CSCI 210: Computer Architecture
Lecture 33: Caches

Stephen Checkoway
Oberlin College
May 13, 2022
Slides from Cynthia Taylor

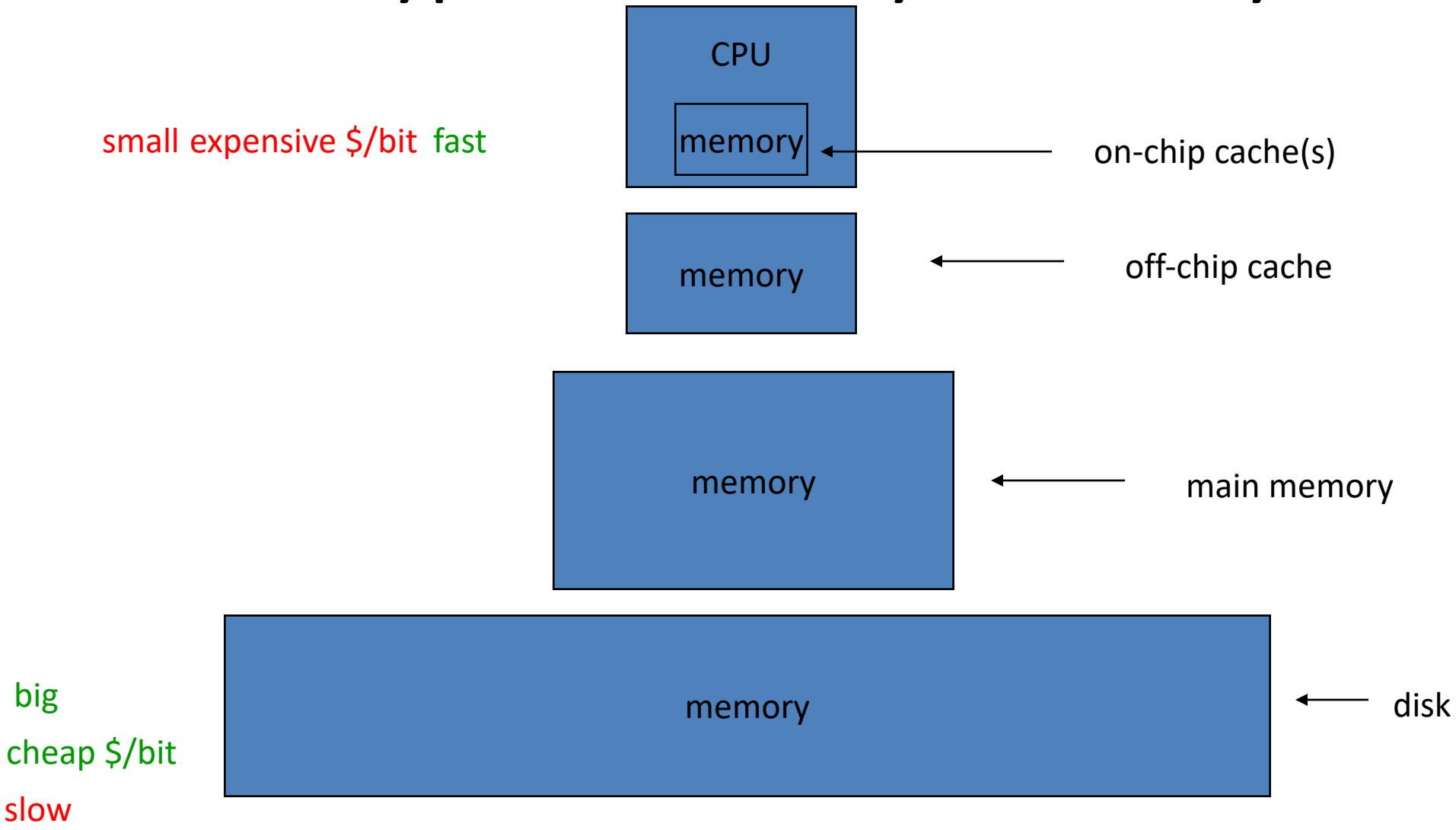
Announcements

- Problem Set 11 due in one week (it’ll be up tonight)
- Cache Lab (final project) due at the end of our scheduled final exam period
- Office Hours today 13:30 – 14:30

Memory

- So far we have only looked at the CPU/datapath
- Now we’re going to look at memory

A typical memory hierarchy



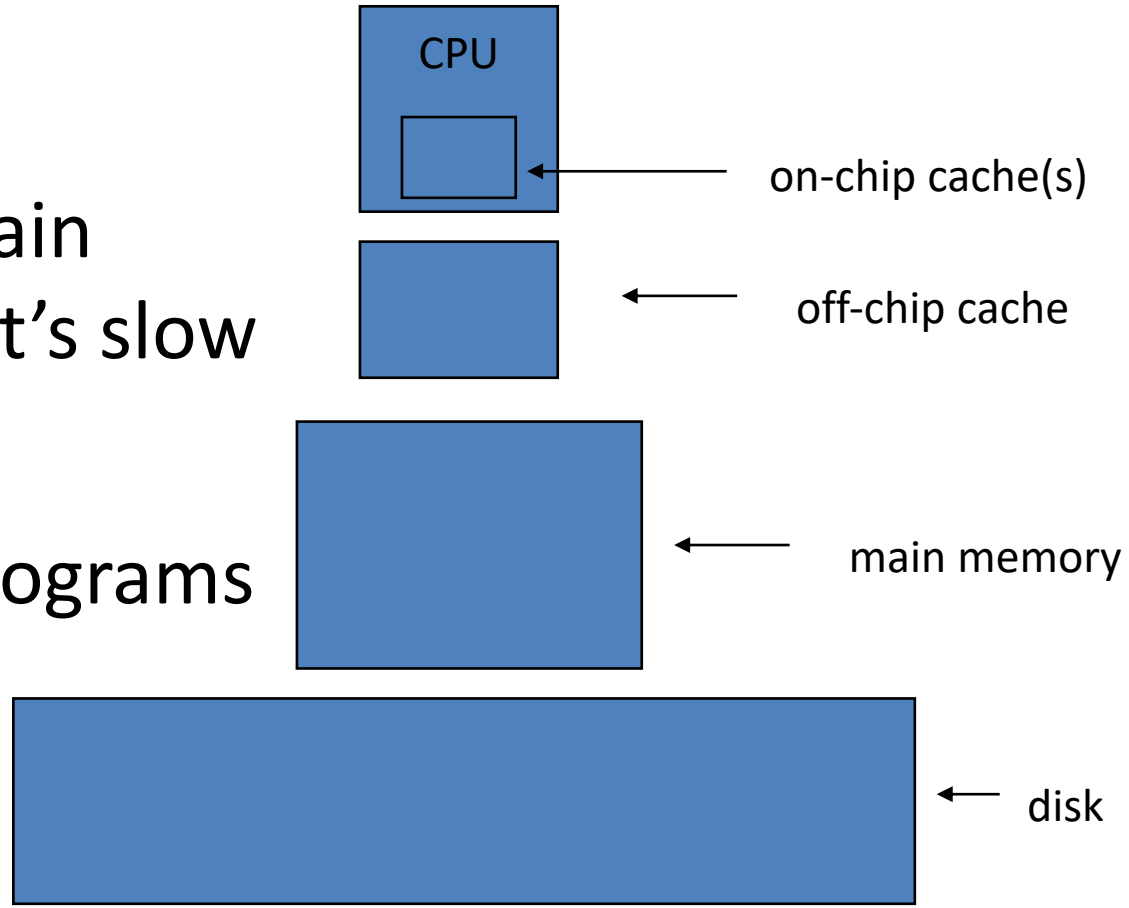
Latency

Table 2.2 Example Time Scale of System Latencies

Event	Latency	Scaled
1 CPU cycle	0.3 ns	1 s
Level 1 cache access	0.9 ns	3 s
Level 2 cache access	2.8 ns	9 s
Level 3 cache access	12.9 ns	43 s
Main memory access (DRAM, from CPU)	120 ns	6 min
Solid-state disk I/O (flash memory)	50–150 µs	2–6 days
Rotational disk I/O	1–10 ms	1–12 months
Internet: San Francisco to New York	40 ms	4 years
Internet: San Francisco to United Kingdom	81 ms	8 years
Internet: San Francisco to Australia	183 ms	19 years
TCP packet retransmit	1–3 s	105–317 years
OS virtualization system reboot	4 s	423 years
SCSI command time-out	30 s	3 millennia
Hardware (HW) virtualization system reboot	40 s	4 millennia
Physical system reboot	5 m	32 millennia

Memory

- Everything is on disk, very few things are in the registers
- Want to avoid going to main memory or disk because it's slow
- Take advantage of how programs actually access memory

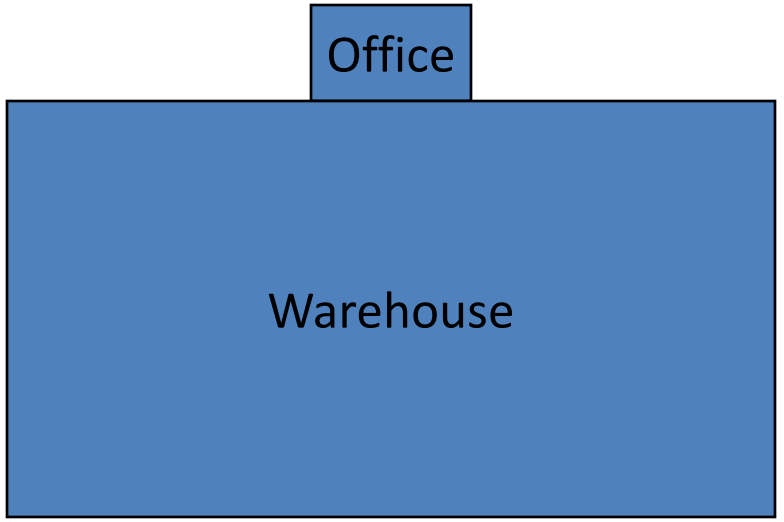


Principle of Locality

- Programs access a small proportion of their address space at any time
- Temporal locality
  - Items accessed recently are likely to be accessed again soon
  - e.g., instructions in a loop, registers spilled to the stack
- Spatial locality
  - Items near those accessed recently are likely to be accessed soon
  - E.g., sequential instruction access, array data

Library

- You have a huge library with EVERY book ever made.
- Getting a book from the library's warehouse takes 15 minutes.
- You can't serve enough people if every checkout takes 15 minutes.
- You have some small shelves in the front office.

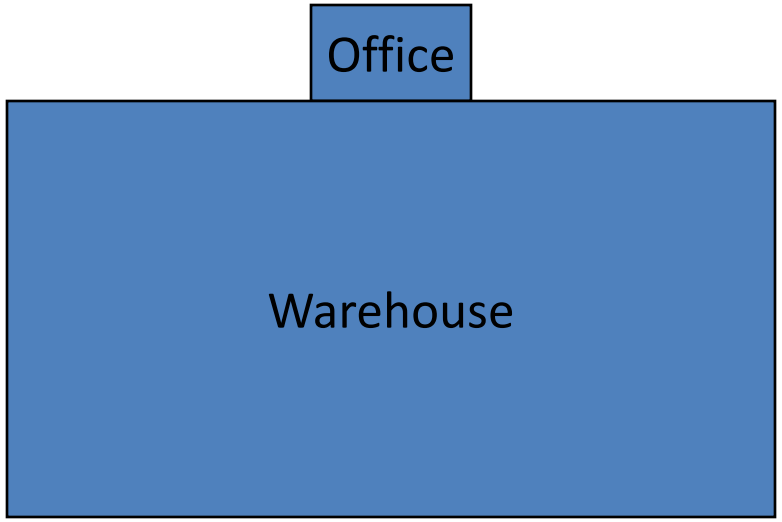


Here are some suggested improvements to the library:

1. Whenever someone checks out a book, keep other copies in the front office for a while in case someone else wants to check out the same book.
2. Watch the trends in books and attempt to guess books that will be checked out soon – put those in the front office.
3. Whenever someone checks out a book in a series, grab the other books in the series and put them in the front.
4. Buy motorcycles to ride in the warehouse to get the books faster

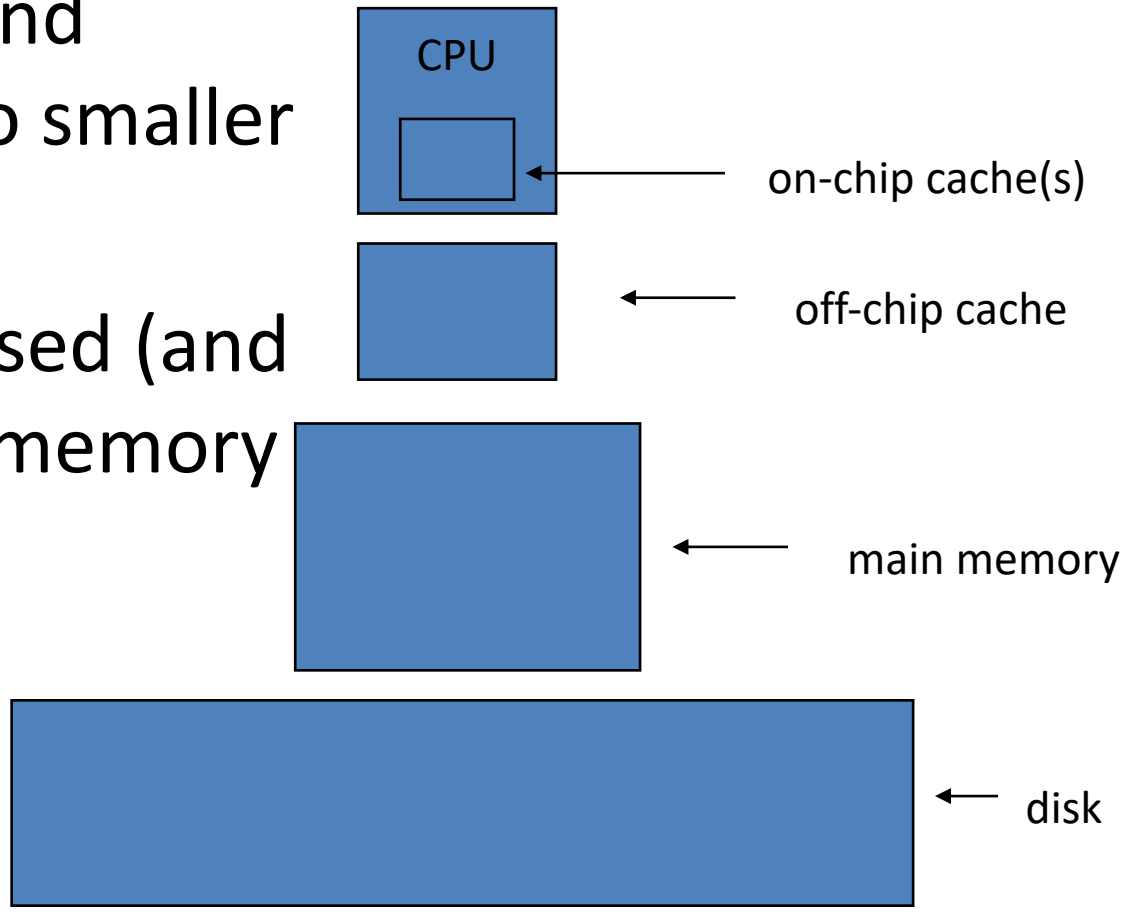
Extending the analogy to locality for caches, which pair of changes most closely matches the analogous cache locality?

Selection	Spatial	Temporal
A	2	1
B	4	2
C	4	3
D	3	1
E	None of the above	



Taking Advantage of Locality

- Store everything on disk
- Copy recently accessed (and nearby) items from disk to smaller main memory
- Copy more recently accessed (and nearby) items from main memory to cache



We know SRAM is very fast, expensive ($/GB), and small. We also know disks are slow, inexpensive ($/GB), and large. Which statement best describes the role of cache when it works.

Selection	Role of caching
A	Locality allows us to keep frequently touched data in SRAM.
B	Locality allows us the illusion of memory as fast as SRAM but as large as a disk.
C	SRAM is too expensive to make large – so it must be small and caching helps use it well.
D	Disks are too slow – we have to have something faster for our processor to access.
E	None of these accurately describes the role of cache.

Reading

- Next lecture: More Caches!
  - Section 6.3

- Problem Set 11 due Friday