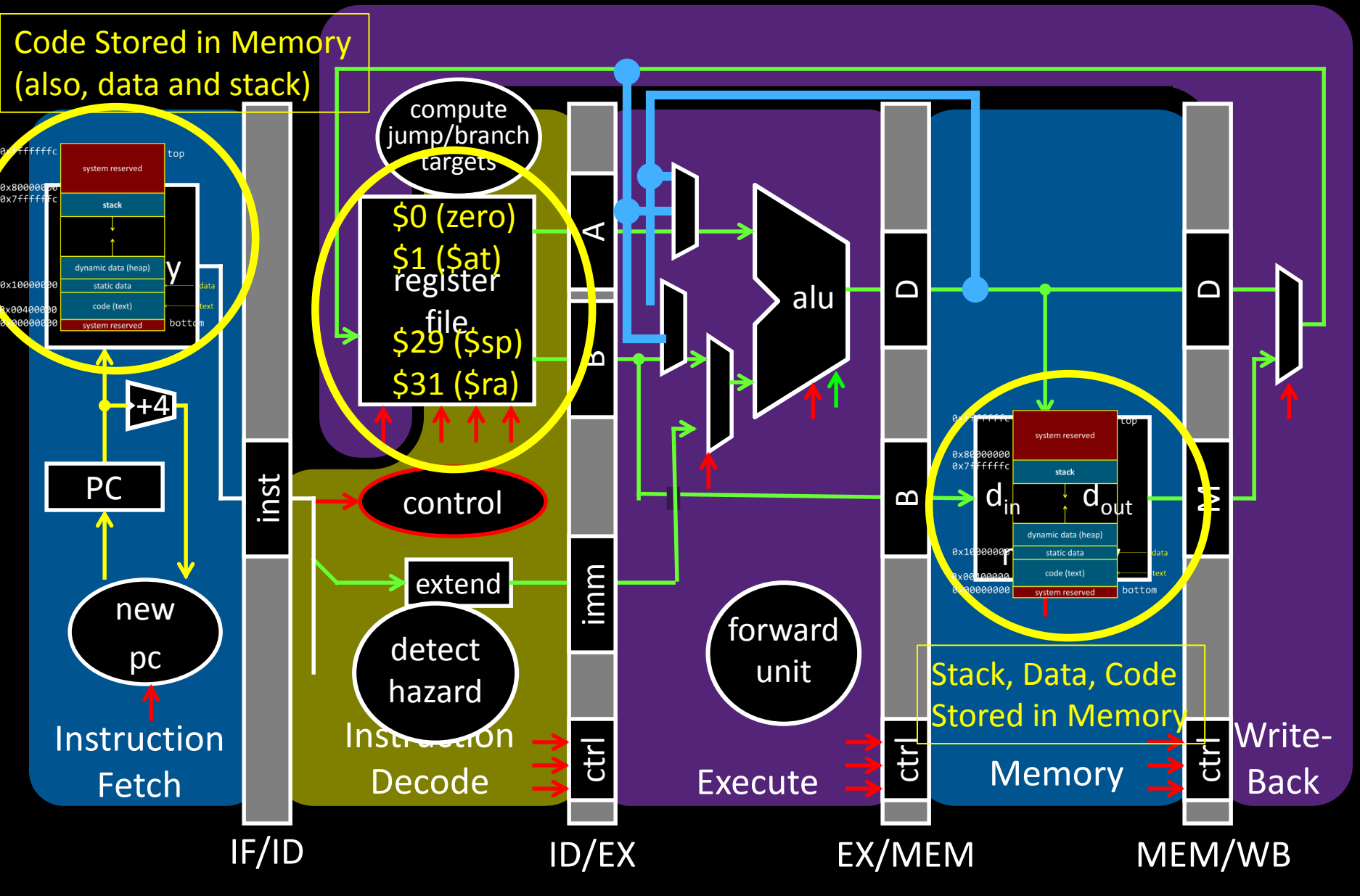
Caches 2

Hakim Weatherspoon
CS 3410, Spring 2013
Computer Science
Cornell University

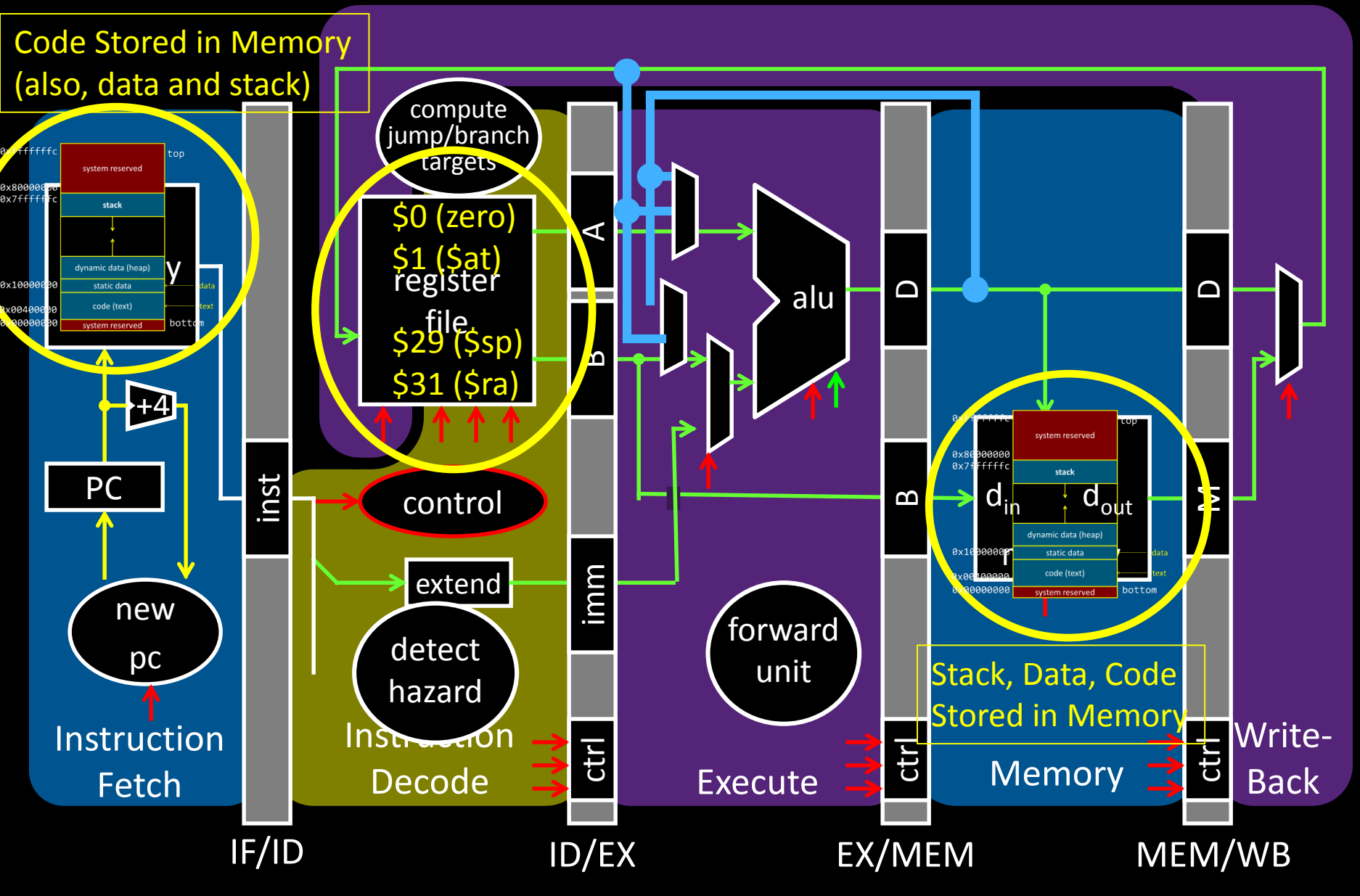
P & H Chapter 5.1-5.3, 5.3-5.4, 5.8, Also, 5.13 & 5.17

Big Picture: Memory



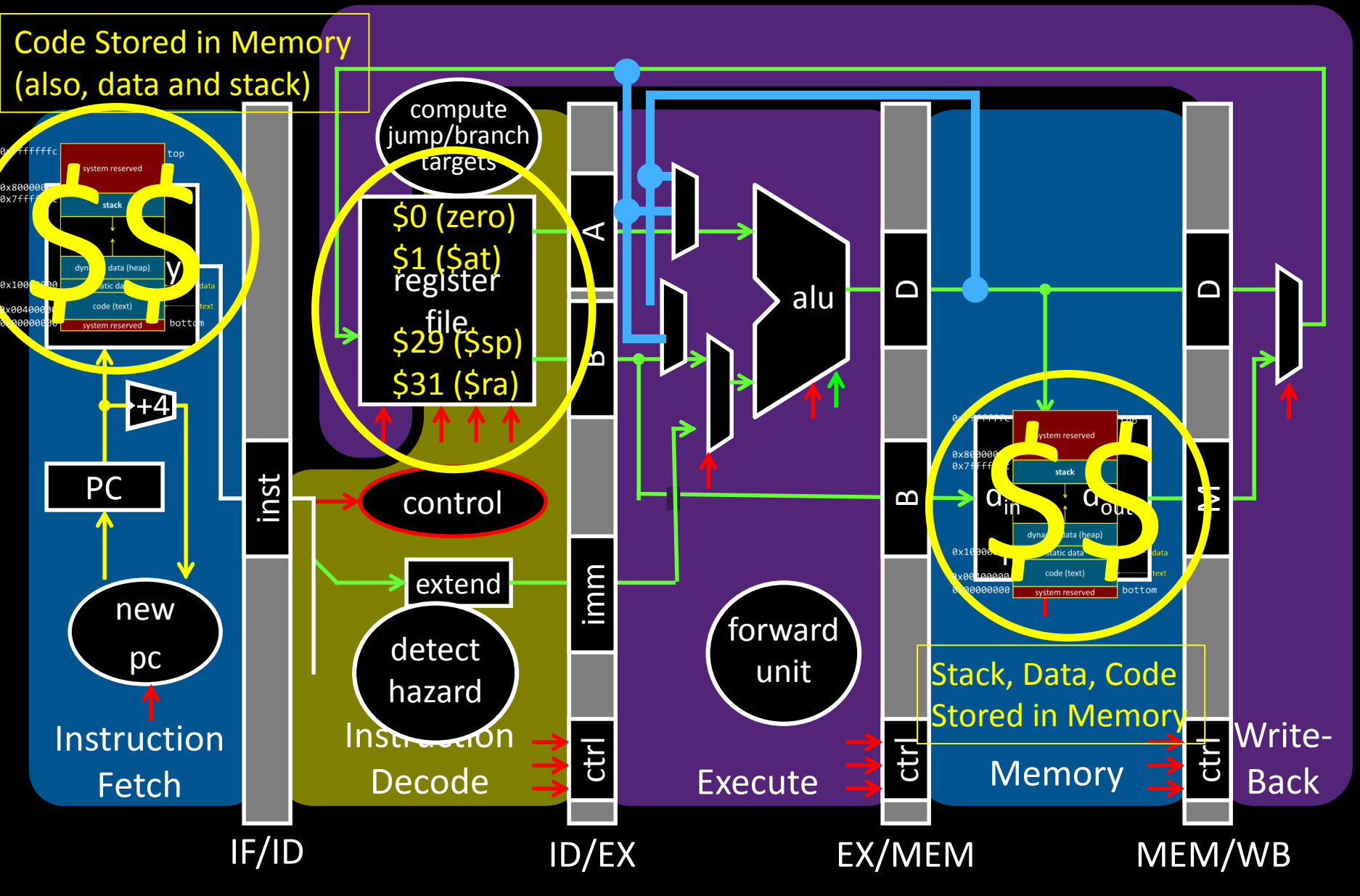
Big Picture: Memory

Memory: big & slow vs Caches: small & fast



Big Picture: Memory

Memory: big & slow vs Caches: small & fast



Big Picture

How do we make the processor fast,
Given that memory is VEEERRRYYYY SLLOOOWWW!!

Big Picture

How do we make the processor fast,
Given that memory is VEEERRRYYYY SLLOOOWWW!!

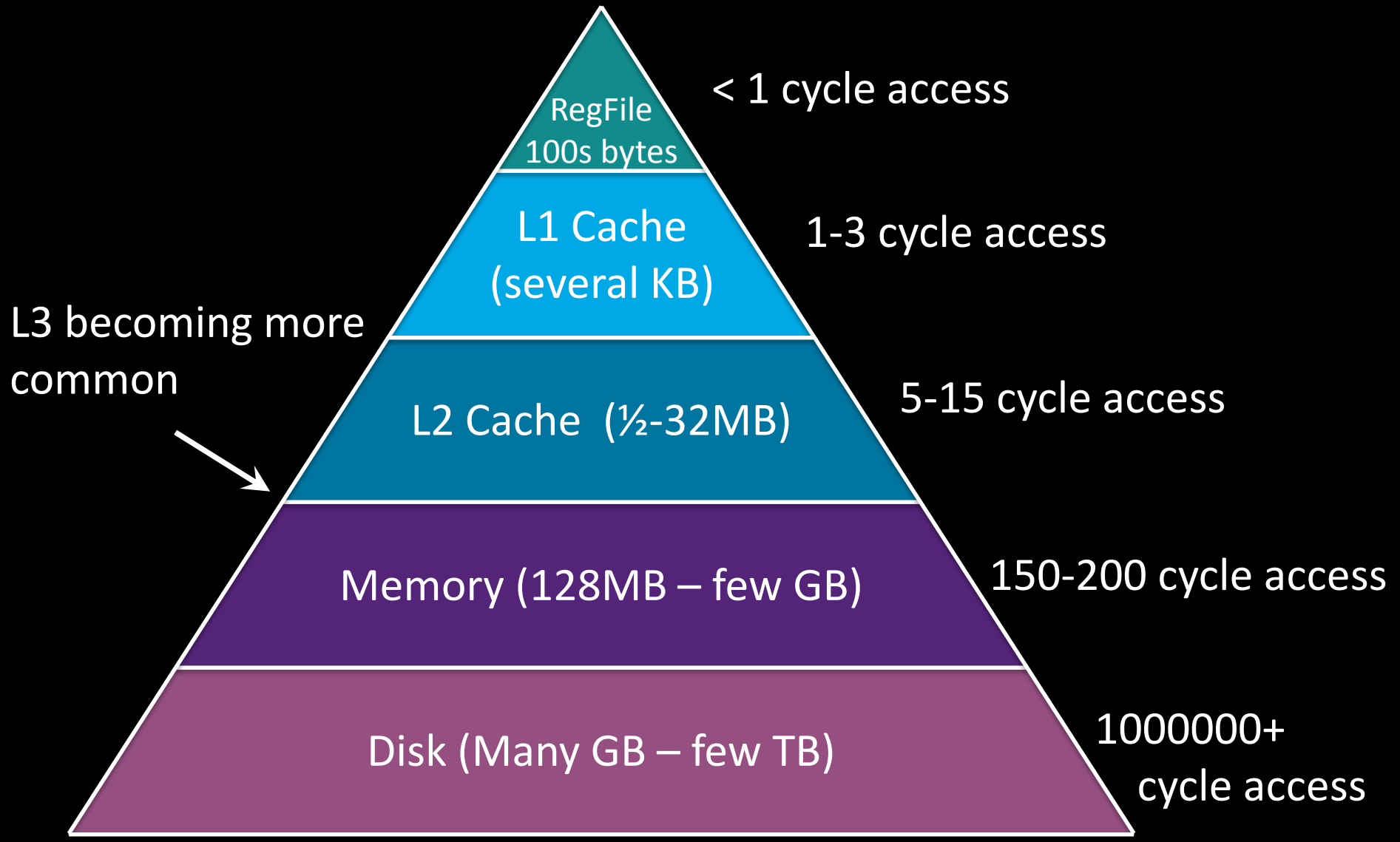
But, insight for Caches

- small working set: 90/10 rule
- can predict future: spatial & temporal locality

Benefits

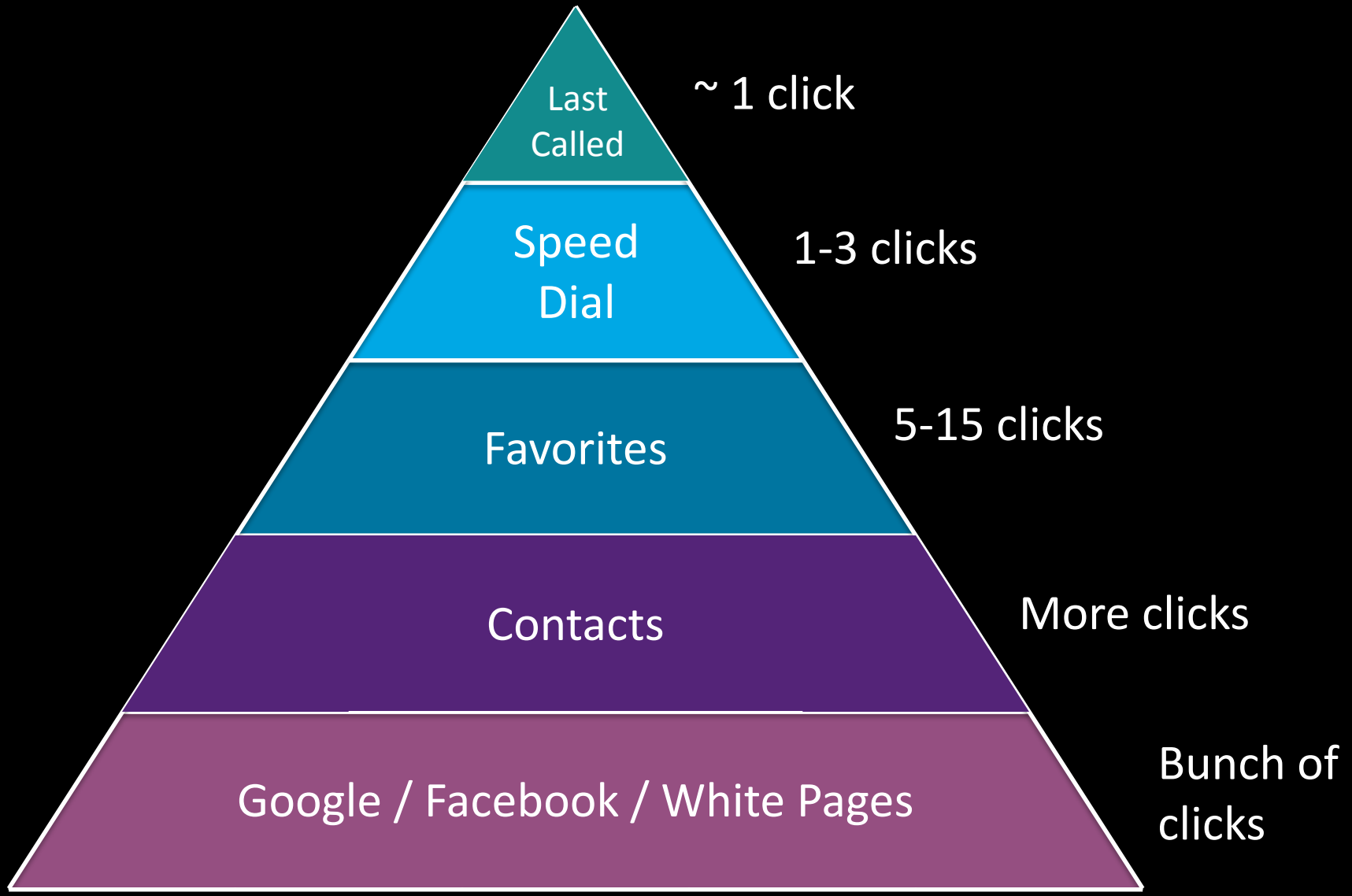
- Abstraction: big & fast memory
  - built from (big & slow memory; DRAM) + (small & fast cache; SRAM)

Memory Hierarchy



*These are rough numbers, mileage may vary.

Just like your Contacts!



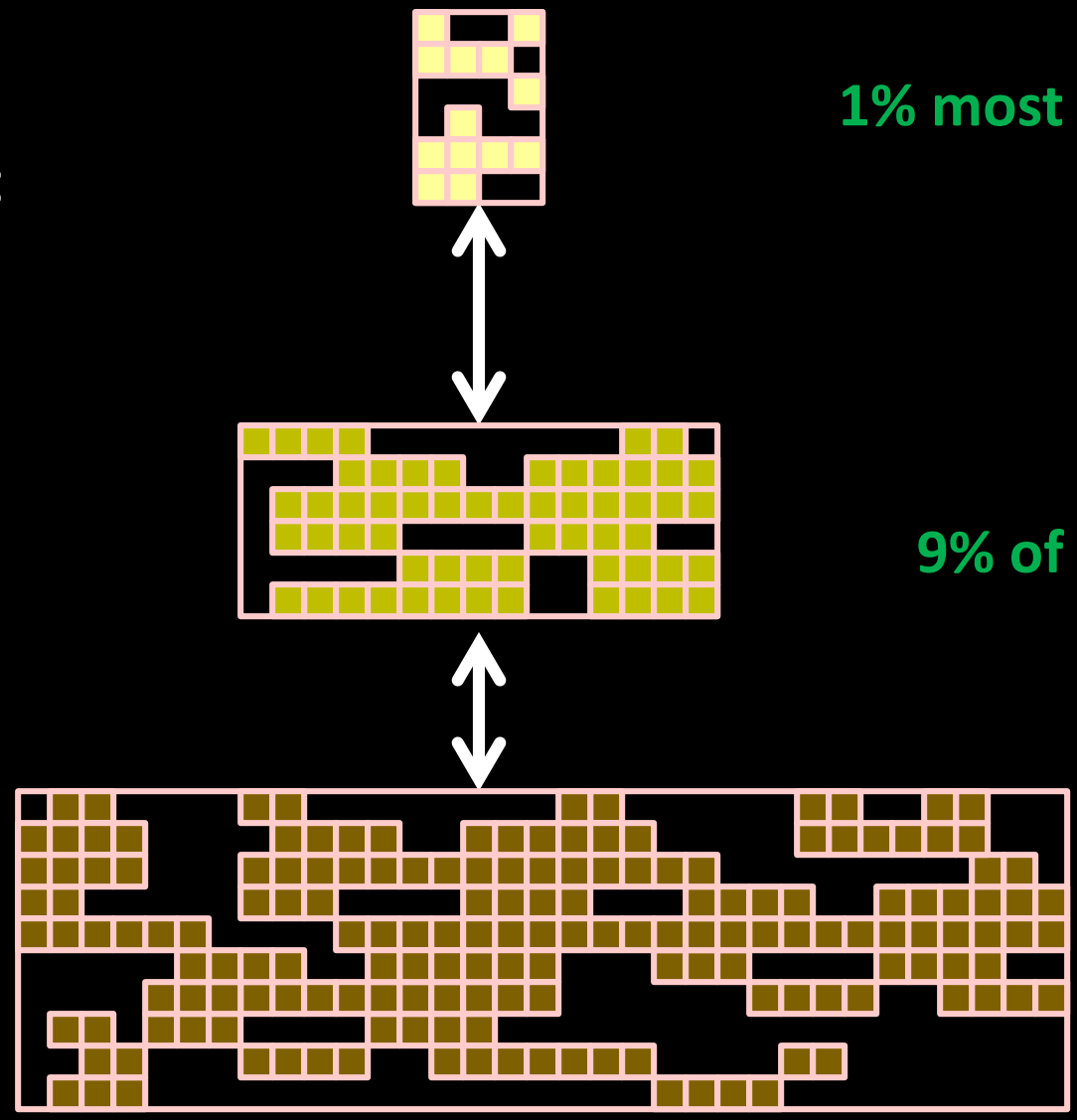
*** Will refer to this analogy using GREEN in the slides.**

Memory Hierarchy

L1 Cache
SRAM-on-chip
1% of data most accessed

L2/L3 Cache
SRAM
9% of data is "active"

Memory DRAM
90% of data inactive (not accessed)



Speed Dial
1% most called people

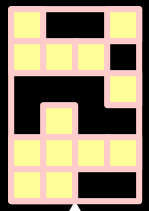
Favorites
9% of people called

Contacts
90% people rarely called

Memory Hierarchy

Memory closer to processor

- small & fast
- stores active data

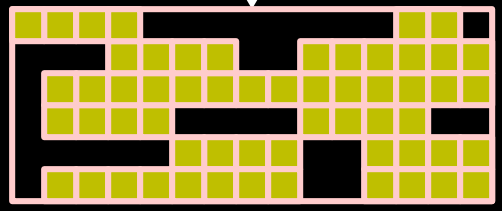


L1 Cache
SRAM-on-chip

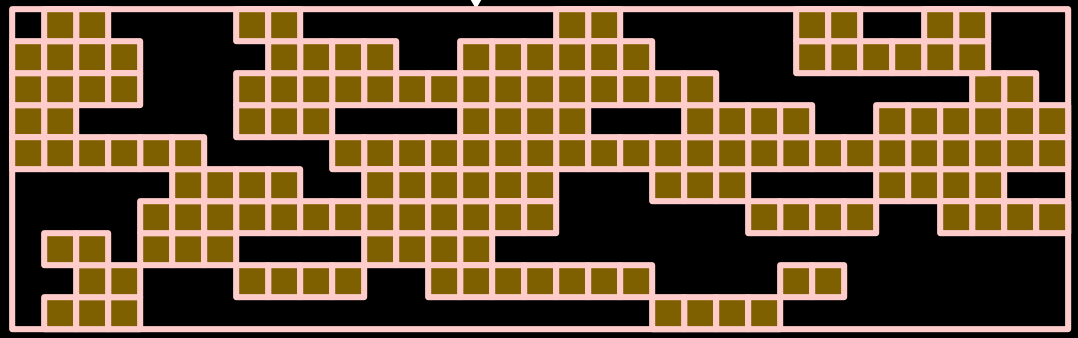
Speed Dial
small & fast

Memory farther from processor

- big & slow
- stores inactive data



L2/L3 Cache
SRAM



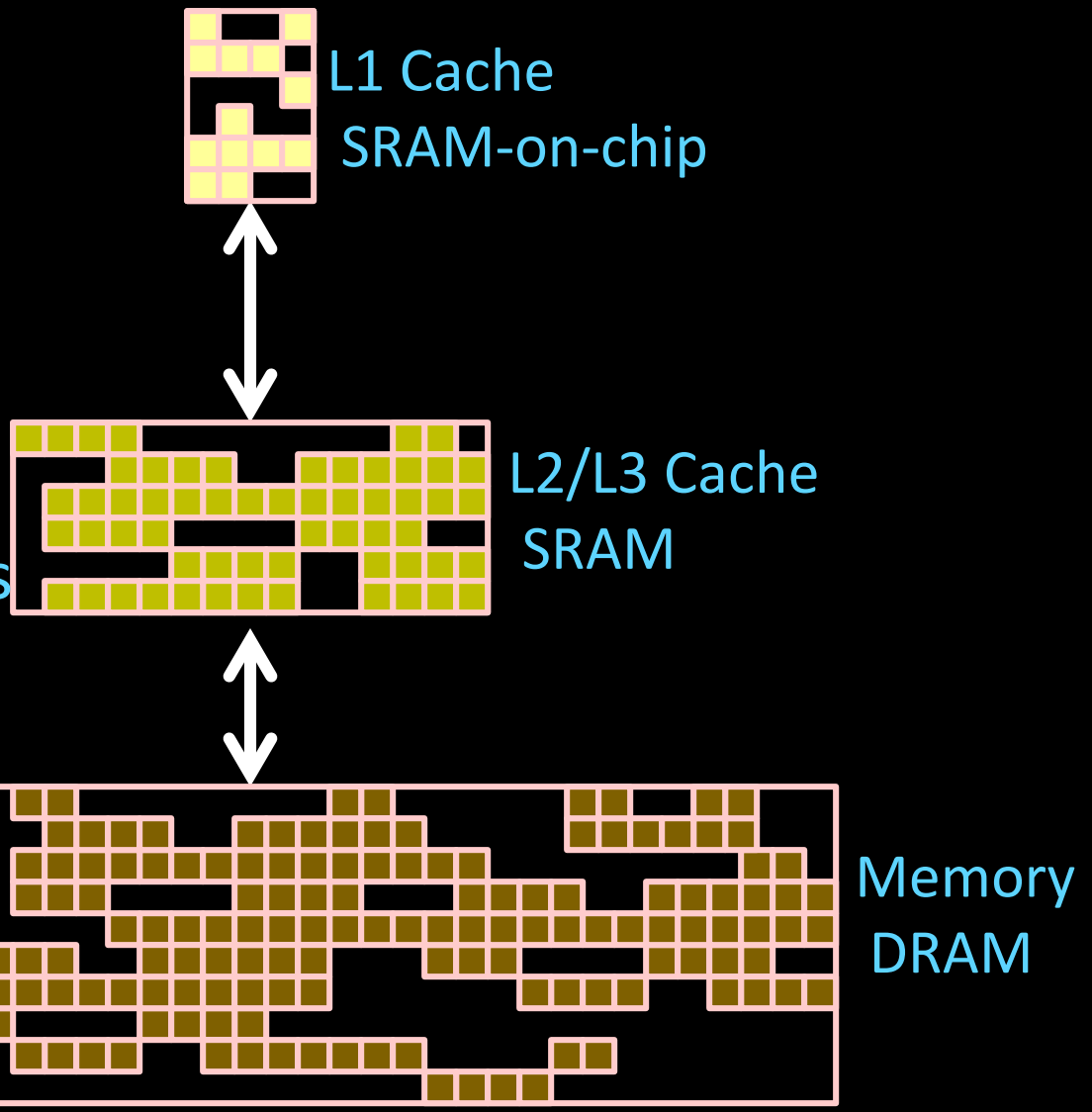
Contact List
big & slow
Memory
DRAM

Memory Hierarchy

Memory closer to processor is fast but small

usually stores subset of memory farther

- "strictly inclusive"



Transfer whole blocks (cache lines):

4kB: disk ↔ RAM

256B: RAM↔L2

64B: L2 ↔ L1

Goals for Today: caches

Comparison of cache architectures:

- Direct Mapped
- Fully Associative
- N-way set associative

Caching Questions

- How does a cache work?
- How effective is the cache (hit rate/miss rate)?
- How large is the cache?
- How fast is the cache (AMAT=average memory access time)

Next time: Writing to the Cache

- Write-through vs Write-back

Next Goal

How do the different cache architectures compare?

- Cache Architecture Tradeoffs?
- Cache Size?
- Cache Hit rate/Performance?

Cache Tradeoffs

A given data block can be placed…

- … in any cache line → Fully Associative

  (a contact maps to *any* speed dial number)

- … in exactly one cache line → Direct Mapped

  (a contact maps to exactly one speed dial number)

- … in a small set of cache lines → Set Associative

  (like direct mapped, a contact maps to exactly one speed dial number, but many different contacts can associate with the same speed dial number at the same time)

Cache Tradeoffs

Direct Mapped		Fully Associative
+ Smaller	Tag Size	Larger –
+ Less	SRAM Overhead	More –
+ Less	Controller Logic	More –
+ Faster	Speed	Slower –
+ Less	Price	More –
+ Very	Scalability	Not Very –
– Lots	# of conflict misses	Zero +
– Low	Hit rate	High +
– Common	Pathological Cases?	?

Cache Tradeoffs

Compromise: Set-associative cache

Like a direct-mapped cache

- Index into a location
- Fast

Like a fully-associative cache

- Can store multiple entries
  - decreases thrashing in cache
- Search in each element

Direct Mapped Cache

- Use the first letter to index!

Speed Dial

2	ABC	Baker, J.
3	DEF	Dunn, A.
4	GHI	Gill, S.
5	JKL	Jones, N.
6	MNO	Mann, B.
7	PQRS	Powell, J.
8	TUV	Taylor, B.
9	WXYZ	Wright, T.

Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Jones, N.	777-777-7777
Lee, V.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

Block Size: 1 Contact Number

Fully Associative Cache

- No index!

Speed Dial

2	Baker, J.
3	Dunn, A.
4	Gill, S.
5	Jones, N.
6	Mann, B.
7	Powell, J.
8	Taylor, B.
9	Wright, T.

Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Jones, N.	777-777-7777
Lee, V.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

Block Size: 1 Contact Number

Fully Associative Cache

- No index!

Speed Dial

2	Mann, B.
3	Dunn, A.
4	Taylor, B.
5	Wright, T.
6	Baker, J.
7	Powell, J.
8	Gill, S.
9	Jones, N.

Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Jones, N.	777-777-7777
Lee, V.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

Block Size: 1 Contact Number

Fully Associative Cache

- No index!
- Use the initial to offset!

Speed Dial

2	Baker, J.	Baker, S.
3	Dunn, A.	Foster, D.
4	Gill, D.	Harris, F.
5	Jones, N.	Lee, V.
6	Mann, B.	Moore, F.
7	Powell, C.	Sam, J.
8	Taylor, B.	Taylor, O.
9	Wright, T.	Zhang, W.

Block Size: 2 Contact Numbers

Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Jones, N.	777-777-7777
Lee, V.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

Fully Associative Cache

- No index!
- Use the initial to offset!

Speed Dial

2	Mann, B.	Moore, F.
3	Powell, C.	Sam, J.
4	Gill, D.	Harris, F.
5	Wright, T.	Zhang, W.
6	Baker, J.	Baker, S.
7	Dunn, A.	Foster, D.
8	Taylor, B.	Taylor, O.
9	Jones, N.	Lee, V.

Block Size: 2 Contact Numbers

Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Jones, N.	777-777-7777
Lee, V.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

N-way Set Associative Cache

- 2-way set associative cache
- 8 sets
- Use the initial to offset!

Speed Dial

2	ABC	Baker, J.	Baker, S.		
3	DEF	Dunn, A.	Foster, D.		
4	GHI	Gill, D.	Harris, F.		
5	JKL	Jones, N.	Lee, V.		
6	MNO	Mann, B.	Moore, F.		
7	PQRS	Powell, C.	Sam, J.		
8	TUV	Taylor, B.	Taylor, O.		
9	WXYZ	Wright, T.	Zhang, W.		

Block Size: 2 Contact Numbers

Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Jones, N.	777-777-7777
Lee, V.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

N-way Set Associative Cache

- 2-way set associative cache
- 8 sets
- Use the initial to offset!

Speed Dial

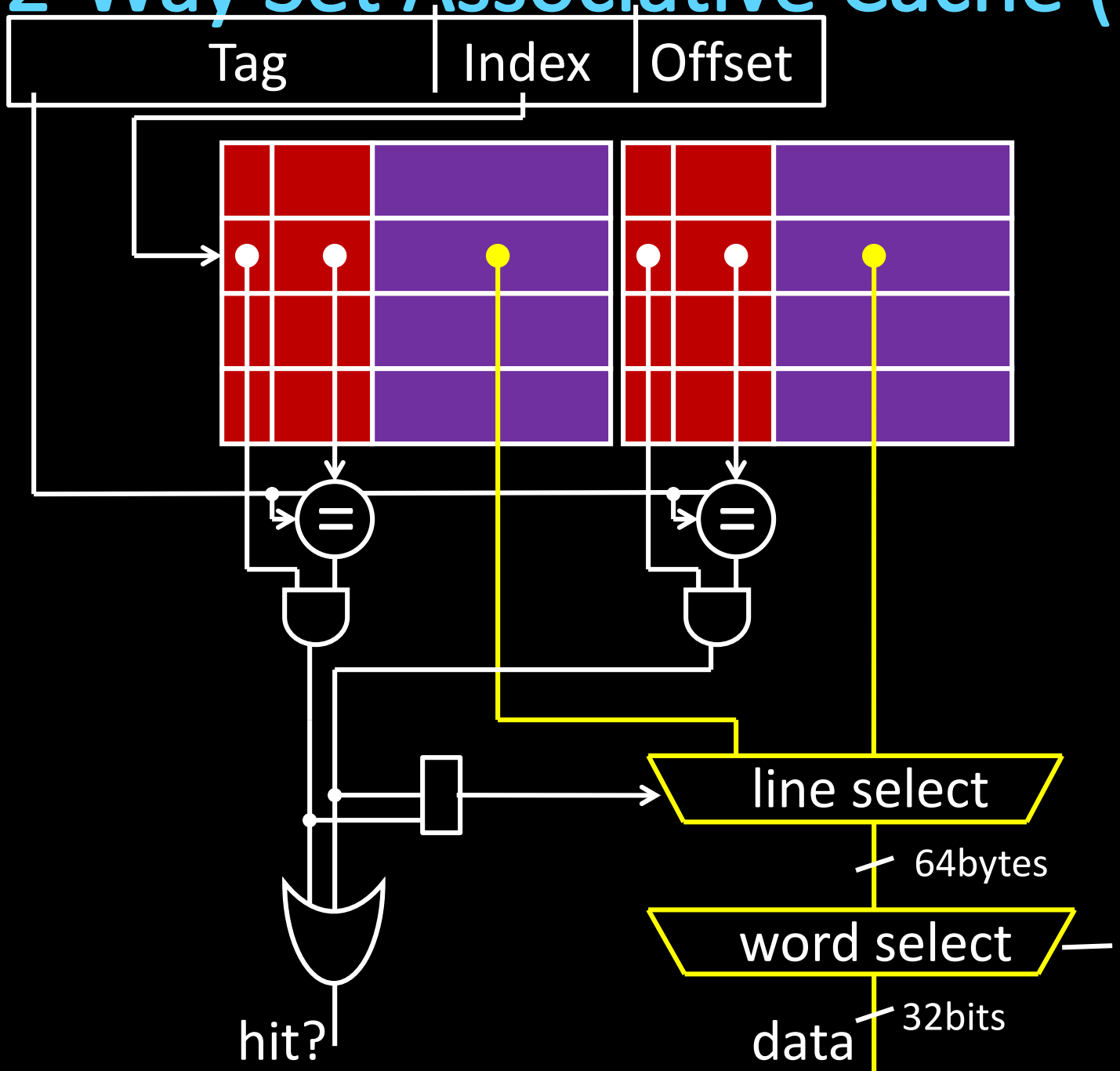
2	ABC	Baker, J.	Baker, S.		
3	DEF	Dunn, A.	Foster, D.		
4	GHI	Gill, D.	Harris, F.	Henry, J.	Isaacs, M.
5	JKL				
6	MNO	Mann, B.	Moore, F.		
7	PQRS	Powell, C.	Sam, J.		
8	TUV	Taylor, B.	Taylor, O.		
9	WXYZ	Wright, T.	Zhang, W.		

Block Size: 2 Contact Numbers

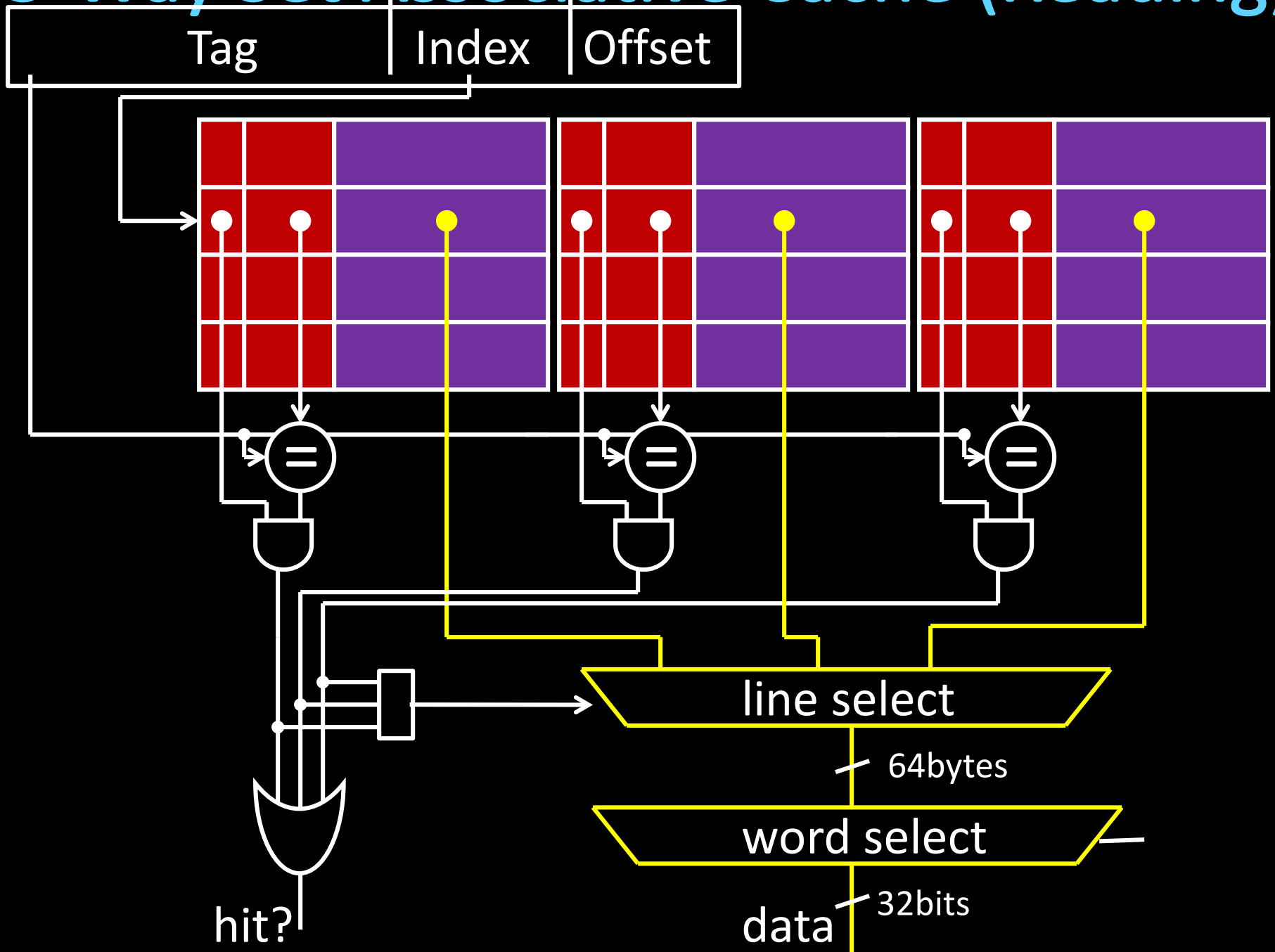
Contacts

Baker, J.	111-111-1111
Baker, S.	222-222-2222
Dunn, A.	333-333-3333
Foster, D.	444-444-4444
Gill, D.	555-555-5555
Harris, F.	666-666-6666
Henry, J.	777-777-7777
Isaacs, M.	888-888-8888
Mann, B.	111-111-1119
Moore, F.	222-222-2229
Powell, C.	333-333-3339
Sam, J.	444-444-4449
Taylor, B.	555-555-5559
Taylor, O.	666-666-6669
Wright, T.	777-777-7779
Zhang, W.	888-888-8889

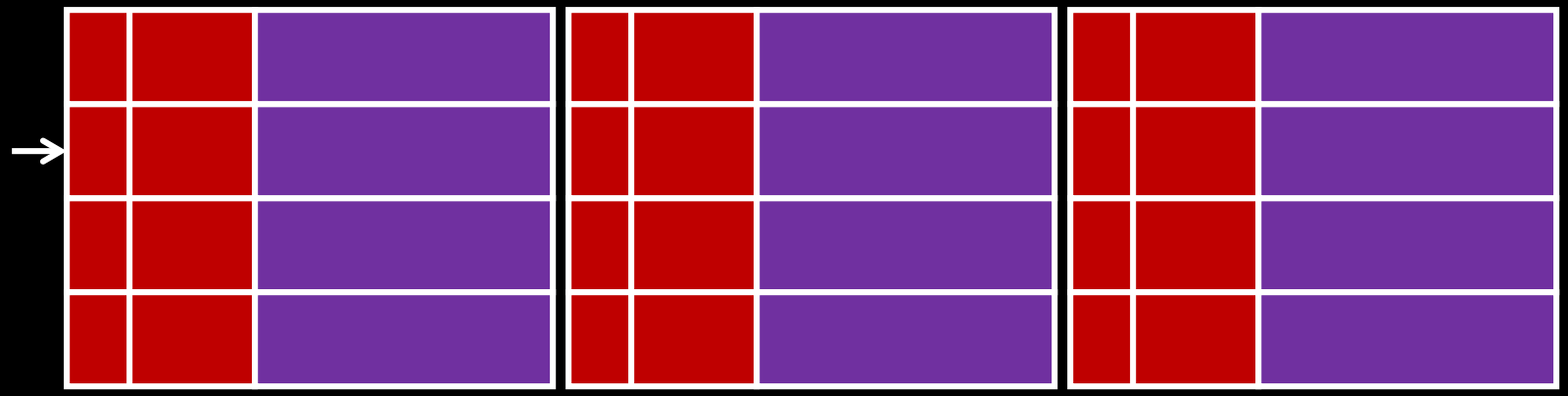
2-Way Set Associative Cache (Reading)



3-Way Set Associative Cache (Reading)



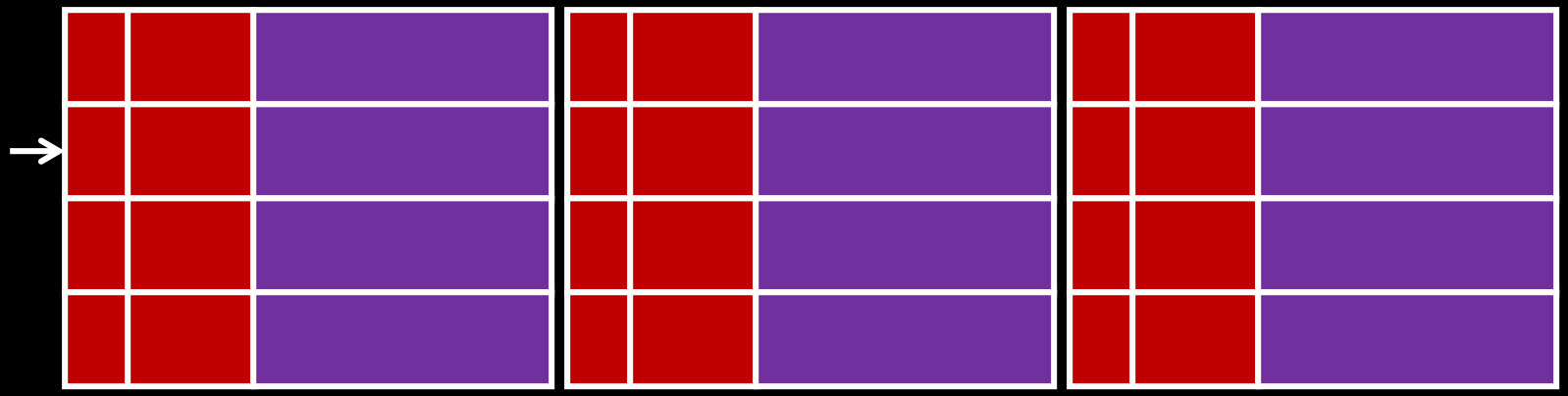
3-Way Set Associative Cache (Reading)



n bit index, *m* bit offset, **N-way Set Associative**

Q: How big is cache (***data only***)?

3-Way Set Associative Cache (Reading)

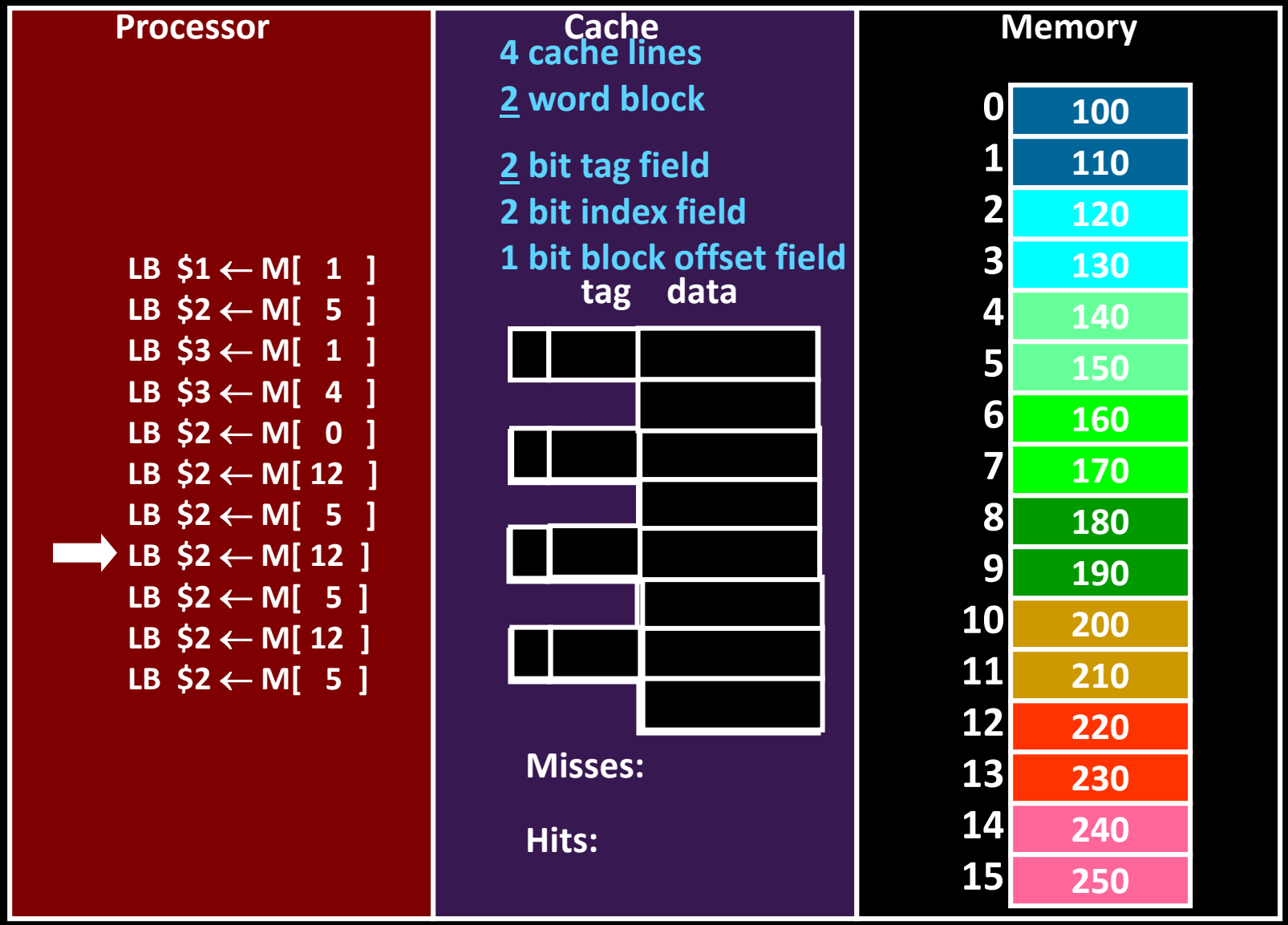


n bit index, *m* bit offset, **N-way Set Associative**

Q: How much SRAM is needed (***data + overhead***)?

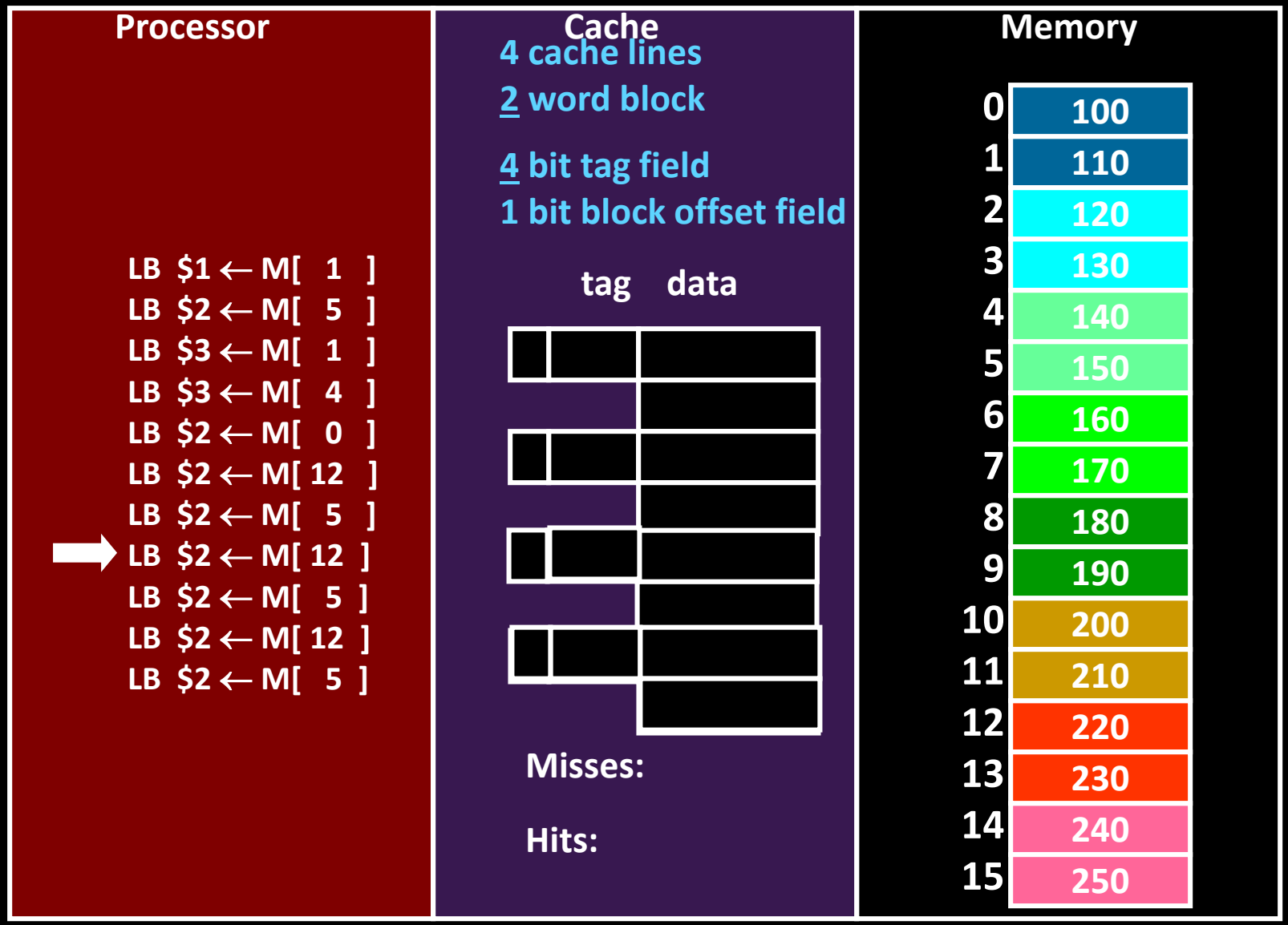
Comparison: Direct Mapped

Using **byte addresses** in this example! Addr Bus = 5 bits



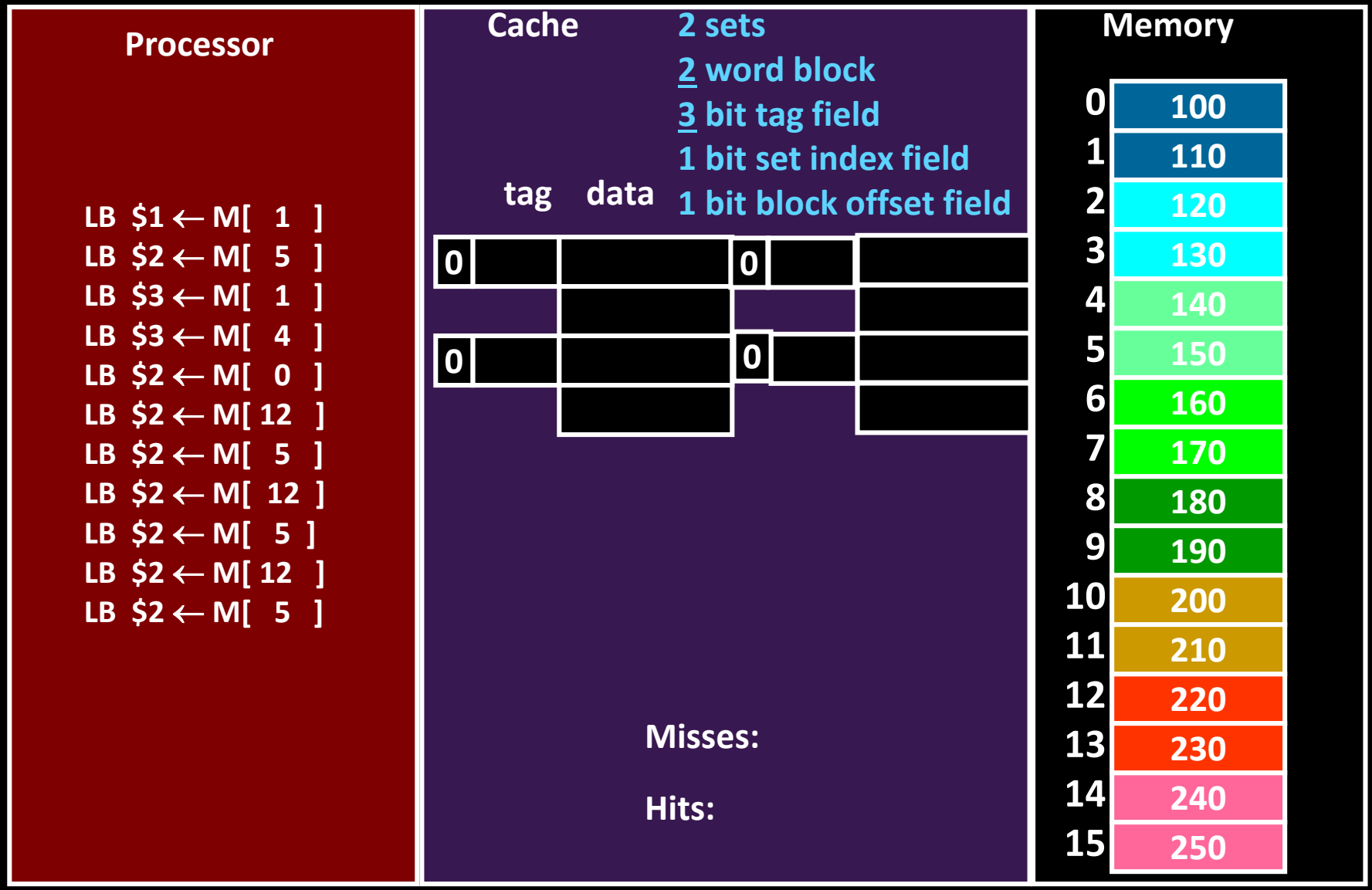
Comparison: Fully Associative

Using **byte addresses** in this example! Addr Bus = 5 bits



Comparison: 2 Way Set Assoc

Using **byte addresses** in this example! Addr Bus = 5 bits



Misses

Cache misses: classification

The line is being referenced for the first time

- Cold (aka Compulsory) Miss

The line was in the cache, but has been evicted…

… because some other access with the same index

- Conflict Miss

… because the cache is too small

- i.e. the *working set* of program is larger than the cache
- Capacity Miss

Takeaway

Direct Mapped → fast but low hit rate

Fully Associative → higher hit cost, but higher hit rate

N-way Set Associative → middleground

Next Goal

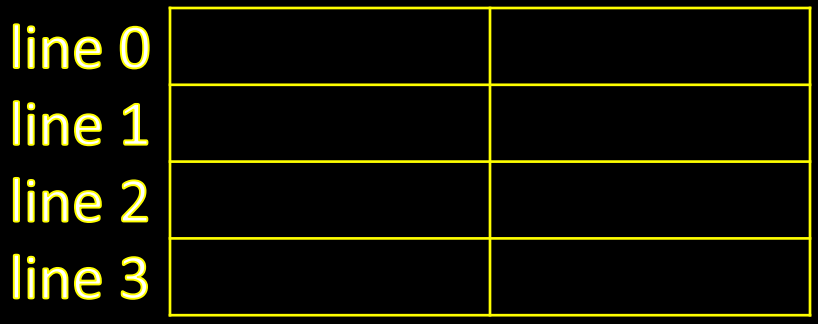
Do larger caches and larger cachelines (aka cache blocks) increase hit rate?

Larger Cachelines

Bigger does not always help

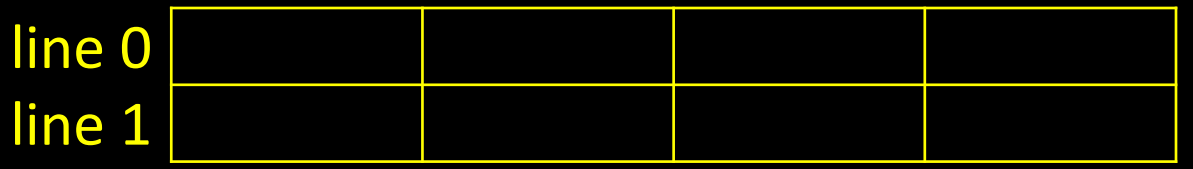
Mem access trace: 0, 16, 1, 17, 2, 18, 3, 19, 4, 20, …

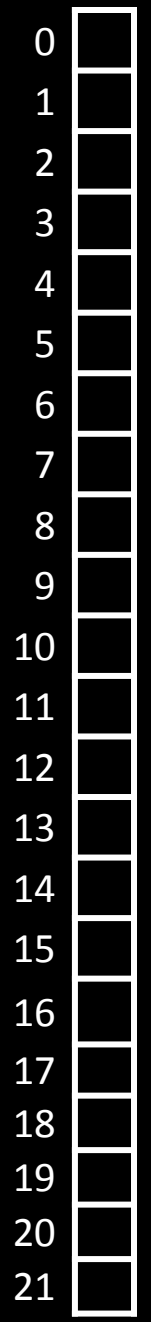
Hit rate with four direct mapped 2-byte cache lines?



With eight 2-byte cache lines?

With two 4-byte cache lines?





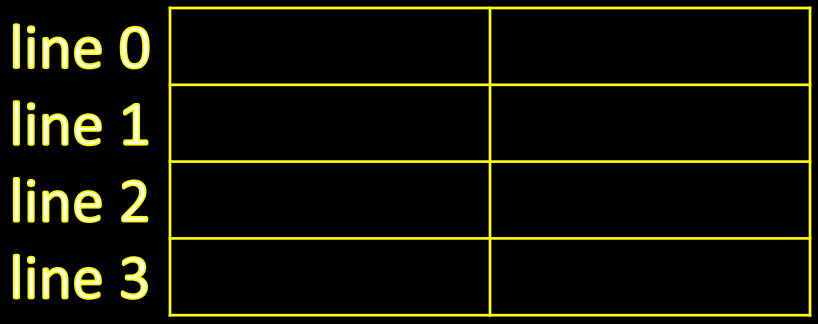
Larger Cachelines

Fully-associative reduces conflict misses...

... assuming good eviction strategy

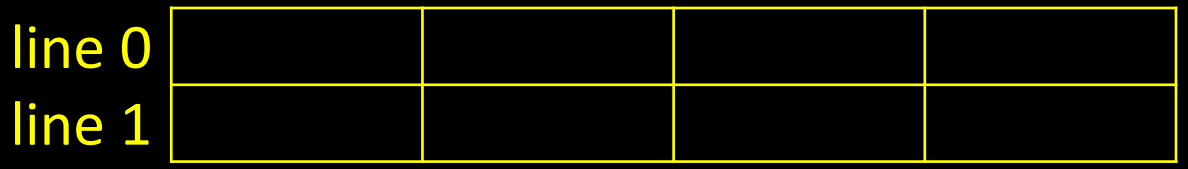
Mem access trace: 0, 16, 1, 17, 2, 18, 3, 19, 4, 20, …

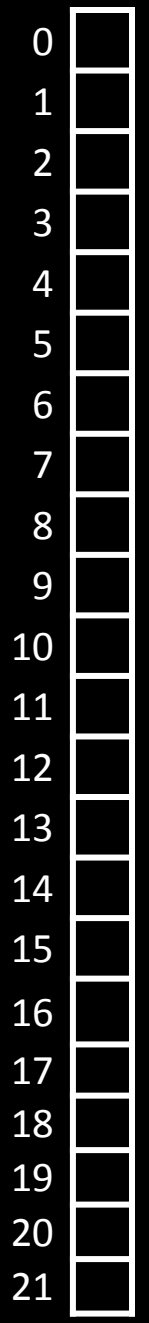
Hit rate with four fully-associative 2-byte cache lines?



… A larger block size can often increase hit rate

With two fully-associative 4-byte cache lines?





Larger Cachelines

… but large block size can still reduce hit rate

Mem access trace: 0, 100, 200, 1, 101, 201, 2, 102, 202,…

Hit rate with four fully-associative 2-byte cache lines?

line 0		
line 1		
line 2		
line 3		

With two fully-associative 4-byte cache lines?

line 0				
line 1				

Misses

Cache misses: classification

Cold (aka Compulsory)

- The line is being referenced for the first time

Capacity

- The line was evicted because the cache was too small
- i.e. the *working set* of program is larger than the cache

Conflict

- The line was evicted because of another access whose index conflicted

Next Goal

How to decide on Cache Organization

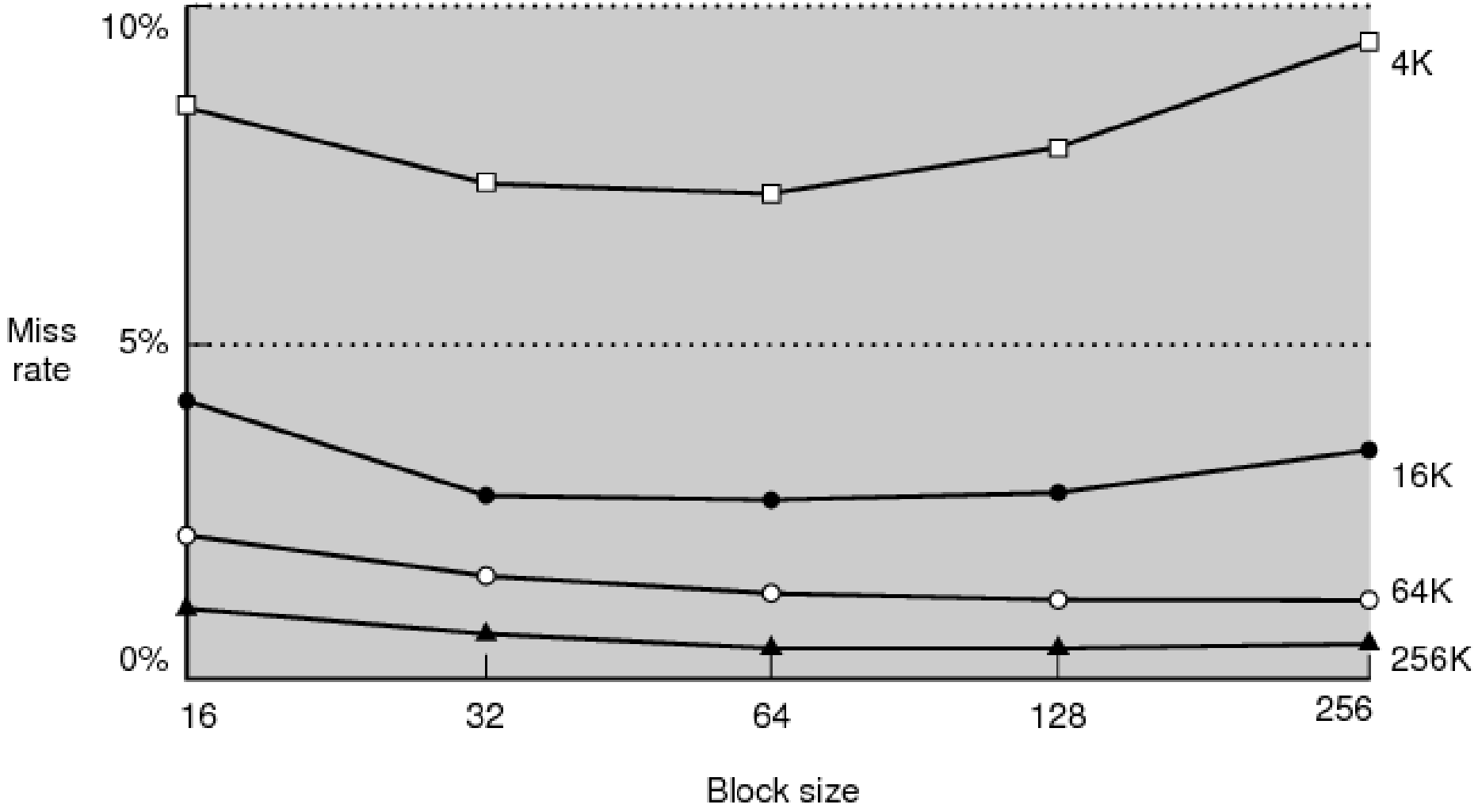
Q: How to decide block size?

A: Try it and see

But: depends on cache size, workload, associativity, …

Experimental approach!

Experimental Results



Tradeoffs

For a given total cache size,

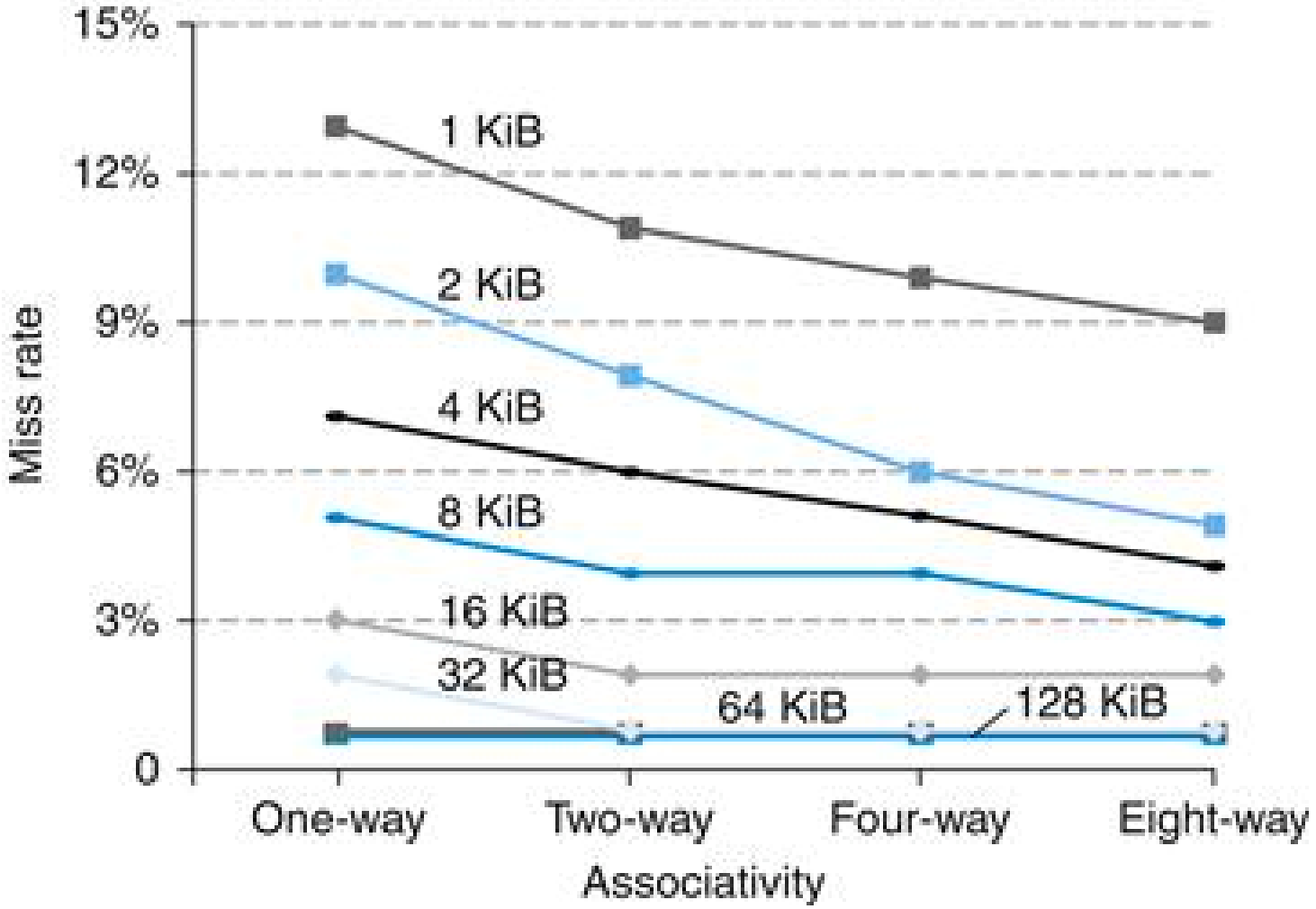
larger block sizes mean….

- fewer lines
- so fewer tags, less overhead
- and fewer cold misses (within-block “prefetching”)

But also…

- fewer blocks available (for scattered accesses!)
- so more conflicts
- and larger miss penalty (time to fetch block)

Associativity



Other Designs

Multilevel caches

Takeaway

Direct Mapped → fast but low hit rate

Fully Associative → higher hit cost, but higher hit rate

N-way Set Associative → middleground

Cacheline size matters. Larger cache lines can increase performance due to prefetching. BUT, can also decrease performance if **working set** size cannot fit in cache.

Next Goal

Performance: What is the average memory access time (AMAT) for a cache?

AMAT = %hit x hit time + % miss x miss time

Cache Performance Example

Average Memory Access Time (AMAT)

Cache Performance (very simplified):

L1 (SRAM): 512 x 64 byte cache lines, direct mapped

Data cost: 3 cycle per word access

Lookup cost: 2 cycle

16 words (i.e. 64 / 4 = 16)

Mem (DRAM): 4GB

Data cost: 50 cycle for first word, plus 3 cycles per subsequent word

Cache Performance Example

Average Memory Access Time (AMAT)

Cache Performance (very simplified):

L1 (SRAM): 512 x 64 byte cache lines, direct mapped

Data cost: 3 cycle per word access

Lookup cost: 2 cycle

16 words (i.e. 64 / 4 = 16)

Mem (DRAM): 4GB

Data cost: 50 cycle for first word, plus 3 cycles per subsequent word

Multi Level Caching

Cache Performance (very simplified):

L1 (SRAM): 512 x 64 byte cache lines, direct mapped

Hit time: 5 cycles

L2 cache: bigger

Hit time = 20 cycles

Mem (DRAM): 4GB

Hit rate: 90% in L1, 90% in L2

Often: L1 fast and direct mapped, L2 bigger and higher associativity

Performance Summary

Average memory access time (AMAT)

depends on cache architecture and size

access time for hit,

miss penalty, miss rate

Cache design a very complex problem:

- Cache size, block size (aka line size)
- Number of ways of set-associativity (1, N, ∞)
- Eviction policy
- Number of levels of caching, parameters for each
- Separate I-cache from D-cache, or Unified cache
- Prefetching policies / instructions
- Write policy

Takeaway

Direct Mapped → fast but low hit rate

Fully Associative → higher hit cost, but higher hit rate

N-way Set Associative → middleground

Cacheline size matters. Larger cache lines can increase performance due to prefetching. BUT, can also decrease performance is **working set** size cannot fit in cache.

Ultimately, cache performance is measured by the average memory access time (AMAT), which depends cache architecture and size, but also the Access time for hit, miss penalty, hit rate

Goals for Today: caches

Comparison of cache architectures:

- Direct Mapped
- Fully Associative
- N-way set associative

Caching Questions

- How does a cache work?
- How effective is the cache (hit rate/miss rate)?
- How large is the cache?
- How fast is the cache (AMAT=average memory access time)

Next time: Writing to the Cache

- Write-through vs Write-back

Next time