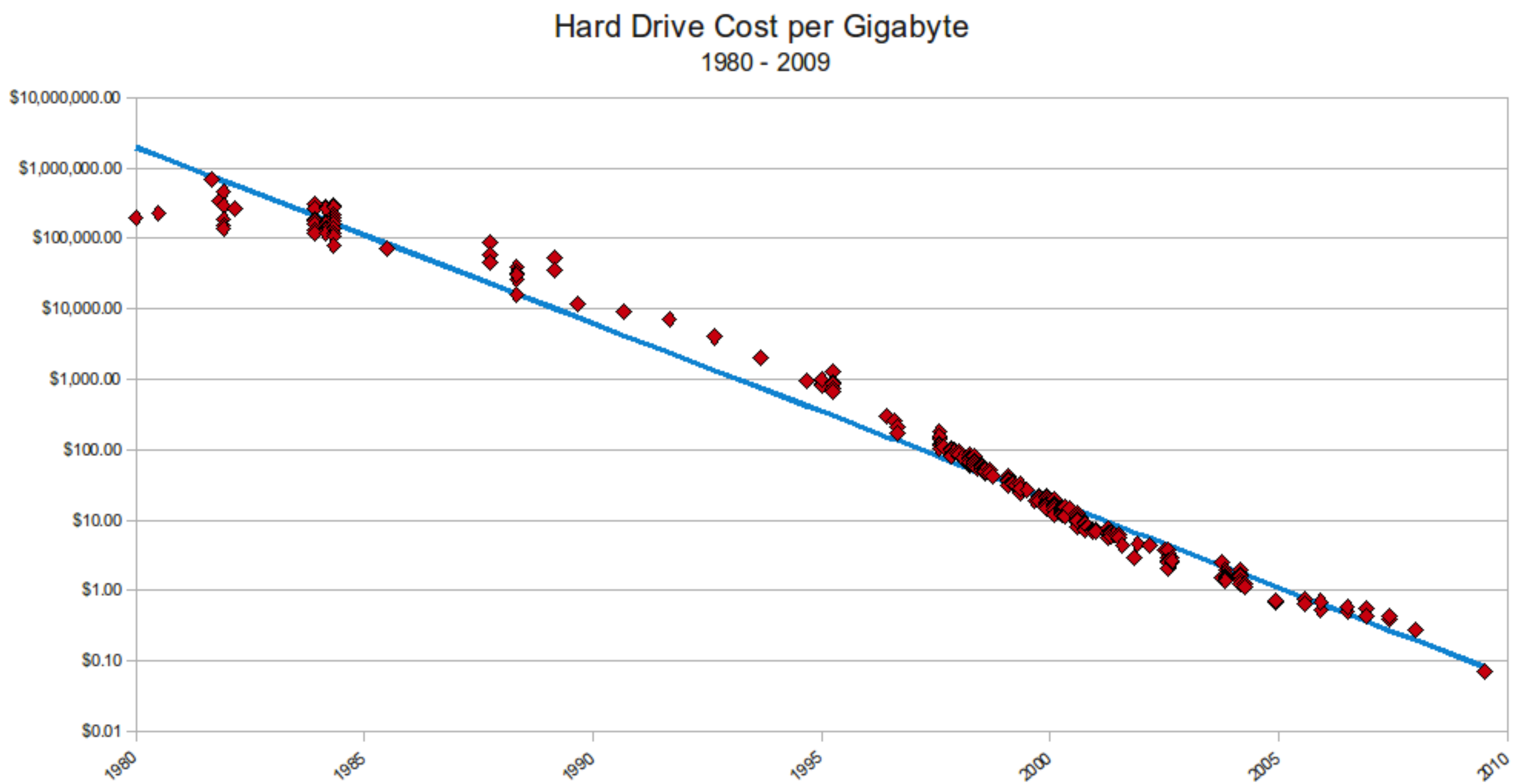
Computer Architecture Trends That Affect Data Analytics

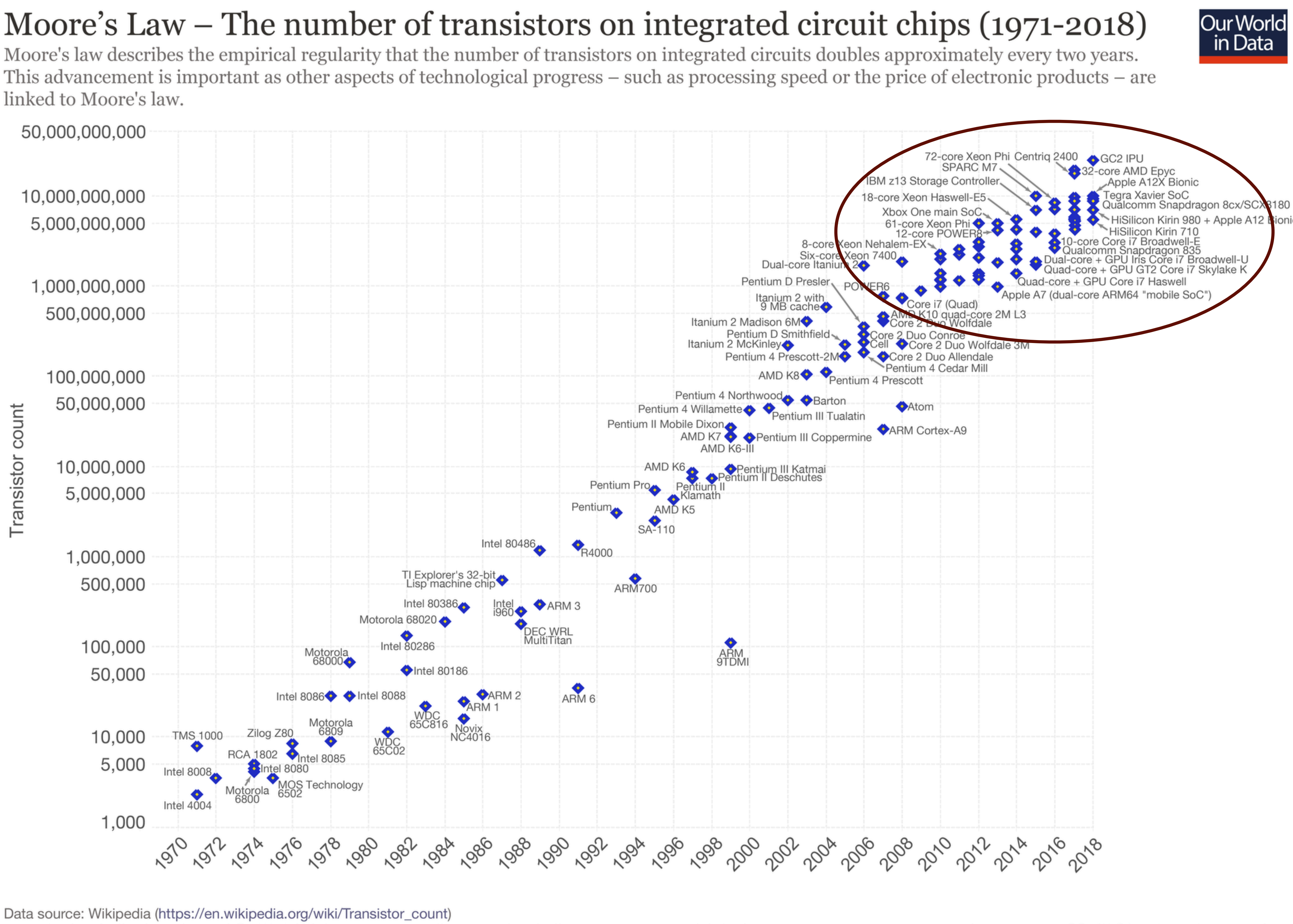
CHIDATA

CHIDATA

Last Lecture



Last Lecture



This Lecture

CPU = Fast

Moving Data = Slow

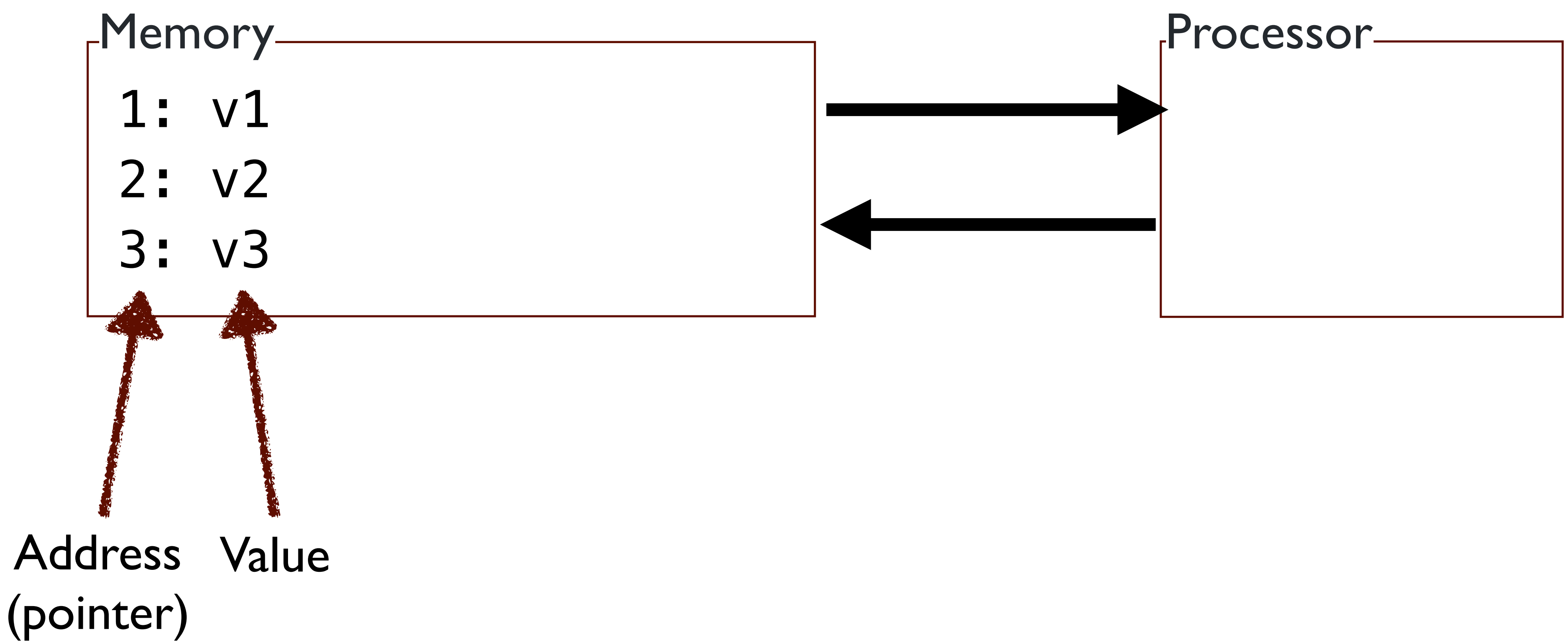
This Lecture

The "storage" story is actually much more complicated.

Data are typically stored in a hierarchy of slow but big to fast but small storage systems.

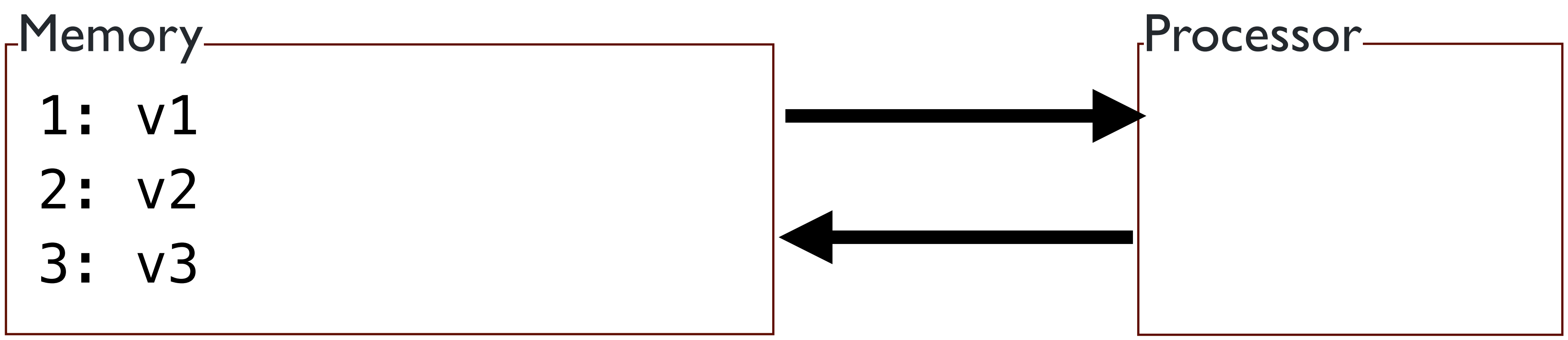
Overview of the changes on the horizon.

A Basic Computing Model



A Basic Computing Model

`read(address)`

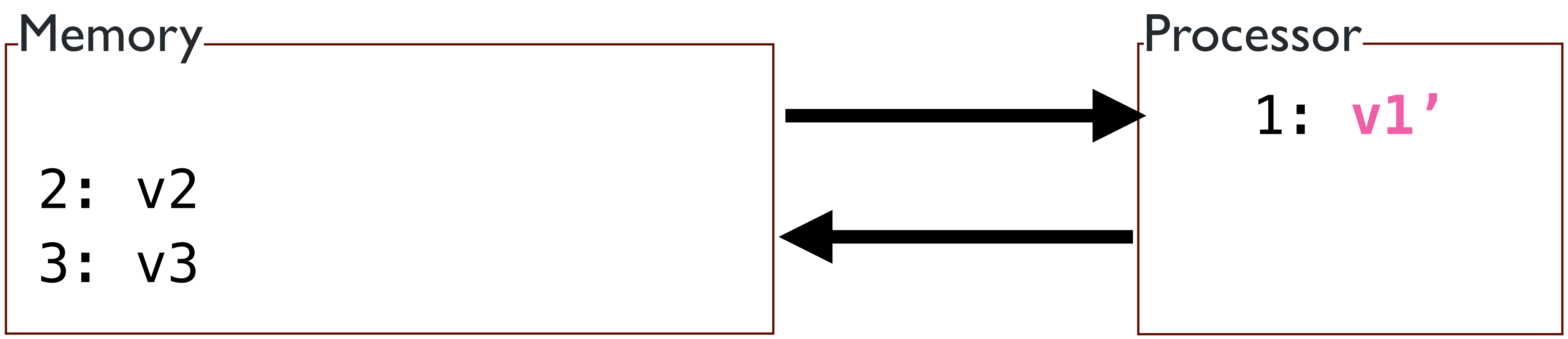


Move data from memory to processor

Processor has a limited amount of native memory

A Basic Computing Model

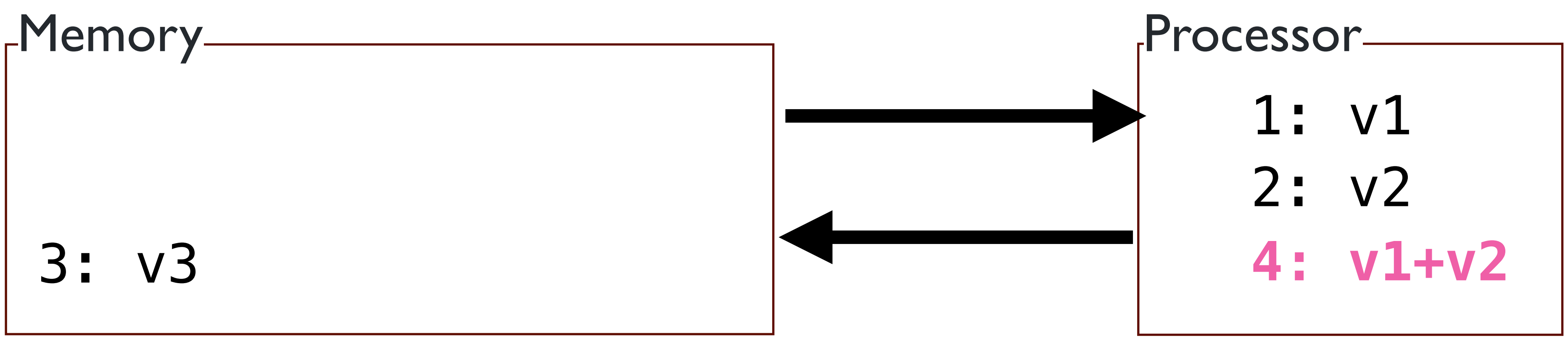
`inst(value)`



Apply a basic instruction to the data

A Basic Computing Model

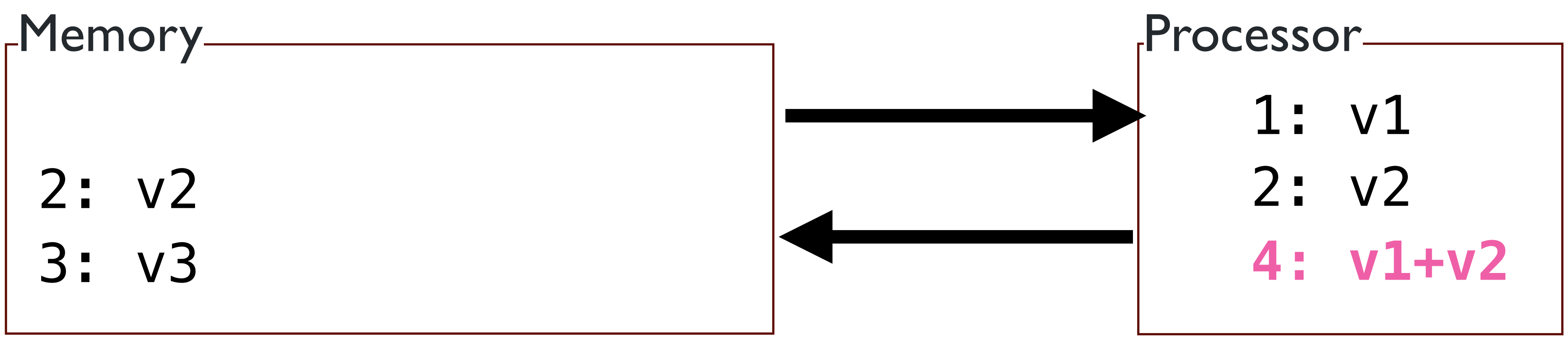
```
inst(value)
```



Apply a basic instruction to the data

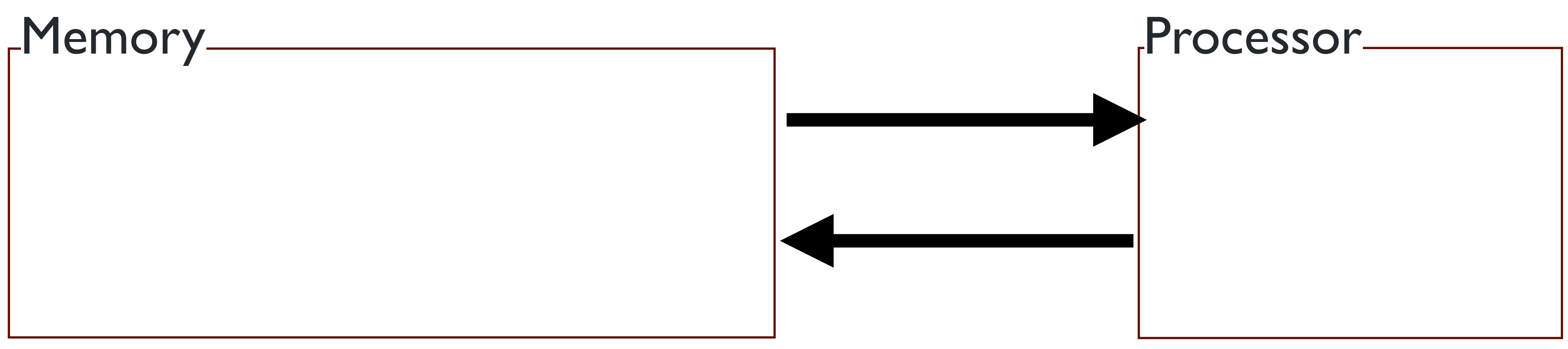
A Basic Computing Model

`write(address)`



Write data back to memory

Run Time



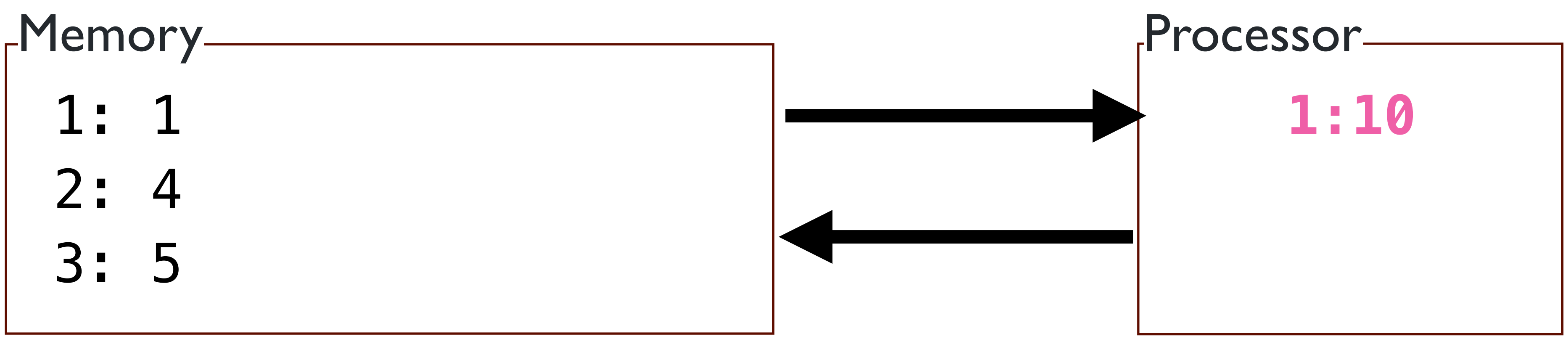
Runtime = I/O Time + Compute Time

Total
read()/write()

Total
inst()

Calculate I/O and CPU Costs

Given a list of numbers [1,4,5,…,6] calculate the sum:



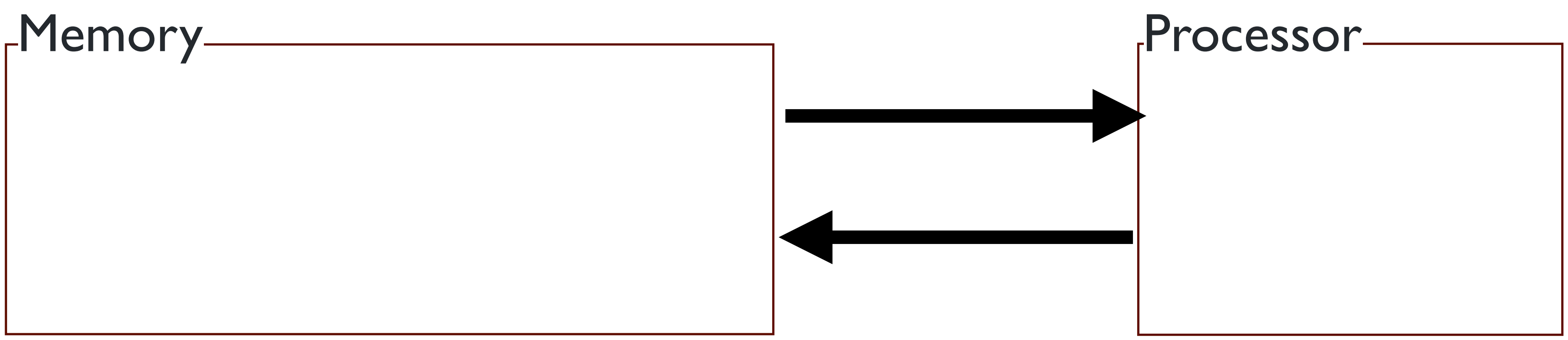
1 read() + 1 sum() per data item

1 write() for final answer

I/O Cost: N+1, CPU Cost: N-1

Calculate I/O and CPU Costs

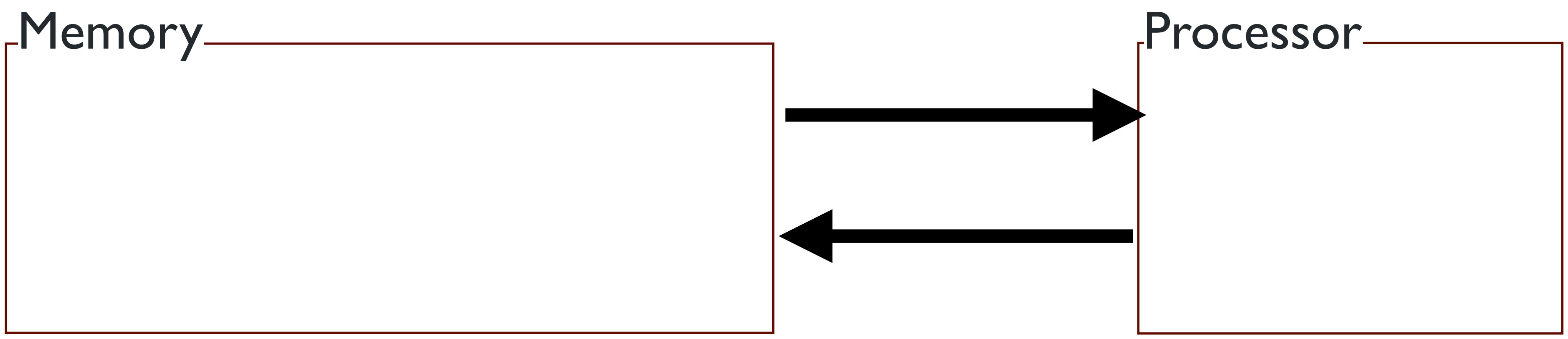
Given a list of numbers [1,4,5,…,6] calculate the sum:



I/O Cost: _N, CPU Cost: _N

Generally care about orders and ignore small constants

Rate-Limiting Operations



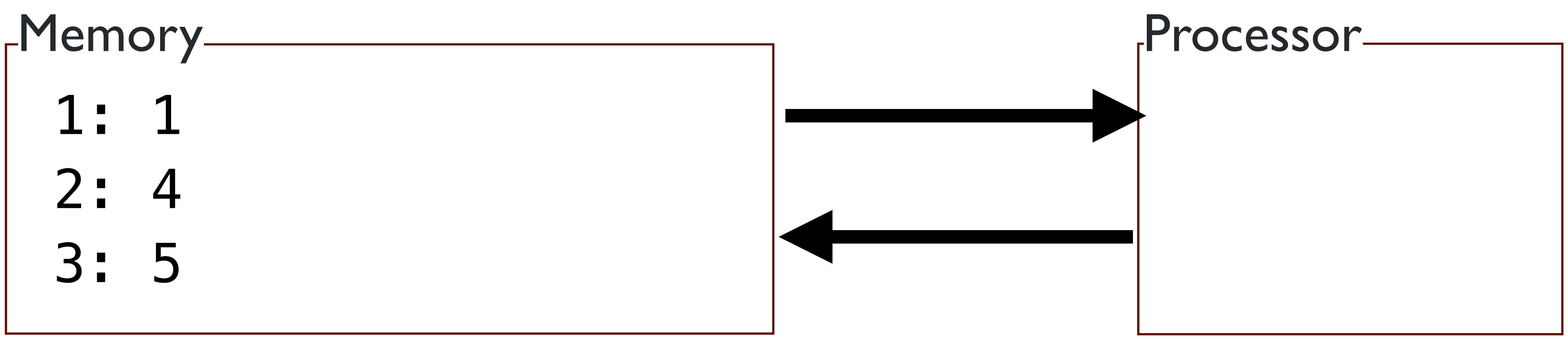
Runtime = I/O Time + Compute Time

Runtime = a*IO Steps + b*CPU Steps

a is usually MUCH bigger than b! (thousands of times!)

Calculate I/O and CPU Costs

Given a list of numbers [1,4,5,…,6] calculate the sum of the even numbers:

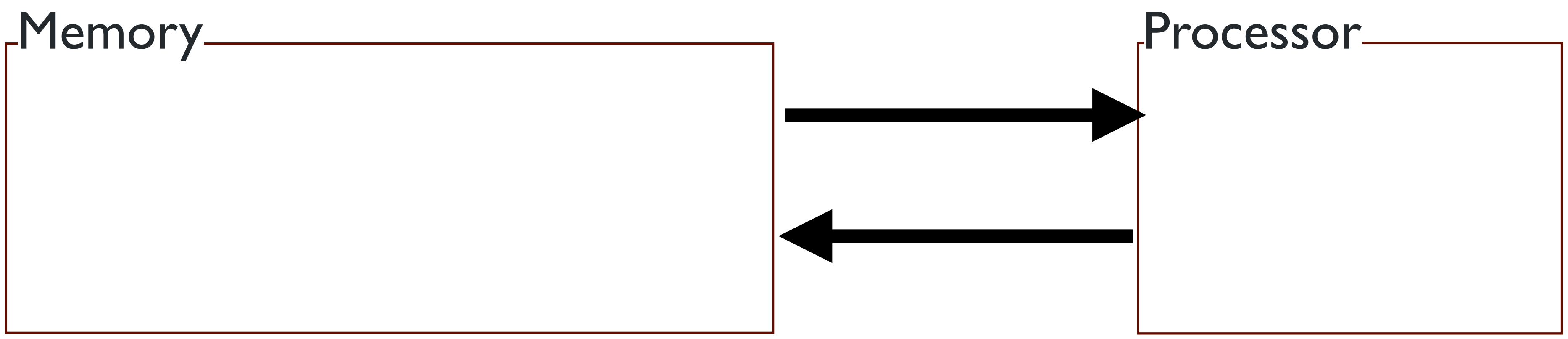


1 read() + 1 mod() + 1 branch() + at most 1 sum() per data item

1 write() for final answer

I/O Cost: N, CPU Cost: **3**N

Rate-Limiting Operations



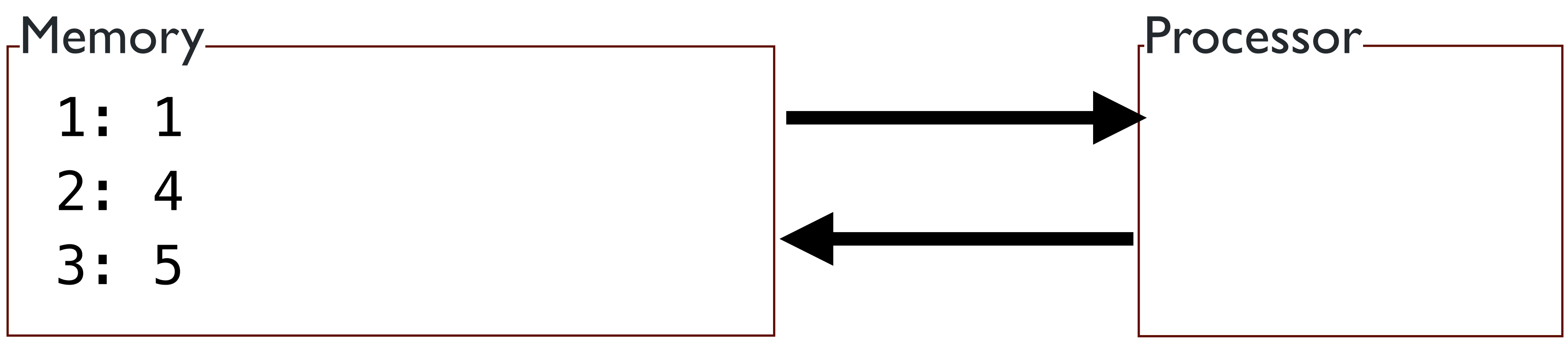
Runtime = I/O Time + Compute Time

"I/O Bound" : I/O Time >> Compute Time

"CPU Bound" : Compute Time >> I/O Time

Calculate I/O and CPU Costs

Given the median of [1,4,5,…,6]



What if the processor can store >N/2 data points?

I/O V.S CPU Bound

"I/O Bound" : Simple operations over lots of data

- Sorting a list of numbers. (Operation: Comparison)
- Summing a list of numbers (Operation: Sum)
- Finding a number in the list that is less than 5 (Operation: Comparison)

"CPU Bound" : Complex repetitive over small working sets

- Matrix multiplication
- Cryptographic hashing
- Image processing

Data Analytics is (usually) I/O Bound!

“I/O Bound” : Simple operations over lots of data

- Sorting a list of numbers. (Operation: Comparison)
- Summing a list of numbers (Operation: Sum)
- Finding a number in the list that is less than 5 (Operation: Comparison)

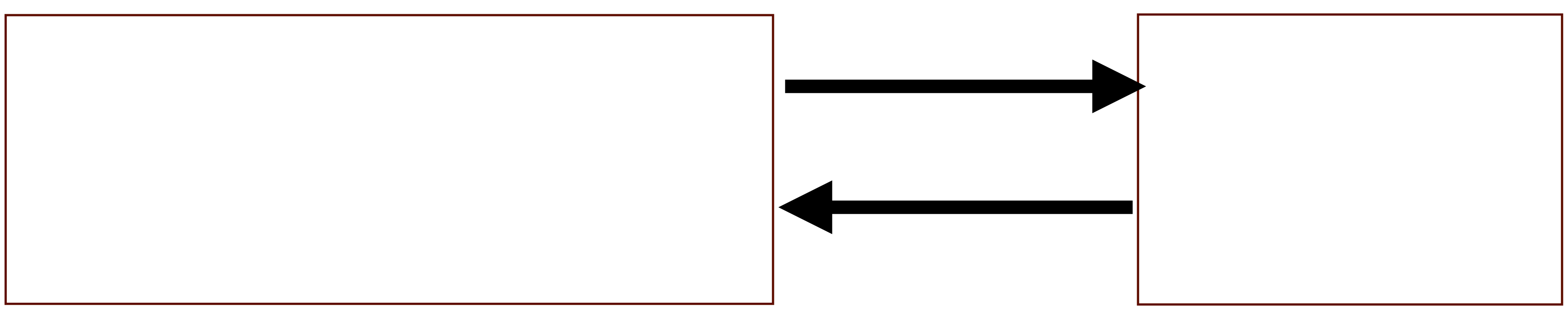
“CPU Bound” : Complex repetitive over small working sets

- Matrix multiplication
- Cryptographic hashing
- Image processing

Optimizing I/O Bound Processes

"I/O Bound" : Simple operations over lots of data

Finding a number in the list that is less than 5 (Operation: Comparison)



Memory

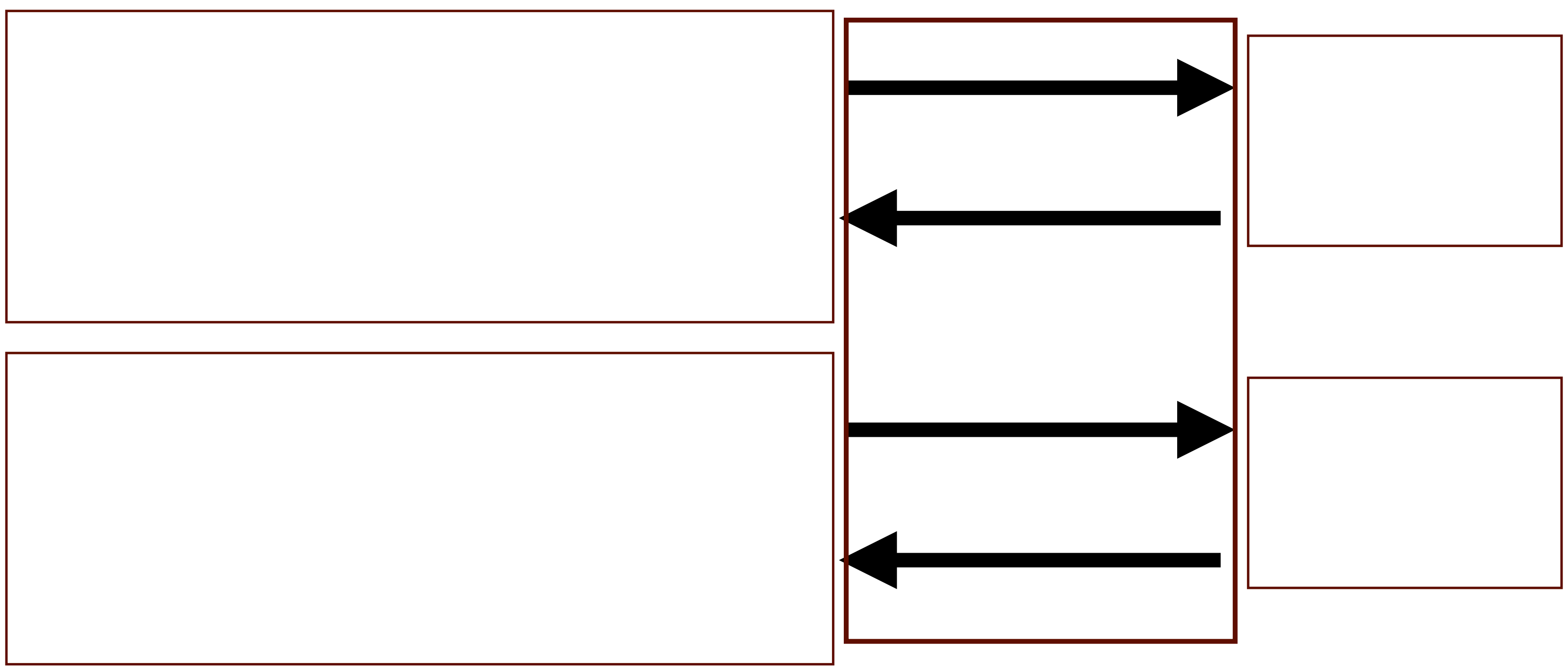
Processor

Not as useful! I/O Cost Still N

Optimizing I/O Bound Processes

“I/O Bound” : Simple operations over lots of data

Finding a number in the list that is less than 5 (Operation: Comparison)



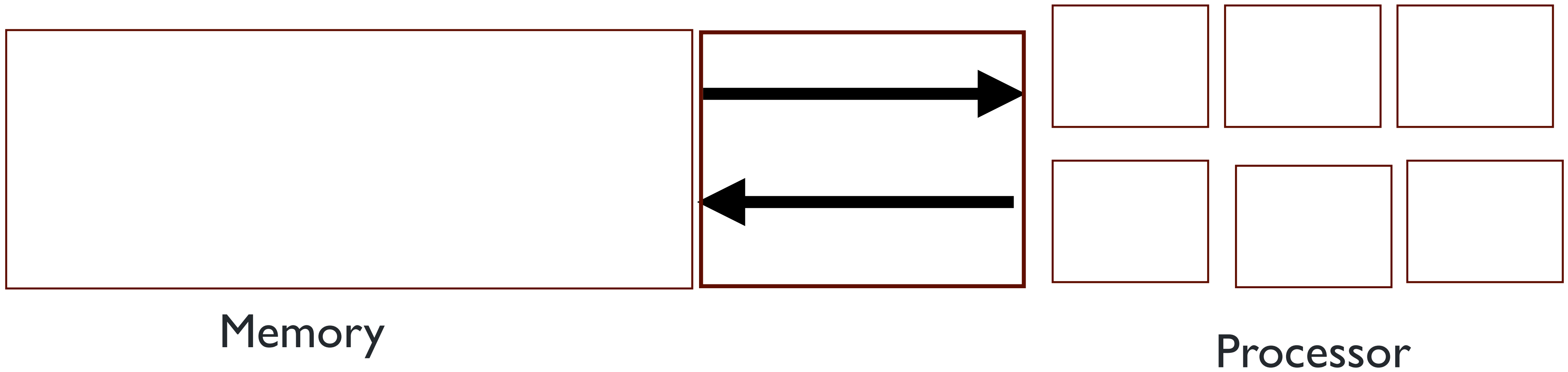
Better Strategy, I/O constant goes dow!

*Equivalent Serial I/O Cost of N/2**

Optimizing I/O Bound Processes

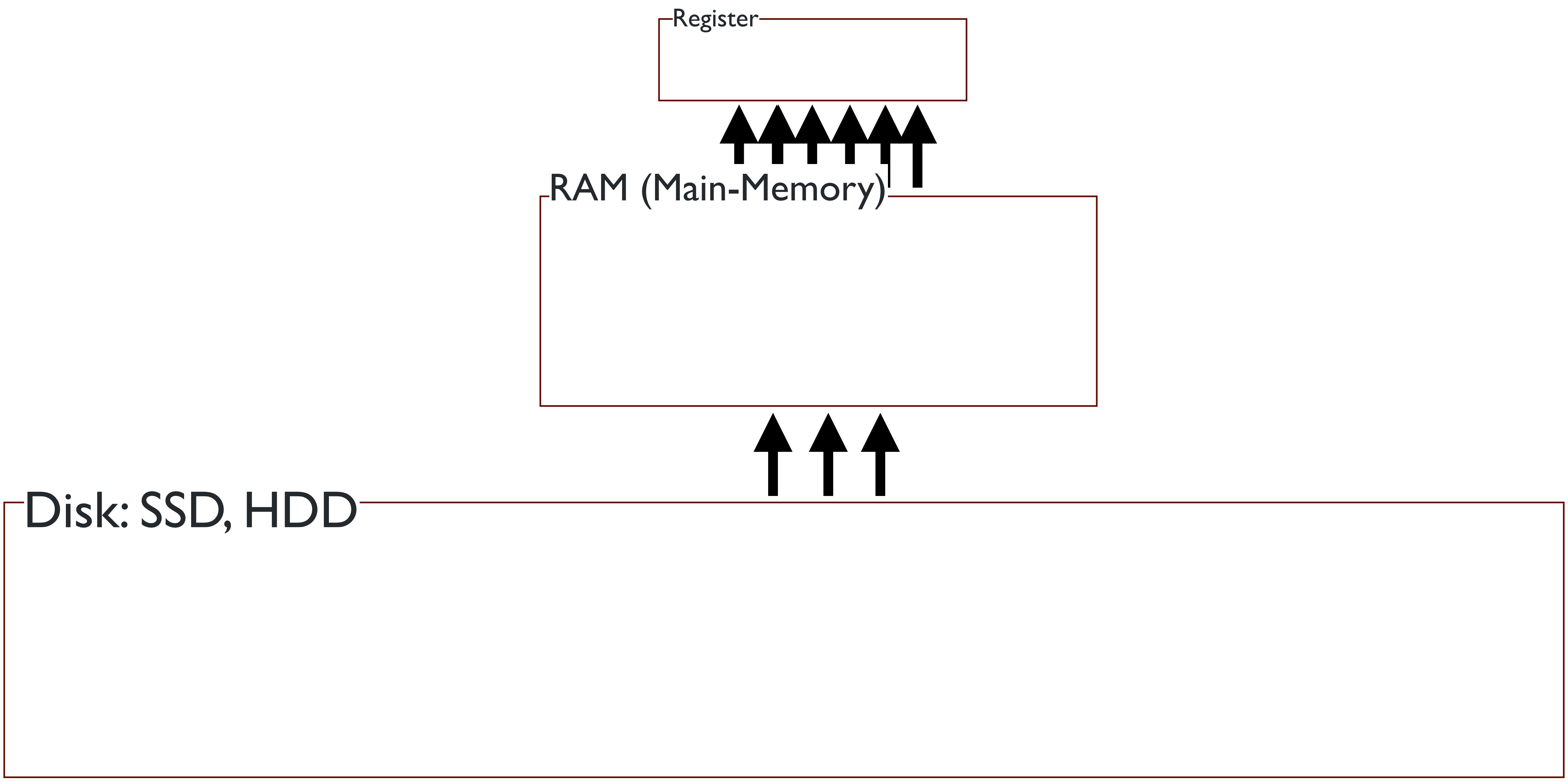
"I/O Bound" : Simple operations over lots of data

Finding a number in the list that is less than 5 (Operation: Comparison)

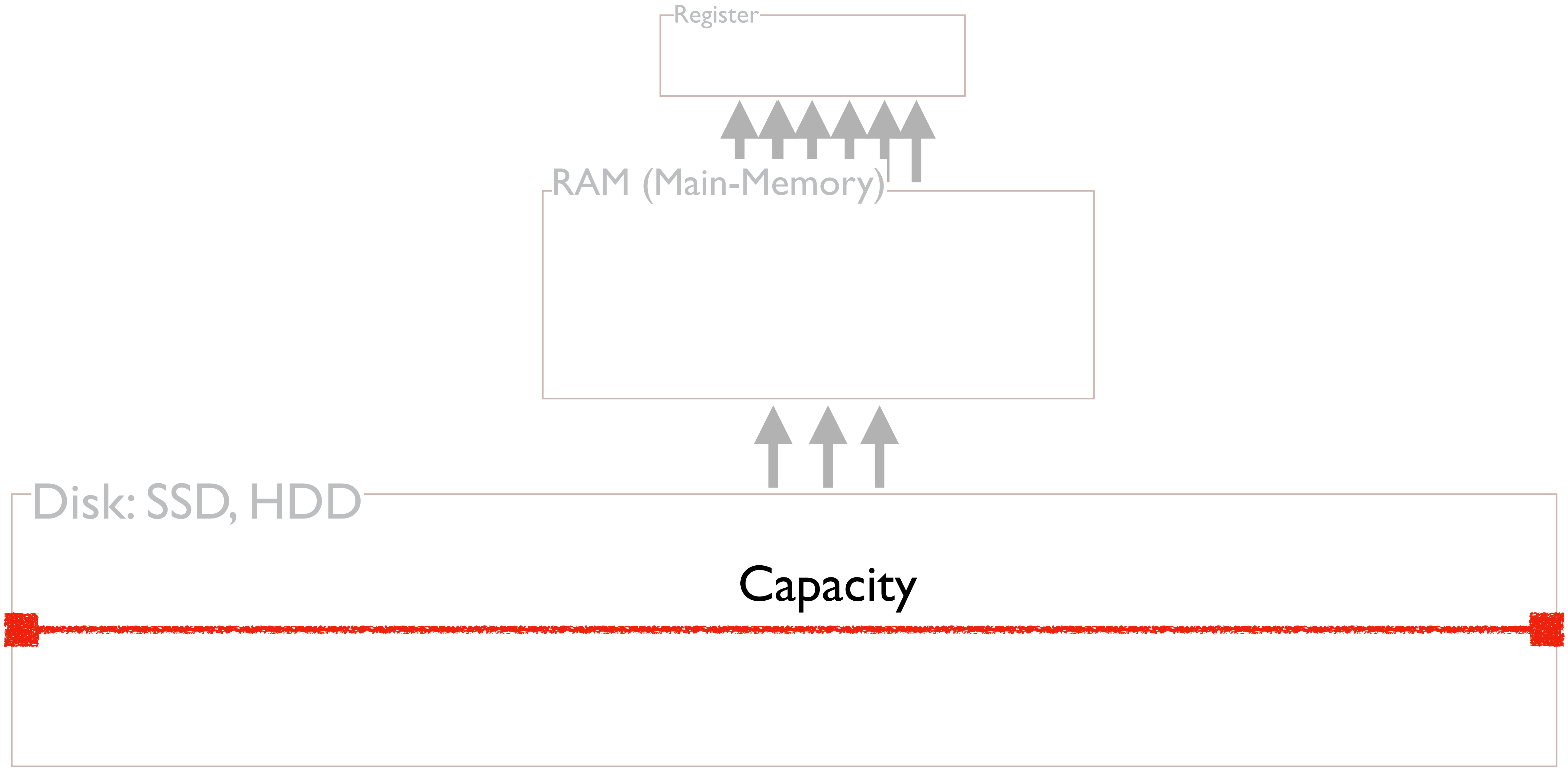


Depends on how much I/O can be parallelized

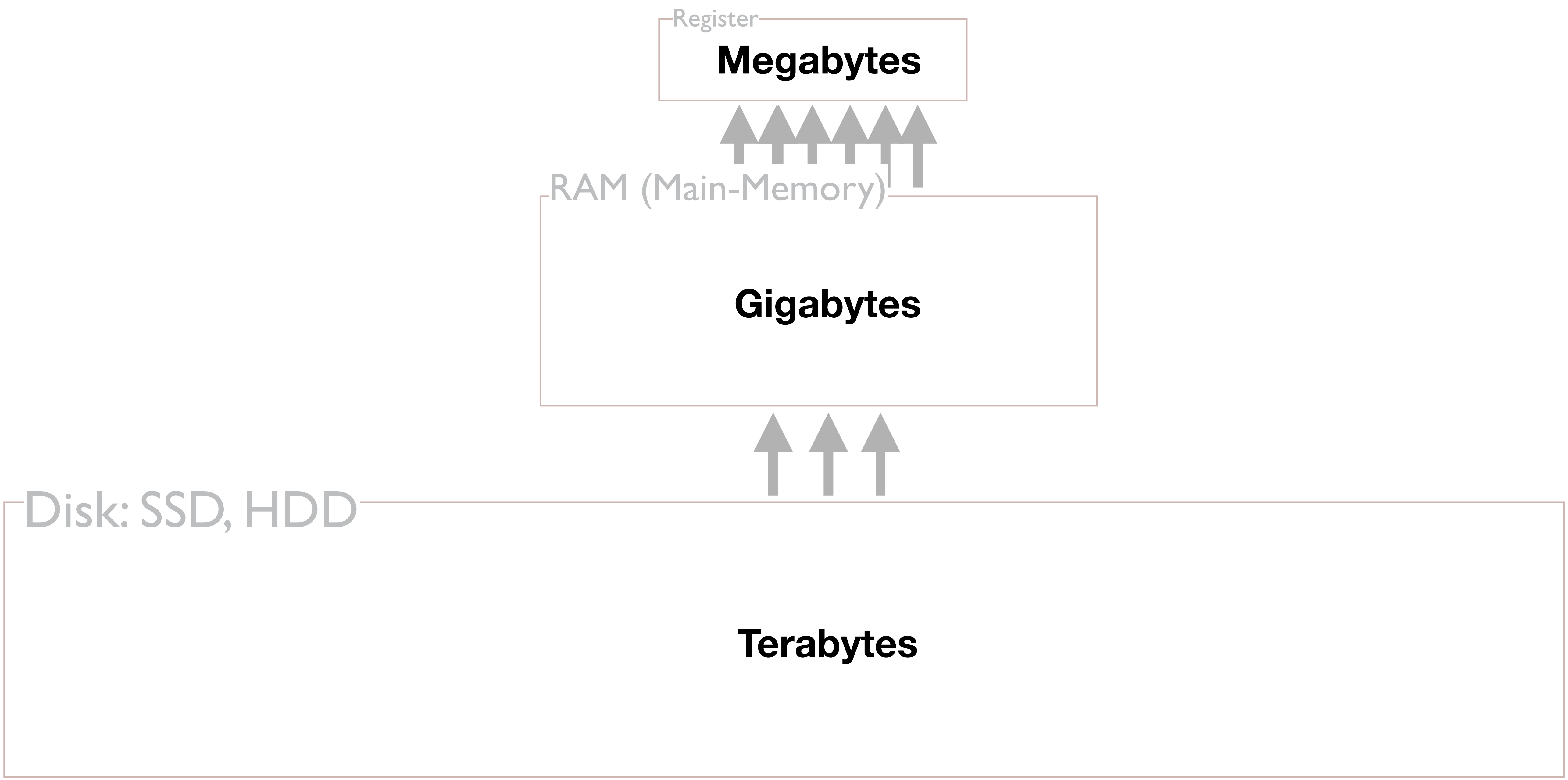
There are more levels!



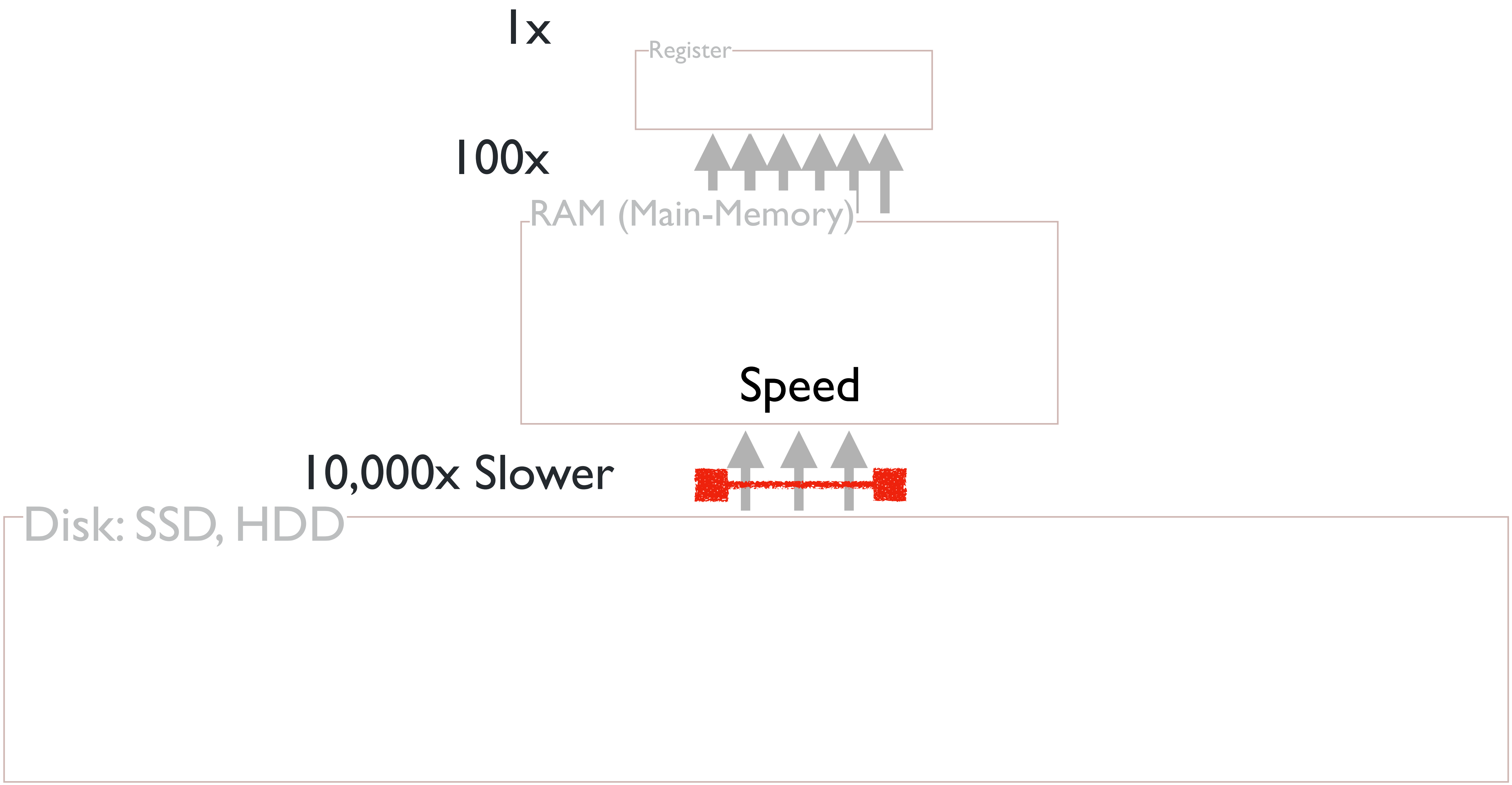
Schematic of Computer Storage



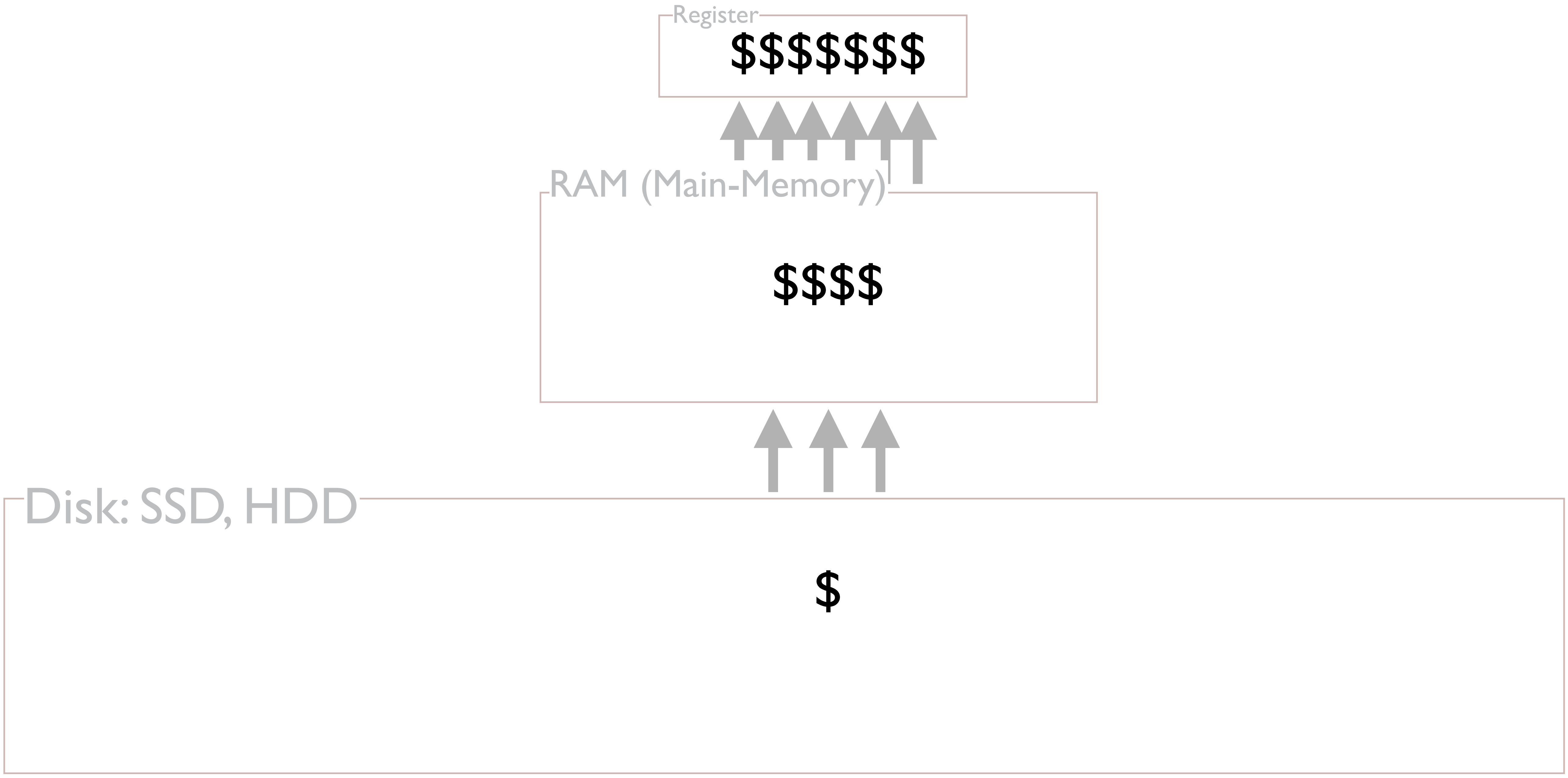
Schematic of Computer Storage



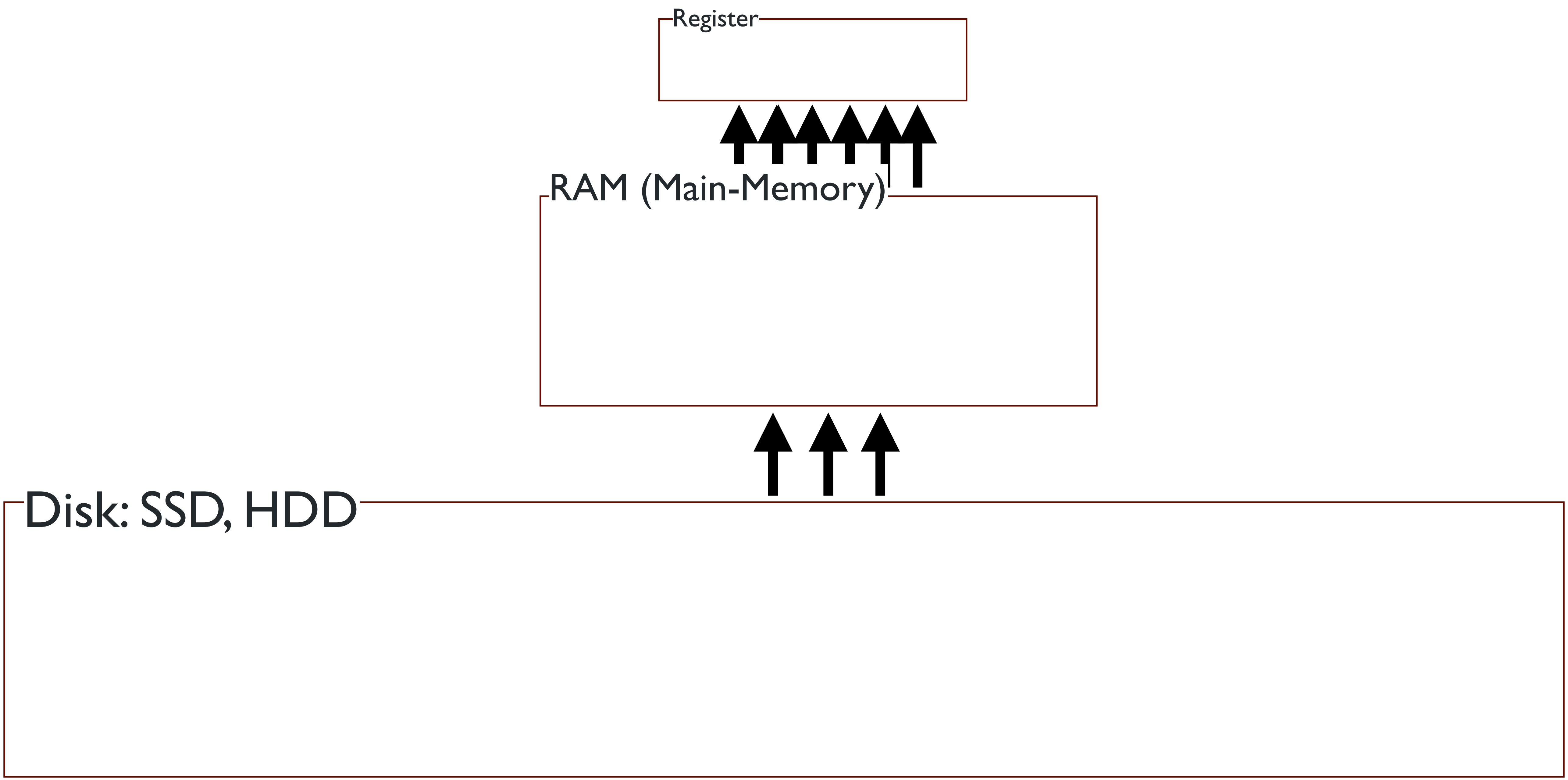
Schematic of Computer Storage



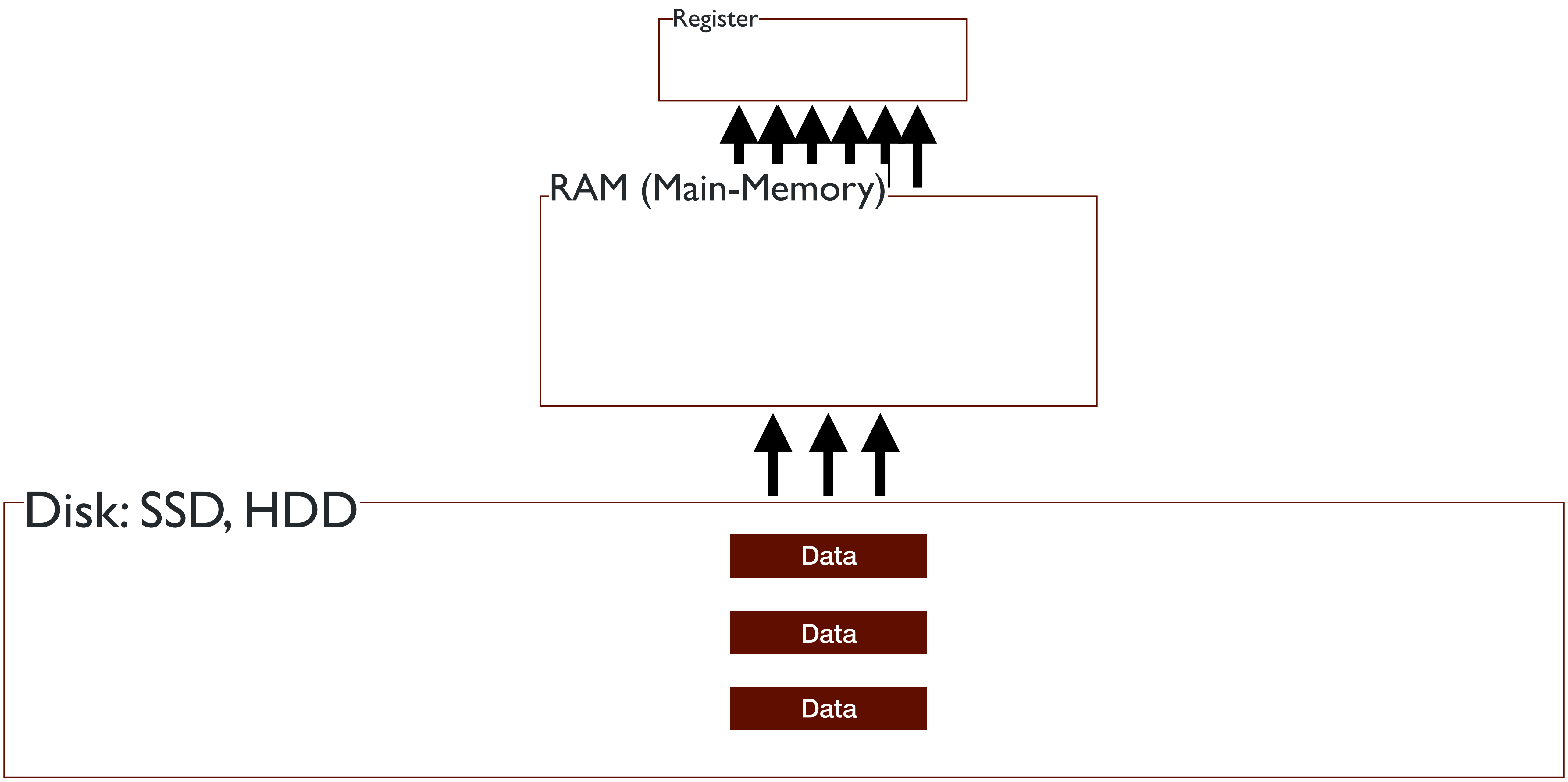
Schematic of Computer Storage



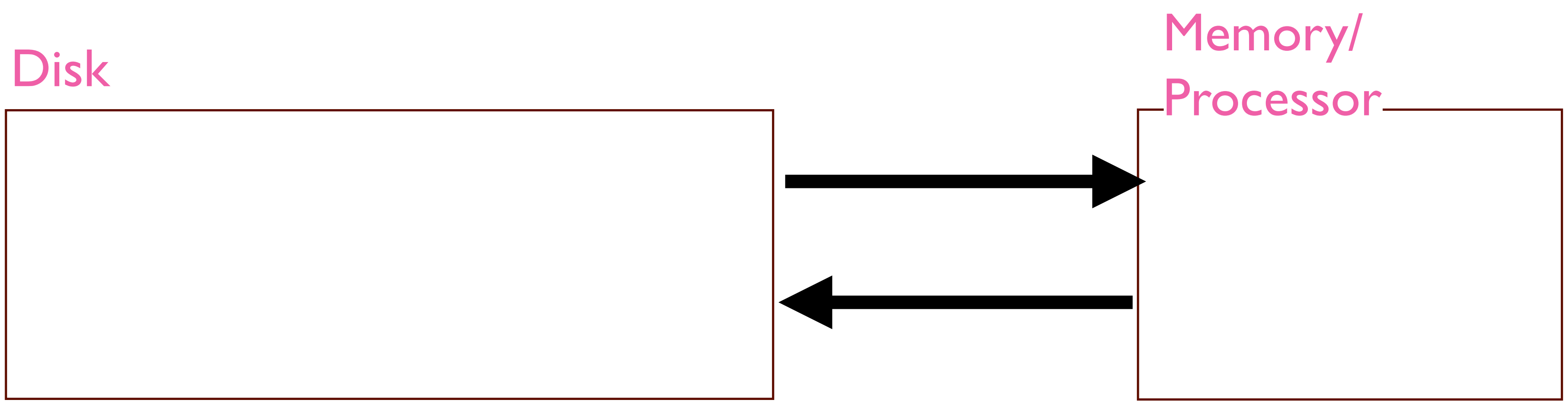
Schematic of Computer Storage



Data Must Move For Analysis



Rate-Limiting Operations

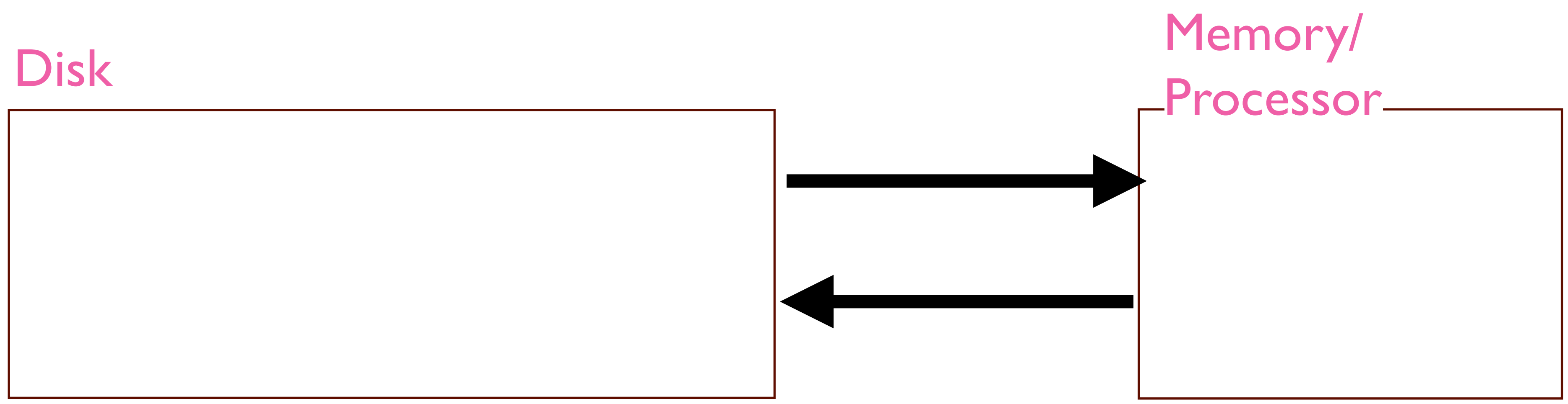


Runtime = I/O Time + Compute Time

"I/O Bound" : I/O Time >> Compute Time

"CPU Bound" : Compute Time >> I/O Time

Rate-Limiting Operations



Given the median of [1,4,5,…,6]

Memory > N/2 => I/O Cost ~ N
Memory < N/2 => I/O Cost ~ N Log N

Effect is called "spilling"

Physical Design

Finding a number in the list that is less than 5 (Operation: Comparison)

Median of a list of numbers

10,000x Slower

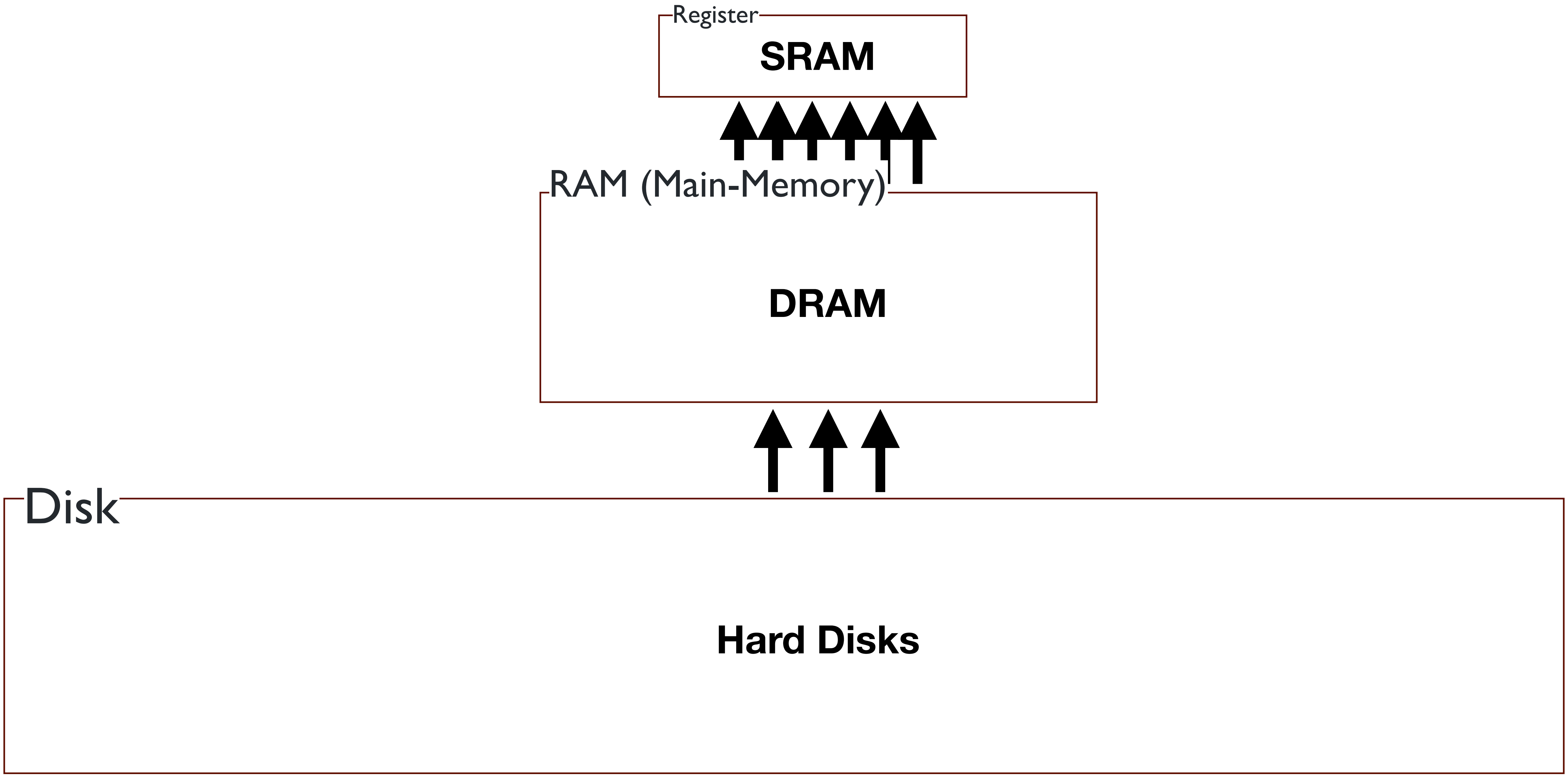
Disk: SSD, HDD

Stored data sorted!

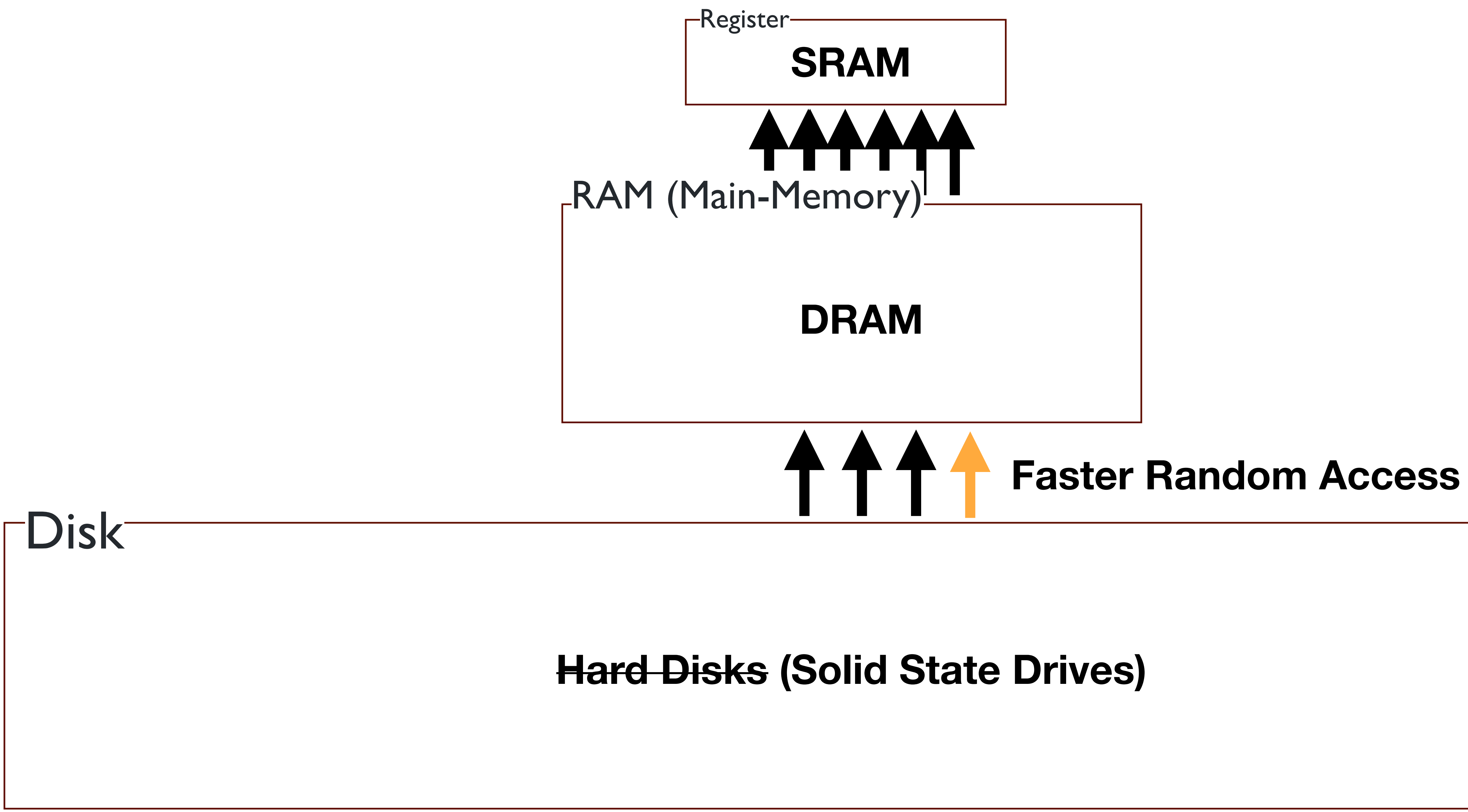
Indexing, Partitioning, and Sorting

Avoid transferring data you don't need

Technologies

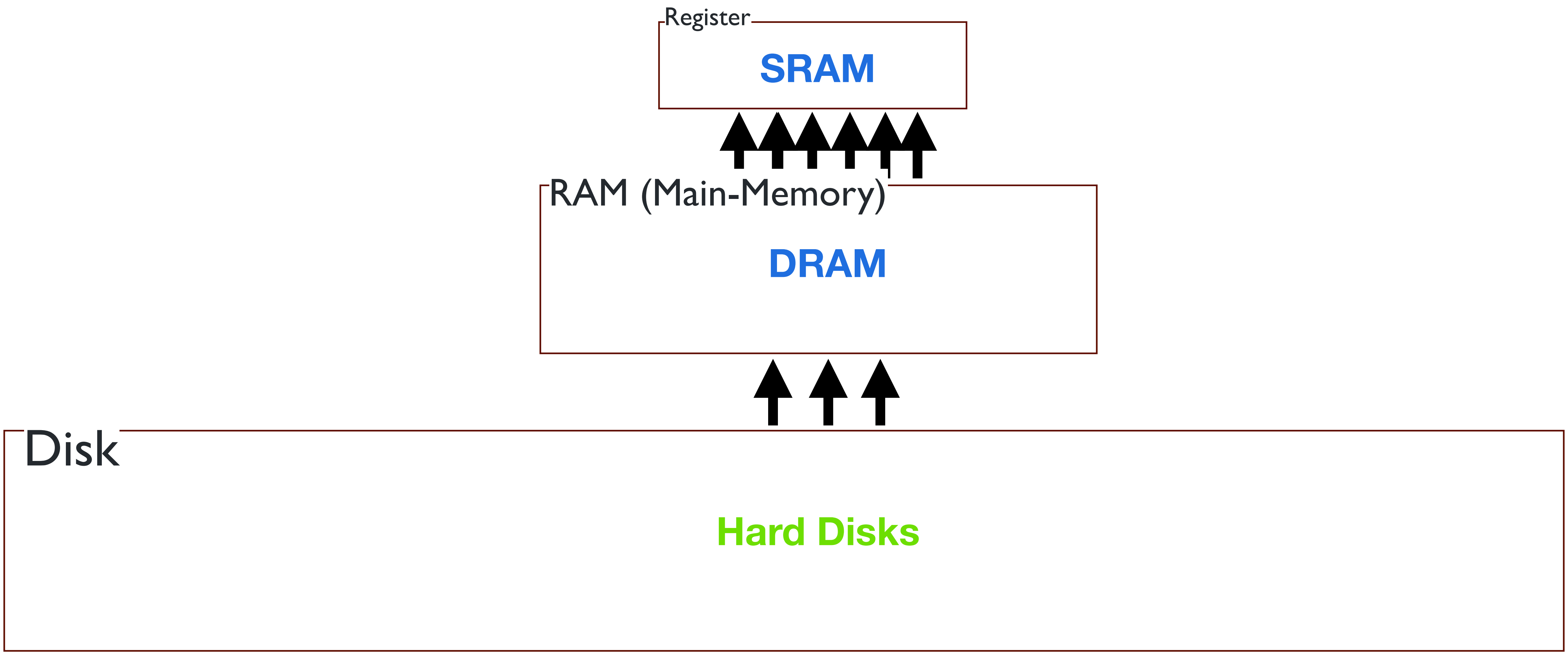


Technologies



Volatile v.s. Non-Volatile

Volatile: Power needs to be applied to store data
Non-Volatile: Data are persistent



Technologies

Register

SRAM

RAM (Main-Memory)

DRAM

$$$$

NVRAM

Non-Volatile RAM (Future…)