Memory Hierarchy

2/18/2016

CS 152 Section 5

Colin Schmidt

Agenda

- Review Memory Hierarchy
- Lab 2 Questions
- Return Quiz 1

Latencies Comparison Numbers

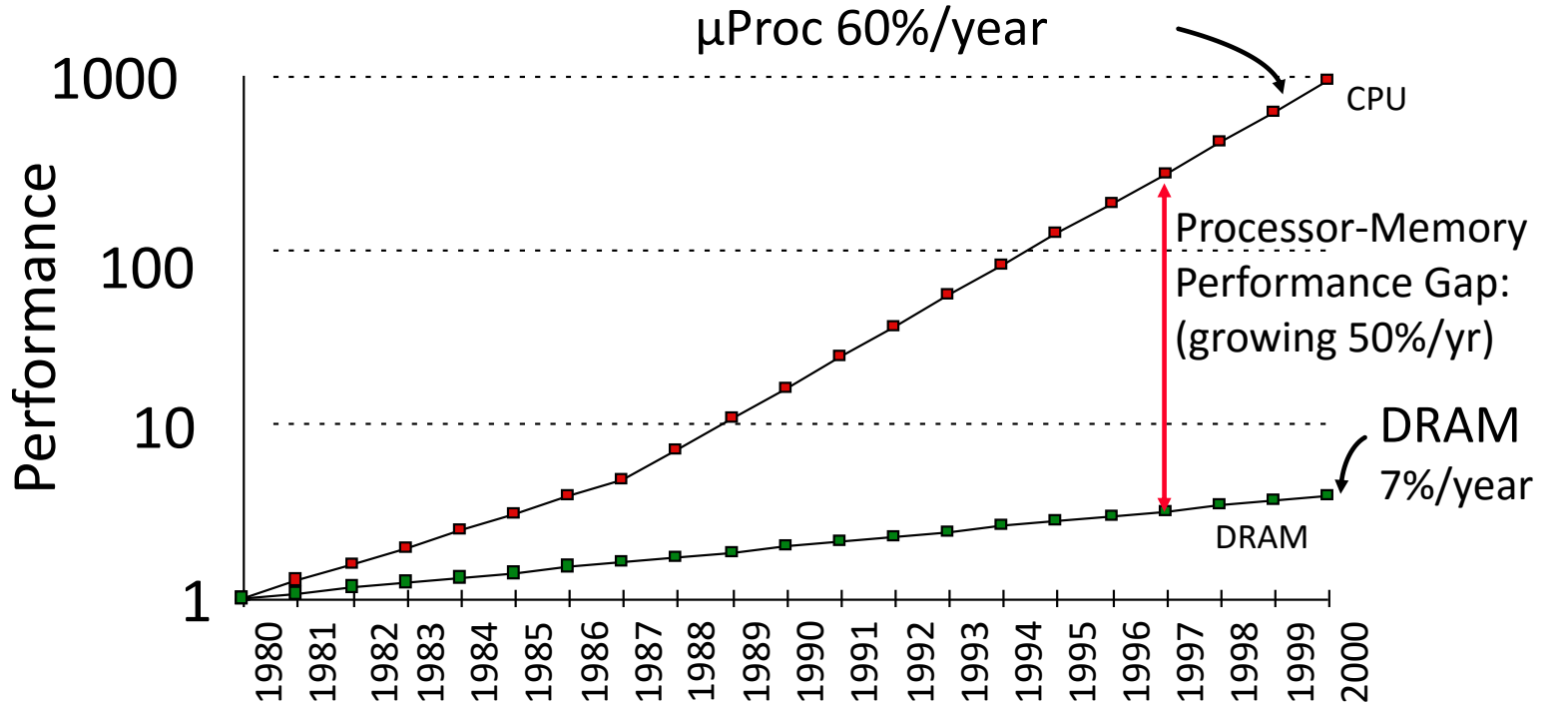
L1 Cache	0.5 ns	
L2 Cache	7 ns	14x L1 cache
Main Memory	100 ns	20x L2, 200x L1
Read 4K randomly from SSD	150,000 ns	
Read 1MB sequentially from memory	250,000 ns	
Read 1MB sequentially from SSD	1,000,000 ns	4x Memory
Disk Seek	10,000,000 ns	
Read 1 MB sequentially from disk	20,000,000 ns	80x memory, 20x SSD

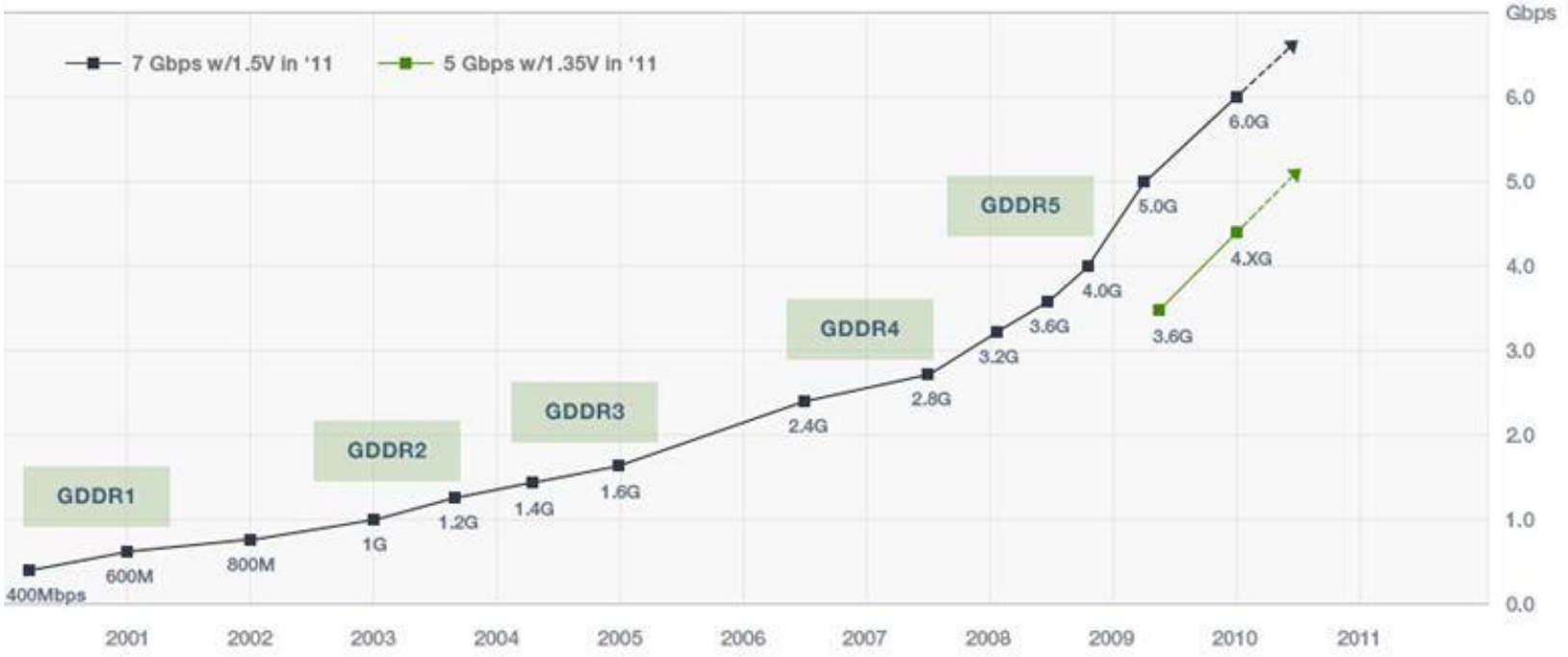
assuming ~1GB/sec SSD

https://gist.github.com/jboner/2841832

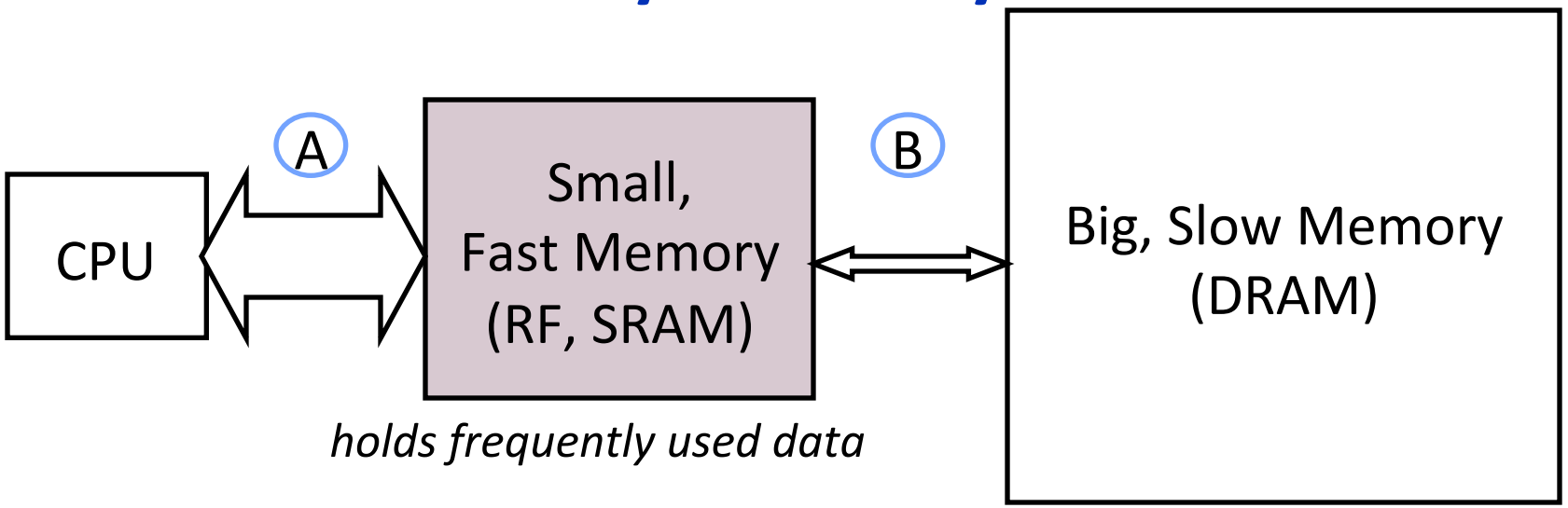
DRAM

- DRAM is complex why put up with it?
  - CHEAP! and very fast compared to disk
- Many steps in modern dram but at a high level just 3
  - RAS, CAS, Precharge
  - Why do we need precharge
    - Capacitor is very small only has a little bit of charge
    - Sense amps turn this small change into 0 or 1
- Most architects just care about avoiding using the DRAM and using the BW effectively when needed


µProc 60%/year
CPU
Processor-Memory
Performance Gap:
(growing 50%/yr)
DRAM
7%/year
DRAM
Performance
1000
100
10
1
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000


7 Gbps w/1.5V in '11
5 Gbps w/1.35V in '11
Gbps
6.0
5.0
4.0
3.0
2.0
1.0
0.0
GDDR1
GDDR2
GDDR3
GDDR4
GDDR5
400Mbps
600M
800M
1G
1.2G
1.4G
1.6G
2.4G
2.8G
3.2G
3.6G
4.0G
5.0G
6.0G
3.6G
4.XG
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

Memory Hierarchy



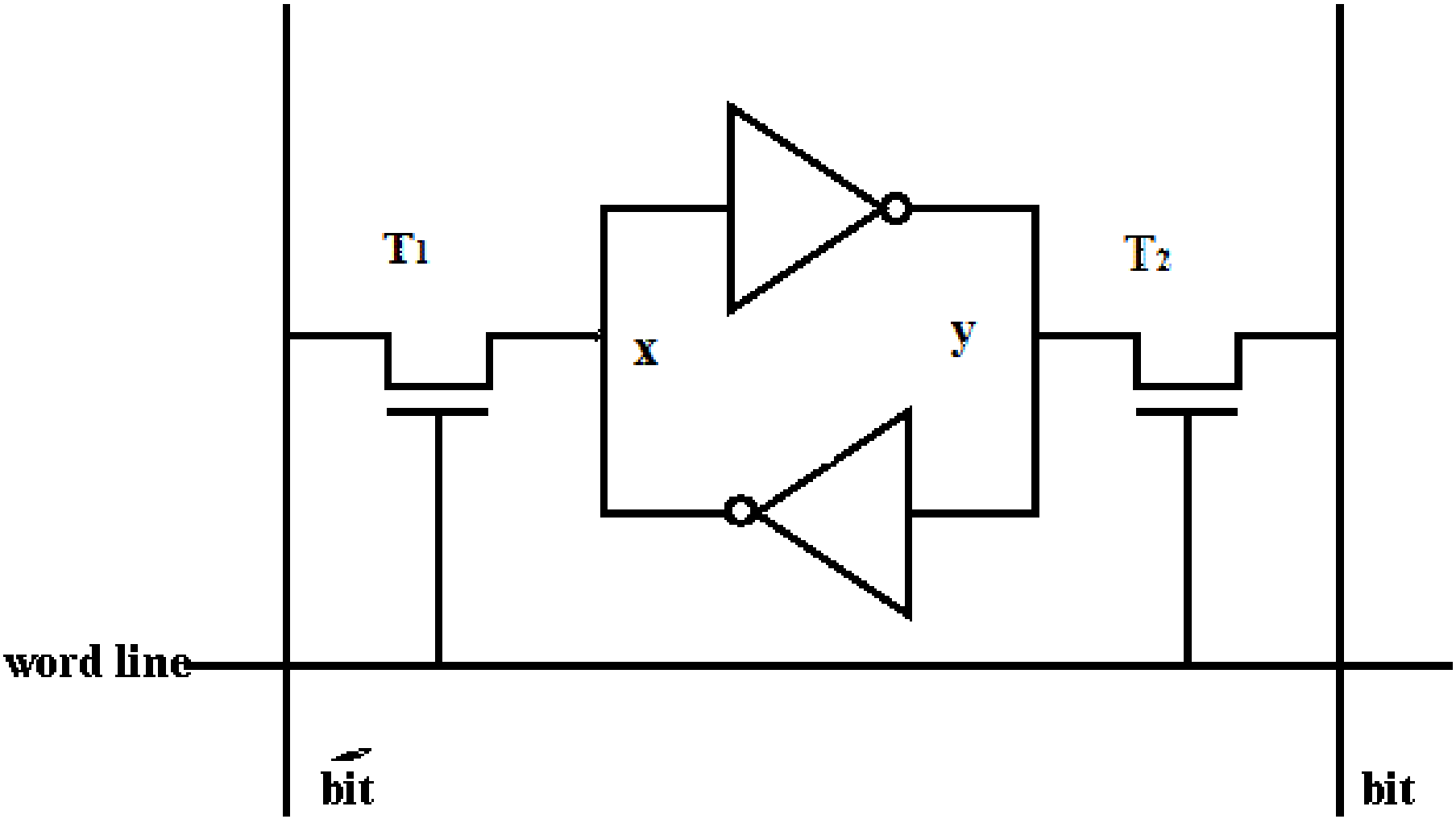
- *capacity*: Register << SRAM << DRAM
- *latency*: Register << SRAM << DRAM
- *bandwidth:* on-chip >> off-chip

On a data access:

if data $\in$ fast memory $\Rightarrow$ low latency access *(SRAM)*

if data $\notin$ fast memory $\Rightarrow$ high latency access *(DRAM)*

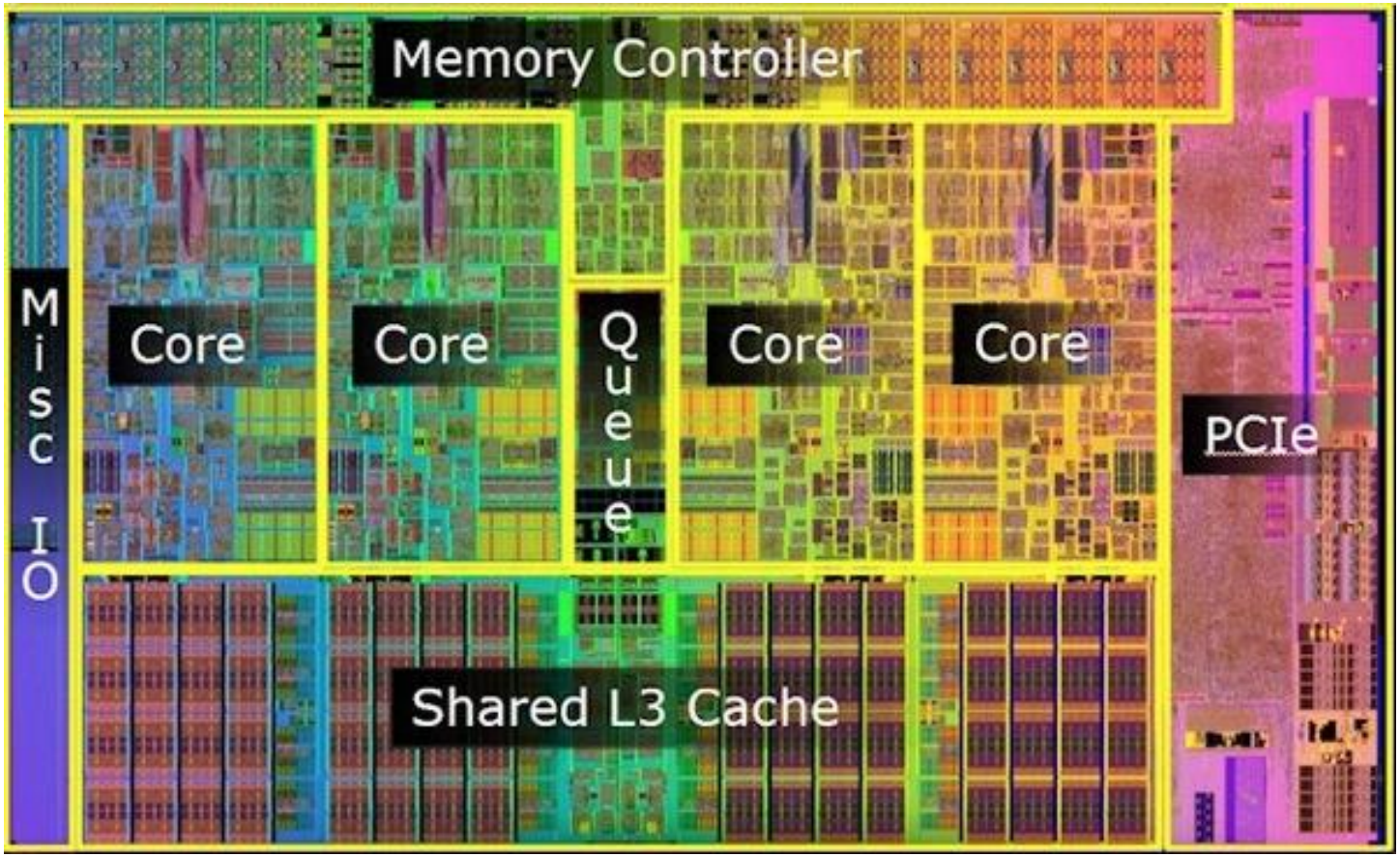
SRAM Cell



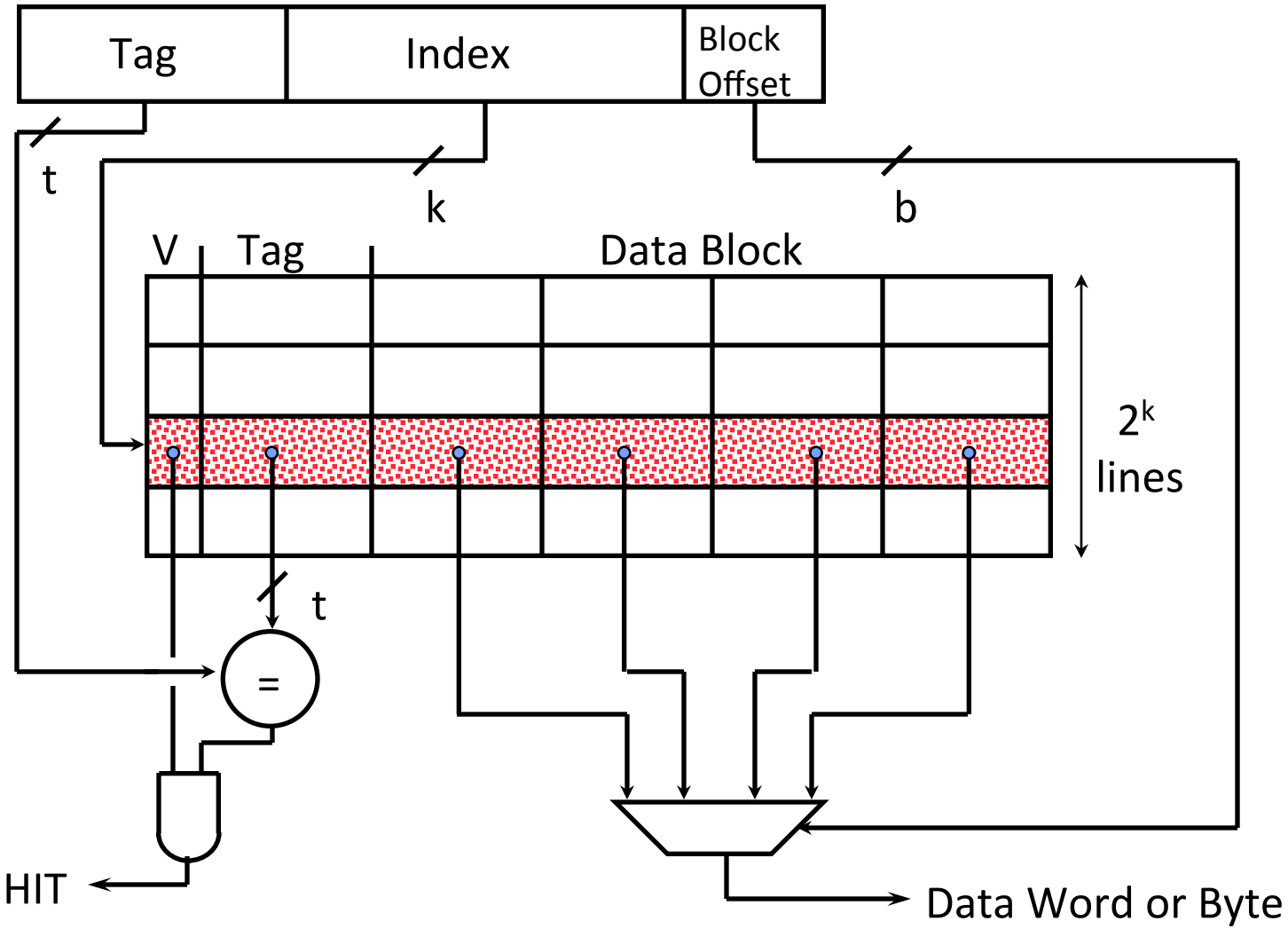
Caches

- Who manages?
  - Hardware
  - How do we know what to cache?
- Locality?
  - Temporal, Spatial
- Basic parts of a cache?
  - Tags, lines
  - How does a Hit work vs a Miss?
- Why do we miss?
  - Compulsory, Capacity, Conflict

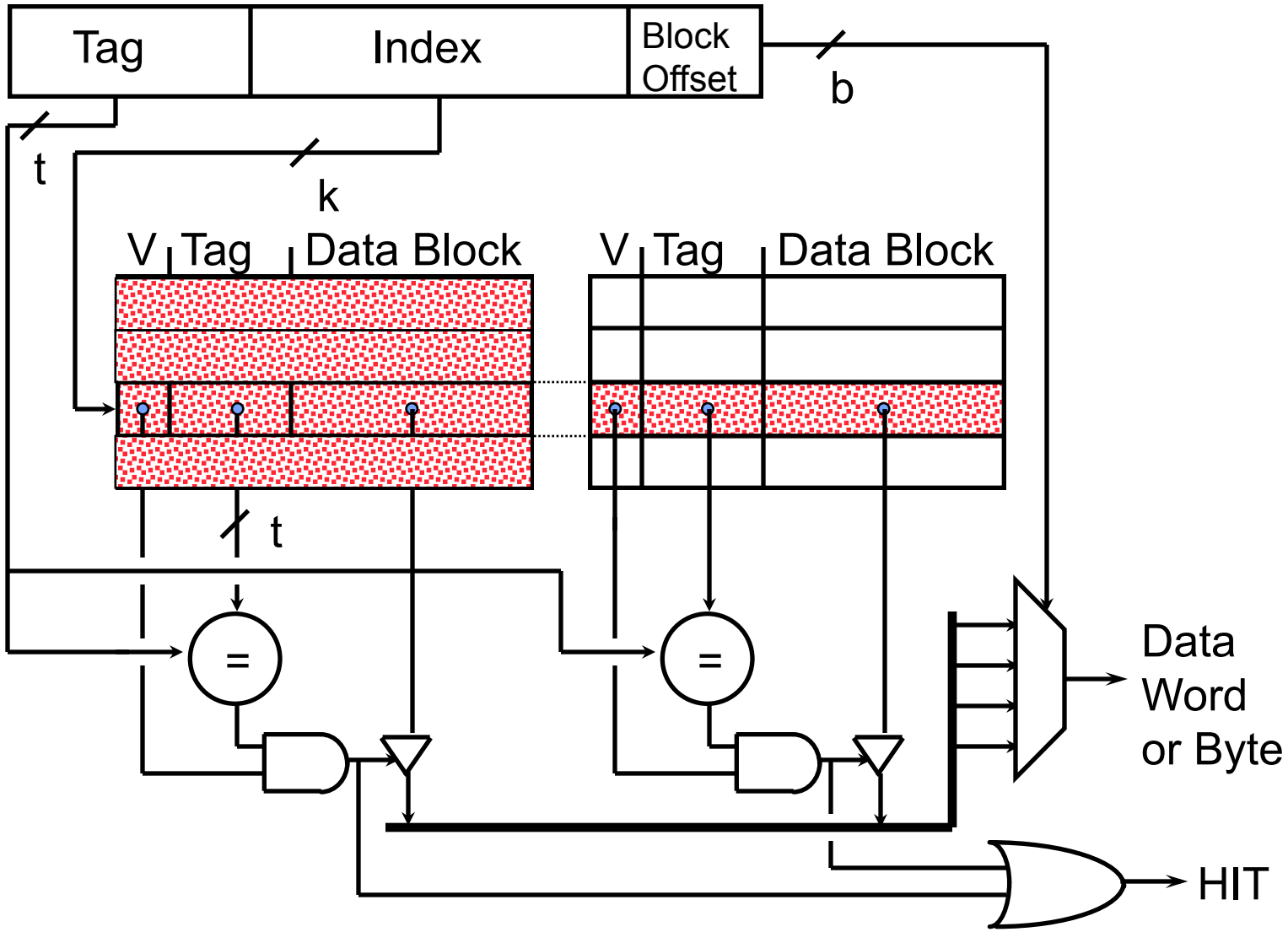
Big Modern Chip



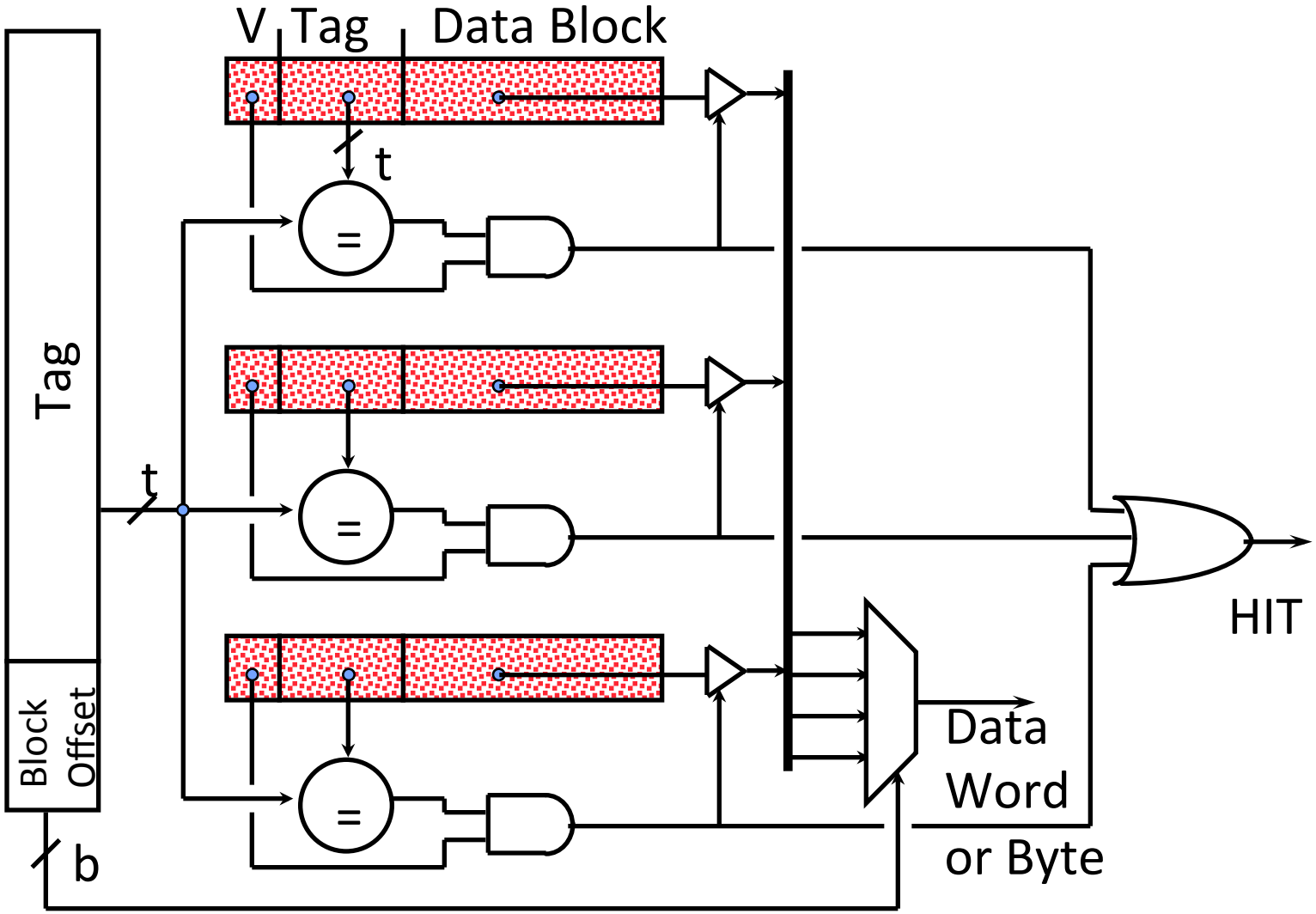
Direct-Mapped Cache



2-Way Set-Associative Cache



Fully Associative Cache



Replacement Policies

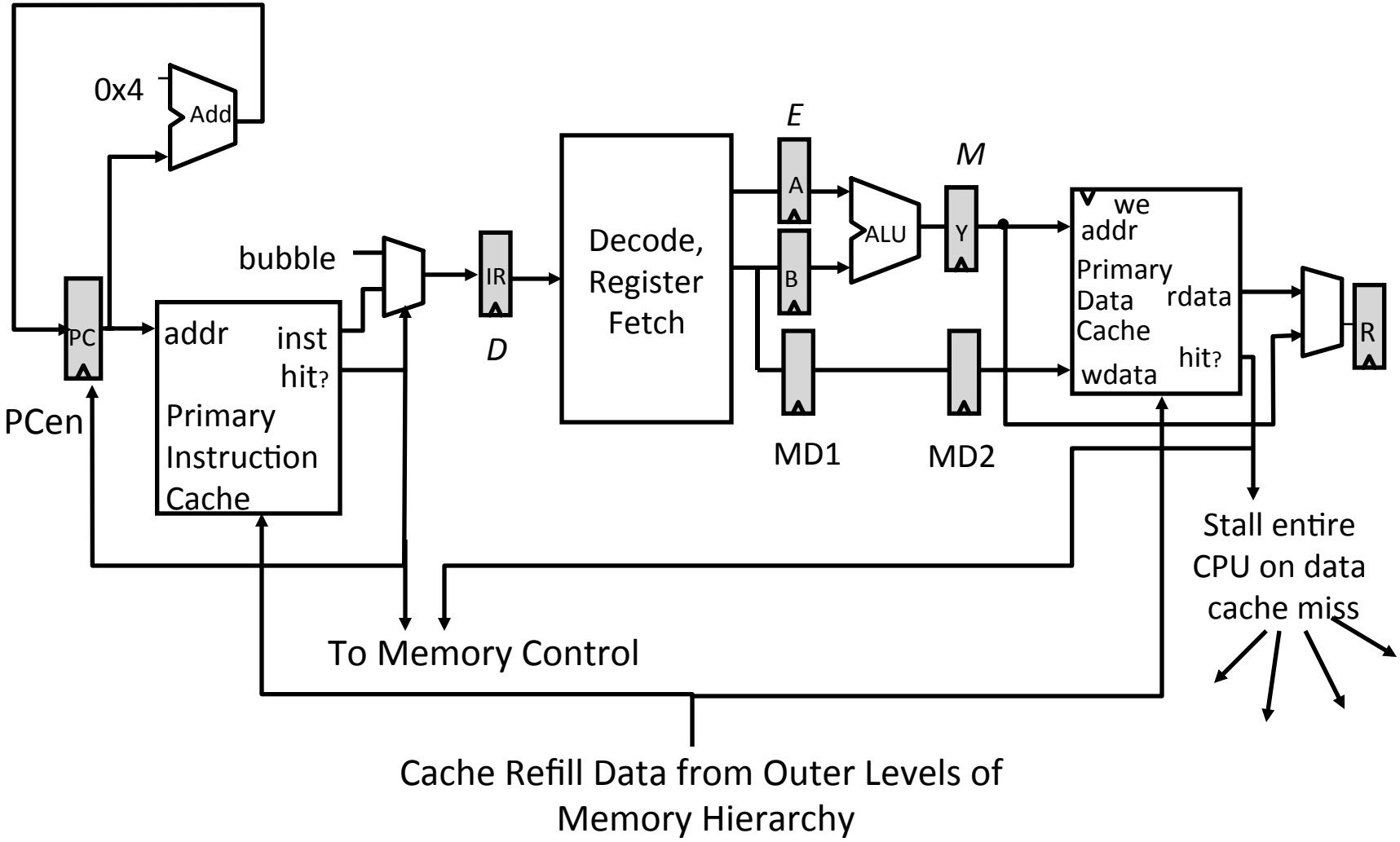
- Random
- LRU
  - Pseudo-LRU
- FIFO
- Can explore in lab 2

AMAT

Average memory access time (AMAT) =
Hit time + Miss rate x Miss penalty

Average memory access time (AMAT) =
Hit time + Miss $rate_1$ x Miss $penalty_1$ +
Miss $rate_2$ x Miss $penalty_2$

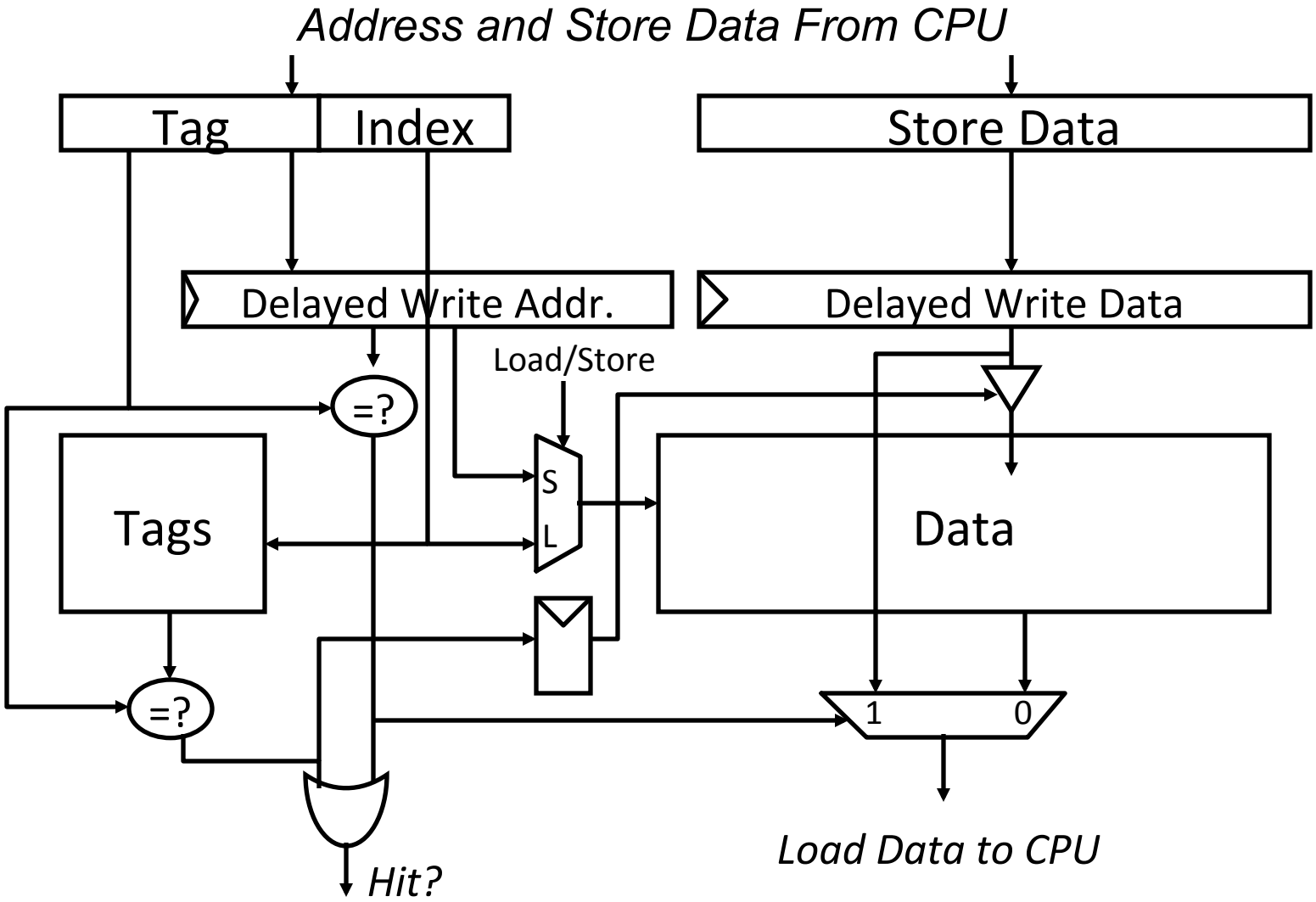
CPU-Cache Interaction
Caches instead of memory blocks
(5-stage pipeline)



Write Policy Choices

- Cache hit:
  - ***write through***: write both cache & memory
    - Generally higher traffic but simpler pipeline & cache design
  - ***write back***: write cache only, memory is written only when the entry is evicted
    - A dirty bit per line further reduces write-back traffic
    - Must handle 0, 1, or 2 accesses to memory for each load/ store
- Cache miss:
  - ***no write allocate***: only write to main memory
  - ***write allocate*** (aka fetch on write): fetch into cache
- Common combinations:
  - write through and no write allocate
  - write back with write allocate

Pipelining Cache Writes



Data from a store hit written into data portion of cache during tag access of subsequent store

Inclusion Policy

- Inclusive multilevel cache:
  - Inner cache can only hold lines also present in outer cache
  - External coherence snoop access need only check outer cache
- Exclusive multilevel caches:
  - Inner cache may hold lines not in 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?

Prefetching

- Speculate on future instruction and data accesses and fetch them into cache(s)
  - Instruction accesses easier to predict than data accesses
- Varieties of prefetching
  - Hardware prefetching
  - Software prefetching
  - Mixed schemes
- What types of misses does prefetching affect?
- Explore in Lab 2

Questions

Lab 2

- Has anyone looked at it?
- Questions?
- Lab 1 Report graded by next discussion

Graded Quiz