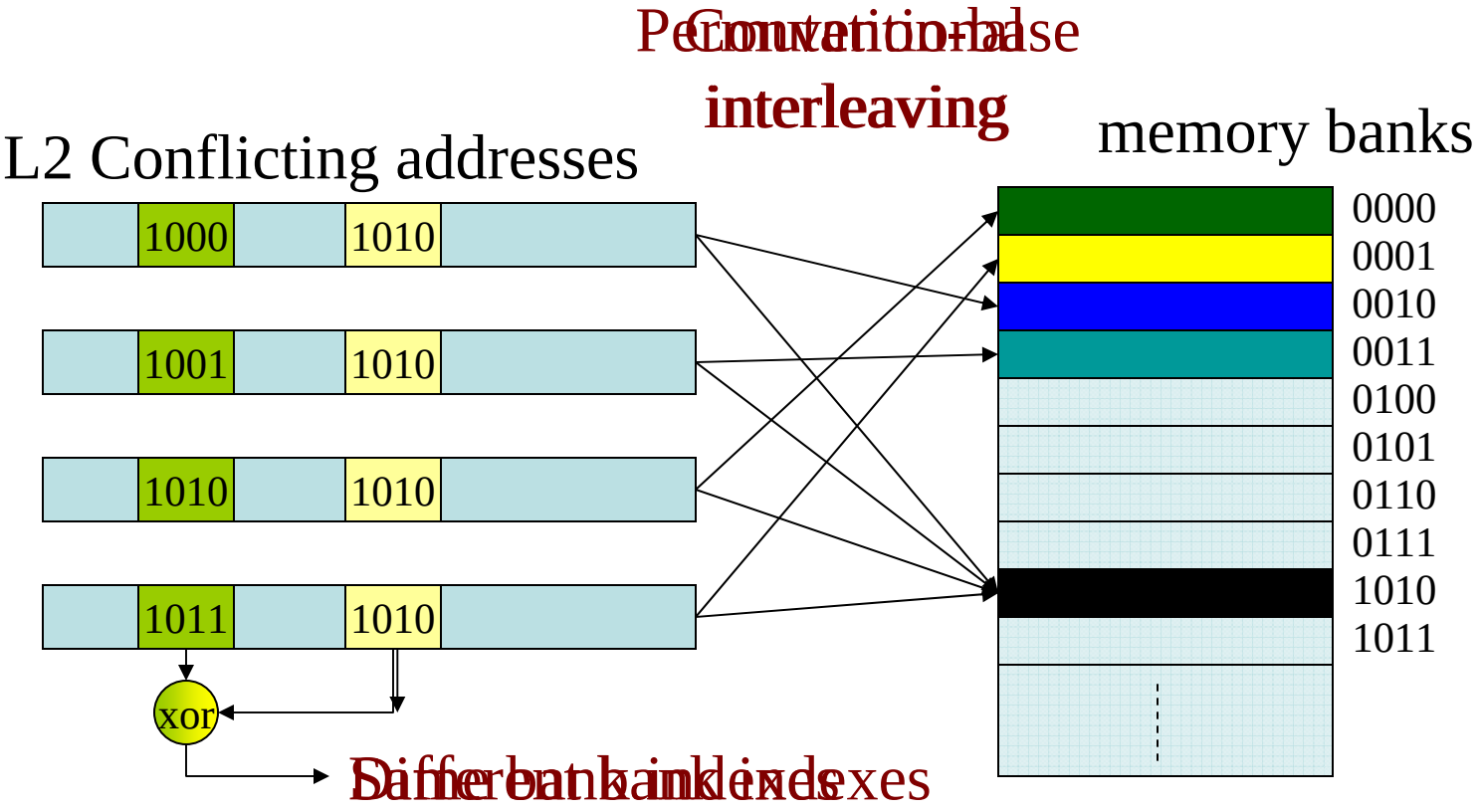


Permutation-based Interleaving



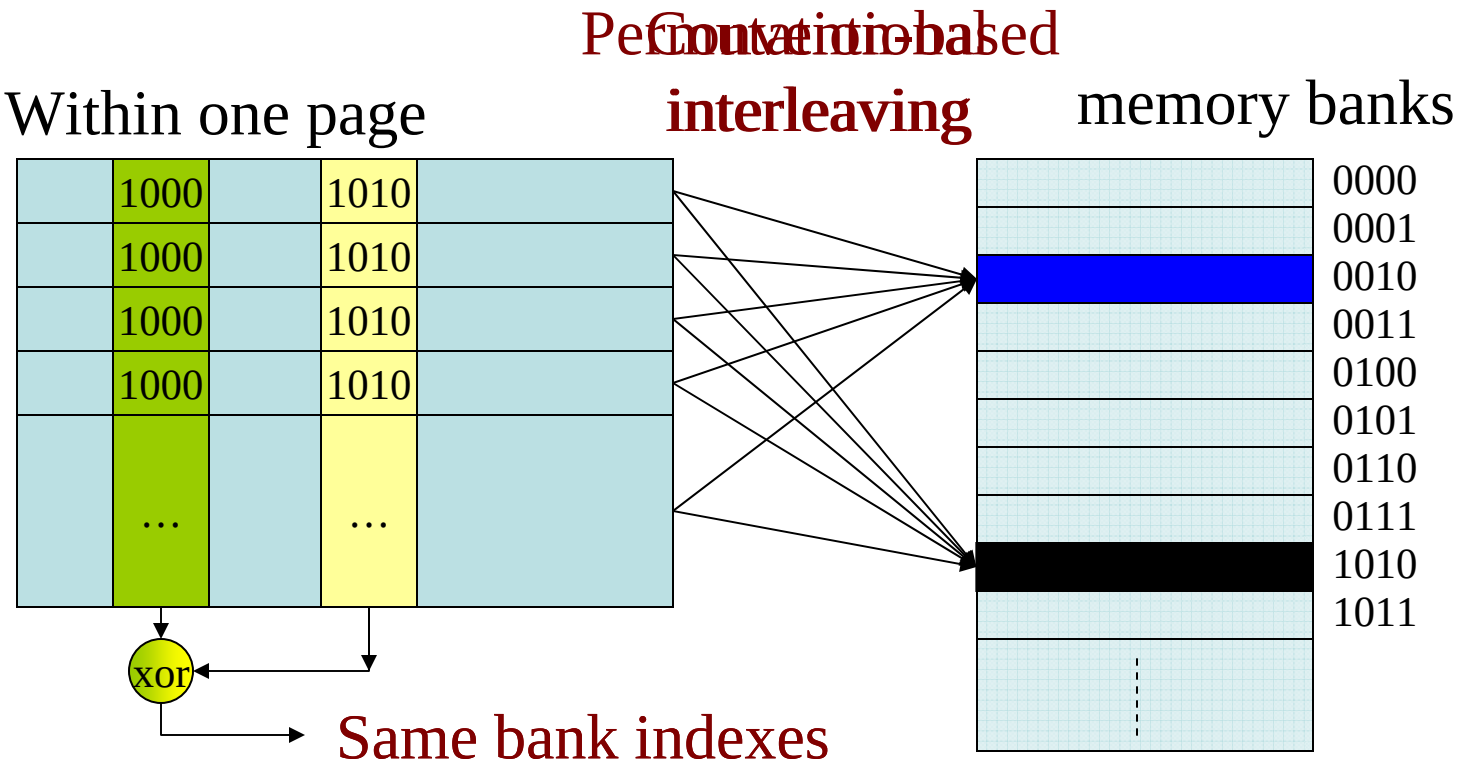
Scheme Properties (1)

- L2-conflicting addresses are distributed onto different banks



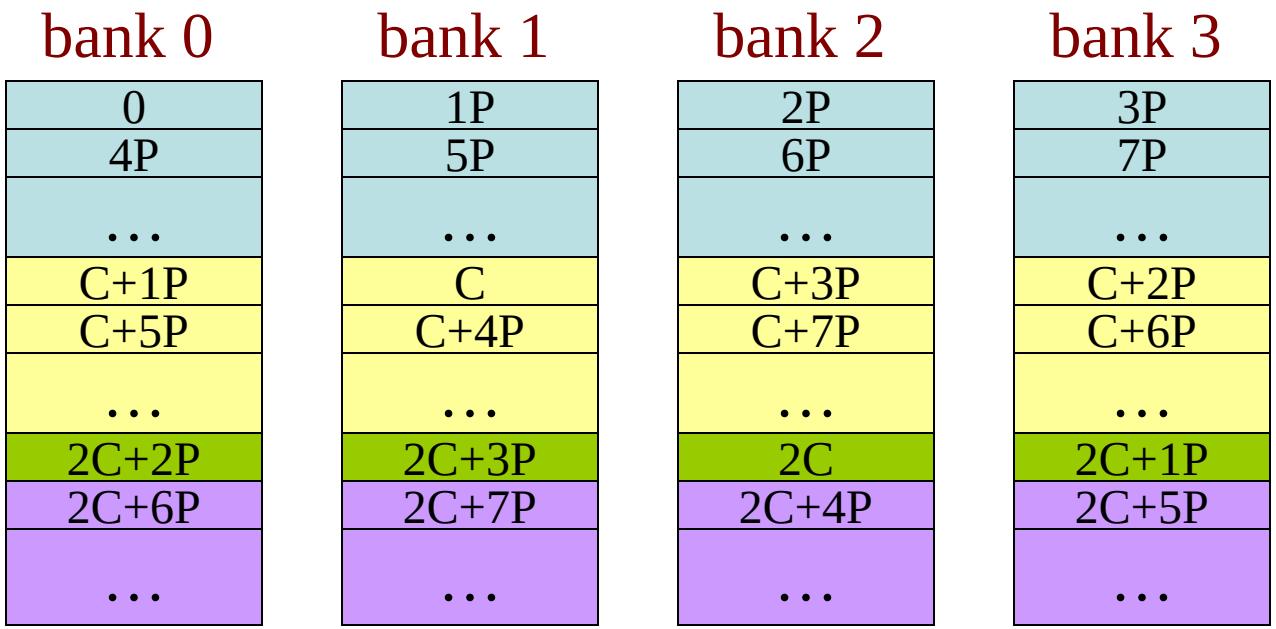
Scheme Properties (2)

- The spatial locality of memory references is preserved.



Scheme Properties (3)

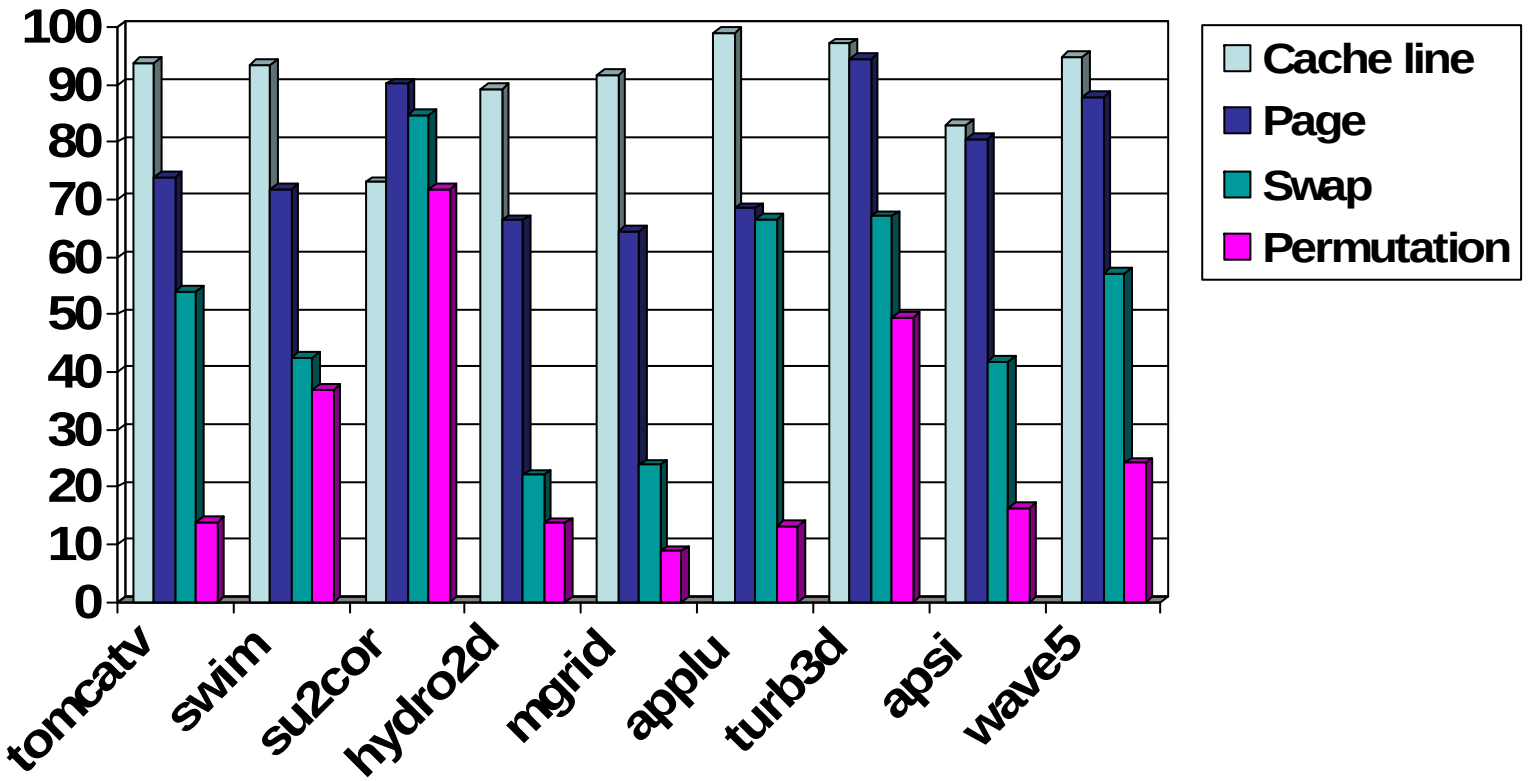
- Pages are uniformly mapped onto ALL memory banks.



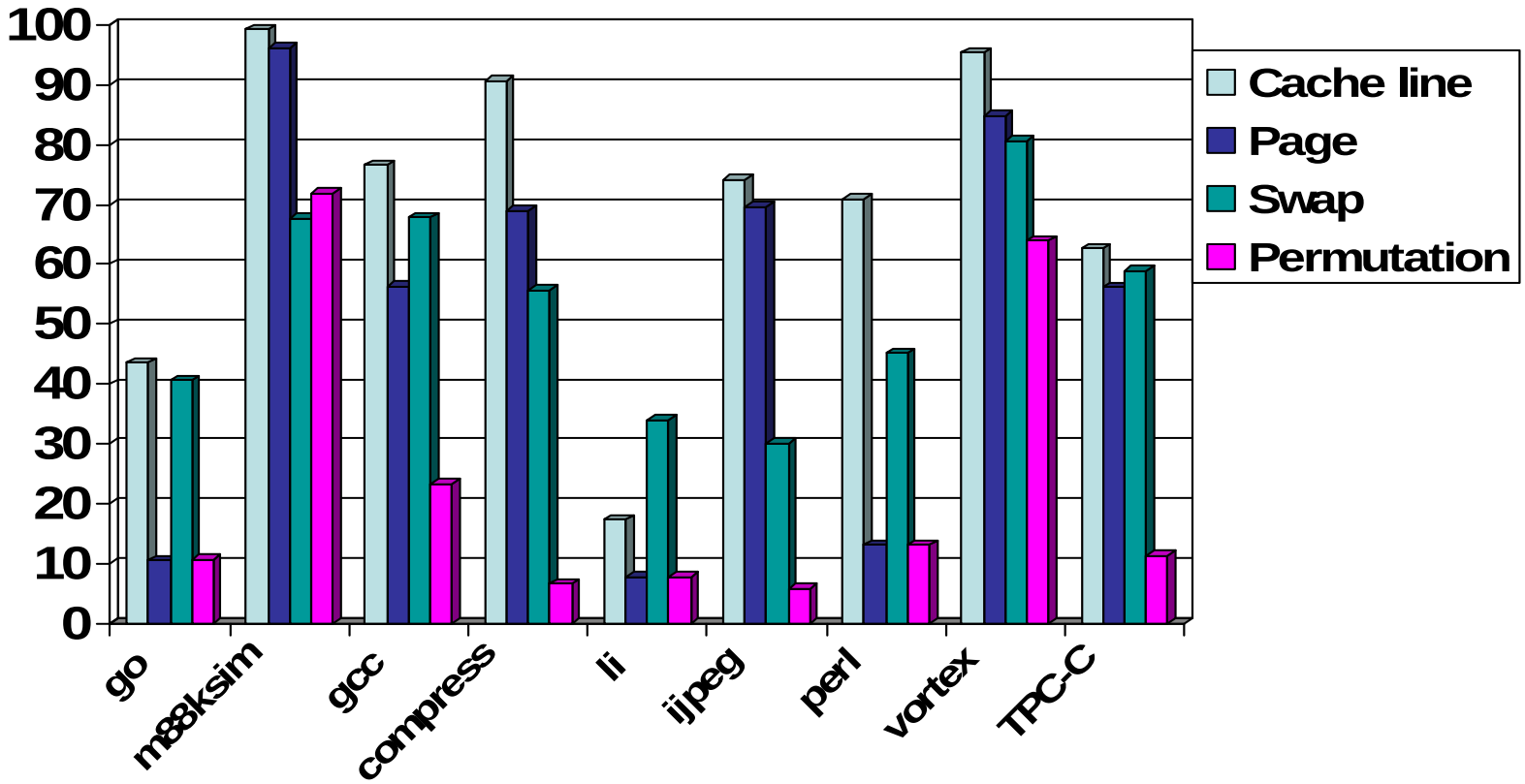
Experimental Environment

- SimpleScalar
- Simulate XP1000
- Processor: 500MHz
- L1 cache: 32 KB inst., 32KB data
- L2 cache: 2 MB, 2-way, 64-byte block
- MSHR: 8 entries
- Memory bus: 32 bytes wide, 83MHz
- Banks: 4-256
- Row buffer size: 1-8KB
- Precharge: 36ns
- Row access: 36ns
- Column access: 24ns

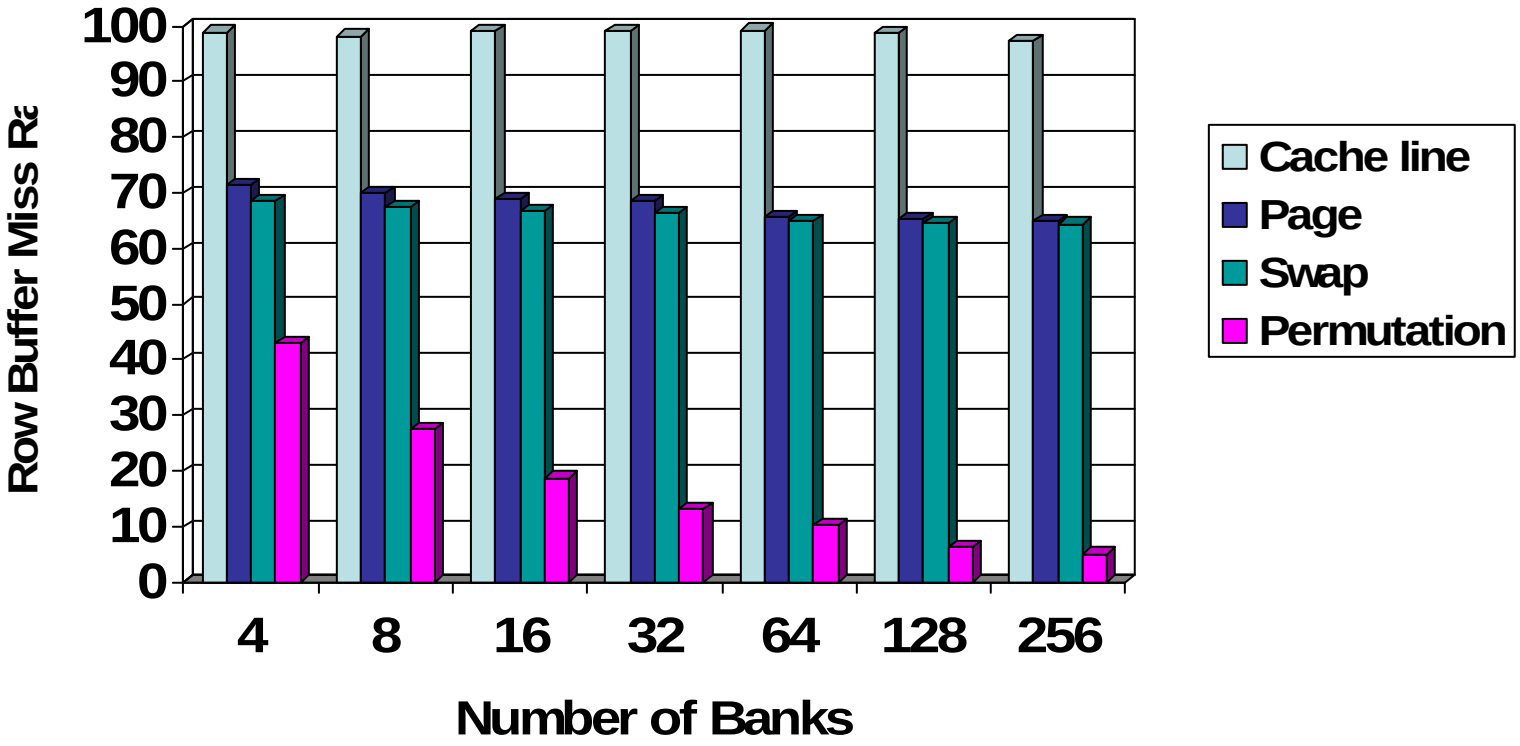
Row-buffer Miss Rate for SPECfp95



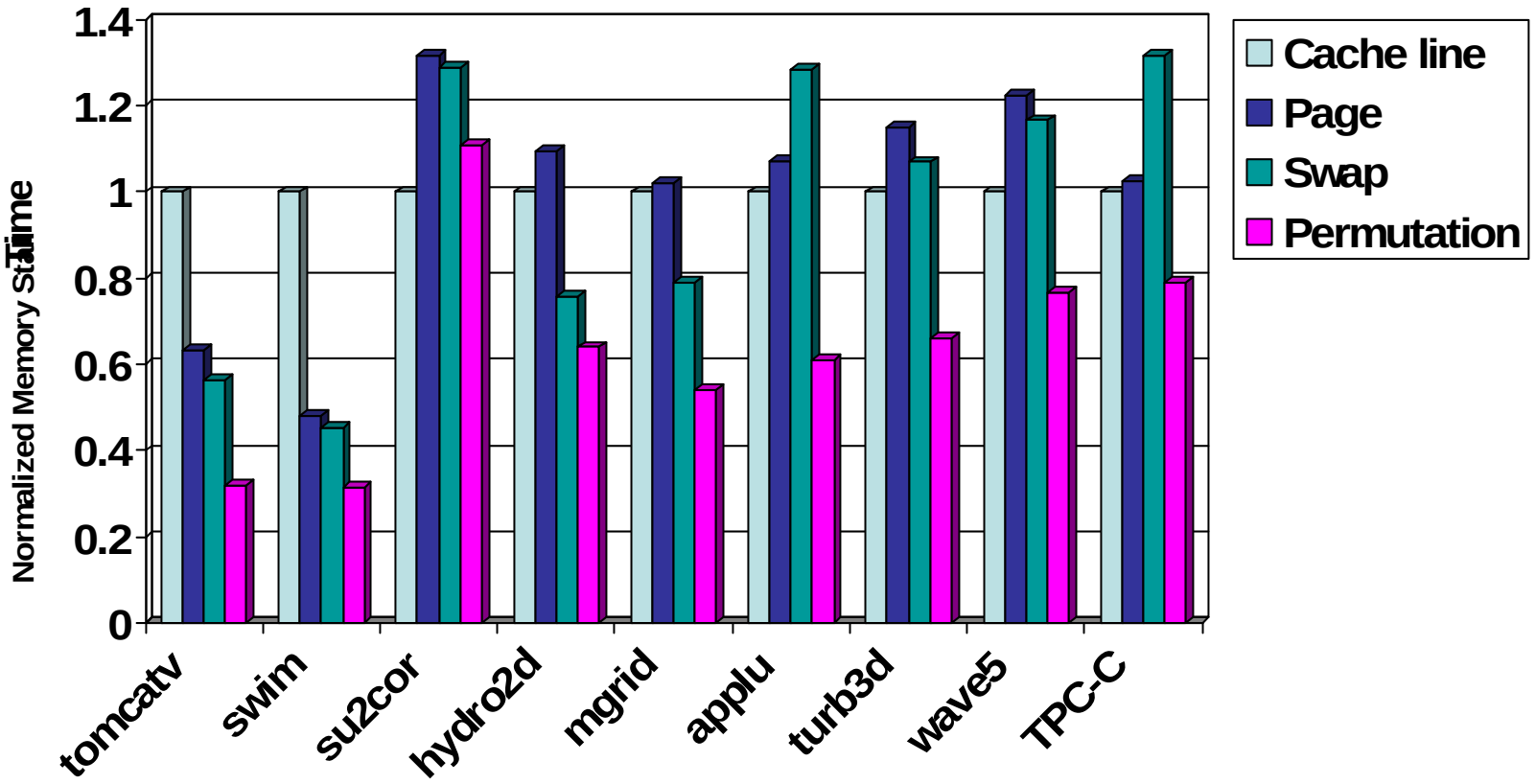
Miss Rate for SPECint95 & TPC-C



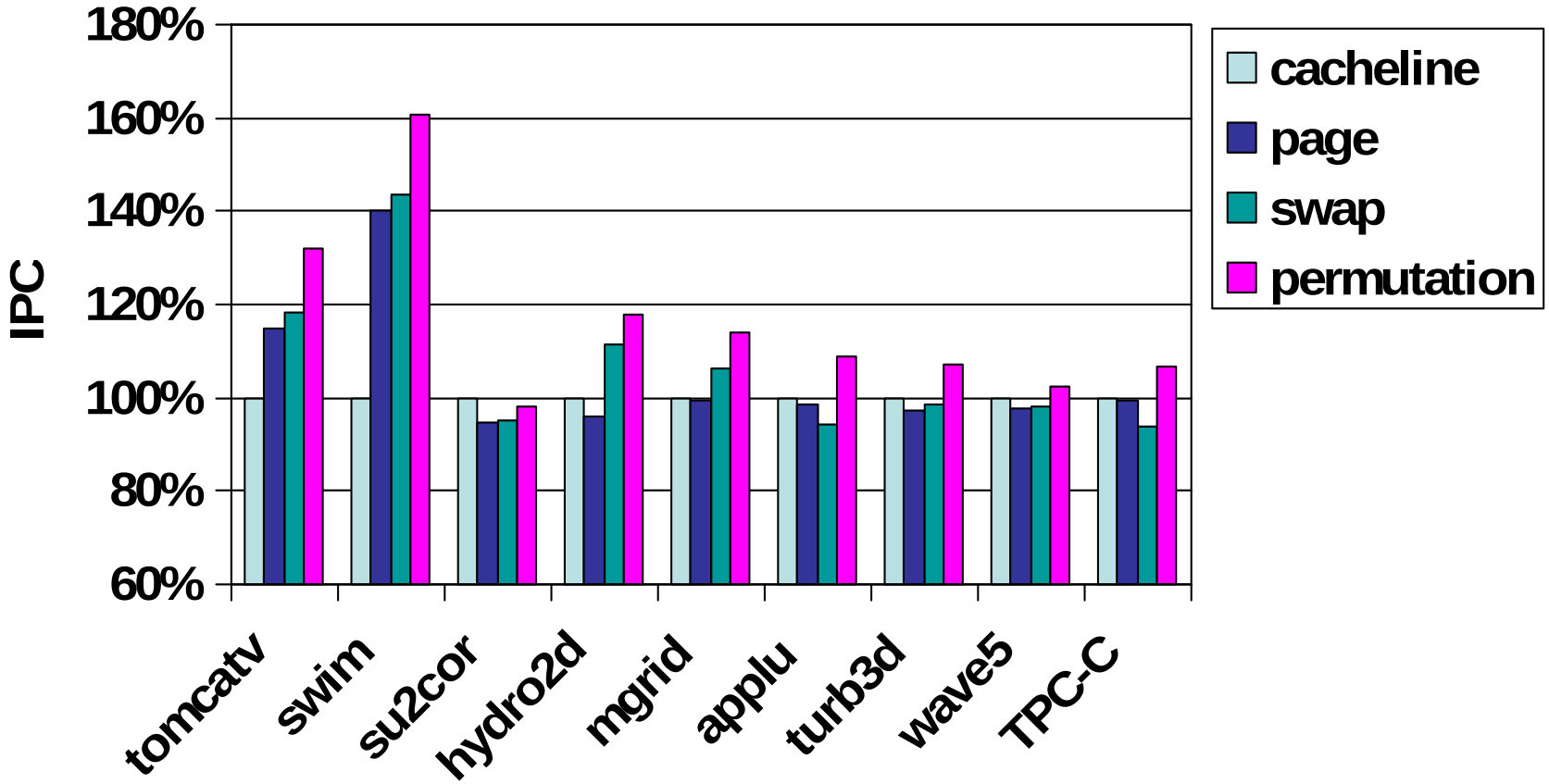
Miss Rate of *Applu*: 2KB Buf. Size



Comparison of Memory Stall Time



Improvement of IPC



Contributions of the Work

- We study interleaving for DRAM
  - DRAM has a row buffer as a natural cache
- We study page interleaving in the context of Superscalar processor
  - Memory stall time is sensitive to both latency and effective bandwidth
  - Cache miss pattern has direct impact on row buffer conflicts and thus the access latency
  - Address mapping conflicts at the cache level, including address conflicts and write-back conflicts, may inevitably propagate to DRAM memory under a standard memory interleaving method, causing significant memory access delays.
- Proposed permutation interleaving technique as a low-cost solution to these conflict problems.

Conclusions

- Row buffer conflicts can significantly increase memory stall time.
- We have analyzed the source of conflicts.
- Our permutation-based page interleaving scheme can effectively reduce row buffer conflicts and exploit data locality.