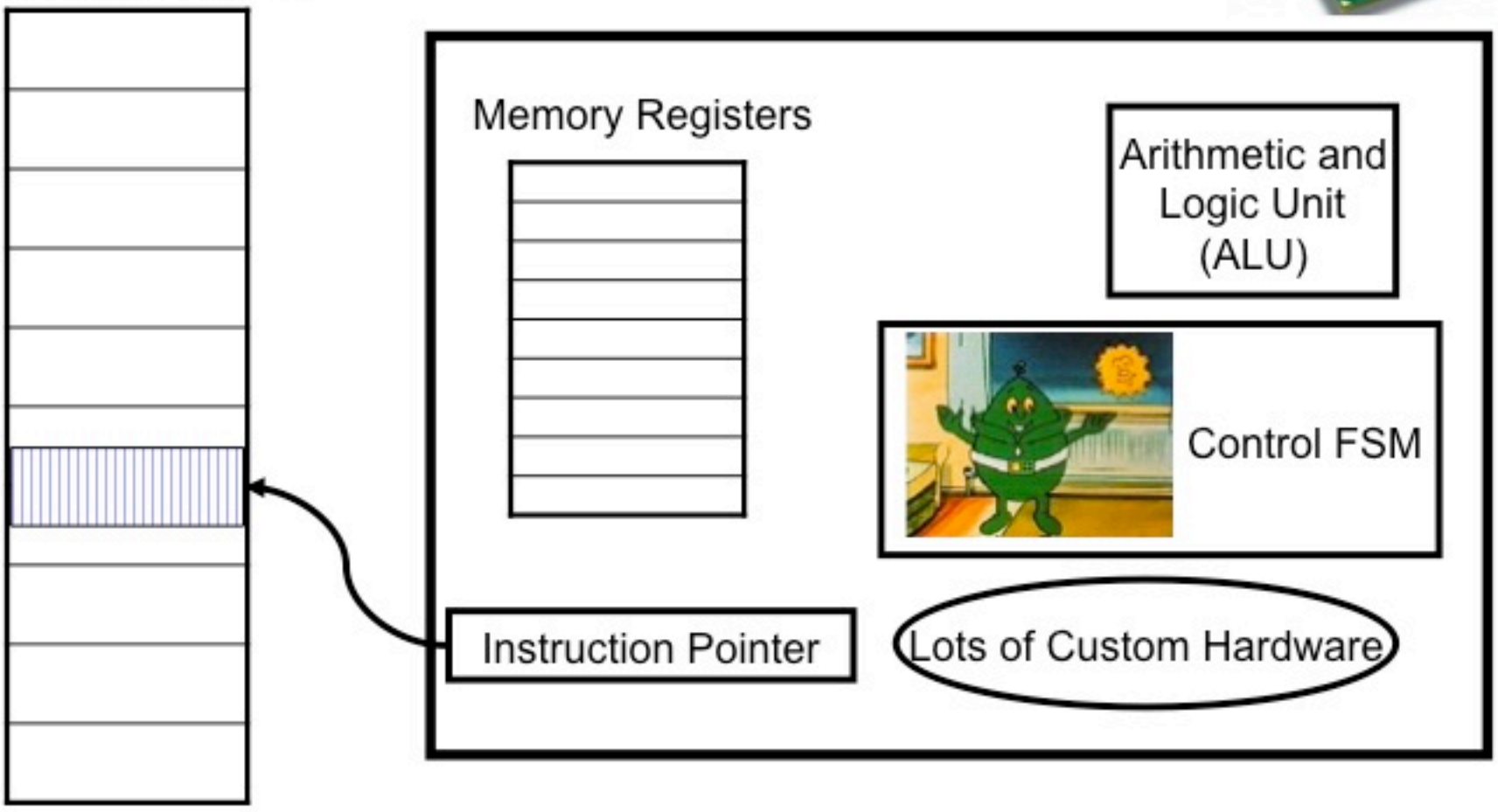
How to streamline your life (lessons from computer architecture).

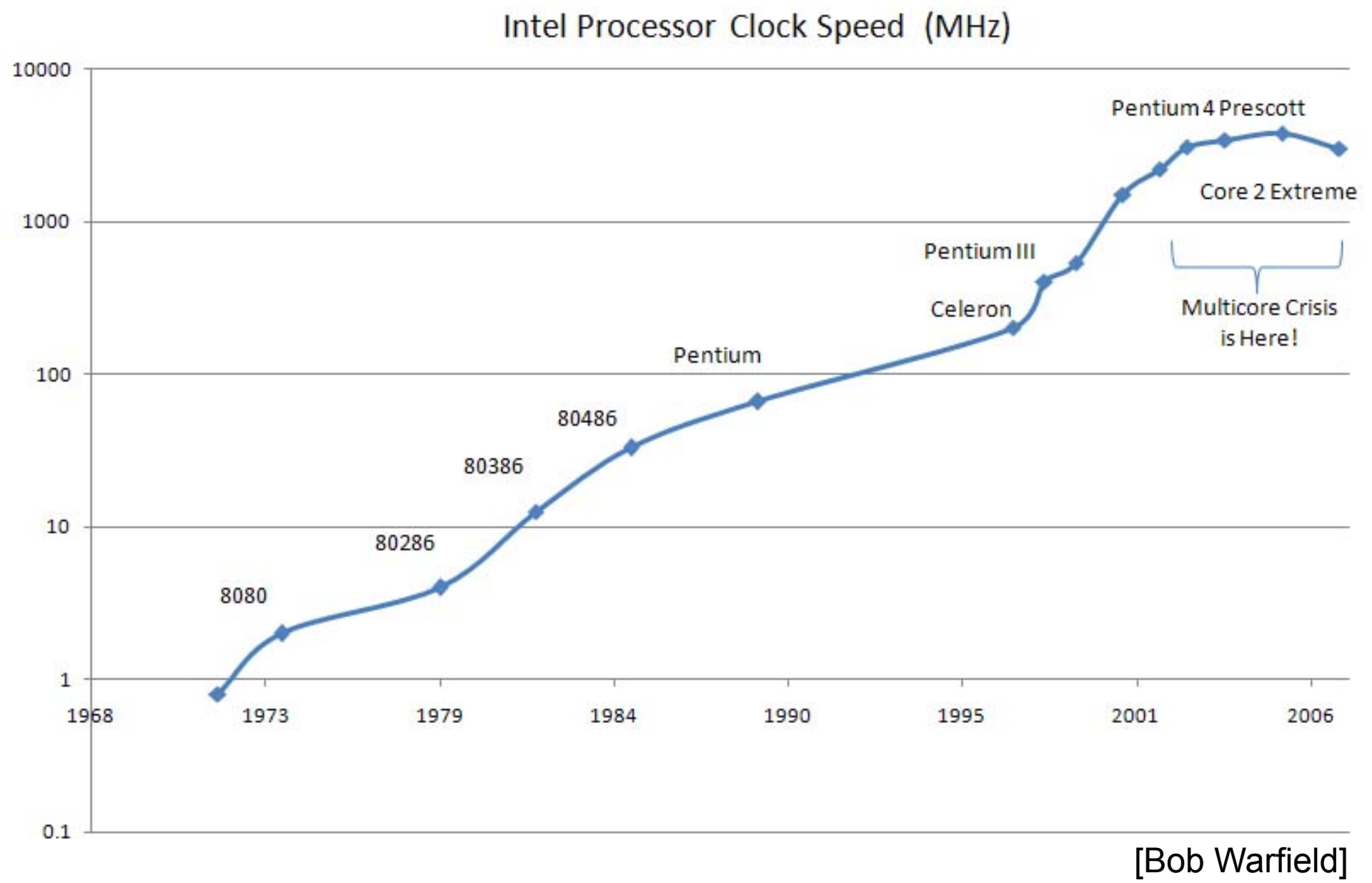
COS 116, Spring 2012

Adam Finkelstein

Greatly simplified view of modern CPUs.

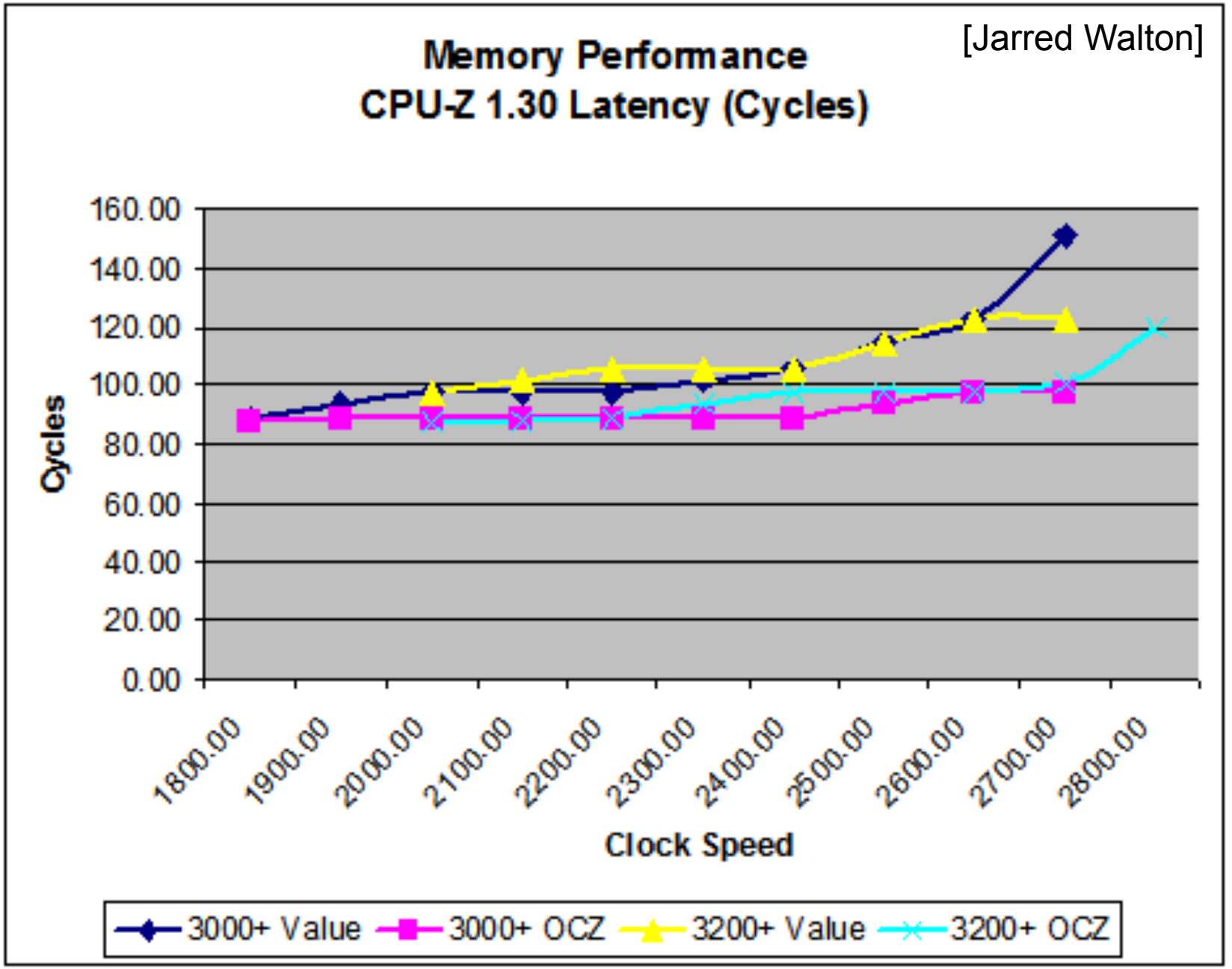
Program (in binary) stored in memory





[Bob Warfield]

[Jarred Walton]

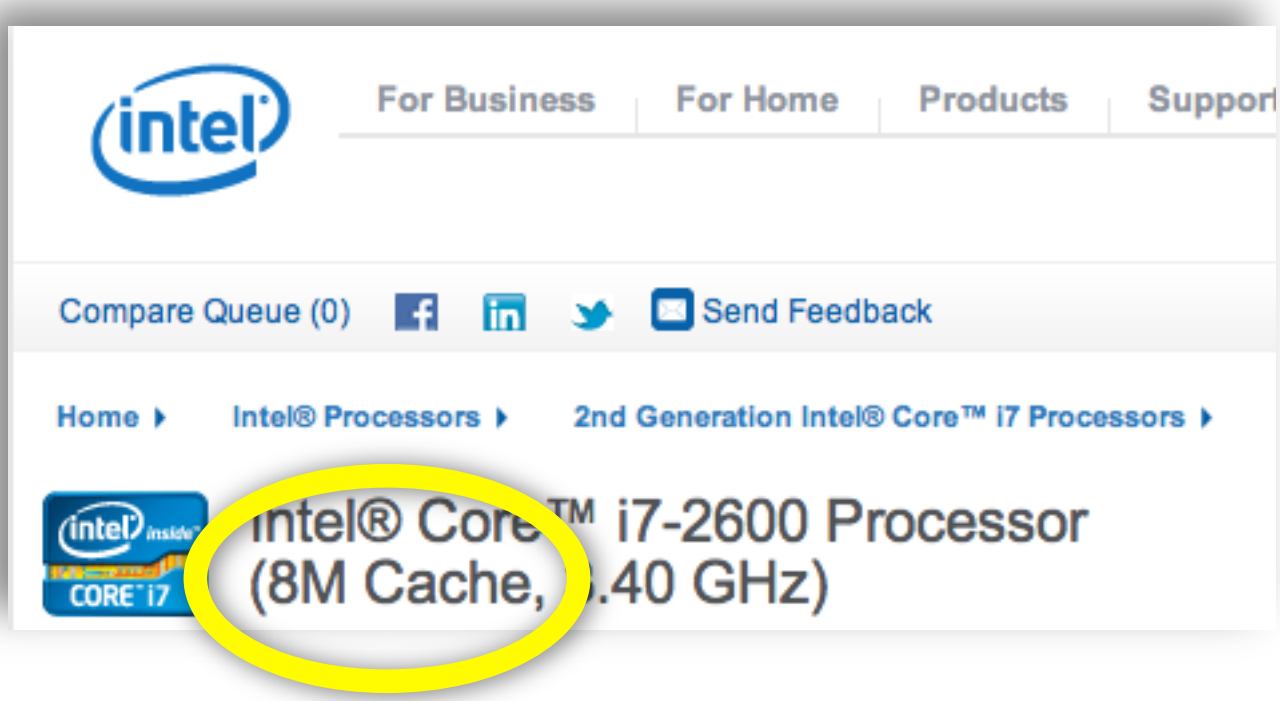


Problem:

Retrieval from memory is relatively slow.

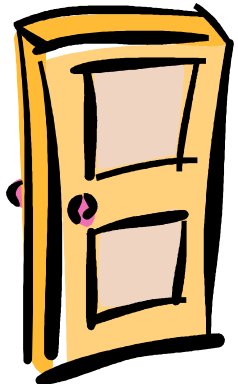
Solution:

"Cache"



the tired librarian

reserves



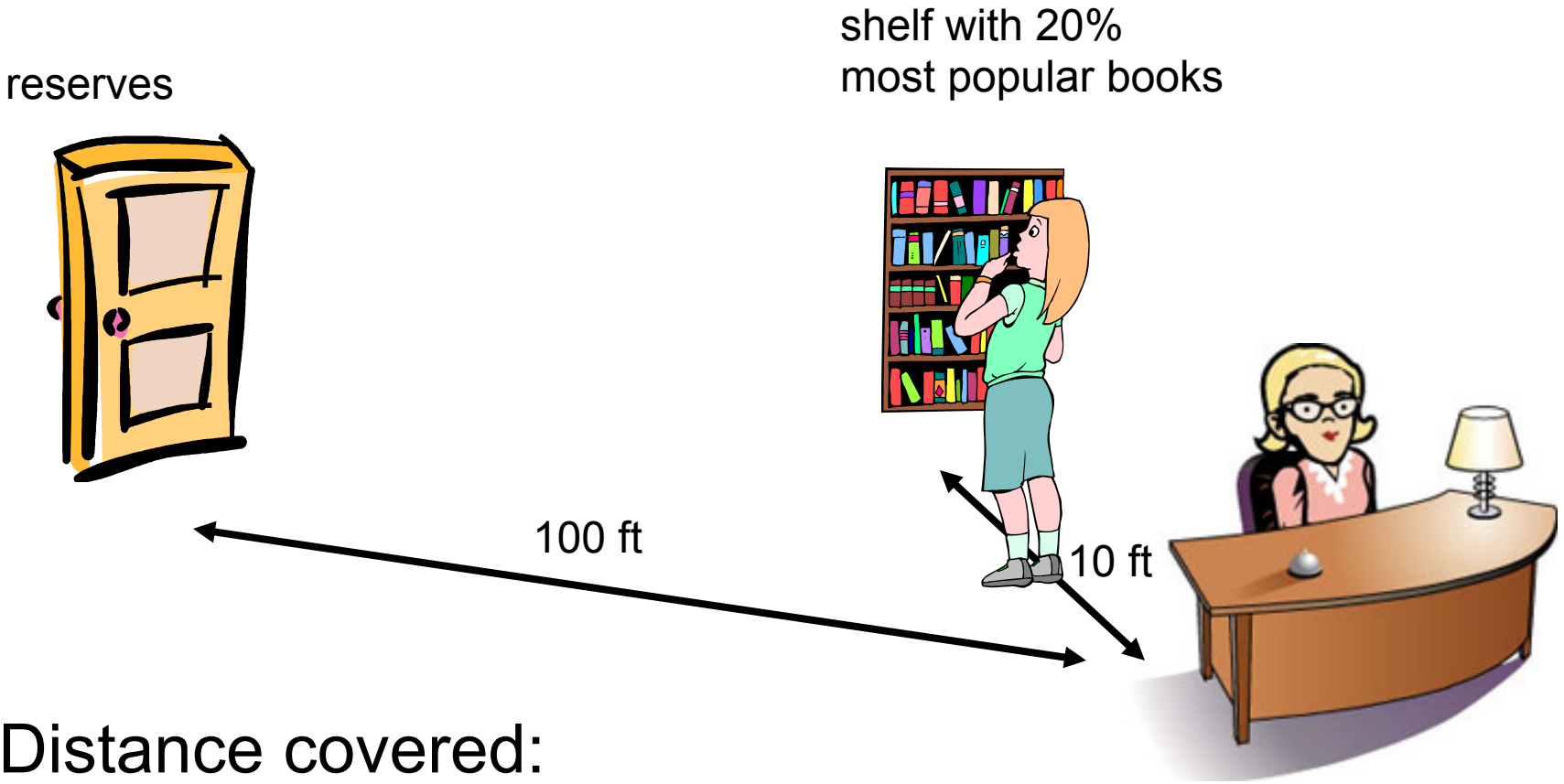
100 ft

- 1000 checkouts or returns per day
- Distance covered:
  100 x 1000 = 100,000 feet ~ 20 miles

“80-20 rule”

- “Pareto principle” Pareto [1906], Juran [1941]
- 80% of wealth held by 20% of the people
- 80% of work done by 20% of organization
- 80% of sales come from 20% of the clients
- 80% of computer crashes from 20% of bugs
- 80% of librarian work comes from 20% of books
- 80% of fetches are for 20% of computer memory

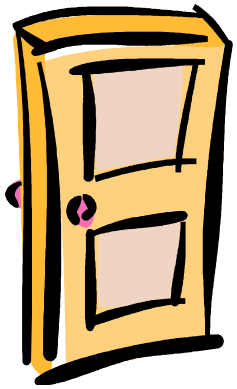
better arrangement



- Distance covered:
  20% (100 x 1000) + 80% (10 x 1000)
  = 28,000 feet ~ 5 miles

even better arrangement

reserves

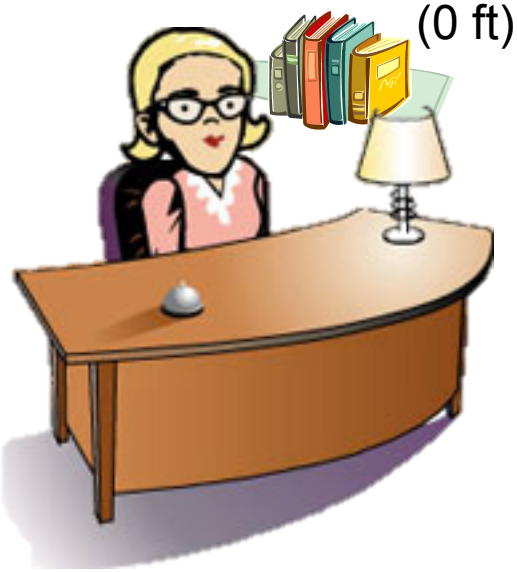


shelf with 20%
most popular books

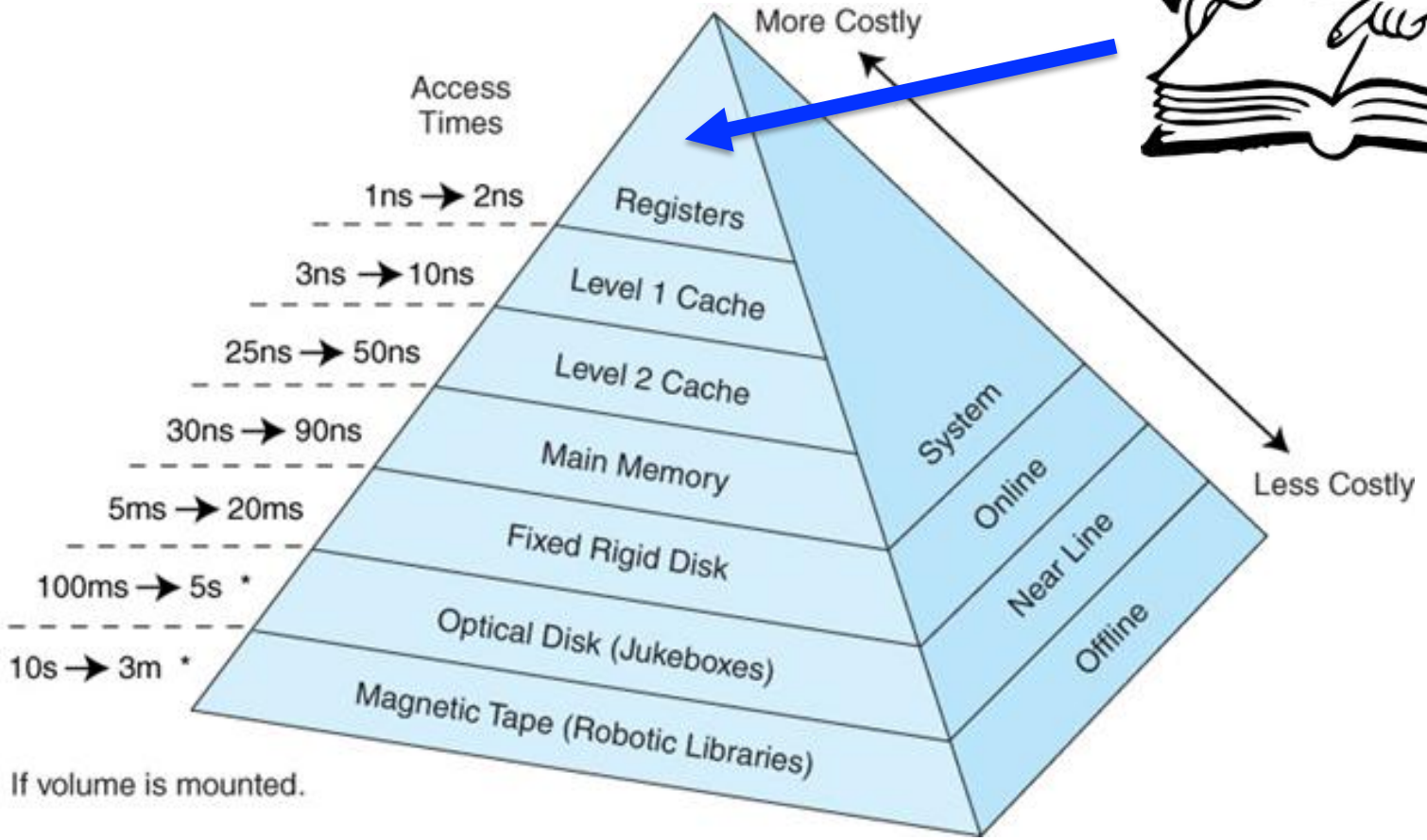
top 4%
(0 ft)

100 ft

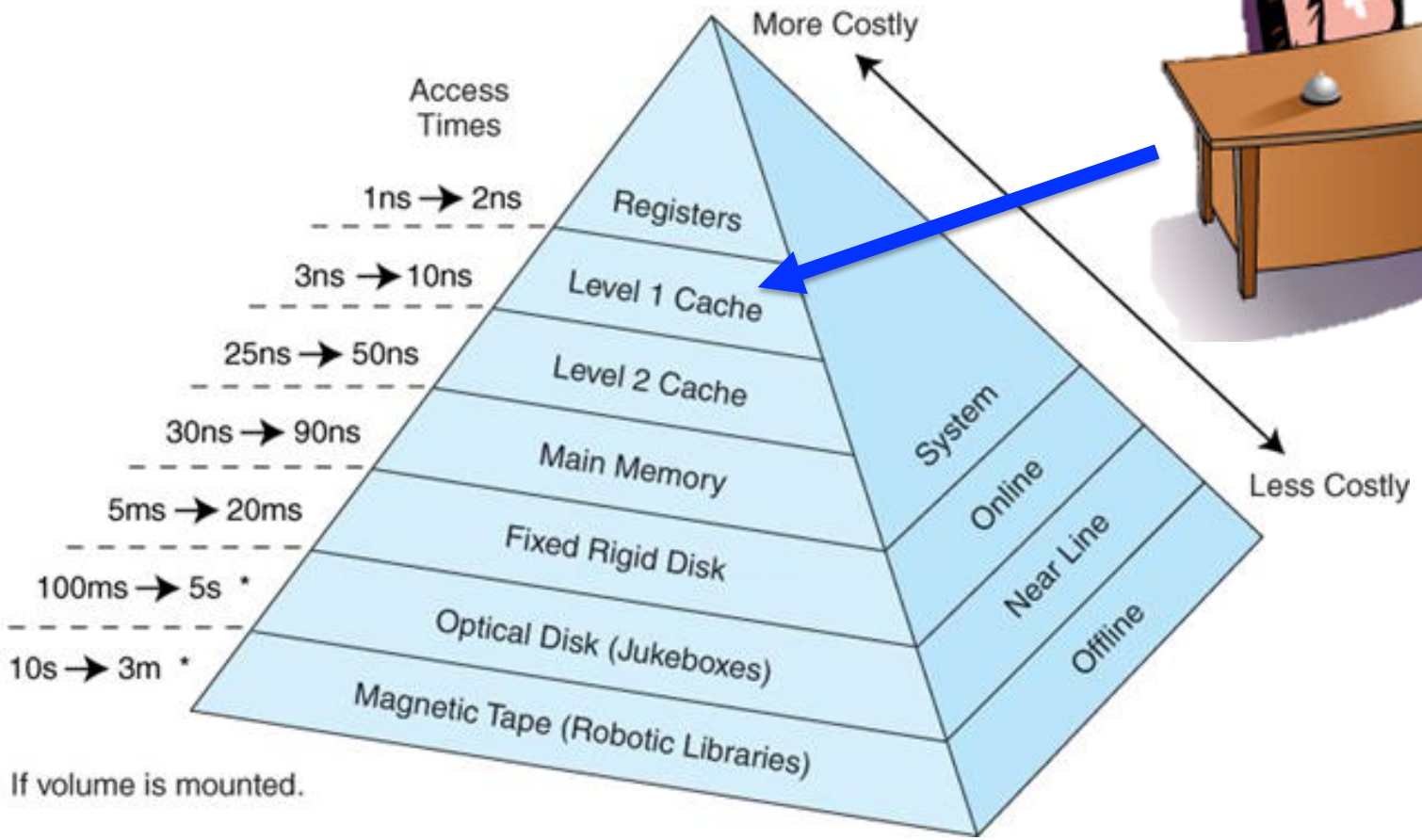
10 ft



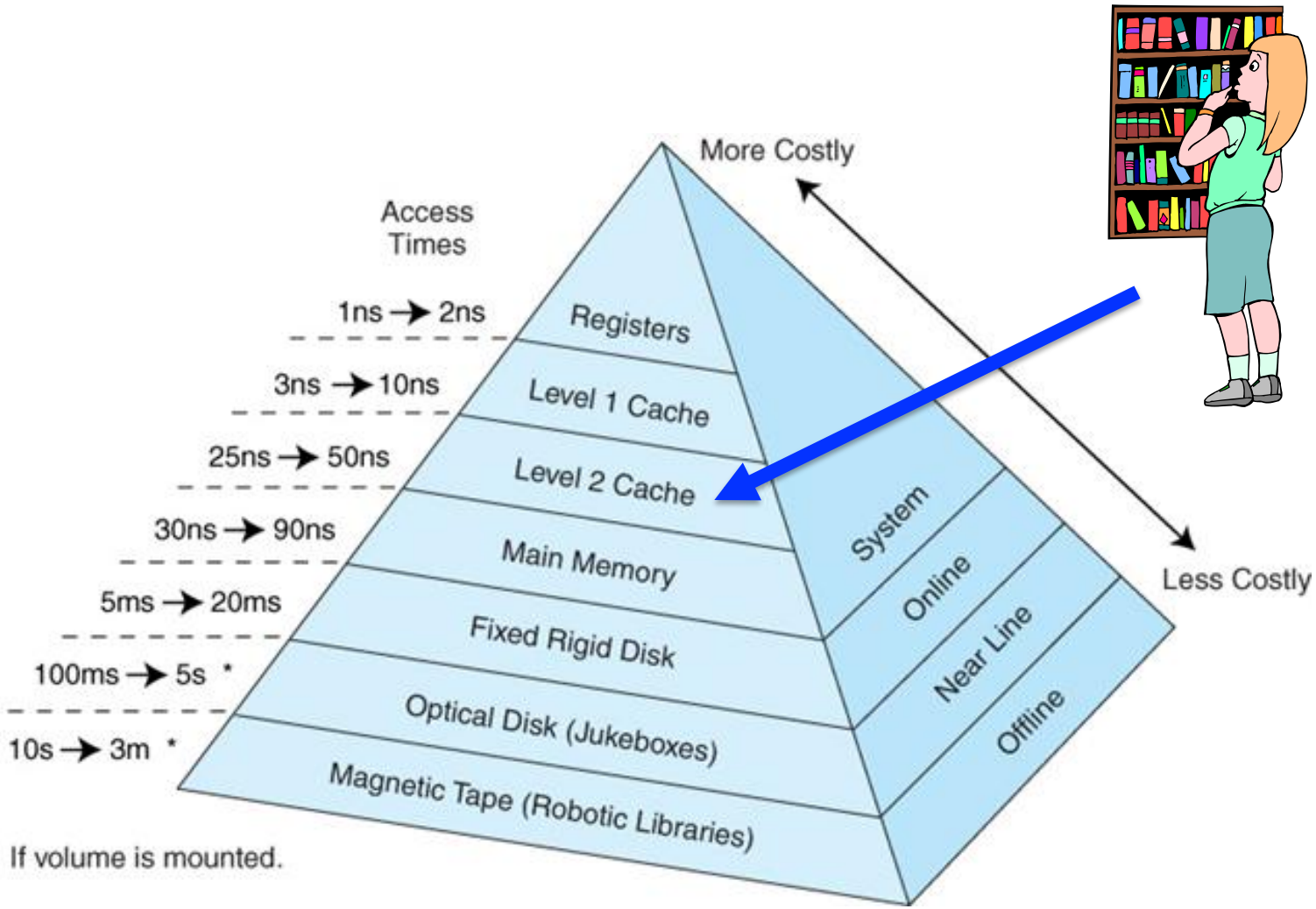
memory cache hierarchy



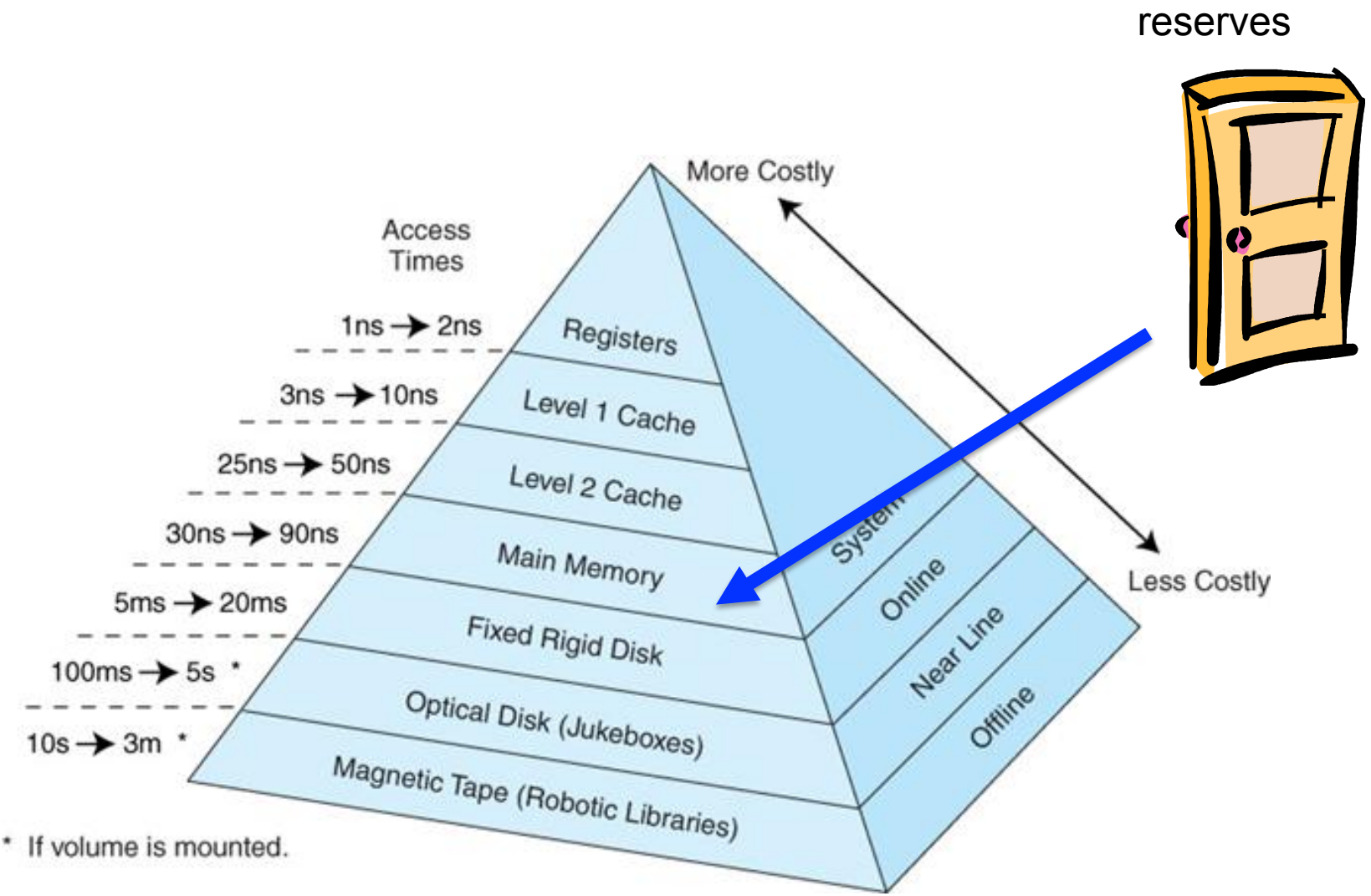
memory cache hierarchy



memory cache hierarchy



memory cache hierarchy



Why this Organization?

- Speed vs cost of various memories

  (as of a few years ago)

	Cost: $ / GB	Speed: GB/s
Hard drive	0.10	0.1
Flash (e.g., SSD, USB stick)	2	0.25
RAM	10	10
On-chip memory for CPU (L2 Cache)	40000	20

Cache benefit

- Performance:
  - Speed is close to that of fastest memory (cache)
  - Overall capacity is that of largest memory (disk)

encyclopedia assignment

1000 questions like these:

- What is the capital city of Albania?
- Who was the fourth Roman Emperor?
- Who is the prime minister of India?
- What is the population of Argentina?
- Which team won the 1968 World Series?
- …etc.

encyclopedia assignment

Does cache help?

- Cache works ok.
- Needed volume is often at librarian's desk.
- Lucky when questions are in same volume.

Discussion Time

- Is the librarian's problem solved?

How to predict the 20% most popular books for next day?

- In general, no easy solution
- In practice, use rules of thumb
  - Example: “Least Recently Used”. When you need to create space on the desk (or shelf), move out the book that was used least recently
  - Many others (LRU is computationally expensive)

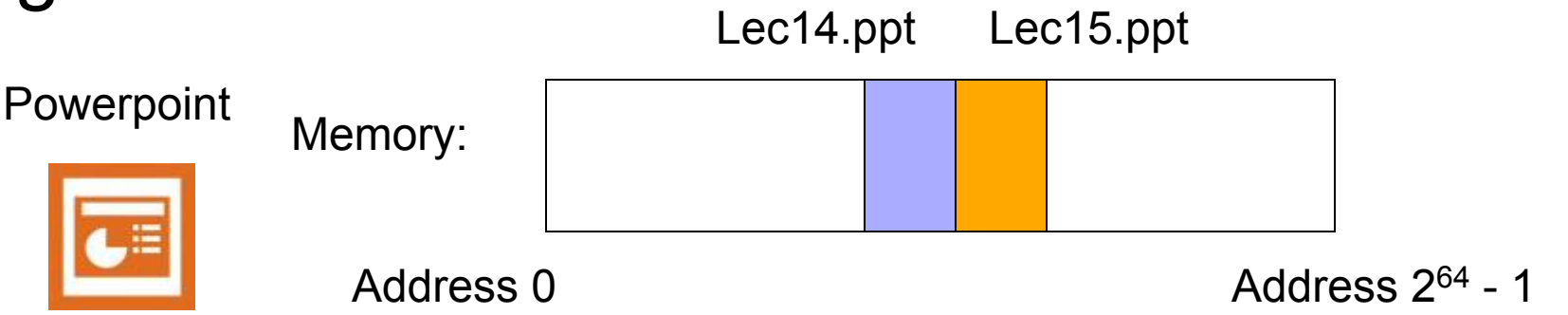
Question

- How does the same program (.exe file) run on different PCs with different memory configurations?

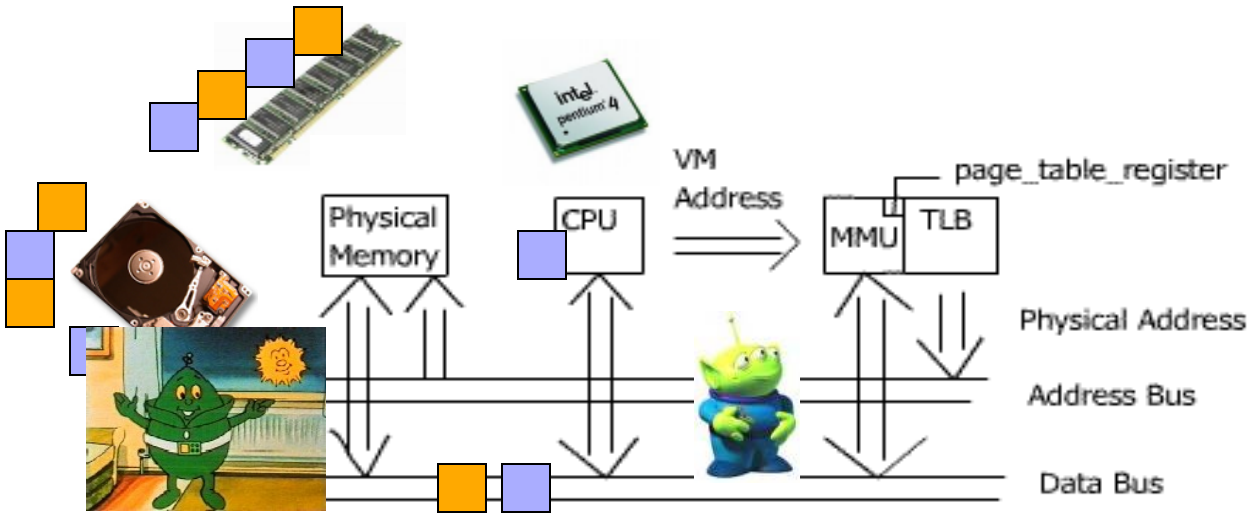
- Answer: “Virtual Memory”
  - All programs live a fiction: allowed to pretend they each have 2^{32} or 2^{64} bytes of memory
  - Illusion is preserved by hardware

Virtual Memory

- Program's view:



- Underneath:

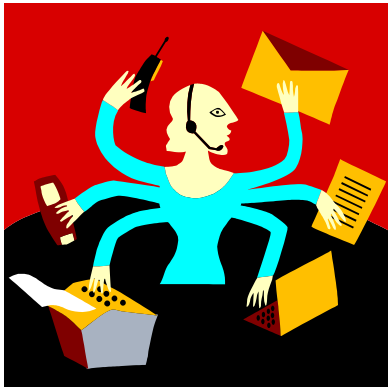


Lesson 2: Multitasking

- “The Multitasking Generation”

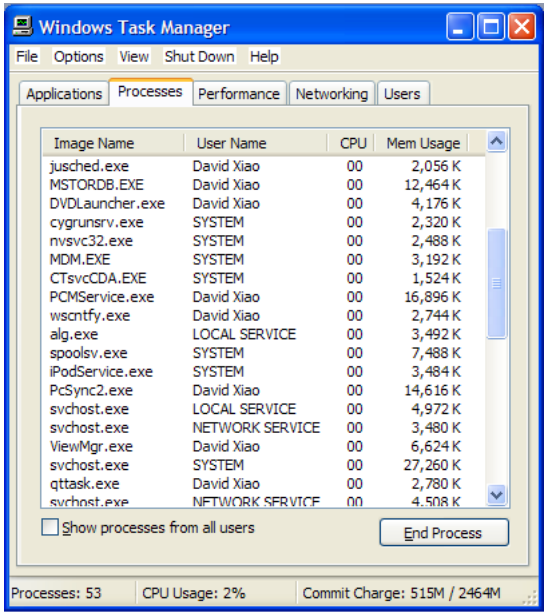
An Evening's Tasks for a Gen-M'er

- Homework
- Listen to music
- Instant Messaging
- Call Mom (goes to bed by 11 PM!)
- Answer phone
- Read a bit more of Joyce's *Ulysses*
- Watch the Daily Show



How do you do it all?!?

Tasks done by my PC last night



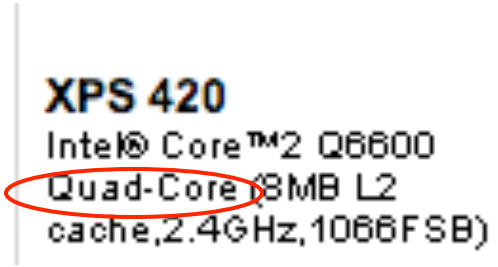
- Word processing
- Play CD
- Download news updates
- Download email
- Run clock
- Hidden tasks: handle network traffic, manage disk and RAM traffic, scheduler, etc.

Managed by “Operating System”
(Windows, Linux, MacOS, etc.)

Multitasking vs. Parallel Processing

Multitasking: A single CPU handles many tasks by switching rapidly among them.
(e.g., all Wintel machines since early 1990s; all Unix machines since the 1970s)

Parallel Processing: Multiple CPUs that do the work of a single CPU. (But, 4 CPUs do not necessarily mean 4x speed.)



Scheduler's objectives

- Fairness
- Timeliness
- Critical tasks processed promptly
- Low overhead

How can one achieve these (often conflicting) goals?