Data Center Technologies

Networking slides, h/t: Vincent Liu

Disk slides, h/t: Garth Gibson

Cloud Computing is Everywhere

Google

amazon

facebook

Google docs

Office 365

Spotify

Cloud Computing is Everywhere

Google

amazon

HESS

Kellogg's

SUNCORP BANK

vodafone

Expedia

DOW JONES

NTT docomo

NOVARTIS

CONDÉ NAST

SIEMENS

Adobe

COMCAST

Spotify

Cloud Computing Benefits

- Elastic
  - Scale up & down based on demand
- Multi-tenancy
  - Multiple independent users share infrastructure
  - Security and resource isolation
  - SLAs on performance & reliability (sometimes)
- Dynamic Management
  - Resiliency: isolate failure of servers and storage
  - Workload movement: move work to other locations

Cloud Service Models

- Software as a Service
  - Provider licenses applications to users as a service
  - E.g., customer relationship management, e-mail, ..
  - Avoid costs of installation, maintenance, patches, …

- Platform as a Service
  - Provider offers platform for building applications
  - E.g., Google's App-Engine
  - Avoid worrying about scalability of platform

Cloud Service Models

- Infrastructure as a Service
  - Provider offers raw computing, storage, and network
  - E.g., Amazon's Elastic Computing Cloud (EC2)
  - Avoid buying servers and estimating resource needs

The Result: Data Centers

Microsoft

Google

Data Centers Are Big

10-100K servers

100s of Petabytes of storage

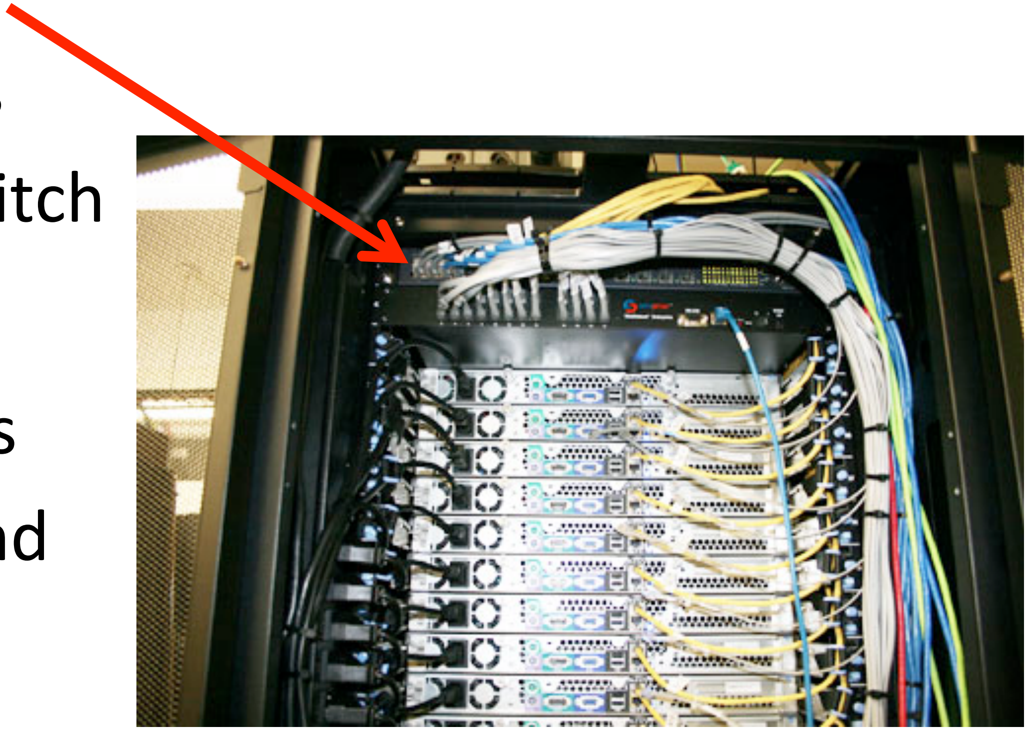
100s of Terabits/s of Bw
(more than core of Internet)

10-100MW of power
(1-2 % of global energy consumption)

100s of millions of dollars

Servers in Racks

- Rack of servers
  - – Commodity servers
  - – And top-of-rack switch
- Modular design
  - – Preconfigured racks
  - – Power, network, and storage cabling

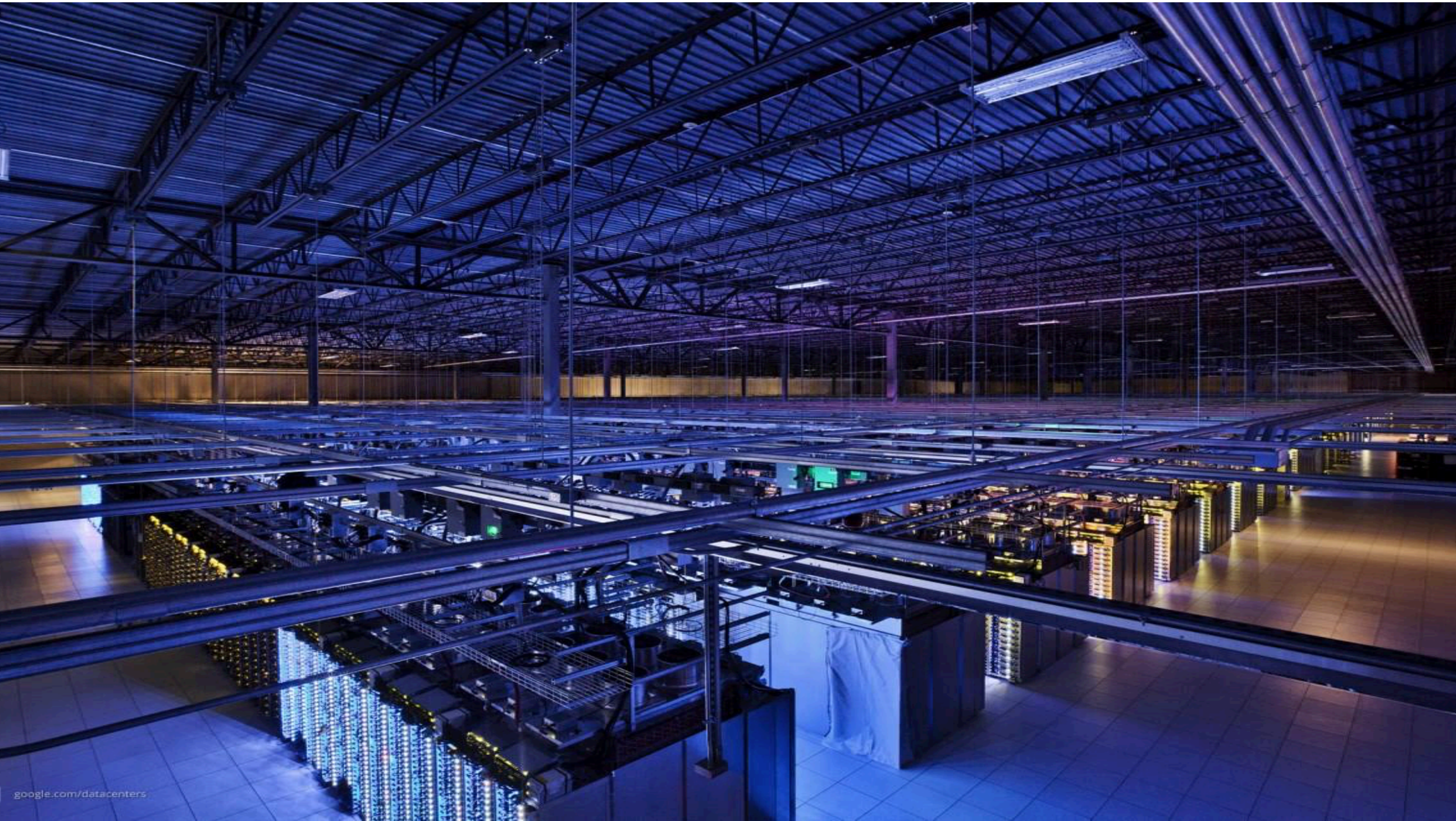


Racks in Rows

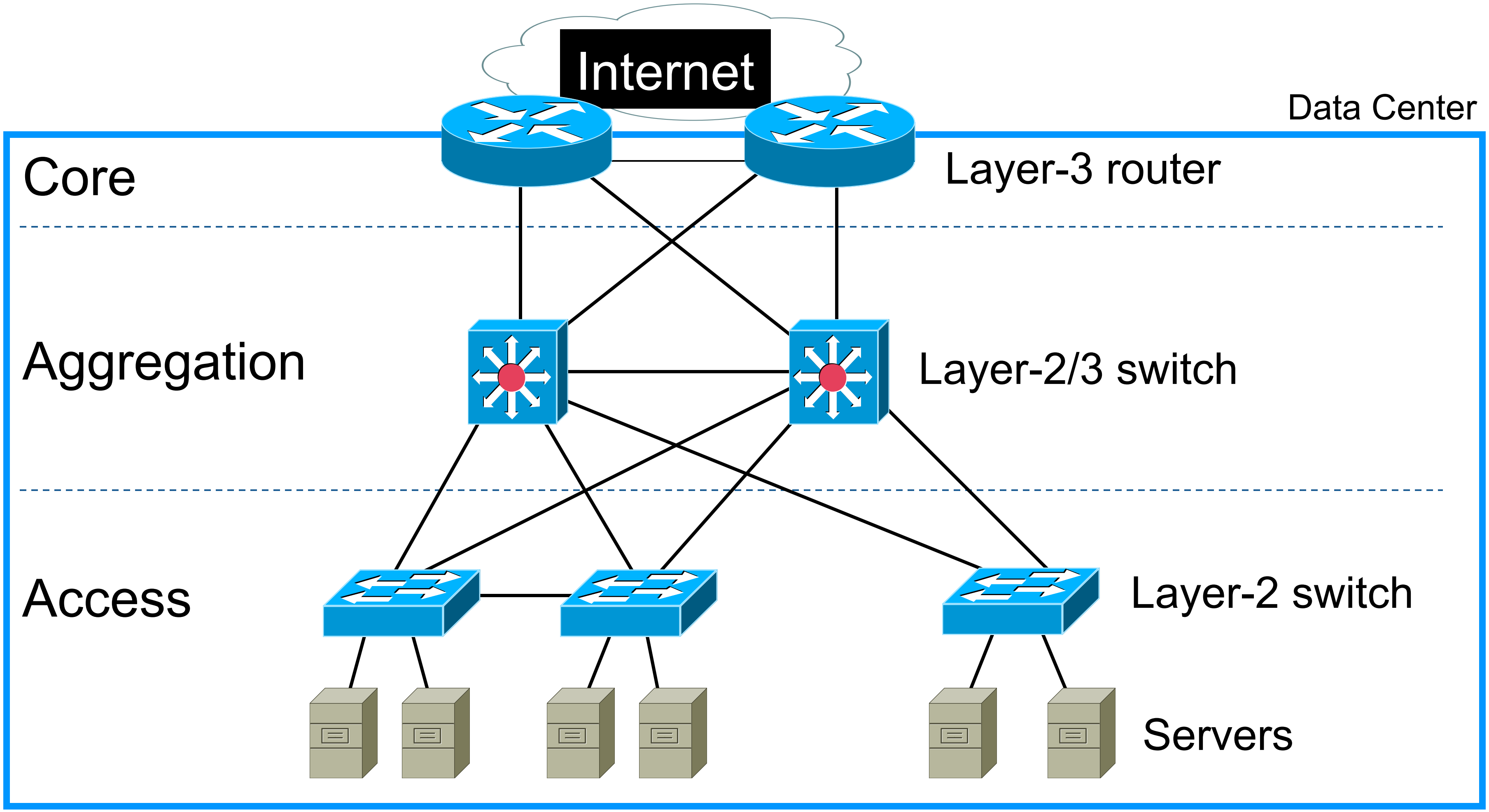
Rows in Hot/Cold Pairs

google.com/datacenters

Hot/Cold Pairs in Data Centers



Early Data Center Networks



Internet

Data Center

Core

Aggregation

Access

Layer-3 router

Layer-2/3 switch

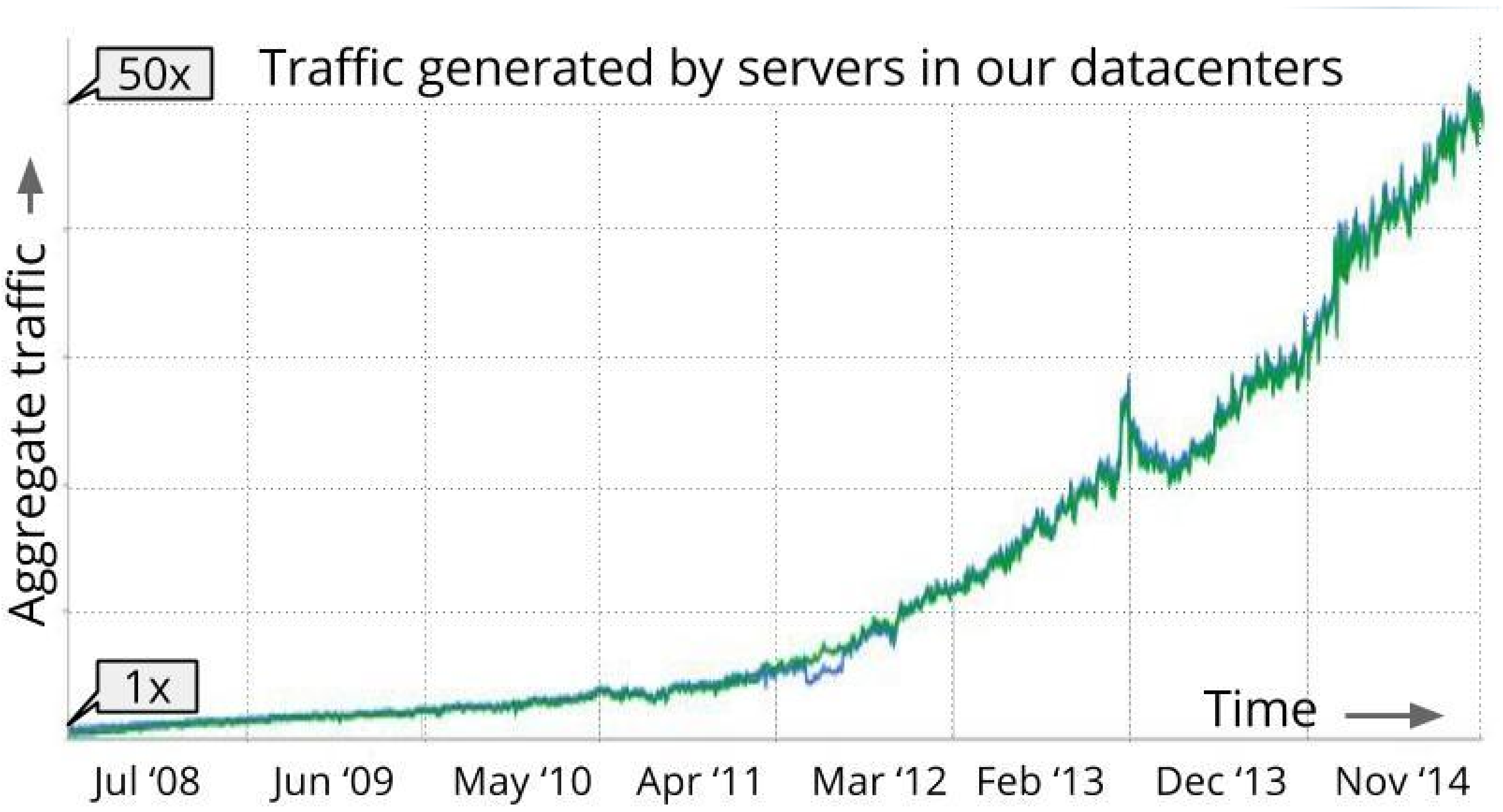
Layer-2 switch

Servers

Problems with Early DC Networks

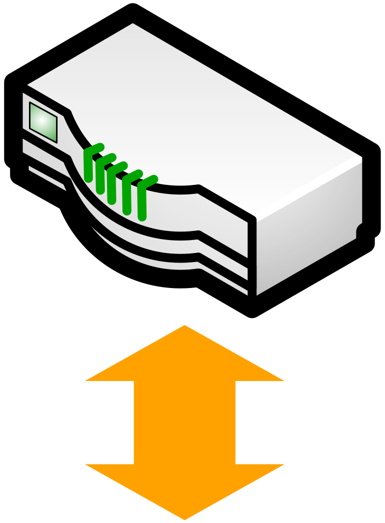
- Cost
  - Core and aggregation routers were high capacity and low volume => expensive
- Fault tolerance
  - Failures of core and aggregation routers cause substantial decrease in network capacity
- Bisection bandwidth across the data center limited by capacity of largest available routers

Data Center Traffic Growth



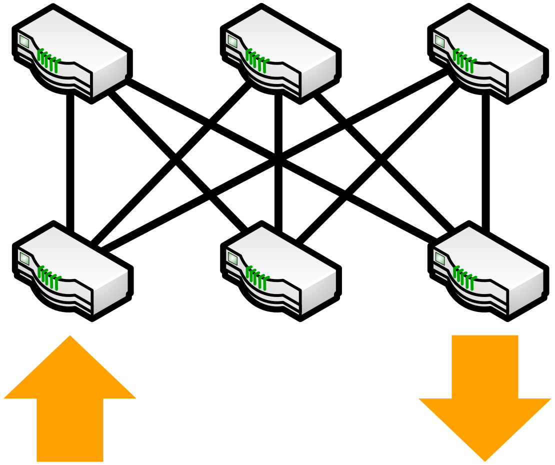
- ✧ Source: "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network", SIGCOMM 2015.

History Lesson: Clos Networks (1953)



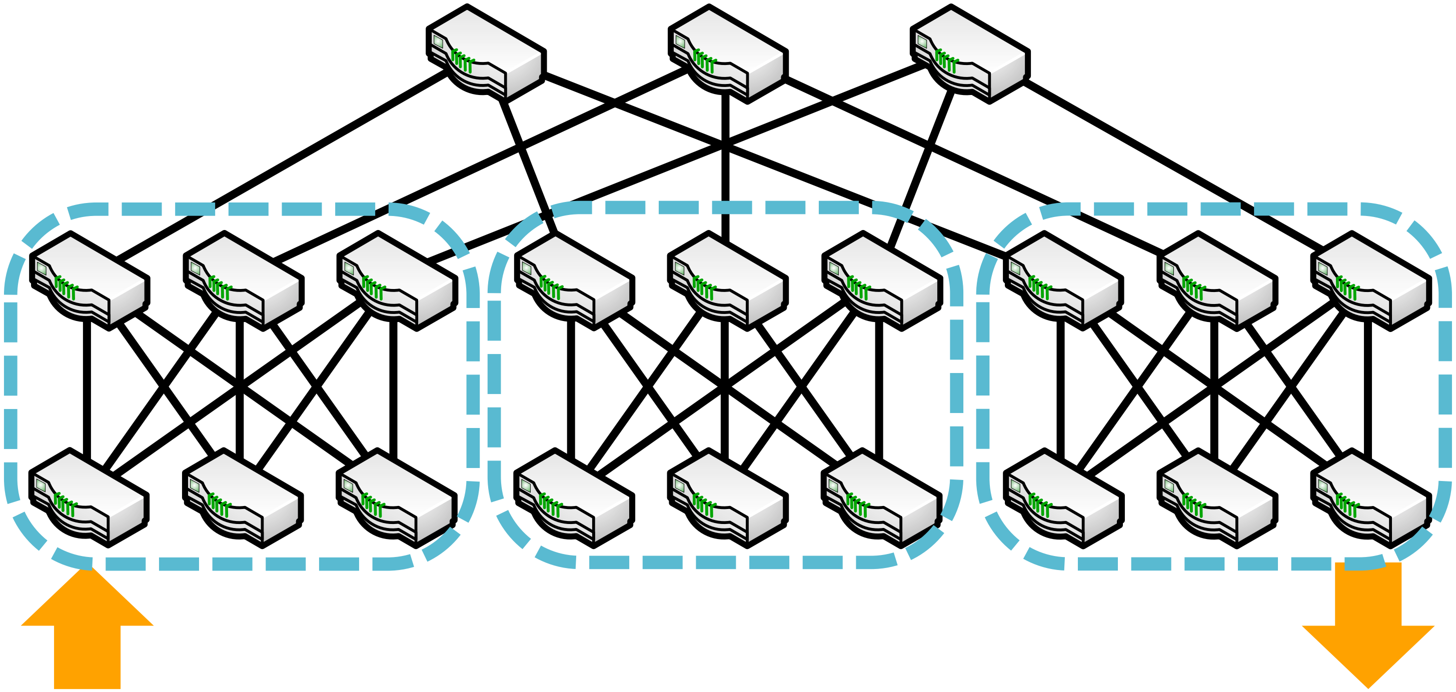
- Emulate a single huge switch with many smaller switches

History Lesson: Clos Networks (1953)



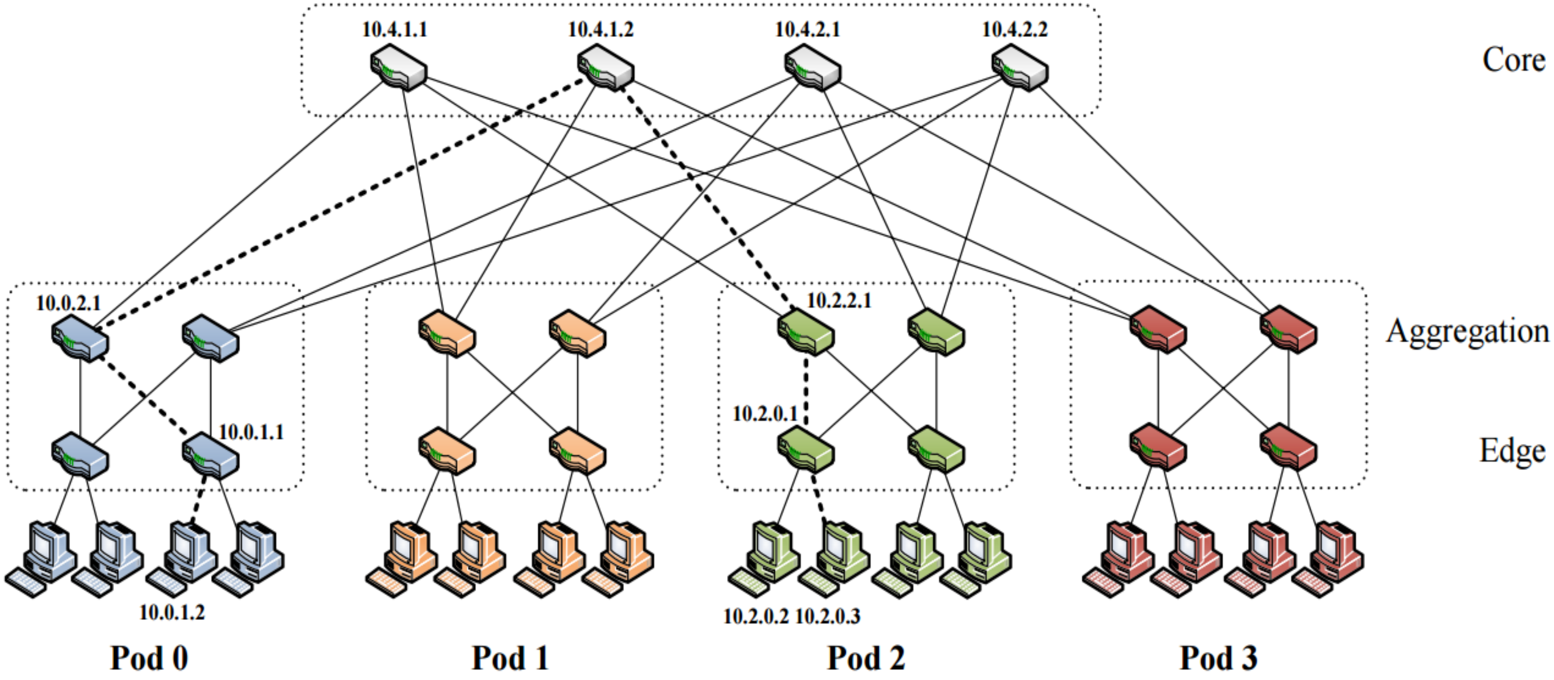
- Emulate a single huge switch with many smaller switches

History Lesson: Clos Networks (1953)



- Emulate a single huge switch with many smaller switches
- Add more layers to scale out

Fat-tree Architecture



Bandwidth oversubscription: thin fat tree at higher levels to reduce cost

Data center networking

- Each physical servers assigned a fixed intranet IP address
  - Ex: 10.0.0.1
- Network address translation to reach virtual machine
  - Migration transparent to network
  - Physical network address invisible to guest OS
- Routing lookup ~ # of data center racks
  - All servers in a rack in same subnet

Multipath Routing

- Lots of available bandwidth, but split across many paths
- TCP dynamics, OS packet handling easier if packets arrive in order
  - In a connection between any pair of servers
- ECMP: hash on packet header to determine route
  - Same for all packets between any pair of servers
  - But: hash collisions, failures, diagnostics, …

Data centers in practice

- End of Dennard scaling
  - Moore's Law: more transistors per chip each year
  - Clock rates decoupled from transistor density
  - # of cores growing slowly (2x/5y for cost-effic configs)
  - power dissipation limits chip density
- Network link bandwidths still scaling
  - 40Gbs server links common, 100Gbps on the way
  - With cut-through, 10-100us latency across DC
- Applications, services scale out across the DC
  - Disaggregated storage, memory

When is data persistent?

- On a single node:
  - In local persistent storage?
  - Many storage devices have DRAM write buffers…
- In a data center:
  - In persistent store on one server?
  - In DRAM on multiple servers?
  - In persistent store on multiple servers?
- Across data centers:
  - In DRAM on a server in multiple data centers?
  - In DRAM on multiple servers in multiple DCs?

Storage Technologies

- Cost/capacity
- Word vs. block access
- Persistence
- Latency (read/write)
- Throughput
- Power drain (in use or when inactive)
- Weight/volume

Volatile Memory: SRAM

- Static RAM (SRAM)
  - Data stored in a transistor flip/flop
  - Bits degrade on poweroff
  - Access latency range: 1 – 10ns
  - Bit density inversely proportional to clock rate
  - Bit density scales with Moore's Law
  - Typical use: on chip cache, high speed access

Volatile Memory: DRAM

- Dynamic RAM (DRAM)
  - Each bit stored in a capacitor
  - 2D/3D array for dense packing
  - 50-100 ns latency for word-level access
  - Bits degrade even when powered, so must be actively refreshed
  - Power drain proportional to storage capacity
  - Bit density scales with Moore's Law
  - Typical use: off-chip volatile random access

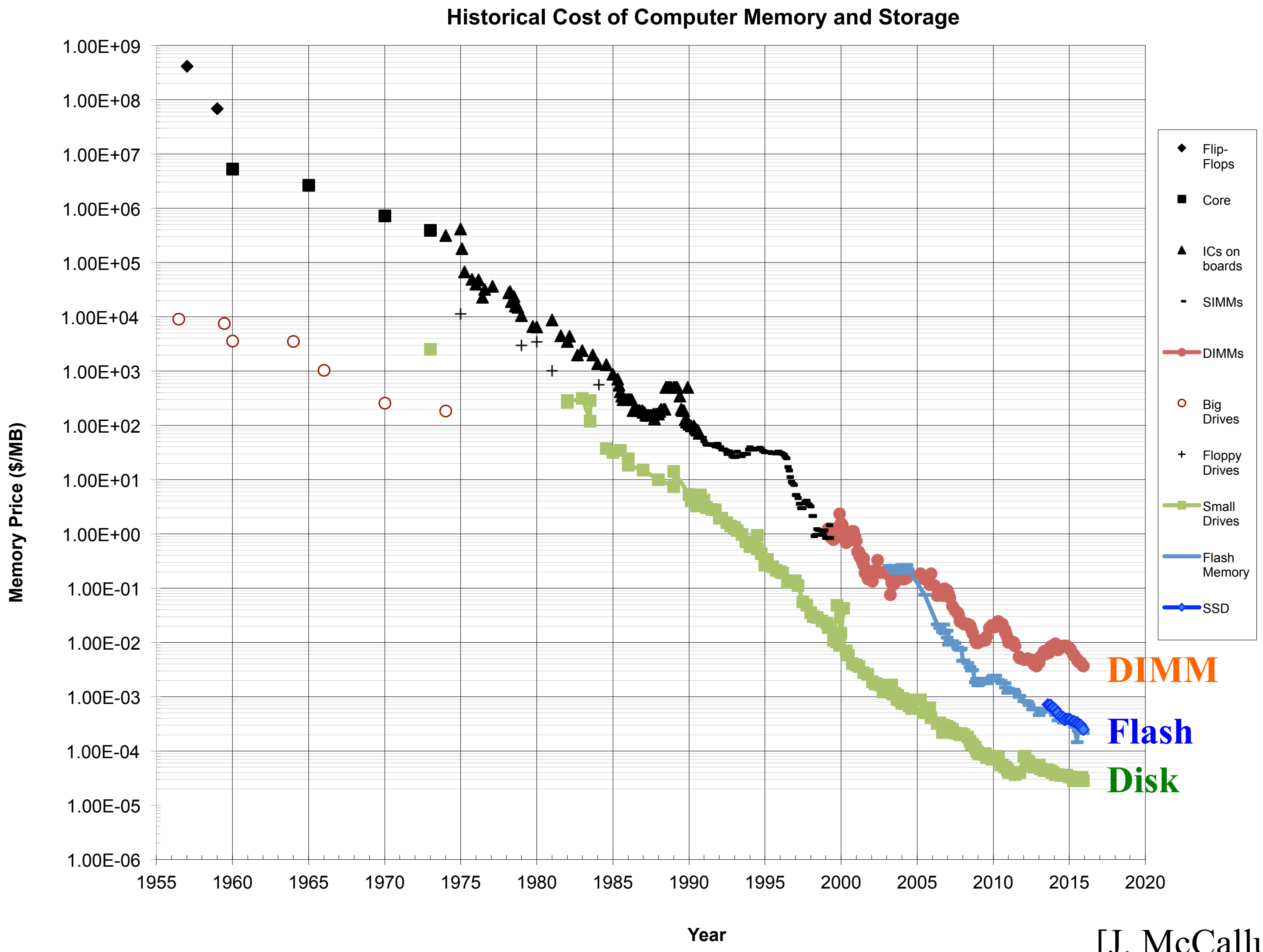
Persistent Memory: Flash

- NAND Flash/Solid State Drive (SSD)
  - Blocks of bits stored persistently in silicon
  - Densely packed in 2-D or 3-D array
  - Blocks remain valid even when unpowered
  - Electrically reprogrammable, for a limited # of times
  - 10-50us block level random read/write
  - Writes must be to a “clean” block, no update in place
  - Erasing only for regions of blocks ~ 256KB
  - Typical use: smartphones, laptops, cloud servers

Persistent Memory: Magnetic Storage

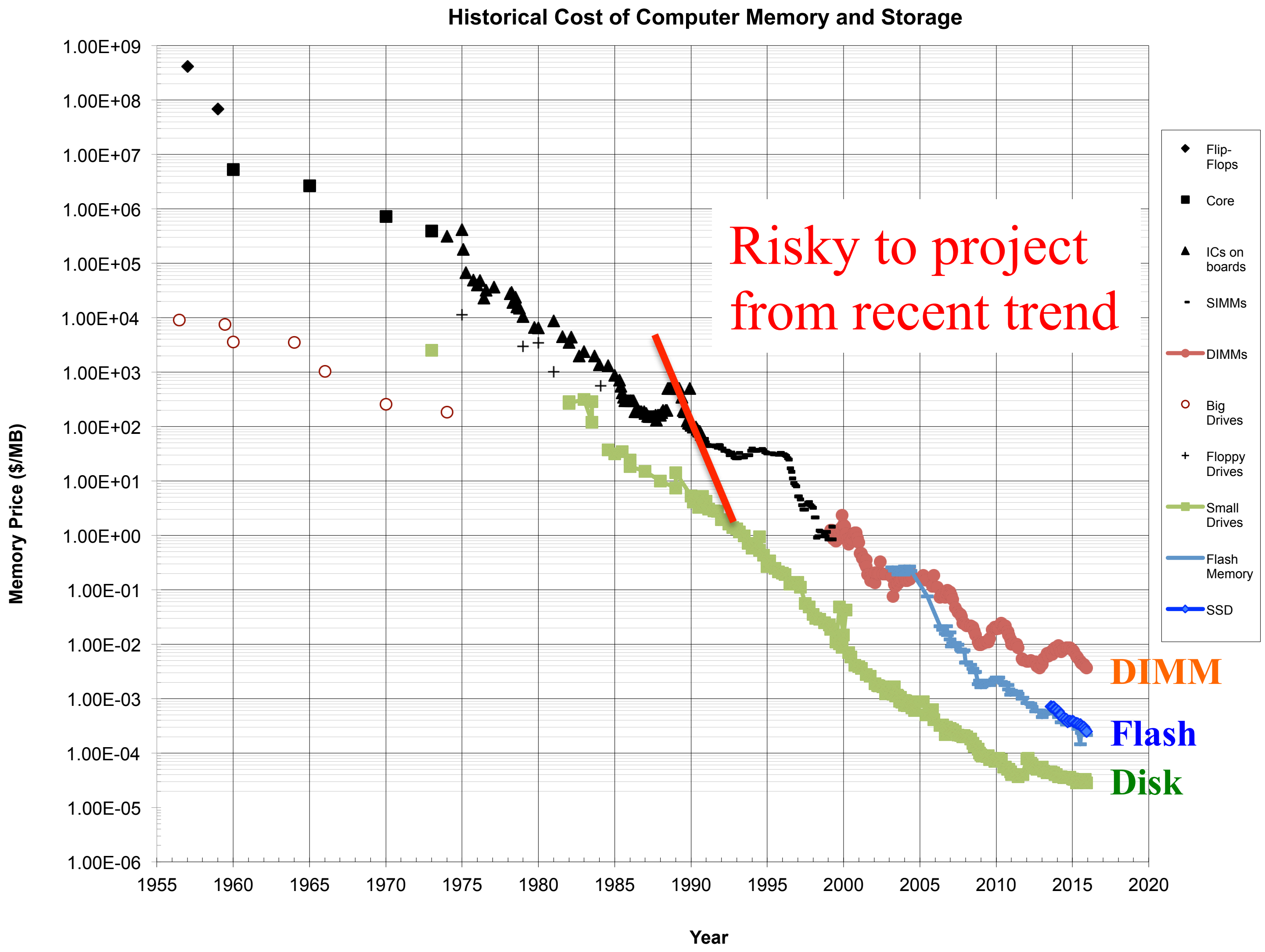
- Bits stored on magnetic surface
  - 1 Tbit per square inch
  - Physical motion needed to read bits off surface
- Magnetic disks
  - Block level random access
  - 10 ms random access latency
  - 150MB/s streaming access
  - Typical use: desktops, data center bulk storage
- Magnetic tapes: archival storage

Memory & storage historical pricing



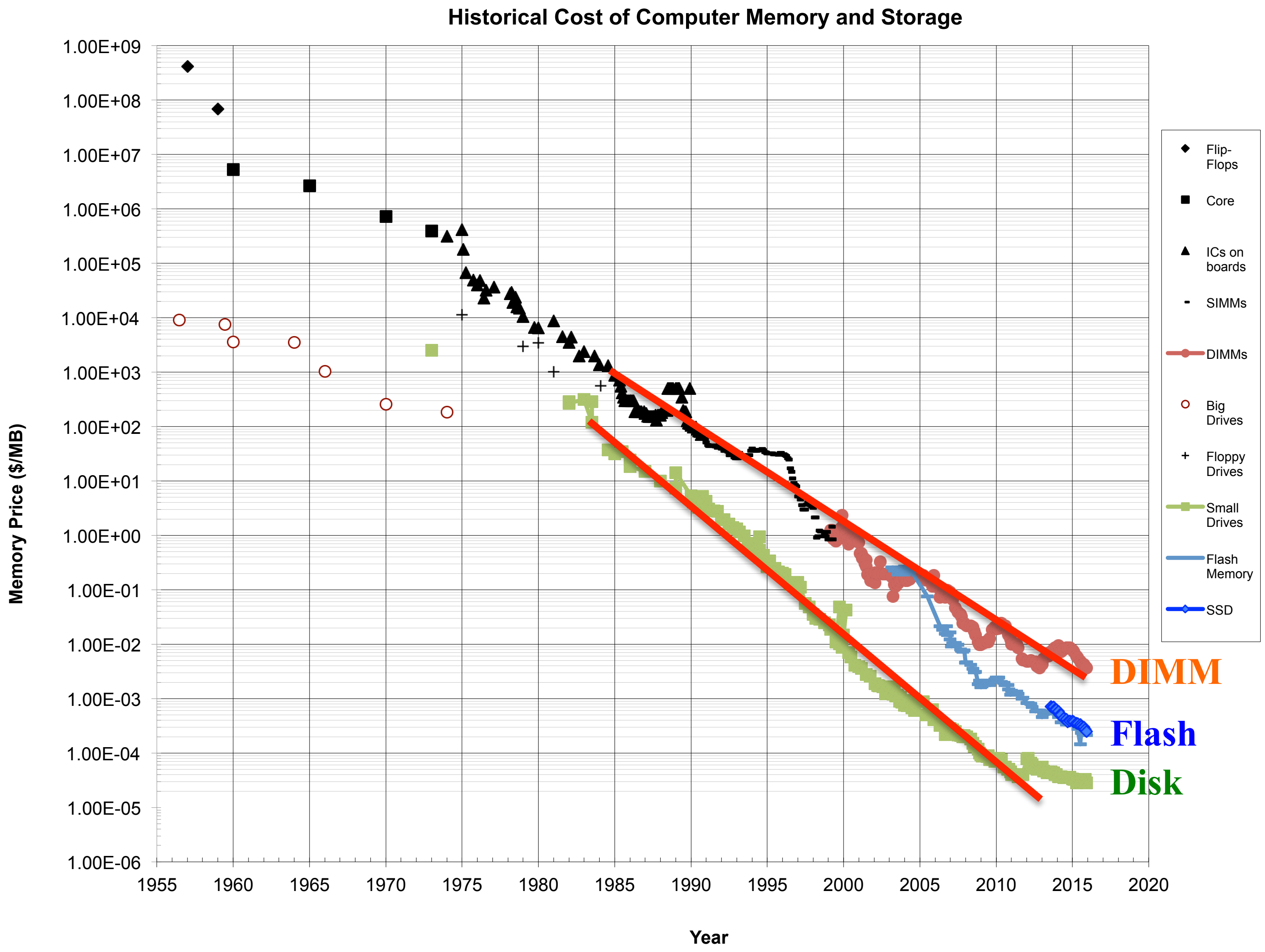
[J. McCallum, jcmit.com]

DRAM & disk pricing, 1991 angst

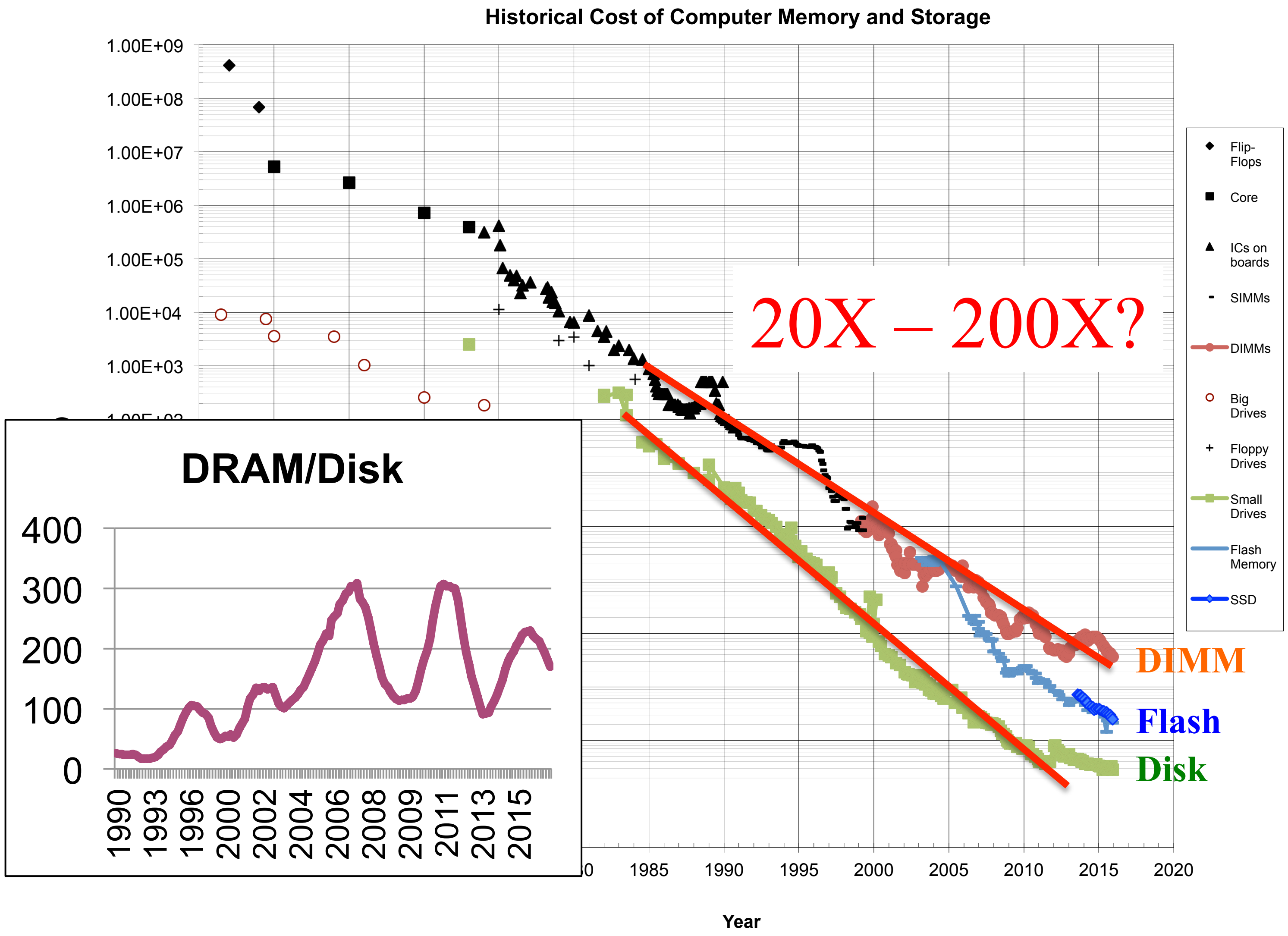


Risky to project from recent trend

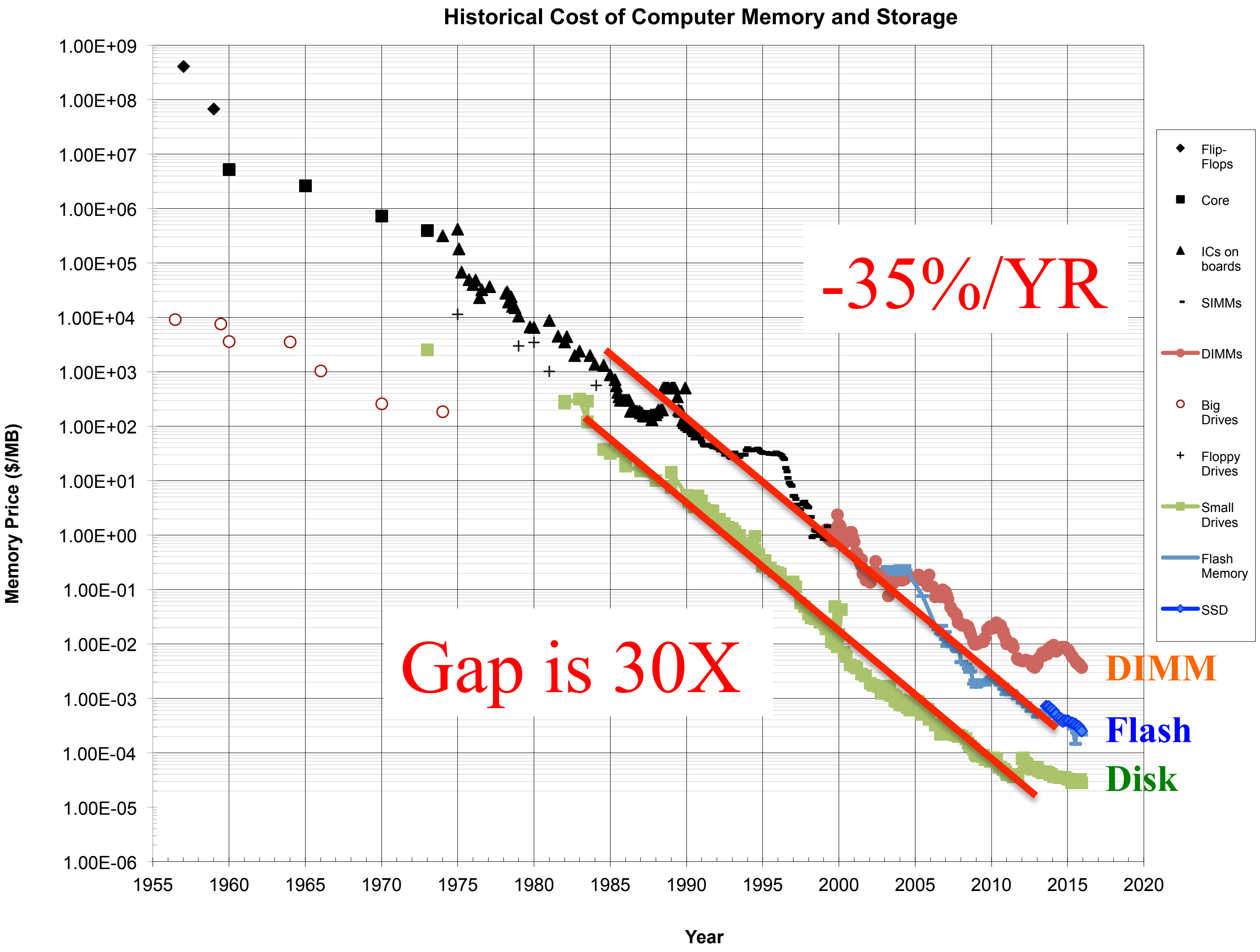
DRAM & disk pricing diverging



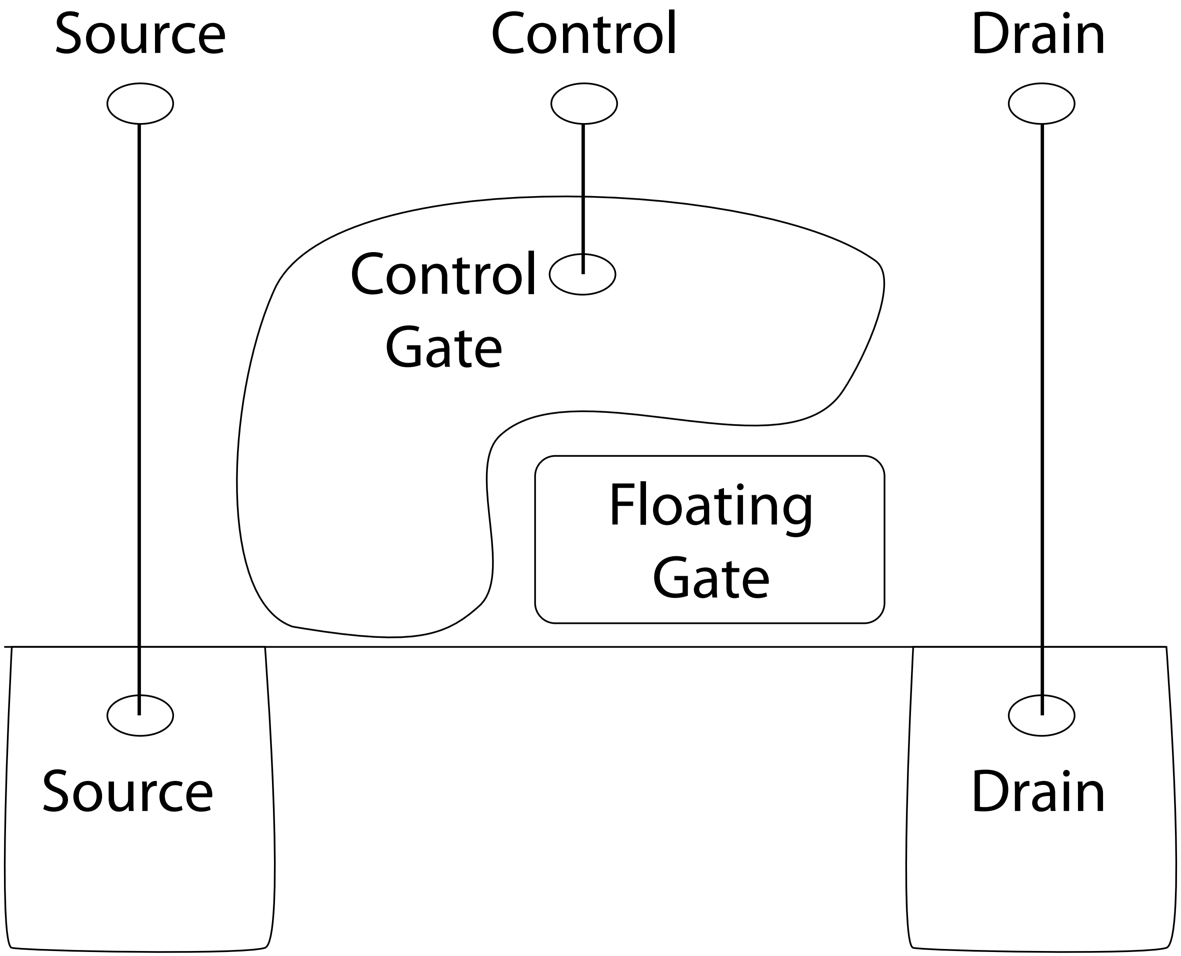
DRAM & disk pricing diverging



Best solid state & disk, Moore's Law?



Flash Memory



Flash Memory

- Basic operation: read/write to 4KB block at a time
  - Latency: 10-50 microseconds
  - Native Command Queueing (NCQ) for concurrent ops
- Blocks arranged in 2-D (soon 3-D) grid
  - Can read/write blocks in different “lanes” concurrently
- Writes must be to “clean” cells
  - Multi-block erasure required before write
  - Erasure block: 128 – 512 KB * # of lanes
  - Erasure time: 1-2 milliseconds
- Limited # of write cycles per block (1000s)

Intel SSD DC P3608 (2016)

Capacity	4 TB
Page Size	4 KB
Bandwidth (Sequential Reads)	5 GB/s
Bandwidth (Sequential Writes)	3 GB/s (peak)
Random 4KB Reads/sec	850 K
Random 4KB Writes/sec	50 K
Endurance	5000 erase/write cycles
Idle/Active Power	11W/20-40W
Interface	NVMe

Question

- Why are random writes so slow?
  - Random write/sec: 50K
  - Random read/sec: 850K
- Why are random writes so fast?
  - 1ms/erase => max 1000 writes/sec

Question

- Is persistence a problem?
  - What if OS writes to the same block repeatedly?
  - What if OS writes in a repeated scan?

- 1B blocks, lifetime 5000 writes/block
- 50K writes/sec (random)
- 750K writes/sec (sequential, peak)

Flash Translation Layer (FTL)

- Map logical block # to physical block #
  - Transparent to operating system
  - Translation stored in flash (along with each block)
  - Translation cached in SRAM/DRAM on device
- On write, put new block anywhere (clean)
- On read, look up translation to find most recent written location

FTL in Operation

FTL Garbage Collection

- Every block write creates an unused block
  - OS can also declare blocks dead (TRIM command)
- What happens when device fills up?
  - Need clean region to write incoming blocks
  - Create new clean region by copying live blocks from some mostly unused region, to clean region
  - Fill remainder with new blocks
  - Erase previous region

FTL Write Amplification

- Number of garbage collection writes/new block
- If device is completely full
  - Potentially need to do full erasure and re-write on every new block write => huge amplification
- Instead, keep 20-30% more physical blocks than logical blocks
  - If random updates, how much write amplification?
  - Are updates random?

Wear Levelling

- Each block can only be written a maximum number of times
  - FTL tracks # of erase/write cycles for each block
  - Unmap blocks that have worn out
- Preferentially
  - Write new blocks into regions with fewer update cycles
  - Clean cold data into regions with more update cycles

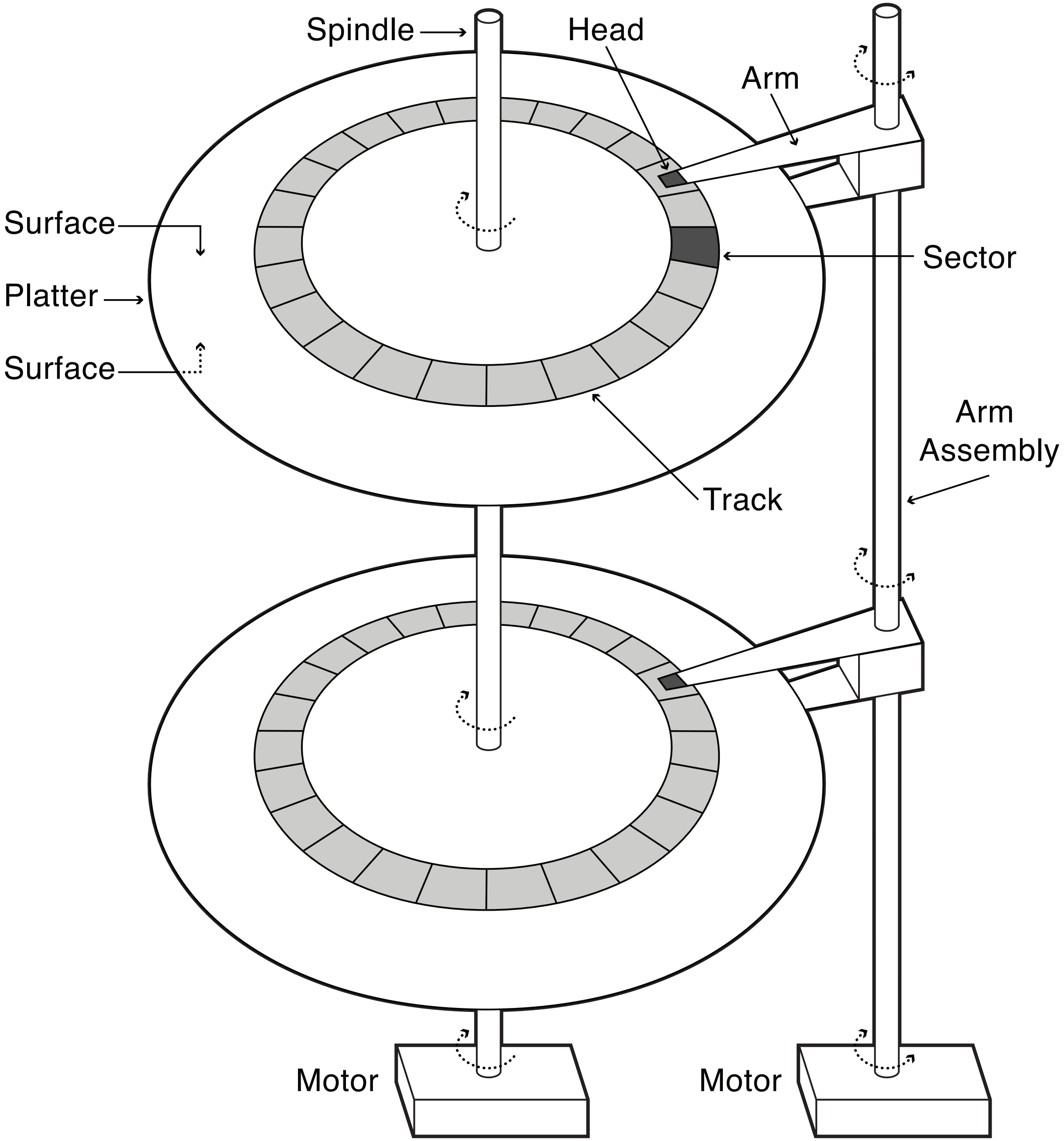
Low Latency Persistence

- Hybrid DRAM/flash devices
  - Commercially available
  - Small DRAM cache in front of flash
  - Capacitor/battery to flush modified data on power outage
  - If PCI (I/O bus) device, ~ 10us writes (request/ response and DMA overheads dominate)
  - If DIMM form factor, -> 100ns reads

Non-flash solid state

- 3D Xpoint, PCM, Memristor, ReRAM
  - Cache block level read/write
  - Latencies ~ 2x DRAM, on memory bus
  - No static power draw
- Low latency persistence
- Low operating power (TCO)
  - Chasing DRAM market share
  - Impact on flash market is uncertain
- Much better endurance than flash
  - With access speeds, direct access w/o wear leveling expires cell in minutes

Magnetic Disk


Spindle
Head
Arm
Surface
Sector
Platter
Surface
Arm
Assembly
Track
Motor
Motor

Disk Tracks

- ~ 1 micron wide
  - Wavelength of light is ~ 0.5 micron
  - Resolution of human eye: 50 microns
  - 100K tracks on a typical 2.5" disk
- Separated by unused guard regions
  - Reduces likelihood neighboring tracks are corrupted during writes (still a small non-zero chance)
- Track length varies across disk
  - Outside: More sectors per track, higher bandwidth
  - Disk is organized into regions of tracks with same # of sectors/track
  - Only outer half of radius is used
    - Most of the disk area in the outer regions of the disk

Sectors

Sectors contain sophisticated error correcting codes

- Disk head magnet has a field wider than track
- Hide corruptions due to neighboring track writes

- Sector sparing
  - Remap bad sectors transparently to spare sectors on the same surface
- Slip sparing
  - Remap all sectors (when there is a bad sector) to preserve sequential behavior
- Track skewing
  - Sector numbers offset from one track to the next, to allow for disk head movement for sequential ops

Disk Performance

Disk Latency =

- Seek Time + Rotation Time + Transfer Time
- Seek Time: time to move disk arm over track (1-20ms)
  - Fine-grained position adjustment necessary for head to “settle”
  - Head switch time ~ track switch time (on modern disks)
- Rotation Time: time to wait for disk to rotate under disk head
  - Disk rotation: 4 – 15ms (depending on price of disk)
  - On average, only need to wait half a rotation
- Transfer Time: time to transfer data onto/off of disk
  - Disk head transfer rate: 100-250MB/s (5-10 usec/sector)
  - Host transfer rate dependent on I/O connector (USB, SATA, …)

HGST Ultrastar He10 (2016)

Capacity	10 TB, 7 platters
Spin Speed	7200 RPM
Sustained Transfer Rate	249 MB/s (read), 225 MB/s (write)
Interface Transfer Rate	1200 MB/s
Seek time (avg)	8 ms (read), 8.6 ms (write)
Rotational latency (avg)	4.16 ms
Cache	256 MB
Idle/Operating Power	6W/9.5W
Bit Error Rate (read)	10^-15

Question

- How long to complete 100 random 4KB disk reads, in FIFO order?

Question

- How long to complete 100 random 4KB disk reads, in FIFO order?
  - Seek: average 8 msec
  - Rotation: average 4.16 msec
  - Transfer: 4KB / 249 MB/s = 16 usec
- 100 * (8 + 4.16 + 0.016) = 1.2 seconds

Question

- How long to complete 100 sequential 4KB disk reads?

Question

- How long to complete 100 sequential 4KB disk reads?
  - Seek Time: 8 ms (to reach first sector)
  - Rotation Time: 4.16 ms (to reach first sector)
  - Transfer Time: 400KB / 249MB/sec = 1.6 ms

Total: 8 + 4.16 + 1.6 = 13.8 ms

  - Might need an extra head or track switch (+1ms)
  - Track buffer may allow some sectors to be read out of order (-2ms)

Question

- How large a transfer is needed to achieve 80% of the max disk transfer rate?

Question

- How large a transfer is needed to achieve 80% of the max disk transfer rate?

  Assume 12.16 ms to reach first sector

  Assume x rotations are needed, 8.5ms/rotation

  Then solve for x:

  0.8 (12.16ms + 8.5ms x) = 8.5ms x

Total: x = 5.7 rotations, 12.1 MB

Disk Scheduling

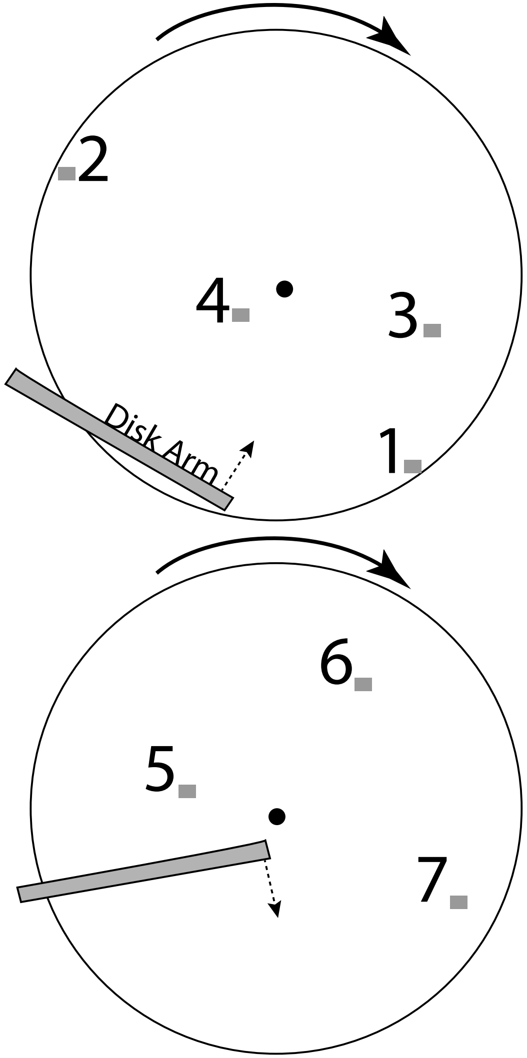
- FIFO
  - Schedule disk operations in order they arrive
  - Downsides?

Disk Scheduling

- Shortest seek time first
  - Not optimal!
    - Suppose cluster of requests at far end of disk
  - Downsides?

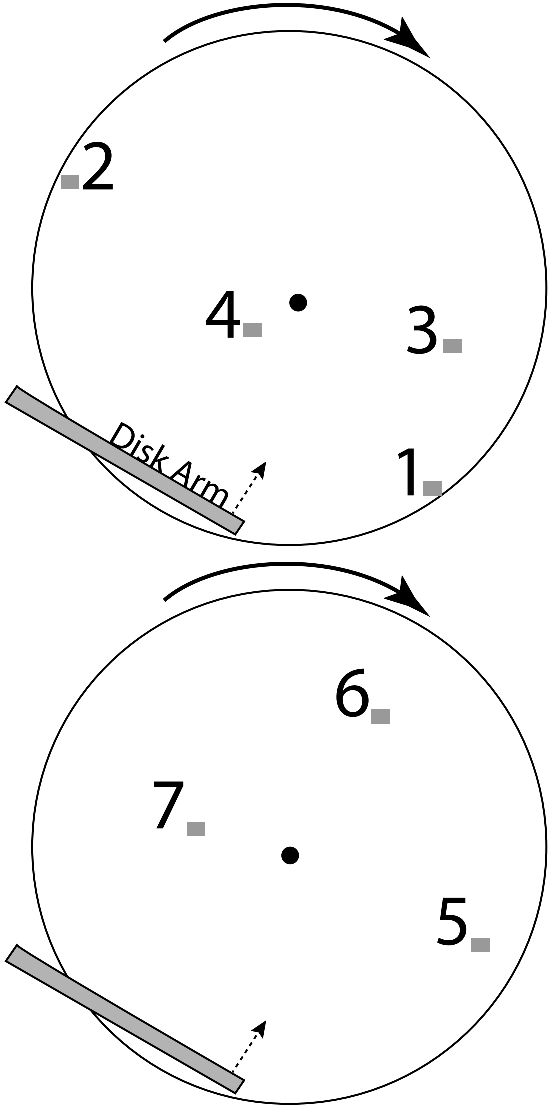
Disk Scheduling

- SCAN: move disk arm in one direction, until all requests satisfied, then reverse direction
- Also called “elevator scheduling”



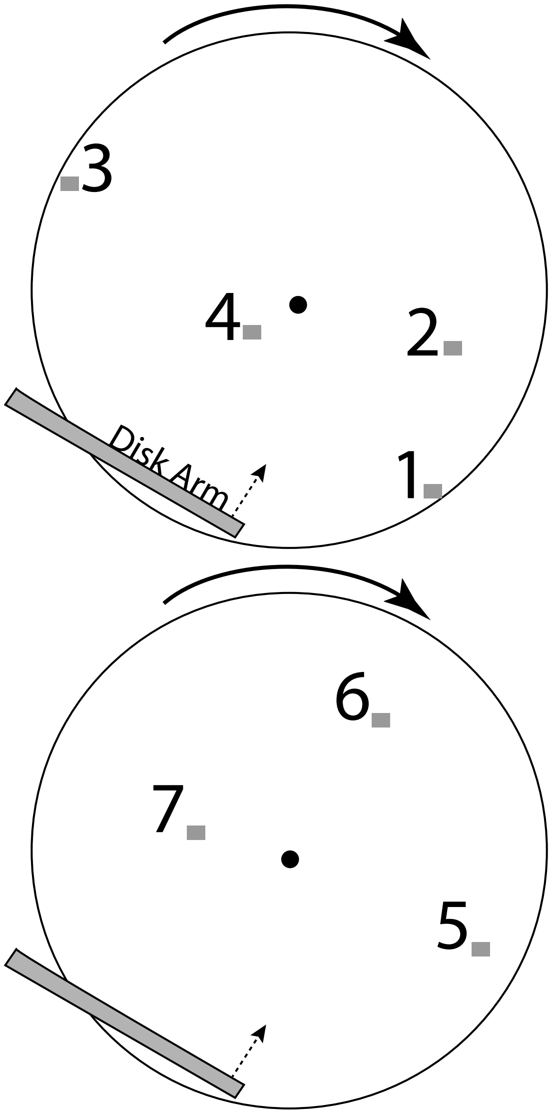
Disk Scheduling

- CSCAN: move disk arm in one direction, until all requests satisfied, then start again from farthest request



Disk Scheduling

- R-CSCAN: CSCAN but take into account that short track switch is < rotational delay



Question

- How long to complete 100 random disk reads, in any order?

Question

- How long to complete 100 random disk reads, in any order?
  - Disk seek: 1ms (most will be short)
  - Rotation: 4.16ms
  - Transfer: 16usec
- Total: 100 * (1 + 4.16 + 0.016) = 0.52 seconds
  - Would be a bit shorter with R-CSCAN
  - vs. 1.2 seconds if FIFO order

Question

- How long to read all of the bytes off of a disk?

Question

- How long to read all of the bytes off of a disk?
  - – Disk capacity: 10TB
  - – Disk bandwidth: 249MB/s (average)
- Transfer time = 40K seconds (12 hours)

Question

- If you read all the data off the disk, how likely will some of the data be corrupted?

Question

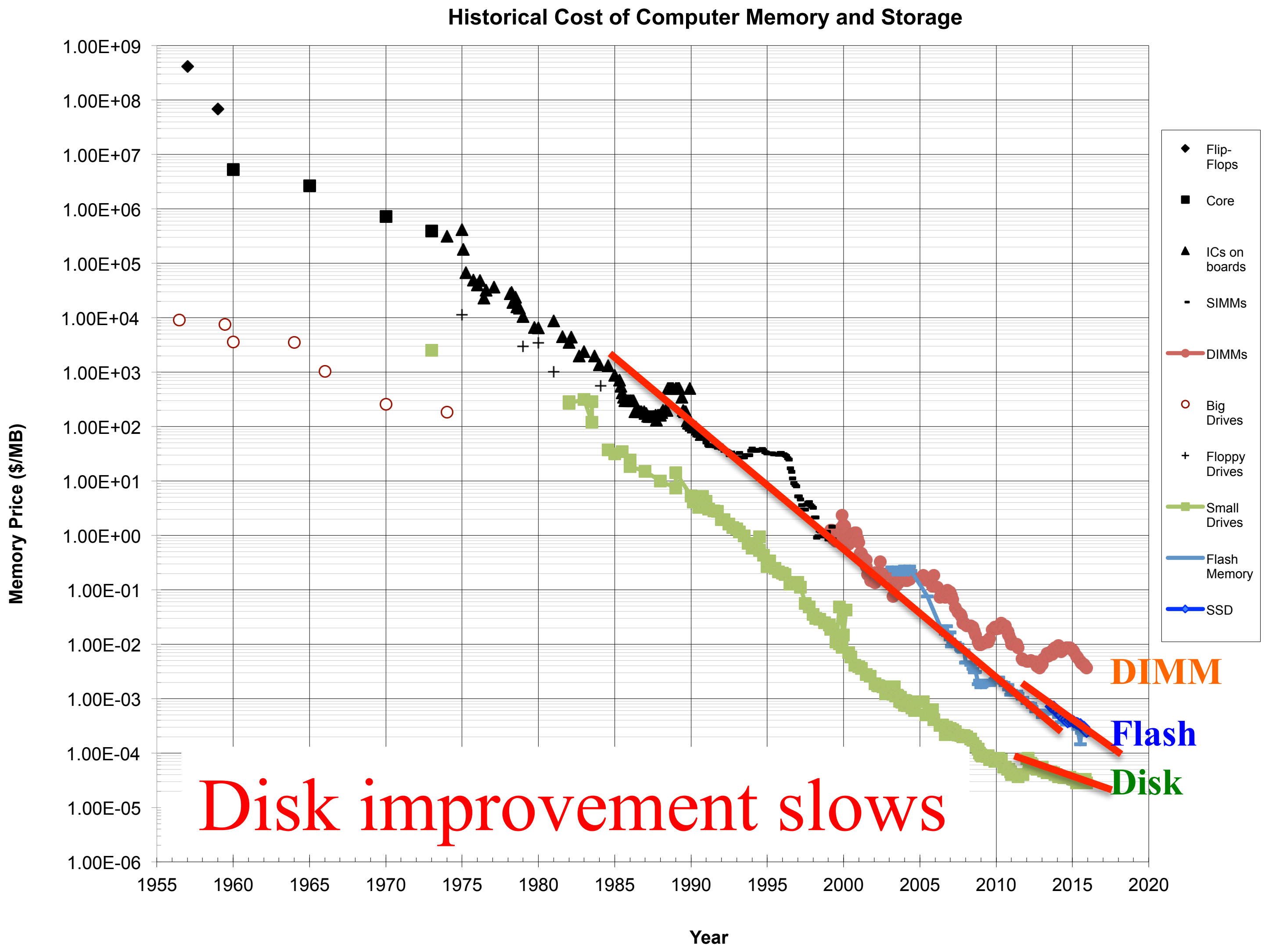
- If you read all the data off the disk, how likely will some of the data be corrupted?

Bit error rate = 10^-15

Bits per disk w/ 10TB = 10^14

=> 10% !!

Flash SSD & disk pricing, recently



DIMM

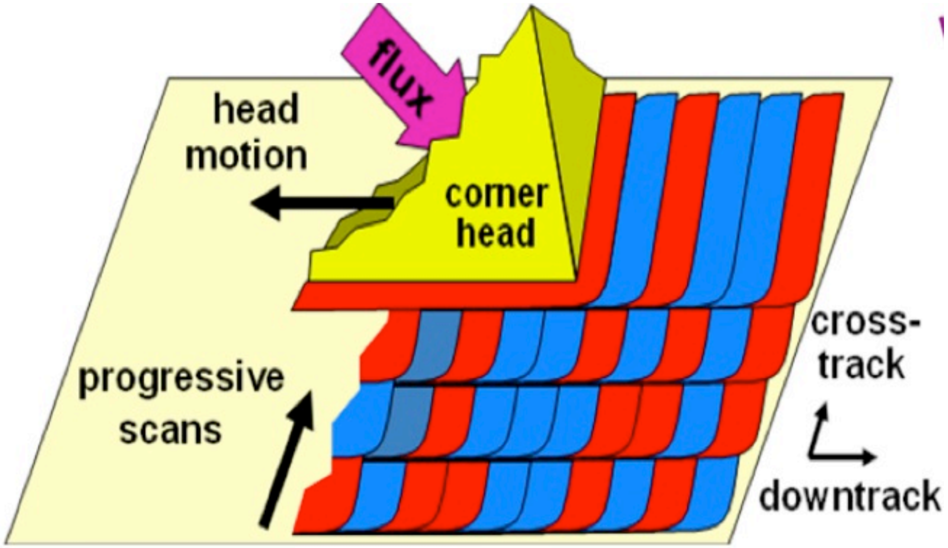
Flash

Disk

Disk improvement slows

Shingled magnetic recording (SMR)

- Uses ~current tech
- Overlap adjacent tracks (no gap)
- More tracks/inch
- No sector overwrite



Wood, Trans. Magnetics., 2009

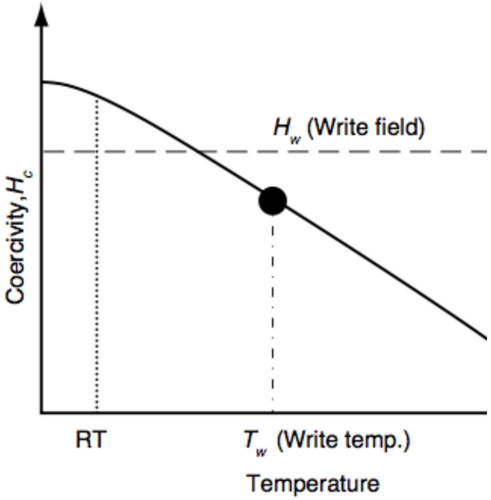
- Two-dimensional magnetic recording (TDMR)
  - Inter-track interference ever worse, data dependent
  - Give up on flying head path staying “in track”
  - Include 2 (then 3) read sensors per head
    - Read multiple “sub-tracks”, signal process to data

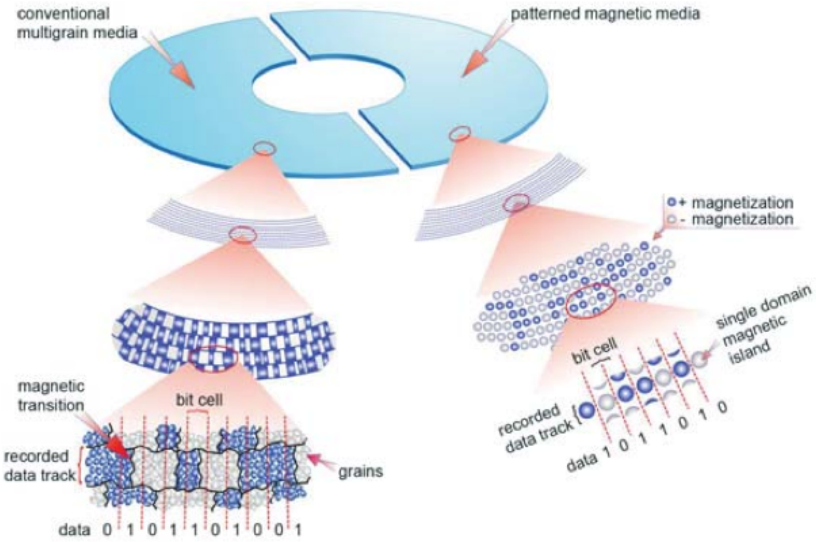
SMR today/TDMR soon

- Hidden behind “Shingle Translation Layer (STL)”
  - Embedded layer that re-writes entire region
  - New blocks go to empty spill region
  - Re-write/coalesce existing regions when mostly empty
- Adding 10% - 30% areal density (not 2X soon)
- Interesting parallel/convergence
  - FTL sequentially writes flash pages in erase block
  - Flash erase block analogous to shingled band

More Changes In Store for Disks

- Heat-Assisted (HAMR)
  - Small bits need high coercivity media to retain orientation
  - High coercivity media is not changed by normal writing
  - Heated media lowers coercivity
  - Include lasers on Rd/Wr head?
- Bit-Patterned (BPM)
  - Small bits retain orientation more easily if bits kept apart
  - Pattern media so only write a single dot per bit
  - Tera-dots per sq. inch?





Still, not looking good for disk

- Driven from margin-rich enterprise apps
- Driven from volume rich mobile
- Big changes in fabrication & materials
- Small number of companies playing
  - Natural disasters can change everything

- How much will cloud storage growth pay?

- Watch for HAMR roll out in next few years