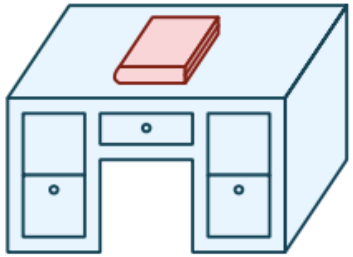
Lecture 13
Fundamental Memory Concepts (Part 1)

Xuan ‘Silvia’ Zhang
Washington University in St. Louis

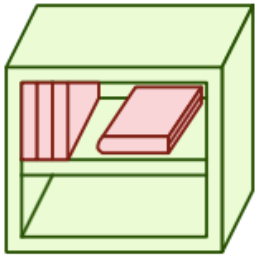
http://classes.engineering.wustl.edu/ese566/

Memory/Library Analogy

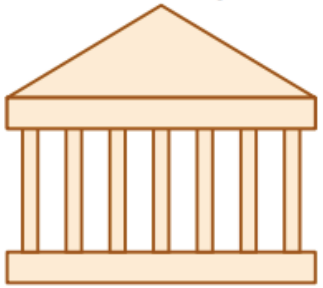
Desk
(can hold one book)



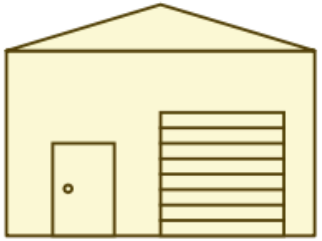
Book Shelf
(can hold a few books)



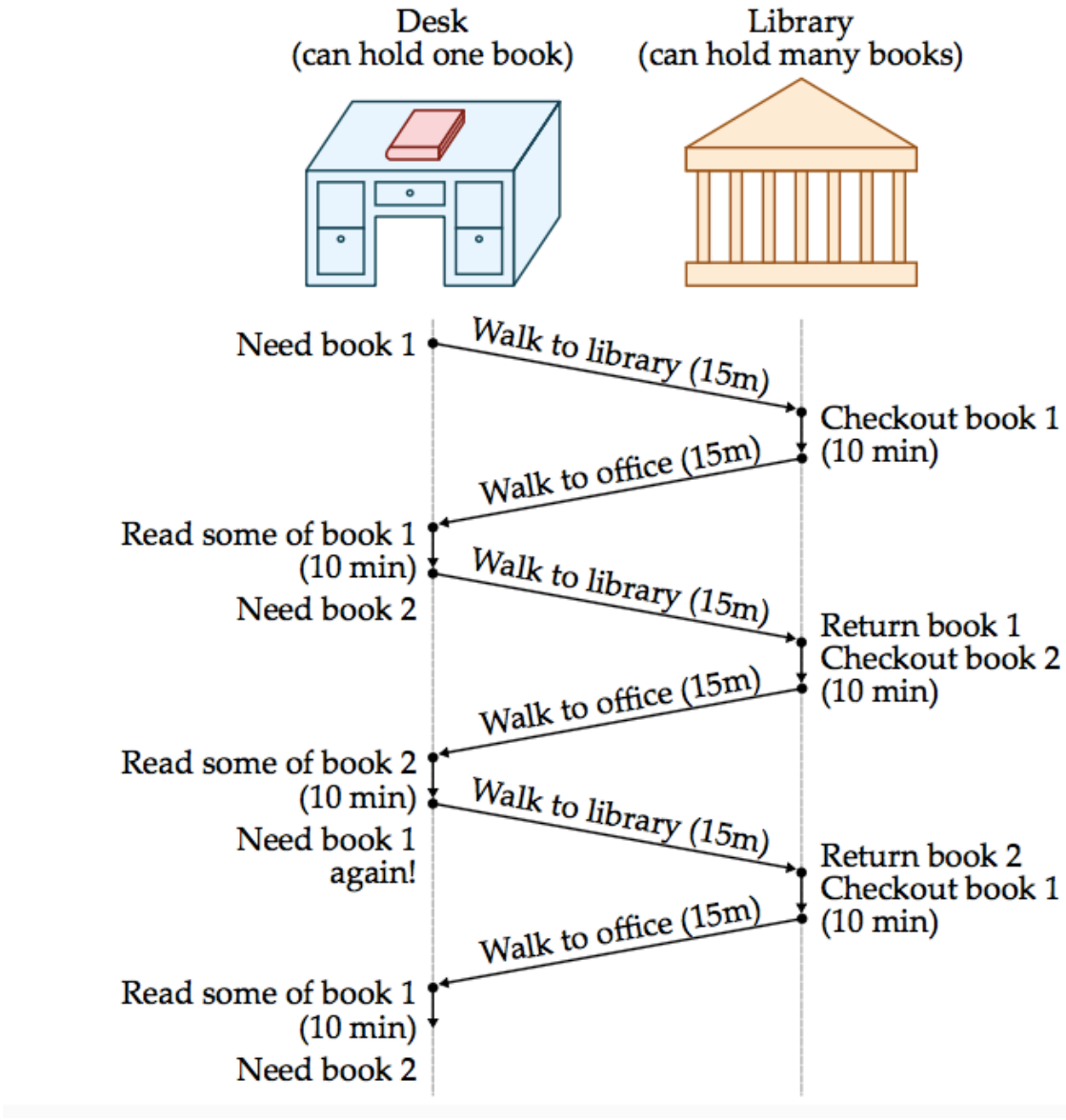
Library
(can hold many books)



Warehouse
(long-term storage)



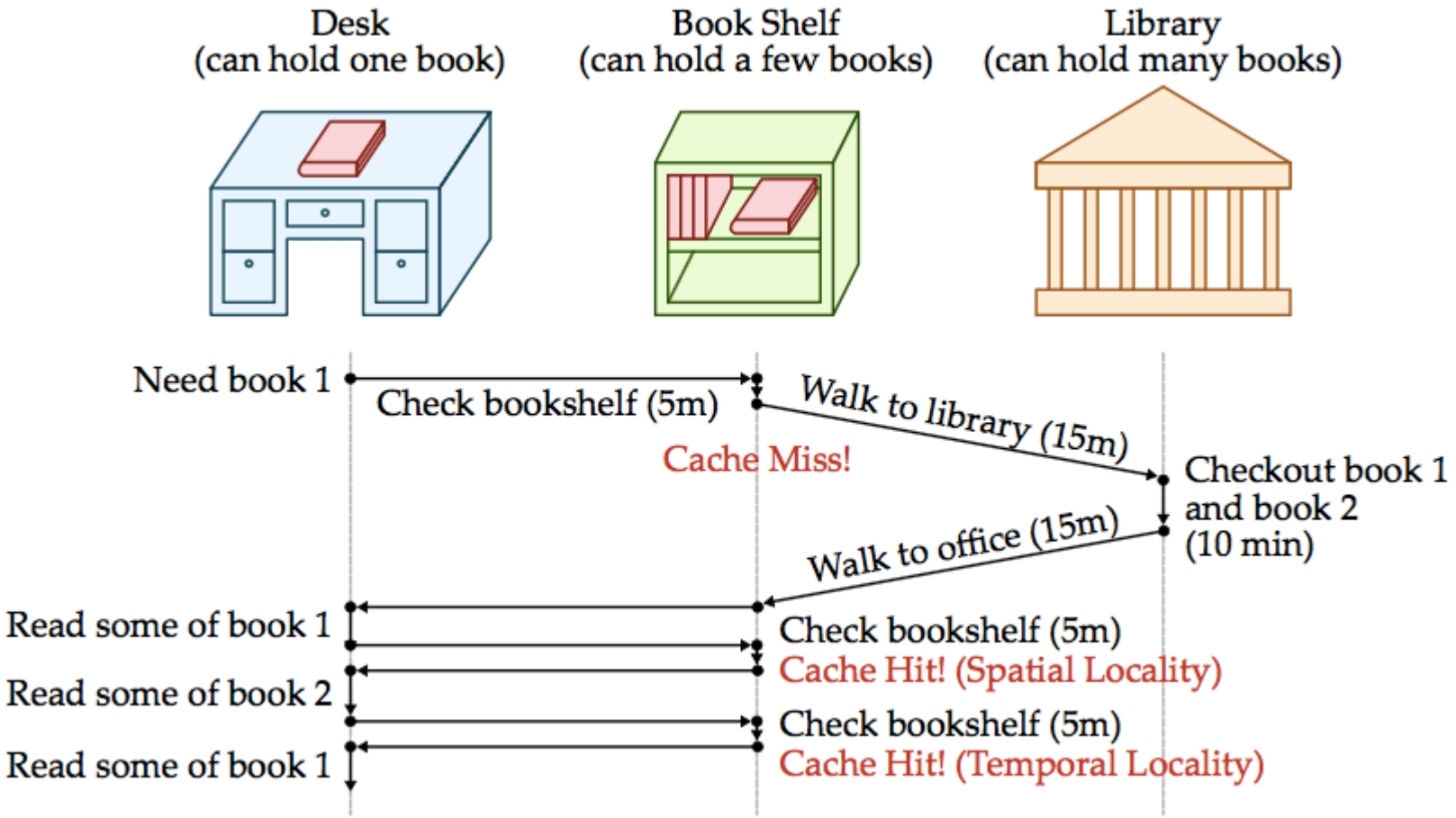
Scenario 1: Desk + Library, No Bookshelf "Cache"



avg. latency:
40 minutes

avg. throughput
(inc. reading):
1.2 books/hour

Scenario 1: Desk + Library with Bookshelf "Cache"



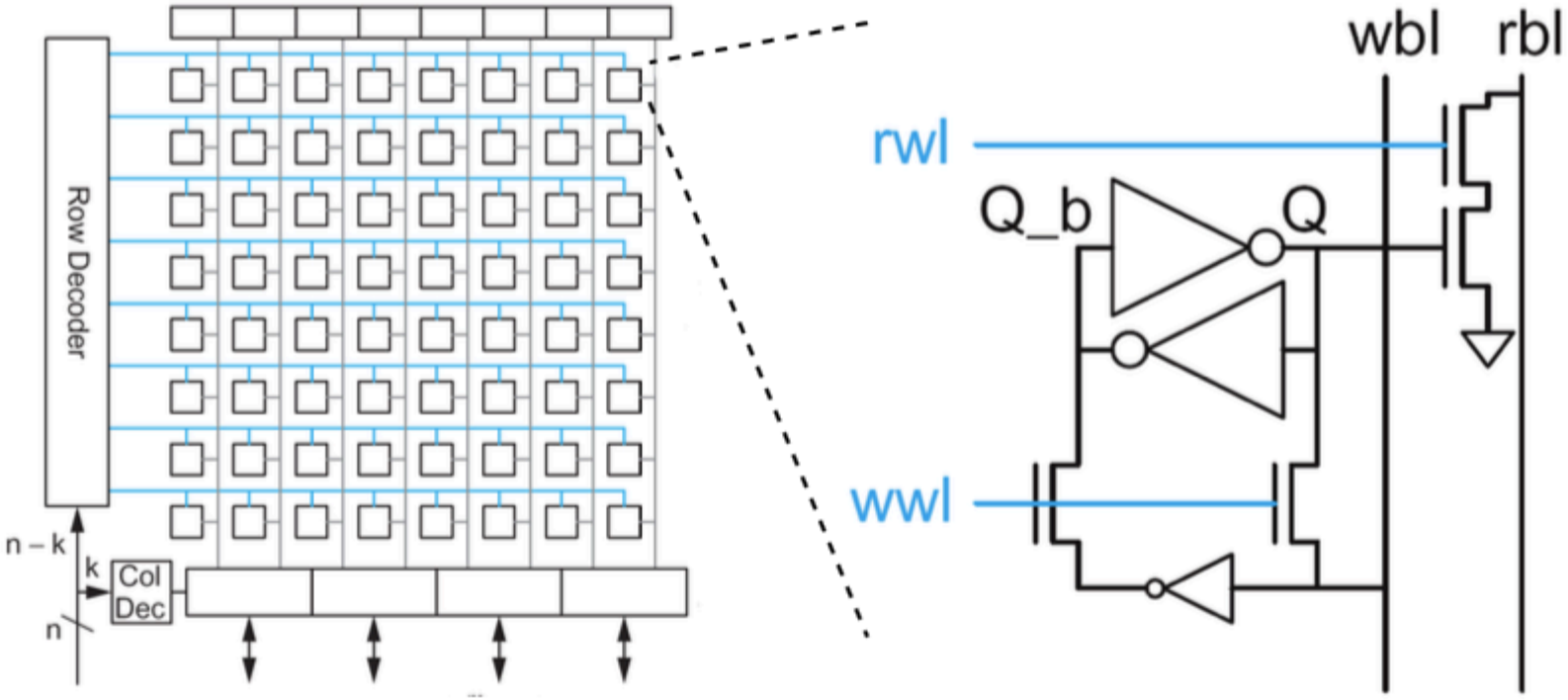
avg. latency: <20 minutes
avg. throughput (inc. reading): 2 books/hour

“Book Storage Hierarchy”

- Storage Blocks
  - bookshelf: low latency, low capacity
  - library: high latency, high capacity
  - warehouse: very high latency, very high capacity
- Bookshelf acts as a small “cache”
  - cache hit: book on the shelf
  - cache miss: need to go to library
- Exploit access pattern to improve access time
  - temporal locality: if we access a book once we are likely to access the same book again in near future
  - spatial locality: if we access a book on a given topic we are likely to access other books on the same topic in the near future

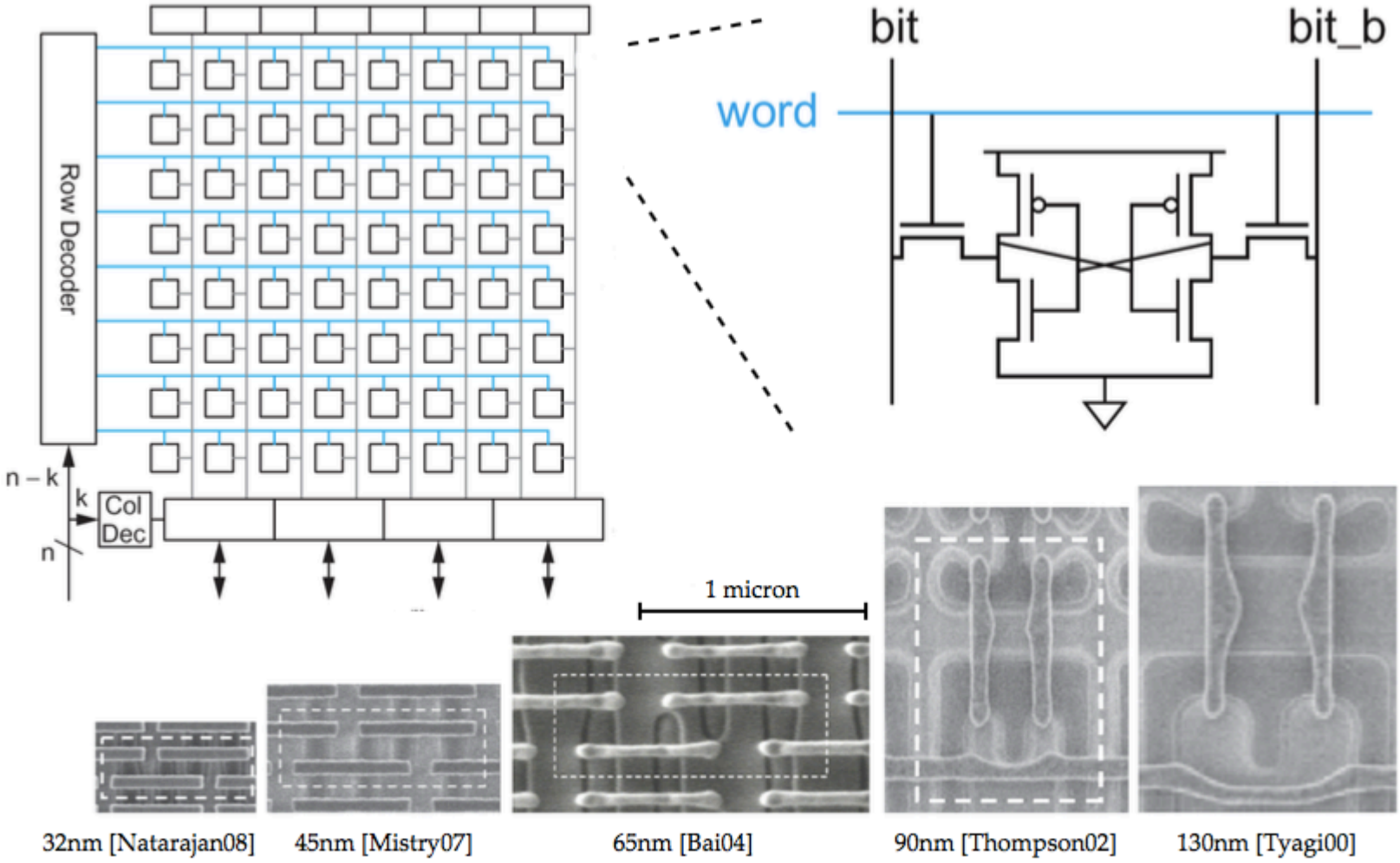
Memory Structure and Technology

- Register Files



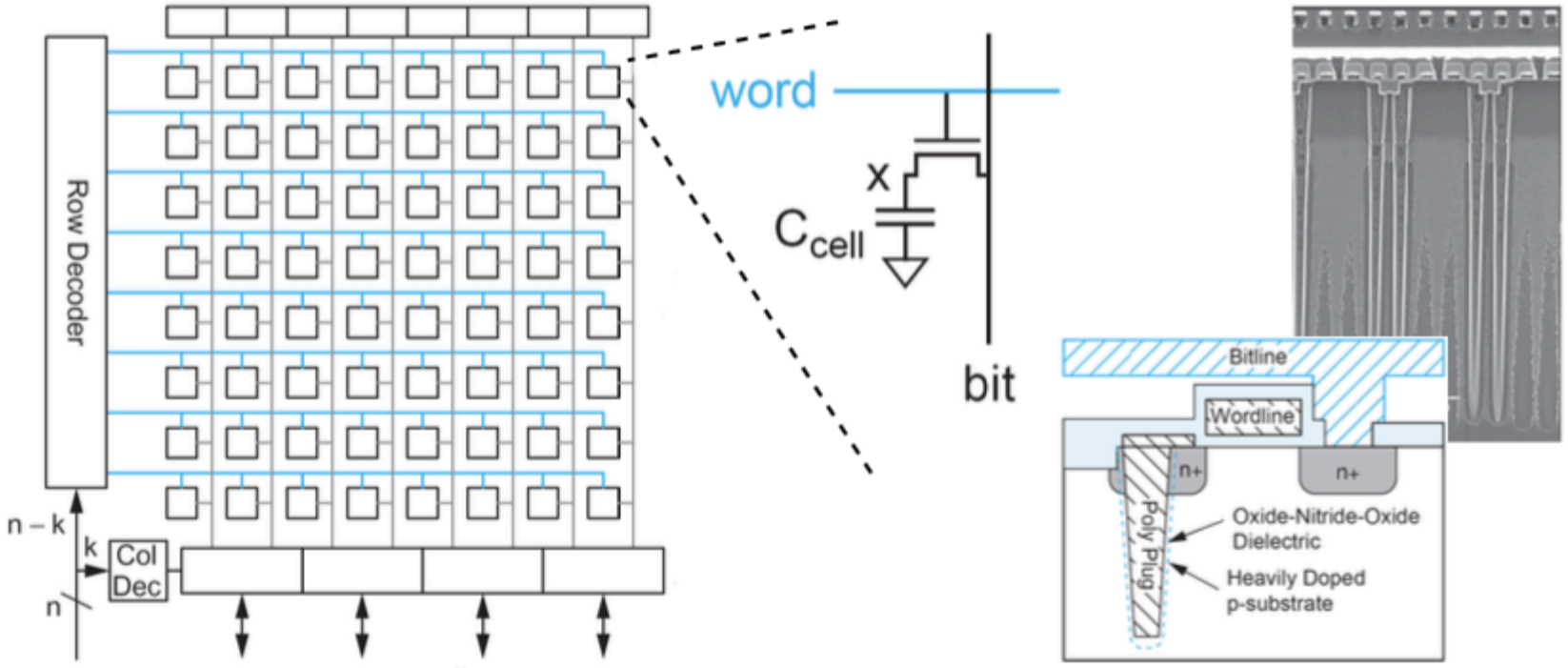
Memory Structure and Technology

- SRAM (cache, on-chip)



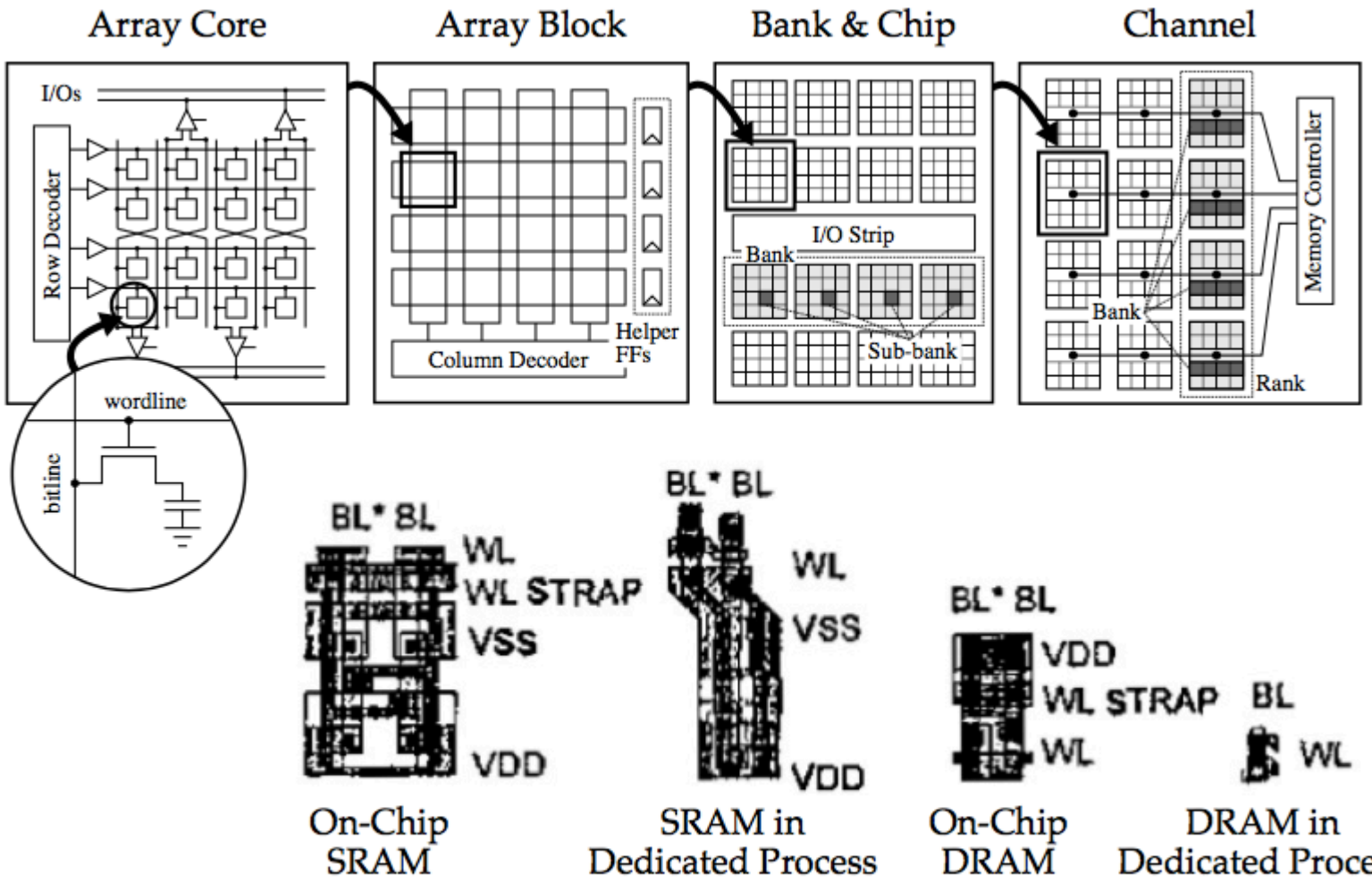
Memory Structure and Technology

- DRAM



Memory Structure and Technology

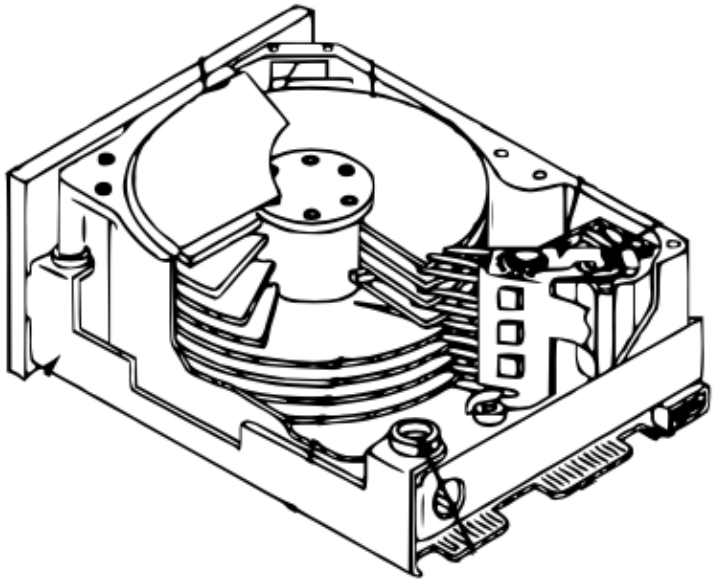
- DRAM



Adapted from [Foss, "Implementing Application-Specific Memory." ISSCC'96]

Memory Structure and Technology

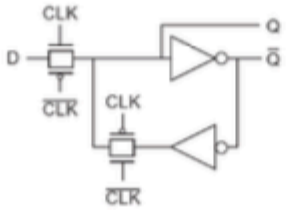
- Disk
  - magnetic hard drives require rotating platters resulting in long random access times with have hardly improved over several decades



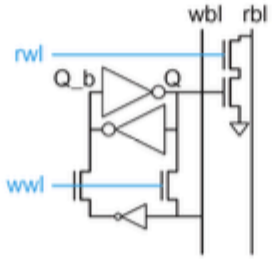
- Flash
  - solid-state drives using flash have 100x lower latencies, but also lower density and higher cost

Memory Technology Trade-offs

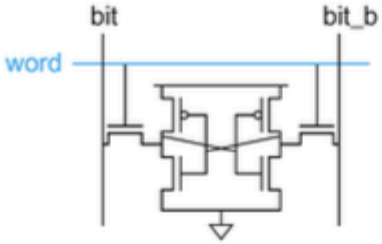
Latches & Registers



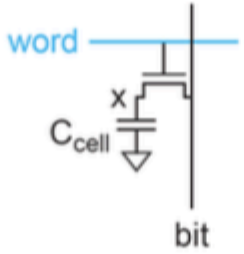
Register Files



SRAM



DRAM



Flash & Disk

Low Capacity
Low Latency
High Bandwidth
(more and wider ports)

High Capacity
High Latency
Low Bandwidth

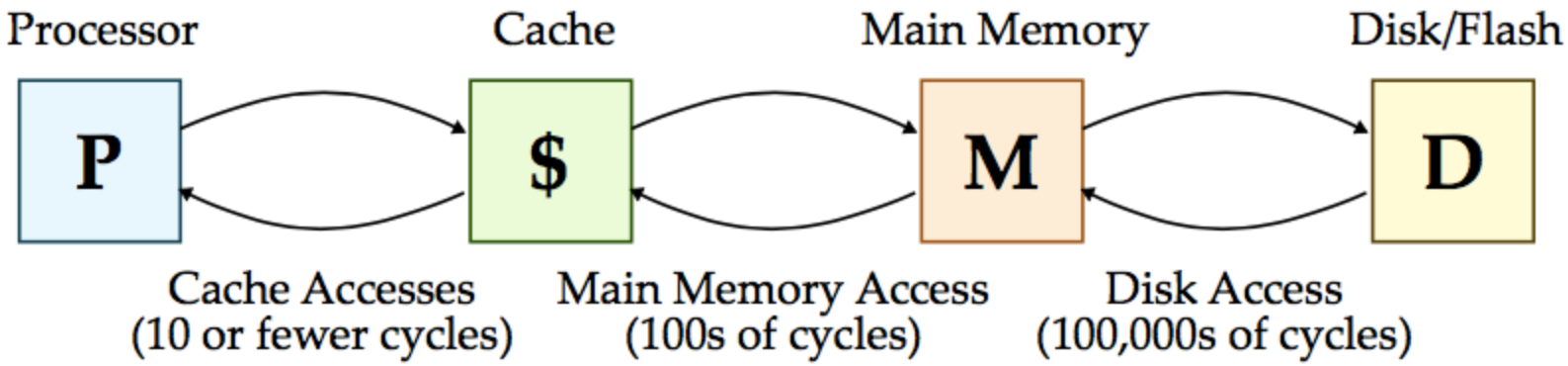
Latency Numbers: every programmers (architect) should know

L1 cache reference	1 ns
Branch mispredict	3 ns
L2 cache reference	4 ns
Mutex lock/unlock	17 ns
Main memory reference	100 ns
Send 2KB over commodity network	250 ns
Compress 1KB with zip	2 us
Read 1MB sequentially from main memory	9 us
SSD random read	16 us
Read 1MB sequentially from SSD	156 us
Round trip in datacenter	500 us
Read 1MB sequentially from disk	2 ms
Disk random read	4 ms
Packet roundtrip from CA to Netherlands	150 ms

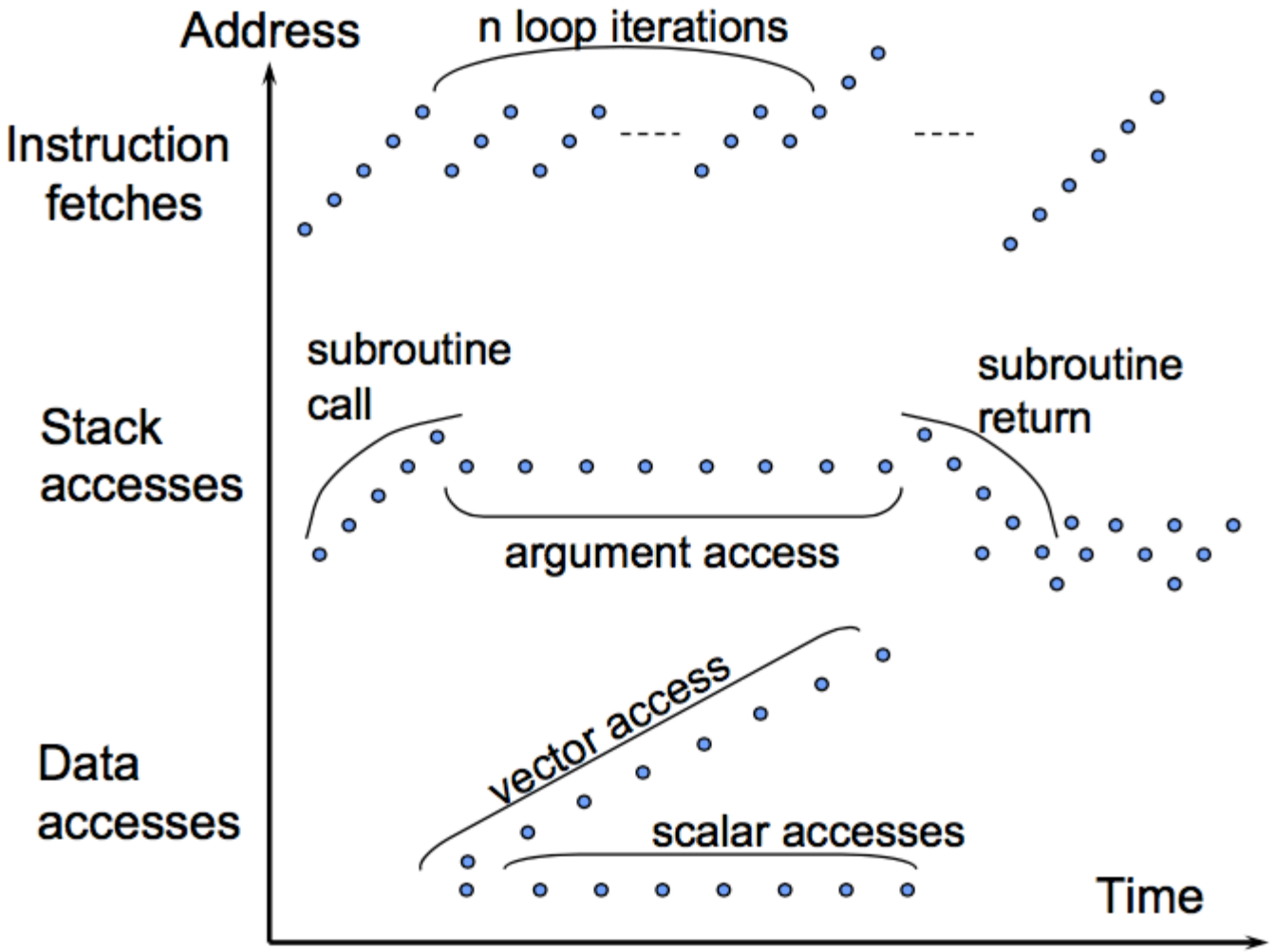
find updated at https://people.eecs.berkeley.edu/~rcs/research/interactive_latency.html

Cache Memories in Computer Architecture

- Three key questions
  - how much data is aggregated in a cache line
  - how do we organize multiple lines in cache
  - what data is replaced to make room for new data when cache is full
- Categorizing misses
- Write policies
- Multi-level cache

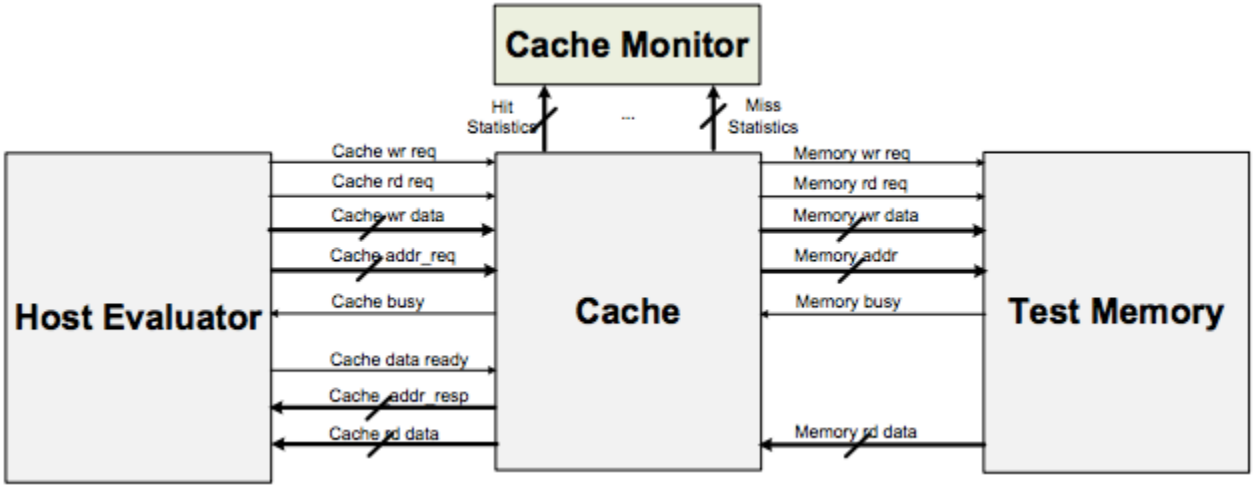


Typical Data Access Pattern
instruction vs data access, temporal vs spatial locality



Lab3: Design a Cache

- Direct-mapped cache (baseline)
- Two-way associative cache
- Write-through vs write-back
- Test bench



- Due on 2/22 at 2:30pm

Questions?

Comments?

Discussion?

Acknowledgement

Cornell University, ECE 4750