Storage Systems

Loïc Guégan
Inf-2201, University of Tromsø
Based on presentations created by: Lars Ailo Bongo, Otto Anshus, Bård Fjukstad, Daniel Stødle And Kai Li and Andy Bavier

Big Data Sources

- Voluntary human produced content
  - Videos, photos, audio…
- Involuntary produced content
  - Online activity logging, tax records…
- Scientific instruments
  - CERN LHC, Sloan Digital Sky Survey, brain simulations, DNA sequencers…

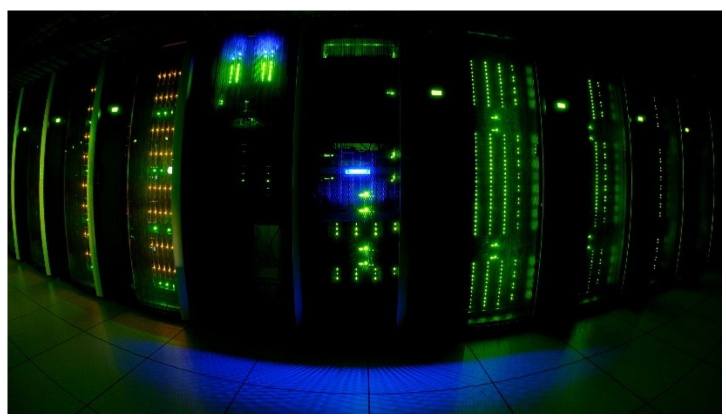
- How big?

Dataset Size

< 16GB

< 1TB

TBs



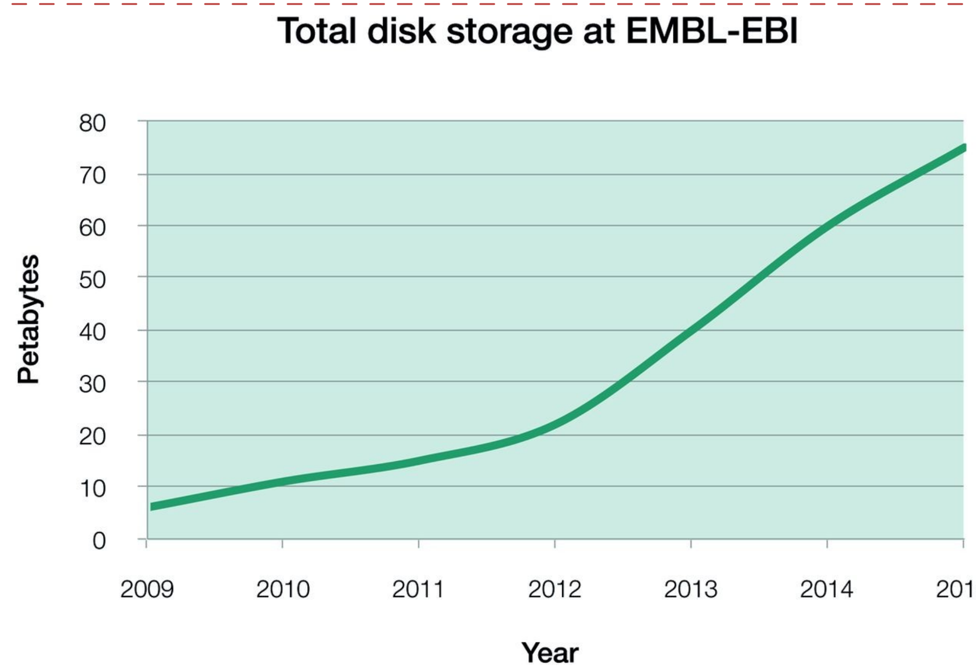
PBs

Data Analysis Framework

The data challenge: Data growth

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady

- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade



Source: Charles E. Cook at al. *Nucl. Acids Res. 2016; 44: D20-D26*

Overview

- Magnetic disks
- Disk arrays
- Flash storage
- DRAM storage
- Storage hierarchy

Storage

- Reliability
  - Archival
  - Reliable
  - Persistent
  - Temporal
- Access pattern
  - Write or read intensive
  - Sequential or random access
  - Low-latency or high throughput
- Cost
- Power

The Memory Hierarchy

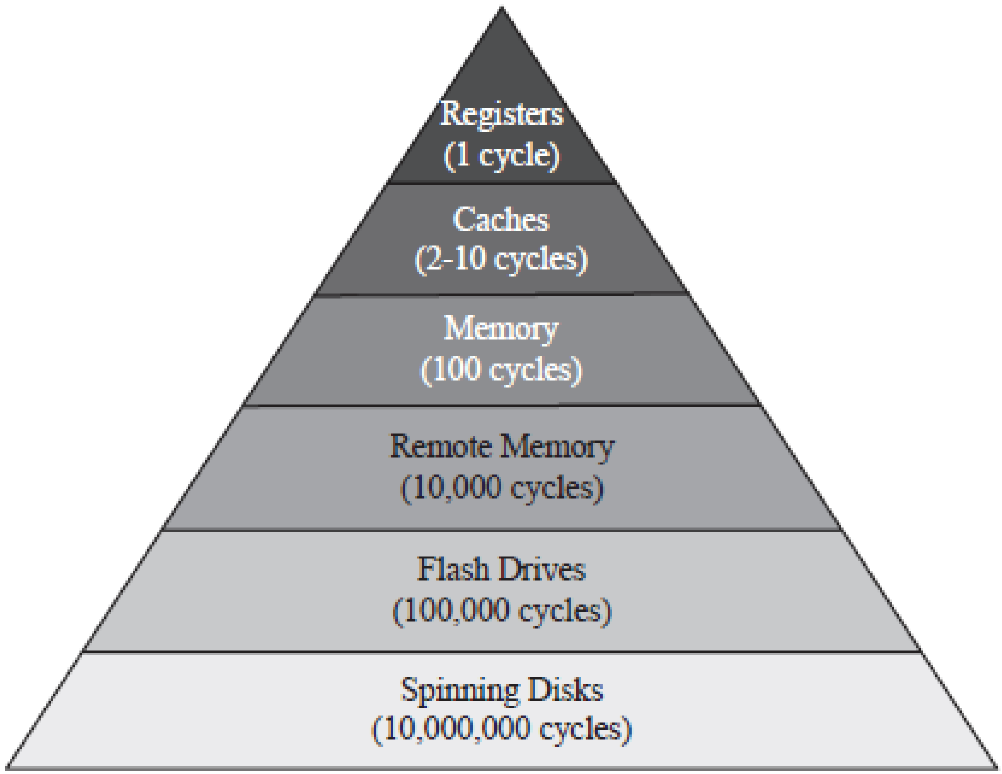


Figure 1. The memory hierarchy. *Each level shows the typical access latency in processor cycles. Note the five-orders-of-magnitude gap between main memory and spinning disks.*

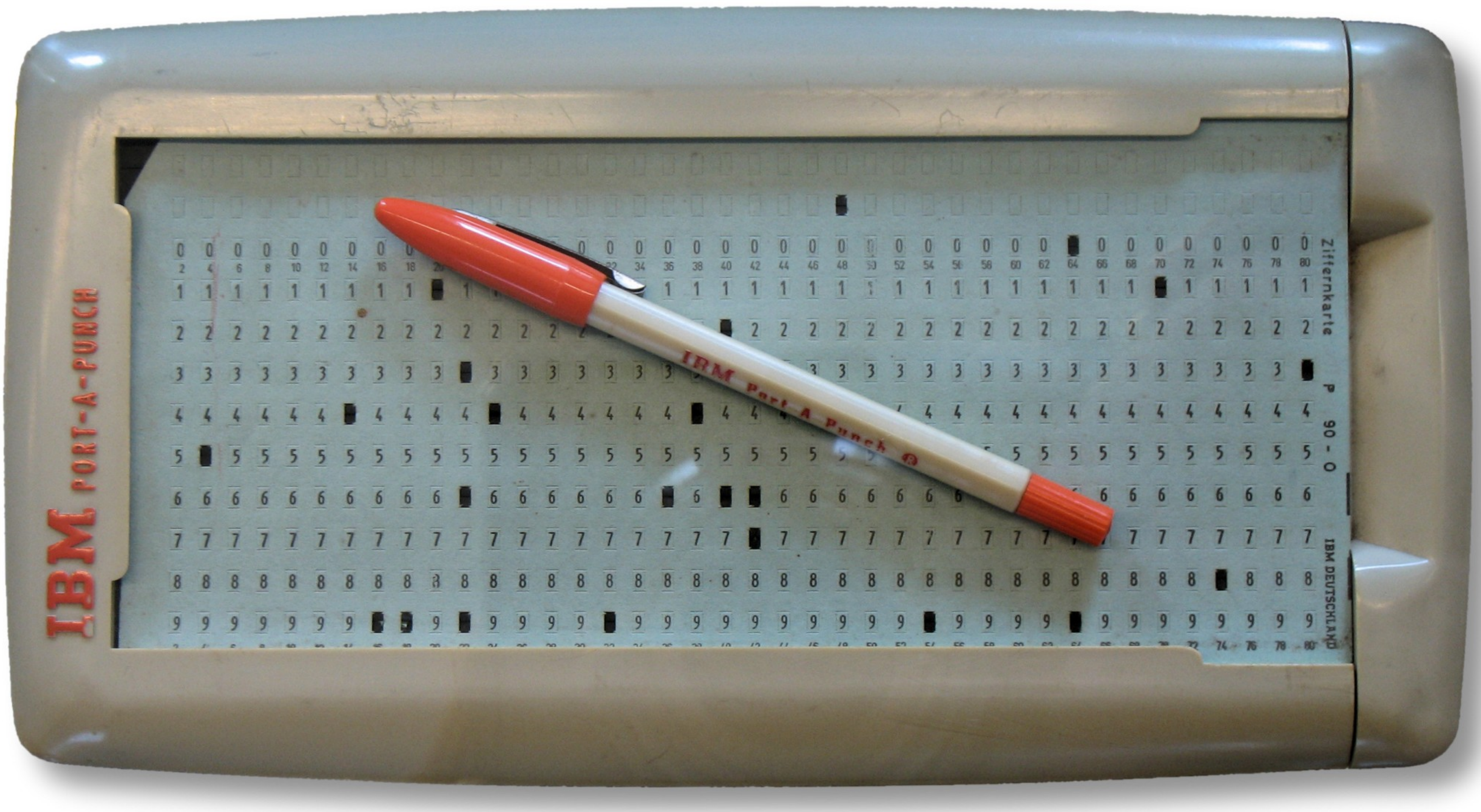
Jiahua He, Arun Jagatheesan, Sandeep Gupta, Jeffrey Bennett, Allan Snavely, "DASH: a Recipe for a Flash-based Data Intensive Supercomputer," sc, pp.1-11, 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis, 2010

Disk vs. Flash vs. DRAM

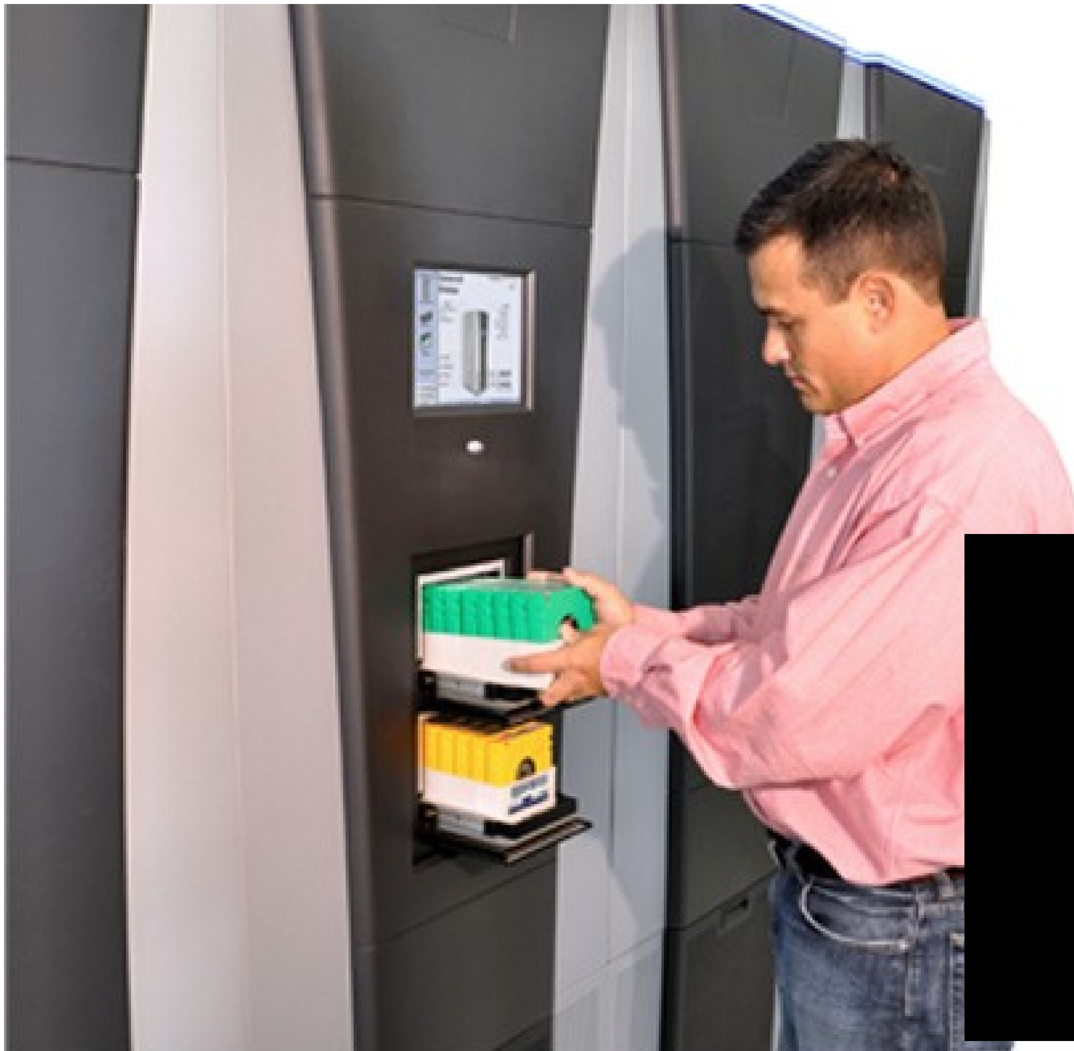
	Disk	Flash	DRAM
Access time (relative)	1	0.01-0.001	0.000001 (1 / 100,000)
Cost (relative)	1	15-25	30-150
Bandwidth (relative)	1	1	80
Bandwidth/ GB (relative)	1		6,000
Bandwidth/ $ (relative)	1		160

Source: Computer Architecture A Quantitative Approach

Punch Cards


IBM PORT-A-PUNCH
IBM Port-A-Punch
Ziffernkarte P 90-0
IBM DEUTSCHLAND

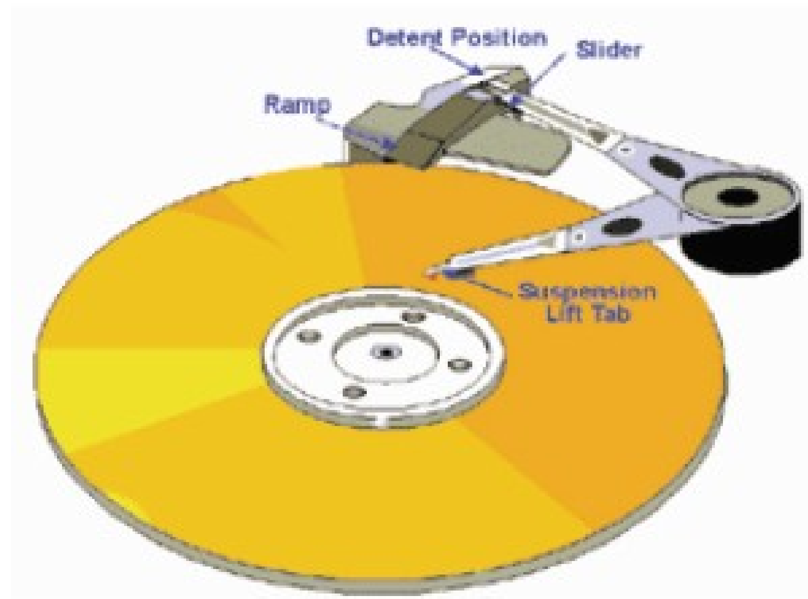
Tape Library



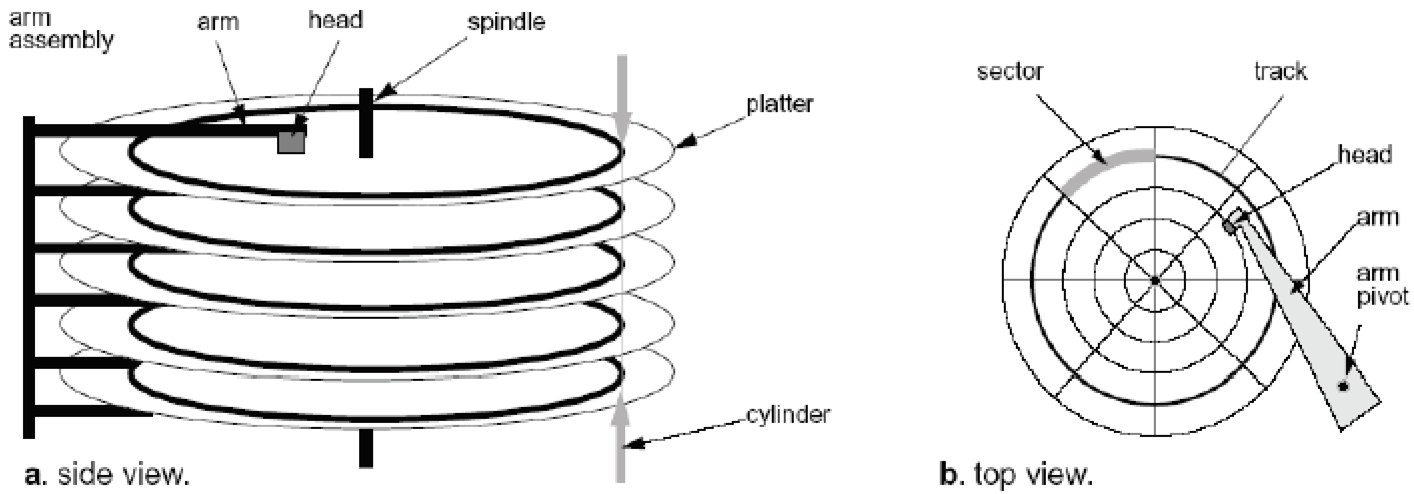
Hard Drive

Disk Arm and Head

- Disk arm
  - A disk arm carries disk heads
- Disk head
  - Mounted on an actuator
  - Read and write on disk surface
- Read/write operation
  - Disk controller receives a command with <track#, sector#>
  - Seek the right cylinder (tracks)
  - Wait until the right sector comes
  - Perform read/write



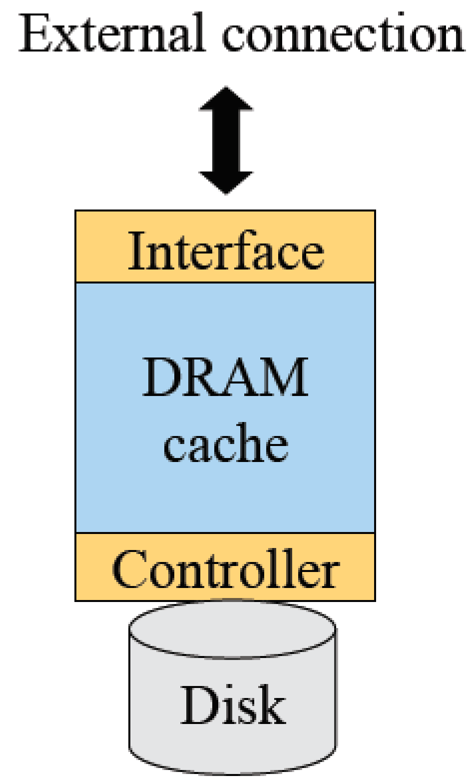
Mechanical Component of A Disk Drive



- Tracks
  - Concentric rings around disk surface, bits laid out serially along each track
- Cylinder
  - A track of the platter, 1000-5000 cylinders per zone, 1 spare per zone
- Sectors
  - Each track is split into arc of track (min unit of transfer)

A Typical Magnetic Disk Controller

- External connection
  - Parallel ATA (aka IDE or EIDE), Serial ATA, SCSI, Serial Attached SCSI (SAS), Fibre Channel, FireWire, USB
- Cache
  - Buffer data between disk and interface
- Controller
  - Read/write operation
  - Cache replacement
  - Failure detection and recovery

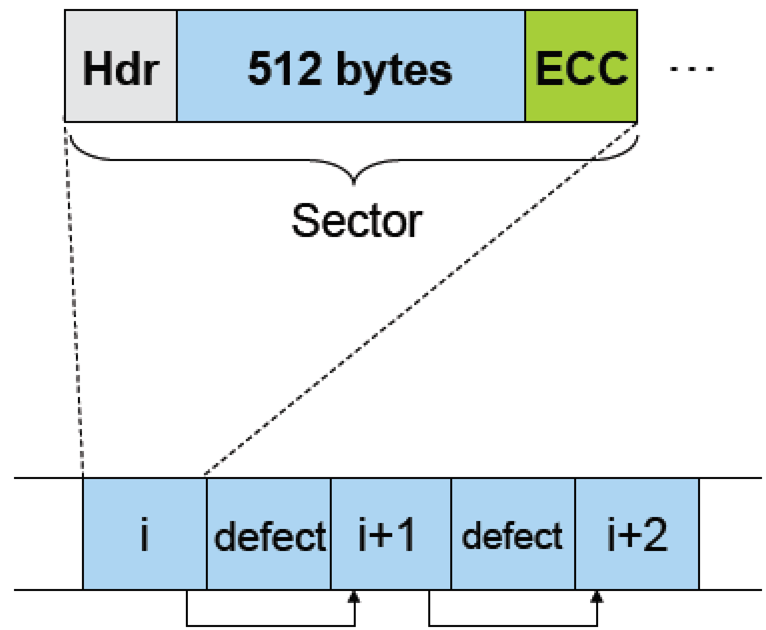


Disk Caching

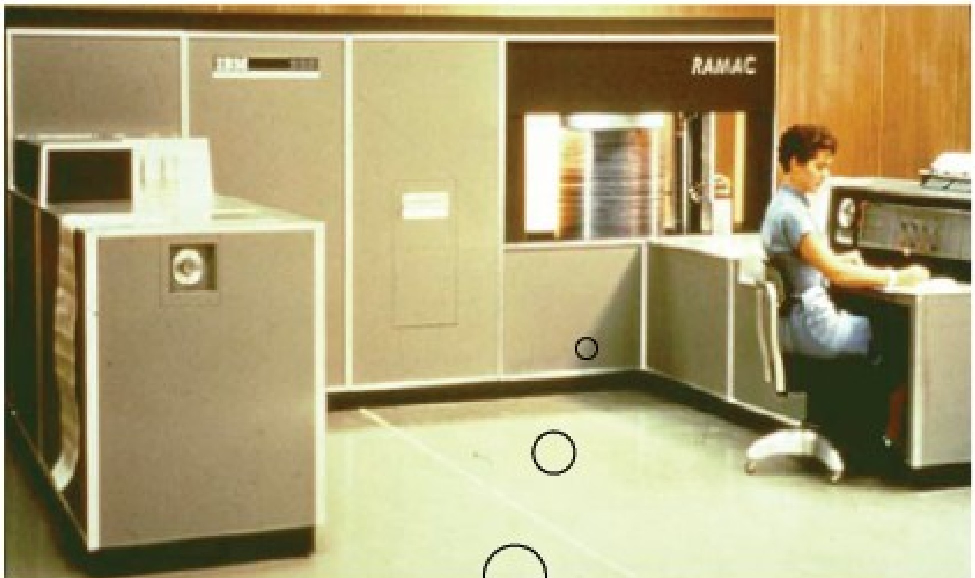
- Method
  - Use DRAM to cache recently accessed blocks
    - Most disks have 32MB
    - Some of the RAM space stores “firmware” (an embedded OS)
  - Blocks are replaced usually in an LRU order
- Pros
  - Good for reads if accesses have locality
- Cons
  - Cost
  - Need to deal with reliable writes

Disk Sectors

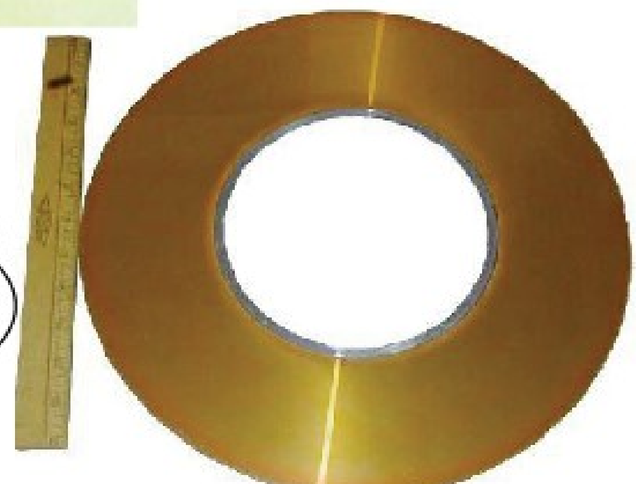
- Where do they come from?
  - Manufacturing process
  - Logical maps to physical
- What is a sector?
  - Header (ID, defect flag, …)
  - Real space (e.g. 512 bytes)
  - Trailer (ECC code)
- What about errors?
  - Detect errors in a sector
  - Correct them with ECC
  - If not recoverable, replace it with a spare
  - Skip bad sectors in the future



Disks Were Large



First Disk:
IBM 305 RAMAC (1956)
5MB capacity
50 disks, each 24”



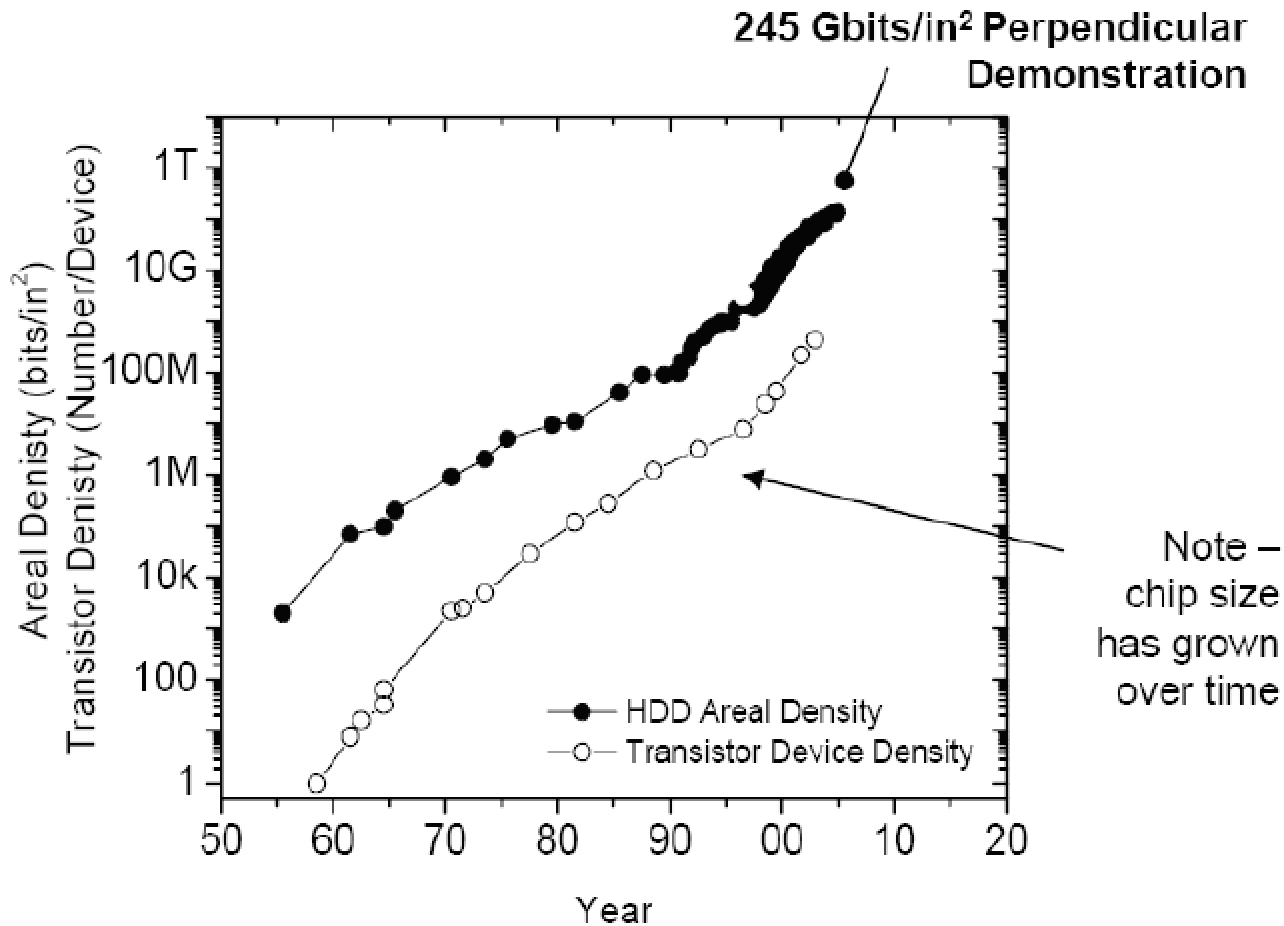
They Are Now Much Smaller

Form factor:
.5-1”× 4”× 5.7”
Storage:
0.5-2TB

Form factor:
.4-.7” × 2.7” × 3.9”
Storage:
60-200GB

Form factor:
.2-.4” × 2.1” × 3.4”
Storage:
1GB-8GB

Areal Density vs. Moore's Law



(Mark Kryder at SNW 2006)

50 Years Later (Mark Kryder at SNW 2006)

	IBM RAMAC (1956)	Seagate Momentus (2006)	Difference
Capacity	5MB	160GB	32,000
Areal Density	2K bits/in^2	130 Gbits/in^2	65,000,000
Disks	50 @ 24” diameter	2 @ 2.5” diameter	1 / 2,300
Price/MB	$1,000	$0.01	1 / 3,200,000
Spindle Speed	1,200 RPM	5,400 RPM	5
Seek Time	600 ms	10 ms	1 / 60
Data Rate	10 KB/s	44 MB/s	4,400
Power	5000 W	2 W	1 / 2,500
Weight	~ 1 ton	4 oz	1 / 9,000

Sample Disk Specs (from Seagate)

	Cheetah 15k.7	Barracuda XT
Capacity		
Formatted capacity (GB)	600	2000
Discs	4	4
Heads	8	8
Sector size (bytes)	512	512
Performance		
External interface	Ultra320 SCSI, FC, S. SCSI	SATA
Spindle speed (RPM)	15,000	7,200
Average latency (msec)	2	4.16
Seek time, read/write (msec)	3.5/3.9	8.5/9.5
Track-to-track read/write (msec)	0.2-0.4	0.8/1.0
Internal transfer (MB/sec)	1,450-2,370	600
Transfer rate (MB/sec)	122-204	138
Cache size (MB)	16	64
Reliability		
Recoverable read errors	1 per 1012 bits	1 per 1010 bits

Disk Performance (2TB disk)

- Seek
  - Time to move disk arm to correct track
  - Position heads over cylinder, typically 3.5-9.5 ms
- Rotational delay
  - Time to wait for a sector to rotate underneath the head
  - Typically 8 - 4 ms (7,200 – 15,000RPM) or 1/2 rotation takes 4 - 2ms
- Transfer
  - Time to move data to / from disk
  - Disk head transfer rate is typically 40-138 MBytes/sec
  - (+ host trasfer rate, higly dependente on chosen I/O interface)
- Performance of transfer 1 KBytes
  - Disk latency = Seek + half rotational delay + transfer (at disk head tranfer rate)
  - So here : 4ms + 2ms + 0.007ms (at 138 MB/s)
  - Disk latency is 6.007 ms. Or 166.47 KBytes/sec

More on Performance

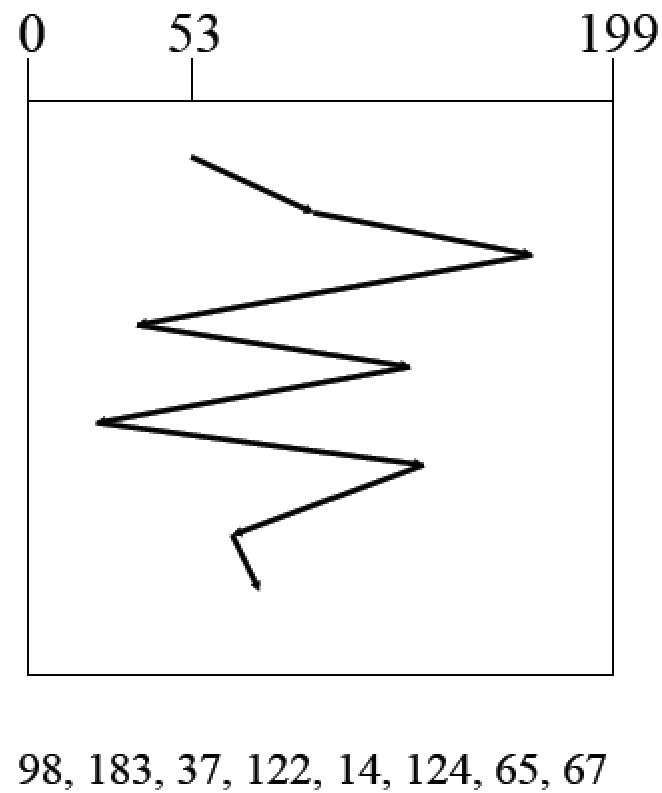
- What transfer size can get 75% of the disk bandwidth?
  - Assume Disk BW = 60MB/sec, 1/2 rotation = 2ms, seek = 4ms

Block Size	% of Disk Transfer Bandwidth
1KBytes	~0.28%
1MBytes	~75%

- Seek and rotational times dominate the cost of small accesses
  - Disk transfer bandwidth are wasted
  - Need algorithms to reduce seek time
- Speed depends on which sectors to access
  - Are outer tracks or inner tracks faster?

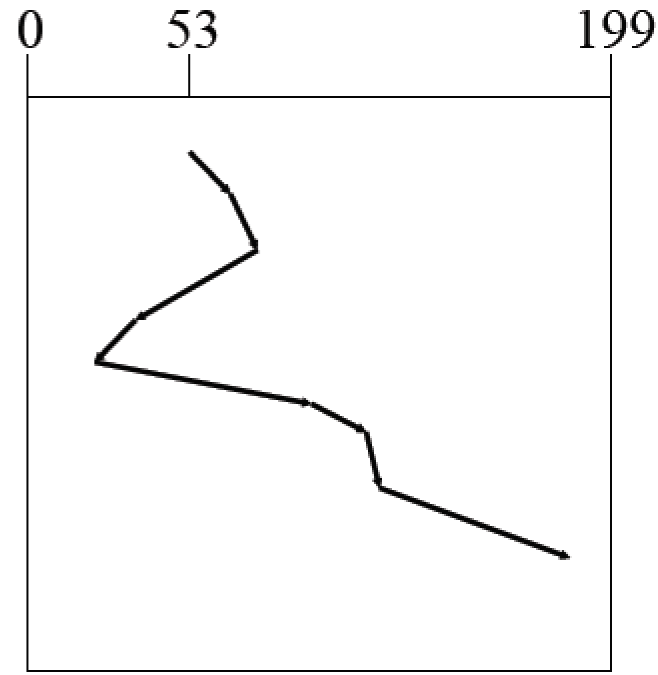
FIFO (FCFS) order

- Method
  - First come first serve
- Pros
  - Fairness among requests
  - In the order applications expect
- Cons
  - Arrival may be on random spots on the disk (long seeks)
  - Wild swing can happen



SSTF (Shortest Seek Time First)

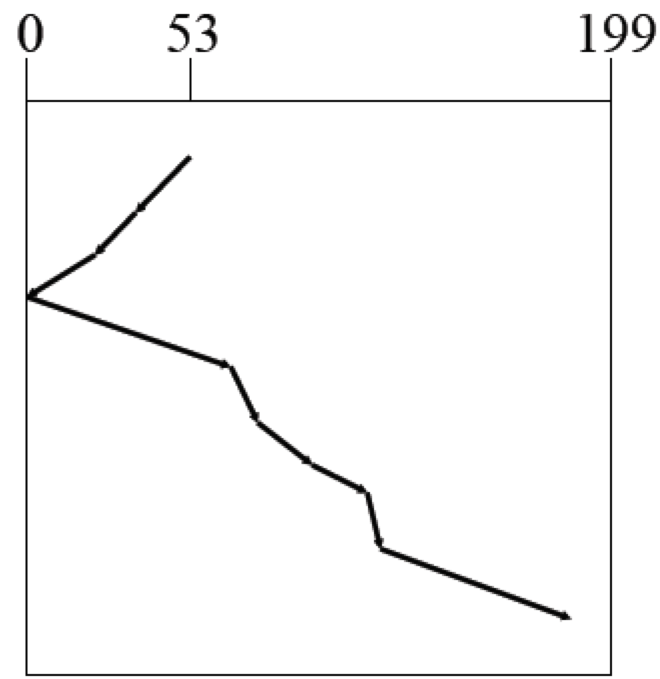
- Method
  - Pick the one closest on disk
  - Rotational delay is in calculation
- Pros
  - Try to minimize seek time
- Cons
  - Starvation
- Question
  - Is SSTF optimal?
  - Can we avoid the starvation?



98, 183, 37, 122, 14, 124, 65, 67
(65, 67, 37, 14, 98, 122, 124, 183)

Elevator (SCAN)

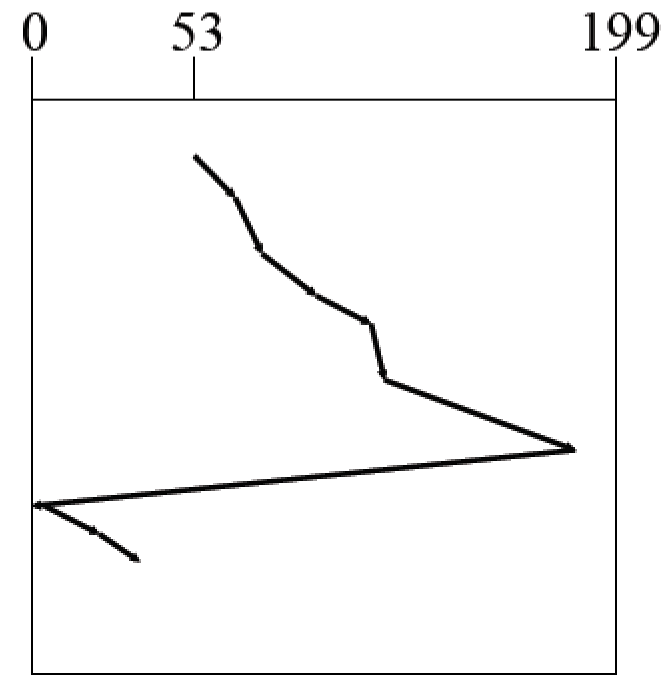
- Method
  - Take the closest request in the direction of travel
  - Real implementations do not go to the end (called LOOK)
- Pros
  - Bounded time for each request
- Cons
  - Request at the other end will take a while



98, 183, 37, 122, 14, 124, 65, 67
(37, 14, 65, 67, 98, 122, 124, 183)

C-SCAN (Circular SCAN)

- Method
  - Like SCAN
  - But, wrap around
  - Real implementation doesn't go to the end (C-LOOK)
- Pros
  - Uniform service time
- Cons
  - Do nothing on the return

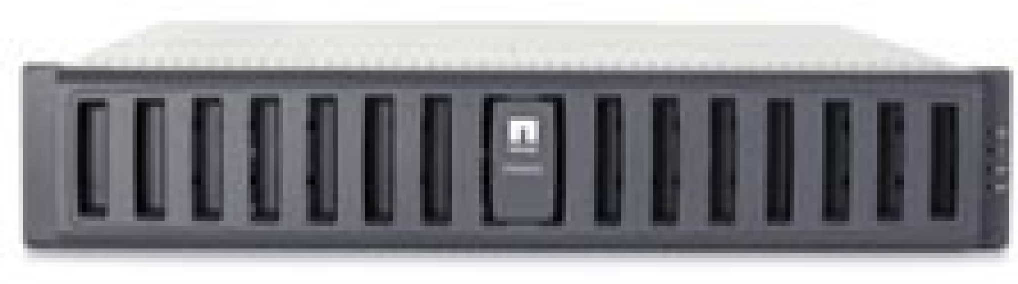


98, 183, 37, 122, 14, 124, 65, 67
(65, 67, 98, 122, 124, 183, 14, 37)

Discussions

- Seek algorithms:
  - FIFO
  - SSTF
  - SCAN
  - C-SCAN
- Disk I/O request buffering
  - How much to requests to buffer?

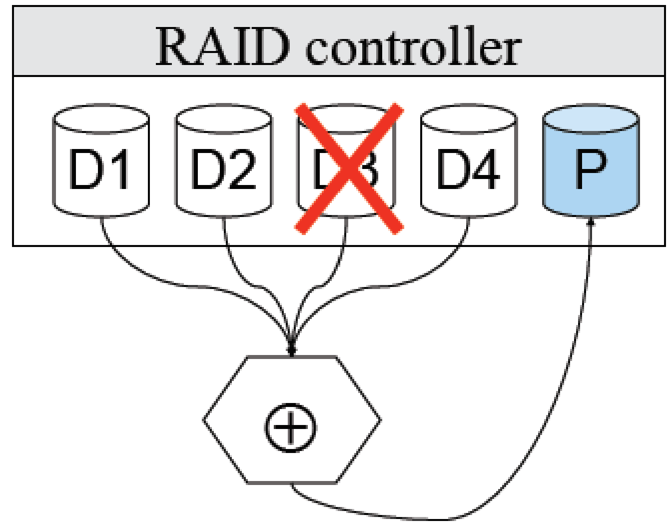
Storage System



- Network connected box with many disks
- Goals
  - Reliability
  - Higher throughput
  - What if there are 1000 disks?

RAID (Redundant Array of Inexpensive Disks)

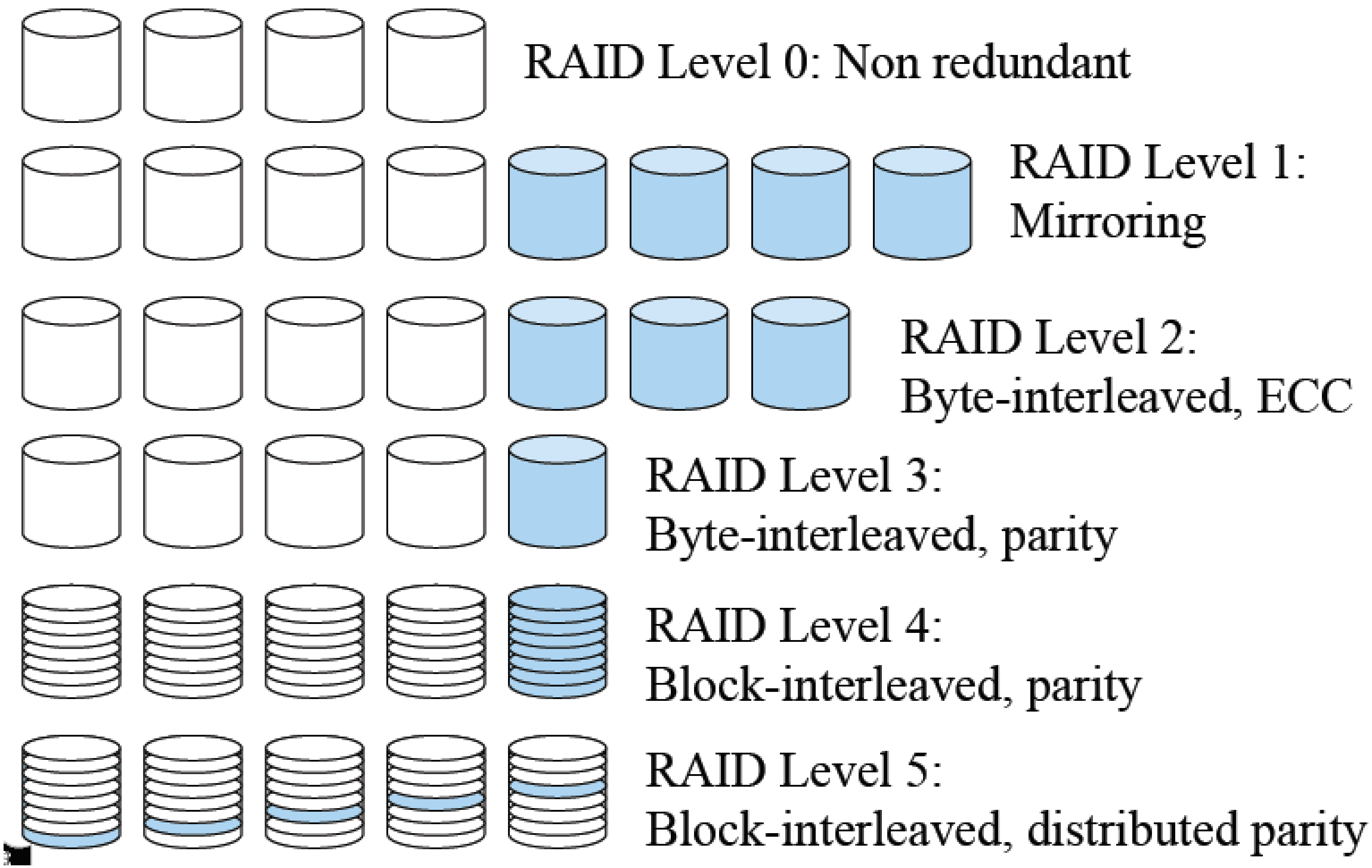
- Main idea
  - Store the error correcting codes on other disks
  - General error correcting codes are too powerful
  - Use XORs or single parity
  - Upon any failure, one can recover the entire block from the spare disk (or any disk) using XORs
- Pros
  - Reliability
  - High bandwidth
- Cons
  - The controller is complex



$$P = D1 \oplus D2 \oplus D3 \oplus D4$$

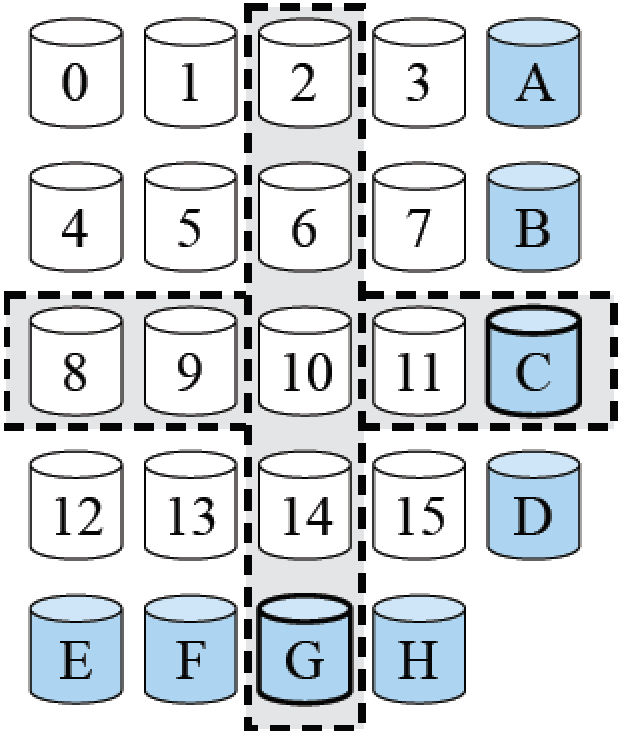
$$D3 = D1 \oplus D2 \oplus P \oplus D4$$

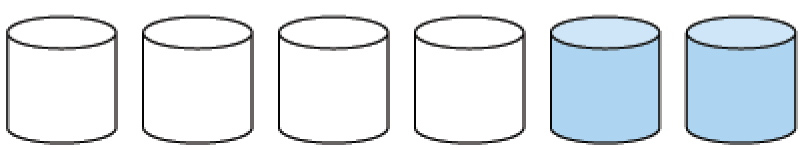
Synopsis of RAID Levels



RAID Level 6 and Beyond

- Goals
  - Better fault tolerance
- Extended Hamming code
- Specialized Eraser Codes
  - IBM Even-Odd, NetApp RAID-DP
- Beyond RAID-6
  - Reed-Solomon codes, using MOD 4 equations
  - Can be generalized to deal with k (>2) disk failures





Dealing with Disk Failures

- What failures
  - Power failures
  - Disk failures
  - Human failures
- What mechanisms required
  - NVRAM for power failures
  - Hot swappable capability
  - Monitoring hardware
- RAID reconstruction
  - Reconstruction during operation
  - What happens if a reconstruction fail?
  - What happens if the OS crashes during a reconstruction

Next Generation: FLASH

- Flash chip density increases on the Moore's law curve
  - 1995 16 Mb NAND flash chips
  - 2005 16 Gb NAND flash chips
  - 2009 64 Gb NAND flash chips
  - Doubled each year since 1995

- Market driven by Phones, Cameras,...

Flash Memory

- NOR
  - Byte addressable
  - Often used for BIOS
  - Much higher price than for NAND
- NAND
  - Dominant for consumer and enterprise devices
  - Single Level Cell (SLC) vs. Multi Level Cell (MLC):
    - SLC is more robust but expensive
    - MLC offers higher density and lower price

NAND Memory Organization

- Organized into a set of *erase blocks (EB)*
- Each *erase block* has a set of *pages*
- Example configuration for a 512 MB NAND device:
  - 4096 EB's, 64 pages per EB, 2112 bytes per page (2KB user data + 64 bytes metadata)
- Read:
  - Random access on any page, multiple times
  - 25-60μs
- Write
  - Data must be written sequentially to pages in an erase block
  - Entire page should be written for best reliability
  - 250-900μσ
- Erase:
  - Entire erase block must be erased before re-writing
  - Up to 3.5ms

What's Wrong With FLASH?

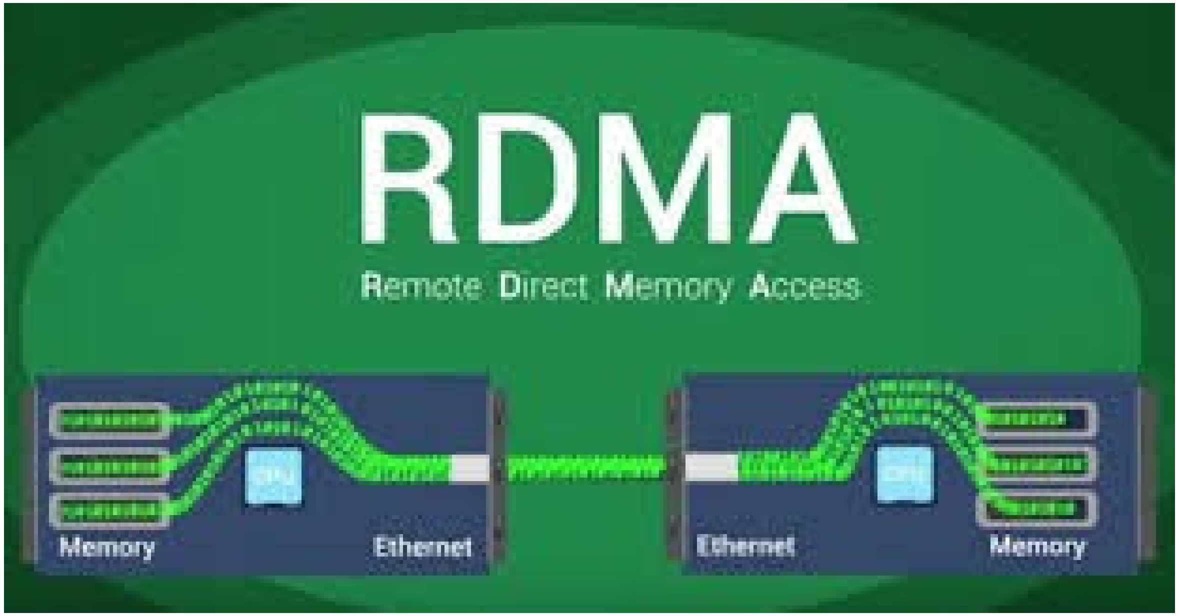
- Expensive: $/GB
  - 2x less than cheap DRAM
  - 50x more than disk today
- Limited lifetime
  - ~100k to 1M writes / page (single cell)
  - ~15k to 1M writes / page (single cell)
  - requires "wear leveling"
- Current performance limitations
  - Slow to write can only write 0's, so erase (set all 1) then write
  - Large (e.g. 128K) segments to erase

Non-volatile DRAM (NVRAM)

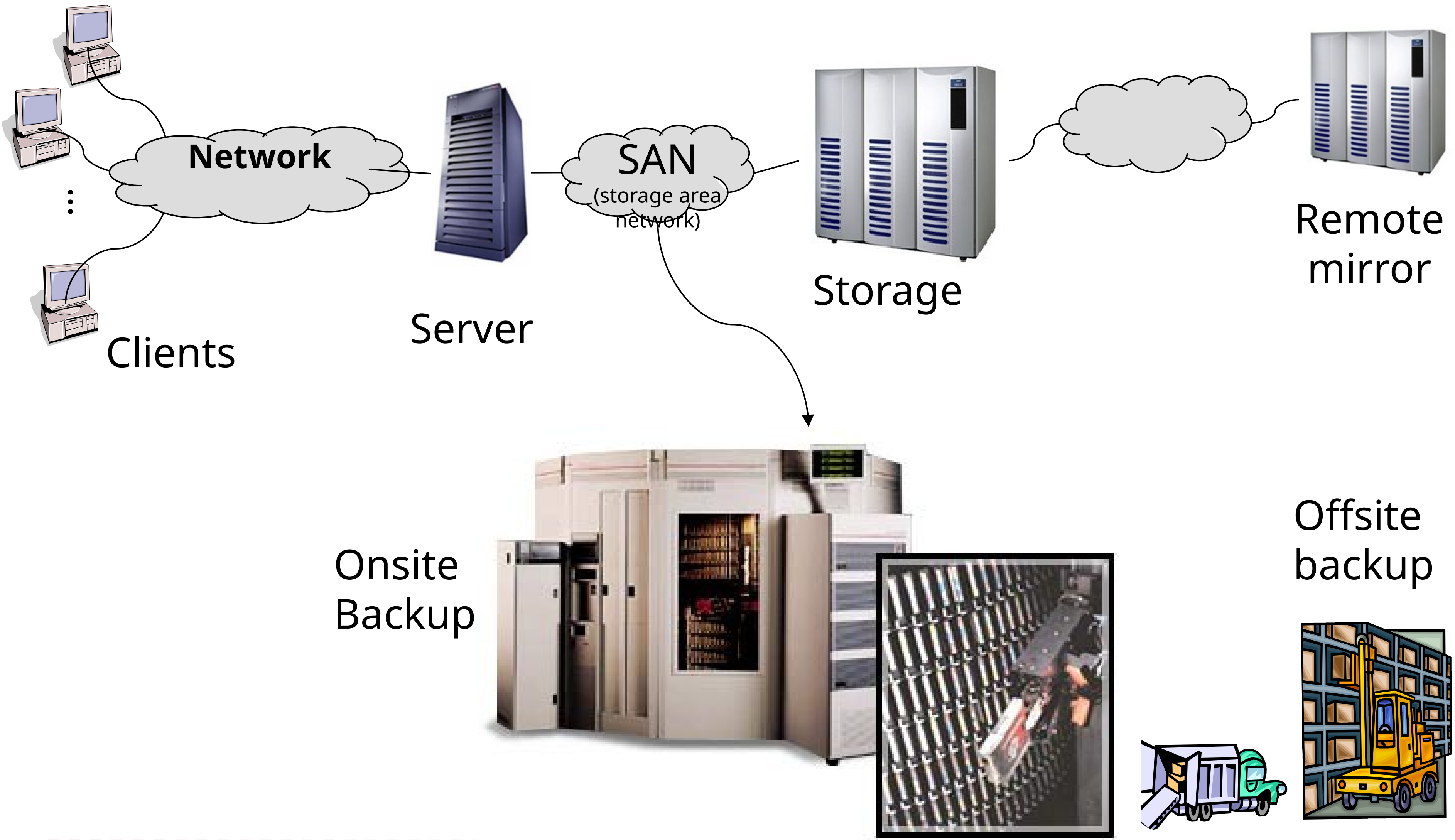
(Netlist Nvvault)

- Battery backed DRAM
  - Backup power during power-out
  - Ordinary DRAM technology
- One part of a storage system
- Expensive
- Targeted at specific application domains such as databases

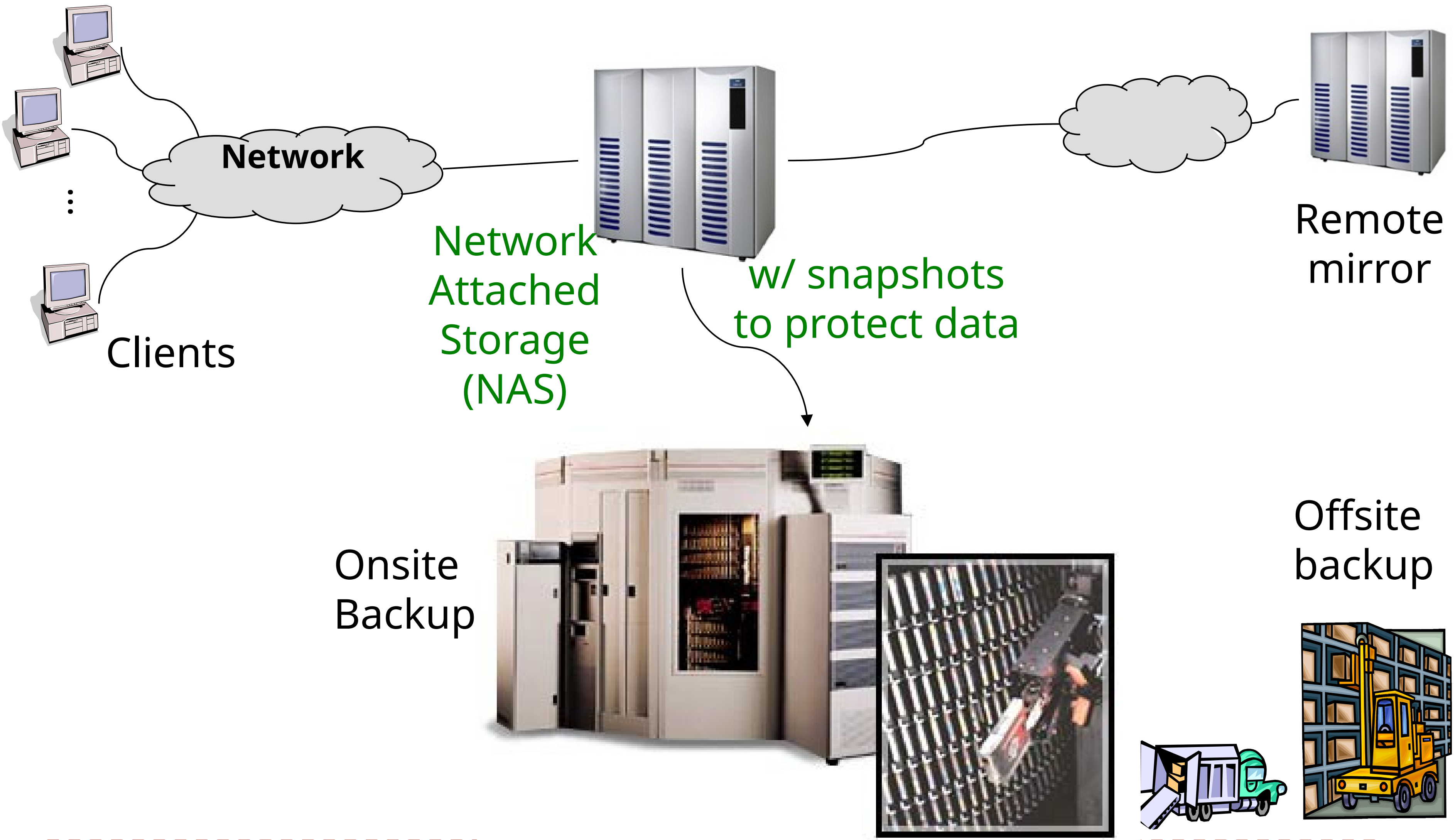
Remote Direct Memory Access



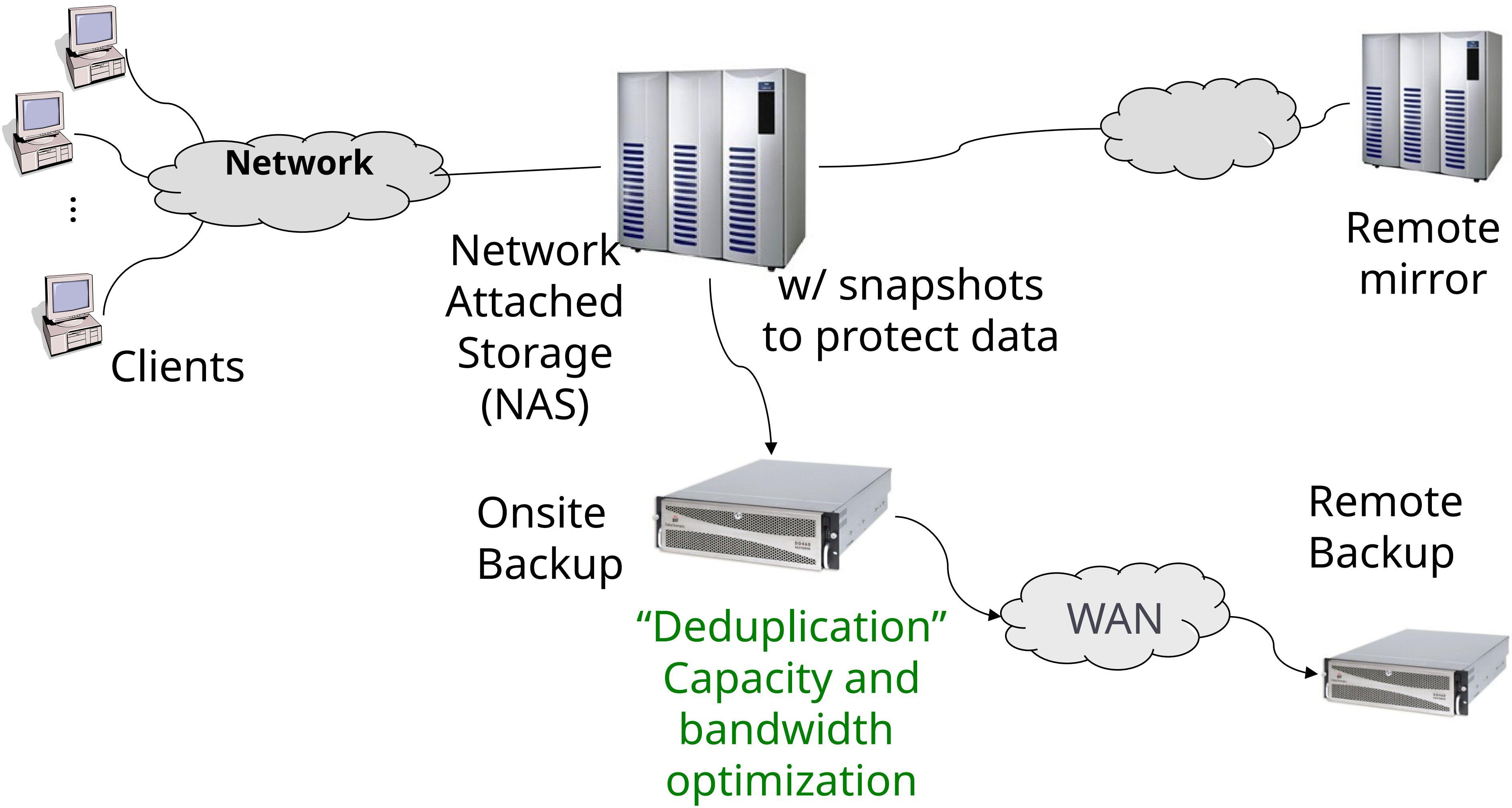
Traditional Storage Center Hierarchy



Evolved Storage Center Hierarchy



Modern Storage Center Hierarchy



Summary

- Disk is complex
  - Architecture, cache, seek algorithms etc
- Disk real density is on Moore’s law curve
- Large data disk accesses needed to achieve good throughput
- RAID adds reliability and increases throughput
- Failures should be considered
- Flash memory is now common