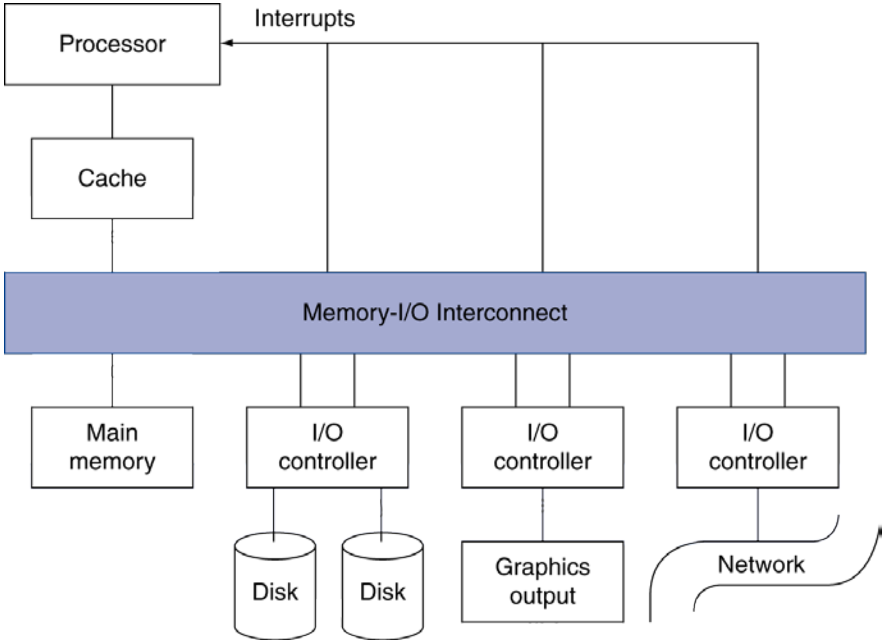
Computer Architecture

Lecture

Storage and Other I/O Topics

Introduction

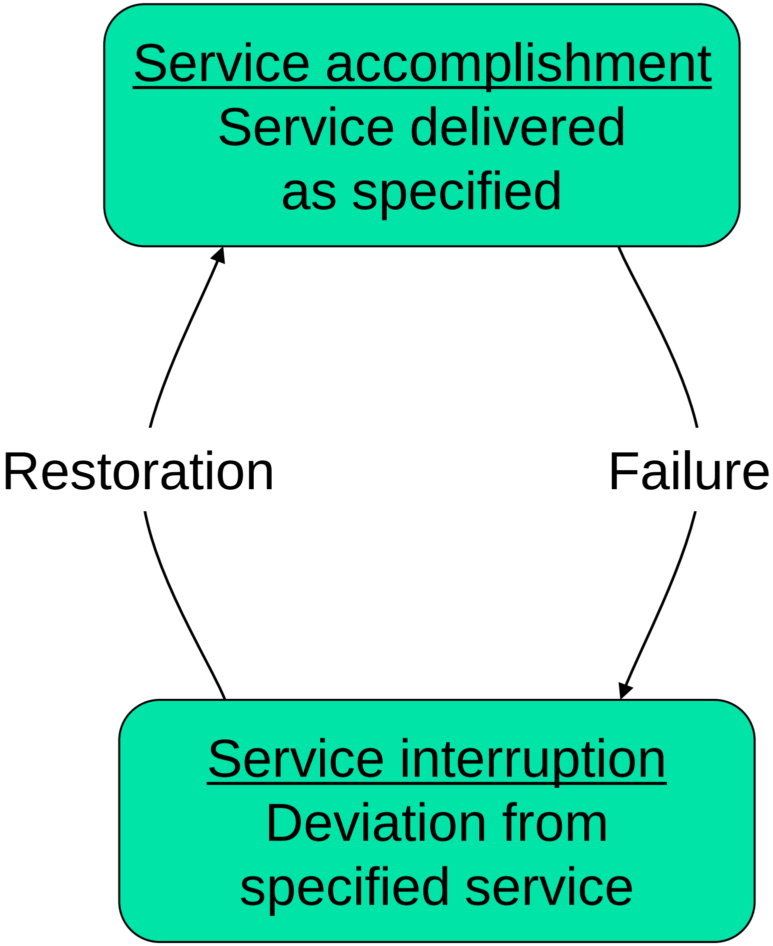
- I/O devices can be characterized by
  - Behaviour: input, output, storage
  - Partner: human or machine
  - Data rate: bytes/sec, transfers/sec
- I/O bus connections



Diverse I/O Devices

Device	Behavior	Partner	Data rate (Mbit/sec)
Keyboard	input	human	0.0001
Mouse	input	human	0.0038
Voice input	input	human	0.2640
Sound input	input	machine	3.0000
Scanner	input	human	3.2000
Voice output	output	human	0.2640
Sound output	output	human	8.0000
Laser printer	output	human	3.2000
Graphics display	output	human	800~8000
Cable Modem	input or output	machine	0.1280~6.0000
Network/ LAN	input or output	machine	100~10000
Network/ wireless LAN	input or output	machine	11~54
Optical disk	storage	machine	80~220
Flash memory	storage	machine	32~200
Magnetic disk	storage	machine	800~3000

Dependability



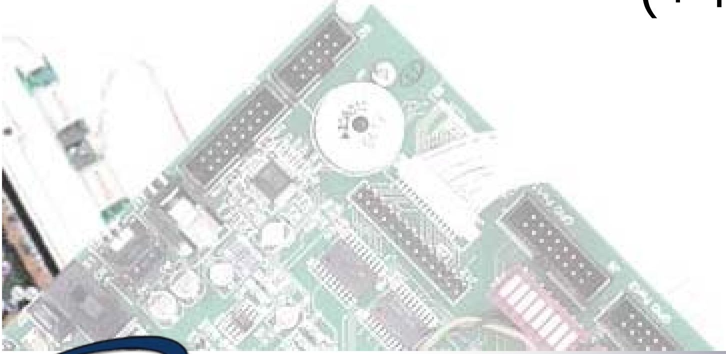
- **Fault: failure of a component**
  - **May or may not lead to system failure**

Dependability Measures

- Reliability: mean time to failure (MTTF)
- Service interruption: mean time to repair (MTTR)
- Mean time between failures
  - MTBF = MTTF + MTTR
- Availability = MTTF / (MTTF + MTTR)
- Improving Availability
  - Increase MTTF: fault avoidance, fault tolerance, fault forecasting
  - Reduce MTTR: improved tools and processes for diagnosis and repair

하드디스크의 실체 및 플래시메모리

민 상 렬 (symin@snu.ac.kr)
(+ Mobile Embedded Systems Lab 연구원들)
서울대학교 공과대학
컴퓨터공학부



Outline

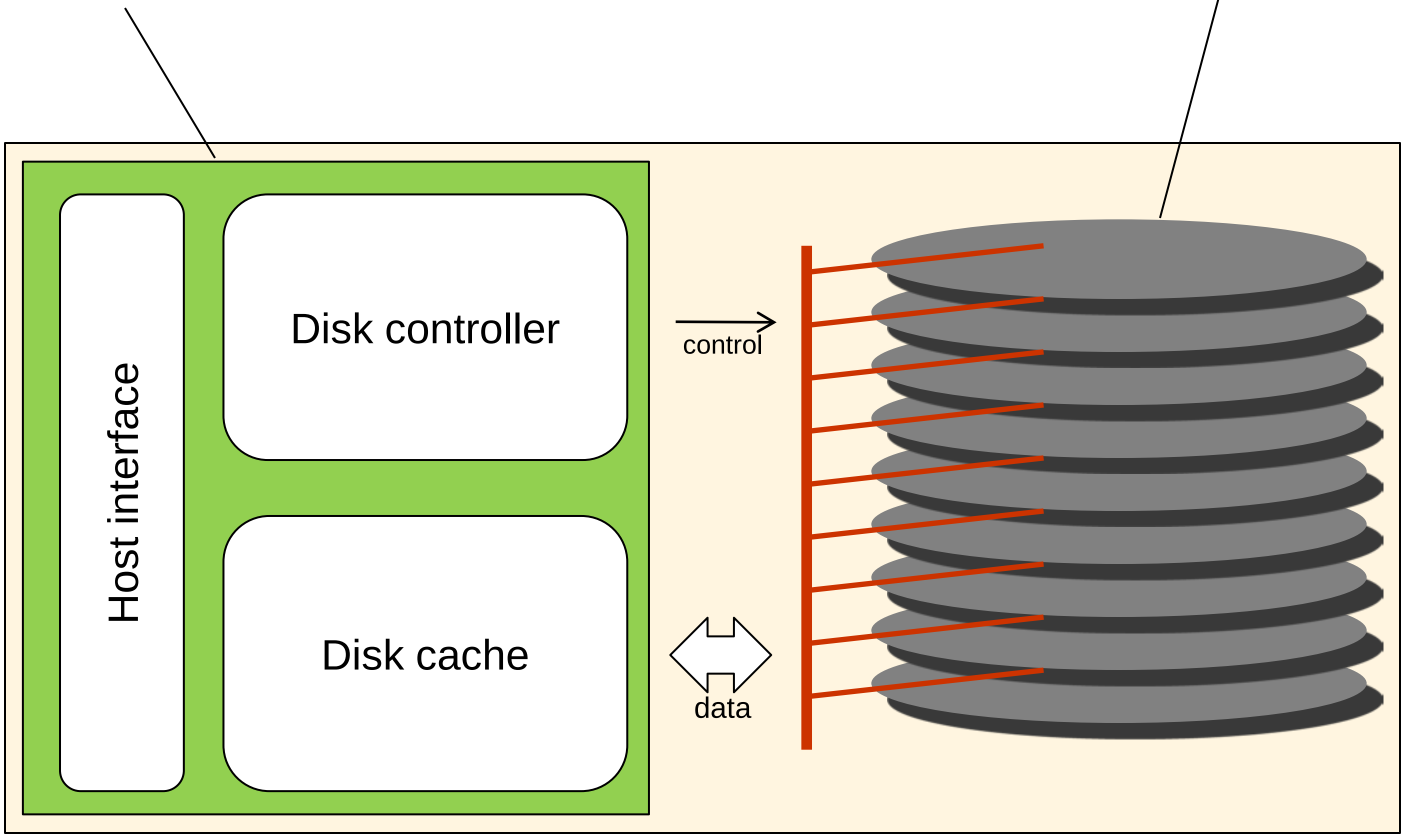
- HDD Basics and Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions

Outline

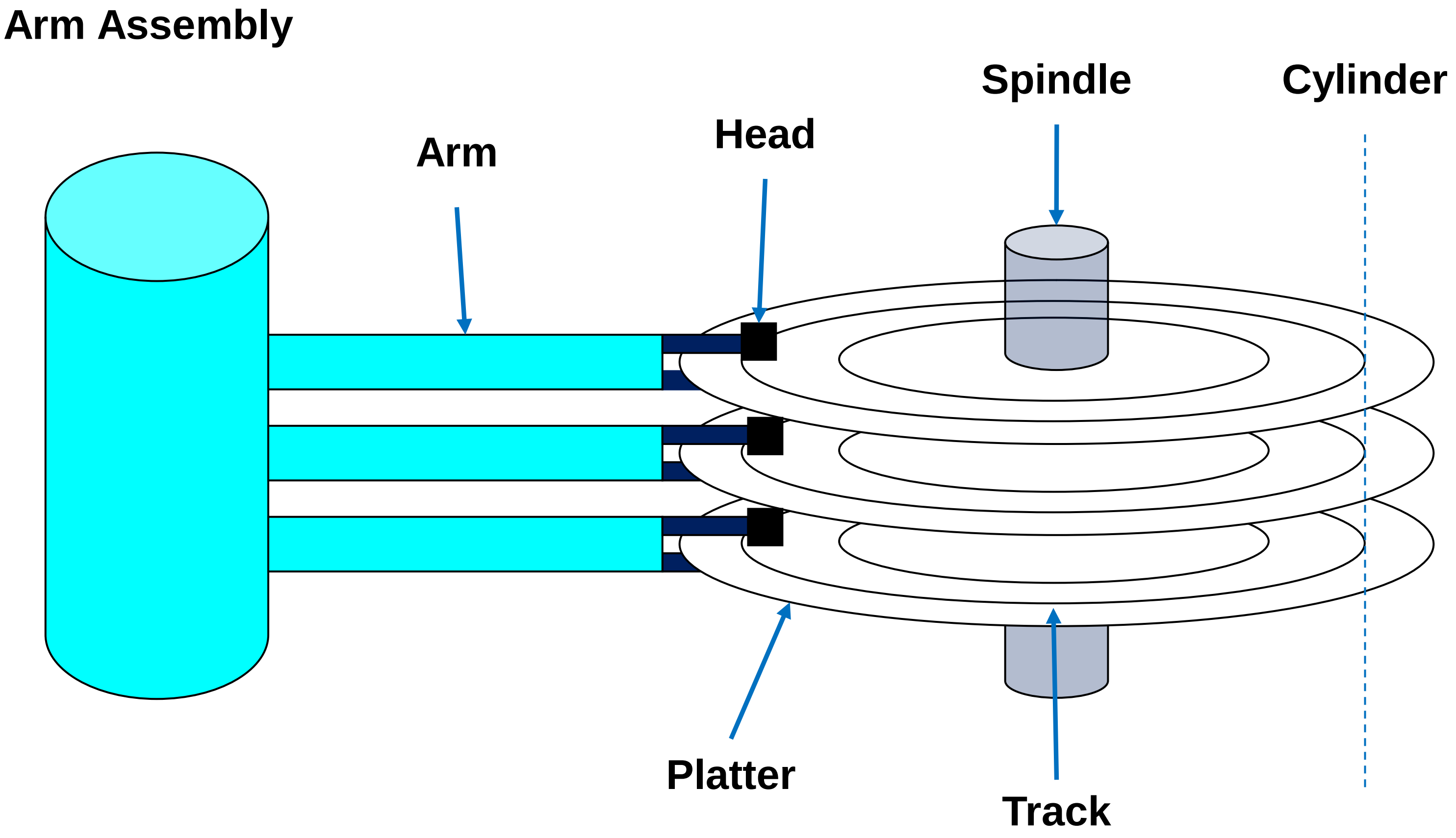
- HDD Basics and Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions

HDD internals

- Electronic components
- Mechanical components



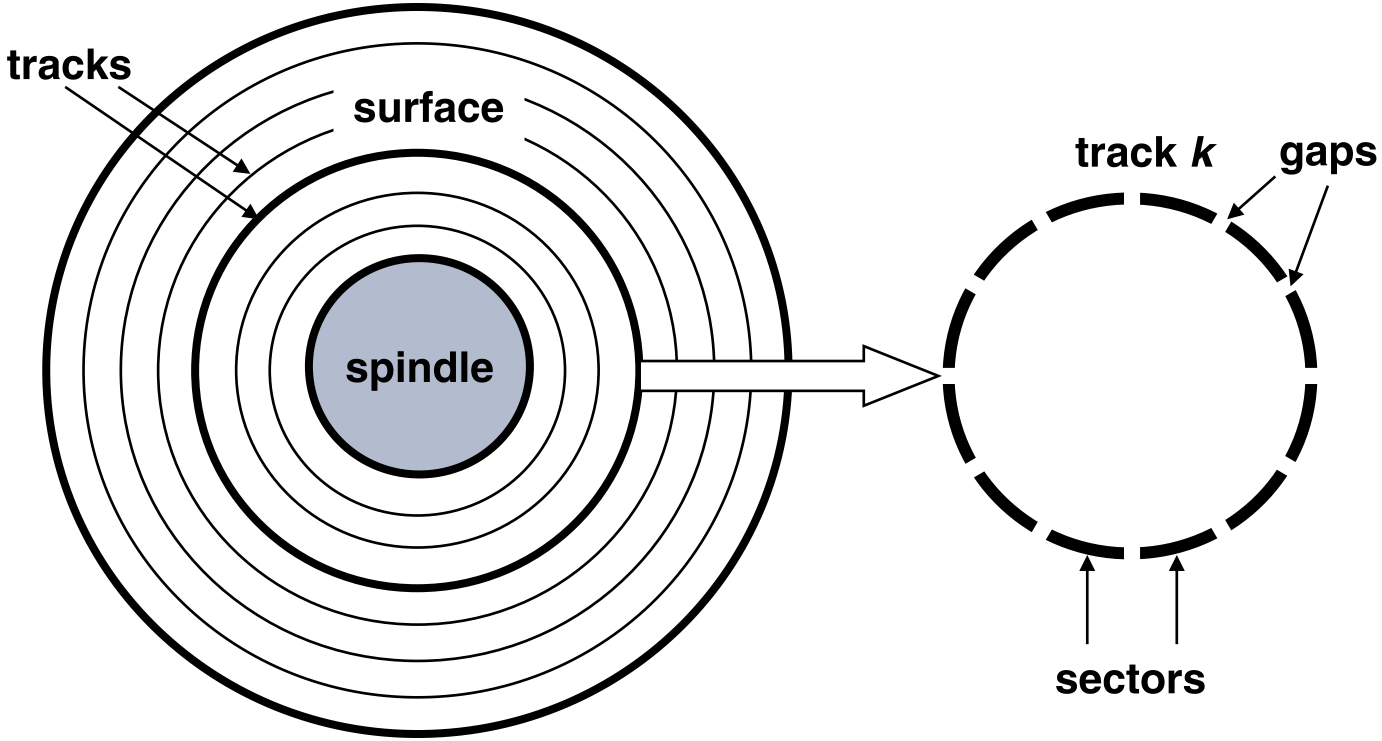
Mechanical components



Source: *"ABCs of Disk Drives,"* Sudhanva Gurumurthi

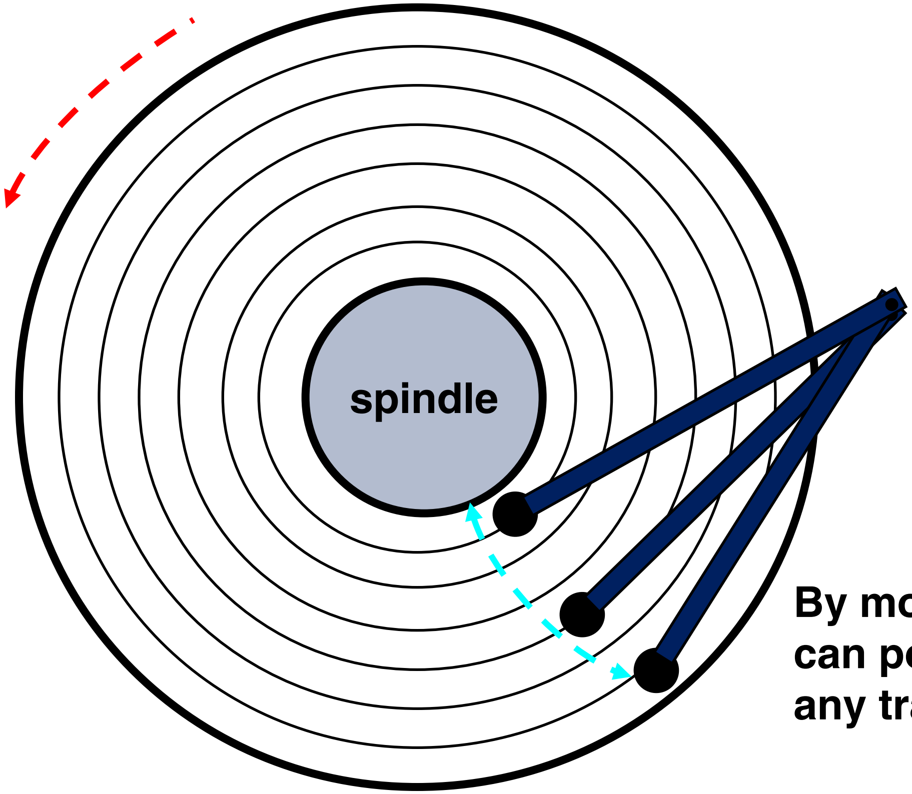
Data layout

- Rotating disks consist of platters, each with two surfaces
- Each surface consists of concentric rings called tracks
- Each track consists of sectors separated by gaps



Source:
"http://camars.kaist.ac.kr/~joon/course/sep562_2006_1/notes/10_11%20Memory_Hierarchy.ppt"

Disk operation



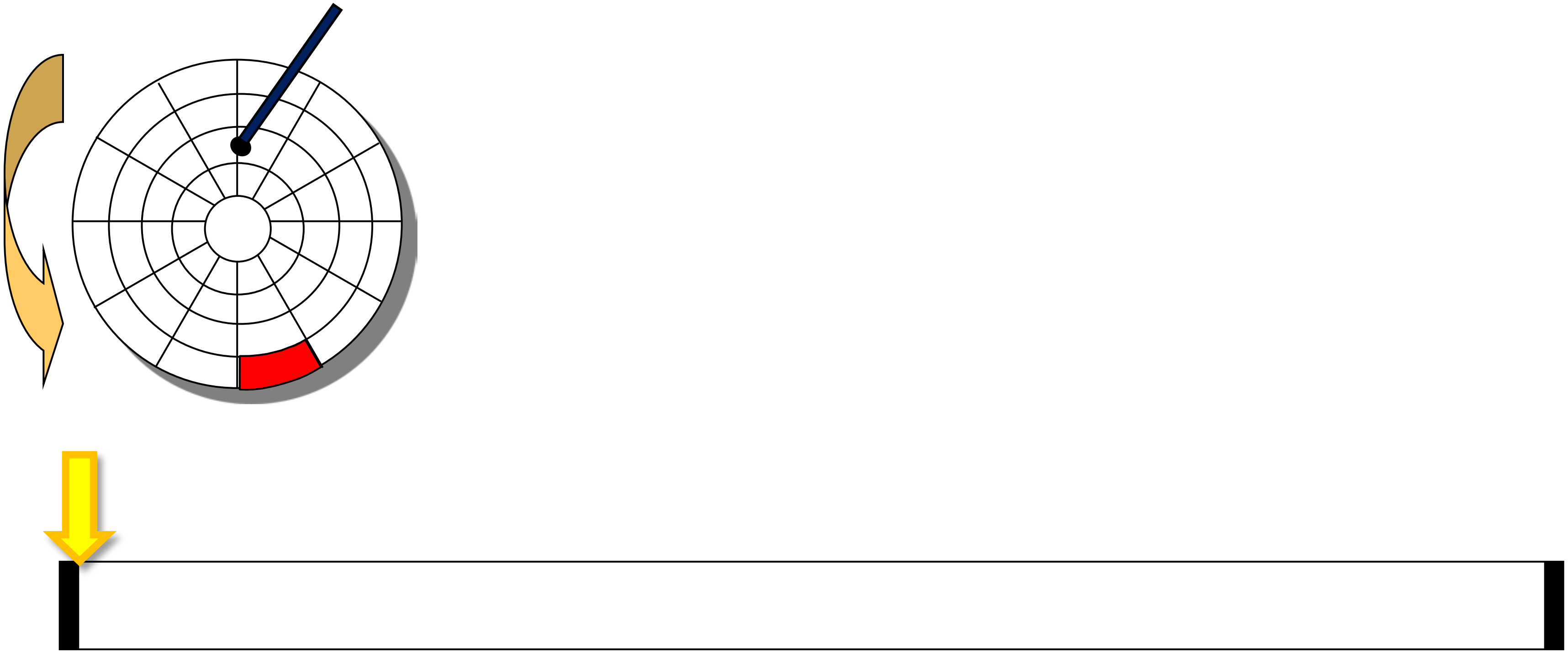
The disk surface spins at a fixed rotational rate

The head is attached to the end of the arm and flies over the disk surface on a thin cushion of air

By moving radially, the arm can position the head over any track

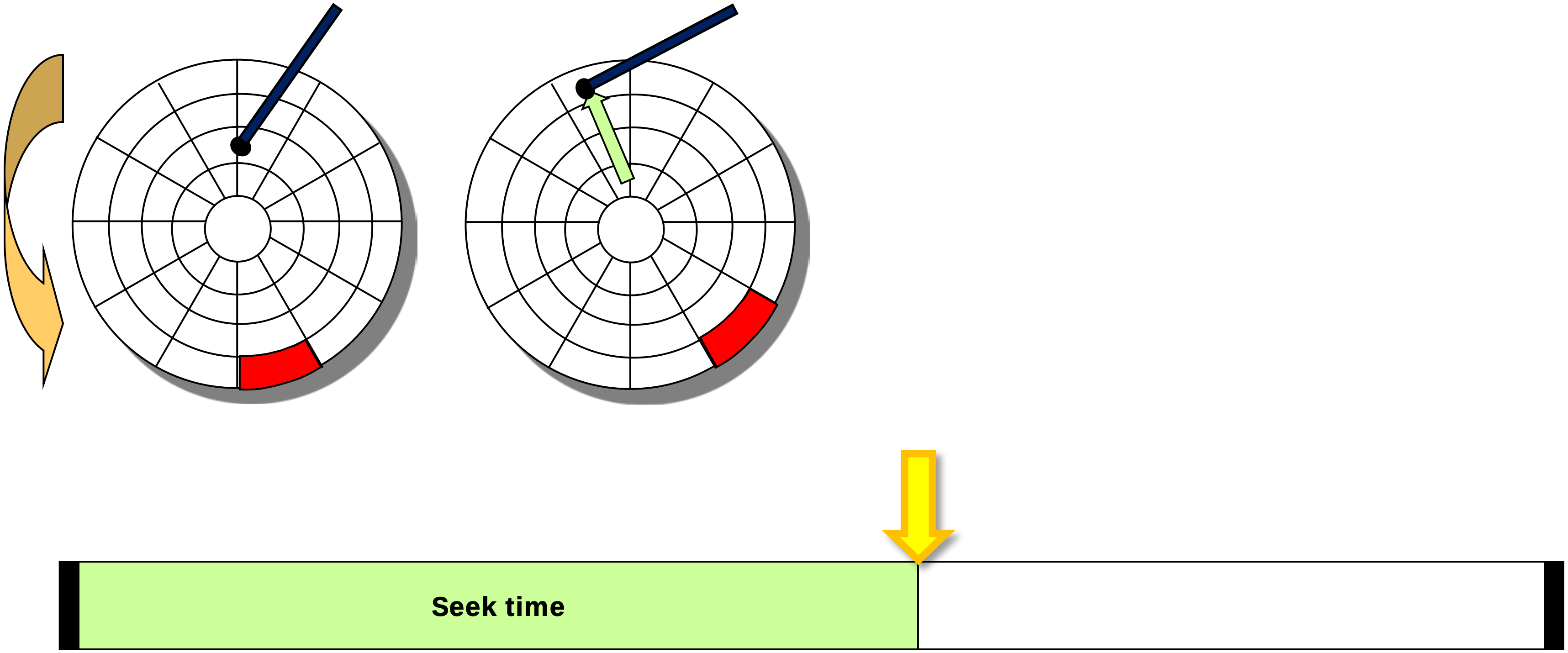
Source:
"http://camars.kaist.ac.kr/~joon/course/sep562_2006_1/notes/10_11%20Memory_Hierarchy.ppt"

Disk operation details



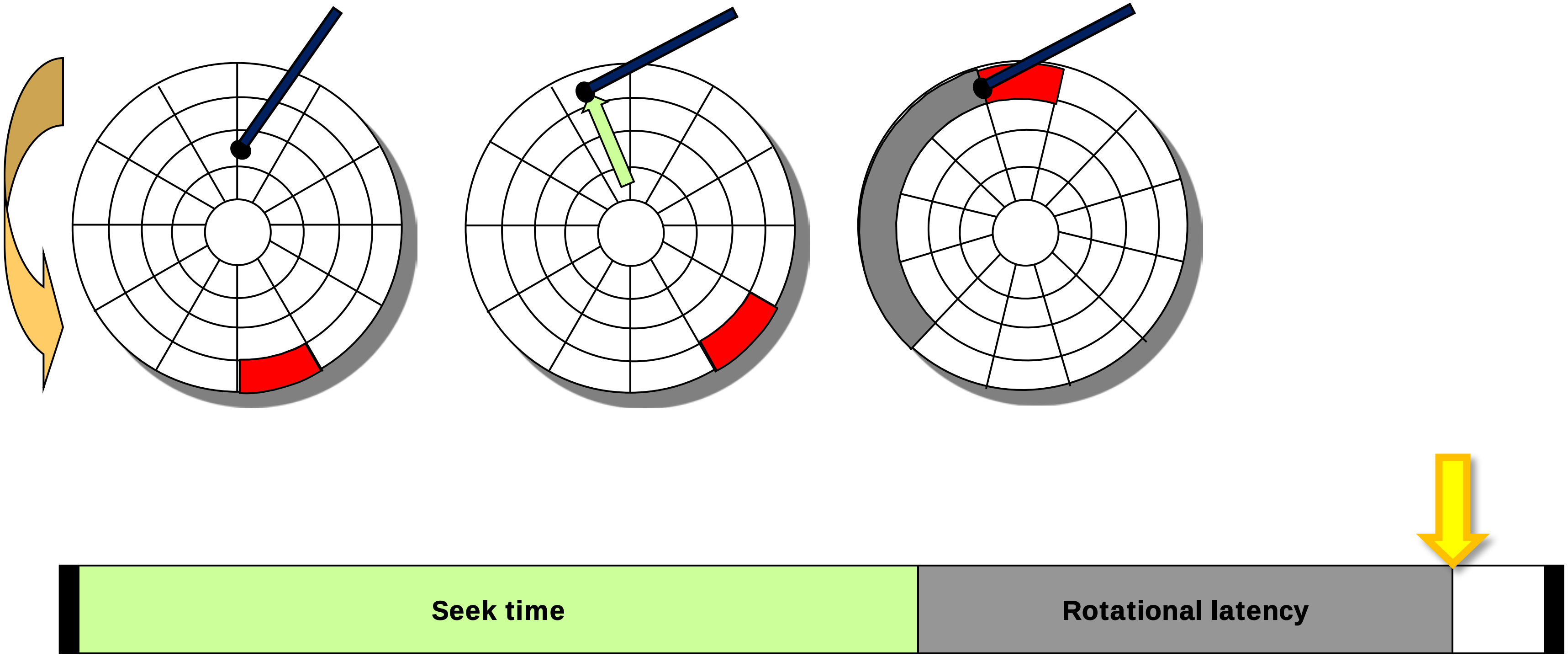
Source: *“http://www.cs.duke.edu/~chase/cps110/slides/files1.ppt”*

Disk operation details



Source: *"http://www.cs.duke.edu/~chase/cps110/slides/files1.ppt"*

Disk operation details

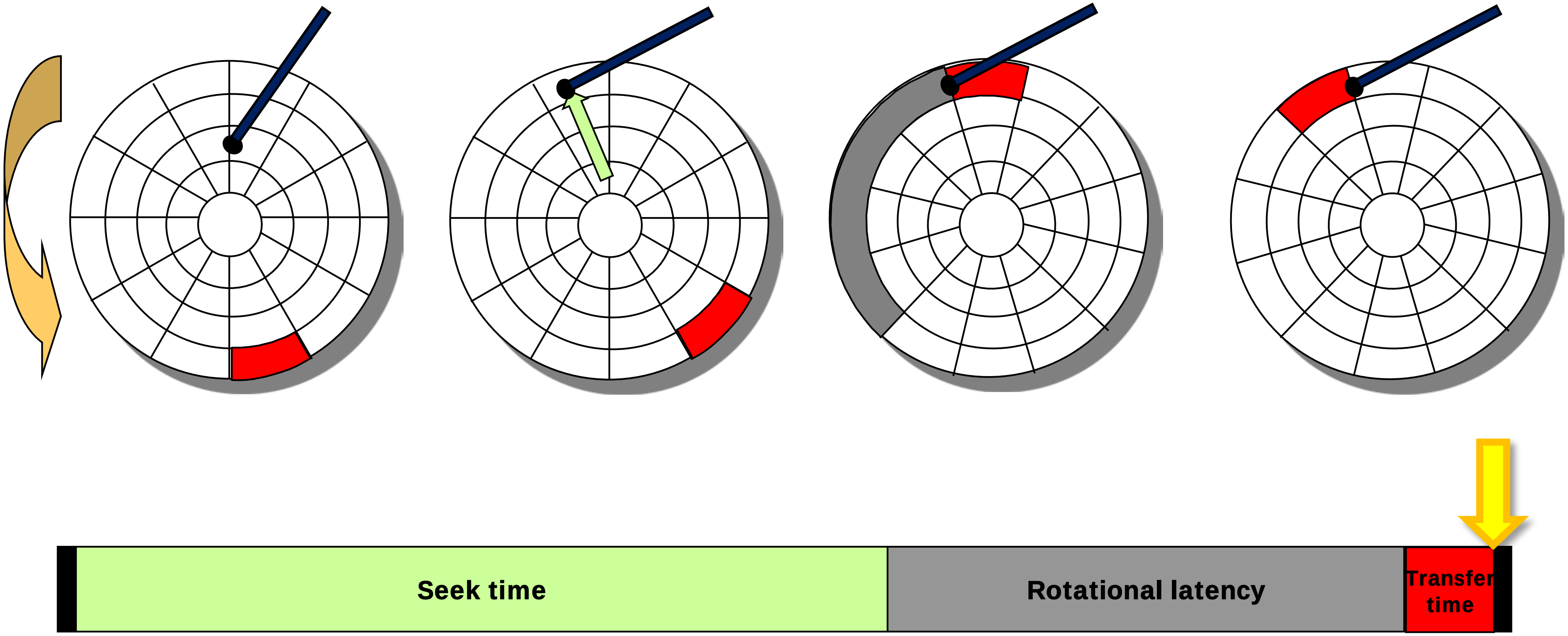


Seek time

Rotational latency

Source: *“http://www.cs.duke.edu/~chase/cps110/slides/files1.ppt”*

Disk operation details



Seek time

Rotational latency

Transfer time

Source: *"http://www.cs.duke.edu/~chase/cps110/slides/files1.ppt"*

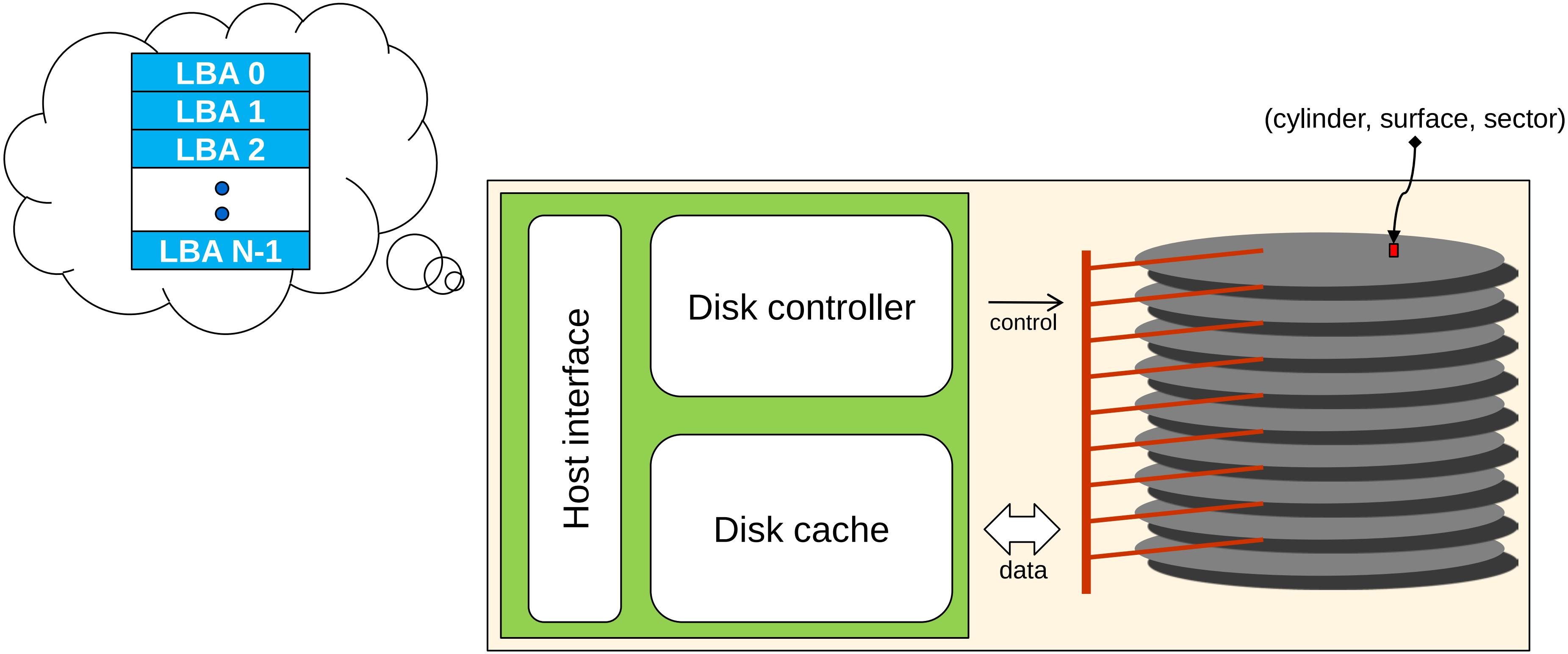
Disk access time

- Disk access time
  - Seek time + Rotational latency + Transfer time

- Seek time
  - Time to position heads over cylinder containing target sector
  - 0 ~ 25 ms
- Rotational latency
  - Time waiting for first bit of target sector to pass under r/w head
  - Full rotation: 4 ~ 12 ms (15000 ~ 5400 RPM)
- Transfer time
  - Time to read the bits in the target sector
  - 1 sector transfer: 1.3 ~ 12.8 us (380 ~ 40 MB/s transfer rate)

Electronic components

- Presenting a simple abstract view of the complex sector geometry



LBA 0

LBA 1

LBA 2

LBA N-1

(cylinder, surface, sector)

Host interface

Disk controller

Disk cache

control

data

Electronic components

- Disk controller
  - Controlling the overall system
  - Major functions
    - Host interface
    - Request translation (LBA ↔ [cylinder, surface, sector])
    - Reliability mechanism (e.g. ECC, bad sector handling)
    - Performance improvement (e.g. request scheduling and disk caching)
    - Power management (e.g. spin down of spindle motor)
  - Typically, embedded processor (such as ARM) + logic circuits

Outline

- HDD Basics and Demo
  - Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions

Demo HDD Specification

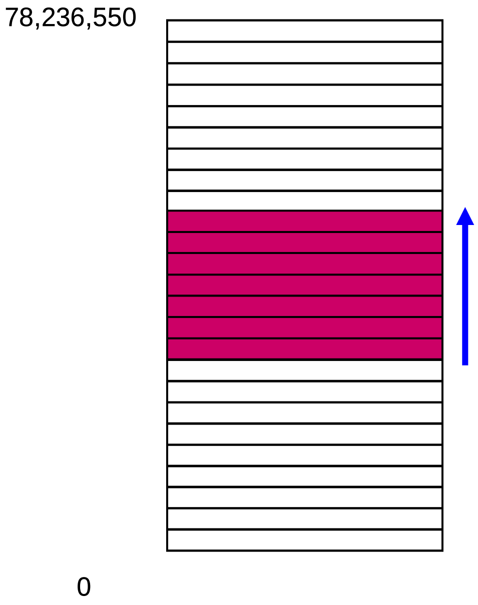
Model Name: SAMSUNG MP0402H (2.5 in)

- Size:
  - total 78,236,550 sectors
  - 40,057,113,600 bytes ≈ 37.30 GB
- Interface: ATA-6 (supports UDMA100)
- Buffer: 8MB DRAM
- Performance brief:
  - Avg. Seek time: 12 ms
  - Avg. Rotational Latency: 5.6 ms (5400 RPM)

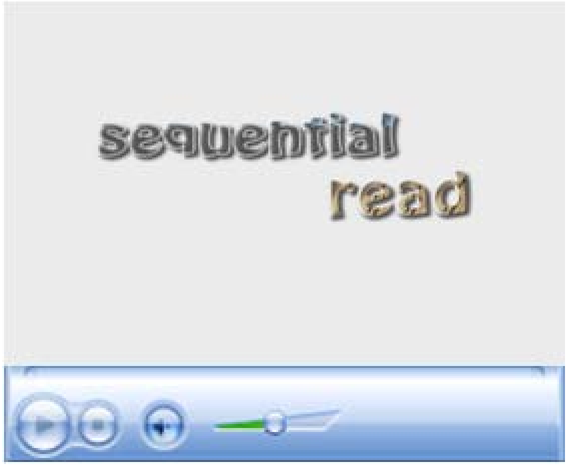
- reference url: http://www.samsung.com/Products/HardDiskDrive/SpinPointMSeries/HardDiskDrive_SpinpointMseries_MP0402H_sp.htm

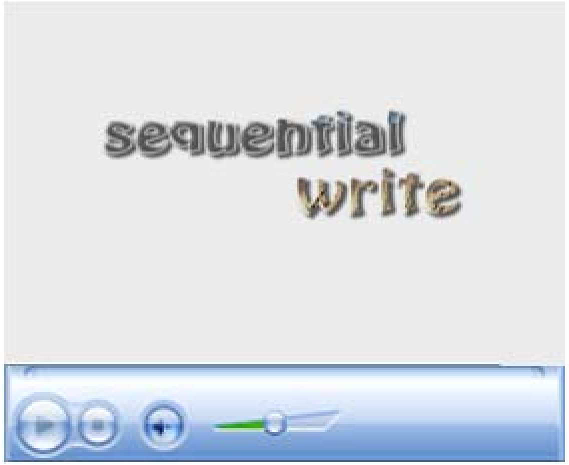
Demo I – Power-on sequence

Demo II – Sequential read/write

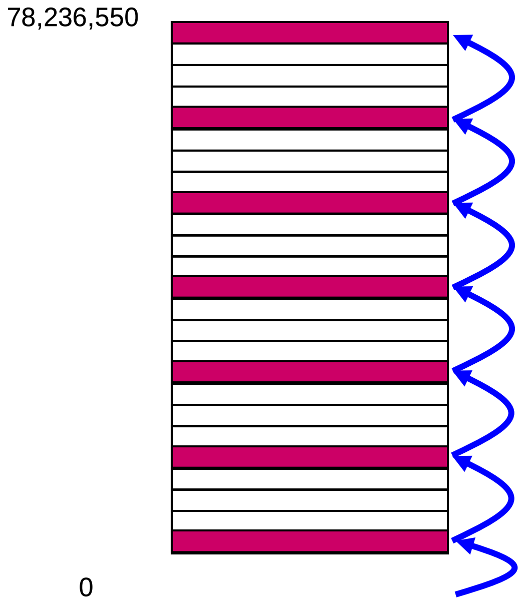


- Access pattern
  - read/write data whose address increases continuously

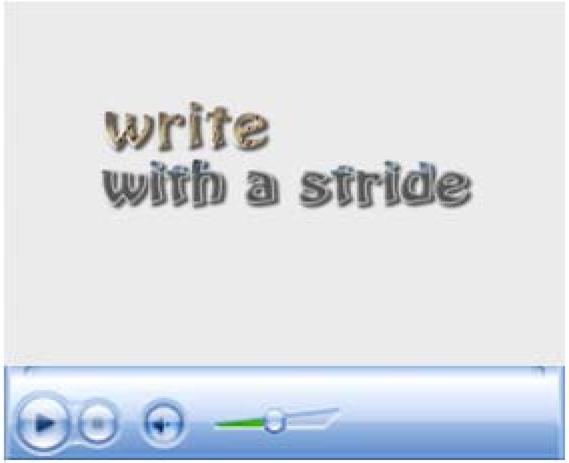




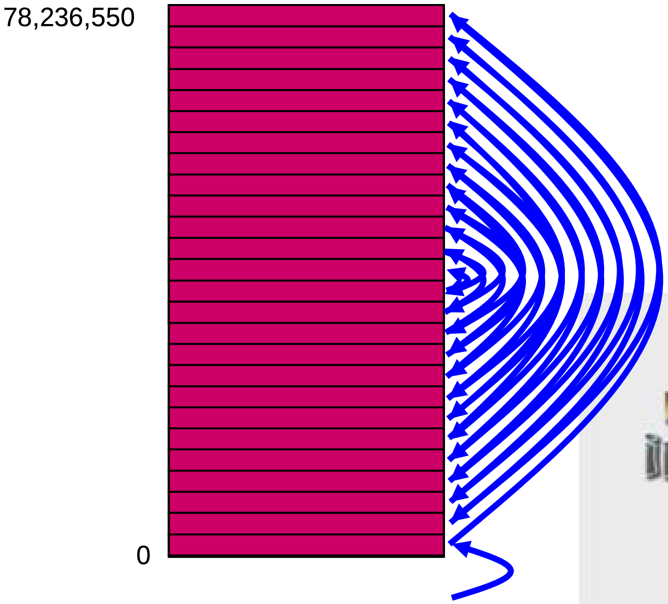
Demo III – Read/Write with a stride



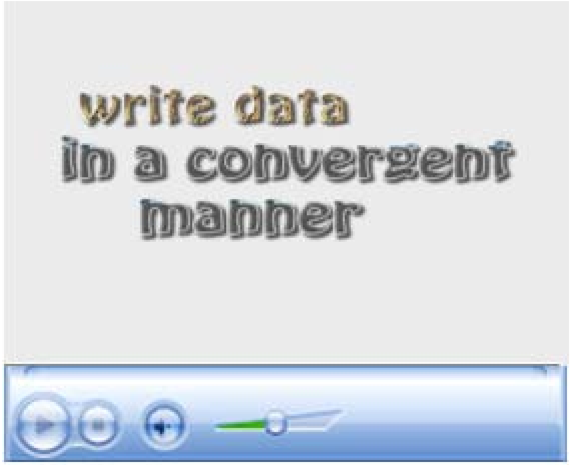
- Access pattern
  - read/write data whose address increases with a regular interval



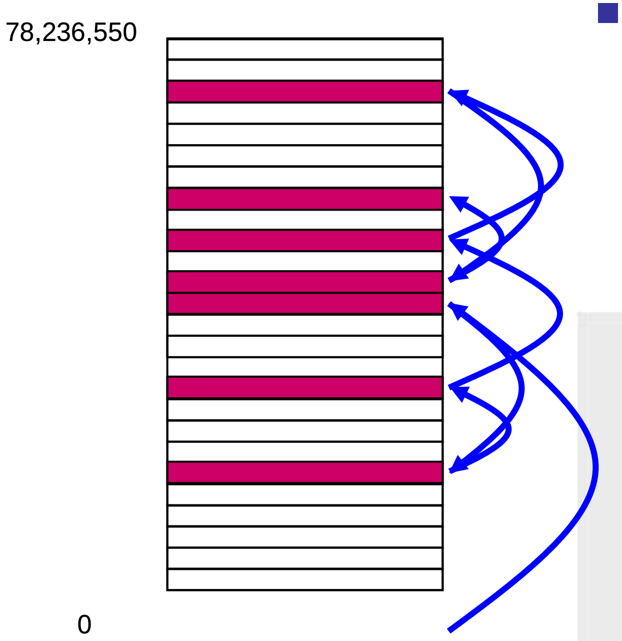
Demo IV – Read/Write in a convergent manner



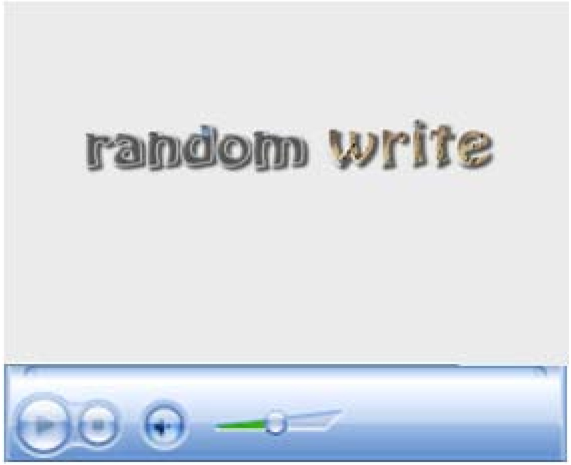
- Access pattern
  - read/write data whose address is not overlapped and is in a convergent manner



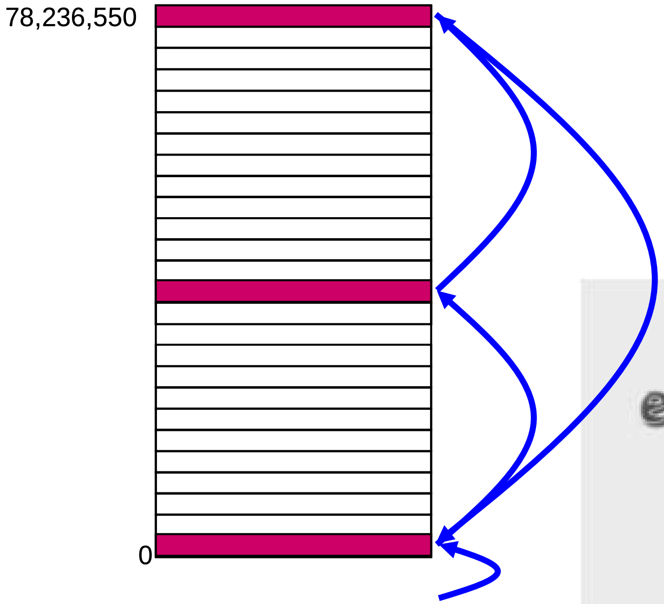
Demo V – Random read/write



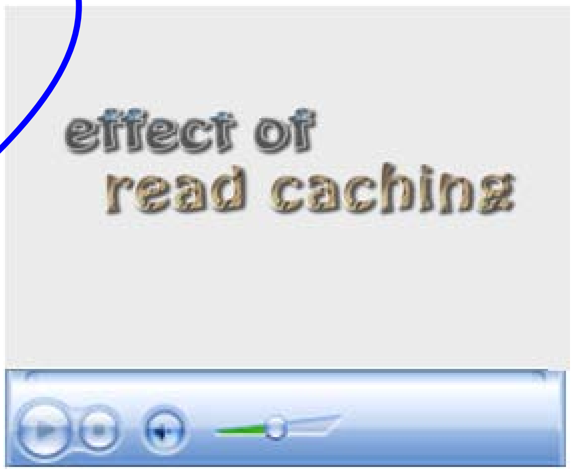
- Access pattern
  - read/write random addresses

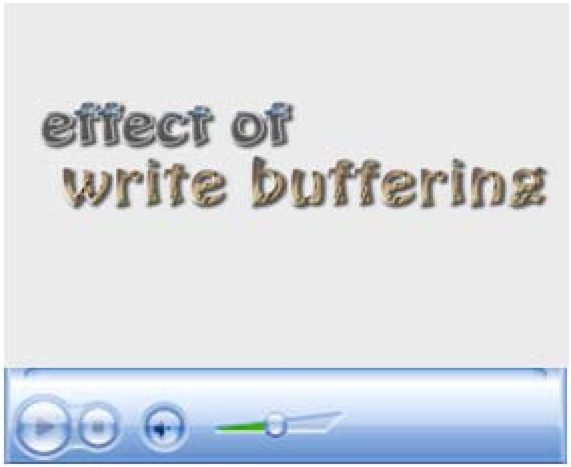


Demo VI – Effect of read caching/write buffering



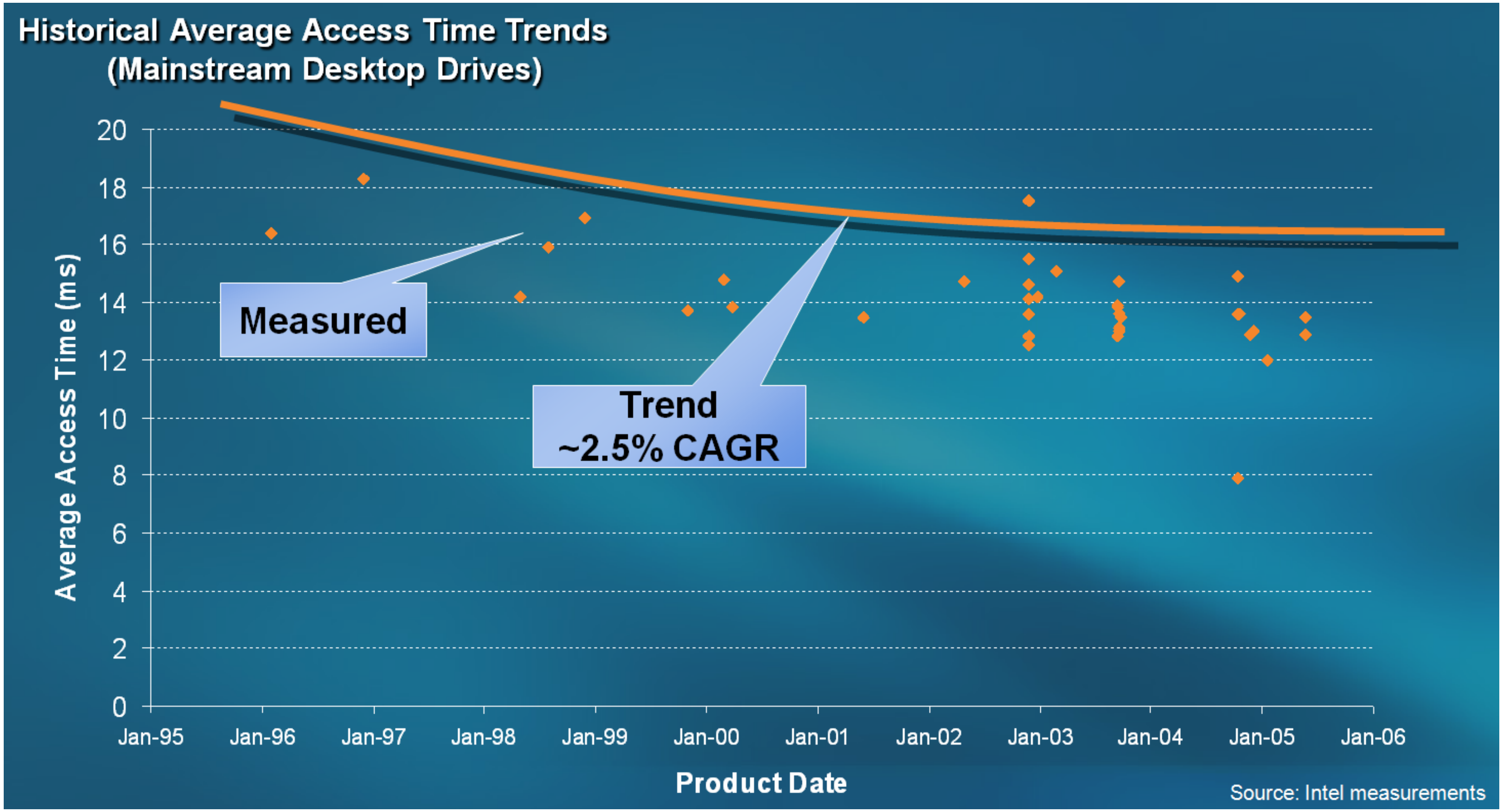
- Access pattern
  - access on the fixed addresses repeatedly





Demo VII – Windows XP start-up

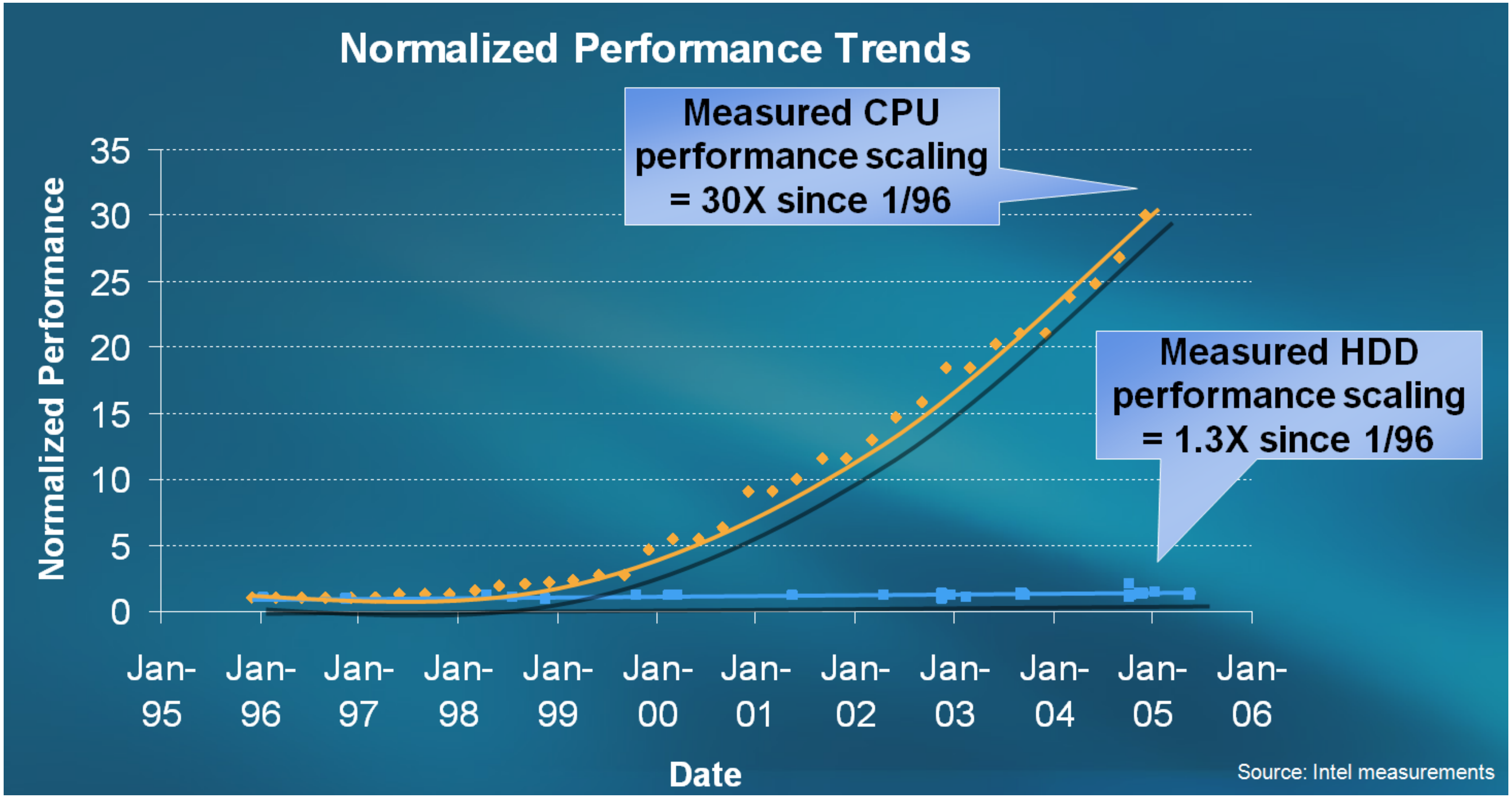
HDD performance trends (1)



Source: Intel measurements

- HDD access time trends are fairly flat due to mechanical nature of device

HDD Performance trends (2)



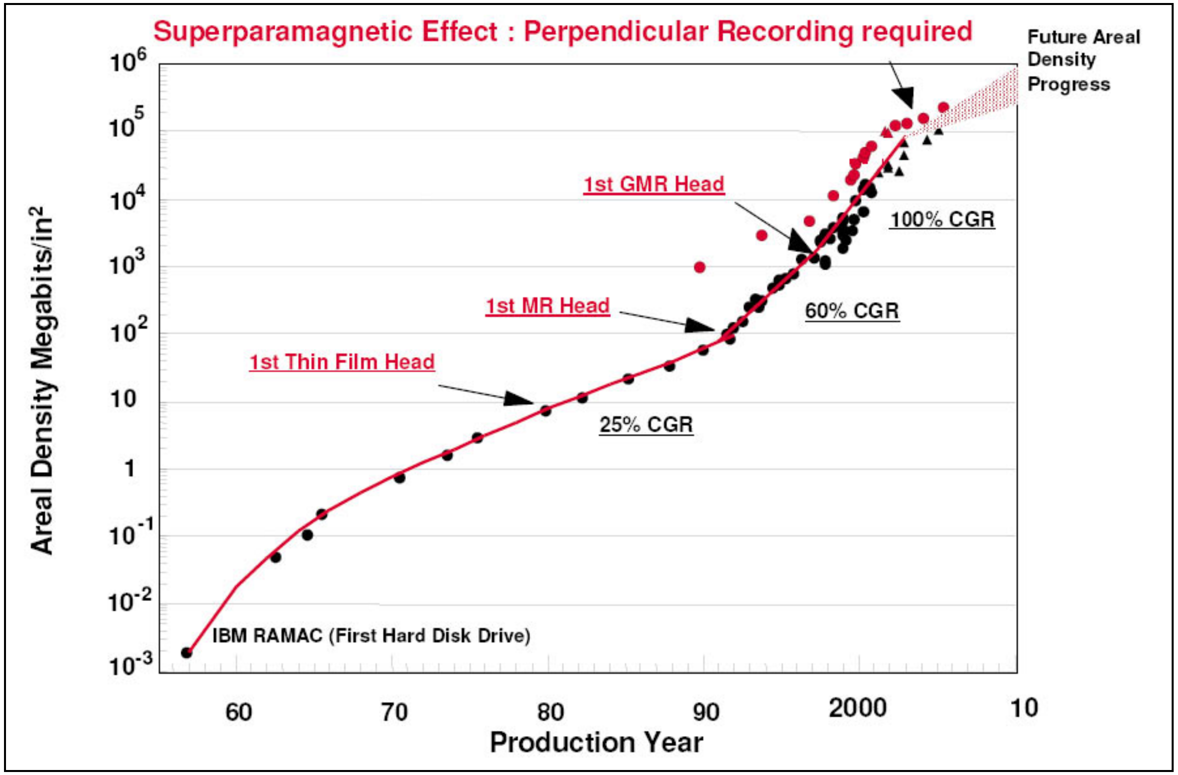
Measured CPU performance scaling = 30X since 1/96

Measured HDD performance scaling = 1.3X since 1/96

Source: Intel measurements

- A workload that was 5% disk bound in ‘96 would be 55% disk bound in ‘05

HDD density trends



Source: Hitachi Global Storage Technologies

HDD Summary

The Ugly

- **Latent sector errors**

The Bad

- **High latency**
- **High power consumption**
- **Low reliability**
- **Large form factor**
- **Limited parallelism**

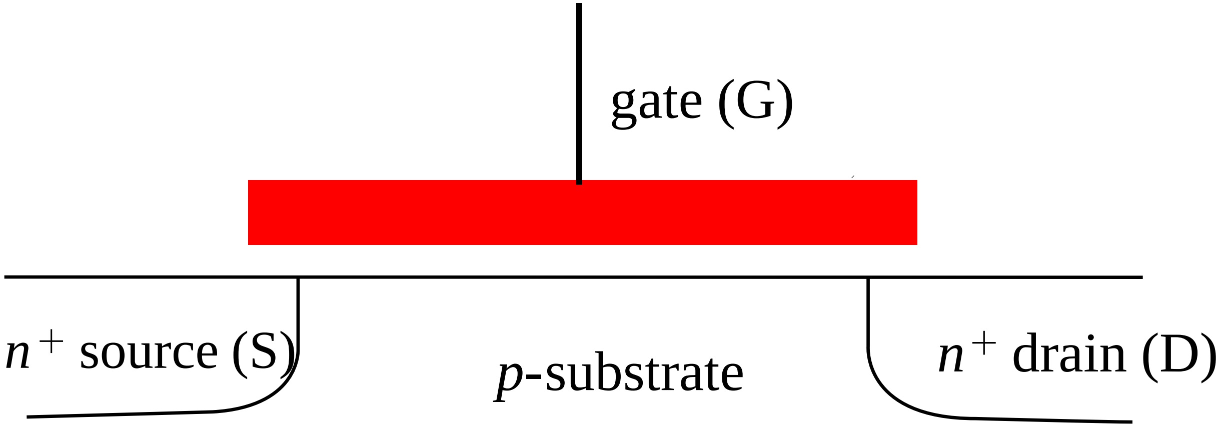
The Good

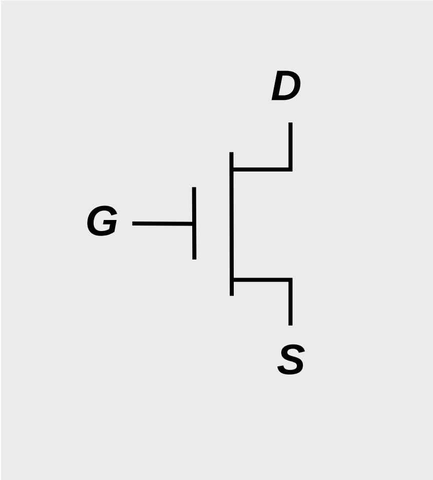
- **High capacity**
- **Low cost**

Outline

- HDD Basics and Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions

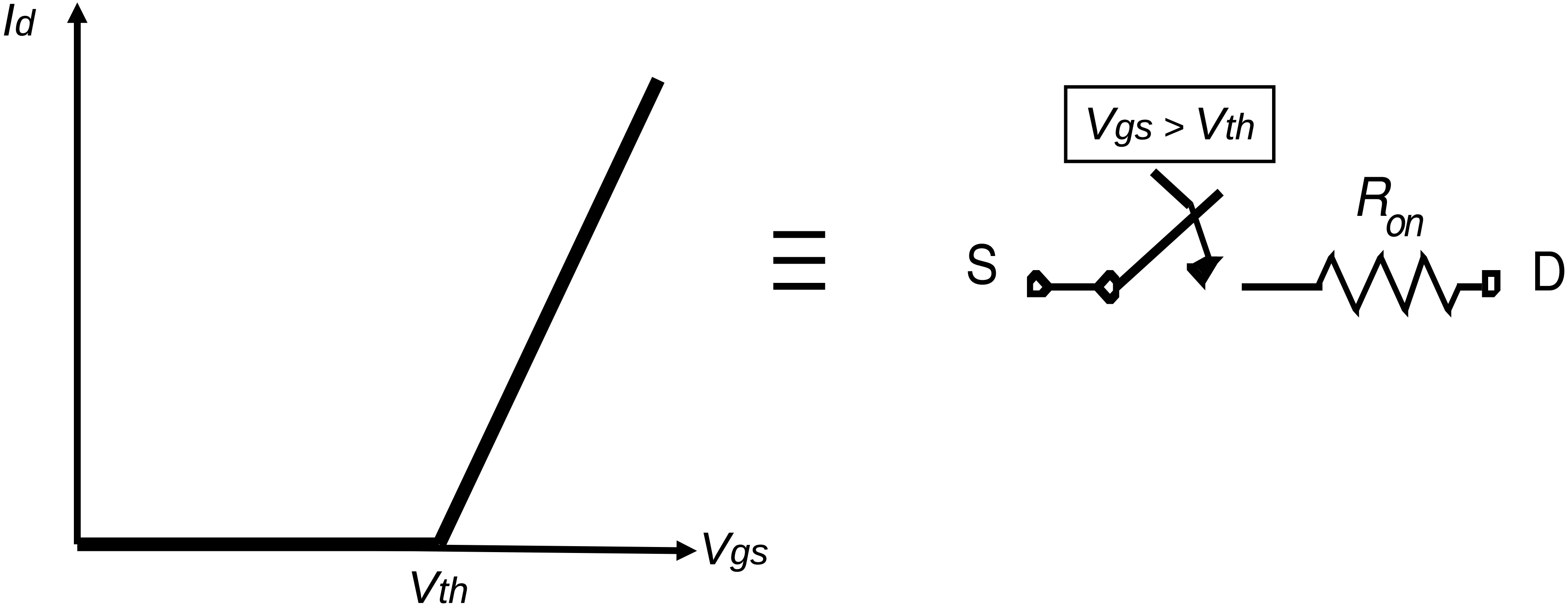
Conventional MOS Transistor



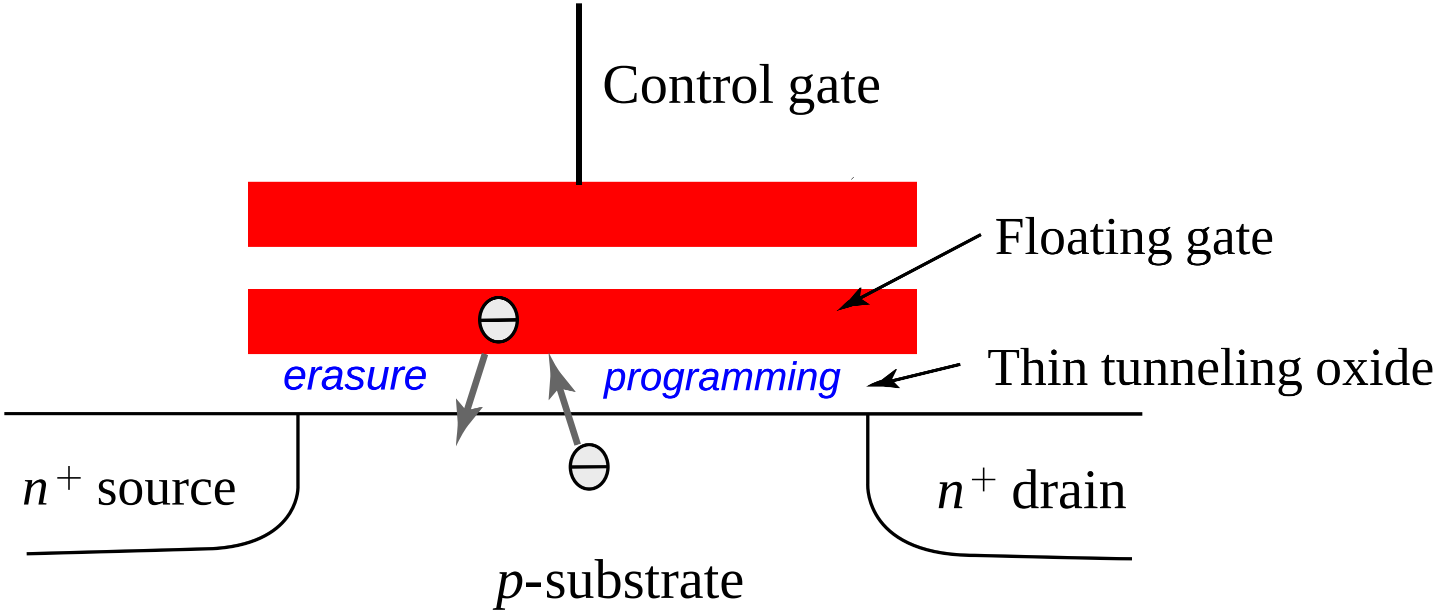


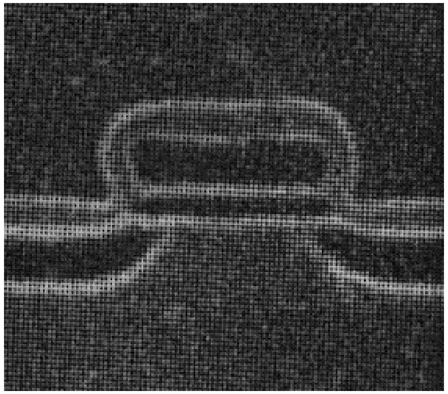
Schematic symbol

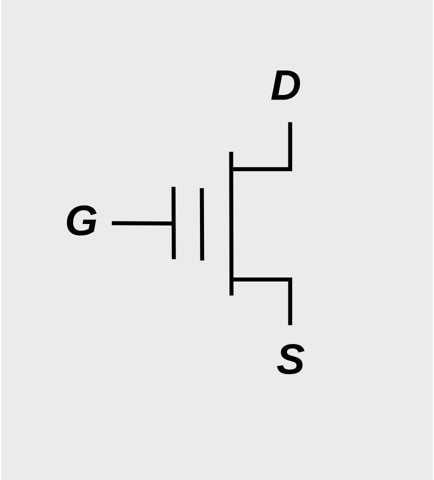
Conventional MOS Transistor: A Constant-Threshold Transistor



Flash Memory

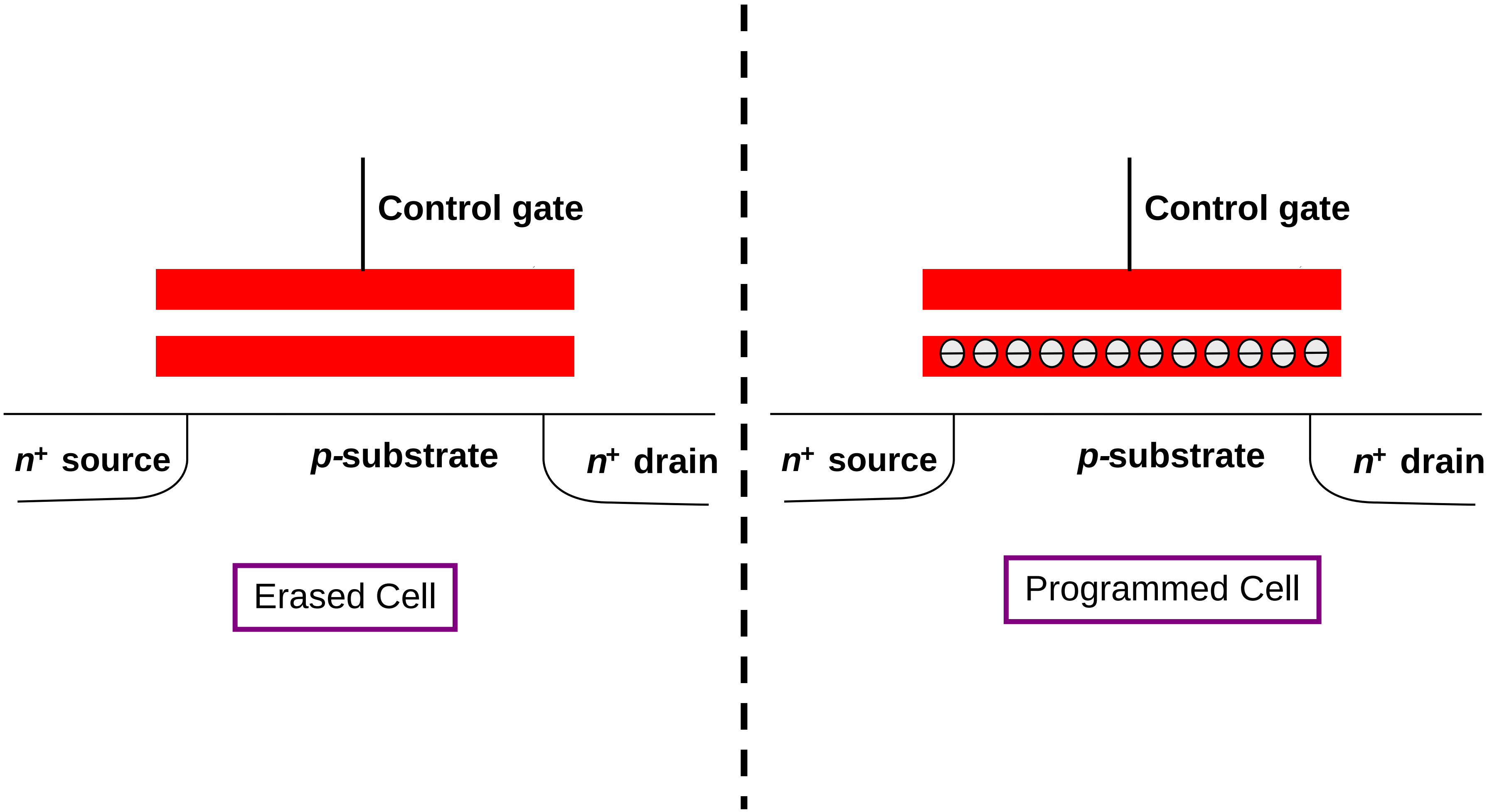




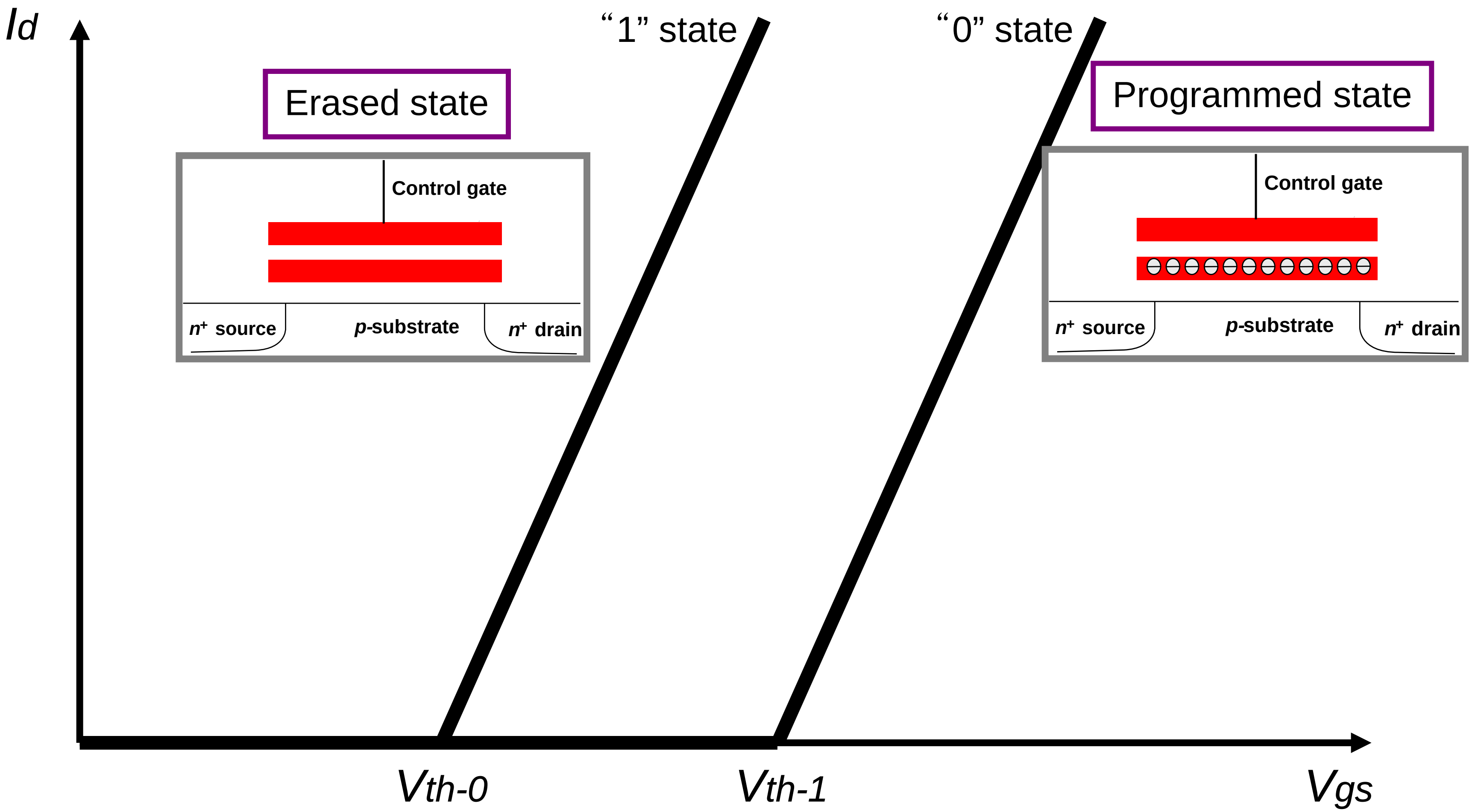


Schematic symbol

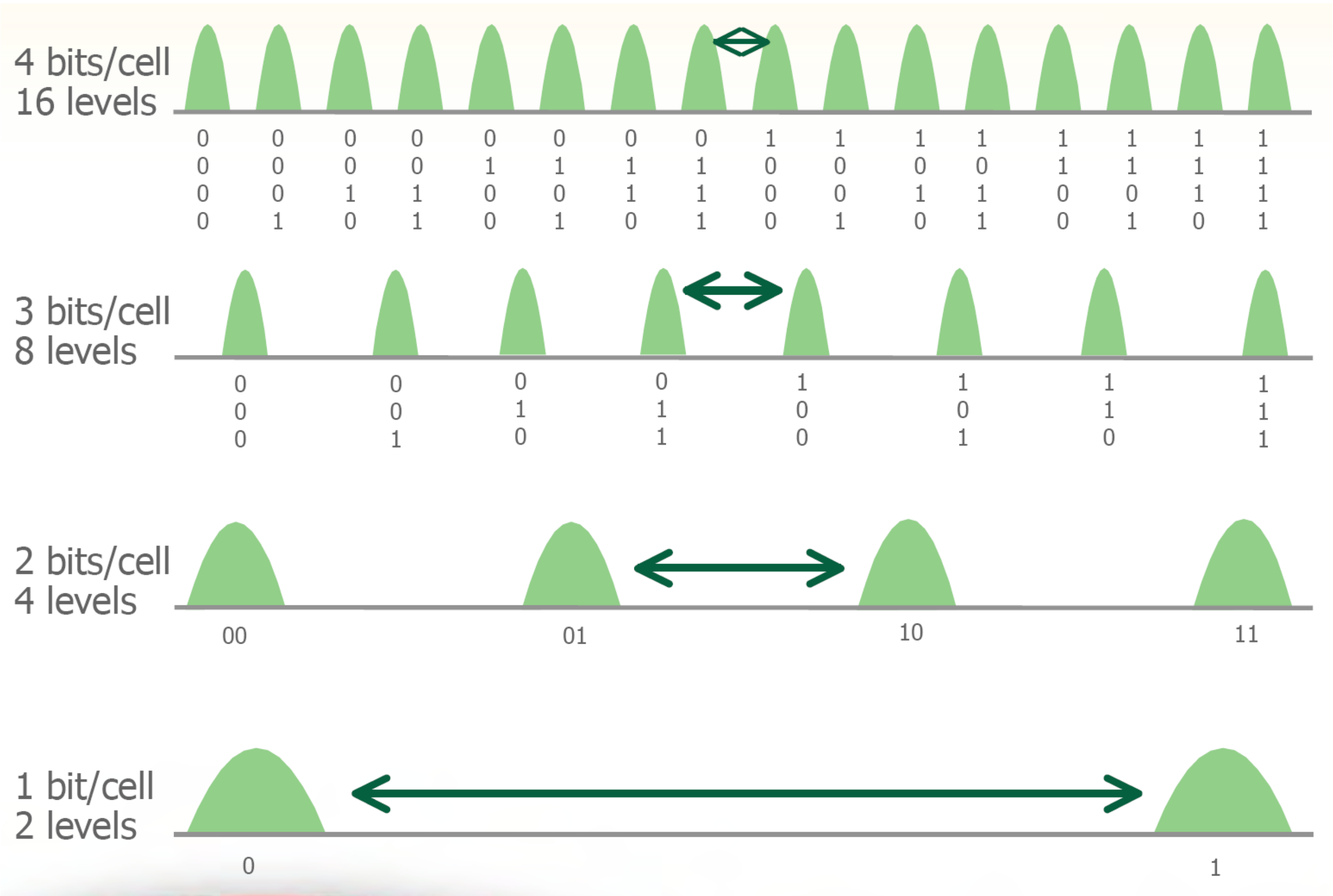
Flash Memory



Flash Memory:
A "Programmable-Threshold" Transistor

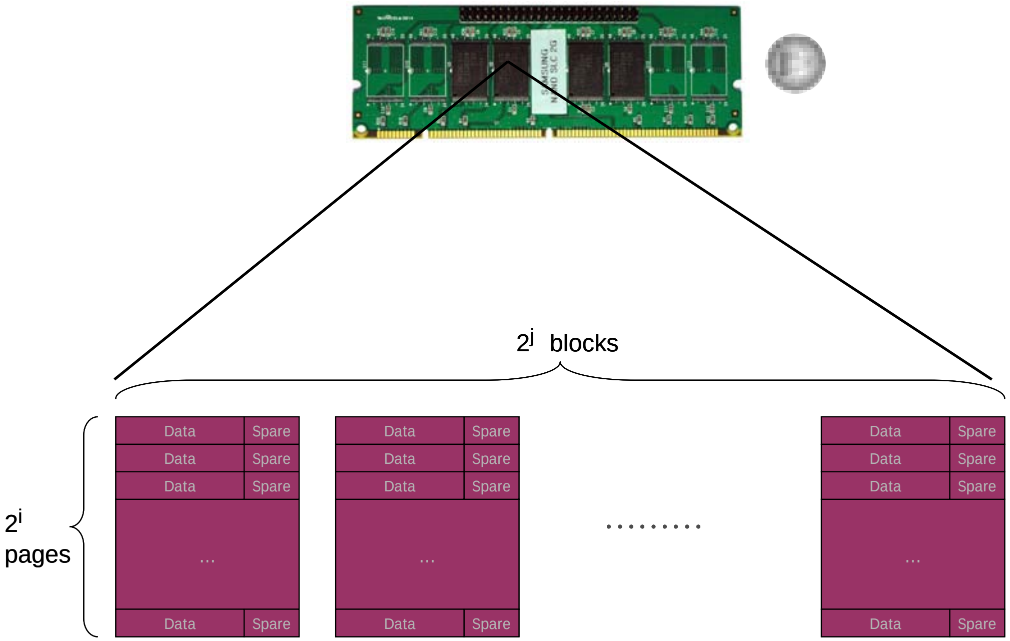


More Bits Per Transistor



Source: Eli Harari (SanDisk), “NAND at Center Stage,” Flash Memory Summit 2007.

NAND Flash Memory Interface

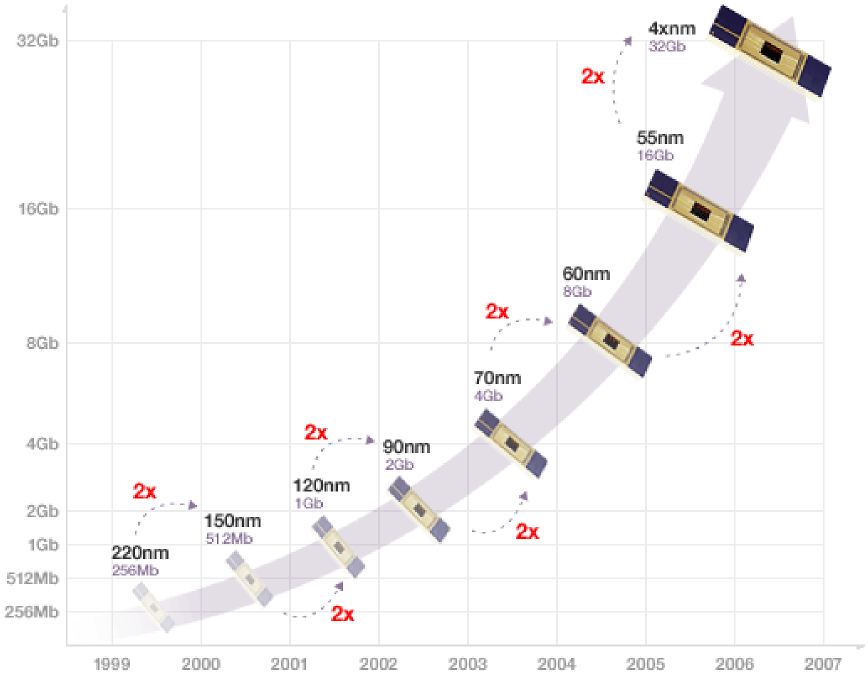


- Read physical page
  - (chip #, block #, page #)
  - ~ 20 us
- Write physical page
  - (chip #, block #, page #)
  - ~ 200 us
- Erase block
  - (chip#, block #)
  - ~ 2 ms

Why (NAND) Flash Memory?

- Advantages of Flash Memory over HDD
  - Low latency
  - Low power consumption
  - Tolerant to shock & vibration
  - Silent operation
  - Small size
  - Abundant parallelism
  - ...

- Single NAND Flash Memory Chip Density Trends



Source: Samsung Electronics

(More) NAND Flash Memory Trends

$ / MB	DRAM	NAND Flash
2000	$0.97	$1.35
2001	0.22	0.43
2002	0.22	0.25
2003	0.17	0.21
2004	0.17	0.10
2005	0.11	0.05
2006	0.096	0.021
2007	0.057	0.012
2008	~0.025	<0.005
CAGR	**-32.1% / yr**	**-50.0% / yr**

Source: Lane Mason (Denali Software), “NAND FlashPoint Platform”

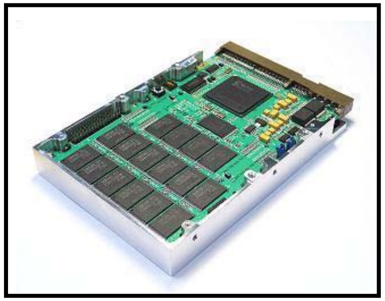
(More) NAND Flash Memory Trends

Millions GB	DRAM	NAND Flash
2000	30	1.1
2001	50	1.6
2002	71	4.6
2003	98	14.6
2004	158	68
2005	240	200
2006	340	600
2007	645	1600
2008	1000	4000
CAGR	**+60.0%/yr**	**+150%/yr**

Source: Lane Mason (Denali Software), "NAND FlashPoint Platform"

Solid State Disk

- Provides an interface identical to a hard disk, but uses flash memory as a storage medium



Identical
Interface

≡

Solid State Disk: Form Factor Agnostic

	Standard FF			Special FF	
	1.8’’	2.5’’	1.0’’	SLIM	So DIMM
Density	4~64GB	4~64GB	4~16GB	4~64GB	8~16GB
Dimension (H x W x T)	78.5x54x8.0	100.2x70x9.5	30x40x4.0	70.6x53.6x: 3.0: 16/32GB 2.5: 4~8GB	53.6x70.6x3.0
Connector	ZIF/IDE 50pin	IDE 44pin	ZIF 35pin	ZIF 40pin	200pin
Weight	44g	46g	TBD	20g	TBD
Market	Notebook	Sub-Note / Tablet	DVC/GPS/UMPC	UMPC	Custom

Source: Jim Elliot (Samsung Electronics), “SSD: The Next Killer App in NAND Flash,” Flash Memory Summit 2007.

Flash memory summary

The Good

- Low latency
- Low power consumption
- High Reliability
- Small form factor
- Massive parallelism

....

FROM *THE DARK NIGHT*

The Bad

- No in-place updating
- Limited endurance
- Bad blocks
- Write disturbance
- Read disturbance

The Ugly

- Retention errors
- Paired page problem

Outline

- HDD Basics and Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions

Storage Trends

Tape Is Dead

Disk Is Tape

- 1 TB disks are available
- 10+ TB disks are predicted in 5 years
- But: ~5..15 **hours to read (sequential)**

  ~15..150 **days to read (random)**
- Need to treat most of disk as Cold-storage archive

Source: Jim Gray (Microsoft), “Tape is Dead, Disk is Tape, Flash is Disk, RAM Locality is King”

Storage Trends

Disk Is Tape

Flash Is Disk

- 1995 16 Mb NAND flash chips

  2005 16 Gb NAND flash chips
- 2012 1 Tb NAND flash chips

  == 128 GB chip

  == 1 TB or 2 TB solid state disk for ~$400

  or 128 GB solid state disk for ~$40

  or 32 GB solid state disk for ~$5

Source: Jim Gray (Microsoft), “Tape is Dead, Disk is Tape, Flash is Disk, RAM Locality is King”

Disk is Tape / Flash is Disk

Poor Reliability ↓
Carnegie Mellon & Google study show up to 8.6% annual failure rate for HDD in controlled environment

Heat ↑
Rotating platters & moving heads need power → produces heat

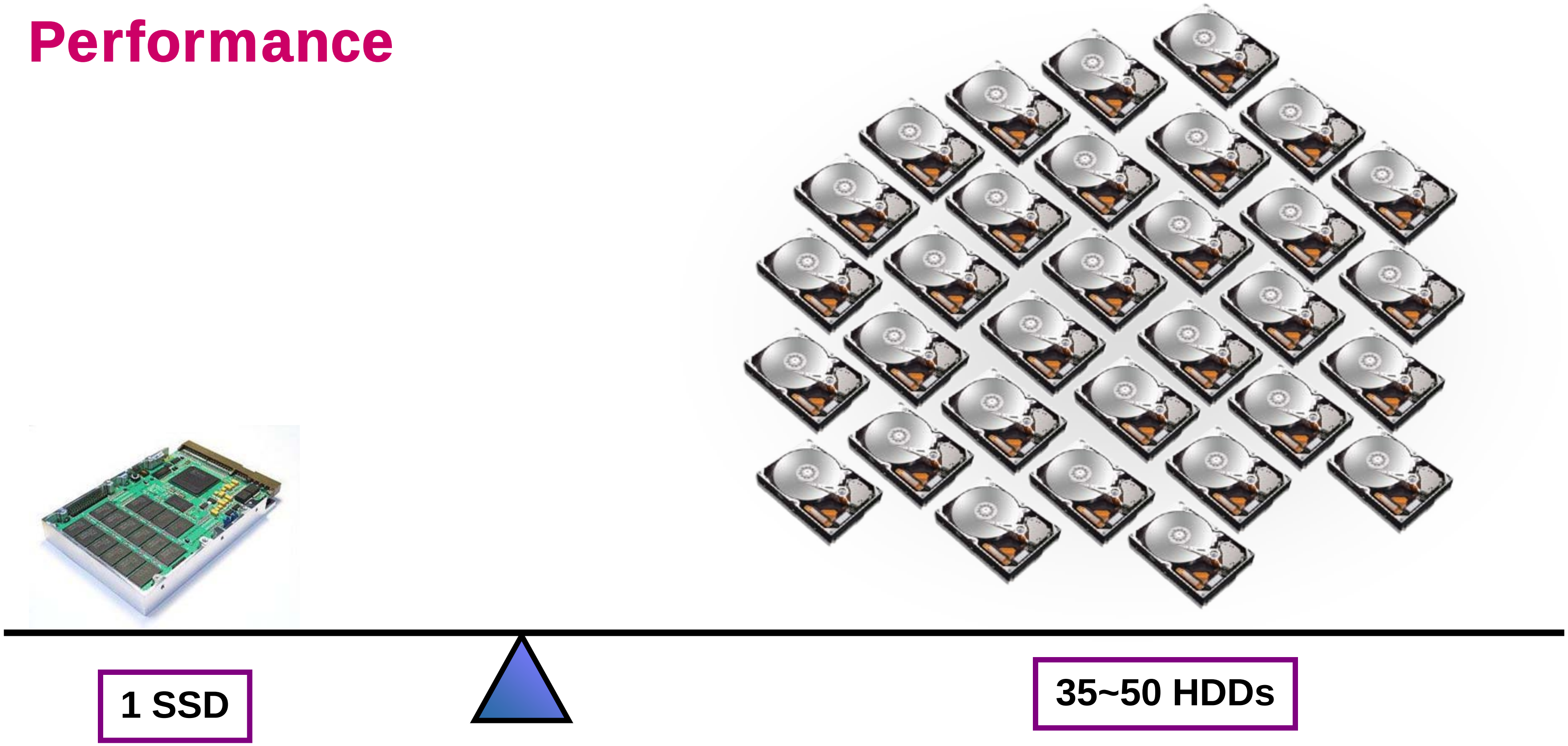
Low Performance
Low IOPS performance → High redundancy to compensate for low performance per drive ↓

High TCO ↑
Initial purchase cost low, but maintenance, space, cooling & replacement will increase TCO substantially

Source: Esther Spanjer (Adtron), "Enterprise SSD: The next killer app," Flash Memory Summit 2007.

Disk is Tape / Flash is Disk

- **Performance**



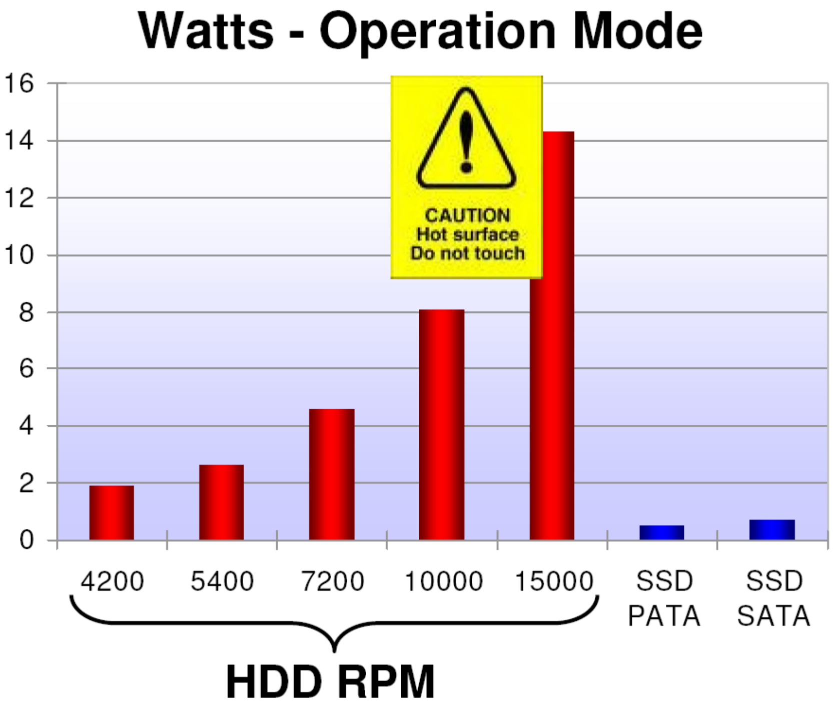
1 SSD

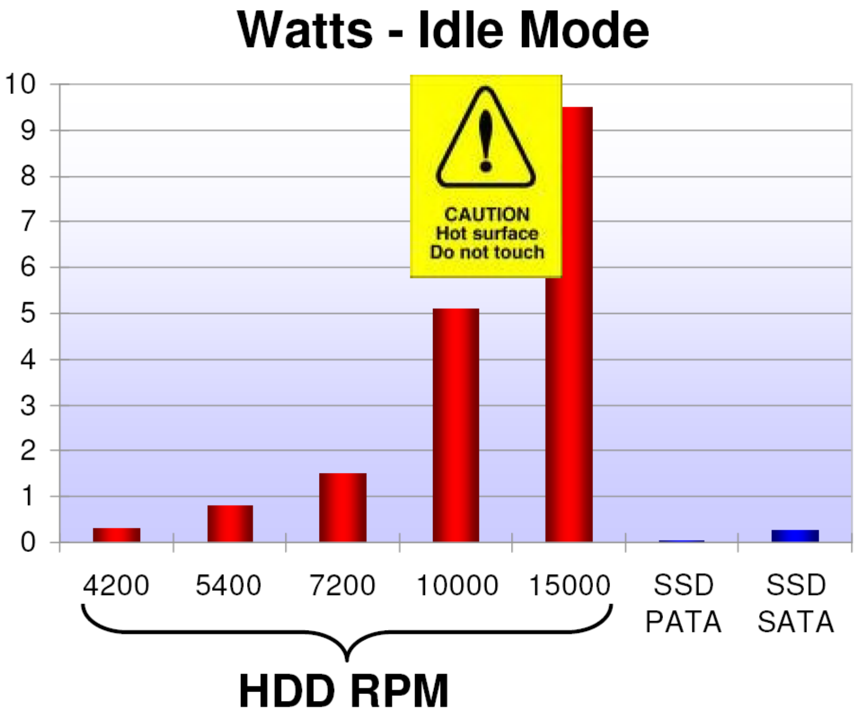
35~50 HDDs

Source: Jim Gray (Microsoft), “Tape is Dead, Disk is Tape, Flash is Disk, RAM Locality is King”

Disk is Tape / Flash is Disk

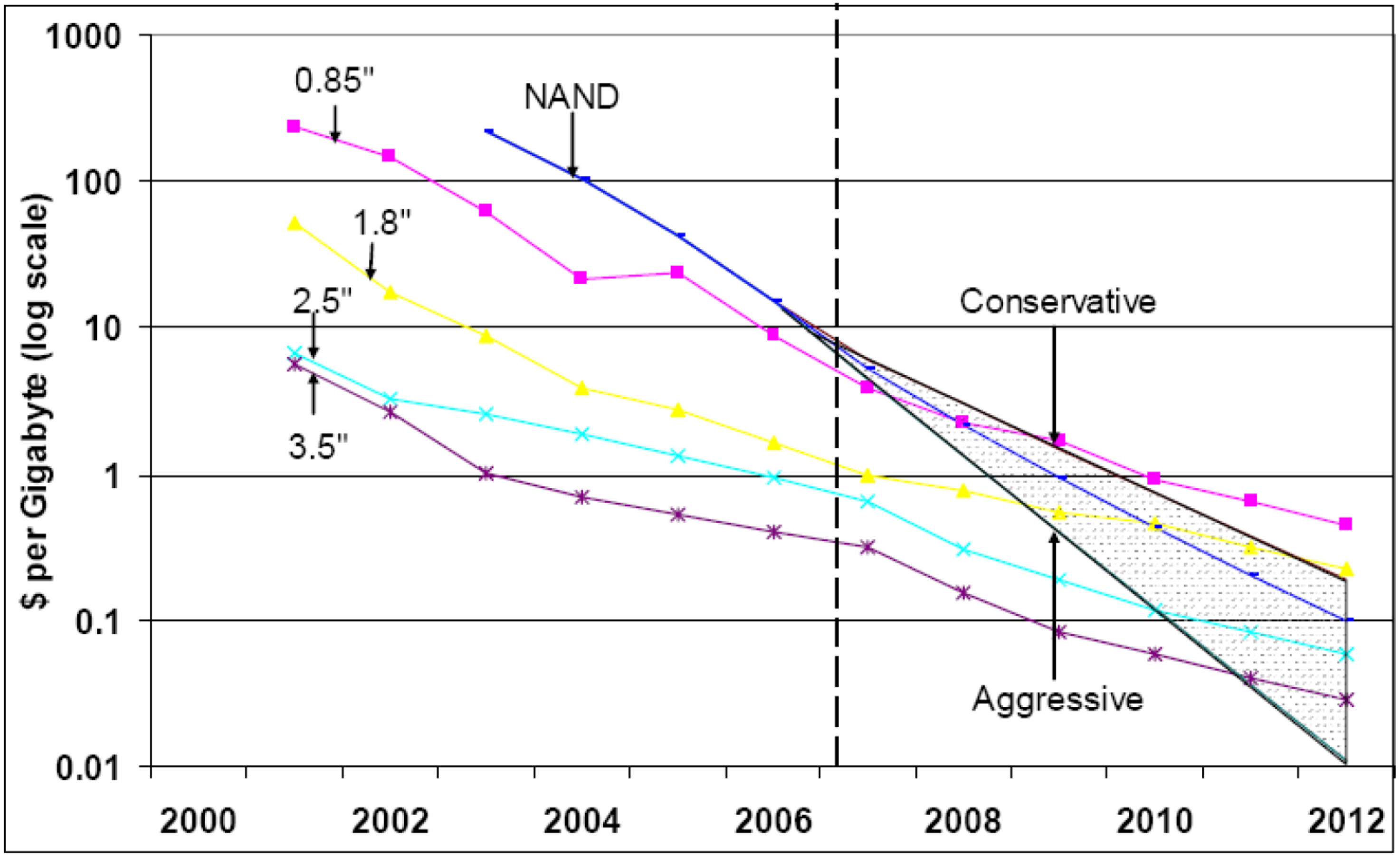
- **Power Consumption**





Source: Jim Elliot (Samsung Electronics), "SSD: The Next Killer App in NAND Flash," Flash Memory Summit 2007.

Future Outlook



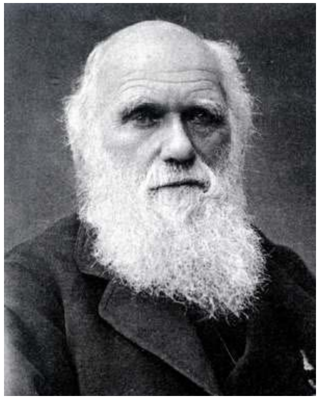
Source: Scott Deutsch (SanDisk), “Bringing Solid State Drives to Mainstream Notebooks,” Flash Memory Summit 2007.

Outline

- HDD Basics and Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions

Conclusions

- In the animal world
  - Survival of the fittest



- In the memory world
  - Survival of the fastest or cheapest

	Volatile	Non-volatile
Fastest	SRAM	FRAM?
Cheapest	DRAM	NAND Flash HDD

Conclusions

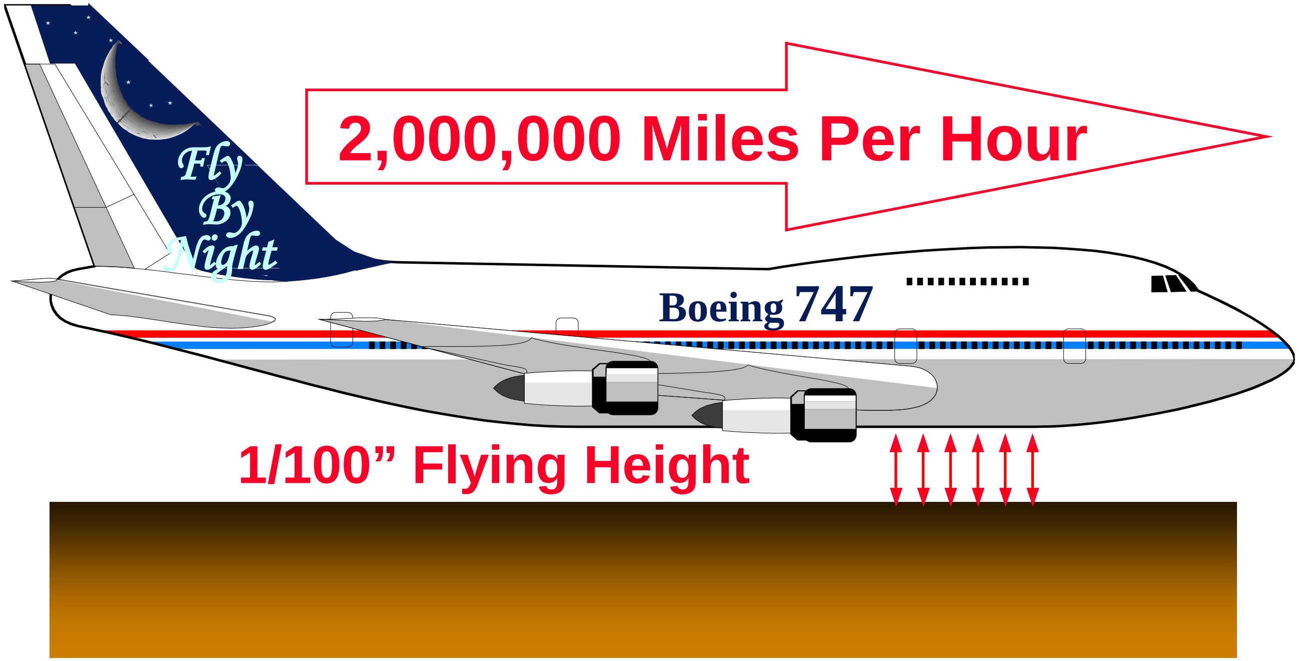
- From the history

	IBM 360/85	IBM 360/91
Clock Rate	80 ns	60 ns
Memory Speed	1040 ns	750 ns
Memory Interleaving	4 way	8 way
Additional Features	Cache Memory	Register Renaming, Out-of-order Execution, *etc*

But, IBM 360/85 faster on 8 of 11 programs!

Source: David Patterson, *et al.*, “A Case for Intelligent DRAM: IRAM”, Hot Chips VIII, August, 1996

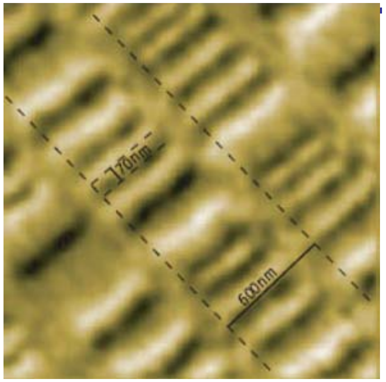
The Ultimate Limit – HDD

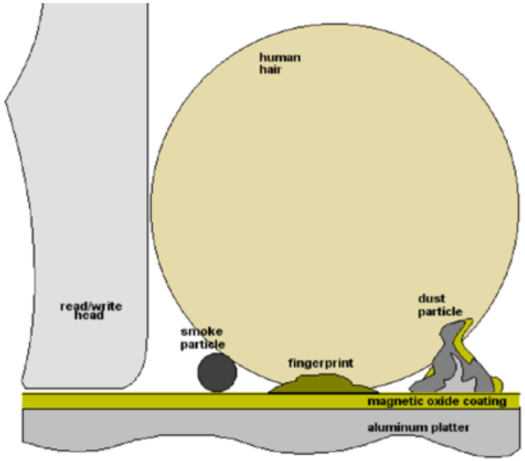


2,000,000 Miles Per Hour

1/100” Flying Height

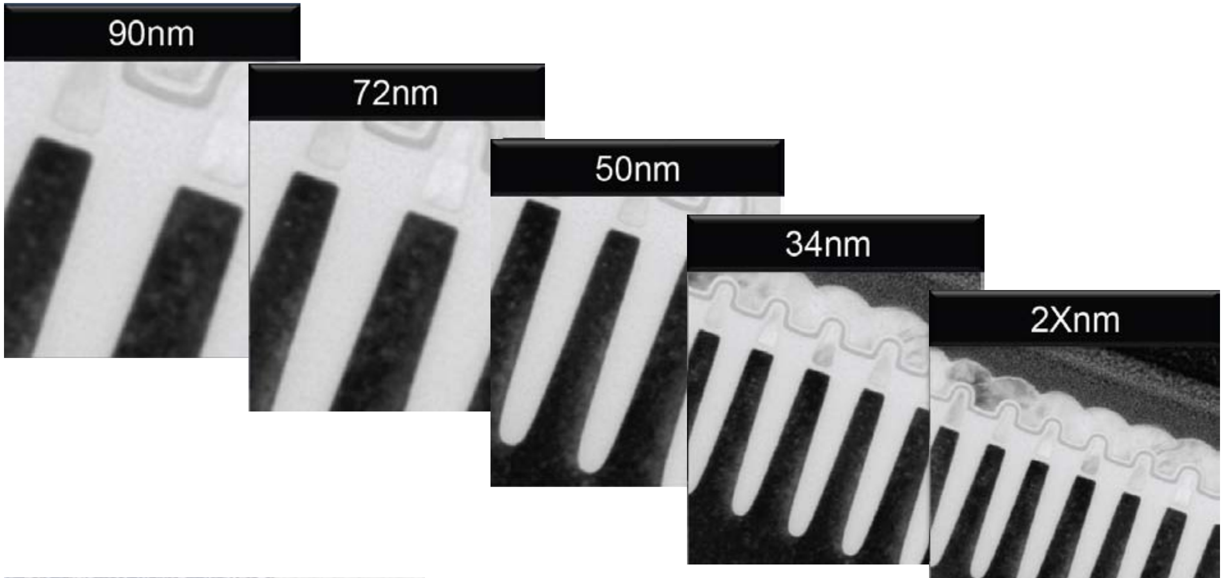
Source: Richard Lary, The New Storage Landscape: Forces shaping the storage economy, 2003.

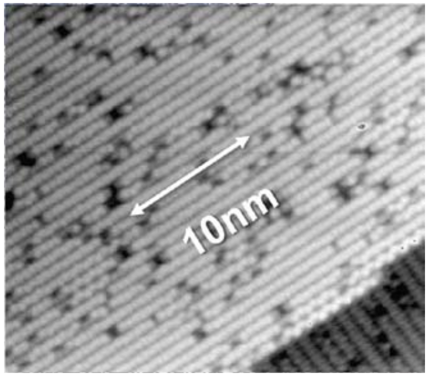




Source: B. Parhami, Dependable Computing: A Multilevel Approach

The Ultimate Limit – Flash Memory





Scanning tunneling microscope image of a silicon surface showing 10 nm is ~20 atoms across

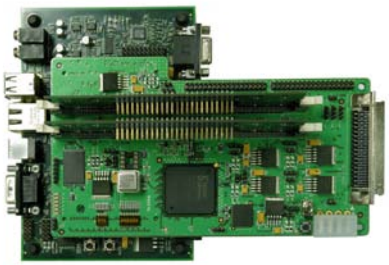
Source: B. Shirley, "The Many Flavors of NAND … and More to Come," Flash Memory Summit 2009

Outline

- HDD Basics and Demo
- Flash Memory Basics and Demo
- Storage Trends
- Conclusions
- (More Demos)

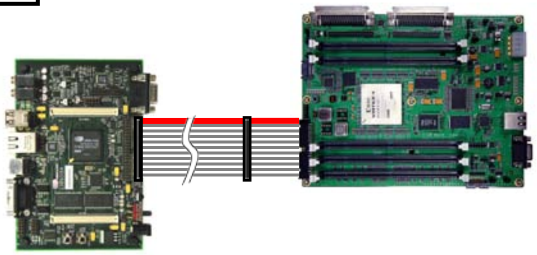
Flash Memory Software Development Platforms

Embedded Platform



Embedded Flash Memory 소프트웨어 솔루션 개발용

SSD Platform



Solid State Disk 소프트웨어 개발용

Flash / NV-RAM Modules

Samsung SLC NAND

RAMTRON FRAM (serial)

Samsung MLC NAND

RAMTRON FRAM (parallel)

Samsung OneNAND

FREESCALE MRAM (parallel)

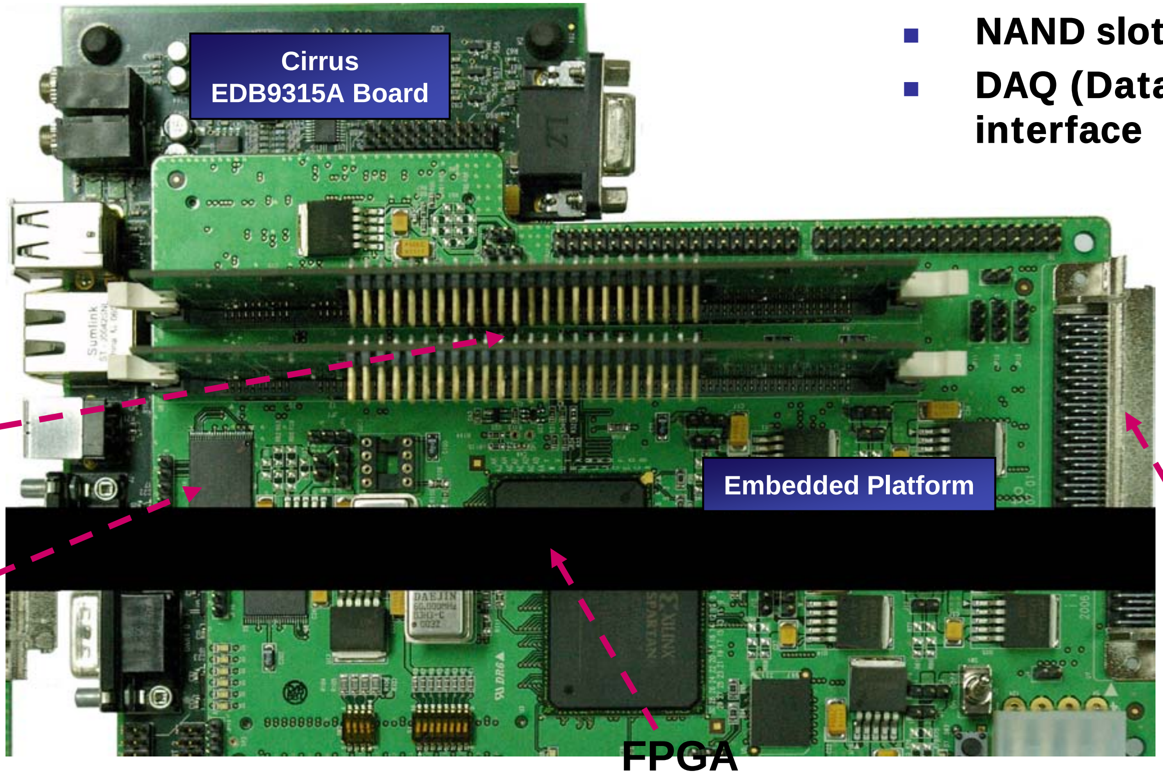
Hynix MLC NAND

Samsung Phase-change RAM

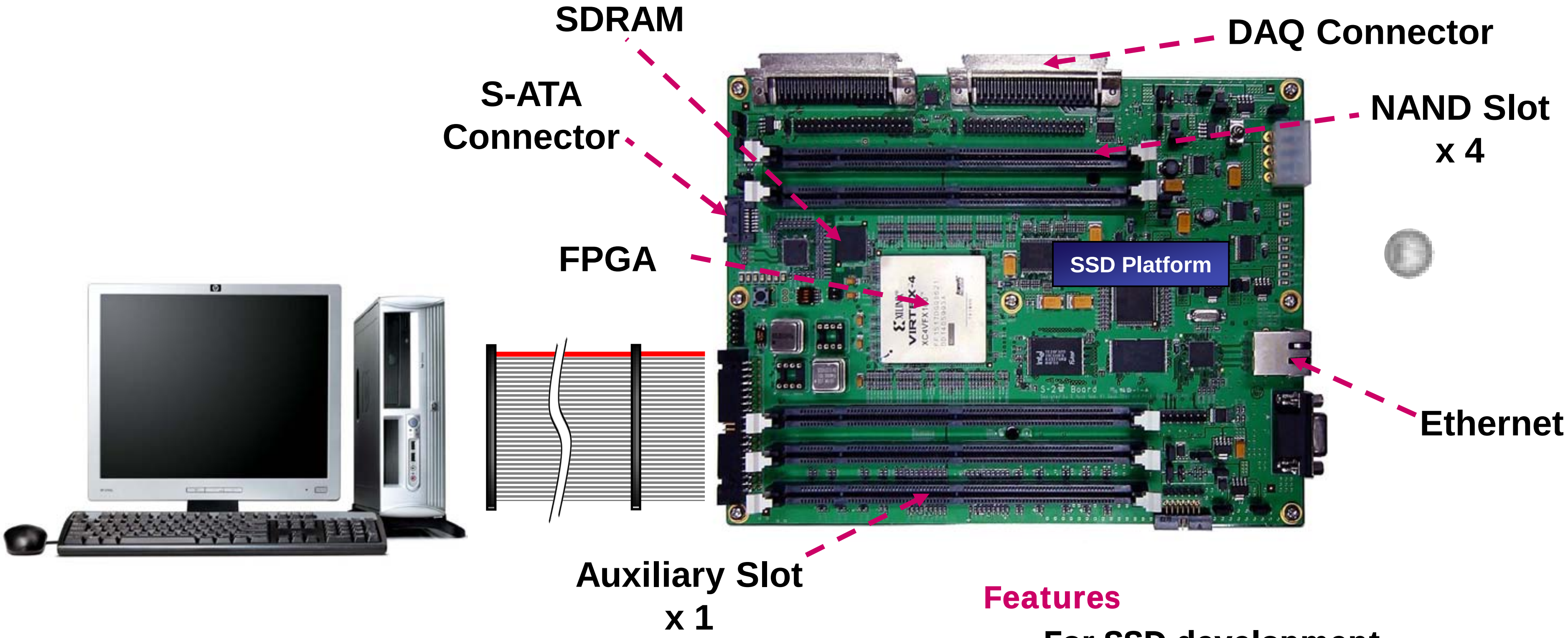
Embedded Platform

Features

- For embedded Flash memory software development
- FPGA-based
- NAND slot x 2
- DAQ (Data Acquisition) interface



SSD (Solid State Disk) Platform



SDRAM

DAQ Connector

S-ATA Connector

NAND Slot x 4

FPGA

SSD Platform

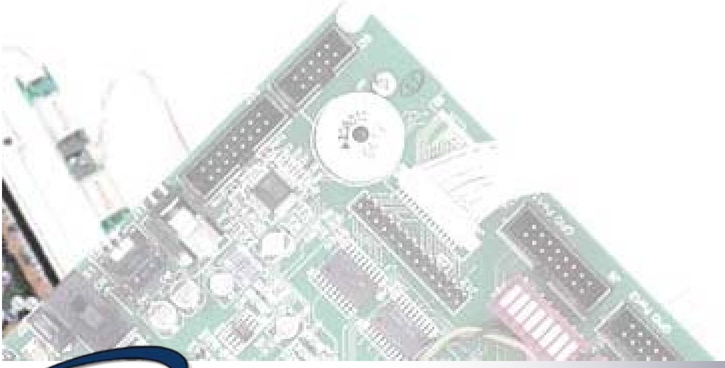
Ethernet

Auxiliary Slot x 1

Features

- **For SSD development**
- **FPGA-based**
- **SSD interface (P-ATA, S-ATA)**
- **NAND slot x 4**

I/O Management



I/O Commands

- I/O devices are managed by I/O controller hardware
  - Transfers data to/from device
  - Synchronizes operations with software
- Command registers
  - Cause device to do something
- Status registers
  - Indicate what the device is doing and occurrence of errors
- Data registers
  - Write: transfer data to a device
  - Read: transfer data from a device

I/O Register Mapping

- Memory mapped I/O
  - Registers are addressed in same space as memory
  - Address decoder distinguishes between them
  - OS uses address translation mechanism to make them only accessible to kernel
- I/O instructions
  - Separate instructions to access I/O registers
  - Can only be executed in kernel mode
  - Example: x86

Polling

- Periodically check I/O status register
  - If device ready, do operation
  - If error, take action
- Common in small or low-performance real-time embedded systems
  - Predictable timing
  - Low hardware cost
- In other systems, wastes CPU time

Interrupts

- When a device is ready or error occurs
  - Controller interrupts CPU
- Interrupt is like an exception
  - But not synchronized to instruction execution
  - Can invoke handler between instructions
  - Cause information often identifies the interrupting device
- Priority interrupts
  - Devices needing more urgent attention get higher priority
  - Can interrupt handler for a lower priority interrupt

I/O Data Transfer

- Polling and interrupt-driven I/O
  - CPU transfers data between memory and I/O data registers
  - Time consuming for high-speed devices
- Direct memory access (DMA)
  - OS provides starting address in memory
  - I/O controller transfers to/from memory autonomously
  - Controller interrupts on completion or error

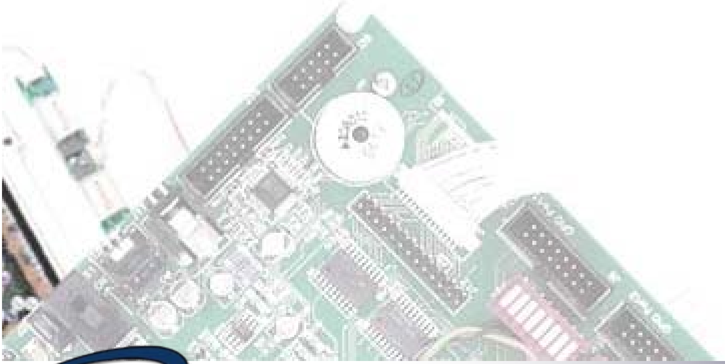
DMA/Cache Interaction

- If DMA writes to a memory block that is cached
  - Cached copy becomes stale
- If write-back cache has dirty block, and DMA reads memory block
  - Reads stale data
- Need to ensure cache coherence
  - Flush blocks from cache if they will be used for DMA
  - Or use non-cacheable memory locations for I/O

DMA/VM Interaction

- OS uses virtual addresses for memory
  - DMA blocks may not be contiguous in physical memory
- Should DMA use virtual addresses?
  - Would require controller to do translation
- If DMA uses physical addresses
  - May need to break transfers into page-sized chunks
  - Or chain multiple transfers
  - Or allocate contiguous physical pages for DMA

RAID (Redundant Array of Inexpensive Disks)

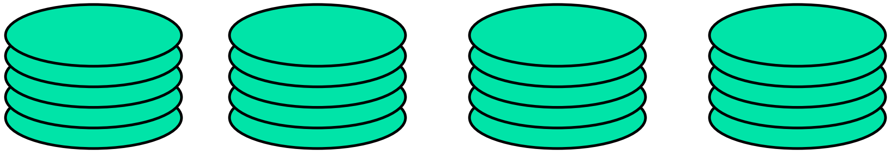


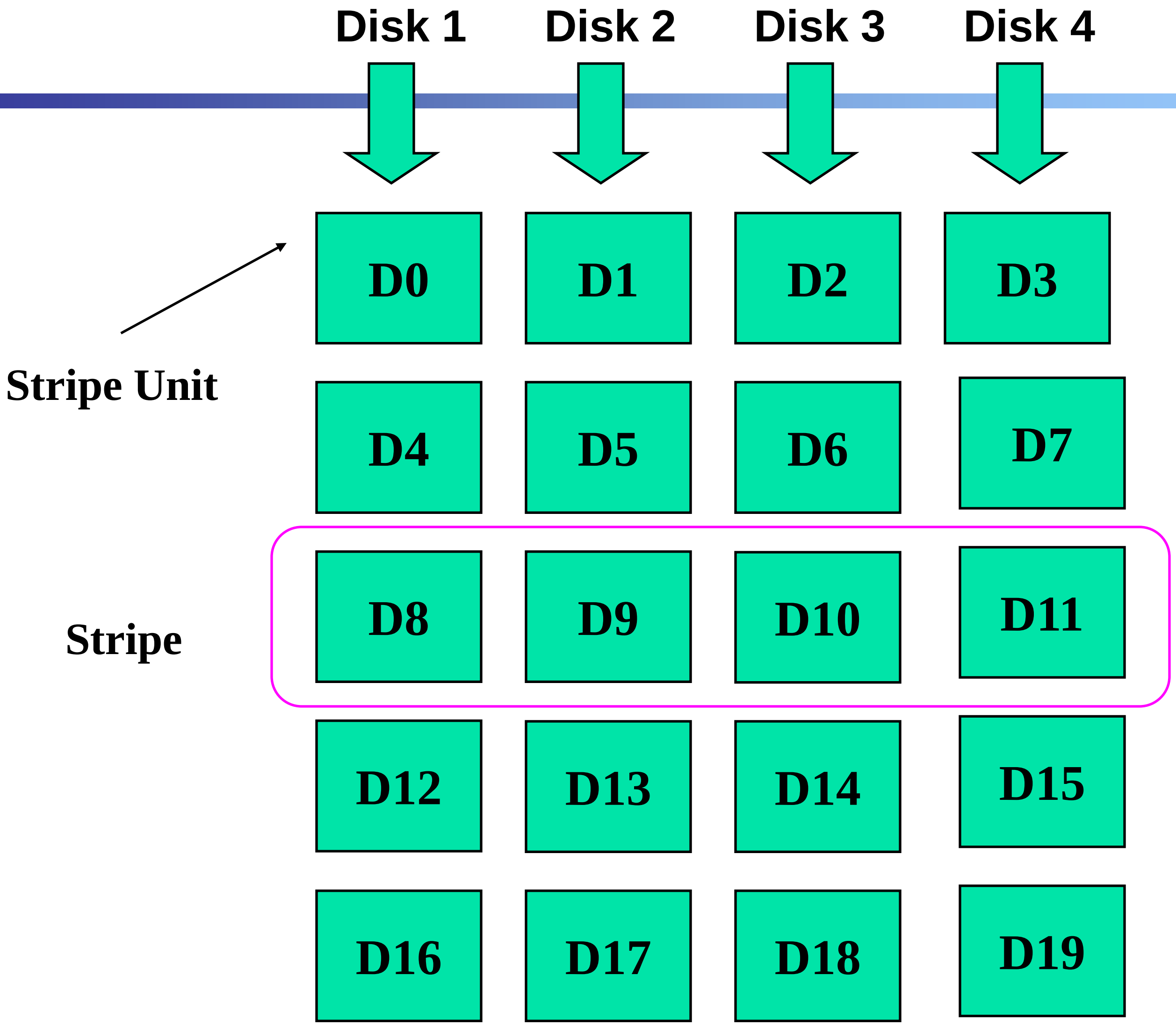
RAID (Redundant Array of Inexpensive Disks)

- Data is Striped for improved performance
  - Distributes data over multiple disks to make them appear as a single fast large disk
  - Allows multiple I/Os to be serviced in parallel
    - Multiple independent requests serviced in parallel
    - A block request may be serviced in parallel by multiple disks
- Data is Redundant for improved reliability
  - Large number of disks in an array lowers the reliability of the array
    - Reliability of N disks = Reliability of 1 disk /N
    - Example:
      - 50,000 hours / 70 disks = 700 hours
      - Disk System MTTF drops from 6 years to 1 month
  - Arrays without redundancy are too unreliable to be useful

From lecture slides by Professor Mazin Yousif

- RAID 0 (Non-redundant)
  - Stripes Data; but does not employ redundancy
  - Lowest cost of any RAID
  - Best Write performance - no redundant information
  - Any single disk failure is catastrophic
  - Used in environments where performance is more important than reliability.

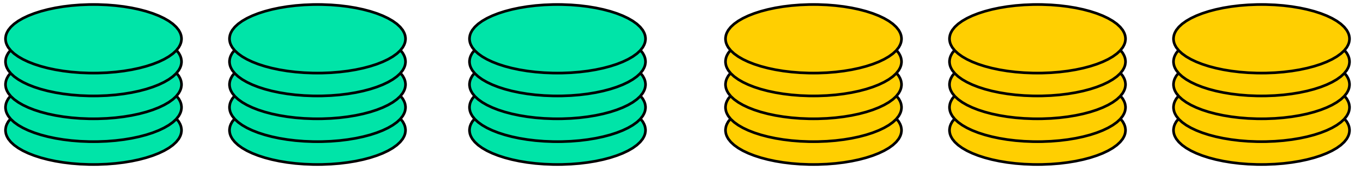




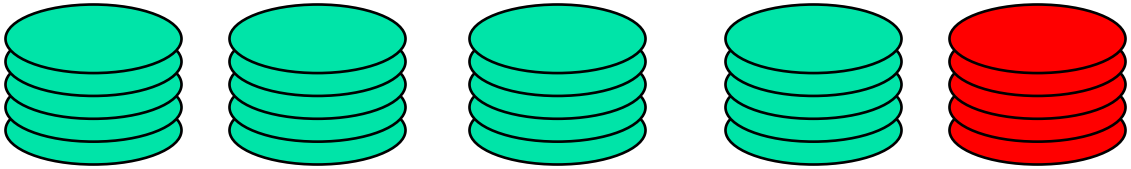
Disk 1	Disk 2	Disk 3	Disk 4
D0	D1	D2	D3
D4	D5	D6	D7
D8	D9	D10	D11
D12	D13	D14	D15
D16	D17	D18	D19

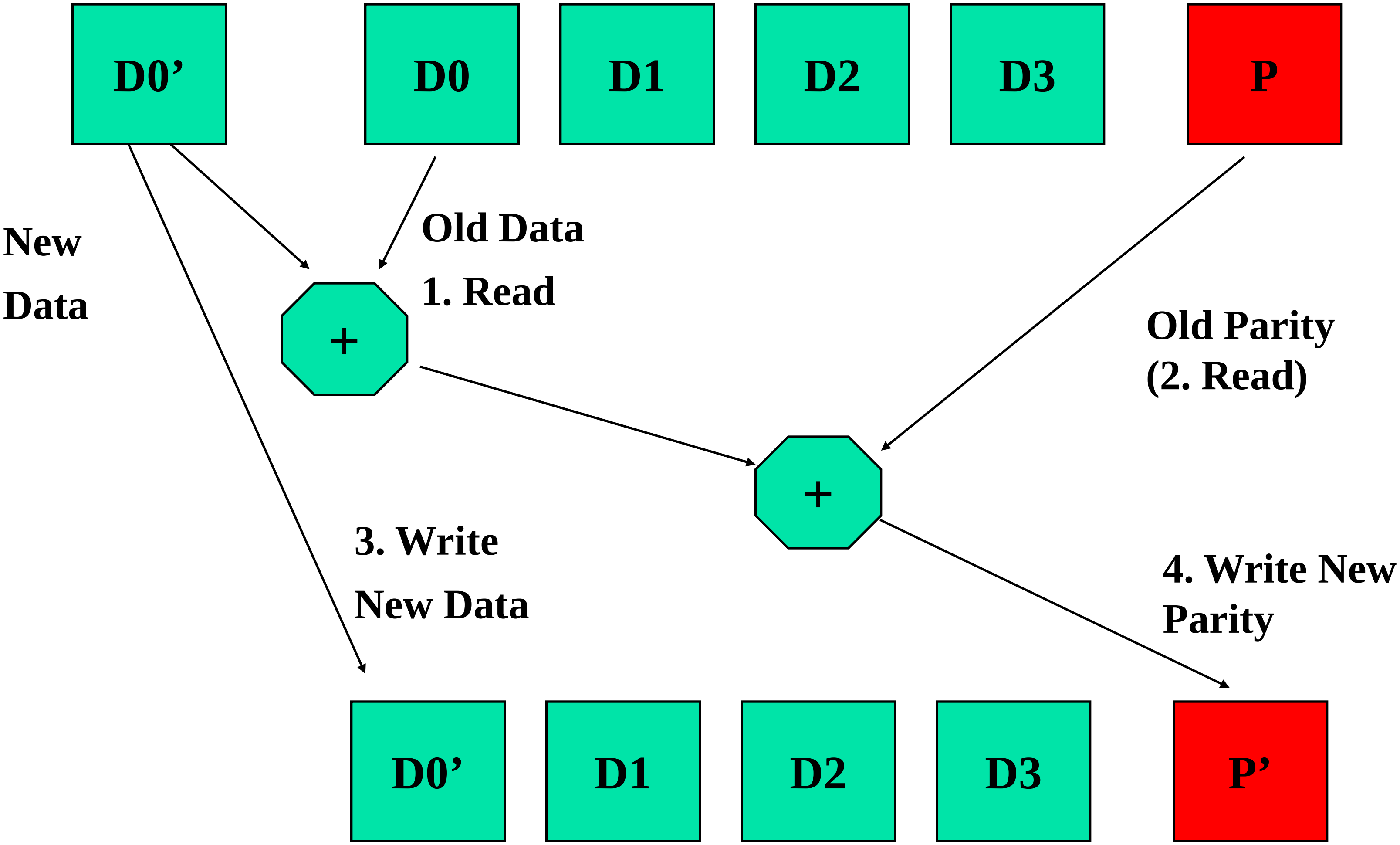
RAID 1 (Mirrored)

- Uses twice as many disks as non-redundant arrays - 100% Capacity Overhead - Two copies of data are maintained
- Data is simultaneously written to both arrays
- Data is read from the array with shorter queuing, seek and rotation delays - Best Read Performance.
- When a disk fails, mirrored copy is still available
- Used in environments where availability and performance (I/O rate) are more important than storage efficiency.

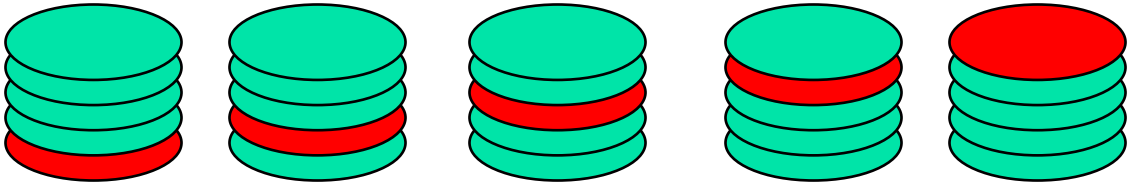


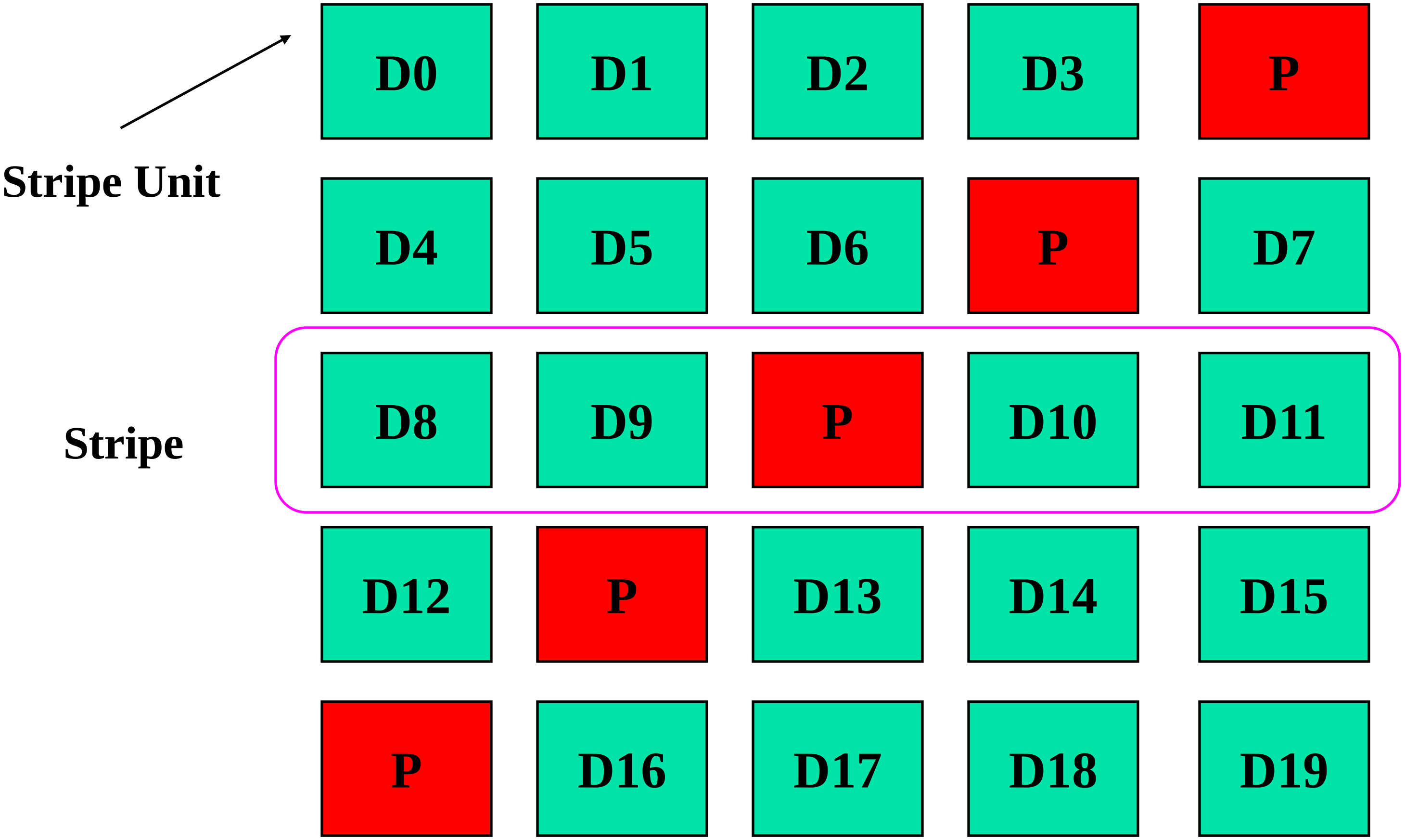
- RAID 4 (Block Interleaved Parity)
  - Similar to bit-interleaved parity disk array; except data is block- interleaved (Striping Units)
  - Write requests update the data block; and the parity block.
  - Generating parity requires 4 I/O accesses (Read/Modify/Write)
  - All writes access the parity disk – parallel service of write requests is not possible




D0’
D0
D1
D2
D3
P
New
Data
Old Data
1. Read
+
Old Parity
(2. Read)
+
3. Write
New Data
4. Write New
Parity
D0’
D1
D2
D3
P’

- RAID 5 (Block-Interleaved Distributed Parity)
  - Eliminates the parity disk bottleneck in RAID 4 - Distributes parity among all the disks
  - Parallel service of write requests is now possible as long as they access disjoint disk



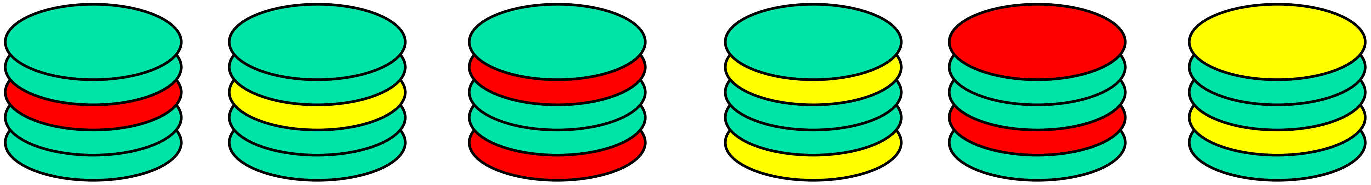


Stripe Unit

Stripe

D0	D1	D2	D3	P
D4	D5	D6	P	D7
D8	D9	P	D10	D11
D12	P	D13	D14	D15
P	D16	D17	D18	D19

- RAID 6 (P + Q Redundancy)
  - Uses Reed-Solomon codes to protect against up to 2 disk failures
  - Two sets of parity P & Q
  - Generating parity requires 6 I/O accesses (Read/Modify/Write) - update both P & Q
  - Used in environments that require stringent reliability



RAID Summary

- RAID can improve performance and availability
  - High availability requires hot swapping
- Assumes independent disk failures
  - Too bad if the building burns down!

Concluding Remarks

- I/O performance measures
  - Throughput, response time
  - Dependability and cost also important
- Buses used to connect CPU, memory, I/O controllers
  - Polling, interrupts, DMA
- RAID
  - Improves performance and dependability