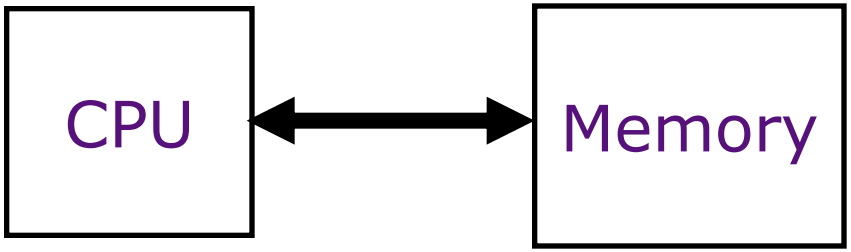
Cache Organization

Daniel Sanchez
Computer Science and Artificial Intelligence Laboratory
M.I.T.

http://www.csg.csail.mit.edu/6.823

CPU-Memory Bottleneck



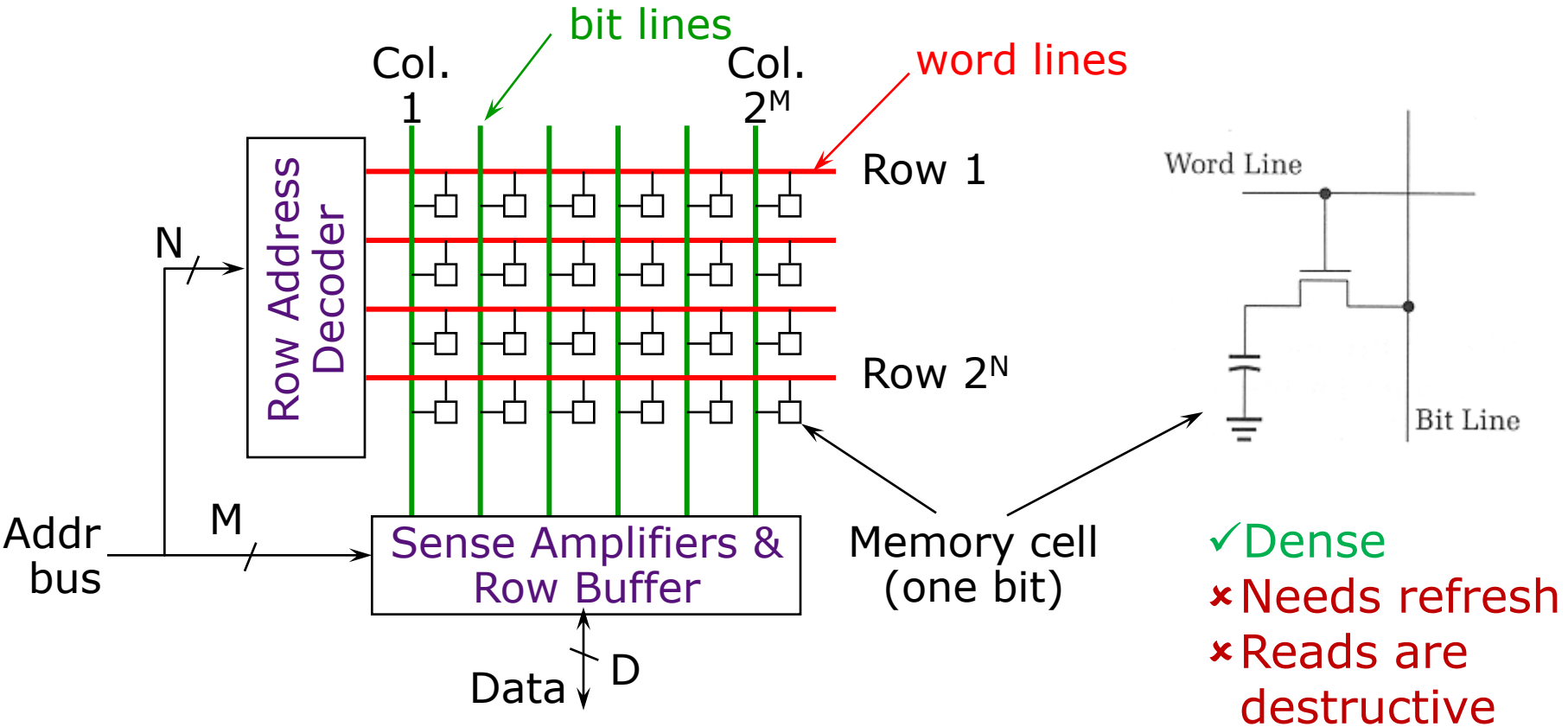
Performance of high-speed computers is usually limited by memory *latency, bandwidth, and energy*

- Latency (time for a single access)
  - Memory access time >> Processor cycle time
- Bandwidth (number of accesses per unit time)
  - if fraction *m* of instructions access memory,
    - $\Rightarrow$ 1+*m* memory references / instruction
    - $\Rightarrow$ CPI = 1 requires 1+*m* memory refs / cycle
  - even if we can hide latency, bandwidth limits throughput!
  - energy/access >> energy/compute op for large memories

Memory Technology

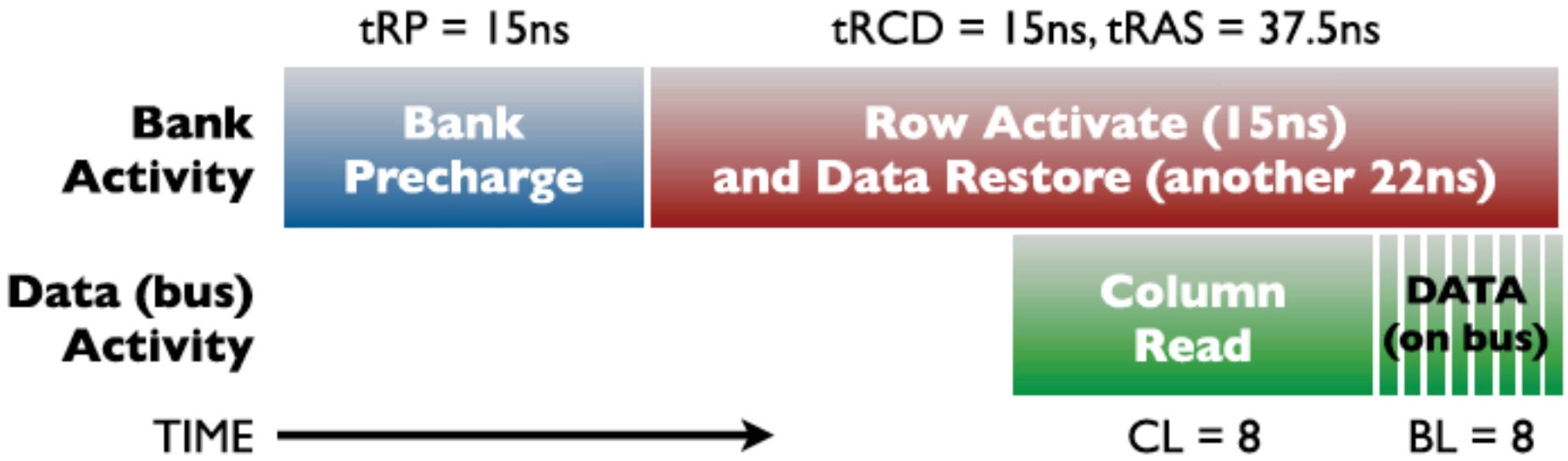
- Early machines used a variety of memory technologies
  - Manchester Mark I used CRT Memory Storage
  - EDSAC used a mercury delay line
- Core memory was first large-scale reliable main memory
  - Invented by Forrester in late 40s at MIT for Whirlwind project
  - Bits stored as magnetization polarity on small ferrite cores threaded onto 2 dimensional grid of wires
- First commercial dynamic RAM (DRAM) was Intel 1103
  - 1Kbit of storage on single chip
  - charge on a capacitor used to hold value
- Semiconductor memory quickly replaced core in 1970s
  - Intel formed to exploit market for semiconductor memory
- Flash memory
  - Slower, but denser than DRAM. Also non-volatile, but with wearout issues
- Emerging memory technologies looking promising for the future
  - e.g., phase-change memory (PCM) is slightly slower than DRAM, but much denser and non-volatile

DRAM Architecture



- Bits stored in 2-dimensional arrays on chip
- Modern chips have around 4 logical banks on each chip
  - Each logical bank physically implemented as many smaller arrays

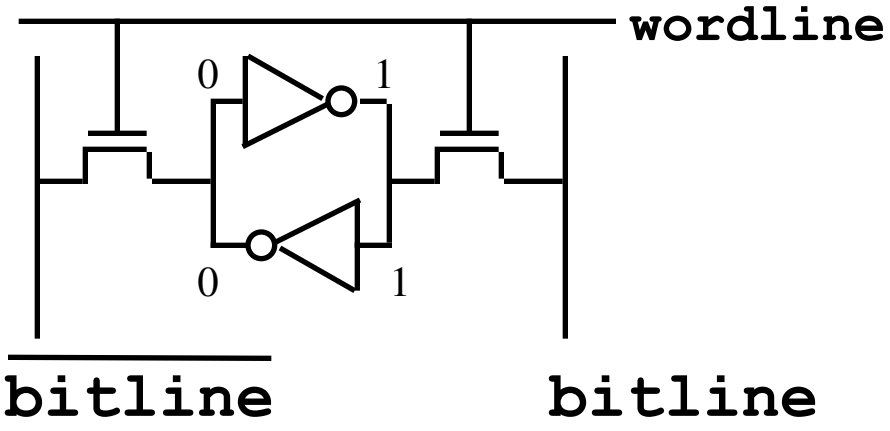
DRAM Timing

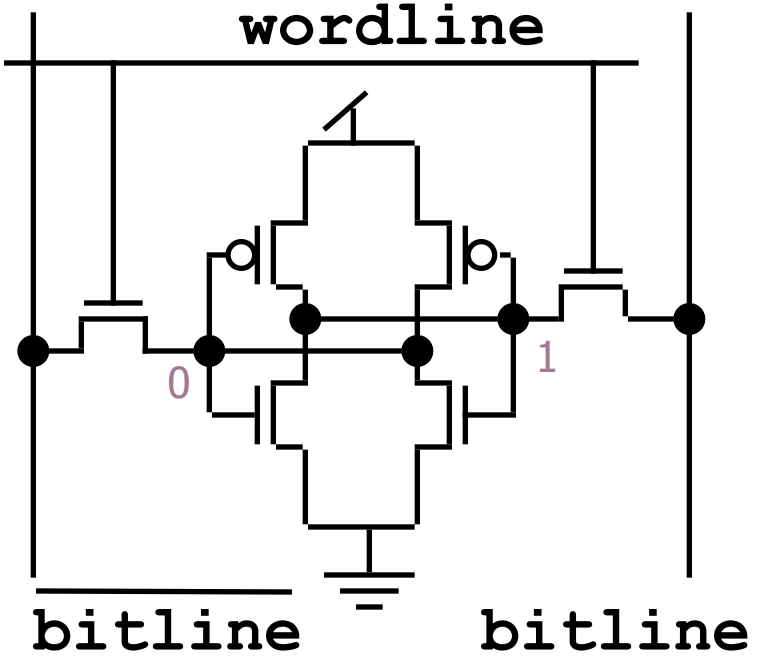


DRAM Spec:
CL, tRCD, tRP, tRAS, e.g., 12-12-12-30

Static RAM Cell

6-Transistor SRAM Cell

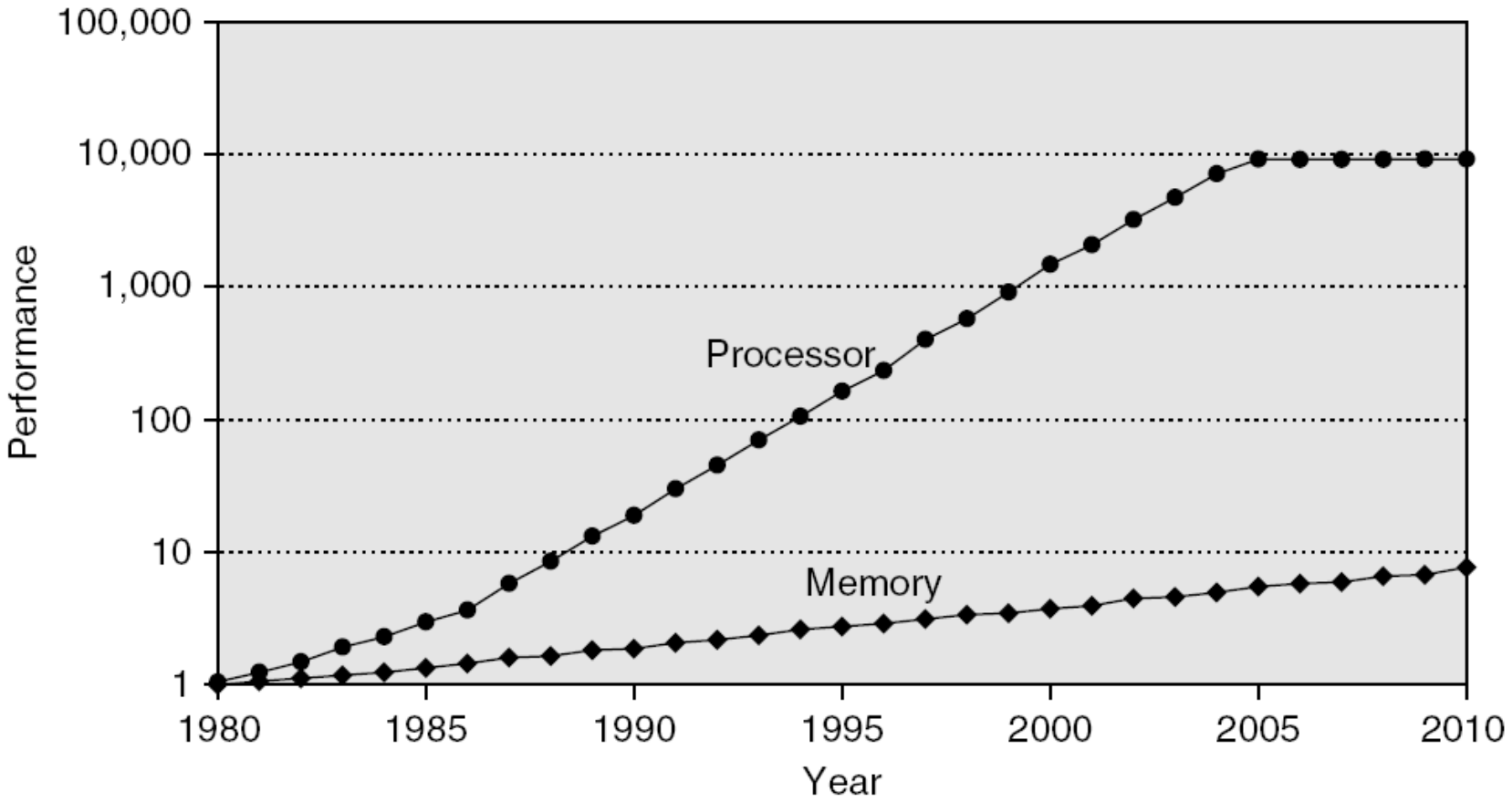




- Write:
  1. Drive bitlines (bit=1, $\overline{\text{bit}}$=0)
  2. Select wordline
- Read:
  1. Precharge bit and $\overline{\text{bit}}$ to Vdd
  2. Select wordline
  3. Cell pulls one line low
  4. Column sense amp detects difference between bit & $\overline{\text{bit}}$

✓No refresh
✓Non-destructive-reads → no row buffer
✗Lower density

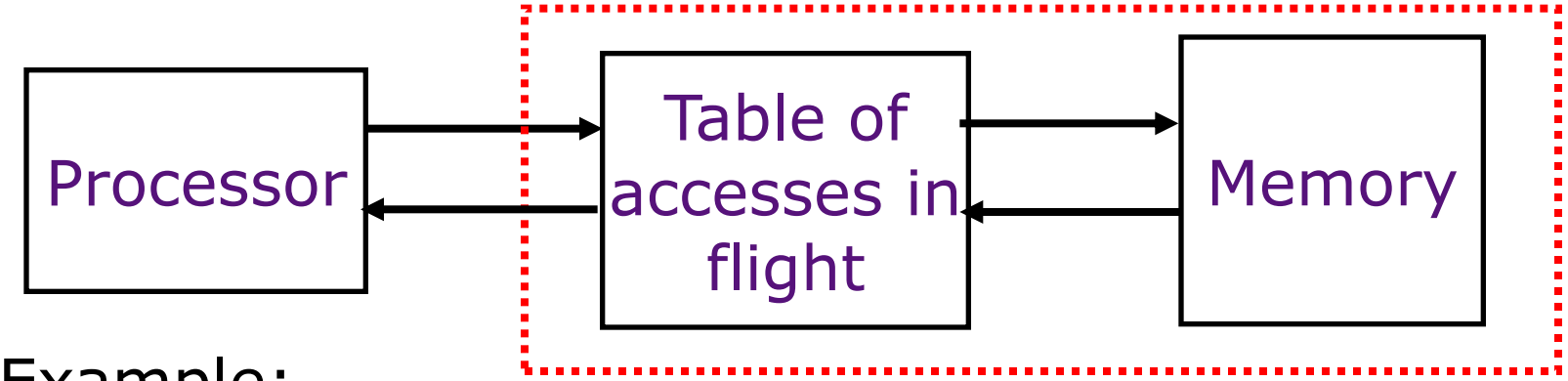
Processor-DRAM Latency Gap



Four-issue 4GHz superscalar accessing 60ns DRAM could execute 960 instructions during time for one memory access!

Little's Law

Throughput (T) = Number in Flight (N) / Latency (L)



Example:

--- Assume infinite-bandwidth memory
--- 100 cycles / memory reference
--- 1 + 0.2 memory references / instruction

$\Rightarrow$ *Table size = 1.2 * 100 = 120 entries*

120 independent memory operations in flight!

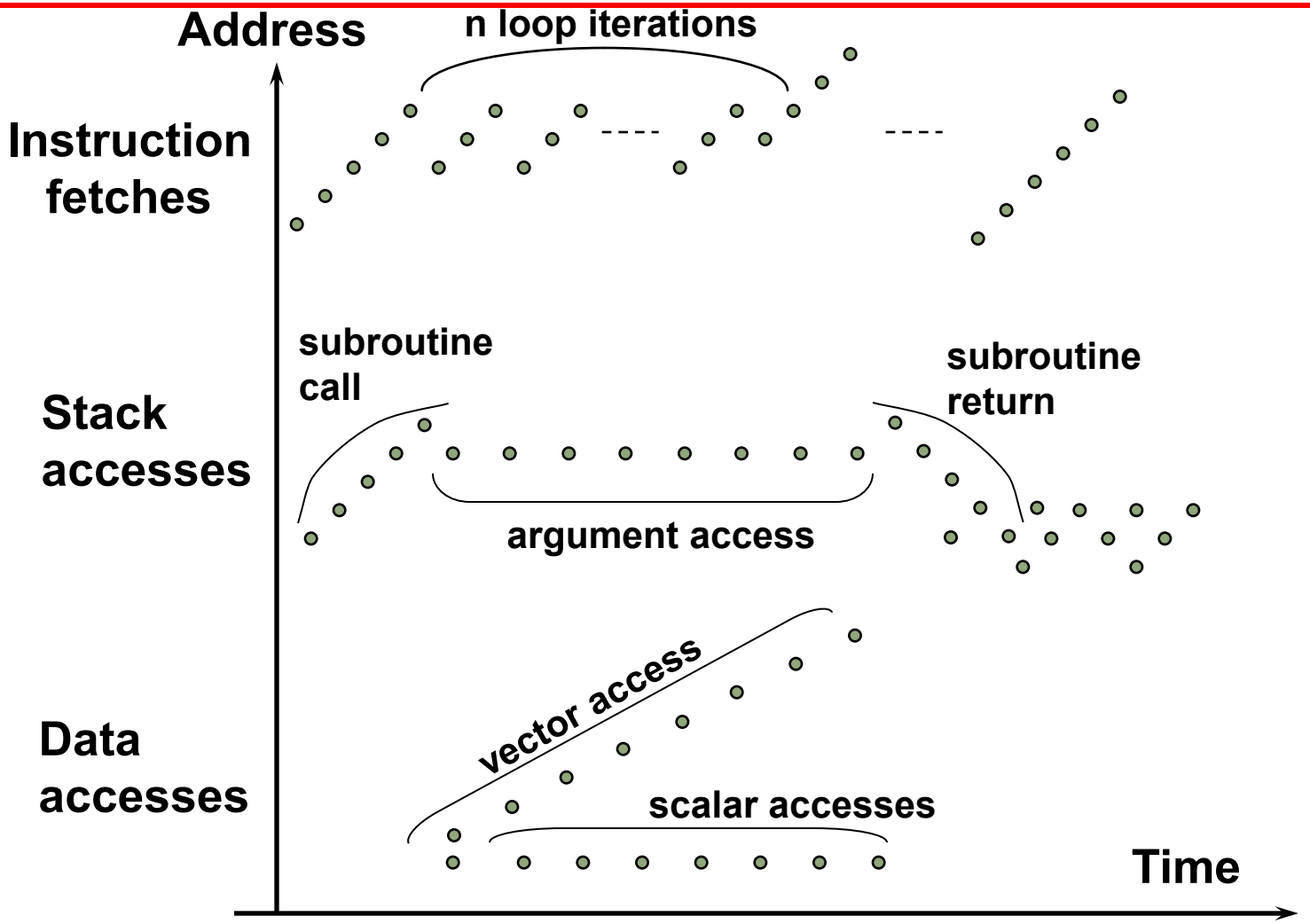
Multilevel Memory

Strategy: Reduce average latency & energy by using caches, small and fast memories that retain recently-accessed data.

Caches work thanks to locality of reference, the empirical observation that the patterns of memory references made by a processor are often highly predictable:

```
                                PC
                    …           96
Loop:   ADD r2, r1, r1          100
        SUBI r3, r3, #1         104
        BNEZ r3, Loop           108
                    …           112
```

Typical Memory Reference Patterns

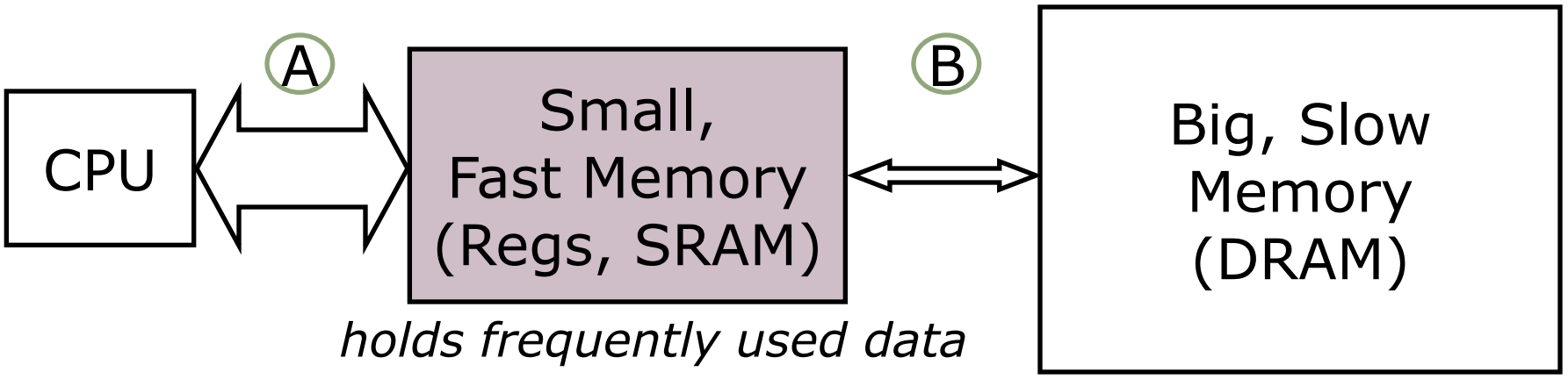


Common Predictable Patterns

Two predictable properties of memory references:

- *Temporal Locality:* If a location is referenced it is likely to be referenced again in the near future.
- *Spatial Locality:* If a location is referenced it is likely that locations near it will be referenced in the near future.

Memory Hierarchy



- *size*: Register << SRAM << DRAM *why?*
- *latency:* Register << SRAM << DRAM *why?*
- *bandwidth:* on-chip >> off-chip *why?*

On a data access:

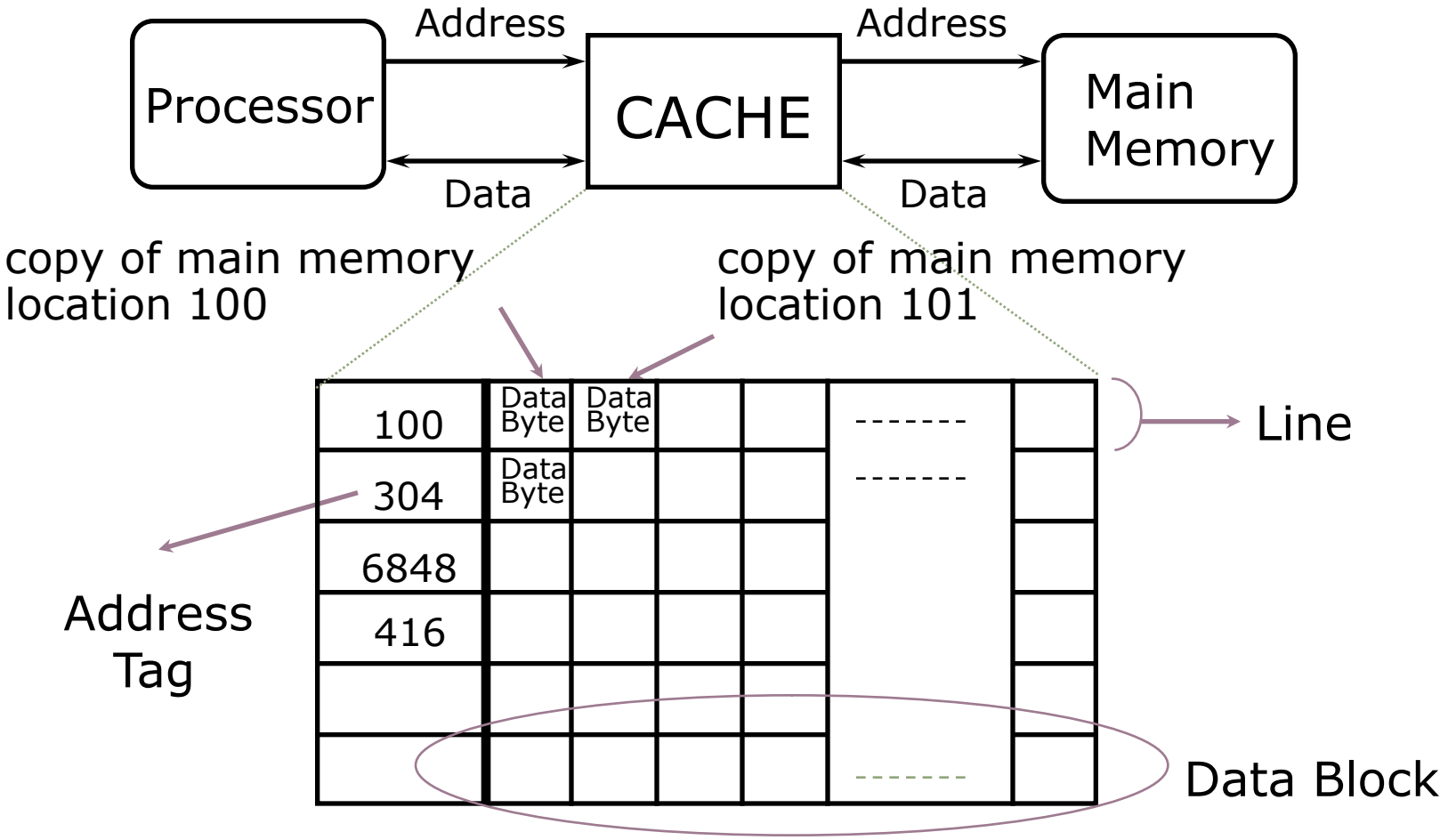
hit (data $\in$ fast memory) $\Rightarrow$ low latency access

miss (data $\notin$ fast memory) $\Rightarrow$ long latency access *(DRAM)*

Management of Memory Hierarchy

- Small/fast storage, e.g., registers
  - Address usually specified in instruction
  - Generally implemented directly as a register file
    - but hardware might do things behind software's back, e.g., stack management, register renaming
- Large/slower storage, e.g., memory
  - Address usually computed from values in register
  - Generally implemented as a cache hierarchy
    - hardware decides what is kept in fast memory
    - but software may provide "hints", e.g., don't cache or prefetch

Inside a Cache

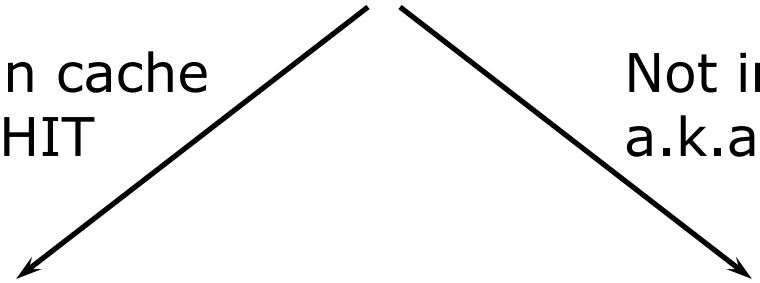


How many bits are needed in tag?

Enough to uniquely identify block

Cache Algorithm (Read)

Look at Processor Address, search cache tags to find match. Then either



Found in cache
a.k.a. HIT

Return copy of data from cache

Not in cache
a.k.a. MISS

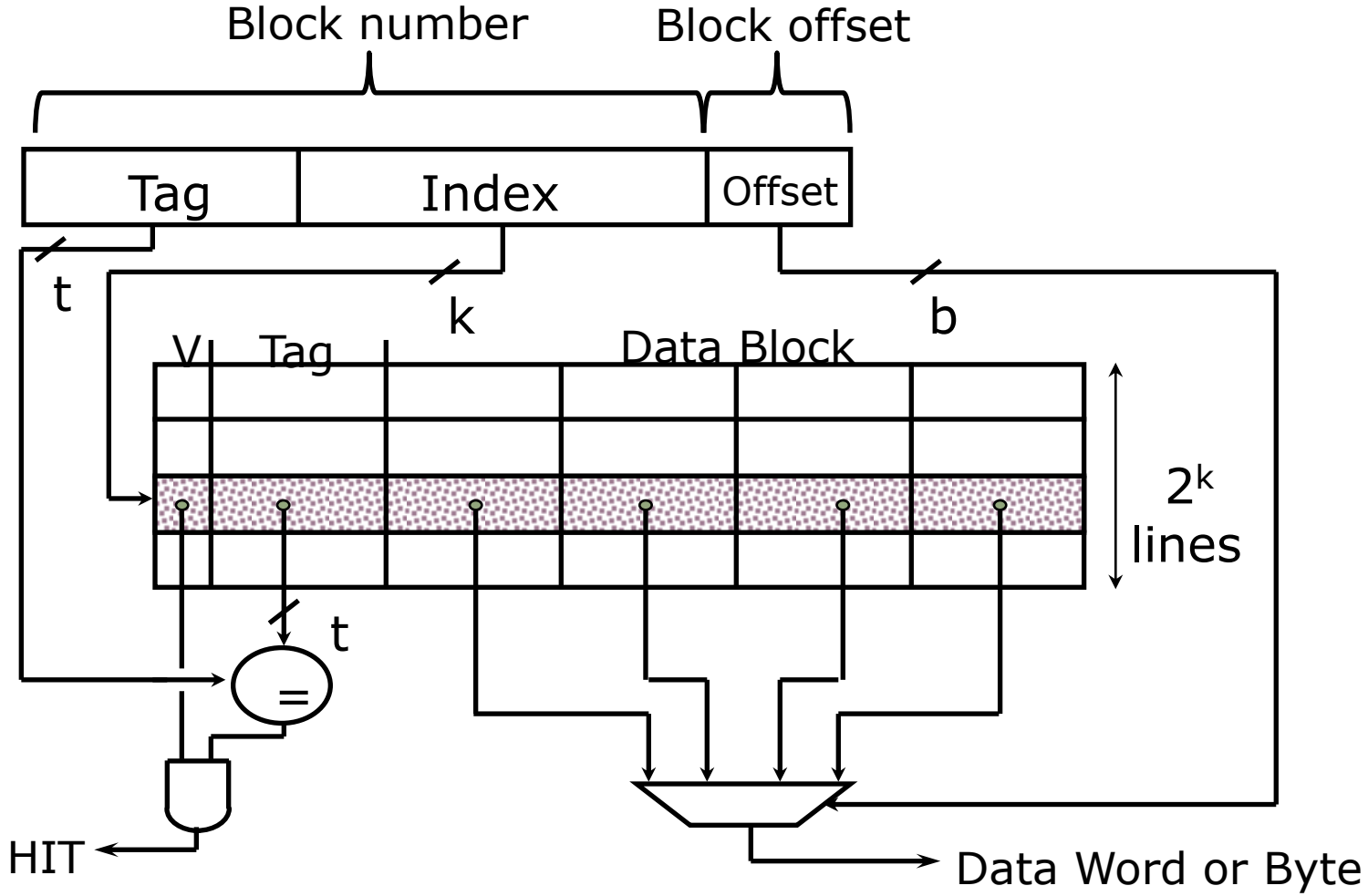
Read block of data from Main Memory

Wait …

Return data to processor and update cache

Q: Which line do we replace?

Direct-Mapped Cache

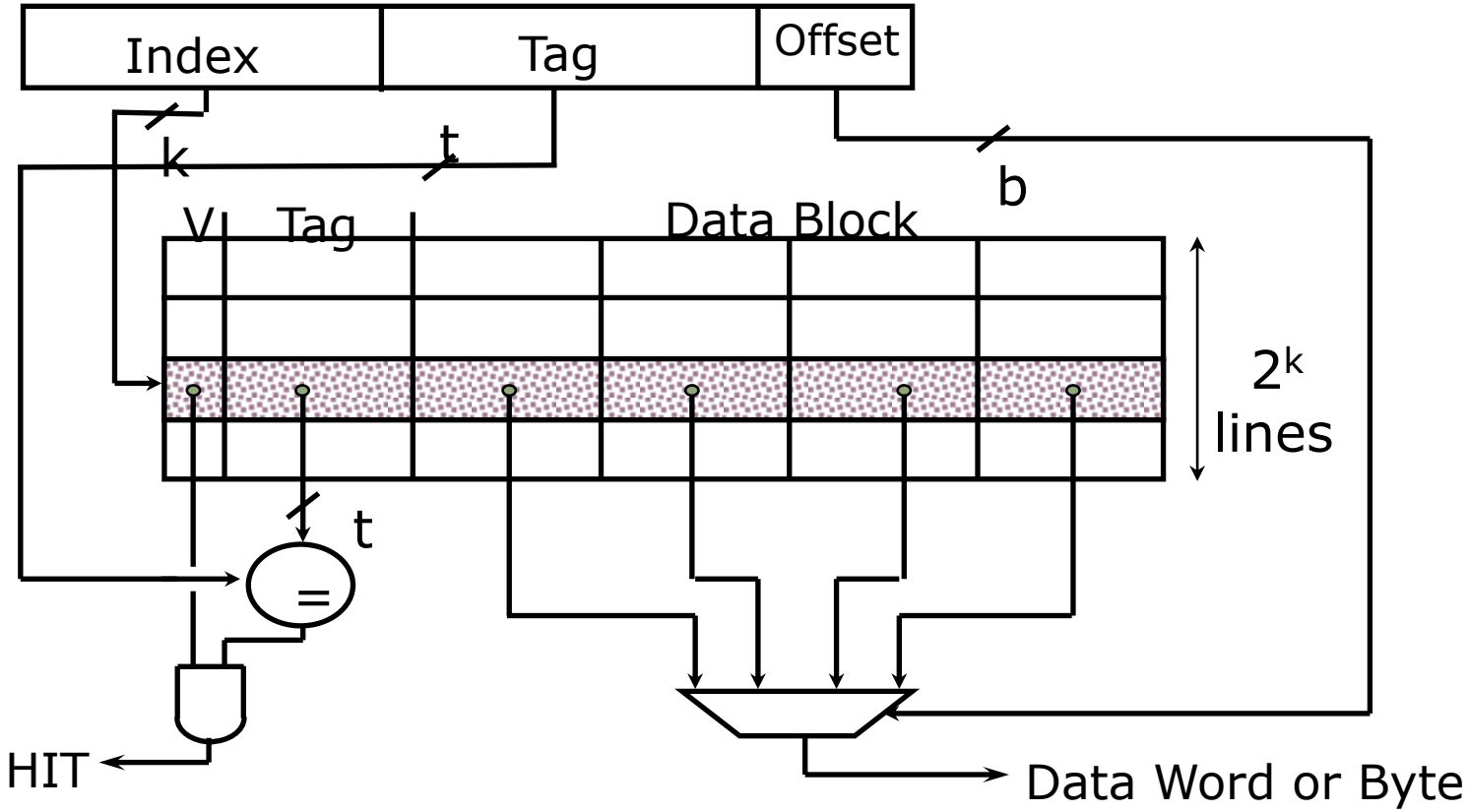


What is a bad reference pattern?

Strided at size of cache

Direct-Mapped Address Selection

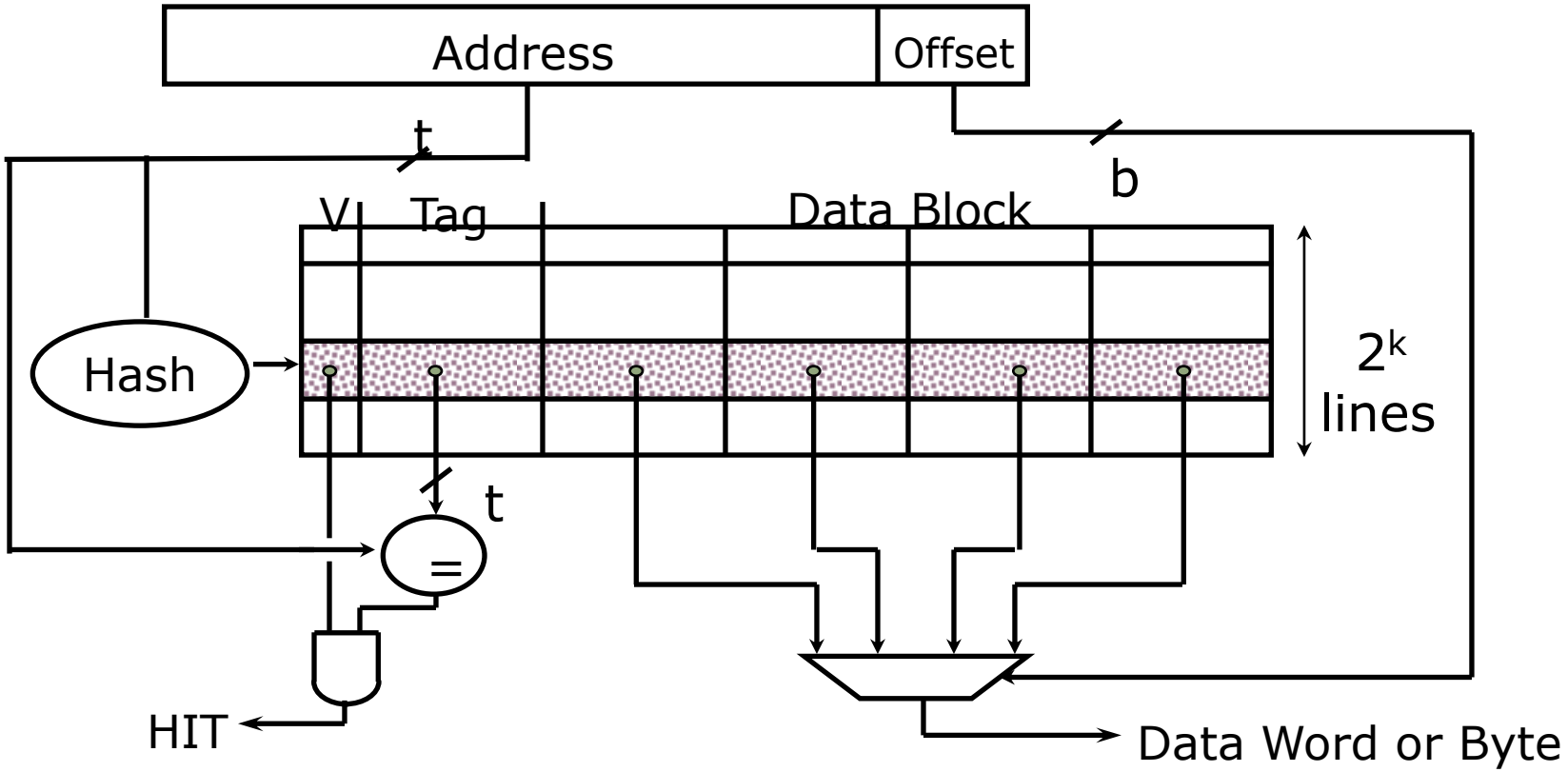
higher-order vs. lower-order address bits



Why might this be undesirable?

Spatially local blocks conflict

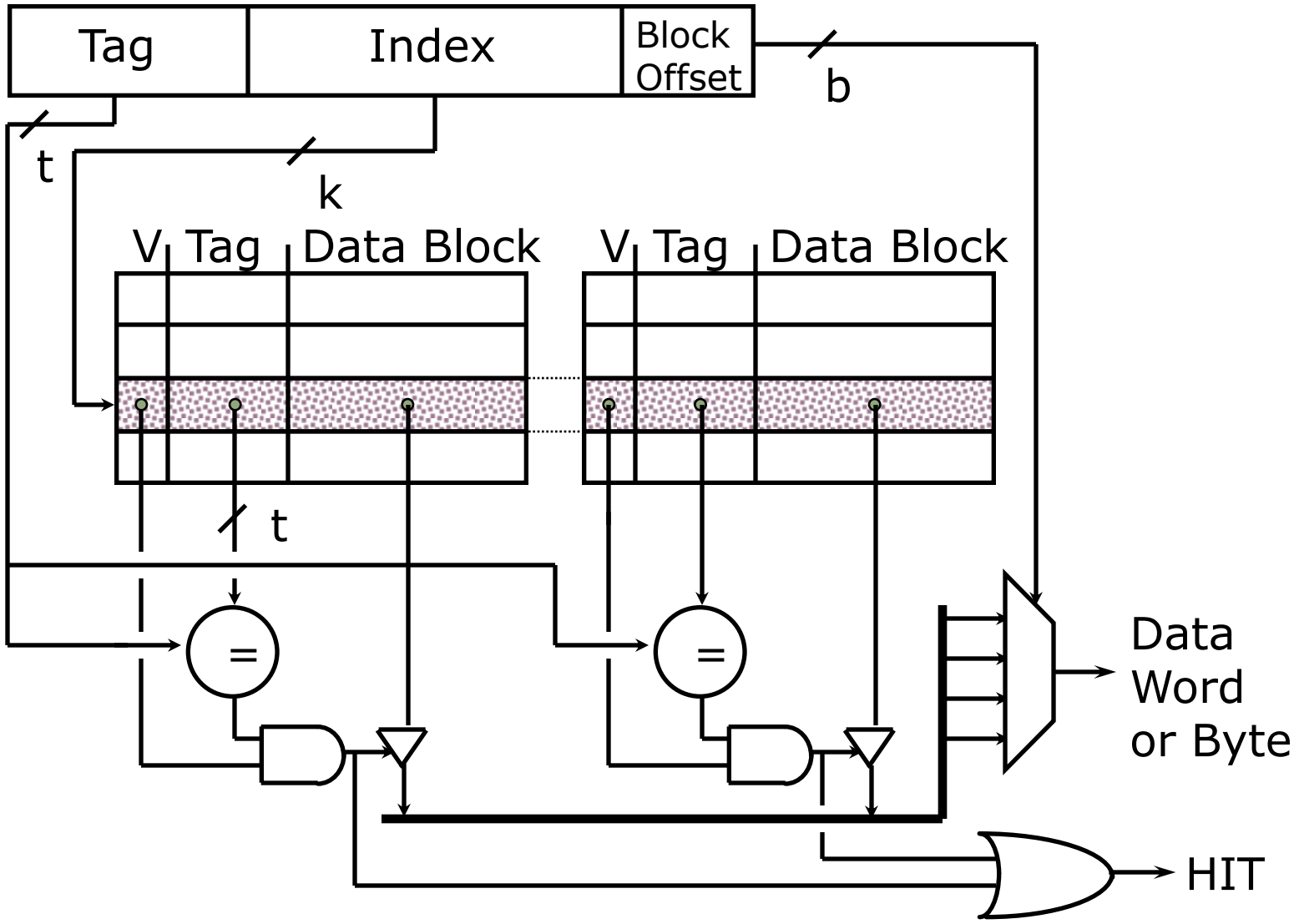
Hashed Address Selection



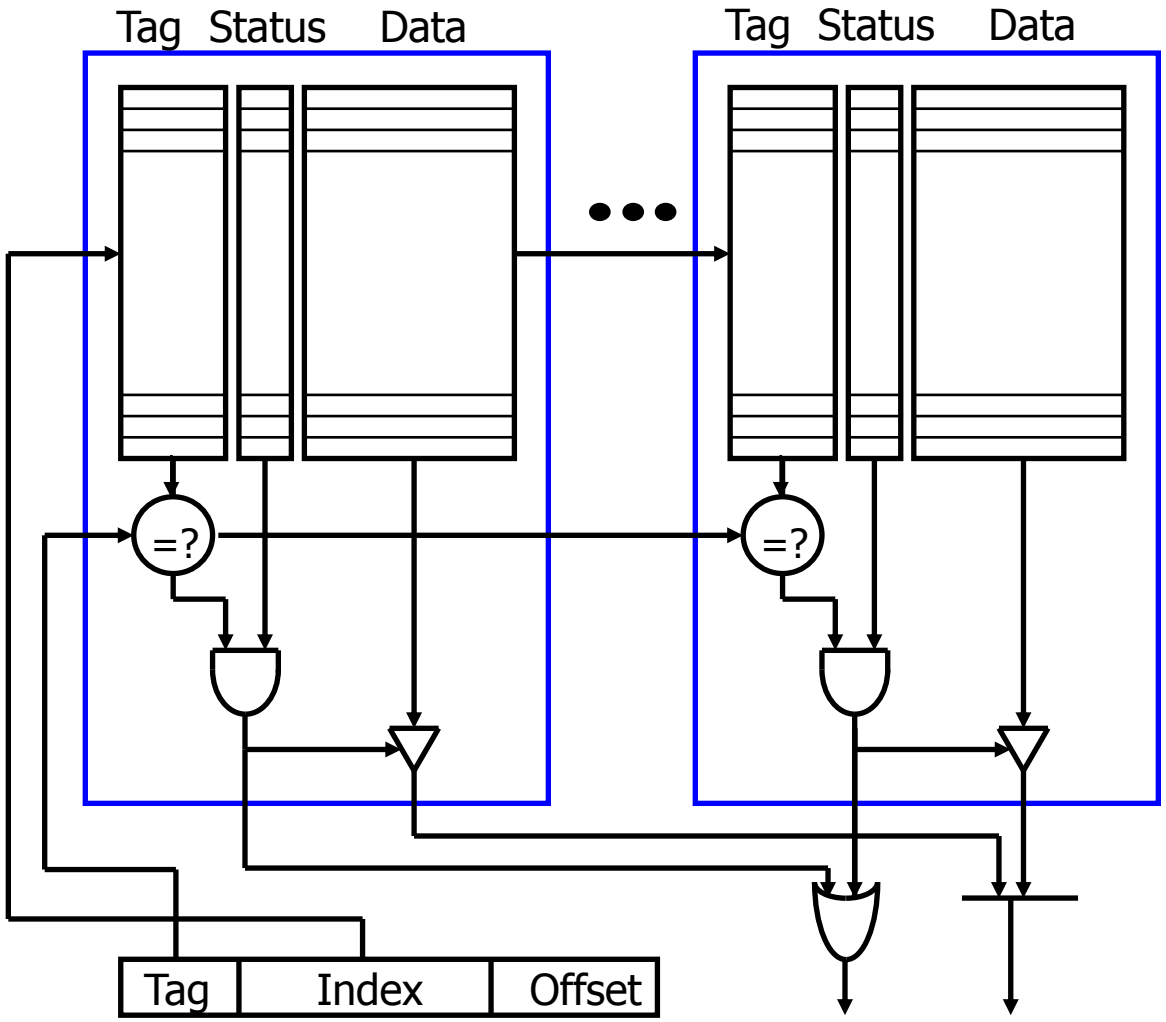
What are the tradeoffs of hashing?

Good: Regular strides don't conflict
Bad: Hash adds latency
Tag is larger

2-Way Set-Associative Cache



Set-Associative RAM-Tag Cache



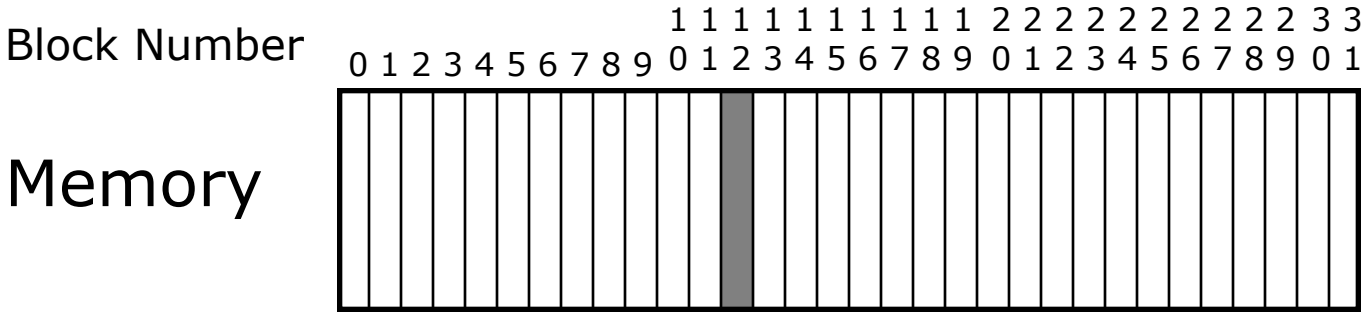
Parallel lookup:

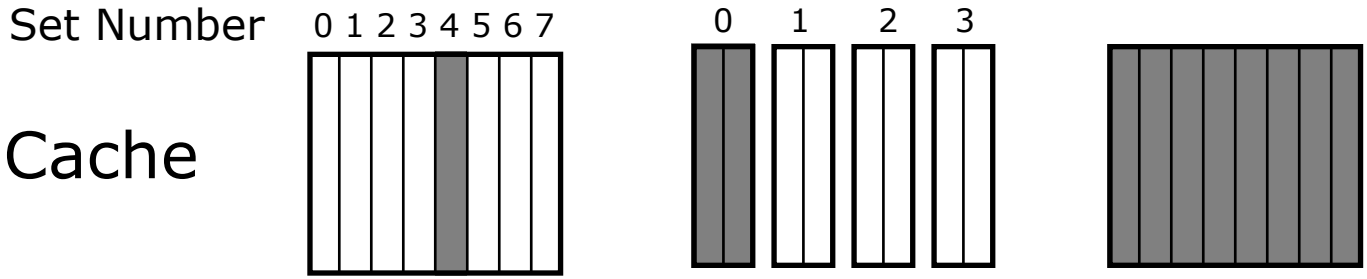
- Tag and data word is read from every way
- Not energy efficient

Serial lookup:

- First read tags, then just read data from selected way
- More energy efficient
- Doubles latency in L1
- OK, for L2 and above, why?

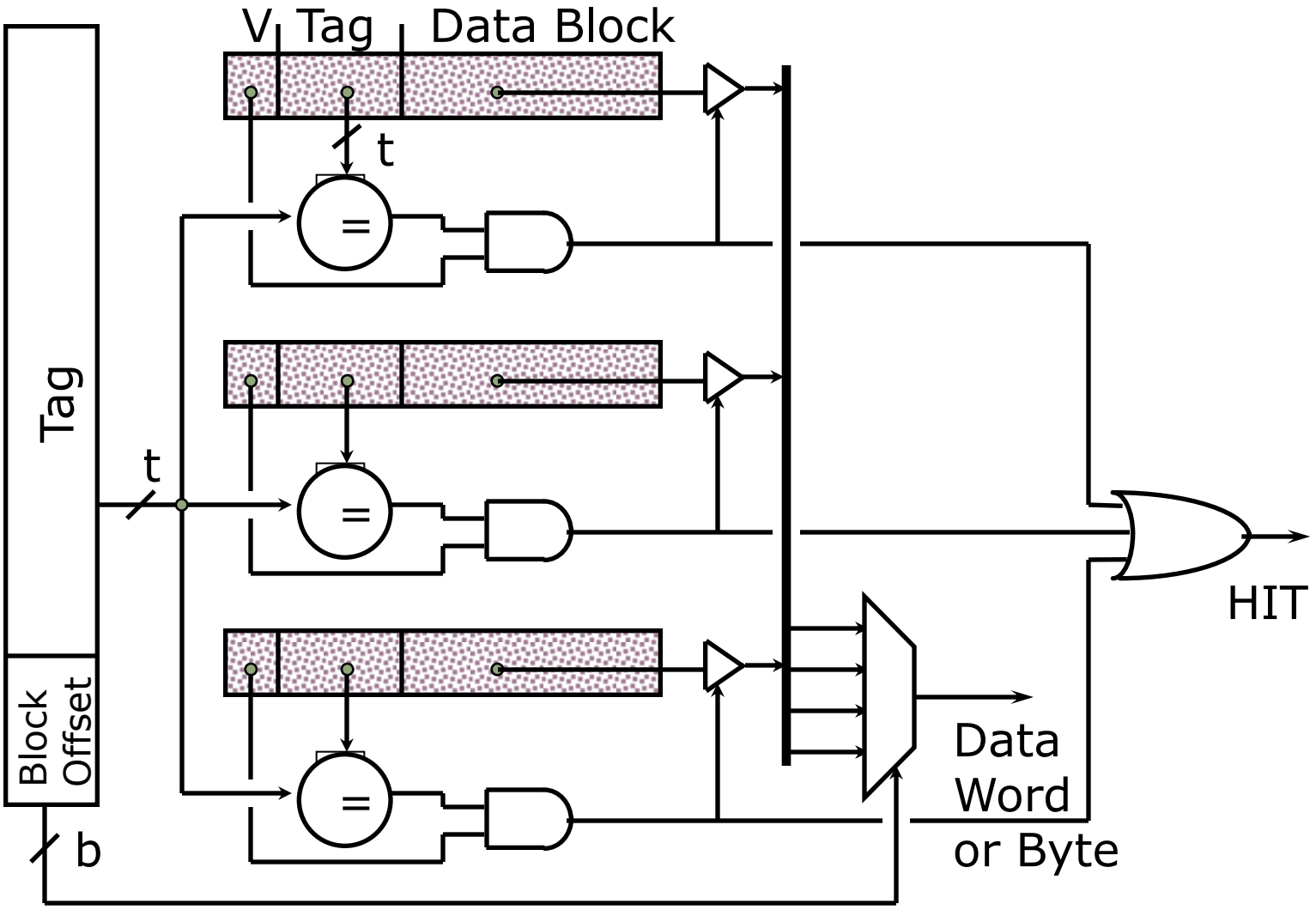
Placement Policy





block 12 can be placed

Fully-Associative Cache



Improving Cache Performance

Average memory access time =
Hit time + Miss rate x Miss penalty

To improve performance:

- reduce the hit time
- reduce the miss rate (e.g., larger, better policy)
- reduce the miss penalty (e.g., L2 cache)

What is the simplest design strategy?

Biggest cache that doesn't increase hit time past 1-2 cycles (approx 8-32KB in modern technology)

[design issues more complex with out-of-order superscalar processors]

Causes for Cache Misses

- *Compulsory:*
  first-reference to a block *a.k.a.* cold start misses
  - misses that would occur even with infinite cache

- *Capacity:*
  cache is too small to hold all data the program needs
  - misses that would occur even with fully-associative cache

- *Conflict:*
  misses from collisions due to limited associativity

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache
Compulsory misses	=	=	↓
Capacity misses	↓	=	↓↑
Conflict misses	↓	↓	?
Hit latency	↑	↑	=
Miss latency	=	=	↑

Block-level Optimizations

- Tags are too large, i.e., too much overhead
  - Simple solution: Larger blocks, but miss penalty could be large.
- Sub-block placement (aka sector cache)
  - A valid bit added to units smaller than the full block, called sub-blocks
  - Only read a sub-block on a miss
  - *If a tag matches, is the word in the cache?*

100	***1***		***1***		***1***		***1***	
300	***1***		***1***		***0***		***0***	
204	***0***		***1***		***0***		***1***	

Replacement Policy

Which block from a set should be evicted?

- Random
- Least Recently Used (LRU)
  - LRU cache state must be updated on every access
  - true implementation only feasible for small sets (2-way)
  - pseudo-LRU binary tree was often used for 4-8 way
- First In, First Out (FIFO) a.k.a. Round-Robin
  - used in highly associative caches
- Not Least Recently Used (NLRU)
  - FIFO with exception for most recently used block or blocks
- One-bit LRU
  - Each way represented by a bit. Set on use, replace first unused.

Re-Reference Interval Prediction

Time

References:

H	F	E	F	B	C	A	F
	0	1	2	3	4	5	6

tag	C	A	F	B
RRI	4	5	0	3

Best candidate?

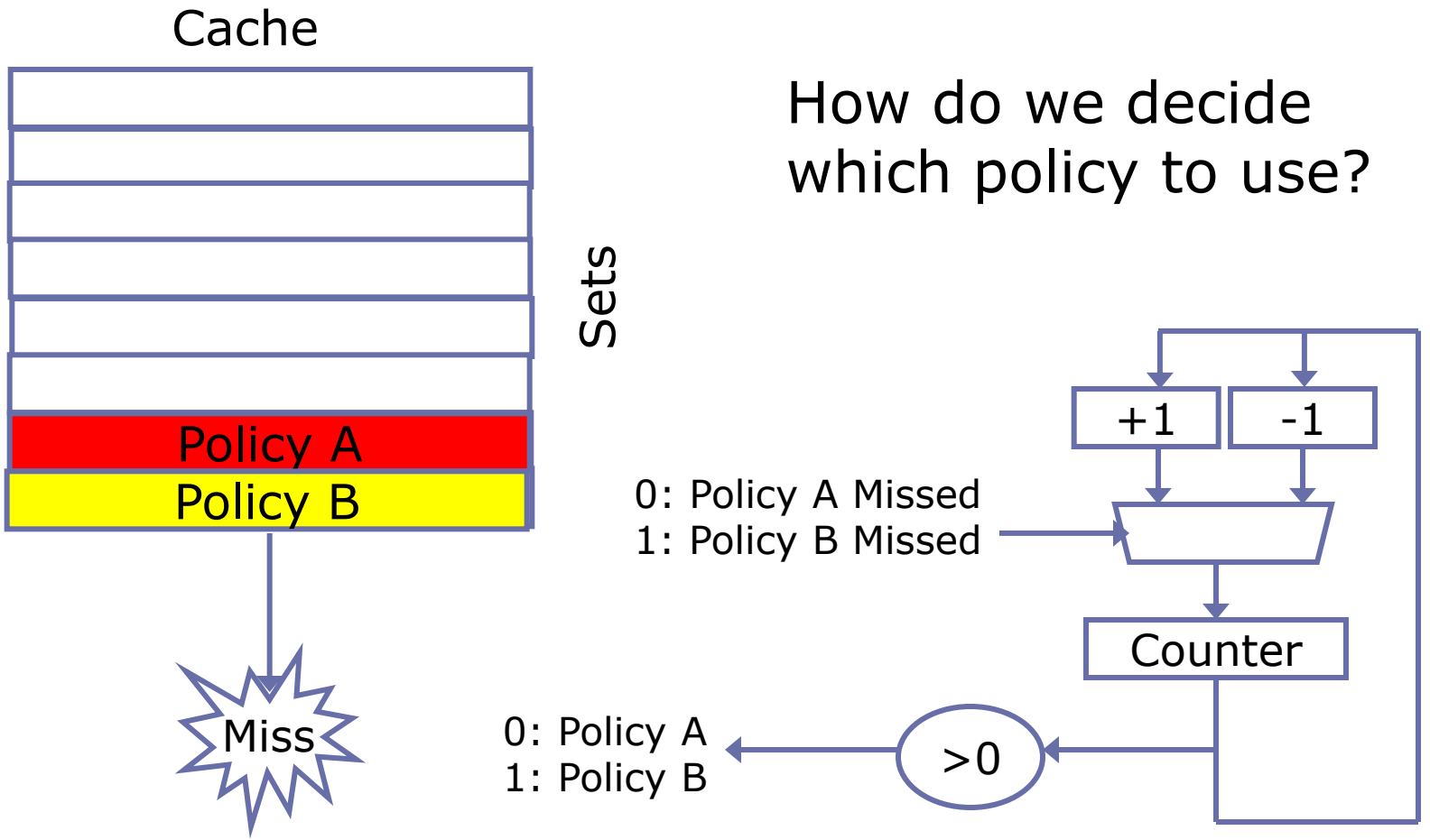
A

tag	C	A	F	B
RRIP	2	3	0	1

RRIP – 2-bit LRU

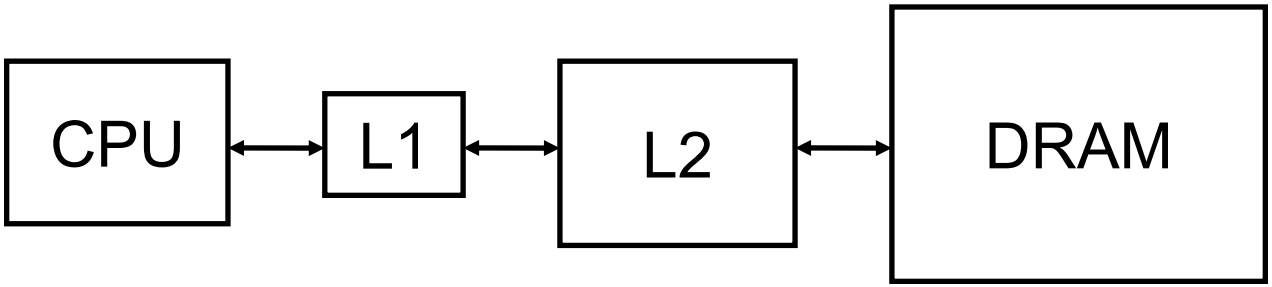
Multiple Replacement Policies

Use the best replacement policy for a program



Multilevel Caches

- A memory cannot be large and fast
- Add level of cache to reduce miss penalty
  - Each level can have longer latency than level above
  - So, increase sizes of cache at each level



Metrics:

Local miss rate = misses in cache/ accesses to cache

Global miss rate = misses in cache / CPU memory accesses

Misses per instruction = misses in cache / number of instructions

Inclusion Policy

- *Inclusive* multilevel cache:
  - Inner cache data must also be in the outer cache
  - External accesses need only check outer cache
  - Most common case
- *Non-inclusive* multilevel cache:
  - Inner cache may hold data not in outer cache
  - On a miss, both inner and outer level store a copy of the data
- *Exclusive* multilevel caches:
  - Inner cache data is not outer cache
  - Swap lines between inner/outer caches on miss
  - Used in AMD Athlon with 64KB primary and 256KB secondary cache

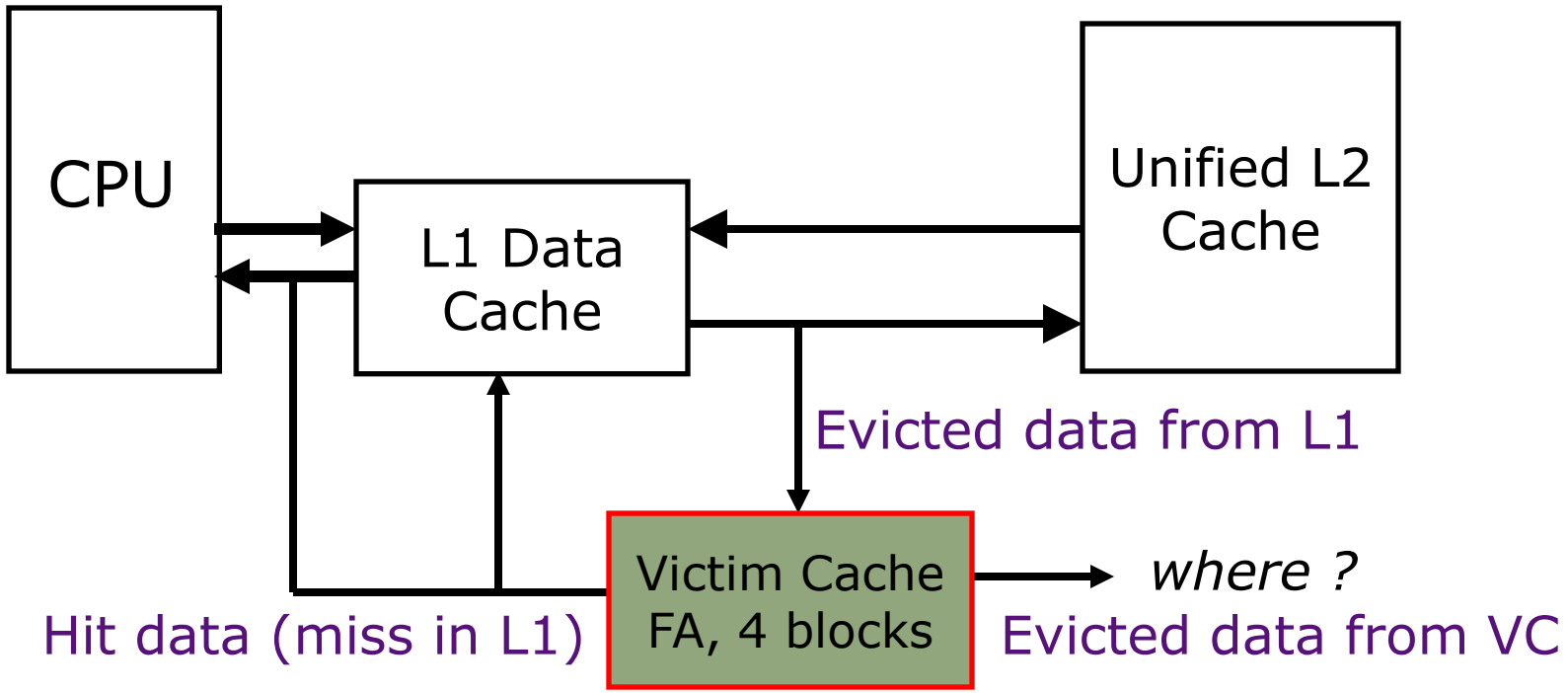
Why choose one type or the other?

Write Policy

- Write-through: Propagate writes to the next level in the hierarchy
  - Keeps next-level cache / memory updated
  - Simple, high-bandwidth

- Write-back: Writes not propagated to next level until we replace the block
  - Contents of next-level cache / memory can be stale
  - Complex, lower bandwidth

Victim Caches (HP 7200)



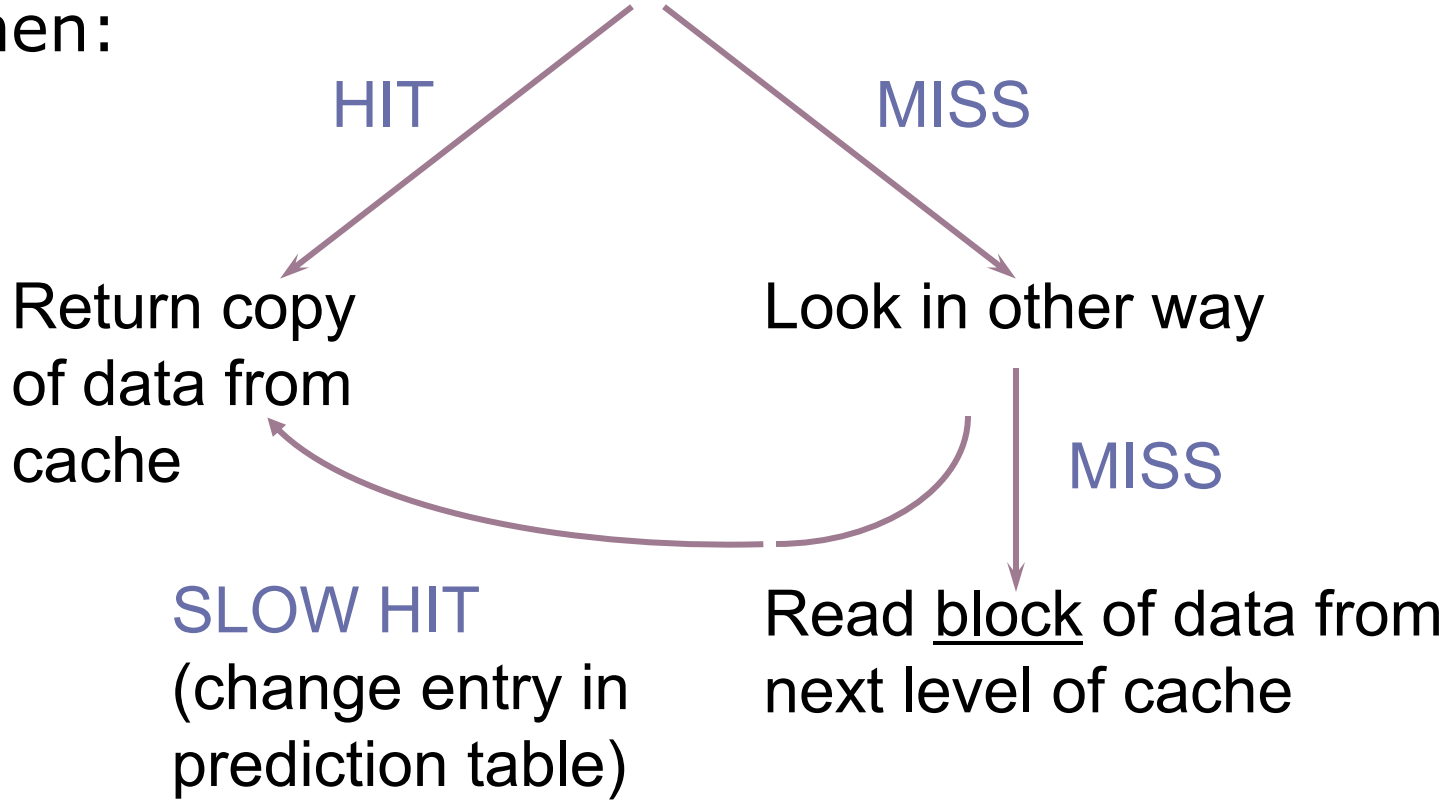
Victim cache: A small associative cache, added to a direct-mapped cache, which holds recently evicted lines

- First look up in direct-mapped cache
- If miss, look in victim cache
- If hit in victim cache, swap hit line with line now evicted from L1
- If miss in victim cache, L1 victim -> VC, VC victim->?

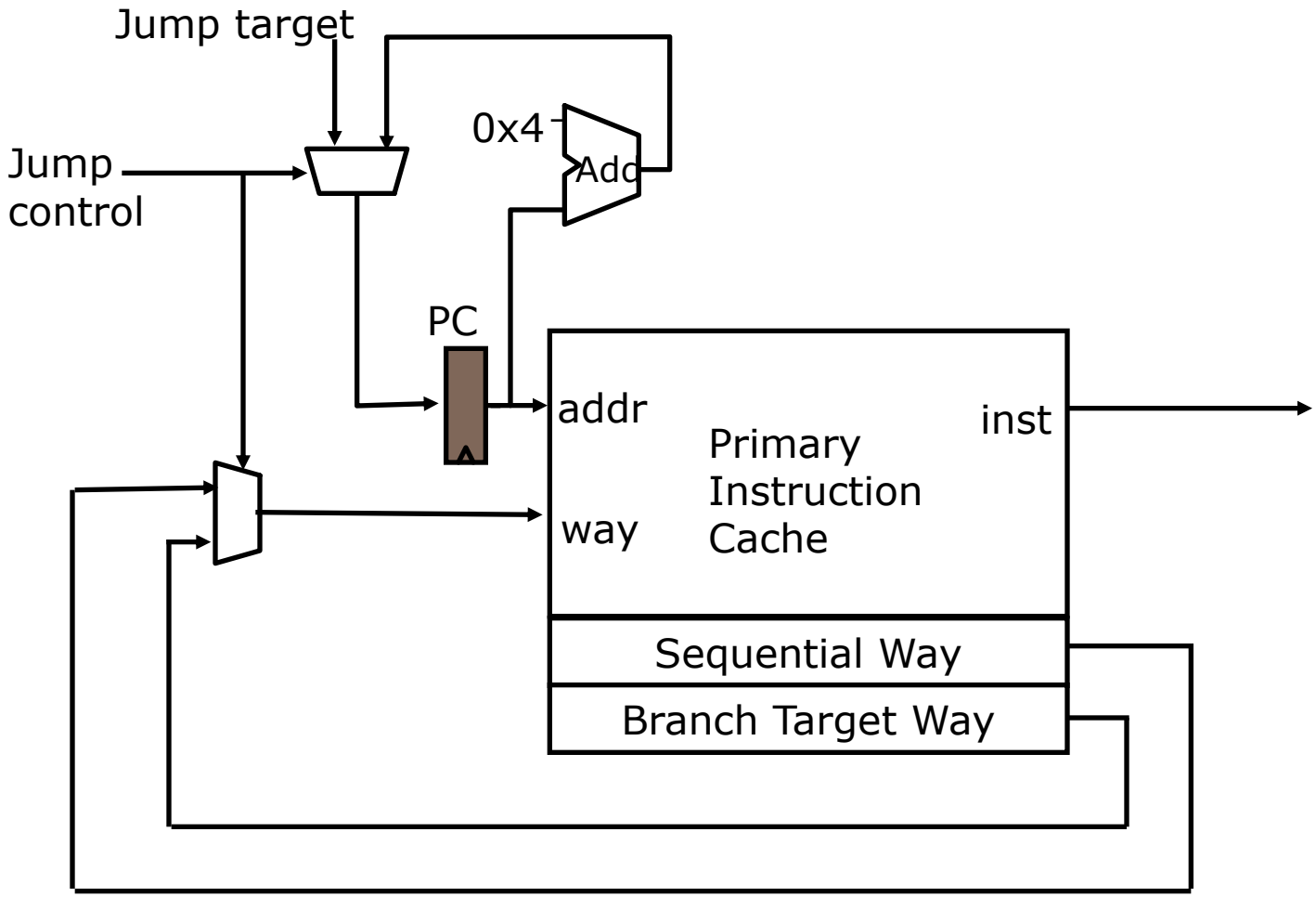
Fast hit time of direct-mapped but with reduced conflict misses

Way Predicting Caches *(MIPS R10000 L2 cache)*

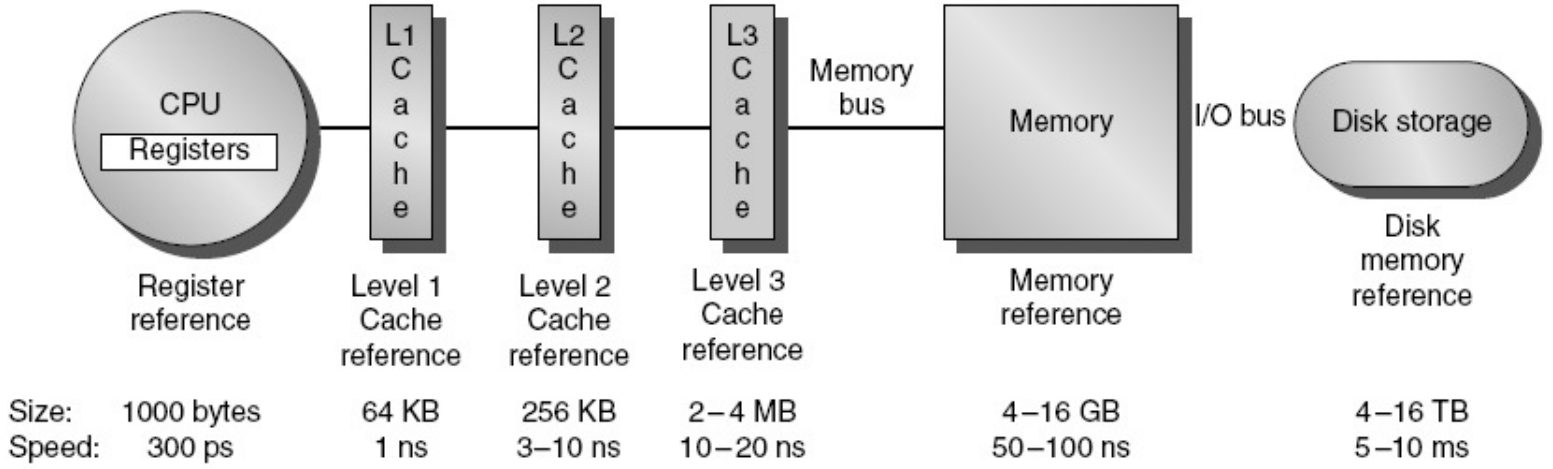
- Use processor address to index into way prediction table
- Look in predicted way at given index, then:



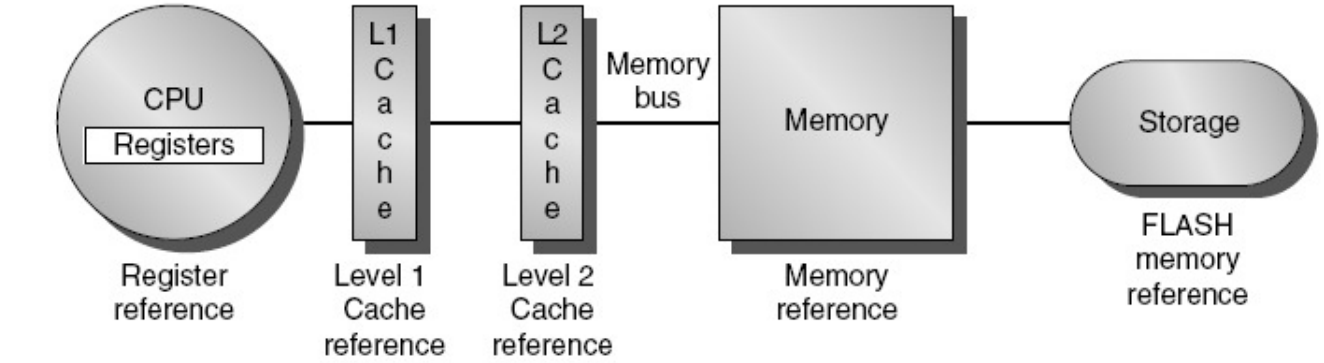
Way Predicting Instruction Cache (Alpha 21264-like)



Typical Memory Hierarchies



(a) Memory hierarchy for server



(b) Memory hierarchy for a personal mobile device

Thank you !

Next lecture – virtual memory