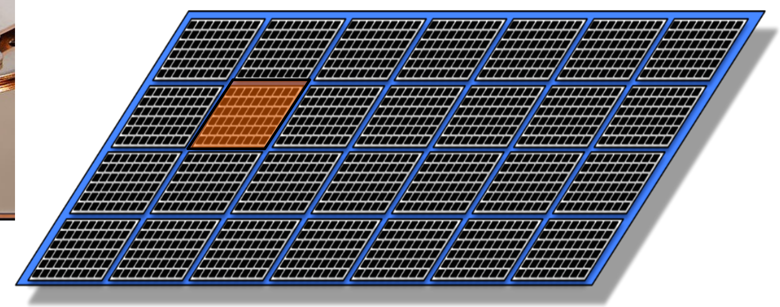
WHERE SITS THE DATA: STORAGE MEDIA

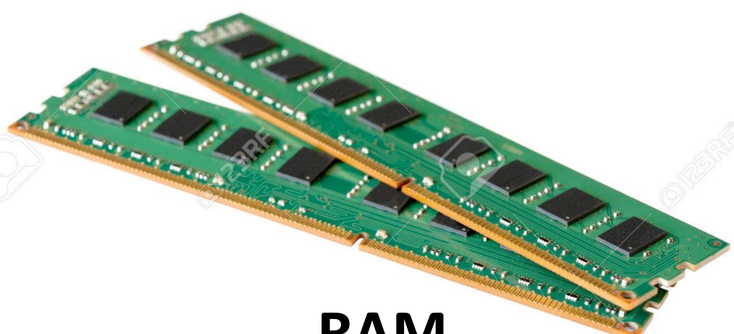
Decades of technology advances

Flash/SSD



HDD

SSD

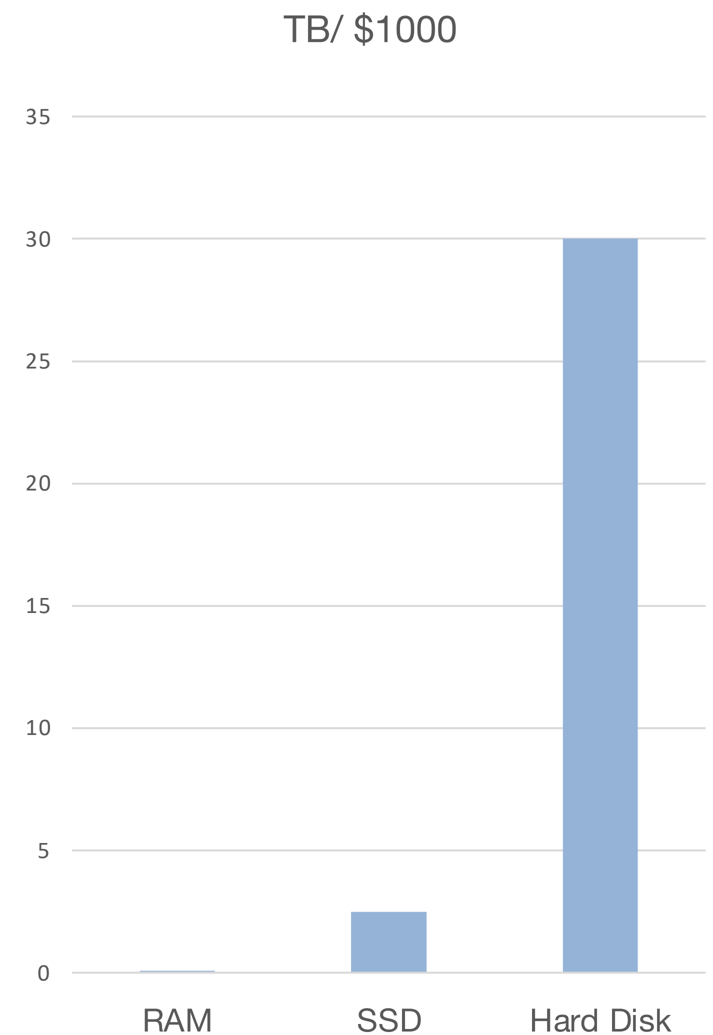


RAM

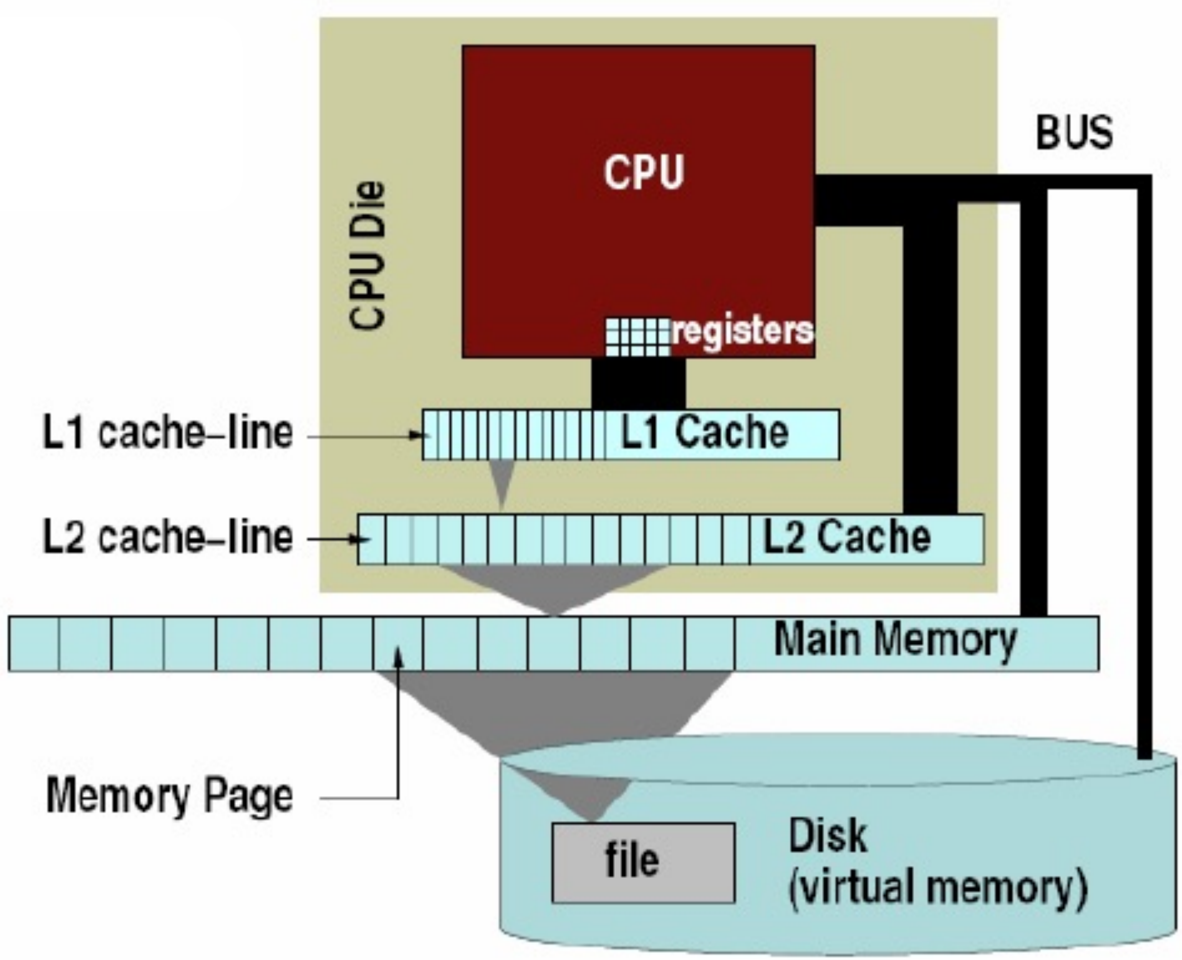
Economics of storage

$1000 can buy:

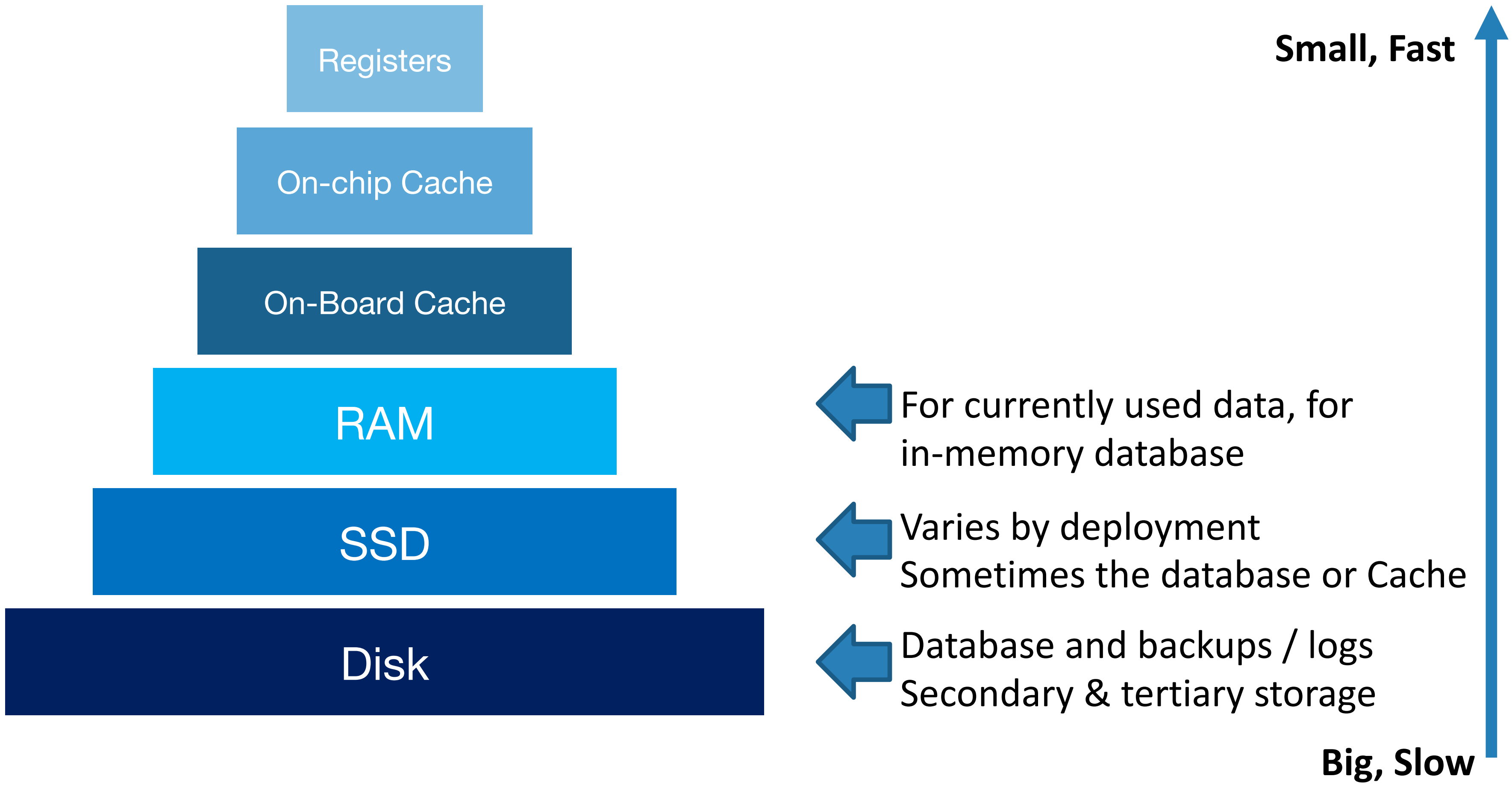
- ~200GB of server-class RAM
- ~2.5TB of enterprise-class SSD
- ~30TB of HDD



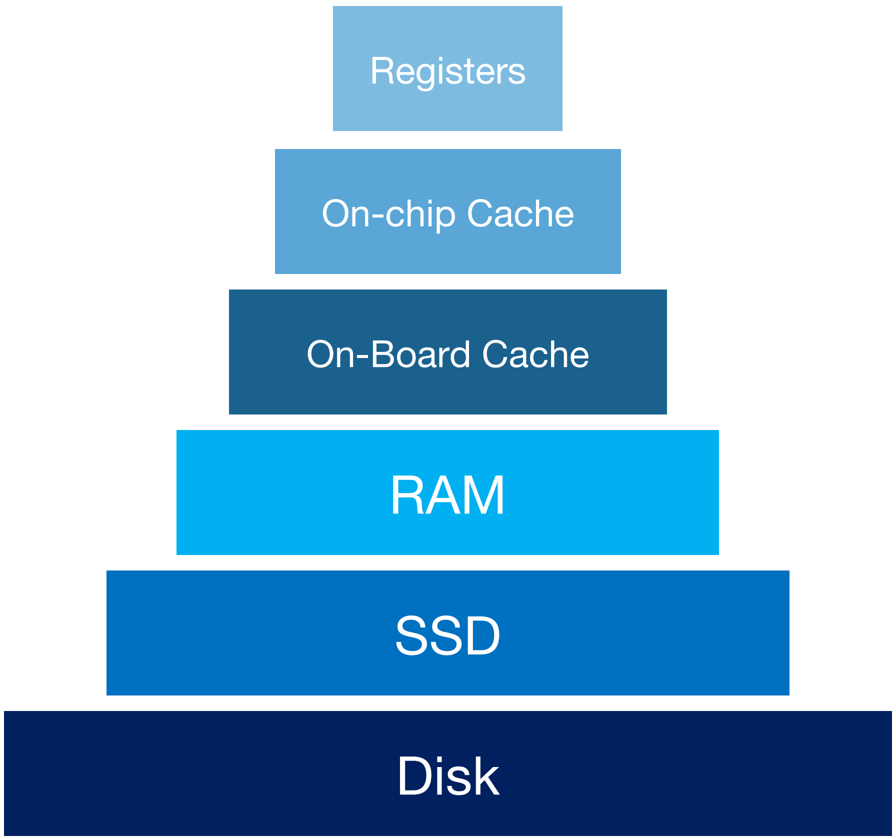
A computer architecture



Storage hierarchy



Storage hierarchy - latencies

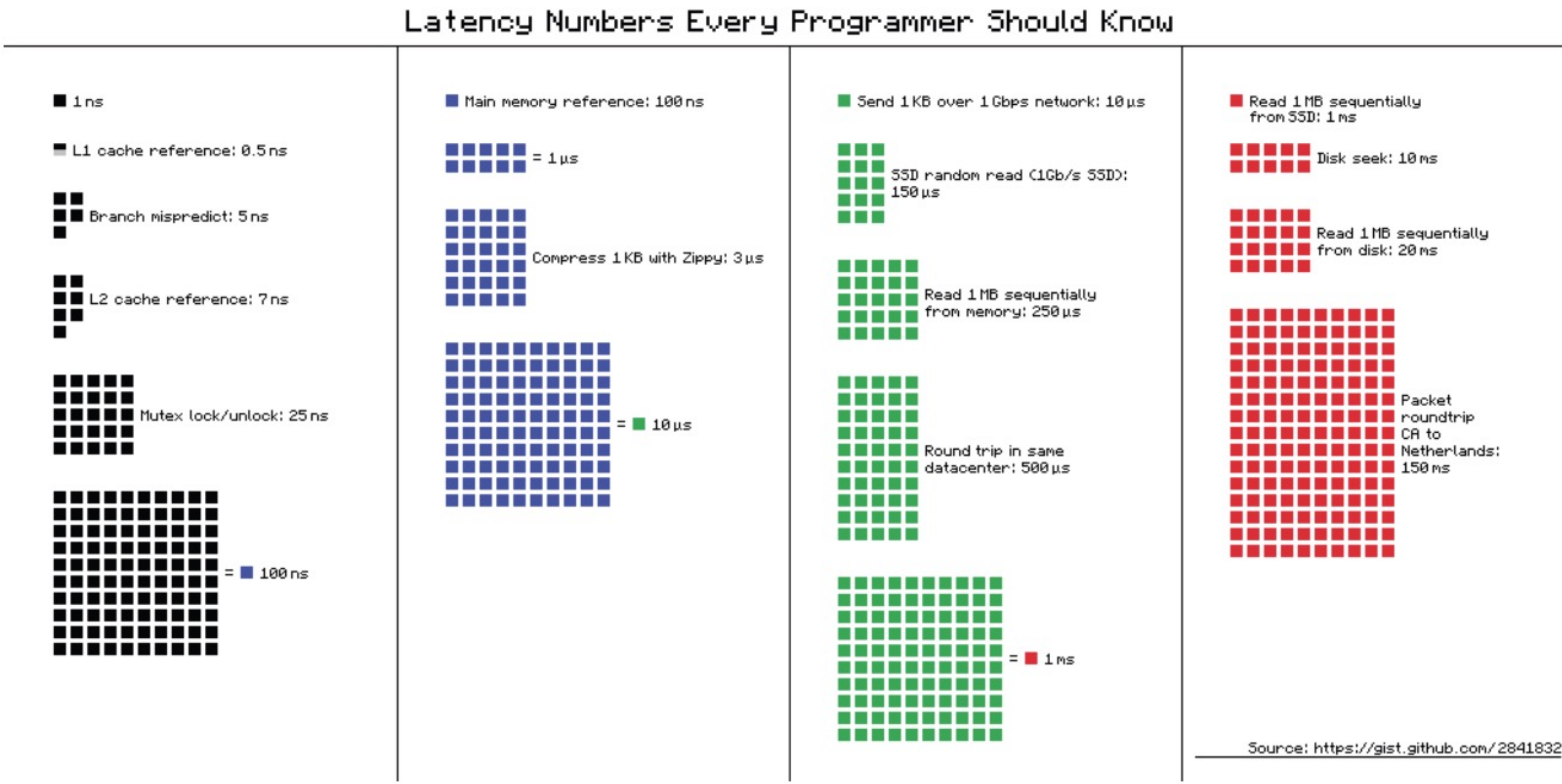


0.5 ns	L1 cache reference
7.0 ns	L2 cache reference
100.0 ns	main memory reference
1,000,000.0 ns	to read 1 MB sequentially
20,000,000.0 ns	to read 1 MB sequentially

Latency numbers every programmer should know

```
L1 cache reference ......................... 0.5 ns
Branch mispredict ............................ 5 ns
L2 cache reference ........................... 7 ns
Mutex lock/unlock ........................... 25 ns
Main memory reference ...................... 100 ns
Compress 1K bytes with Zippy ............. 3,000 ns  =   3 µs
Send 2K bytes over 1 Gbps network ....... 20,000 ns  =  20 µs
SSD random read ........................ 150,000 ns  = 150 µs
Read 1 MB sequentially from memory ..... 250,000 ns  = 250 µs
Round trip within same datacenter ...... 500,000 ns  = 0.5 ms
Read 1 MB sequentially from SSD* ..... 1,000,000 ns  =   1 ms
Disk seek ........................... 10,000,000 ns  =  10 ms
Read 1 MB sequentially from disk .... 20,000,000 ns  =  20 ms
Send packet CA->Netherlands->CA .... 150,000,000 ns  = 150 ms
```

Assuming ~1GB/sec SSD



Visual chart provided by ayshen

Data by Jeff Dean

Originally by Peter Norvig

L1 cache reference	0.5 s	One heart beat (0.5 s)
Branch mispredict	5 s	Yawn
L2 cache reference	7 s	Long yawn
Mutex lock/unlock	25 s	Making a coffee

Hour:

Main memory reference	100 s	Brushing your teeth
Compress 1K bytes with Zippy	50 min	One episode of a TV show (including ad breaks)

Day:

Send 2K bytes over 1 Gbps network	5.5 hr	From lunch to end of work day

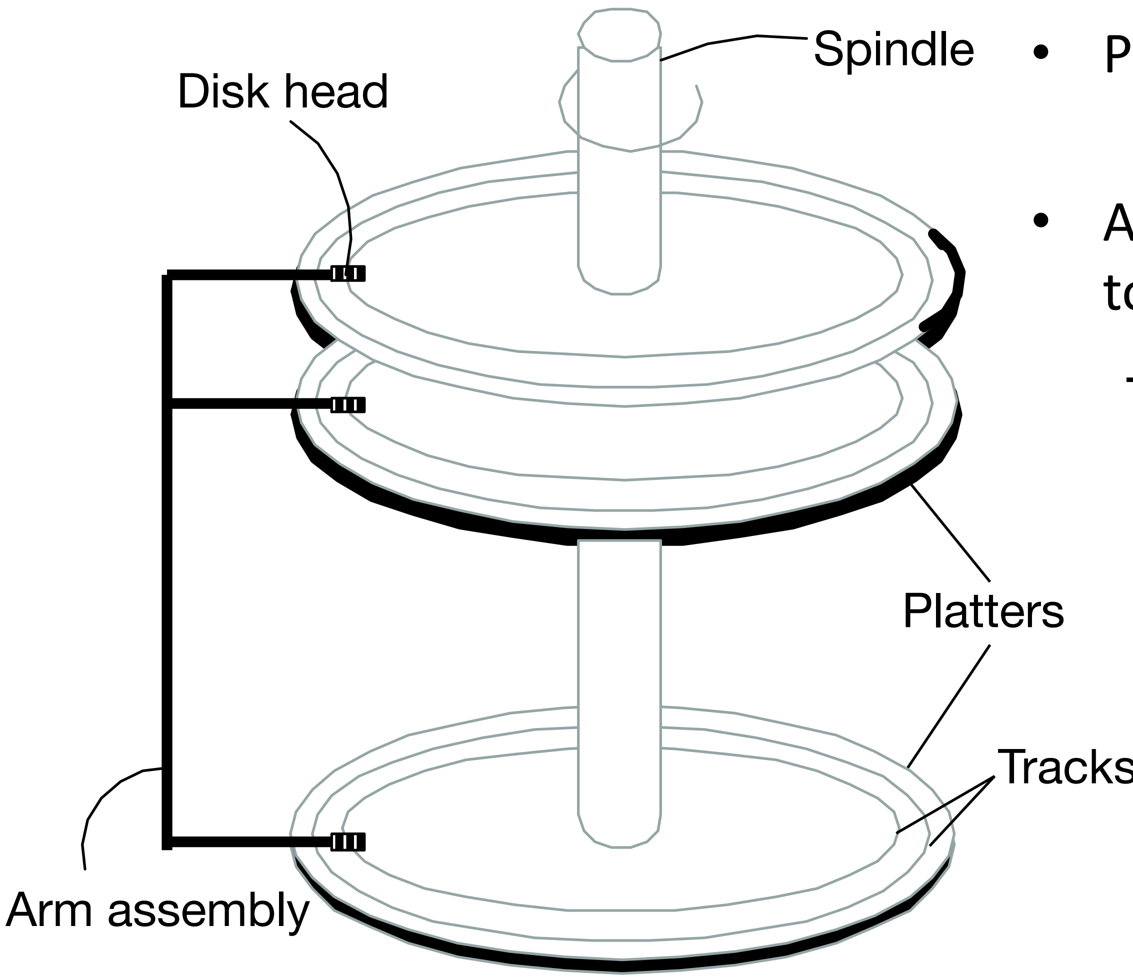
Week

SSD random read	1.7 days	A normal weekend
Read 1 MB sequentially from memory	2.9 days	A long weekend
Round trip within same datacenter	5.8 days	A medium vacation
Read 1 MB sequentially from SSD	11.6 days	Waiting for almost 2 weeks for a delivery

Year

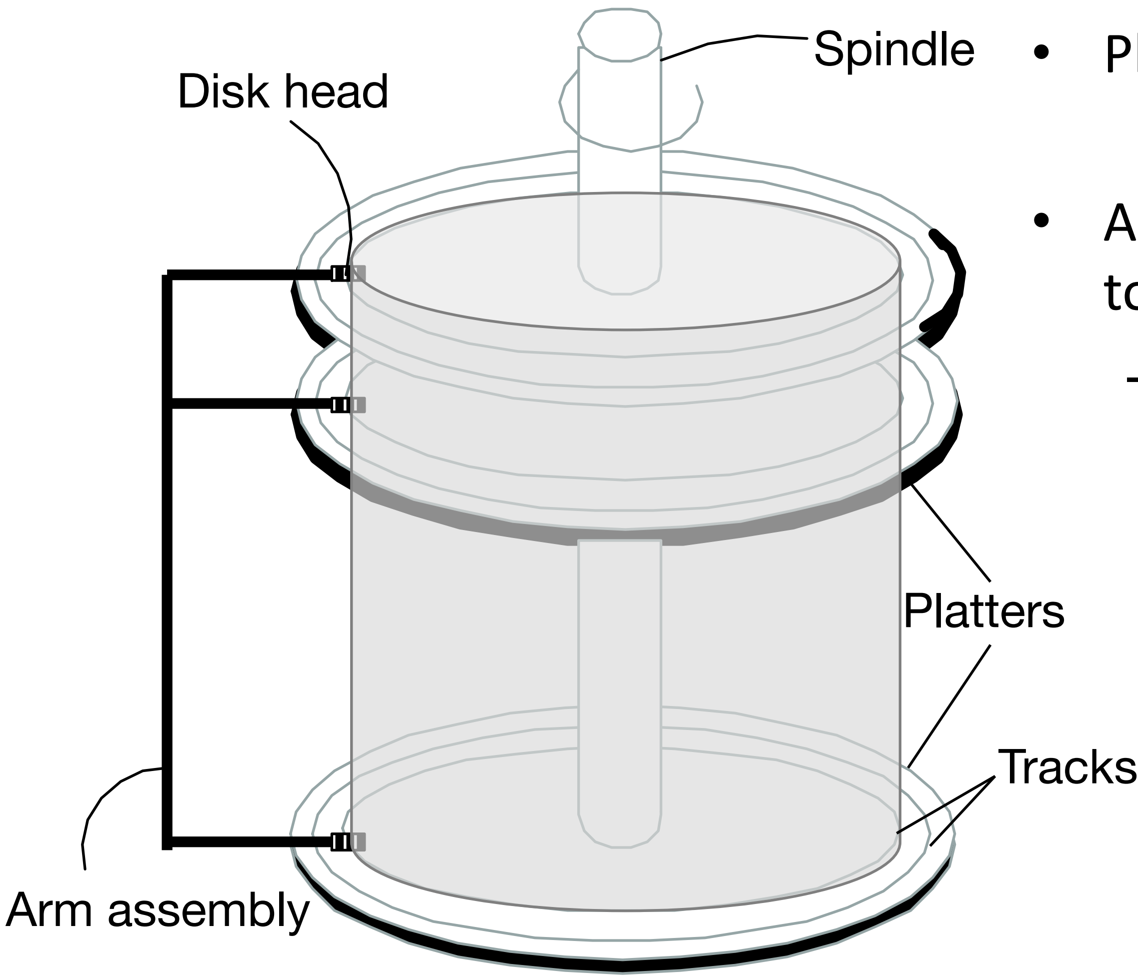
Disk seek	16.5 weeks	A semester in university
Read 1 MB sequentially from disk	7.8 months	Almost producing a new human being
The above 2 together	1 year	

Components of a spinning disk drive (HDD)



- Platters spin (say 15000 rpm)
- Arm assembly moved in / out to place head on desired track
  - Tracks under heads make a “cylinder” (imaginary)

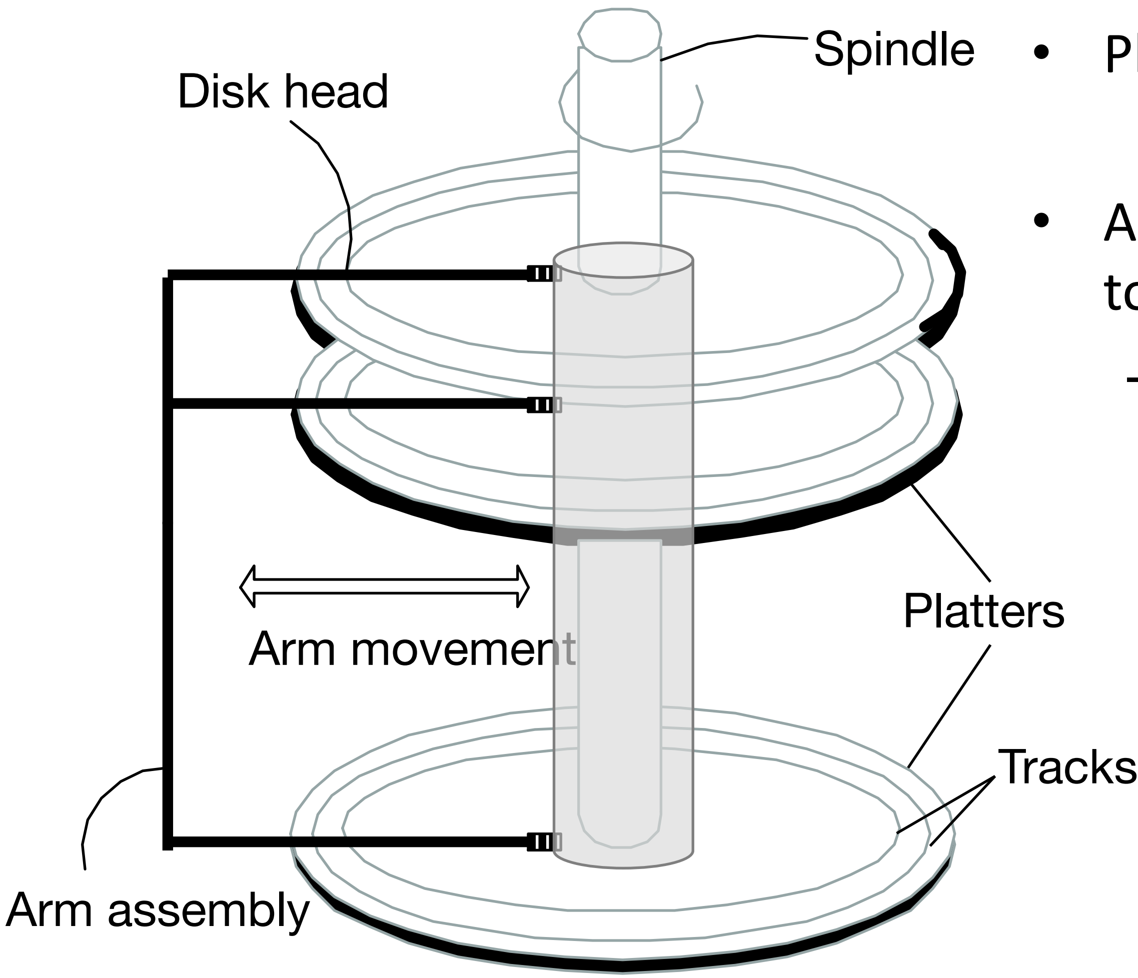
Components of a spinning disk drive (HDD)



- Platters spin (say 15000 rpm)

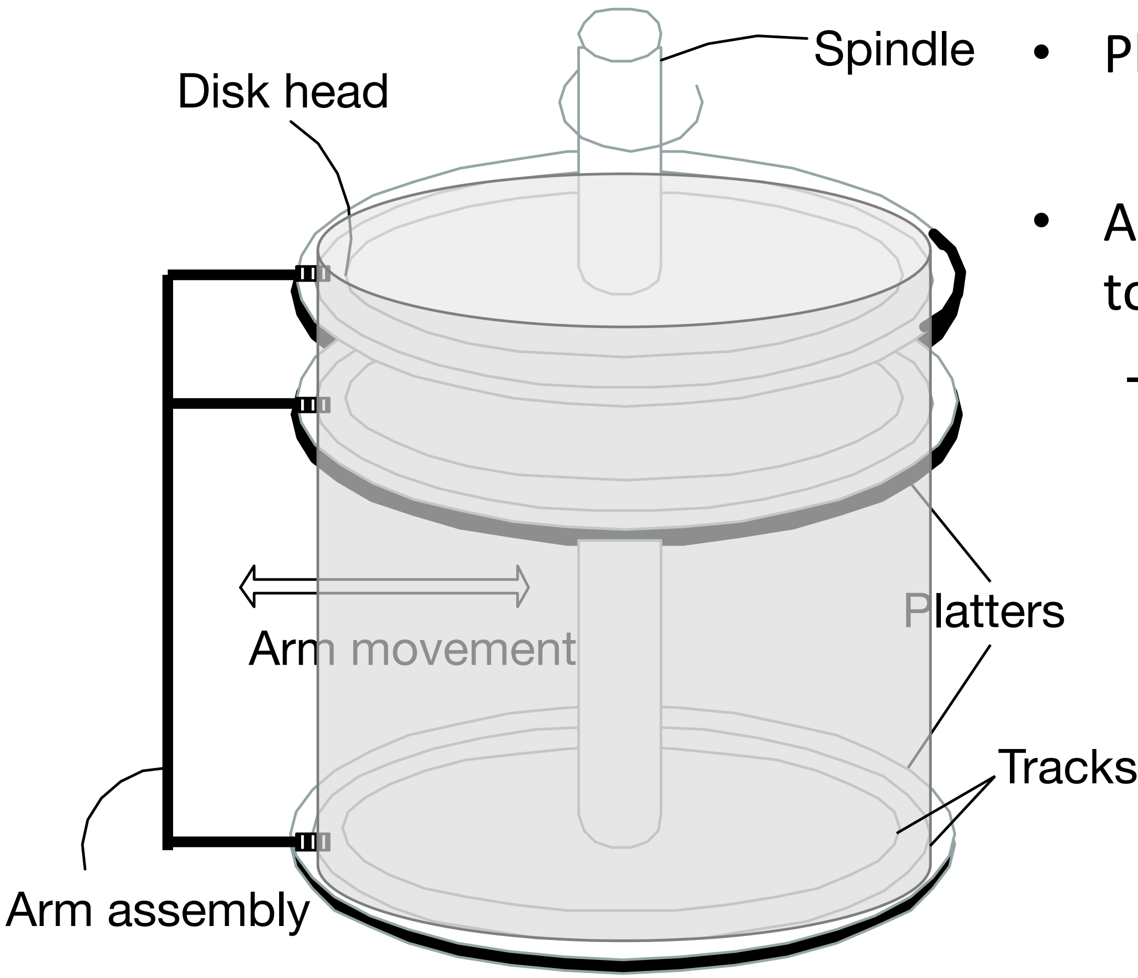
- Arm assembly moved in / out to place head on desired track
  - Tracks under heads make a "cylinder" (imaginary)

Components of a spinning disk drive (HDD)



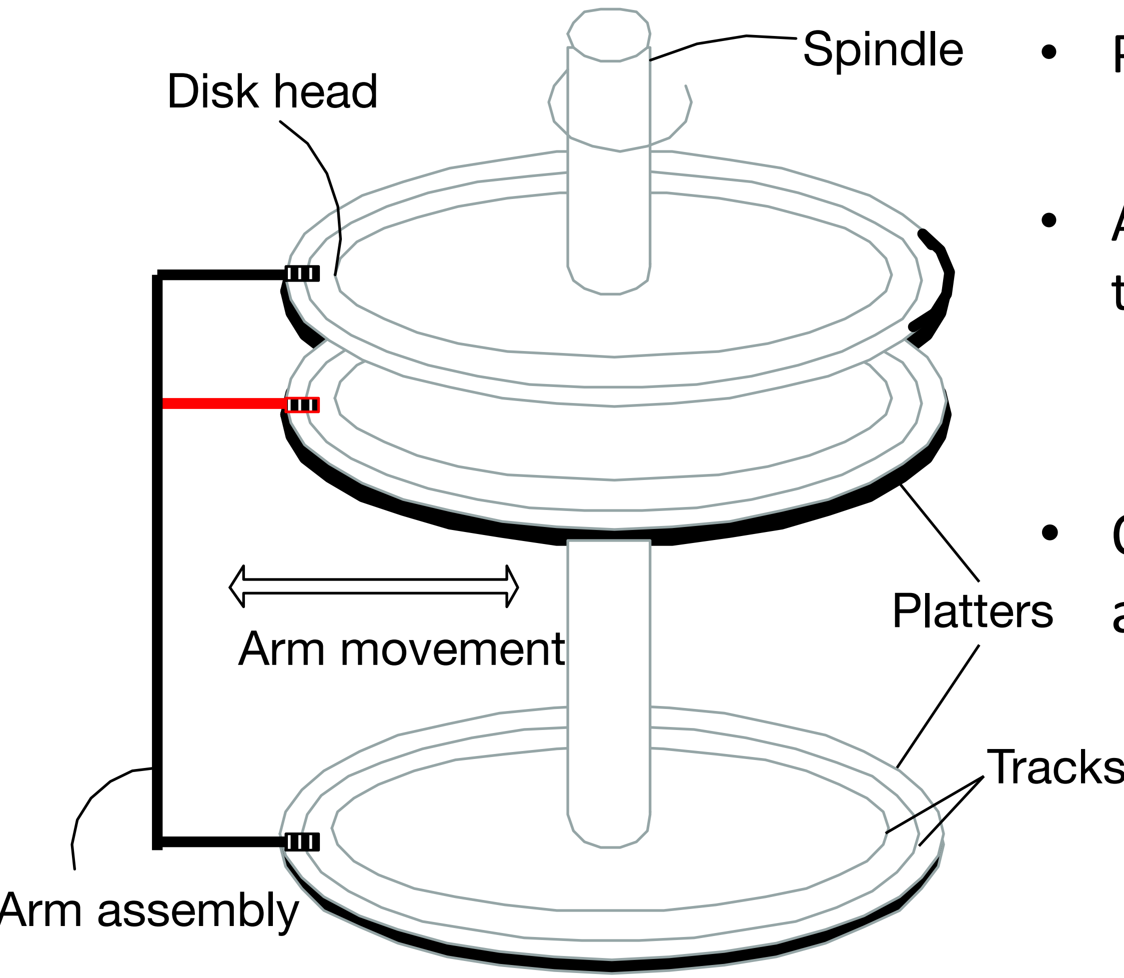
- Platters spin (say 15000 rpm)
- Arm assembly moved in / out to place head on desired track
  - Tracks under heads make a "cylinder" (imaginary)

Components of a spinning disk drive (HDD)



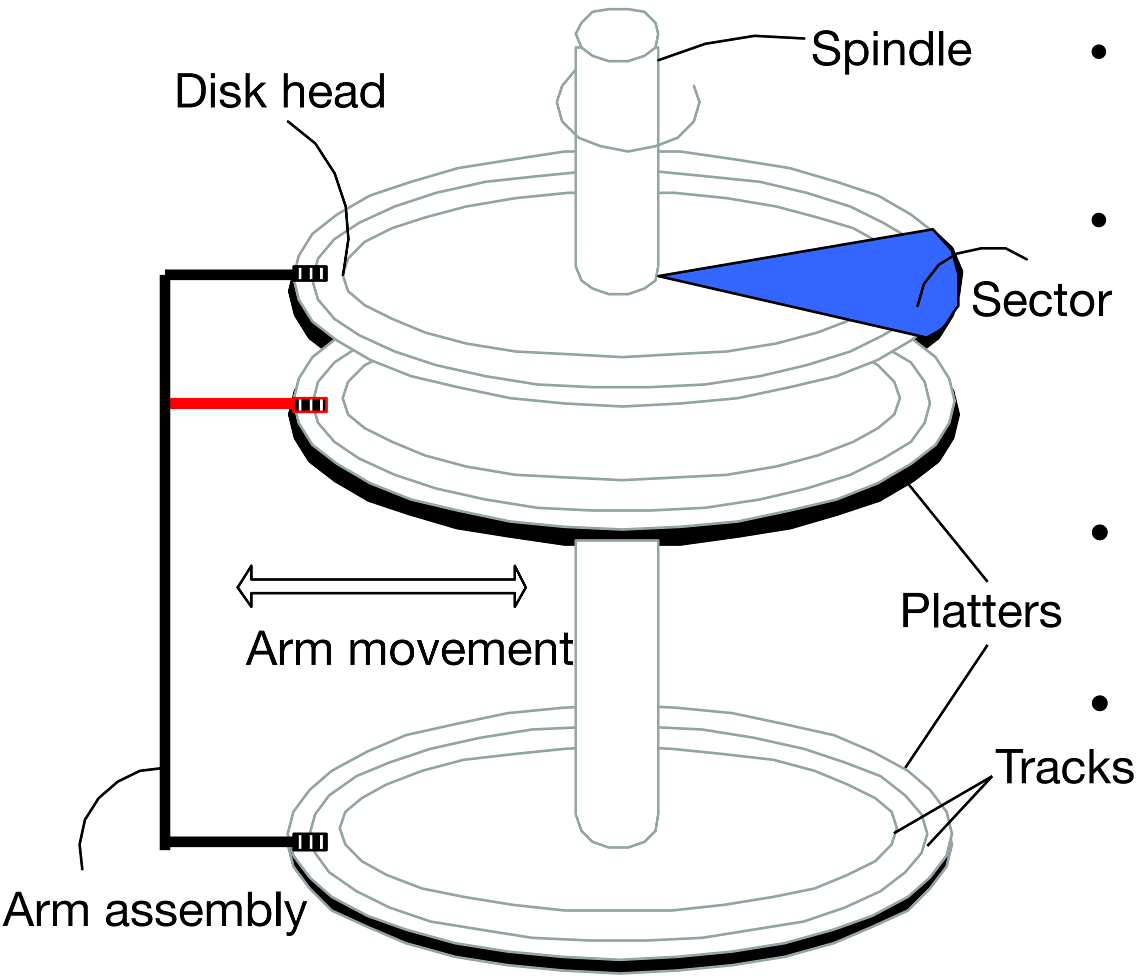
- Platters spin (say 15000 rpm)
- Arm assembly moved in / out to place head on desired track
  - Tracks under heads make a "cylinder" (imaginary)

Components of a spinning disk drive (HDD)



- Platters spin (say 15000 rpm)
- Arm assembly moved in / out to place head on desired track
  - Tracks under heads make a "cylinder" (imaginary)
- Only one head reads/writes at any one time

Components of a spinning disk drive (HDD)



- Platters spin (say 15000 rpm)
- Arm assembly moved in / out to place head on desired track
  - Tracks under heads make a "cylinder" (imaginary)
- Only one head reads/writes at any one time
- Block/page size is a multiple of (fixed) sector size

Accessing a disk page

- Time to access (read/write) a disk block:
  - **seek time** (moving arms to position disk head on track)
    - ~2-3 ms on average
  - **rotational delay** (waiting for block to rotate under head)
    - ~0-4 ms (15000 RPM)
  - **transfer time** (actually moving data to/from disk surface)
    - ~0.25 ms per 64KB page

- Key to lower I/O cost: reduce seek/rotational delays

Conclusion on spinning disks

- A lot of databases still use such disks
  - Spinning disks are a mechanical anachronism!

- Major implications
  - No byte-level access. Instead, an API:
    - READ: transfer “page” of data from disk to RAM
    - WRITE: transfer “page” of data from RAM to disk
  - Both API calls are very, very slow
    - Plan carefully

Quick quizz

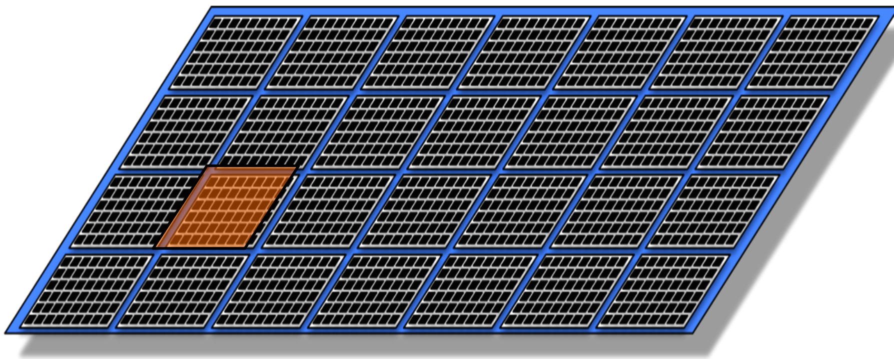
Which of the following operations is slowest ?

A. Access all the blocks found on a single track in order
B. Access an equivalent amount of data but on randomly-located disk blocks
C. Transfer an entire track of data into RAM
D. Overwrite an entire track-worth of data found in RAM with 0s
E. They are all the same

A	B	C	D	E

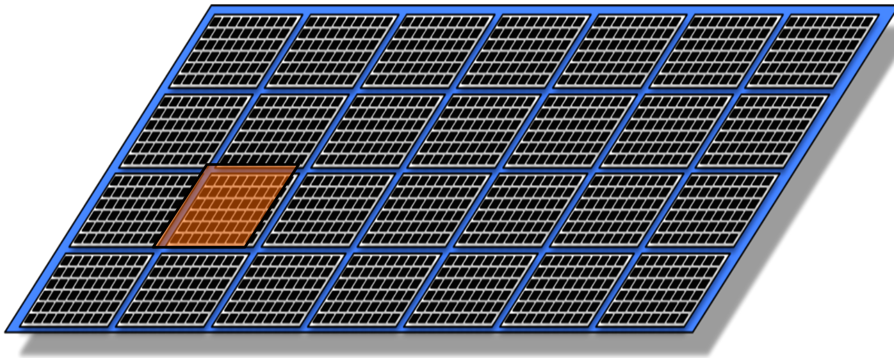
Flash disk (SSD)

- Issues in current generation (NAND)
  - Fine-grain reads (4-8K reads), coarse-grain writes (1-2 MB writes)
  - Only 2k-3k erasures before failure, so keep moving hot write units around ("wear leveling")
  - Write amplification: big units, need to reorganize for wear & garbage collection



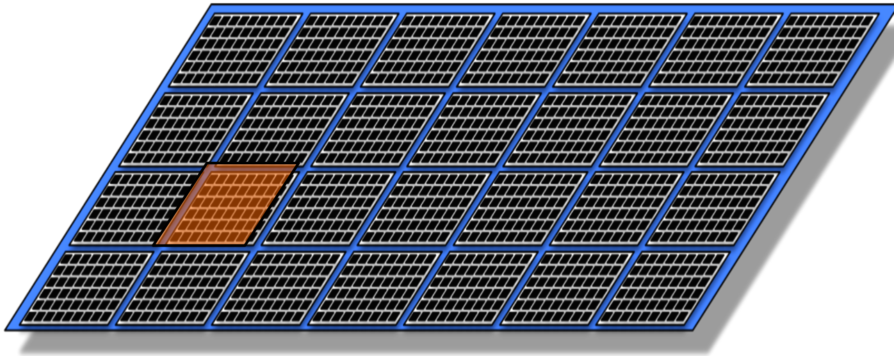
Flash disk (SSD)

- Issues in current generation (NAND)
  – Fine-grain reads (4-8K reads), coarse-grain writes (1-2 MB writes)
  – Only 2k-3k erasures before failure, so keep moving hot write units around ("wear leveling")
  – Write amplification: big units, need to reorganize for wear & garbage collection
- So… read is fast and *predictable*
  – Single read access time: 0.03 ms
  – 4KB random reads: ~500MB/sec
  – Sequential reads: ~525MB/sec
  – 64K: 0.12 ms



Notes on flash disk (SSD)

- Issues in current generation (NAND)
  - Fine-grain reads (4-8K reads), coarse-grain writes (1-2 MB writes)
  - Only 2k-3k erasures before failure, so keep moving hot write units around ("wear leveling")
  - Write amplification: big units, need to reorganize for wear & garbage collection
- So… read is fast and *predictable*
  - Single read access time: 0.03 ms
  - 4KB random reads: ~500MB/sec
  - Sequential reads: ~525MB/sec
  - 64K: 0.12 ms
- But... write is not! Slower for random
  - Single write access time: 0.03 ms
  - 4KB random writes: ~120 MB/sec
  - Sequential writes: ~480 MB/sec



Quick quizz

- Why might SSD writes take a long time
  A. In order to postpone media failure
  B. The transfer size is very large
  C. SSD media suffers from write amplification
  D. Writing a single unit might require writing multiple ones, if we need to move units around
  E. All of the above

A	B	C	D	E

Flash is faster than spinning disk, but…

- SSD is faster than HDD
  - Can be 1-10x the bandwidth (bytes / sec) of ideal HDD
    - Note: Ideal HDD performance hard to achieve. Expect 10-100x bandwidth for non-sequential read
- “Locality” matters for both
  - Reading / writing two “far away” blocks on disk requires slow seek / rotation delay
  - Writing two “far away” blocks on SSD may require writing multiple much larger units
  - High-end flash drives are getting much better at this
- Spinning disk offers about 10x the capacity per $

Storage debate

- Many significant DBs are not really that big
  - Daily weather, round the globe, 1929-2009: 20GB
  - 2000 US Census: 200GB
  - 2009 English Wikipedia: 14GB
- But data sizes grow faster than Moore's Law
  - "Big Data" is real: can generate & archive data cheaply and easily
    - Boeing 787 generates ½ TB of data per flight
    - Walmart handles 1M transactions/hour, maintains 2.5 PB data warehouse
- So… what is the role of disk, flash, RAM ?
  - The subject of some debate !

Bottom line

- Very large DBs: relatively traditional
  - Disk still the best cost / MB by a lot
  - SSDs improve performance and *performance variance*
  - Both still have good price per performance tradeoff

- Smaller DB story is changing quickly
  - Entry cost for disk is not cheap, so flash wins at the low end
  - Many interesting databases may even fit in RAM !
  - Emergence of persistent memory ?