01076114
องค์ประกอบและสถาปัตยกรรมคอมพิวเตอร์
Computer Organization and Architecture

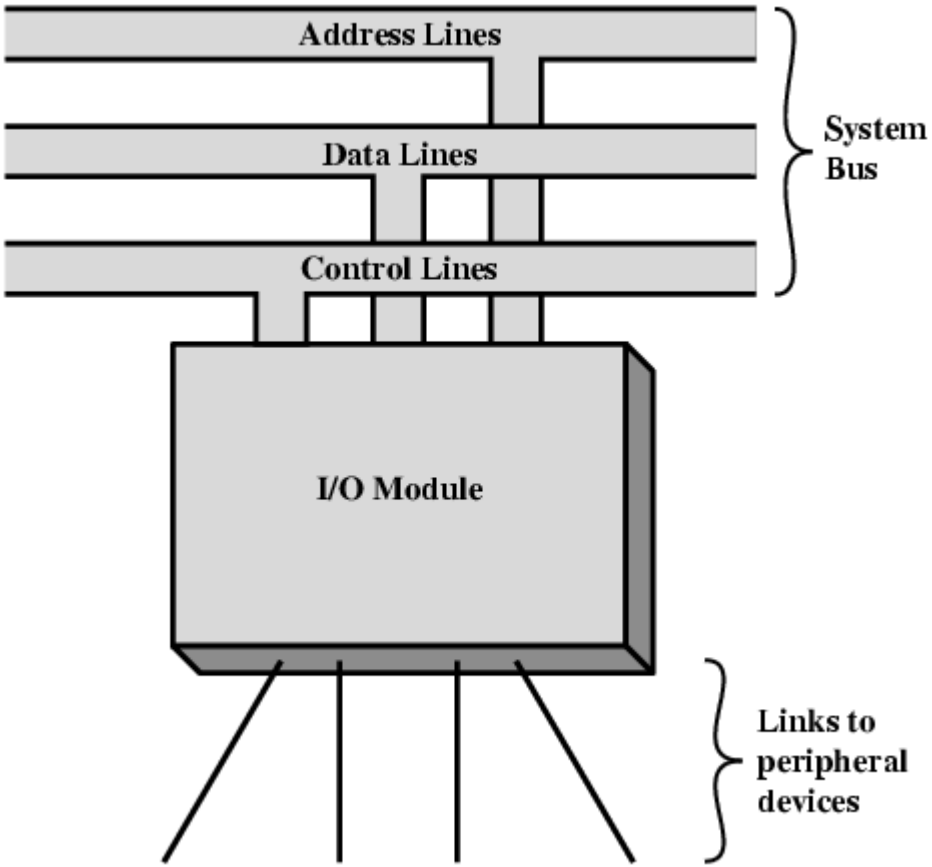
Storage and I/O

Storage and I/O

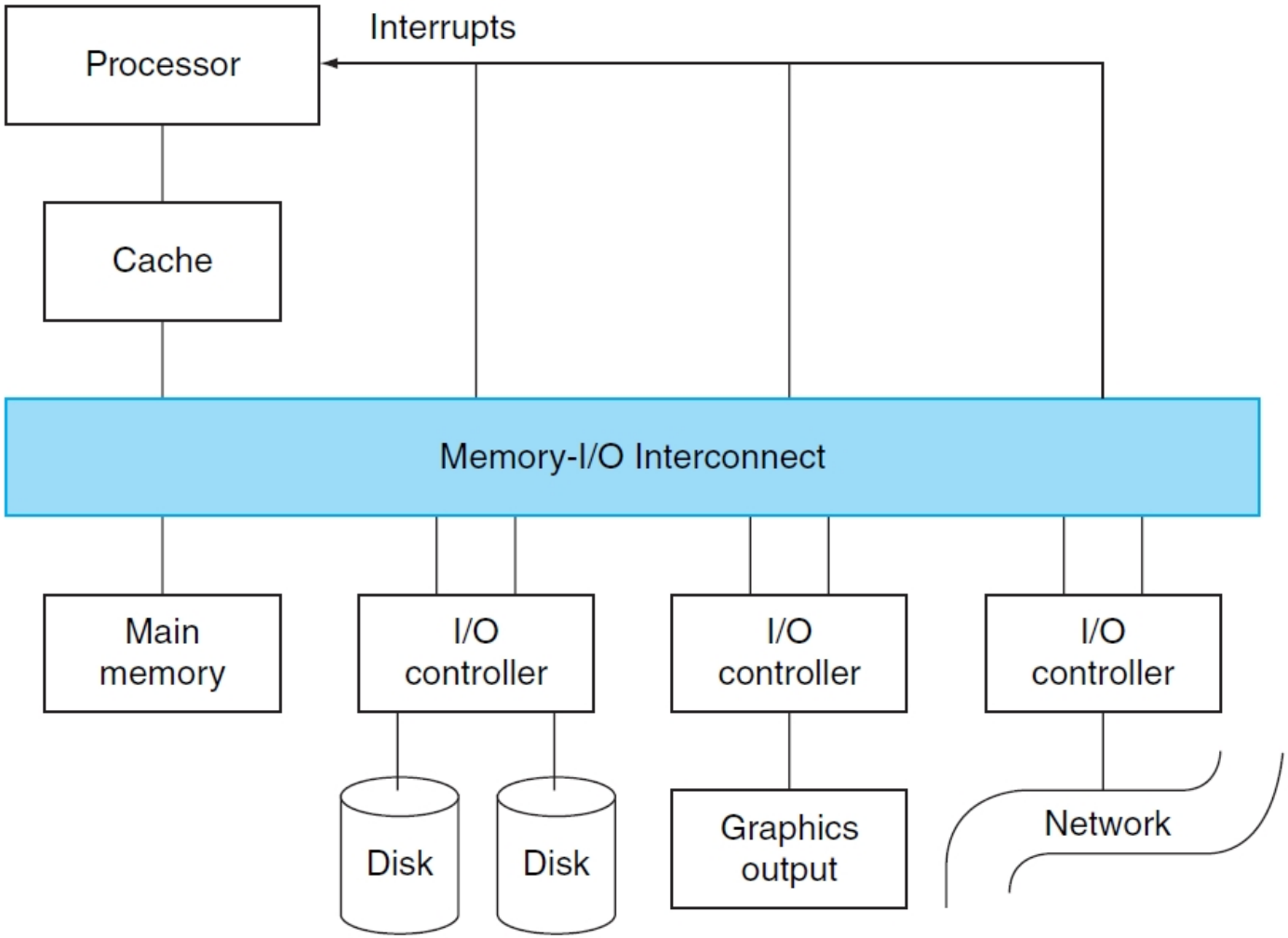
- Storage and I/O เป็นส่วนประกอบของระบบคอมพิวเตอร์ที่สำคัญ
  - ต้องมีความมั่นคงสูง แม้เจอความเสียหาย เช่น ไฟดับ ก็ต้องไม่ทำให้ข้อมูลหาย
  - ต้องมีกลไกในการ recover จากความเสียหายต่างๆ
  - ต้องรองรับอุปกรณ์ I/O หลากหลาย ให้สามารถทำงานร่วมกันได้
  - อุปกรณ์ I/O หนึ่งสามารถทำงานร่วมกับระบบคอมพิวเตอร์ที่หลากหลายได้
  - ต้องรองรับอุปกรณ์ที่มีความเร็วที่แตกต่างกันมาก เช่น คีย์บอร์ด, Mouse ที่ต้องการส่งข้อมูลในหลักร้อยไบต์ต่อวินาที ไปจนถึงฮาร์ดดิสก์ที่ต้องการส่งข้อมูลในระดับหลายร้อย MB ต่อวินาที

Generic Model of I/O Module

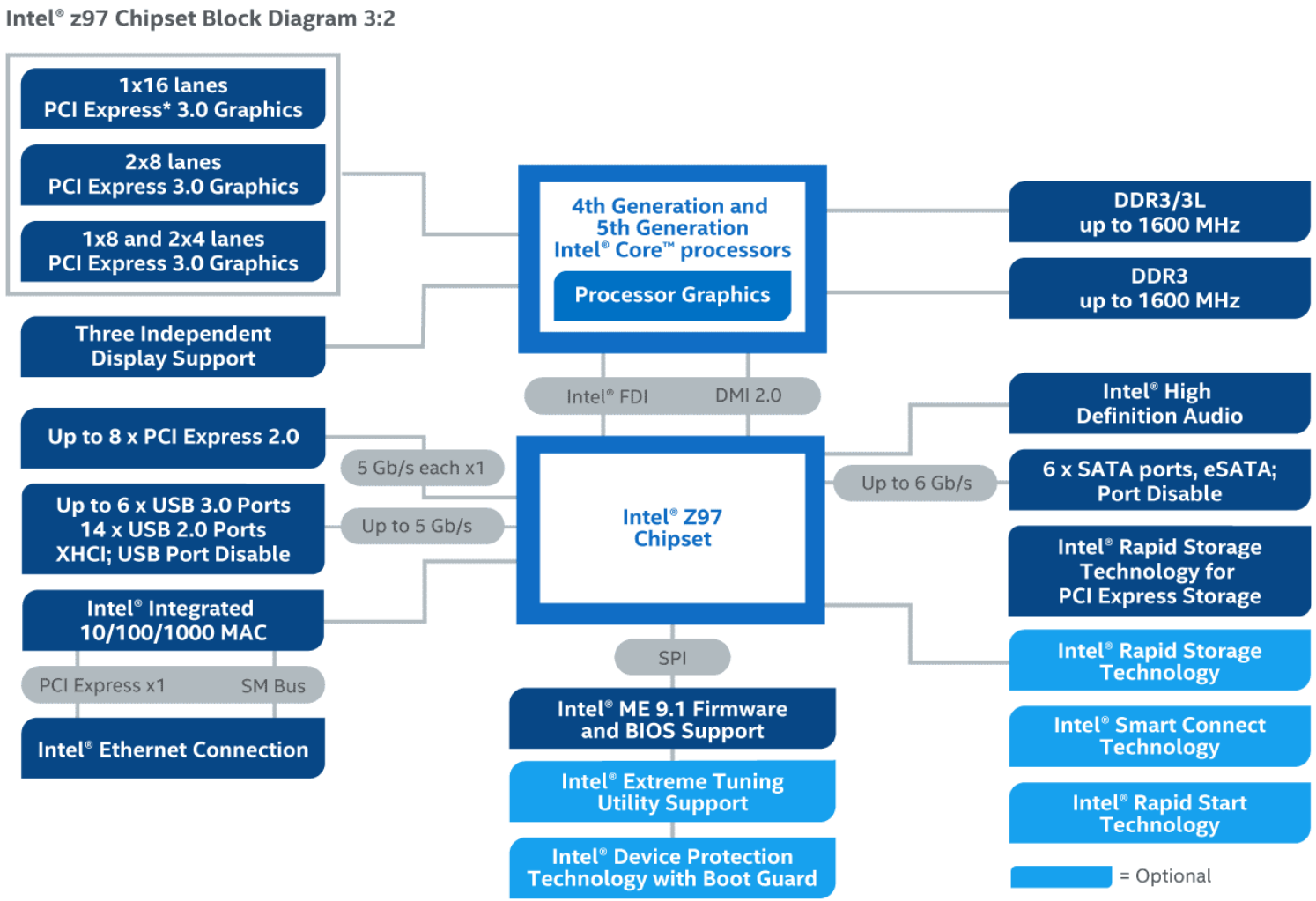
- ในการติดต่อระหว่าง Processor กับ I/O จะกระทำผ่าน Address Bus, Data Bus และ Control
- โดยมี I/O Module ประเภทต่างๆ
  - Hard drive
  - Network
  - Display
  - Low bandwidth



Collection of I/O devices



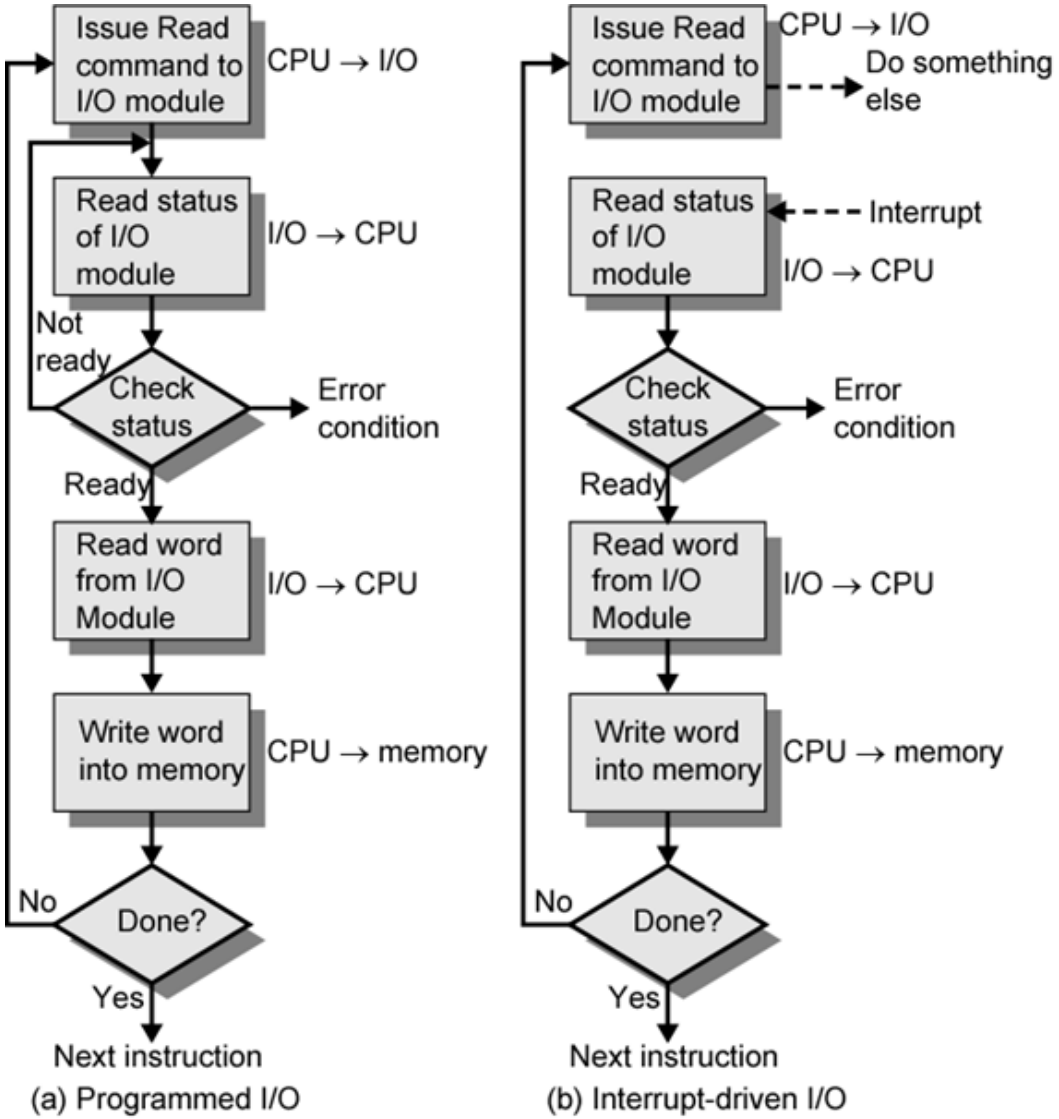
Intel z97 chipset



Diversity of 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	800.0000–8000.0000
Cable modem	Input or output	Machine	0.1280–6.0000
Network/LAN	Input or output	Machine	100.0000–10000.0000
Network/wireless LAN	Input or output	Machine	11.0000–54.0000
Optical disk	Storage	Machine	80.0000–220.0000
Magnetic tape	Storage	Machine	5.0000–120.0000
Flash memory	Storage	Machine	32.0000–200.0000
Magnetic disk	Storage	Machine	800.0000–3000.0000

Programmed I/O vs Interrupt-driven I/O



Programmed I/O

- CPU จะควบคุม I/O โดยตรง
  - ตรวจสอบสถานะการทำงาน
  - ส่งคำสั่ง Read/Write
  - รับส่งข้อมูล
- หลังจากส่งคำสั่งแล้ว CPU จะรอจนกว่าการทำงานของ I/O จะเสร็จสิ้น
  - เช่น การรับคีย์บอร์ด ก็จะรอรับจนกว่าผู้ใช้จะกดปุ่มคีย์บอร์ด
- การทำงานแบบนี้ จะสิ้นเปลืองเวลาในการรออย่างมาก
- บางครั้งเรียกการทำงานแบบนี้ว่า Polling
- ปัจจุบันยังมีการใช้งานกับ Embedded System ขนาดเล็กเท่านั้น

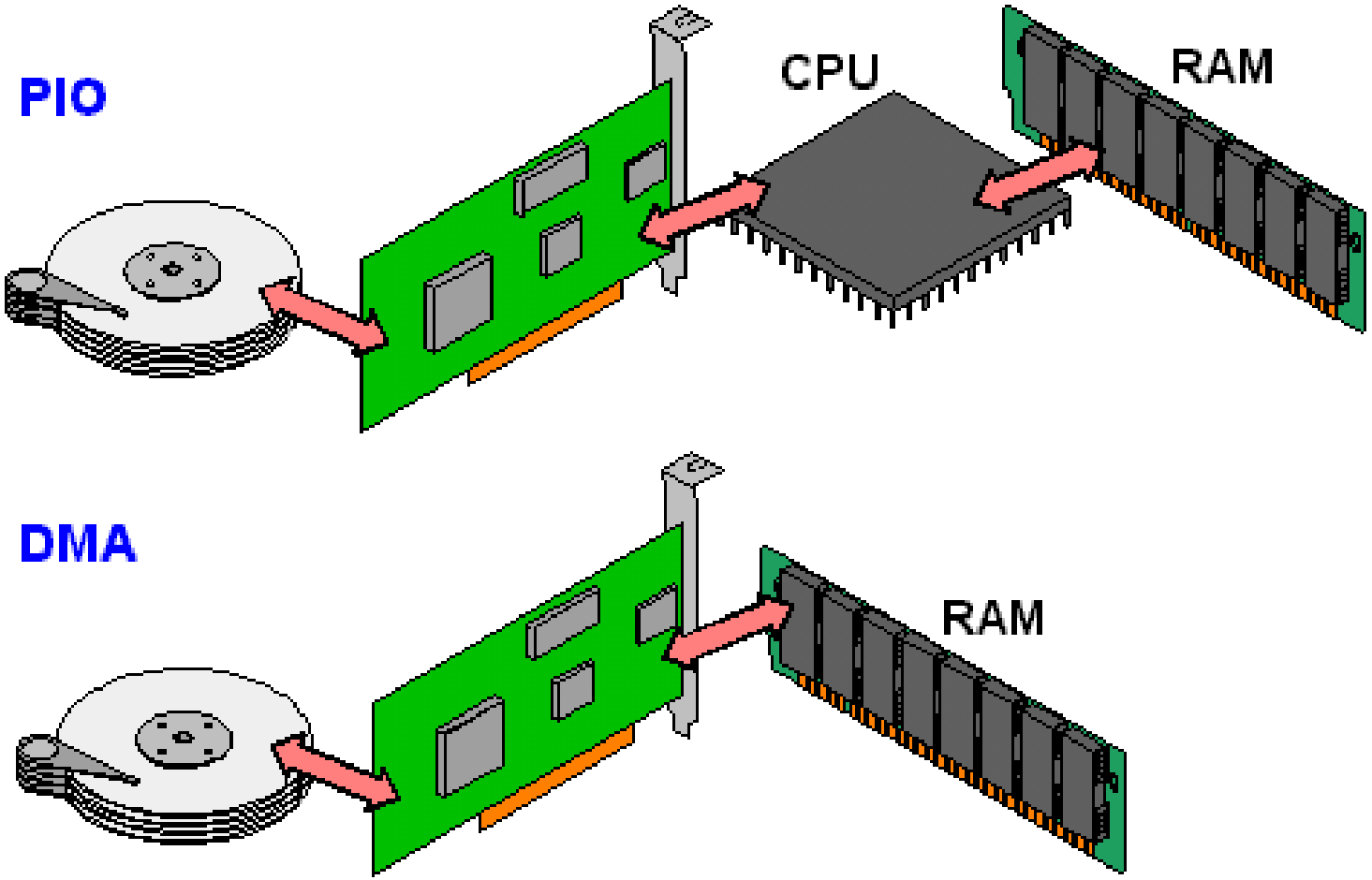
Interrupt driven I/O

- CPU ยังคงควบคุมการทำงานต่างๆ ได้แก่ ตรวจสอบสถานะการทำงาน, ส่งคำสั่ง Read/Write และ รับส่งข้อมูล แต่จะลดบทบาทลง
- หลังจากการส่งคำสั่งแล้ว I/O Module จะรับหน้าที่ทำงานส่วนที่เหลือแทน โดยเฉพาะการรอให้การทำงานเสร็จสิ้น ทำให้ CPU สามารถหันไปทำงานอื่นได้
- I/O Module จะคอย polling อุปกรณ์ว่าทำงานเสร็จสิ้นแล้วหรือไม่ เช่น เมื่อรับคีย์บอร์ด ก็ตรวจสอบว่ามีการกดคีย์บอร์ดเสร็จแล้วหรือไม่
- เมื่อการทำงานเสร็จสิ้น I/O Module จะ Interrupt CPU เพื่อแจ้งว่าการทำงานเสร็จสิ้นแล้ว ให้ CPU มารับข้อมูลไป
- การทำงานลักษณะนี้ มีความซับซ้อนกว่า แต่เป็นการใช้เวลาของ CPU ได้คุ้มค่ามากกว่า

Direct Memory Access (DMA)

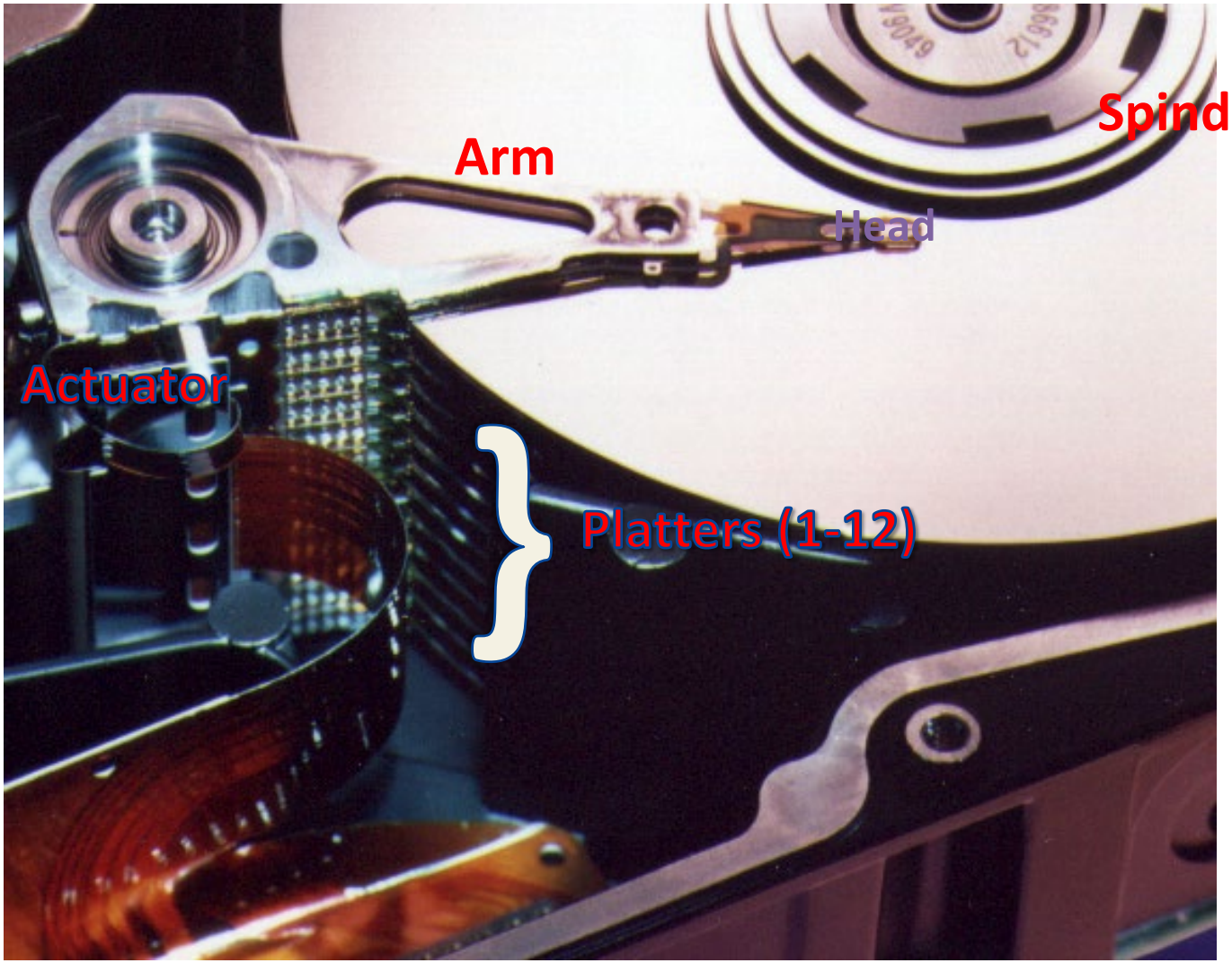
- ใช้กับอุปกรณ์ที่ส่งข้อมูลคราวละมากๆ
- ลดการทำงานของ CPU
- เพิ่มความเร็วในการโอนถ่ายข้อมูล โดยอุปกรณ์สามารถจะติดต่อกับหน่วยความจำได้โดยตรงโดยไม่ผ่าน CPU
- ต้องมี Hardware เพิ่ม : DMA Controller
  - กำหนดอุปกรณ์ที่จะส่ง
  - กำหนดตำแหน่งที่จะรับ
  - กำหนดขนาดข้อมูล
  - กำหนดทิศทางของข้อมูล

PIO vs. DMA

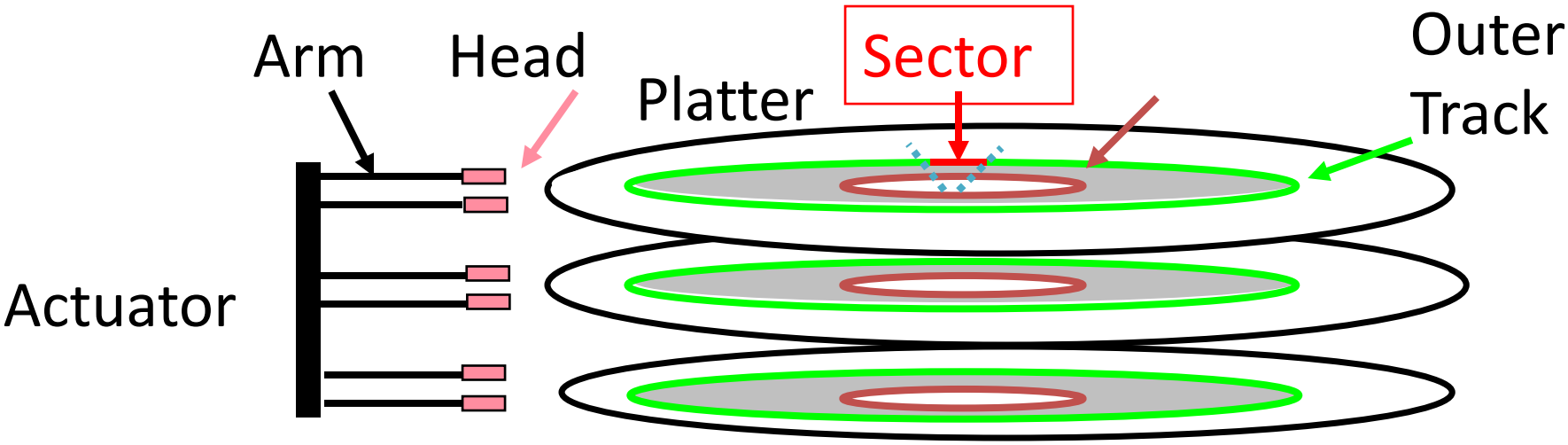


Disk Storage : magnetic disk

Disk Storage : magnetic disk



Disk Storage : magnetic disk



- แผ่นจานเรียกว่า platter (มักจะมีหลายจาน) มักจะมีข้อมูลบันทึกทั้งสองด้าน
- ข้อมูลจะบันทึกใน tracks ซึ่งจะแบ่งออกเป็น sectors (เช่น 512 Bytes) (100-500 sector ต่อ track)
- Actuator จะเลื่อนหัวอ่าน (head) ไปยัง track ที่ต้องการ เรียกว่า “seek” และรอให้ sector ที่ต้องการอ่านหมุนจนมาอยู่ใต้หัวอ่าน จากนั้นจึงจะอ่านหรือเขียน
- ฮาร์ดดิสก์ที่นิยมในปัจจุบัน มีขนาด 3.5 นิ้ว และ 2.5 นิ้ว แต่ความนิยมก็ลดลงเป็นลำดับเนื่องจาก SSD มีราคาถูกลง

Disk Storage : magnetic disk

- Disk Access Time = Seek Time + Rotation Time + Transfer Time + Controller Overhead
  - Seek Time = ระยะเวลาที่หัวอ่านเลื่อนมาจนถึง track ที่ต้องการ
  - Rotation Time = ระยะเวลาที่ดิสก์หมุนจนตำแหน่งที่ sector แรกที่ต้องการอ่าน เคลื่อนมาอยู่ใต้หัวอ่าน
  - Transfer Time = ระยะเวลาที่ข้อมูลอ่านจากหัวอ่านไปที่ Controller

Disk Storage : magnetic disk

- Rotation time
  - เนื่องจาก rotation time หายาก จึงมักใช้ค่าเวลาการหมุนครึ่งรอบเป็น rotation time
  - เช่น 5400 RPM = 5400/60 round per second = 90 round per second
  - Rotation time = 0.5 / rotation = 0.5/90 = 0.0056 sec. = 56 ms.

- จงหา rotation time ของฮาร์ดดิสก์ความเร็ว 7,200 rpm และ 15,000 rpm

Example

- กำหนดให้ HDD ตัวหนึ่งมีอัตราการหมุน 15,000 RPM, มี average seek time = 4 ms., มี transfer rate 100 MB/sec, มี controller overhead = 0.2 ms
- ให้หา average time ในการอ่าน/เขียน 1 sector (512 byte)
- Disk Access Time = Seek Time + Rotation Time + Transfer Time + Controller Overhead
- = 4 + 0.5/15000 RPM + 0.5 KB / 100 MB/s
  - 0.5/15000 RPM = 0.5/250 RPS = 2 ms
  - 0.5 KB/100 MB/s = (0.5/100x1000)*1000 ms = 0.005 ms
- = 4 + 2 + 0.005 + 0.2 = 6.2 ms

Exercise

- กำหนดให้ HDD ตัวหนึ่งมีอัตราการหมุน 7,200 RPM, มี average seek time = 12 ms., มี transfer rate 128 MB/sec, มี controller overhead = 0.2 ms
- ใช้เวลาในการอ่าน 500 sector แบบต่อเนื่อง (sequential) และแบบสุ่ม (random) ใช้เวลาเท่าไร

Disk Performance : Case Study

- Ex. Random access workload of 500 read requests
  - Disk access time = seek time + rotation time + transfer time
    - Seek time : avg. seek time = 10.5 ms
    - Rotation Time : ½ x 7200 RPM rotate once ≅ 4.15 ms
    - Transfer Time : bandwidth 54 MB/s :- 512 b = 9.5 μs
  - Total (10.5 + 4.15 + .0095) * 500 = 7.33 seconds
- Ex. Sequential access workload of 500 read requests
  - Disk access time = seek time + rotation time + transfer time
    - Seek time : avg. seek time = 10.5 ms
    - Rotation Time : ½ x 7200 RPM rotate once ≅ 4.15 ms
    - Transfer time of 500 sectors
      - Outer tracks = 500 x 512 byte/sector x 1 / 54 MB/s = 4.8ms
      - Inner tracks = 500 x 512 byte/sector x 1 / 128 MB/s = 2 ms
  - Total 10.5+4.15+2 = 16.7 ms, 10.5+4.15+4.8 = 19.5 ms

Reliability & Availability

- อุปกรณ์ I/O โดยเฉพาะอุปกรณ์เก็บข้อมูลมักจะมีตัวเลขที่บอกความคงทนต่อการใช้งาน
  - MTTF : mean time to failure
  - MTTR : mean time to repair
  - MTBF : Mean time between failures (MTTF + MTTR)
- Availability = MTTF / (MTBF)

Disk Storage : magnetic disk

Characteristics	Seagate ST33000655SS	Seagate ST31000340NS	Seagate ST973451SS	Seagate ST9160821AS
Disk diameter (inches)	3.50	3.50	2.50	2.50
Formatted data capacity (GB)	147	1000	73	160
Number of disk surfaces (heads)	2	4	2	2
Rotation speed (RPM)	15,000	7200	15,000	5400
Internal disk cache size (MB)	16	32	16	8
External interface, bandwidth (MB/sec)	SAS, 375	SATA, 375	SAS, 375	SATA, 150
Sustained transfer rate (MB/sec)	73–125	105	79–112	44
Minimum seek (read/write) (ms)	0.2/0.4	0.8/1.0	0.2/0.4	1.5/2.0
Average seek read/write (ms)	3.5/4.0	8.5/9.5	2.9/3.3	12.5/13.0
Mean time to failure (MTTF) (hours)	1,400,000 @ 25°C	1,200,000 @ 25°C	1,600,000 @ 25°C	—
Annual failure rate (AFR) (percent)	0.62%	0.73%	0.55%	—
Contact start-stop cycles	—	50,000	—	>600,000
Warranty (years)	5	5	5	5
Nonrecoverable read errors per bits read	<1 sector per 10^{16}	<1 sector per 10^{15}	<1 sector per 10^{16}	<1 sector per 10^{14}
Temperature, shock (operating)	5°–55°C, 60 G	5°–55°C, 63 G	5°–55°C, 60 G	0°–60°C, 350 G
Size: dimensions (in.), weight (pounds)	1.0" × 4.0" × 5.8", 1.5 lbs	1.0" × 4.0" × 5.8", 1.4 lbs	0.6" × 2.8" × 3.9", 0.5 lbs	0.4" × 2.8" × 3.9", 0.2 lbs
Power: operating/idle/ standby (watts)	15/11/—	11/8/1	8/5.8/—	1.9/0.6/0.2
GB/cu. in., GB/watt	6 GB/cu.in., 10 GB/W	43 GB/cu.in., 91 GB/W	11 GB/cu.in., 9 GB/W	37 GB/cu.in., 84 GB/W
Price in 2008, $/GB	~ $250, ~ $1.70/GB	~ $275, ~ $0.30/GB	~ $350, ~ $5.00/GB	~ $100, ~ $0.60/GB

Flash storage

- เกิดจากพัฒนาการของหน่วยความจำ flash

Toshiba flash
2 GB

Samsung flash
16 GB

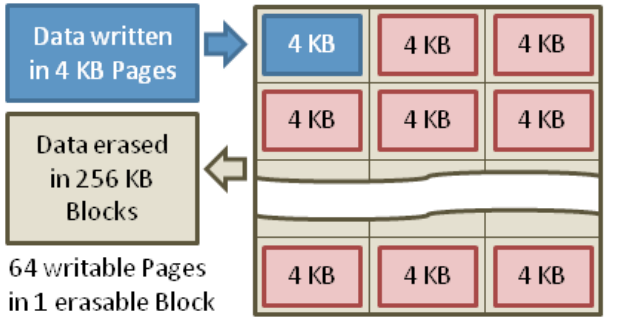
Toshiba 1.8-inch HDD
80, 120, 160 GB

Toshiba flash
32, 64 GB

shuffle nano classic touch

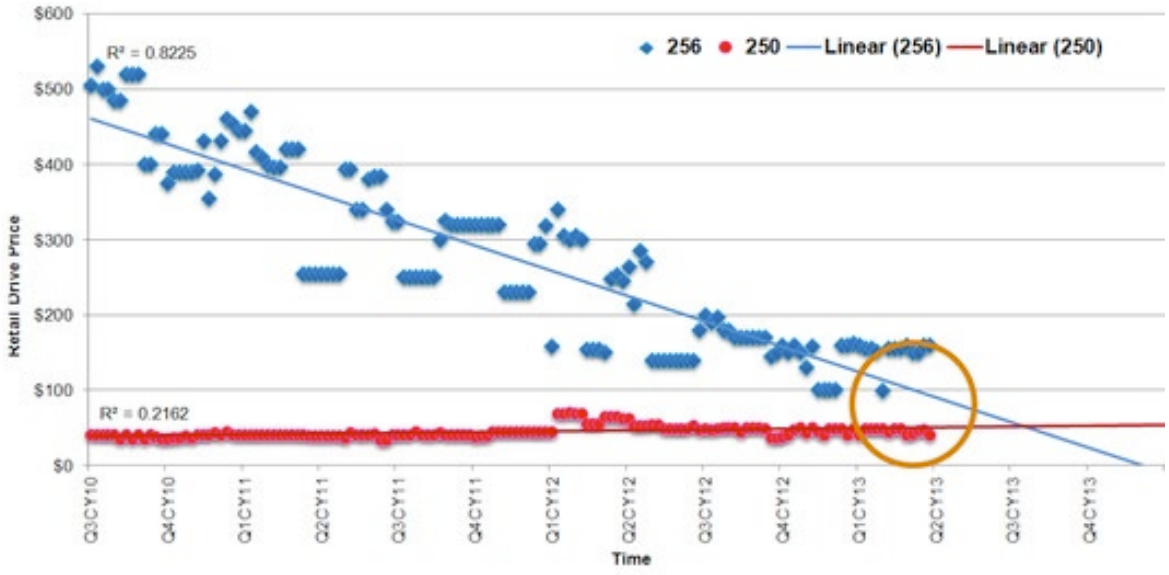
Flash storage

- ใช้สารกึ่งตัวนำ ทำให้ไม่มีส่วนเคลื่อนไหว => มีความทนทาน
  - ไม่มี seek time และ rotation time
- มีความเร็ว 100–1000 เท่าของฮาร์ดดิสก์
  - Reading
    - SSD : Read 4 KB Page: ~25 µsec
    - SATA: 300-600MB/s => ~4 x10^3 b / 400 x 10^6 bps => 10 us
  - Writing
    - (~200 µs – 1.7 ms)
    - ในการเขียนจะเขียนได้เฉพาะ page ที่ว่าง ดังนั้นหากมีข้อมูลอยู่จะต้องลบก่อน ~1.5ms
    - Controller จะต้องเก็บ pool of empty blocks เอาไว้
- ใช้พลังงานต่ำ และ น้ำหนักเบา
- มีปัญหาเรื่อง wear leveling คือ มีจำนวนครั้งของการ write จำกัด (10,000 – 100,000)



Typical NAND Flash Pages and Blocks

HDD vs SSD Comparison



Price Crossover Point for HDD and SSD

	2012	2013	2014	2015E	2016F	2017F
HDD	0.09	0.08	0.07	0.06	0.06	0.06
2.5" SSD	0.99	0.68	0.55	0.39	0.24	0.17

SSD vs HDD

Usually 10 000 or 15 000 rpm SAS drives

0.1 ms	**Access times** SSDs exhibit virtually no access time	**5.5 ~ 8.0 ms**
SSDs deliver at least **6000 io/s**	**Random I/O Performance** SSDs are at least 15 times faster than HDDs	HDDs reach up to **400 io/s**
SSDs have a failure rate of less than **0.5 %**	**Reliability** This makes SSDs 4 - 10 times more reliable	HDD''s failure rate fluctuates between **2 ~ 5 %**
SSDs consume between **2 & 5 watts**	**Energy savings** This means that on a large server like ours, approximately 100 watts are saved	HDDs consume between **6 & 15 watts**
SSDs have an average I/O wait of **1 %**	**CPU Power** You will have an extra 6% of CPU power for other operations	HDDs' average I/O wait is about **7 %**
the average service time for an I/O request while running a backup remains below **20 ms**	**Input/Output request times** SSDs allow for much faster data access	the I/O request time with HDDs during backup rises up to **400~500 ms**
SSD backups take about **6 hours**	**Backup Rates** SSDs allows for 3 - 5 times faster backups for your data	HDD backups take up to **20~24 hours**

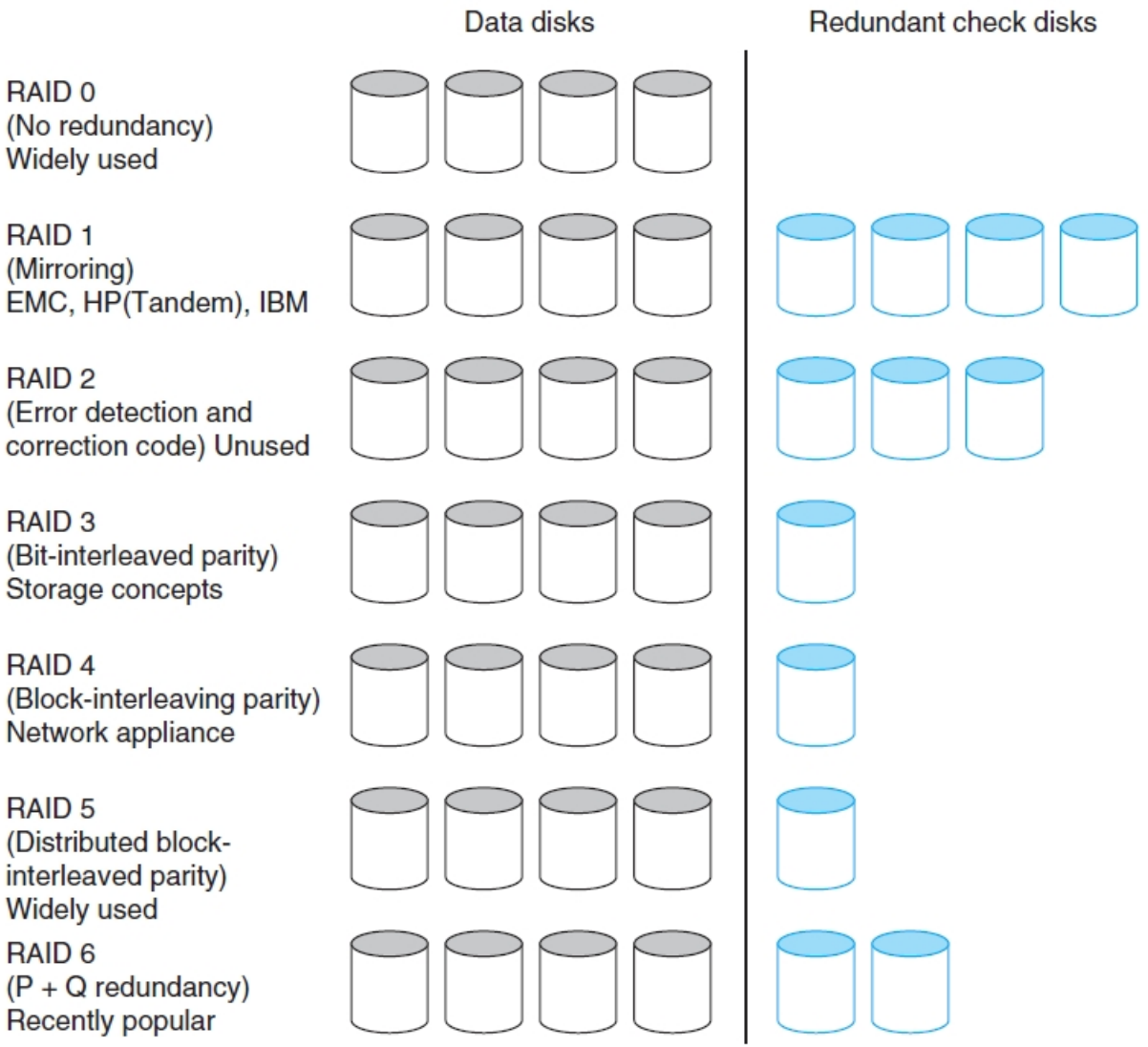
SSD prices drop much faster than HDD

I/O Protocol

Characteristic	Firewire (1394)	USB 2.0	PCI Express	Serial ATA	Serial Attached SCSI
Intended use	External	External	Internal	Internal	External
Devices per channel	63	127	1	1	4
Basic data width (signals)	4	2	2 per lane	4	4
Theoretical peak bandwidth	50 MB/sec (Firewire 400) or 100 MB/sec (Firewire 800)	0.2 MB/sec (low speed), 1.5 MB/sec (full speed), or 60 MB/sec (high speed)	250 MB/sec per lane (1x); PCIe cards come as 1x, 2x, 4x, 8x, 16x, or 32x	300 MB/sec	300 MB/sec
Hot pluggable	Yes	Yes	Depends on form factor	Yes	Yes
Maximum bus length (copper wire)	4.5 meters	5 meters	0.5 meters	1 meter	8 meters
Standard name	IEEE 1394, 1394b	USB Implementors Forum	PCI-SIG	SATA-IO	T10 committee

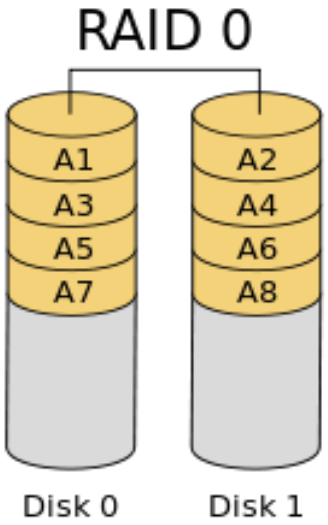
RAID

- Redundant Arrays of Inexpensive Disks



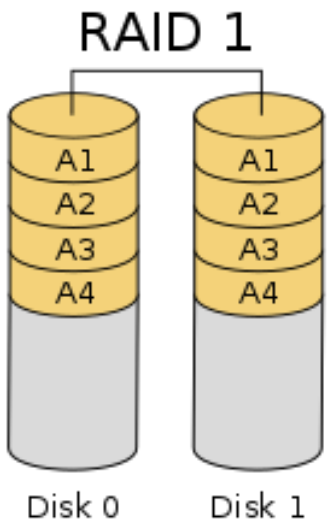
RAID 0

- No Redundancy (RAID 0)
- ข้อมูลจะถูกเก็บในดิสก์หลายๆ ลูก เรียกว่า striping เช่น หากมีดิสก์ 2 ลูกต่อแบบ RAID 0 บล็อกที่ 0 จะเก็บที่ Disc 0 และบล็อกที่ 1 จะเก็บที่ Disc 1
- เมื่อมองจาก Software Level แล้วดิสก์ทั้ง 2 ลูก จะเหมือนกับดิสต์ตัวเดียวกัน
- ข้อดี : เร็วขึ้น และ มีขนาดของ Hard disk ใหญ่ขึ้น
- ข้อเสีย : หาก Hard disk ลูกใดเสียหาย จะสูญเสียข้อมูลทั้งหมด
- ปัจจุบัน ไม่นิยมใช้งานแล้ว



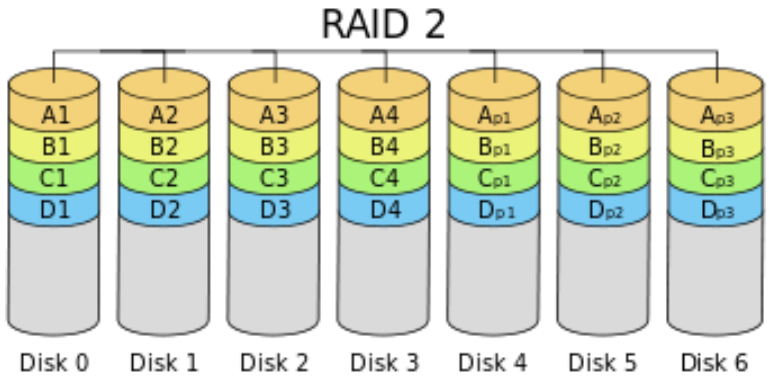
RAID 1

- Mirroring (RAID 1)
- ข้อมูลจะเก็บใน Hard Disk จำนวน 2 ชุดที่เหมือนกัน จึงเรียกว่า mirroring หรือ *shadowing*
- เมื่อมีการเขียนข้อมูล จะเขียนลงใน Hard disk ทั้ง 2 ชุด พร้อมๆ กัน แต่การอ่านจะอ่านจาก Disk เพียงชุดเดียว
- ข้อดี : มีข้อมูล 2 ชุด ทำให้เกิด Fault tolerance มีความมั่นคงของข้อมูลสูง หาก Hard disk เสียไป 1 ลูก ก็ยังสามารถใช้งานได้อยู่ และทำให้การอ่านข้อมูลเร็วขึ้น
- ข้อเสีย : มีราคาแพง เพราะต้องเปลืองค่า storage อีก 1 เท่า



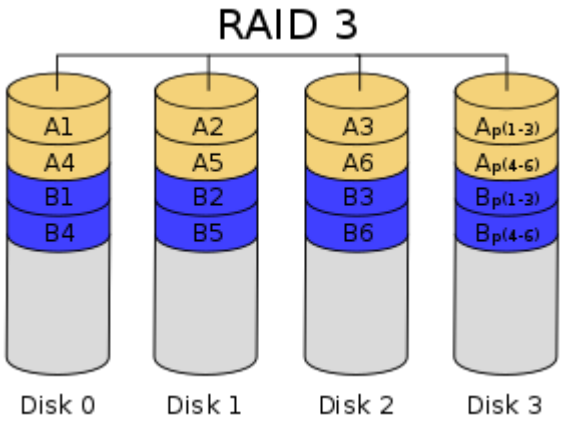
RAID 2

- Error Detecting and Correcting Code (RAID 2)
- เป็นการนำแนวคิดของ ECC มาใช้งาน (เดิมที ECC มีการนำมาใช้งานกับหน่วยความจำ โดยเรียกว่าหน่วยความจำชนิด ECC โดยการเพิ่มบิตเข้ามา 3 บิต รวมเป็น 11 บิตต่อ 1 ไบต์ ทำให้เมื่อบิตใดบิตหนึ่งของหน่วยความจำเสียหายไป สามารถจะ correction กลับมาได้ มักนิยมใช้กับเครื่อง Server)
- จะเพิ่มฮาร์ดดิสก์จำนวน 3 ลูก โดยทำ Error Detecting and Correcting ในระดับบิต
- ข้อดี : สามารถ Correction ข้อมูลได้หากเสียหายบางส่วน
- ข้อเสีย : เปลือง Hard disk ประสิทธิภาพต่ำ
- ปัจจุบัน ไม่มีการใช้งาน

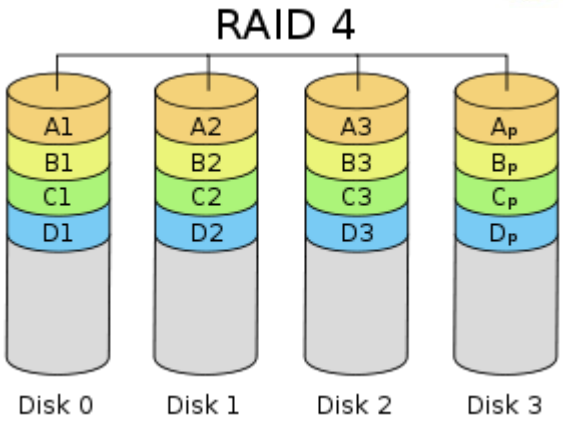


RAID 3

- Byte Level Parity (RAID 3)
- ใช้ฮาร์ดดิสก์เพียง 1 ลูกในสำหรับเก็บ redundant Information เพื่อใช้ในการกู้คืนข้อมูล หาก hard disk บางลูกมีความเสียหาย การอ่านหรือเขียน จะต้องอ่านหรือเขียนทุกลูกพร้อมๆ กัน
- ข้อดี : สามารถ Correction ข้อมูลได้หากเสียหายบางส่วน ความเร็วเพิ่มขึ้น และใช้ Hard disk น้อยกว่า RAID 2
- ข้อเสีย : ใช้เวลาในการอ่านมากขึ้น เนื่องจากต้องคอยตรวจสอบความถูกต้อง เกิดคอขวดที่ redundant disk



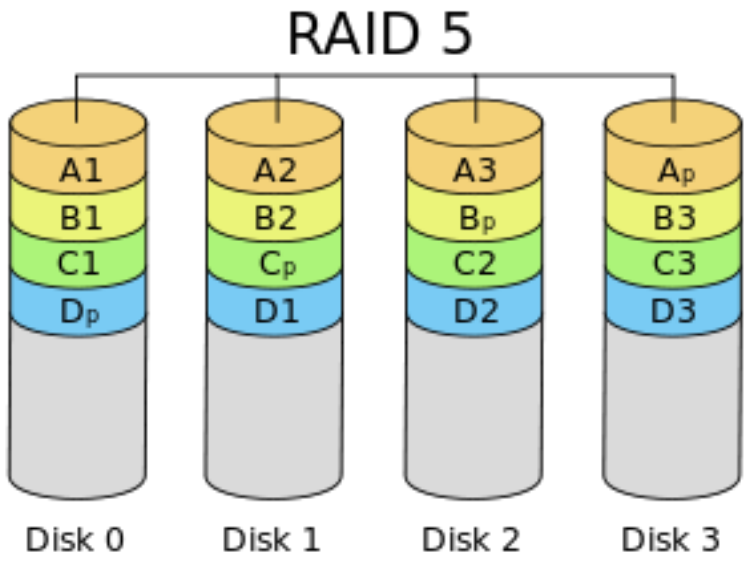
RAID 4



- Block-Interleaved Parity (RAID 4)
- คล้ายกับ RAID 3 แต่จะทำเป็น Block level
- จากรูปตัวอย่างมี Hard disk จำนวน 3 ลูก โดยมี redundant check จำนวน 1 ลูก โดยข้อมูลในดิสก์ redundant check นี้จะสร้างจากข้อมูลใน hard disk ทั้ง 4 ลูก ในขณะที่ disk 1 จะอ่านข้อมูลขนาดเล็ก (เช่น 1 sector) ก็จะมีการอ่าน disk 1 และ redundant check disk ทำให้ในขณะเดียวกัน สามารถจะอ่าน disk 2 หรือ disk 3 ในเวลาเดียวกันได้ ทำให้ประสิทธิภาพการทำงานสูงขึ้น
- ข้อดี : สามารถ Correction ข้อมูลได้หากเสียหายบางส่วน ทำงานได้เร็วขึ้น
- ข้อเสีย : เกิดคอขวดที่ redundant disk (on write)

RAID 5

- Distributed Block-Interleaved Parity (RAID 5)
- ทำงานคล้าย RAID-4 แต่นำ Parity Block มากระจายในทุกฮาร์ดดิสก์ ทำให้ลดคอขวดที่ Parity Disk ไป
- เป็นเทคโนโลยีที่นิยมใช้มากที่สุดในปัจจุบัน



For your attention