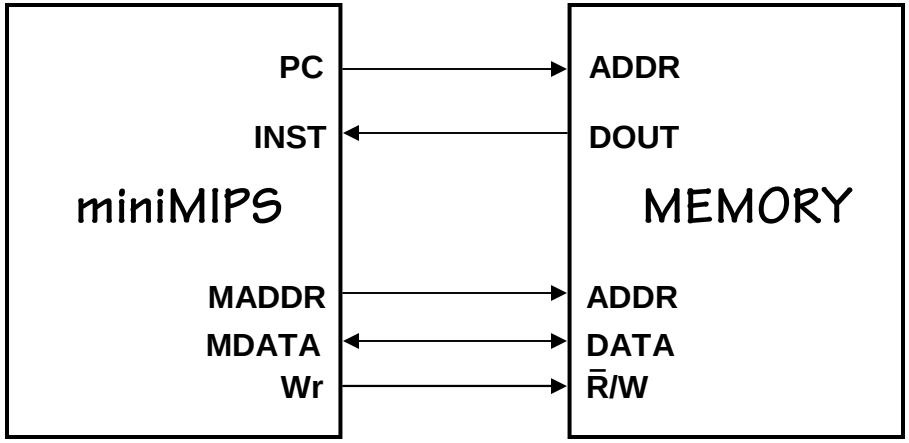
Memory Hierarchy

- Memory Flavors
- Principle of Locality
- Program Traces
- Memory Hierarchies
- Associativity

(Study Chapter 7)

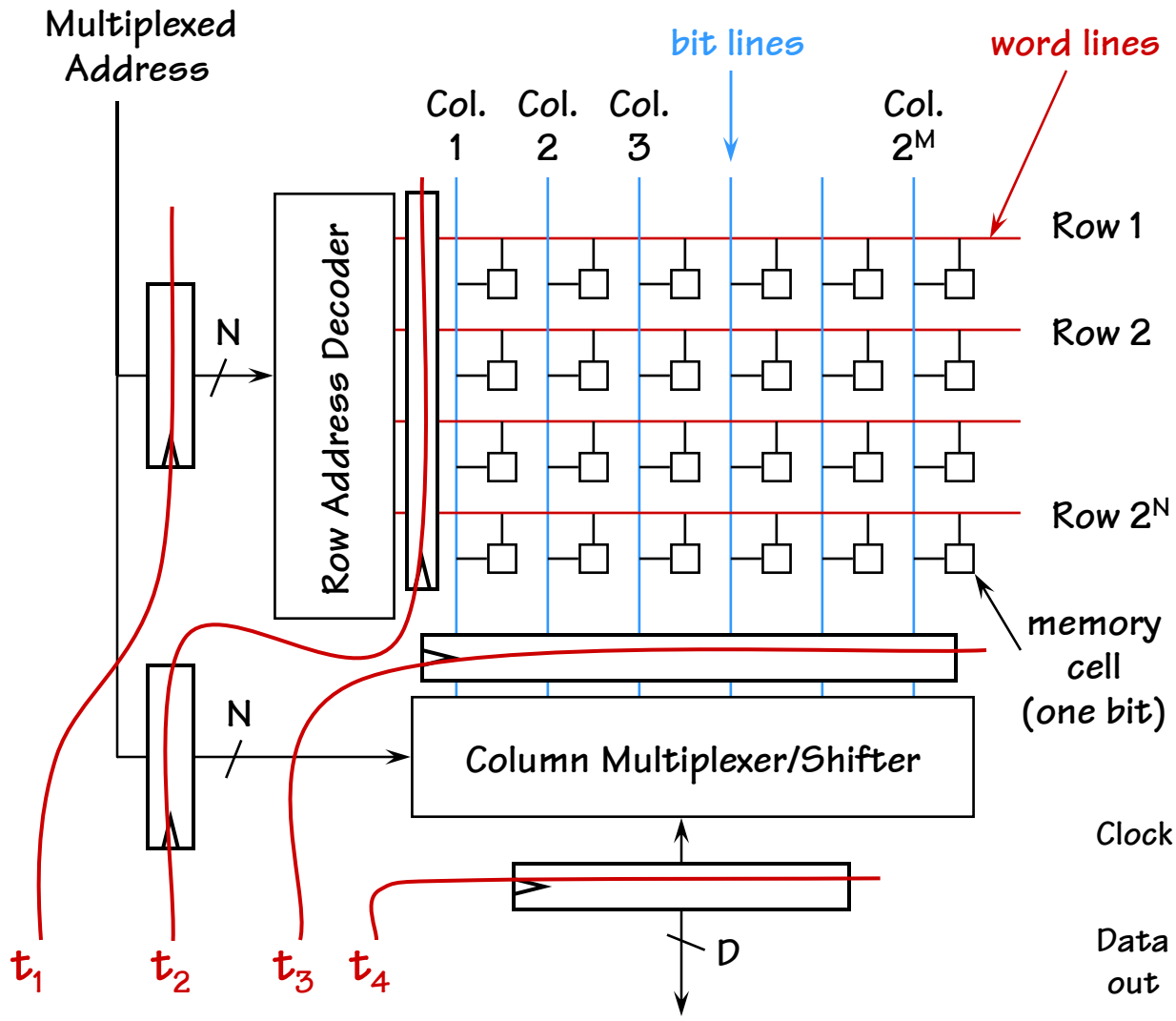
What Do We Want in a Memory?



	Capacity	Latency	Cost
Register	1000's of bits	10 ps	$$$$
SRAM	100's Kbytes	0.2 ns	$$$
DRAM	100's Mbytes	5 ns	$
Hard disk*	100's Gbytes	10 ms	¢
Want?	1 Gbyte	0.2 ns	cheap

* non-volatile

Tricks for Increasing Throughput

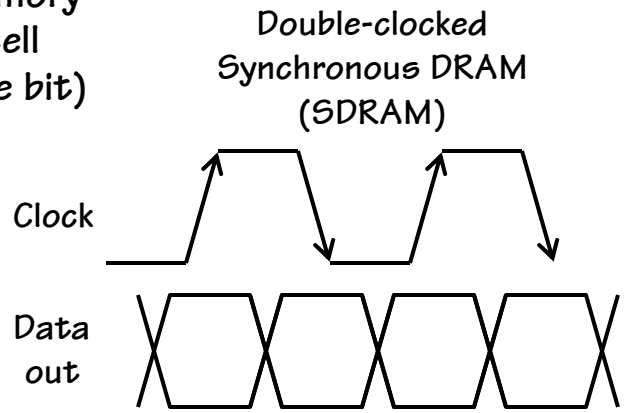


The first thing that should pop into you mind when asked to speed up a digital design…

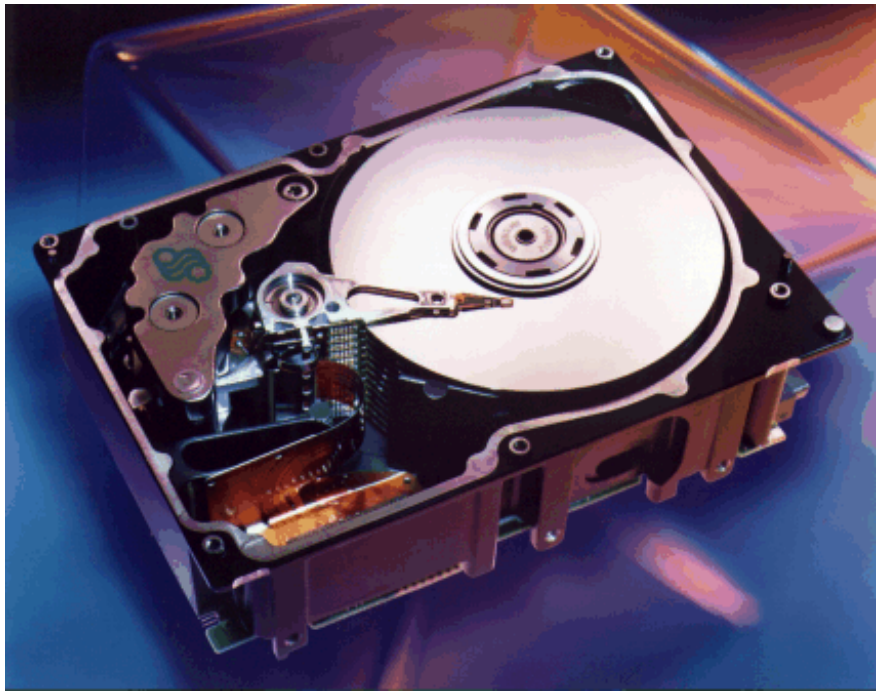
PIPELINING

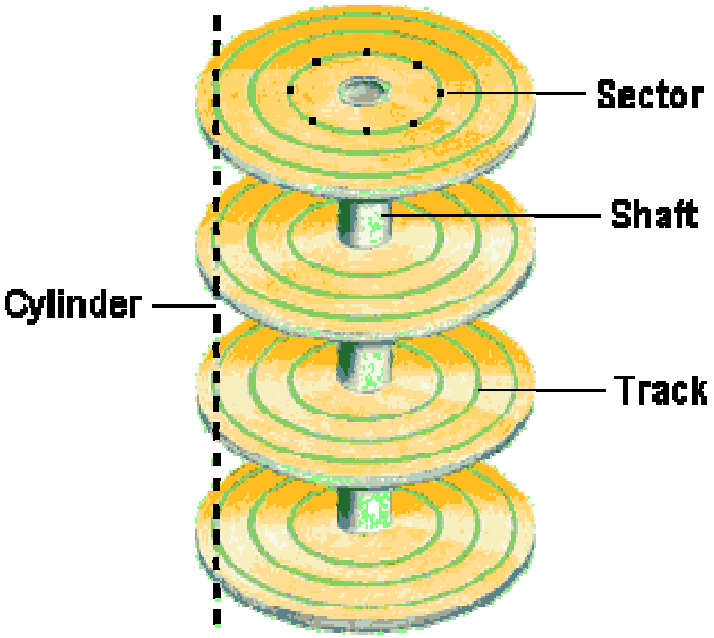
Synchronous DRAM (SDRAM)

($100 per Gbyte)



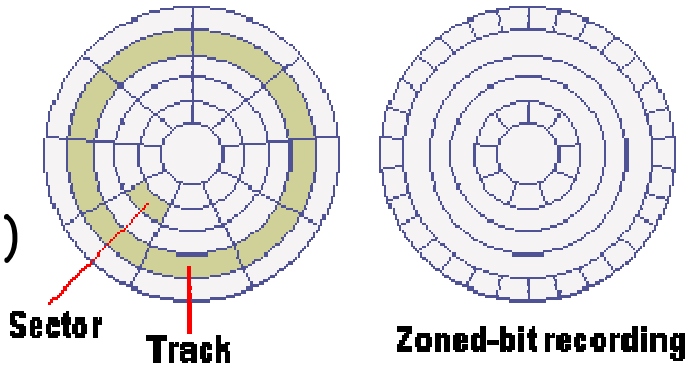
Hard Disk Drives





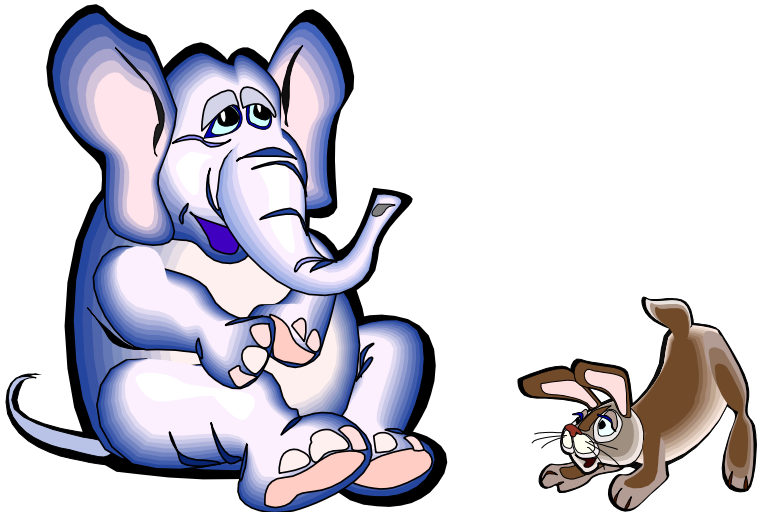
Typical high-end drive:

- Average latency = 4 ms (7200 rpm)
- Average seek time = 8.5 ms
- Transfer rate = 300 Mbytes/s (SATA)
- Capacity = 250 G byte
- Cost = $85 (33¢ Gbyte)



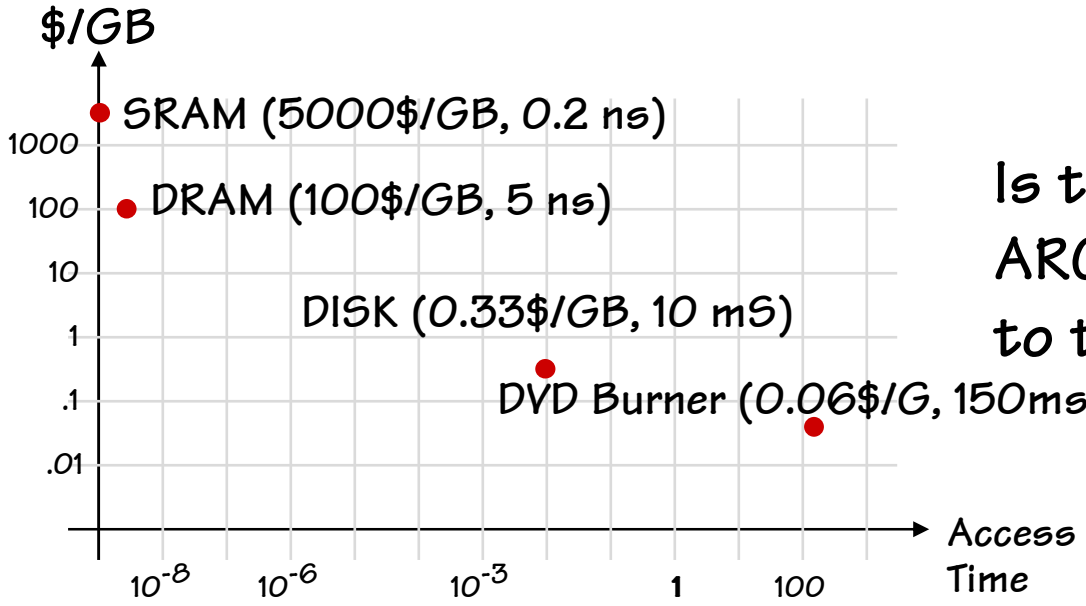
figures from www.pctechguide.com

Quantity vs Quality…



Your memory system can be
- BIG and SLOW… or
- SMALL and FAST.

We've explored a range of device-design trade-offs.



Is there an ARCHITECTURAL solution to this DELIMA?

Managing Memory via Programming

- In reality, systems are built with a mixture of all these various memory types

- How do we make the most effective use of each memory?
- We could push all of these issues off to programmers
  - Keep most frequently used variables and stack in SRAM
  - Keep large data structures (arrays, lists, etc) in DRAM
  - Keep bigger data structures on disk (databases) on DISK
- It is harder than you think… data usage evolves over a program's execution

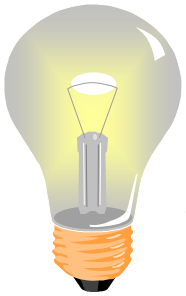
Best of Both Worlds

What we REALLY want: A BIG, FAST memory!
(Keep everything within instant access)

We'd like to have a memory system that
- PERFORMS like 2 GBytes of SRAM; but
- COSTS like 512 MBytes of slow memory.

SURPRISE: We can (nearly) get our wish!

KEY: Use a hierarchy of memory technologies:



Key IDEA

- Keep the most often-used data in a small, fast SRAM (often local to CPU chip)
- Refer to Main Memory only rarely, for remaining data.

The reason this strategy works: LOCALITY

Locality of Reference:

Reference to location X at time t implies that reference to location $X+\Delta X$ at time $t+\Delta t$ becomes more probable as ΔX and Δt approach zero.

Cache

cache (kash)

n.

A hiding place used especially for storing provisions.

A place for concealment and safekeeping, as of valuables.

The store of goods or valuables concealed in a hiding place.

Computer Science. A fast storage buffer in the central processing unit of a computer. In this sense, also called cache memory.

v. tr. cached, cach·ing, cach·es.

To hide or store in a cache.

Cache Analogy

You are writing a term paper at a table in the library

As you work you realize you need a book

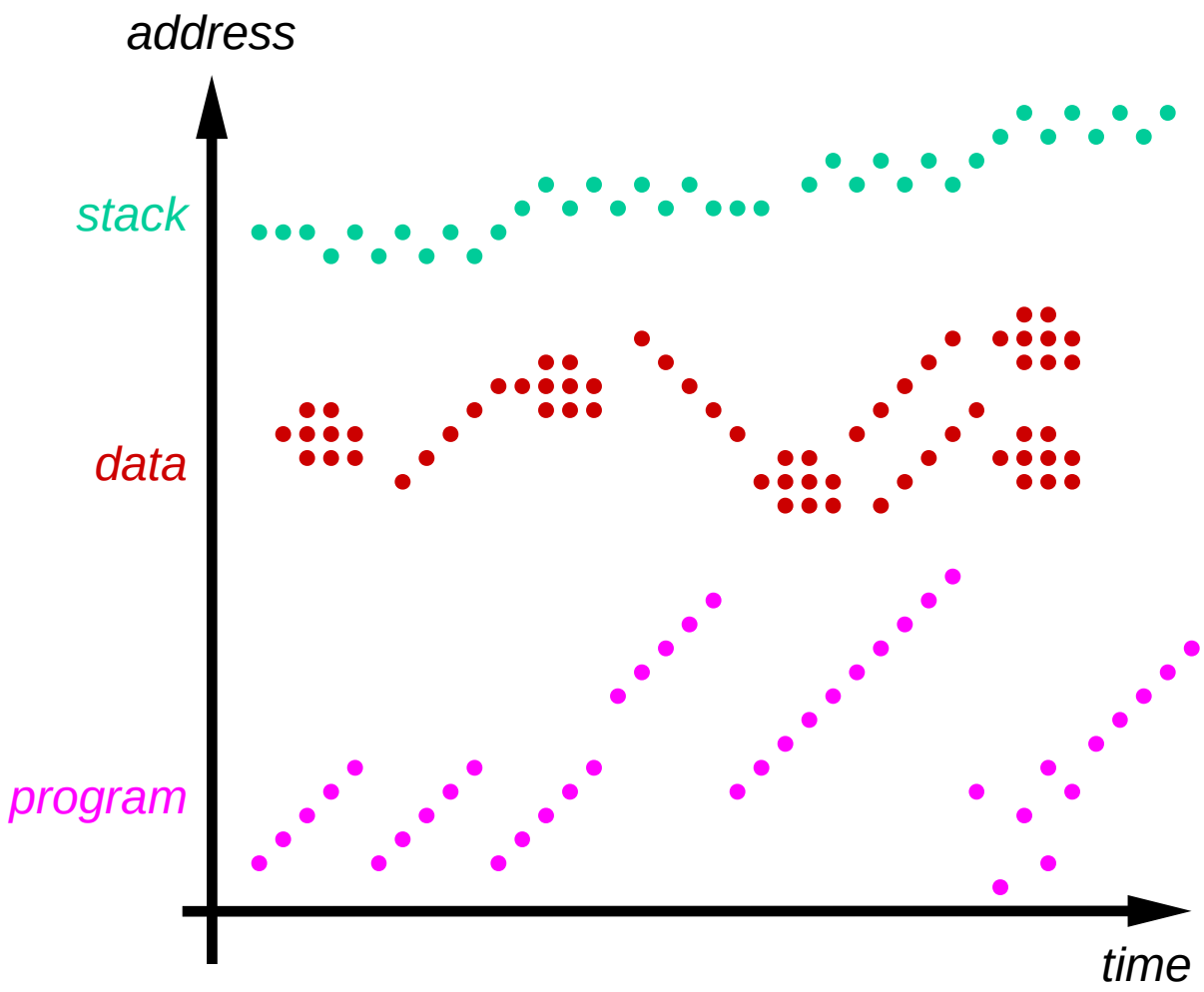
You stop writing, fetch the reference, continue writing

You don't immediately return the book, maybe you'll need it again

Soon you have a few books at your table and no longer have to fetch more books

The table is a CACHE for the rest of the library

Typical Memory Reference Patterns

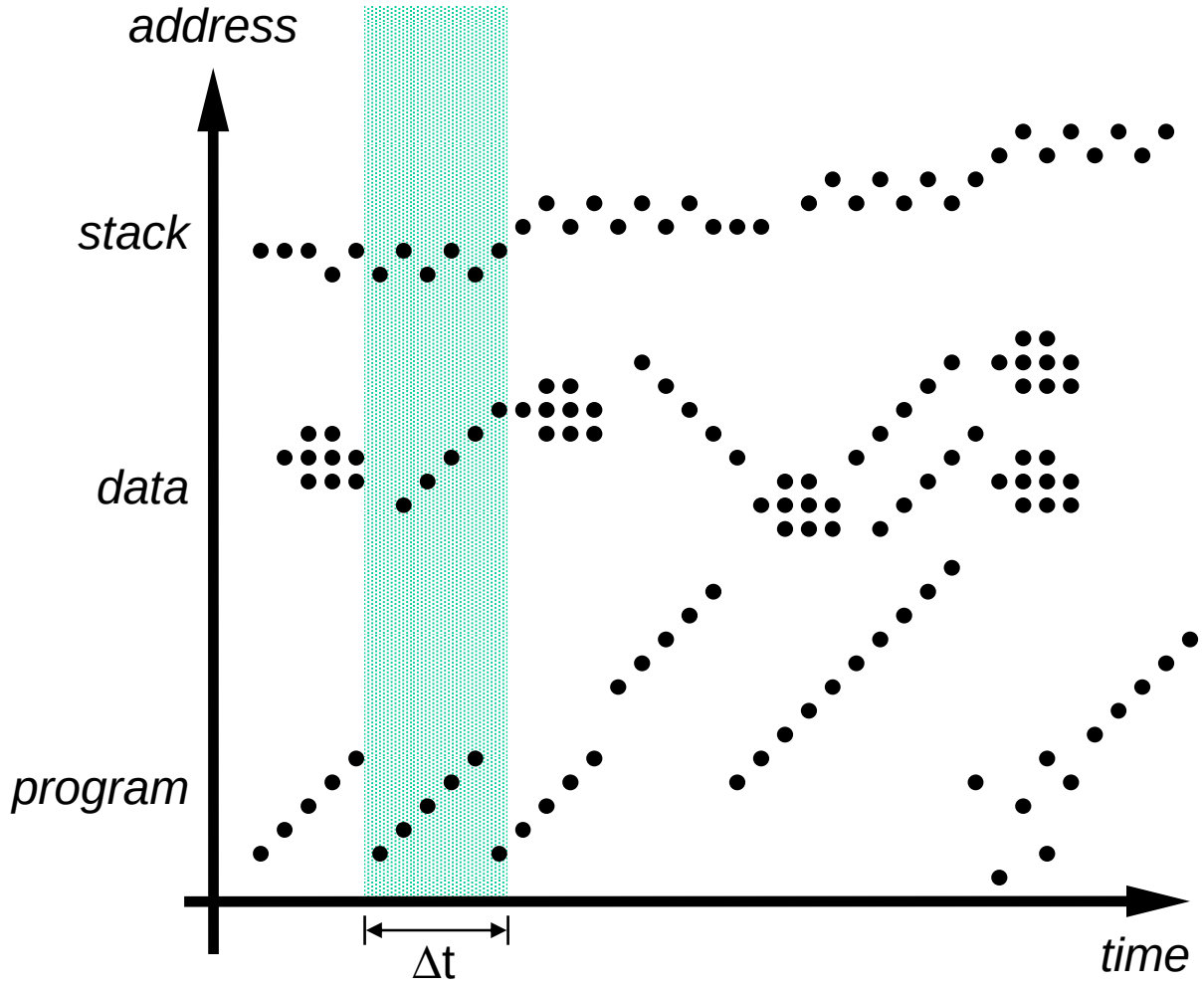


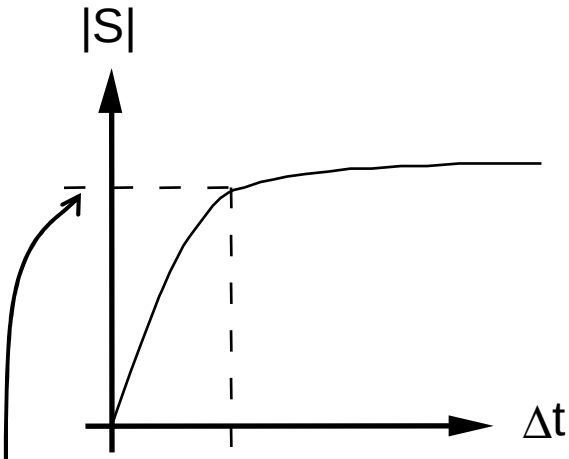
MEMORY TRACE –
A temporal sequence of memory references (addresses) from a real program.

TEMPORAL LOCALITY –
If an item is referenced, it will tend to be referenced again soon

SPATIAL LOCALITY –
If an item is referenced, nearby items will tend to be referenced soon.

Working Set





S is the set of locations accessed during Δt.

Working set: a set S which changes slowly w.r.t. access time.

Working set size, |S|

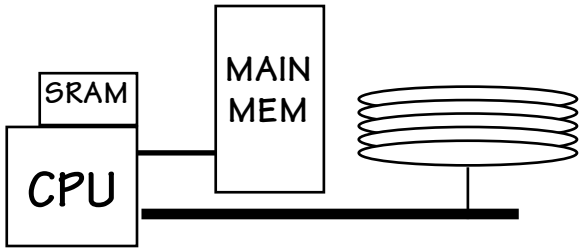
Exploiting the Memory Hierarchy

Approach 1 (Cray, others): Expose Hierarchy

- Registers, Main Memory,

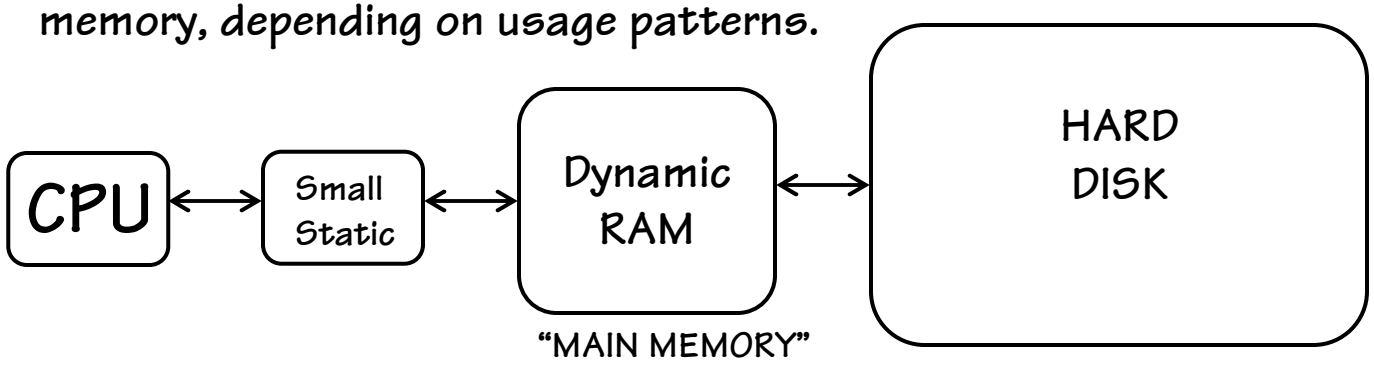
  Disk each available as storage alternatives;

- Tell programmers: "Use them cleverly"



Approach 2: Hide Hierarchy

- Programming model: SINGLE kind of memory, single address space.
- Machine AUTOMATICALLY assigns locations to fast or slow memory, depending on usage patterns.

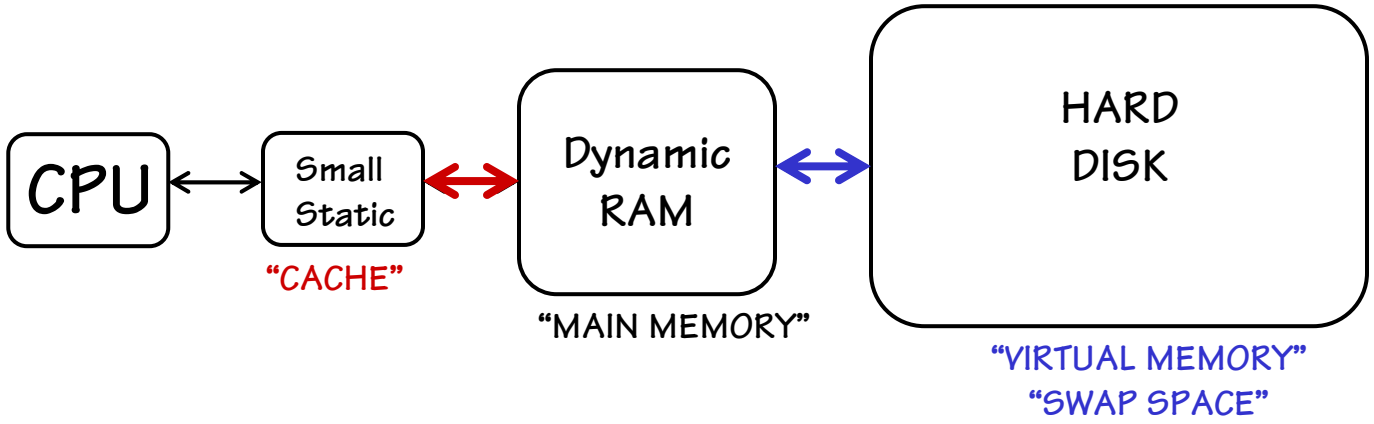


Why We Care

CPU performance is dominated by memory performance.

More significant than:

ISA, circuit optimization, pipelining, super-scalar, etc



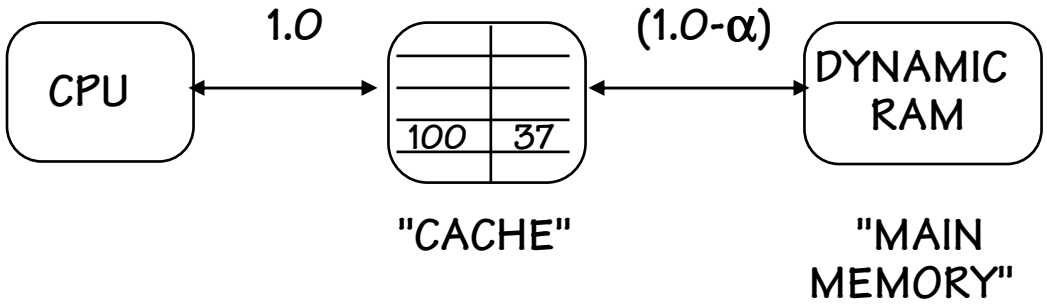
TRICK #1: How to make slow MAIN MEMORY appear faster than it is.

Technique: CACHEING – Next 2 Lectures

TRICK #2: How to make a small MAIN MEMORY appear bigger than it is.

Technique: VIRTUAL MEMORY – Lecture after that

The Cache Idea:
Program-Transparent Memory Hierarchy



Cache contains TEMPORARY COPIES of selected main memory locations... eg. Mem[100] = 37

GOALS:

1) Improve the average access time

α HIT RATIO: Fraction of refs found in CACHE.

(1-α) MISS RATIO: Remaining references.

$$t_{ave} = \alpha t_c + (1-\alpha)(t_c + t_m) = t_c + (1-\alpha)t_m$$

Challenge: To make the hit ratio as high as possible.

2) Transparency (compatibility, programming ease)

How High of a Hit Ratio?

Suppose we can easily build an on-chip static memory with a 0.8 nS access time, but the fastest dynamic memories that we can buy for main memory have an average access time of 10 nS. How high of a hit rate do we need to sustain an average access time of 1 nS?

$$\alpha = 1 - \frac{t_{ave} - t_c}{t_m} = 1 - \frac{1 - 0.8}{10} = 98\%$$

WOW, a cache really needs to be good?

Cache

Sits between CPU and main memory

Very fast table that stores a *TAG* and *DATA*

TAG is the memory address

DATA is a copy of memory at the address given by TAG

Tag	Data
1000	17
1040	1
1032	97
1008	11

Memory

1000	17
1004	23
1008	11
1012	5
1016	29
1020	38
1024	44
1028	99
1032	97
1036	25
1040	1
1044	4

Cache Access

On load we look in the TAG entries for the address we're loading

Found à a *HIT*, return the DATA

Not Found à a *MISS*, go to memory for the data and put it and the address (TAG) in the cache

Tag	Data
1000	17
1040	1
1032	97
1008	11

Memory

1000	17
1004	23
1008	11
1012	5
1016	29
1020	38
1024	44
1028	99
1032	97
1036	25
1040	1
1044	4

Cache Lines

Usually get more data than requested (Why?)

a *LINE* is the unit of memory stored in the cache

usually much bigger than 1 word, 32 bytes per line is common

bigger LINE means fewer misses because of spatial locality

but bigger LINE means longer time on miss

Tag	Data	
1000	17	23
1040	1	4
1032	97	25
1008	11	5

Memory

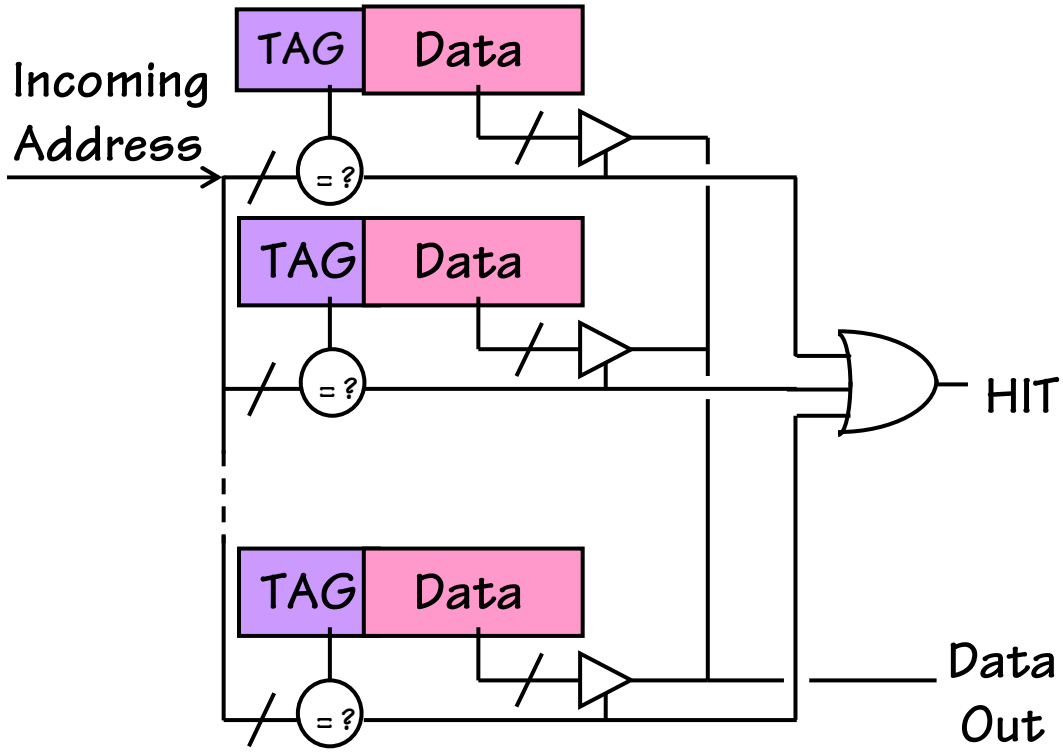
1000	17
1004	23
1008	11
1012	5
1016	29
1020	38
1024	44
1028	99
1032	97
1036	25
1040	1
1044	4

Finding the TAG in the Cache

A 1MByte cache may have 32k different lines each of 32 bytes

We can't afford to sequentially search the 32k different tags

ASSOCIATIVE memory uses hardware to compare the address to the tags in parallel but it is expensive and 1MByte is thus unlikely



Finding the TAG in the Cache

A 1MByte cache may have 32k different lines each of 32 bytes

We can't afford to sequentially search the 32k different tags

ASSOCIATIVE memory uses hardware to compare the address to the tags in parallel but it is expensive and 1MByte is thus unlikely

DIRECT MAPPED CACHE computes the cache entry from the address

- multiple addresses map to the same cache line
- use TAG to determine if right

Choose some bits from the address to determine the Cache line

- low 5 bits determine which byte within the line
- we need 15 bits to determine which of the 32k different lines has the data
- which of the 32 – 5 = 27 remaining bits should we use?

Direct-Mapping Example

- With 8 byte lines, the bottom 3 bits determine the byte within the line
- With 4 cache lines, the next 2 bits determine which line to use

1024d = 10000**00**000b à line = 00b = 0d

1000d = 01111**01**000b à line = 01b = 1d

1040d = 10000**10**000b à line = 10b = 2d

Tag	Data	
1024	44	99
1000	17	23
1040	1	4
1016	29	38

Memory

1000	17
1004	23
1008	11
1012	5
1016	29
1020	38
1024	44
1028	99
1032	97
1036	25
1040	1
1044	4

Direct Mapping Miss

•What happens when we now ask for address 1008?

1008d = 0111111**10**000b à line = 10b = 2d

but earlier we put 1040d there...

1040d = 1000001**10**000b à line = 10b = 2d

Tag	Data	
1024	44	99
1000	17	23
1008	11	5
1016	29	38

Memory

1000	17
1004	23
1008	11
1012	5
1016	29
1020	38
1024	44
1028	99
1032	97
1036	25
1040	1
1044	4

Miss Penalty and Rate

The *MISS PENALTY* is the time it takes to read the memory if it isn't in the cache

50 to 100 cycles is common.

The MISS RATE is the fraction of accesses which MISS

The HIT RATE is the fraction of accesses which HIT

MISS RATE + HIT RATE = 1

Suppose a particular cache has a MISS PENALTY of 100 cycles and a HIT RATE of 95%. The CPI for load on HIT is 5 but on a MISS it is 105. What is the average CPI for load?

Average CPI = 10 5 * 0.95 + 105 * 0.05 = 10

Suppose MISS PENALTY = 120 cycles?

then CPI = 11 (slower memory doesn't hurt much)

Some Associativity can help

Direct-Mapped caches are very common but can cause problems...

SET ASSOCIATIVITY can help.

Multiple Direct-mapped caches, then compare multiple TAGS

- 2-way set associative = 2 direct mapped + 2 TAG comparisons
- 4-way set associative = 4 direct mapped + 4 TAG comparisons

Now array size == power of 2 doesn't get us in trouble

But

- slower
- less memory in same area
- maybe direct mapped wins...

What about store?

What happens in the cache on a store?

- WRITE BACK CACHE à put it in the cache, write on replacement
- WRITE THROUGH CACHE à put in cache and in memory

What happens on store and a MISS?

- WRITE BACK will fetch the line into cache
- WRITE THROUGH might just put it in memory

Cache Questions = Cash Questions

What lies between Fully Associate and Direct-Mapped?

When I put something new into the cache, what data gets thrown out?

How many processor words should there be per tag?

When I write to cache, should I also write to memory?

What do I do when a write misses cache, should space in cache be allocated for the written address.

What if I have INPUT/OUTPUT devices located at certain memory addresses, do we cache them?

Answers: Stay Tuned

5 to go!