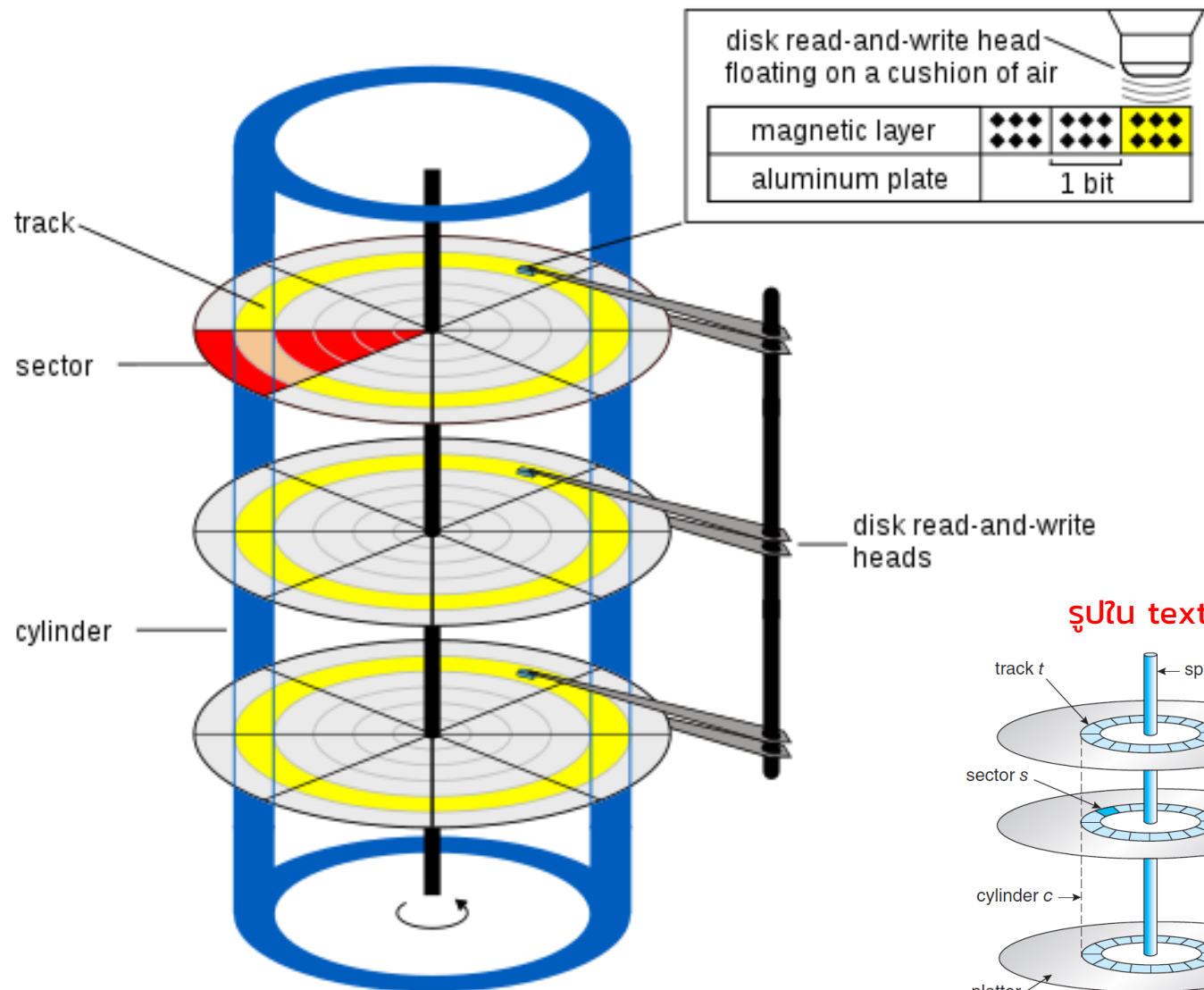
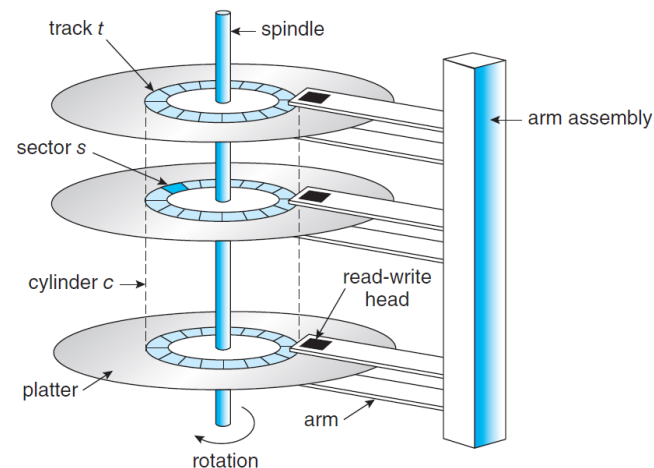


Mass-Storage Structure



รูปใน textbook เข้าใจยาก





Platter
Disk arm
Tracks
Sectors
Cylinder
Rotations per minute (RPM)
54k 72k 10k 15k
Transfer rate
Positioning time
Random-access time
Seek time
Rotational latency
Head crash
Removable
Effective transfer rate
Nonvolatile memory (NVM)
Solid-state disk (SSD)
USB drive

รูปจากอินเทอร์เน็ต

Seagate – จะผลิตฮาร์ดดิสก์งานแม่เหล็กความจุ 50 TB ในปี 2026 ส่วนขนาด 18/20 TB ภายในปี 2020

12 Nov 19 - By EDDY I

<https://notebookspec.com/seagate-will-have-50tb-hdds-by-2026-18tb-and-20tb-launch-in-2020/501531/>



Modern disk drives = large one-dimensional array of logical blocks.
A logical block is the smallest unit of transfer.
The size of a logical block is usually 512 bytes (from factory),
but can be low-level formatted to 1024 bytes (less header and
trailer, but more internal fragmentation).

Block (sector)

Header
Data (512 bytes)
Trailer

Constant linear velocity - CD-ROM, DVD-ROM



ความเร็วเชิงเส้น (m/s) คงที่ เวลาหัวอ่านเลื่อนเข้าไป
อ่านข้อมูลที่ track ข้างใน แขนจะหมุนเร็วขึ้น (rpm)
เพื่อให้ความเร็วเชิงเส้นคงที่

Constant angular velocity - hard disk



ใช้ความเร็วรอบ (rpm) คงที่เสมอ เปลืองไฟ
data density decreases in outer tracks

Magnetic tapes ใช้ backup ข้อมูลเท่านั้น



Magnetic Tapes เป็นแบบ sequential access จุได้ถึง 580 TB เหมาะกับ data backup ที่ต้องเก็บข้อมูลไว้เป็นเวลานานหลายปี

IBM's Tale of the Tape

Nearly 70 years of tape innovation: reliable, secure & energy efficient for Hybrid Clouds

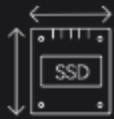


	2006	2010	2014	2015	2017	2020
Areal Density (bits per sq inch)	6.67 Billion	29.5 Billion	85.9 Billion	123 Billion	201 Billion	317 Billion
Cartridge Capacity (Terabytes)	8	35	154	220	330	580
# of Books Stored*	8 Million	35 Million	154 Million	220 Million	330 Million	580 Million
Track Width	1.5 µm	0.45 µm	0.177 µm	0.140 µm	103 nm	56.2 nm
Linear Density (bits per inch)	400'000	518'000	600'000	680'000	818'000	702'000
Tape Material	Barium Ferrite	Barium Ferrite	Barium Ferrite	Barium Ferrite	Sputtered Media	Strontium Ferrite
Tape Thickness (micrometers - µm)	6.1	5.9	4.3	4.3	4.7	4.3
Tape Length (meters)	890	917	1255	1255	1098	1255

© Copyright IBM Corporation 2020. IBM and the IBM logo are trademarks of IBM Corp. registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml

* assumes 1MB of text data per book





Form Factor

M.2



Capacity

250GB, 500GB, 1TB



Sequential Read Speed

Up to 3,500 MB/s



Sequential Write Speed

Up to 3,000 MB/s

M.2 = 22 millimeters wide and 60 mm or 80 mm long

NVM = controller + flash NAND die semiconductor

NAND semiconductors have some characteristics that present their own storage and reliability challenges. For example, they can be read and written in a “page” increment (similar to a sector), but data cannot be overwritten—rather, the NAND cells have to be erased first. The erasure, which occurs in a “block” increment that is several pages in size, takes much more time than a read (the fastest operation) or a write (slower than read, but much faster than erase). Helping the situation is that NVM flash devices are composed of many die, with many datapaths to each die, so operations can happen in parallel (each using a datapath). NAND semiconductors also deteriorate with every erase cycle, and after approximately 100,000 program-erase cycles (the specific number varies depending on the medium), the cells no longer retain data. Because of the write wear, and because there are no moving parts, NAND NVM lifespan is not measured in years but in **Drive Writes Per Day (DWPD)**. That measure is how many times the drive capacity can be written per day before the drive fails. For example, a 1 TB NAND drive with a 5 DWPD rating is expected to have 5 TB per day written to it for the warranty period without failure.

Flash Translation Layer (FTL)

valid page	valid page	invalid page	invalid page
invalid page	valid page	invalid page	valid page

เลเยอร์ (layer) ข้างบน เช่น OS
จะมองเห็นแค่ valid page เท่านั้น

1 page = 4KB และ 1 block = 128 pages เขียนได้แค่ทีละ page และเขียนทับ page เดิมไม่ได้ ต้องลบก่อน แต่ต้องลบทั้ง block

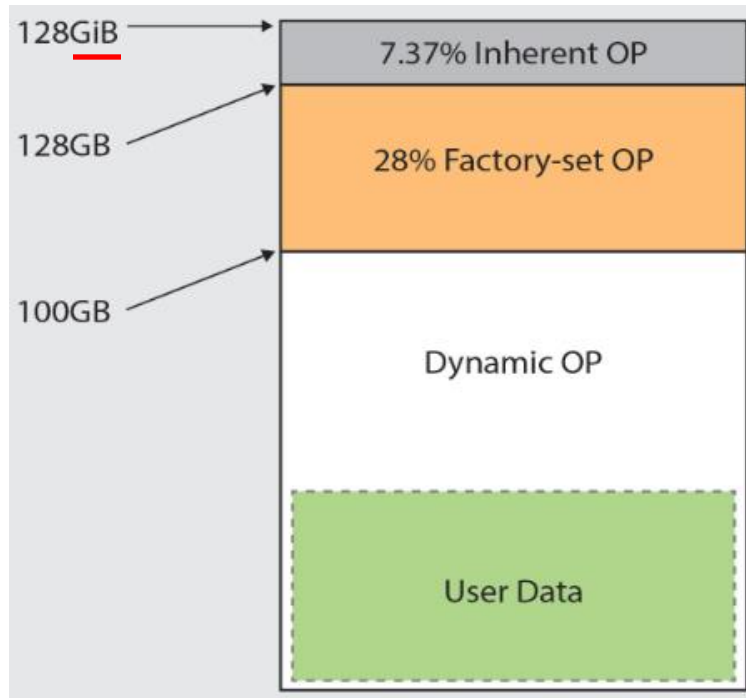
ถ้าเขียน (อัปเดต) page เดิมซ้ำ ๆ จะต้องใช้ page ใหม่เสมอ พอหมด block ก็ต้องขึ้น block ใหม่ ทำให้ block หมดไปเร็วมาก จะลบ invalid page ก็ไม่ได้ เพราะใน block มี valid page ปนอยู่ด้วย (รูปข้างบน)

Garbage collector จะทำงานเป็นระยะ ๆ โดยอ่าน valid page จาก block หนึ่งขึ้นมา แล้วนำไปเก็บไว้ใน over-provisioning space เพื่อจะได้ลบ block นั้นได้ (ต้องลบทีเดียวทั้ง block)

Wear leveling เอา block ที่โดนลบน้อยครั้งที่สุดมาใช้ก่อน เพื่อให้สึกหรอเท่า ๆ กัน

SSD Over-Provisioning And Its Benefits

Many people are blissfully unaware that one **gigabyte (GB)** is precisely 1,000,000,000 bytes, and one **gibibyte (GiB)** is precisely $2^{30} = 1,073,741,824$ bytes, or about 7.37% more than a GB. Many people are also blissfully unaware that storage is properly measured in gigabytes, whereas memory is properly measured in gibibytes. Even though SSDs are built from NAND flash memory chips, they are marketed as storage devices, and SSD manufacturers reserve the extra 7.37% of memory space as a provision for background activities such as garbage collection. In practice, an SSD's performance begins to **decline after it reaches about 50% full**. This is why some manufacturers reduce the amount of capacity available to the user and set it aside as additional over-provisioning.



SSD Over-provisioning Example
Allocation of NAND Flash Memory Inside A
100GB SSD with 28% OP

Over-provision คือ ให้มาเกินที่จะใช้จริง
เวลาซื้อถ้า user data เท่ากัน เช่น 100 GB
ต้องดู OP space ว่ายี่ห้อไหนให้มากกว่า

Volatile Memory

- RAM drives / RAM disks สร้างขึ้นโดย OS เมื่อเริ่ม boot ยังหา disk ไม่เจอ
- ผู้ใช้สามารถสร้าง RAM drive สำหรับการใช้งานทั่วไปได้

ITEMS	DDR4	DDR5
Frequency	1600~3200Mbps	3200~8400Mbps
Density	2Gb, 4Gb, 8Gb, 16Gb	8Gb, 16Gb, 24Gb, 32Gb, 64Gb
On die ECC	No	Yes
Bank	16banks	32banks
VDD/VDDQ	1.2V	1.1V
VPP	2.5V	1.8V
BL	8	16
DFE	No	Yes
Same bank refresh	No	Yes

ihavecpu.com/article/392/มาแน่นอนปีหน้า-sk-hynix-ปล่อยรายละเอียดแรม-ddr5-ความเร็วบัสสูงถึง-8400-mhz

<https://linuxhint.com/create-ramdisk-linux/>

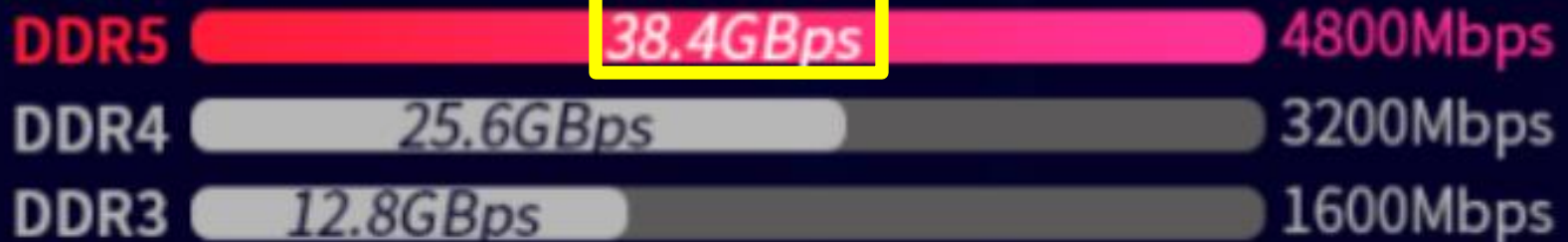
Speed

- Transfer data 1.5x faster at a 38% higher bandwidth than DDR4

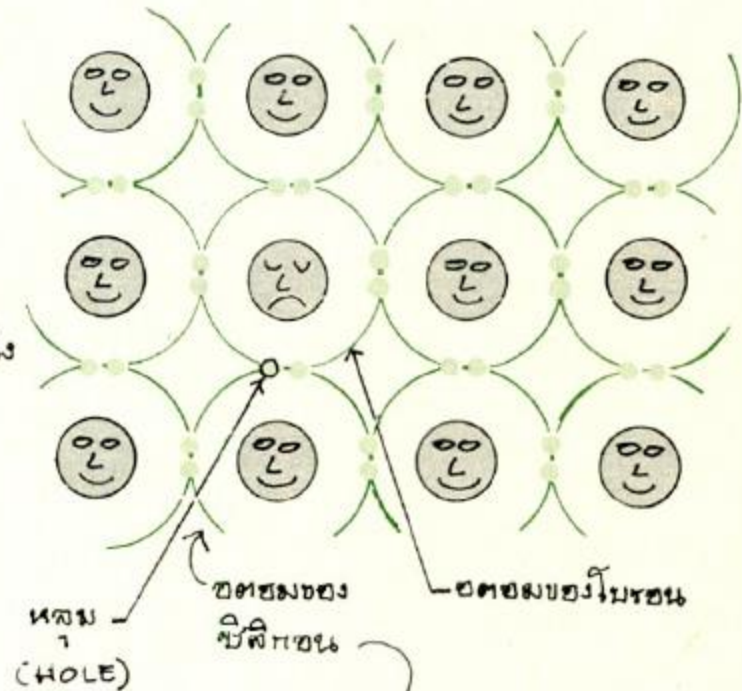


เปรียบเทียบโดยการนำมาวิ่งแข่งกัน

Bus Speed

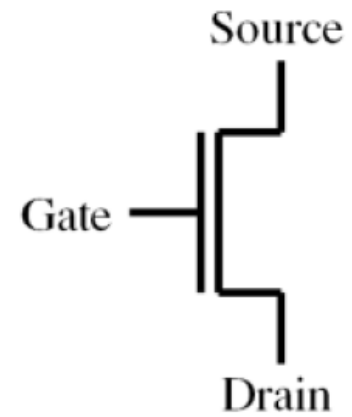
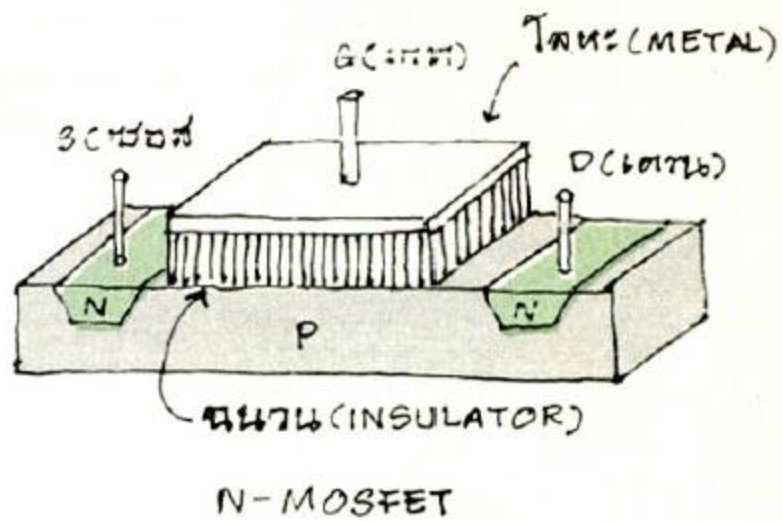


- ชนิดกึ่งตัวนำแบบ P (P-TYPE SILICON) - ผลิตจากซิลิคอนบริสุทธิ์โดยการเติมธาตุที่มีวาเลนซ์ 3 อะตอม เช่น โบรอน (B) หรือ แอลูมิเนียม (Al) เข้าไปแทนที่ซิลิคอนบางส่วน ทำให้เกิดพันธะโควาเลนต์ที่ไม่สมบูรณ์ เกิดเป็นช่องว่างที่เรียกว่า โฮล (HOLE) ซึ่งโฮลเหล่านี้สามารถเคลื่อนที่ผ่านผลึกซิลิคอนได้เหมือนอนุภาคที่มีประจุลบ



- ชนิดกึ่งตัวนำแบบ N (N-TYPE SILICON) - ผลิตจากซิลิคอนบริสุทธิ์โดยการเติมธาตุที่มีวาเลนซ์ 5 อะตอม เช่น ฟอสฟอรัส (P) หรือ อาร์เซนิก (As) เข้าไปแทนที่ซิลิคอนบางส่วน ทำให้เกิดพันธะโควาเลนต์ที่ไม่สมบูรณ์ เกิดเป็นอิเล็กตรอนอิสระที่เคลื่อนที่ผ่านผลึกซิลิคอนได้เหมือนอนุภาคที่มีประจุลบ





N-channel MOSFET

□ การทำงานของ MOSFET - MOSFET ทั้งหลาย จะไม่เหมือนกัน JFET เกาของ MOSFET

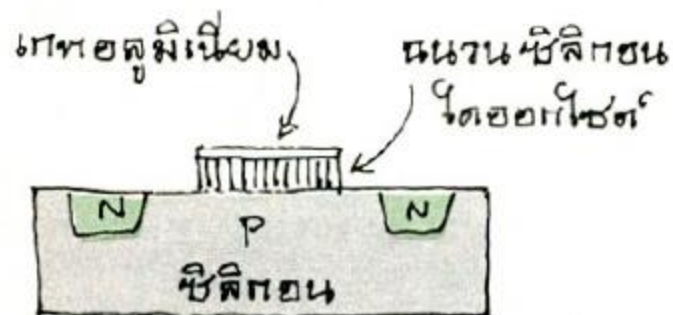
จะไม่ต่อกันทางไฟฟ้า กับชั้น ช่องแคบ ด้าน เชน

ชั้นฉนวนเหมือนแก้ว ของซิลิกอนไดออกไซด์

ไซด์ (SILICON DIOXIDE) จะช่วยเกา

โลหะ ที่สัมผัสกับส่วนที่เหลืของ

ทารานชีส์เตอร์ ออกจากกัน



เมื่อเกาของ MOSFET มีแรงเคลือบ

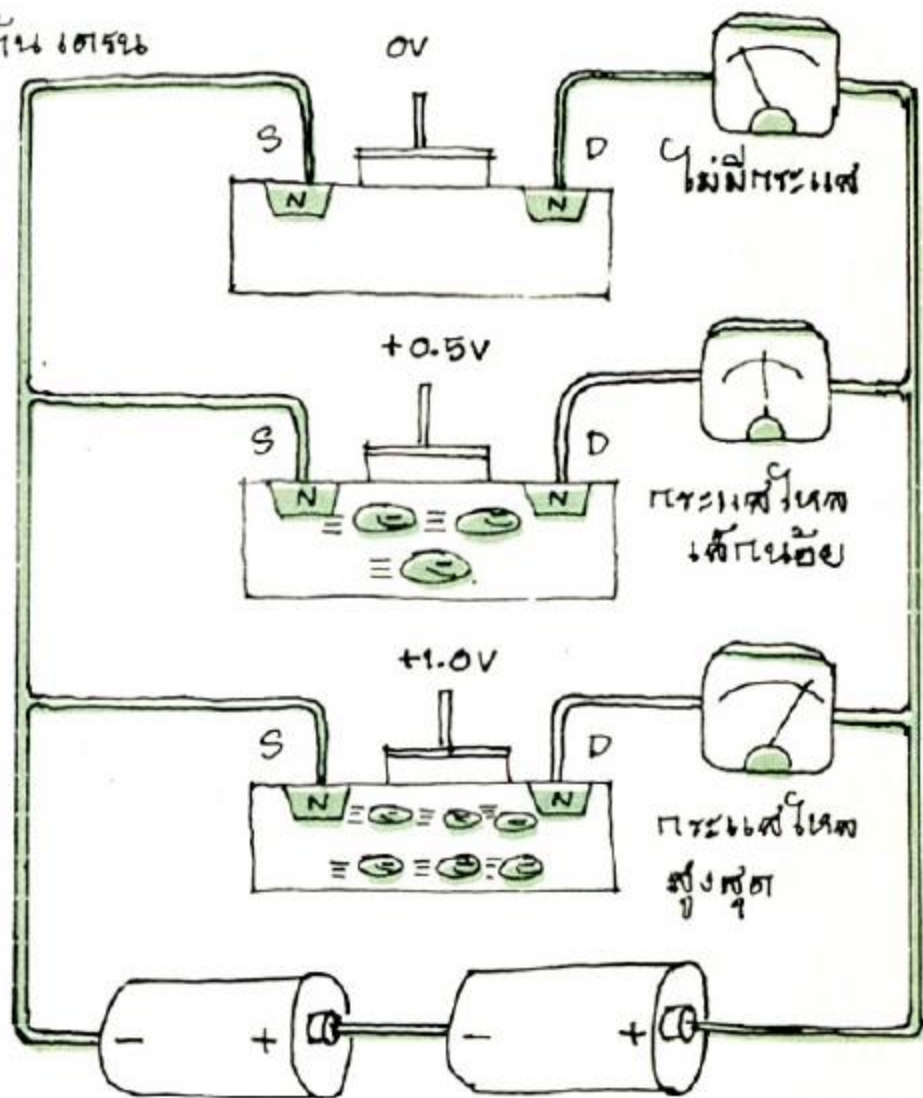
เป็นบวก จะดึงดูด (ATTRACT)

อิเล็กตรอนเข้าไปสู่ใต้เกา จะทำให้มี

กระแสไหลจาก N-แชนแนล ผ่าน

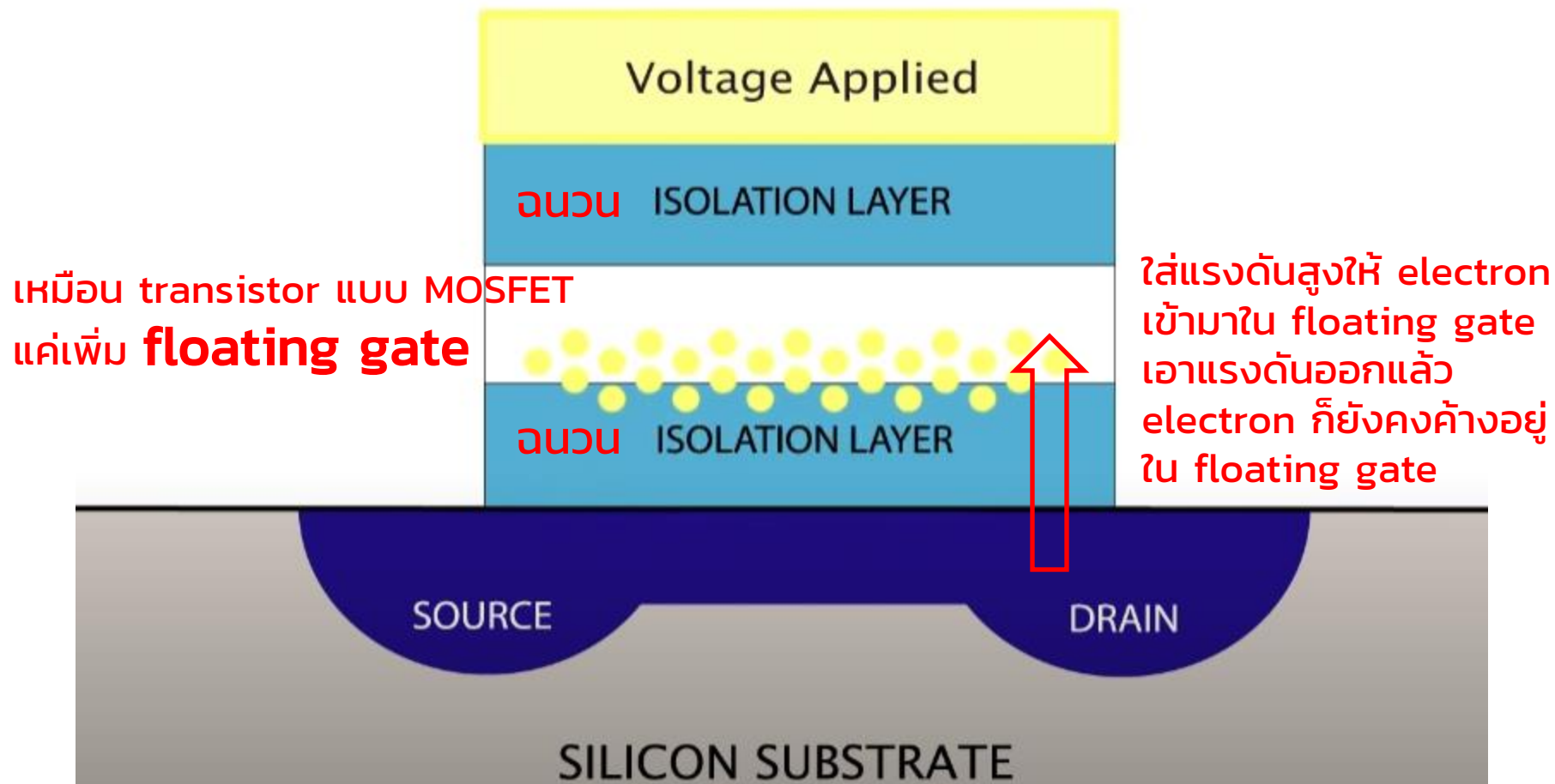
P-แชนแนล ระหว่างช่องแคบ เชน เชน

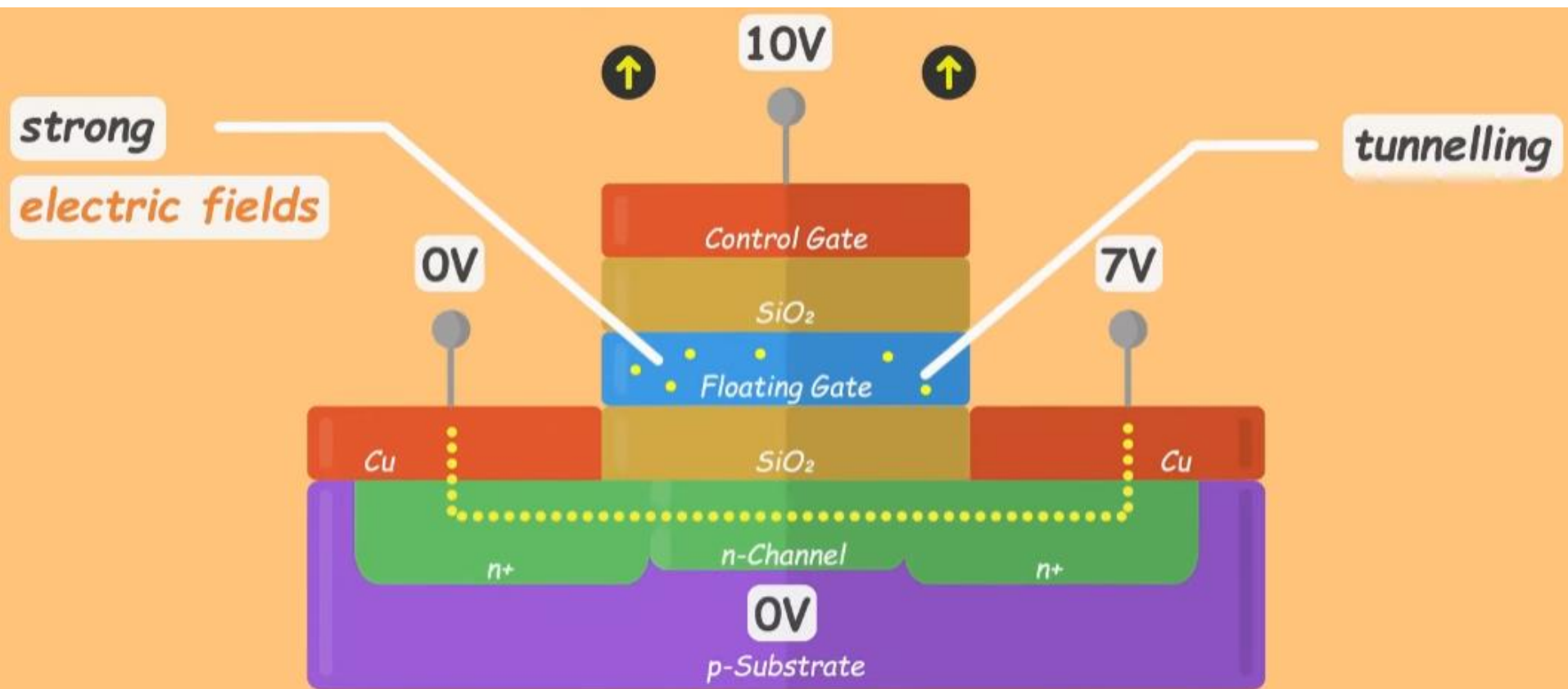
แรงเคลือบที่เกา จะเป็นตัวกำหนดค่าการไหลของ แชนแนล



Programming NAND Flash Cells

When the floating gate is charged it is programmed to a binary value of 0

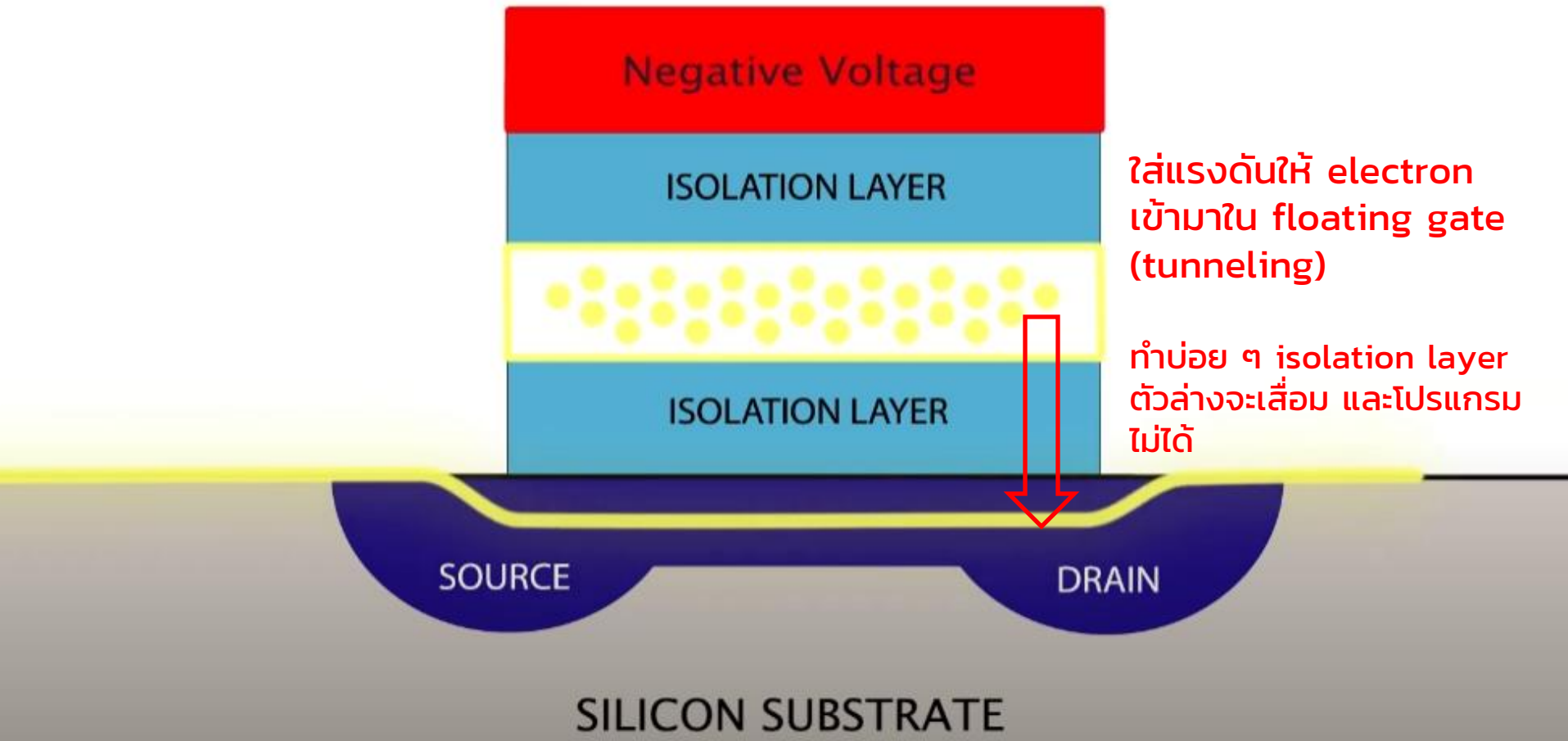




Since the FG is surrounded by highly resistive material, the charge contained in it remains unchanged for long periods of time, nowadays typically **longer than 10 years**.

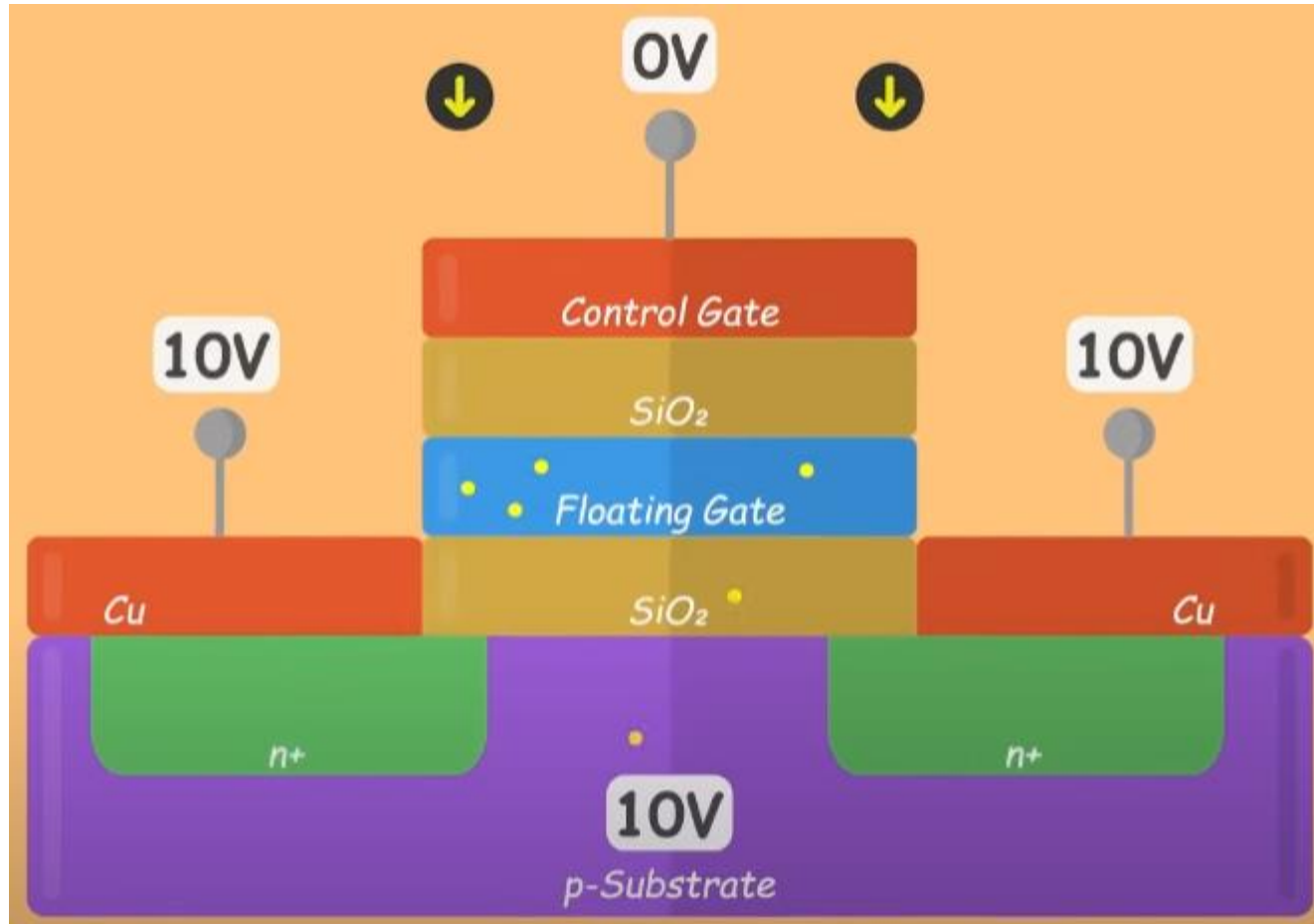
Erasing a NAND flash cell

When the floating gate is not charged it is in an erased state and has a binary value of 1



วิศวกรญี่ปุ่นเรียกว่า flash (cell) เพราะว่ามันลบได้เร็วมาก เหมือนแสง flash ของกล้องถ่ายรูป

<https://www.youtube.com/watch?v=AO7CNNZmtTw>



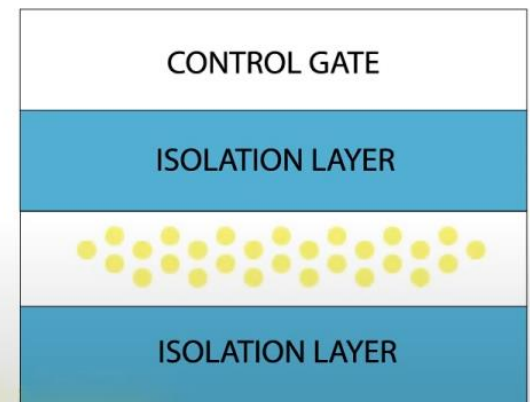
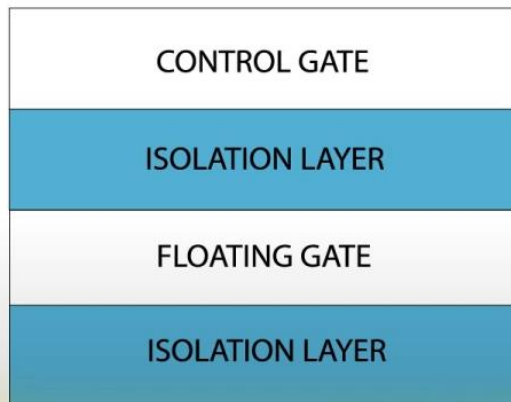
Reading NAND Flash Cells

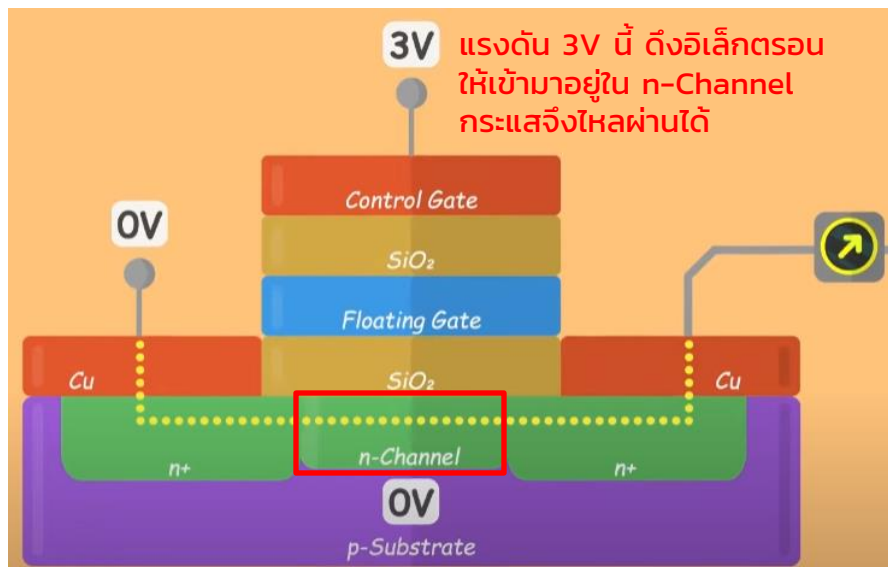
A reference voltage is applied across the source and drain to measure the current thresholds

‘1’

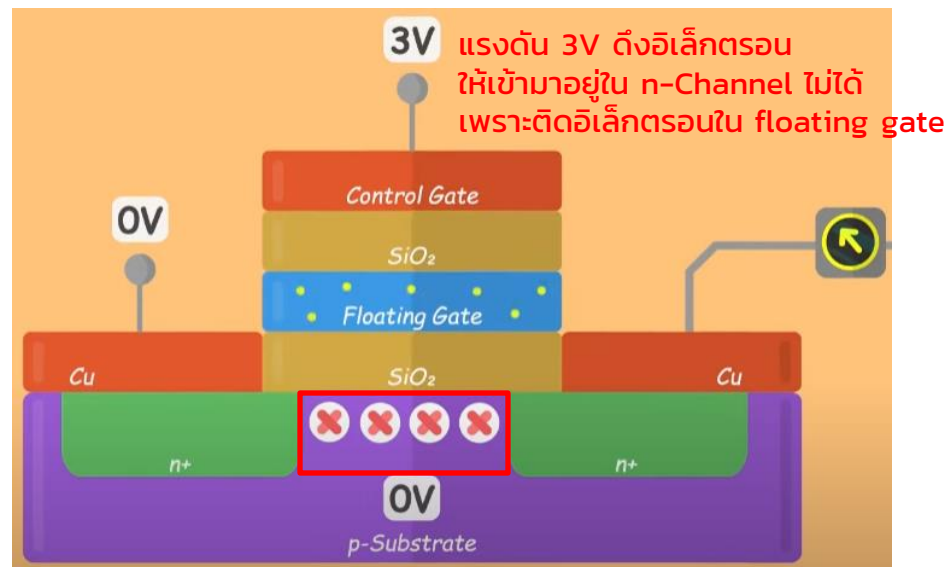
‘0’

ดูคำอธิบาย
ในหน้าถัดไป





1



0

Disk Attachment

1. Host-Attached Storage (local I/O port)

IDE, **Serial ATA (SATA)**

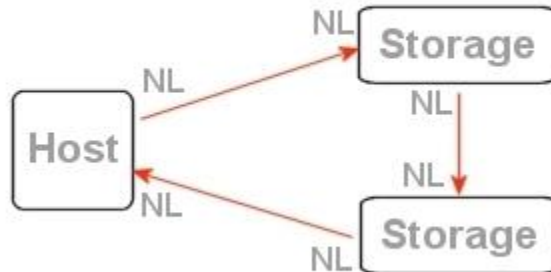
SCSI (อ่านว่า scuzzy), FC for high-end workstations

1 SCSI initiator + 15 SCSI targets per bus

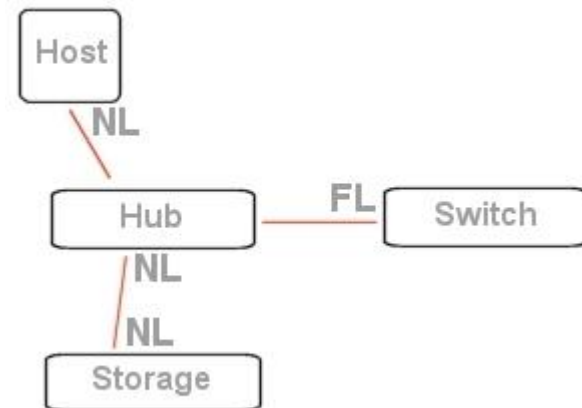
FC (Fiber channel)

FC-AL (arbitrated loop) 126 devices and controllers.

Arbitrated Loop



Public Loop with Hub



2. Network-Attached Storage (**NAS**) uu LAN

3. Storage-Area Network (**SAN**) uu private network ใยแก้วนำแสง

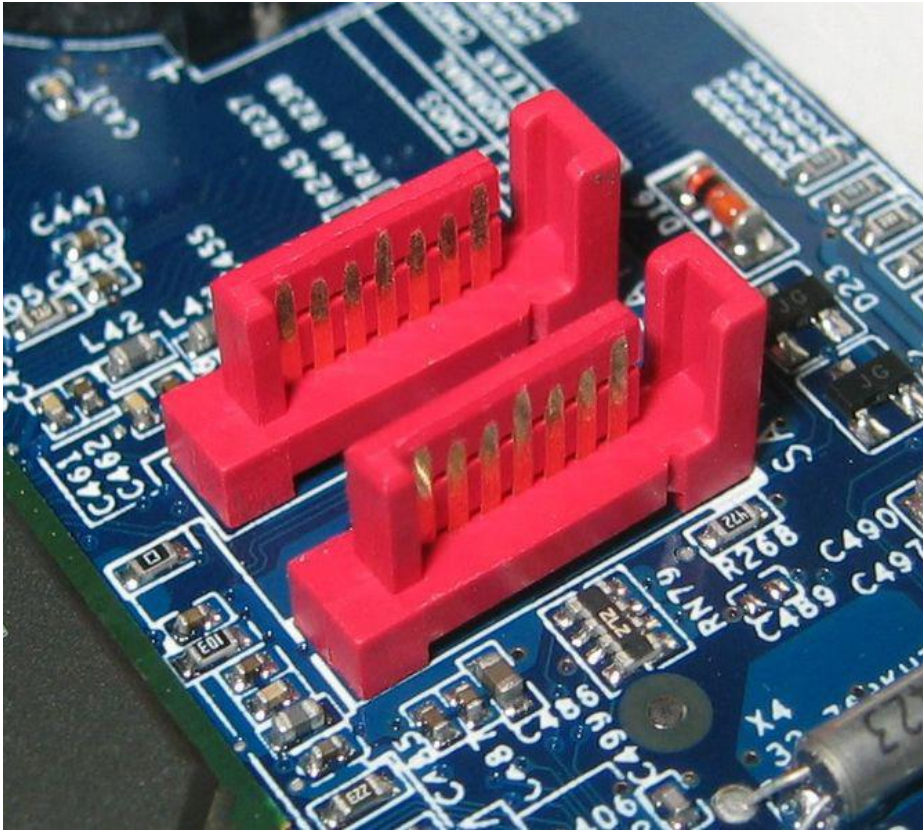
Fiber Channel ได้ความเร็วสูงสุด
16 Gb/s หรือ 2 GB/s



HP 416728-001 300GB Fiber Channel hard disk drive – 15,000 RPM,
with fiber channel arbitrated loop (FC-AL-2) protocol

Can be installed into the HP M5314A FC 14 bay disk drive enclosure

SATA



SATA is a point-to-point serial system.
จะต่อพ่วงดิสก์หลาย ๆ ตัว บนสายเส้นเดียวไม่ได้

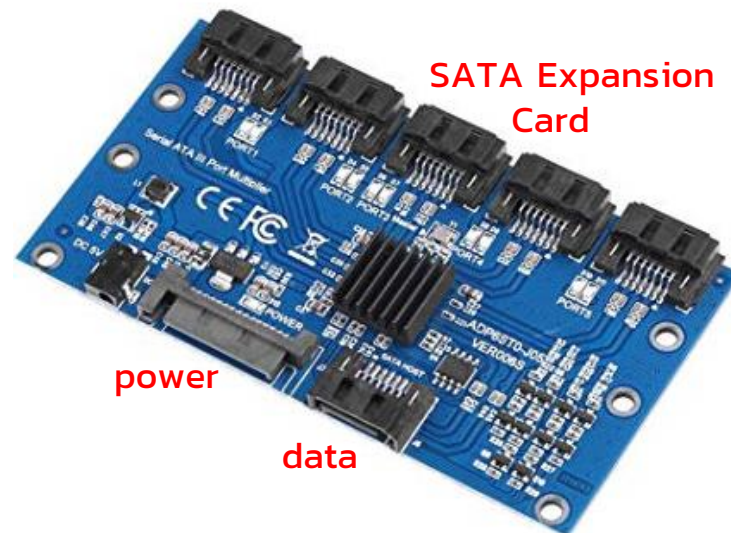




1 port พ่วงดิสก์หลายตัวไม่ได้ (ตรงข้ามกับ SCSI)
แต่ใช้ hub หรือ expansion card ได้



SATA Hub



SATA Expansion
Card

power

data

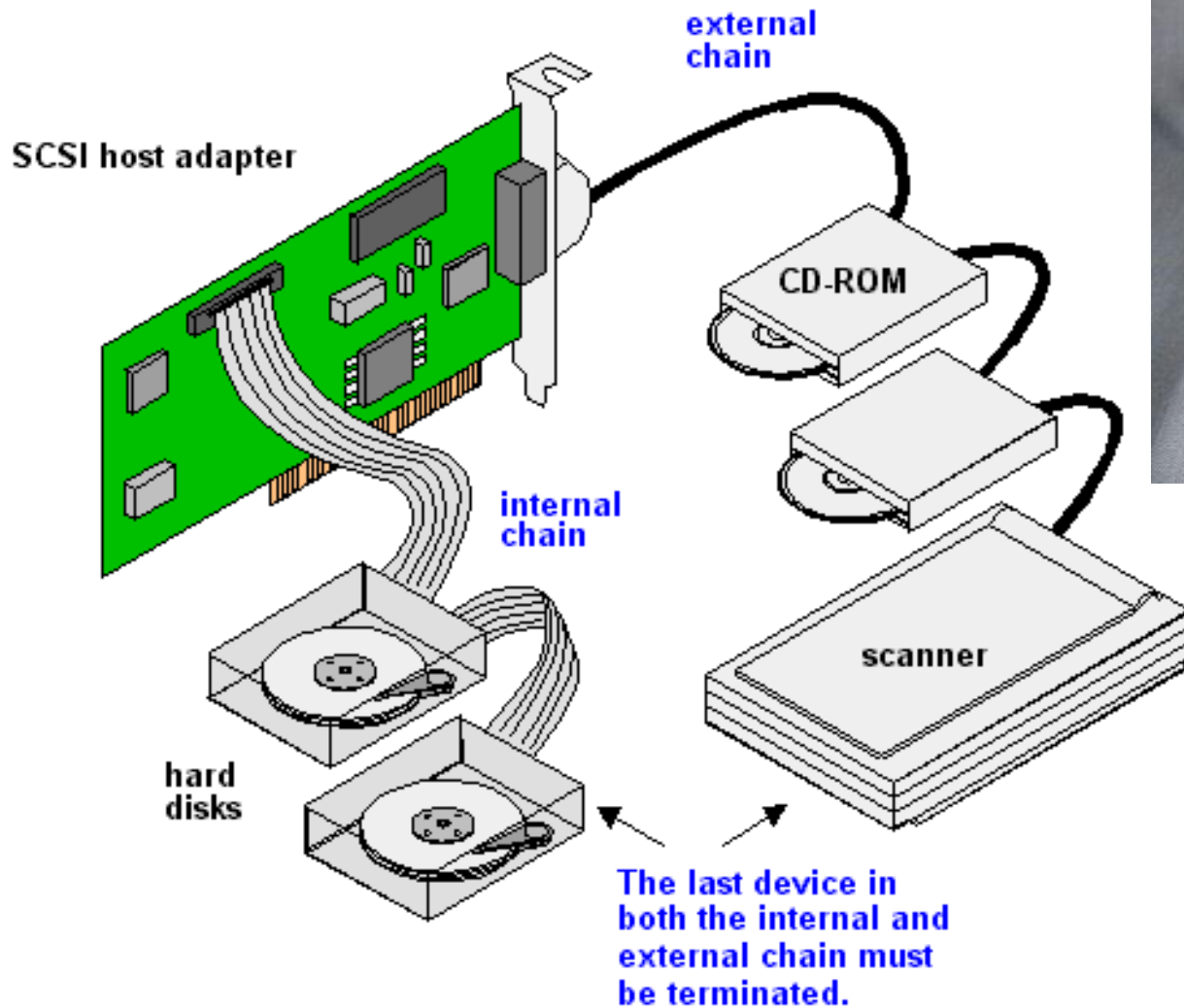
Comparison	SATA I	SATA II	SATA III
1. Bandwidth	1.5 Gb/s	3.0 Gb/s	6.0 Gb/s
2. Actual Speed	150 Mb/s	300 Mb/s	600 Mb/s
3. Backward Compatibility	-	Yes	Yes
4. Forward Compatibility	Yes	Yes	Yes (SATA 3.1 and 3.2)
5. NCQ	No	Yes	Yes
6. Best Use case	Optical drives like CD and DVD	Mechanical Hard drives	SATA SSDs
7. Released	2003	2004	2009

** Note that a gigabit (Gb) is not the same as a gigabyte (GB). 1 GB = 8 Gb.*

SCSI

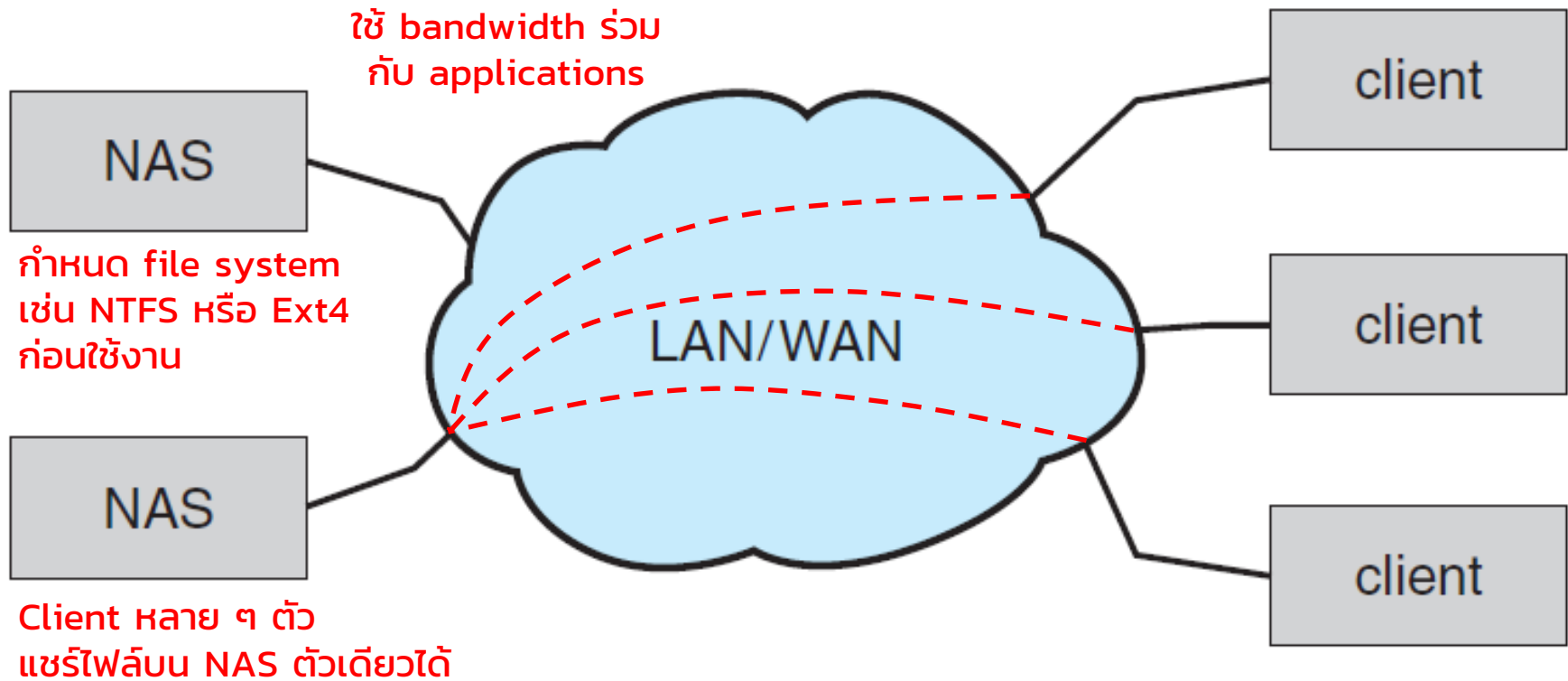
มีทั้งแบบ onboard หรือ PCI card

From Computer Desktop Encyclopedia
© 1998 The Computer Language Co. Inc.



Current SCSI Standards

Technology name	Maximum cable length (meters)	Maximum speed (Mbps)	Maximum number of devices
SCSI-I	6 Single Ended	5	8
Fast SCSI	3 Single Ended	10	8
Fast Wide SCSI	3 Single Ended	20	16
Ultra SCSI	1.5 Single Ended	20	8
Ultra SCSI	3 Single Ended	20	4
Wide Ultra SCSI	-	40	16
Wide Ultra SCSI	1.5 Single Ended	40	8
Wide Ultra SCSI	3 Single Ended	40	4
Ultra2 SCSI	12 LVD/25 HVD	40	8
Wide Ultra2 SCSI	12 LVD/25 HVD	80	16
Ultra3 SCSI (Ultra160 SCSI)	12 LVD	160	16
Ultra320 SCSI	12 LVD	320	16
Ultra640 SCSI	10 LVD	640	16

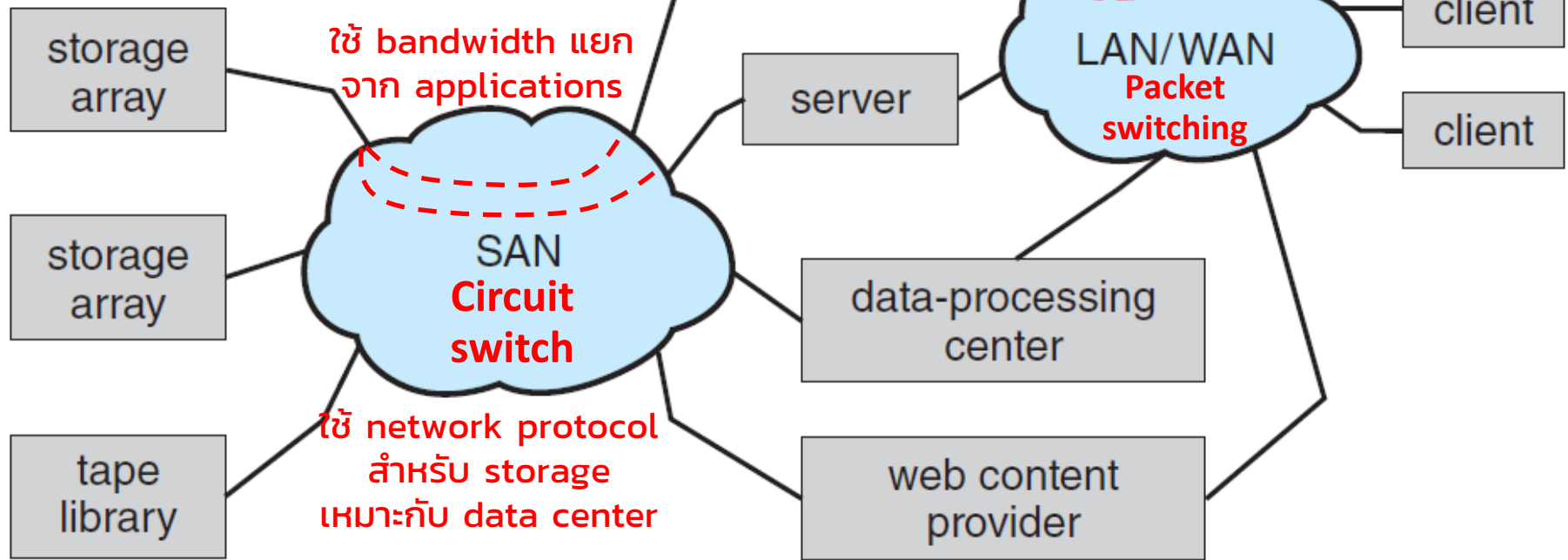


เหมาะกับการใช้งานตามบ้านหรือองค์กรขนาดเล็ก

สาย LAN



ให้ raw blocks กับ server
server ไปกำหนด file system เอง



NAS กับ SAN ต่างกันยังไง ? NAS ให้บริการ file-system space แต่ SAN ให้บริการ free blocks

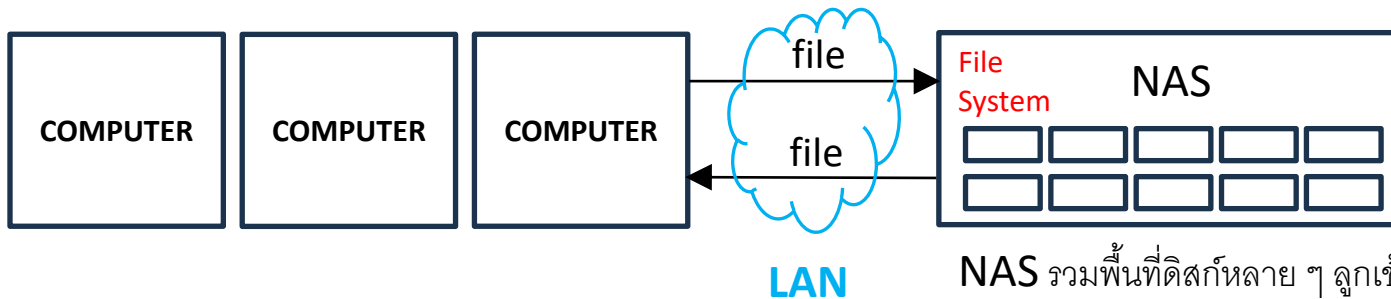
<https://www.youtube.com/watch?v=3yZDDrOJKVc>

Thick vs. Thin Provisioning

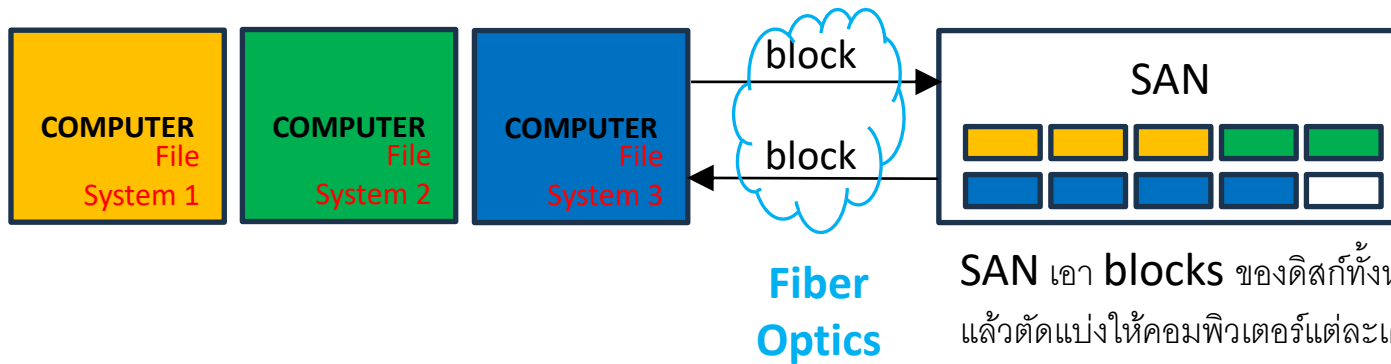
<https://www.youtube.com/watch?v=Ks6XgT69ZLc>



File system (FS) = data structure ที่ใช้ในการเก็บ files และ block

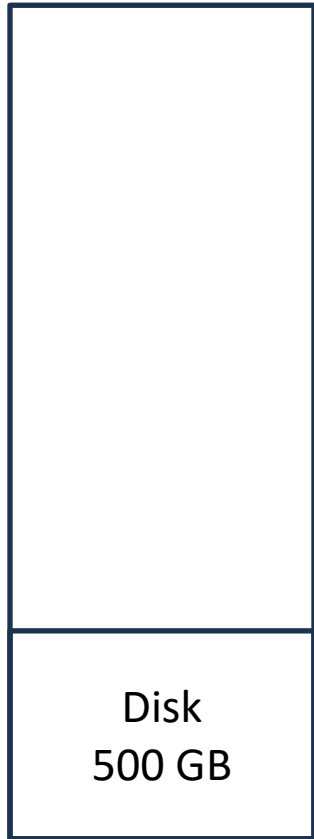


NAS รวมพื้นที่ดิสก์หลาย ๆ ลูกเข้าด้วยกัน
NAS OS เป็นคนสร้าง file system อันเดียว
แชร์ให้กับคอมพิวเตอร์หลาย ๆ เครื่อง
(เลือก FS จาก NAS OS)



SAN เอา blocks ของดิสก์ทั้งหมดมาเทรวมกัน
แล้วตัดแบ่งให้คอมพิวเตอร์แต่ละเครื่อง (ไม่แชร์กัน)
COMPUTER OS เป็นคนสร้าง file system
(เลือก FS จาก COMPUTER OS)

THICK



THIN



นิยมใช้กับ SAN และ VM

ผู้ให้บริการ cloud นิยมทำ thin-provisioning

แต่ถ้าเราเป็นลูกค้าควรจะเห็นสัญญาแบบ thick-provisioning

SAN จะให้พื้นที่ดิสก์เท่าที่ใช้เก็บข้อมูลจริง ทำให้มีพื้นที่เหลือไปสร้าง
VM เพิ่มได้ คิดค่าใช้จ่ายเต็มจำนวนพื้นที่ แต่ถ้าปริมาณข้อมูลเพิ่มขึ้น
ต้องหาดิสก์มาเพิ่มให้ทัน ต้องมีสถิติการใช้งานของลูกค้า

Thick ได้ vm 4 ตัว Thin ได้ vm 200 ตัว

Disk Scheduling

1. FCFS scheduling
- ~~2. SSTF scheduling~~
3. SCAN scheduling
4. C-SCAN scheduling
- ~~5. LOOK/C-LOOK scheduling~~

Note:

FCFS = First Come First Serve

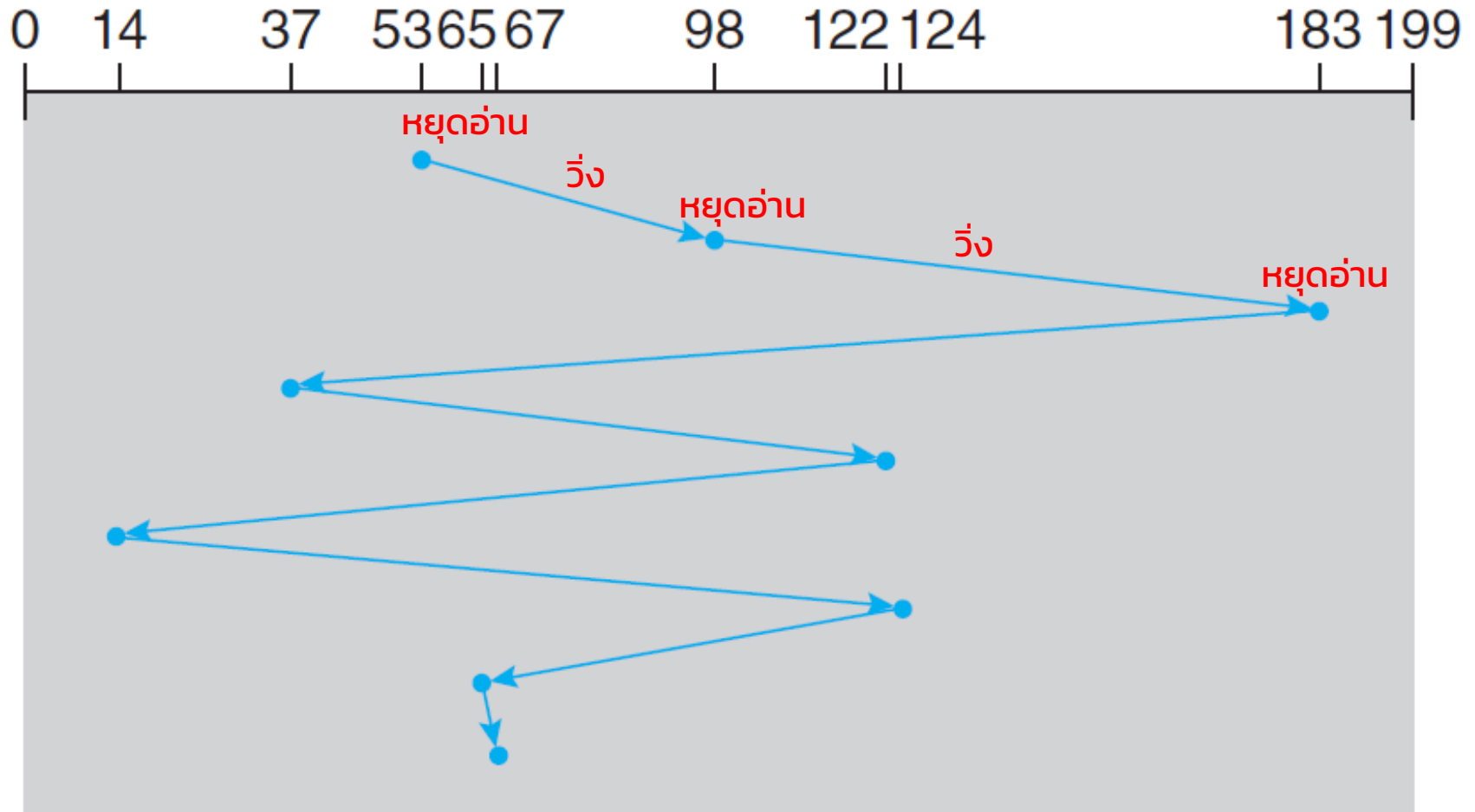
SSTF = Shortest Seek Time First

C = Circular

FCFS

queue = 98, 183, 37, 122, 14, 124, 65, 67

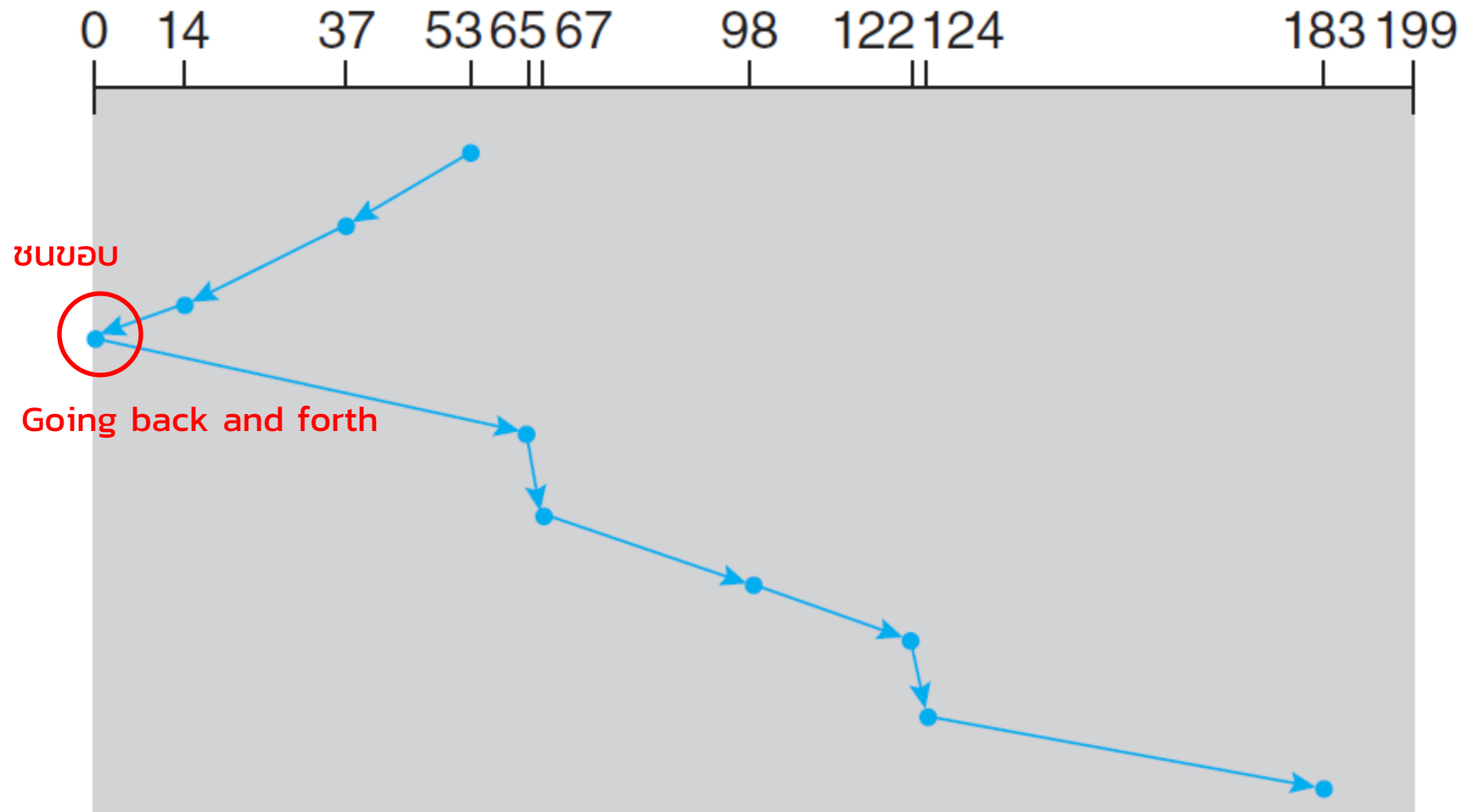
head starts at 53 (cylinder ดูรูปในหน้าแรก)



SCAN

queue = 98, 183, 37, 122, 14, 124, 65, 67

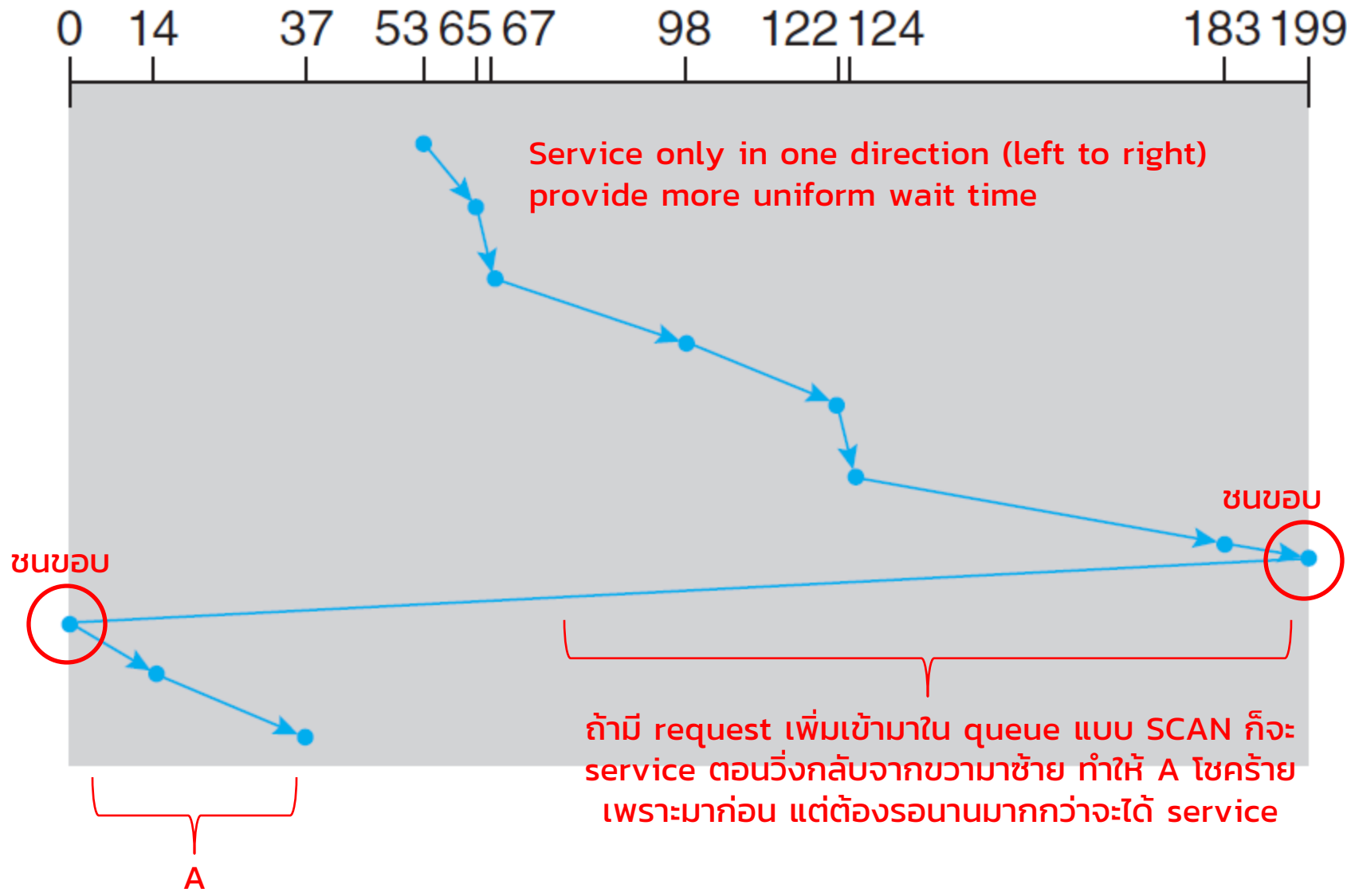
head starts at 53



C-SCAN คือไม่ต้องวิ่งไปชนขอบ คล้าย ๆ กับลิฟต์

C-SCAN

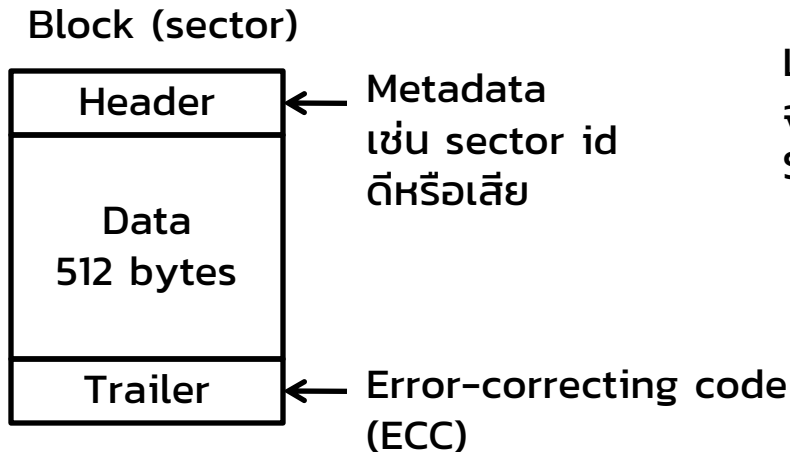
queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53



Disk Management

1. Disk Formatting

- Physical formatting เช่น กำหนด block size เป็น 512 หรือ 1,024 bytes
- Logical formatting เช่น กำหนด partition เป็น drive C: และ D:



Logical block 0, 1, 2, 3, 4, ... (ที่ OS เห็น)
จะถูก map กับ defect-free sector เช่น
Sector 475, 476, 980, 981, 982, ...

Soft Error:

สามารถ recover ได้จาก ECC

Hard Error:

ไม่สามารถ recover ได้จาก ECC
(bad sector)

Read:

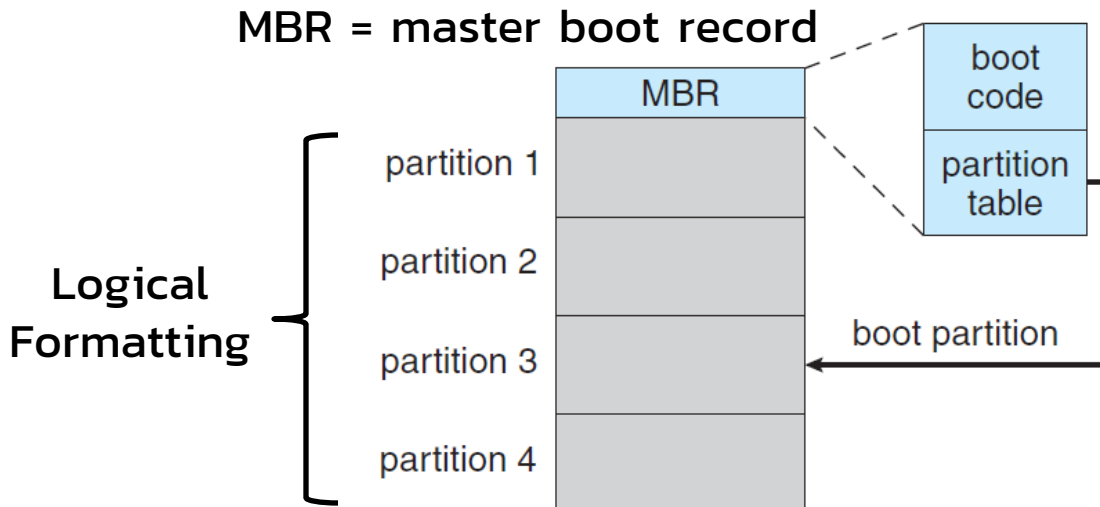
disk controller คำนวณ code ใหม่ และ
เช็คว่าตรงกับ code ใน trailer หรือไม่ ถ้าไม่
อาจจะลองพยายามอ่านใหม่อีกครั้ง ถ้าเกิน n
ครั้งไม่สำเร็จจะ error

Write:

disk controller คำนวณ code ใหม่ และ
เขียนทับลงใน trailer

Disk Management

2. Boot Block



Bootting steps:

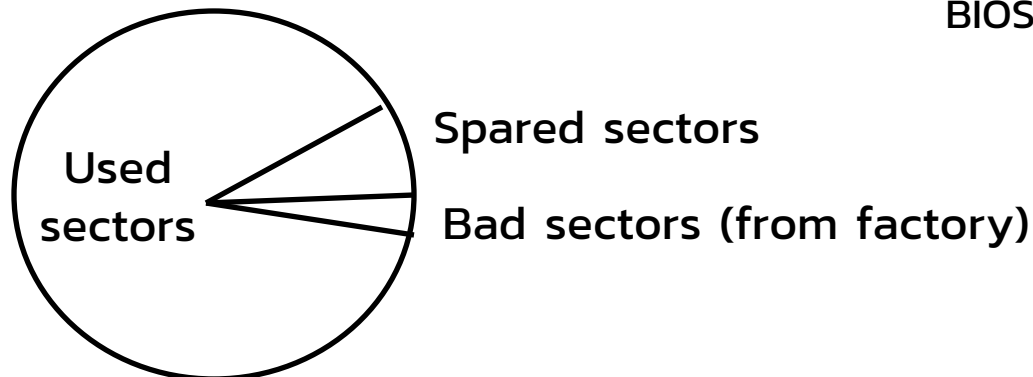
1. โหลด first program (BIOS) จาก ROM
2. First program จะไปอ่าน boot code จากใน MBR
3. Boot code จะอ่าน partition table ได้และต้องรู้ว่าไปโหลดโปรแกรมต่อไปมาจาก partition ใด
4. โหลด OS kernel จาก boot partition

MBR พัฒนาไปเป็น GPT

GPT = GUID Partition Table

BIOS พัฒนาไปเป็น UEFI

3. Bad Blocks

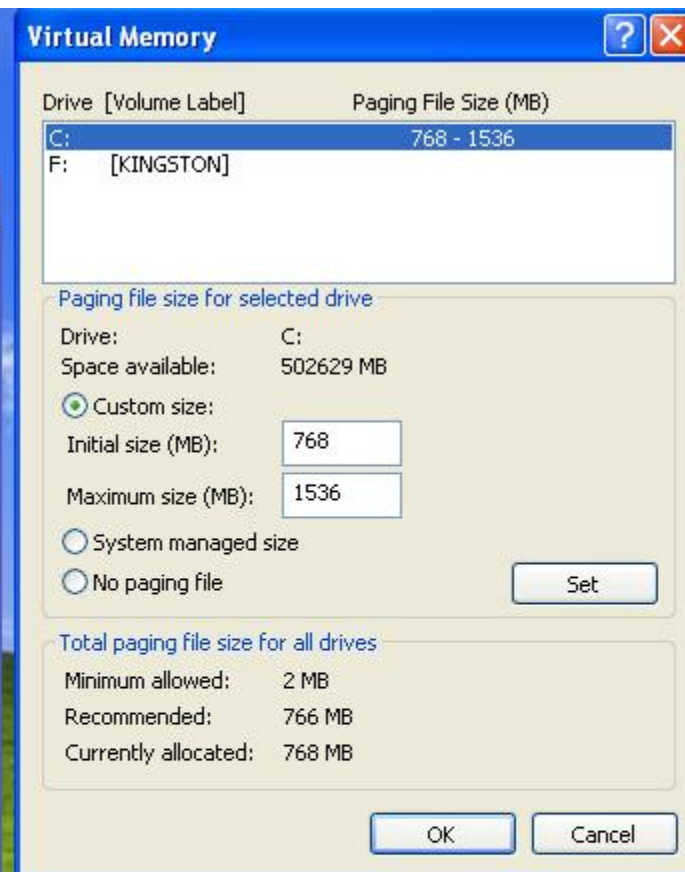
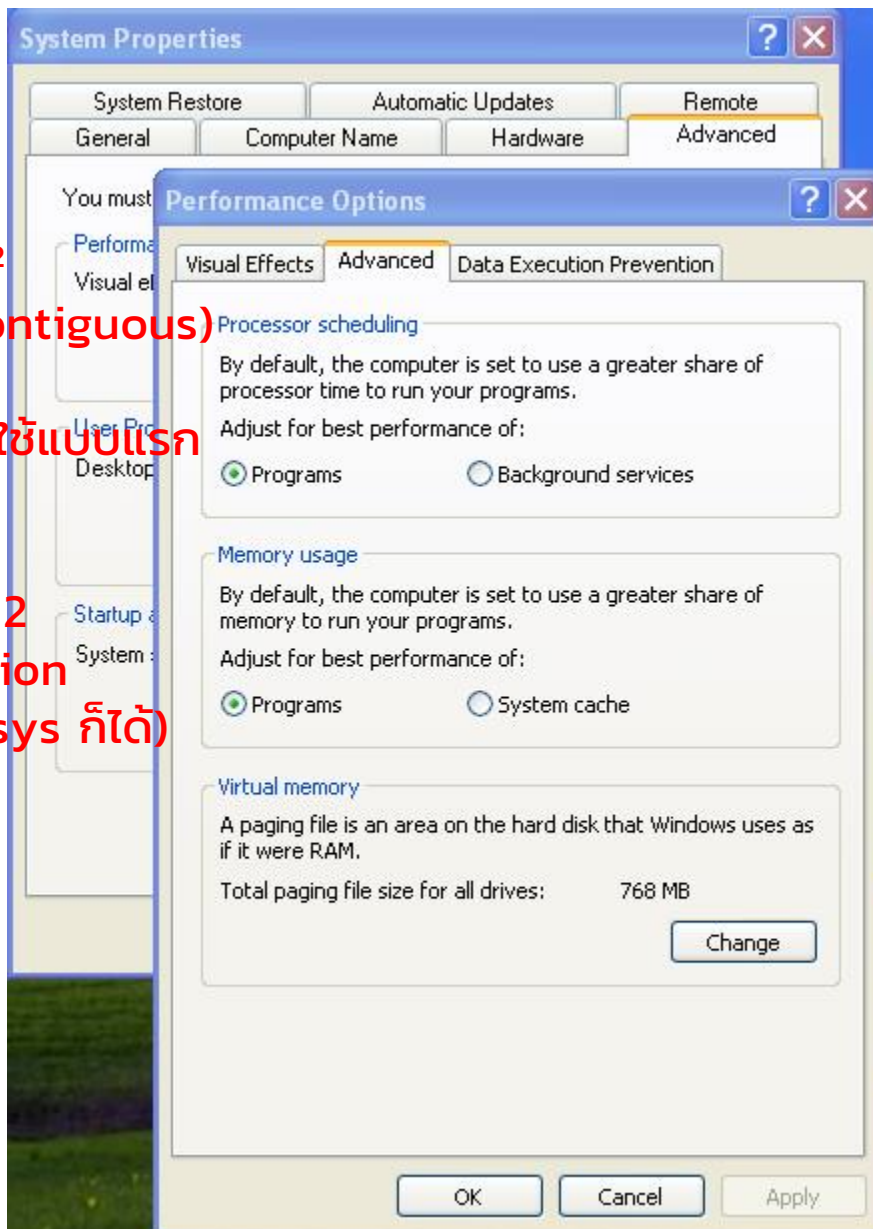


Swap-Space Management (swapping = paging)

อยู่บน
file system¹
(แซร์, ขยายได้)
หรือ
raw partition²
(perf ดีกว่า, contiguous)

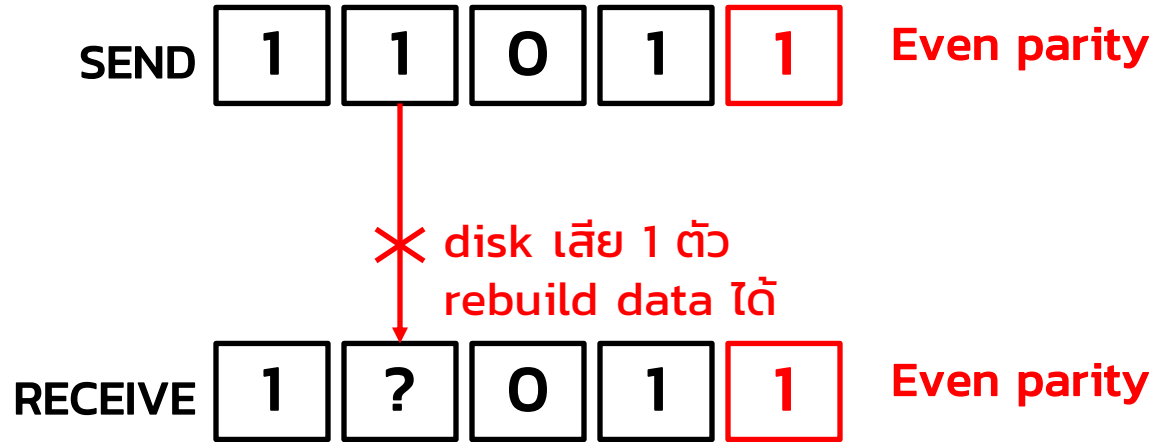
Windows XP ใช้แบบแรก
มี page file

Linux ใช้ แบบ 2
มี swap partition
(หรือจะใช้ file sys ก็ได้)



RAID (Redundant Array of Inexpensive Disks)

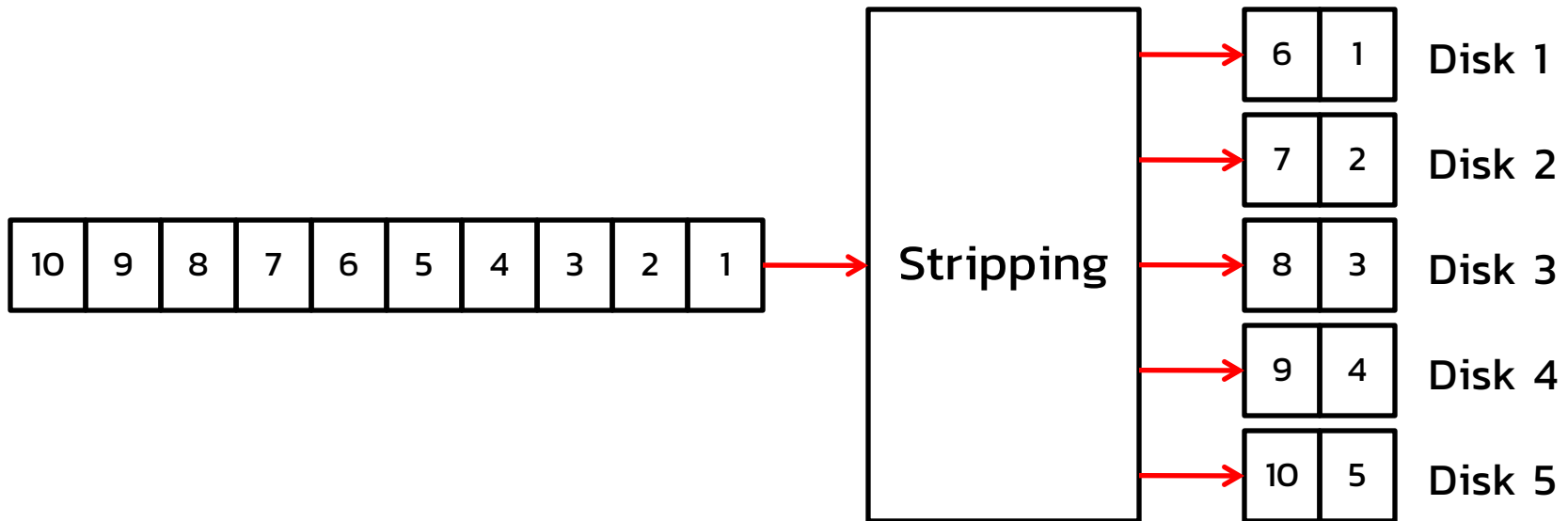
Error detection (parity bit)



Error detected, need retransmission (ผิดได้ไม่เกิน 1 บิต หรือจำนวนคี่)

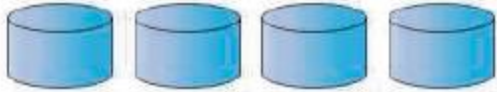
RAID (Redundant Array of Inexpensive Disks)

Data striping (bit/byte/block)



ไม่มี redundant
แต่ทำให้อ่านเขียนข้อมูลได้เร็วขึ้น

RAID0

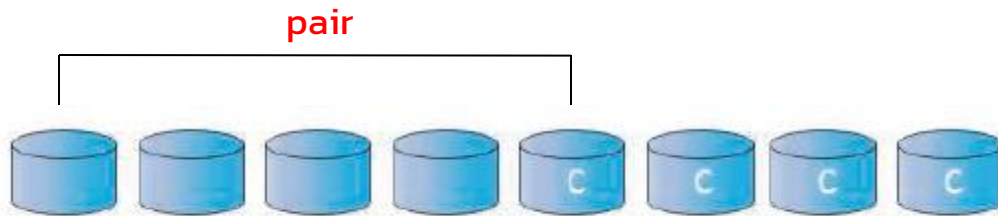


(a) RAID 0: non-redundant striping.

Striping

ทำให้อ่านเขียนได้เร็วขึ้น

RAID1

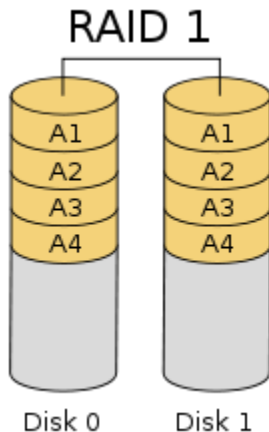


(b) RAID 1: mirrored disks.

Mirroring

มีข้อมูลซ้ำกัน 2 ชุด อ่านได้เร็วขึ้น

ป้องกันการเสียหายของ disk ตัวใดตัวหนึ่ง



ป้องกัน disk ตัวใดตัวหนึ่งเสีย

แต่ถ้าเสียพร้อมกันทั้งสองตัวก็ช่วยไม่ได้

RAID5



(d) RAID 5: block-interleaved distributed parity.

Distribute parity blocks

Reduce the workload of P-disk

เพราะทุกครั้งต้องเขียน P-disk ไม่งั้น P-disk จะเสียก่อน

RAID6



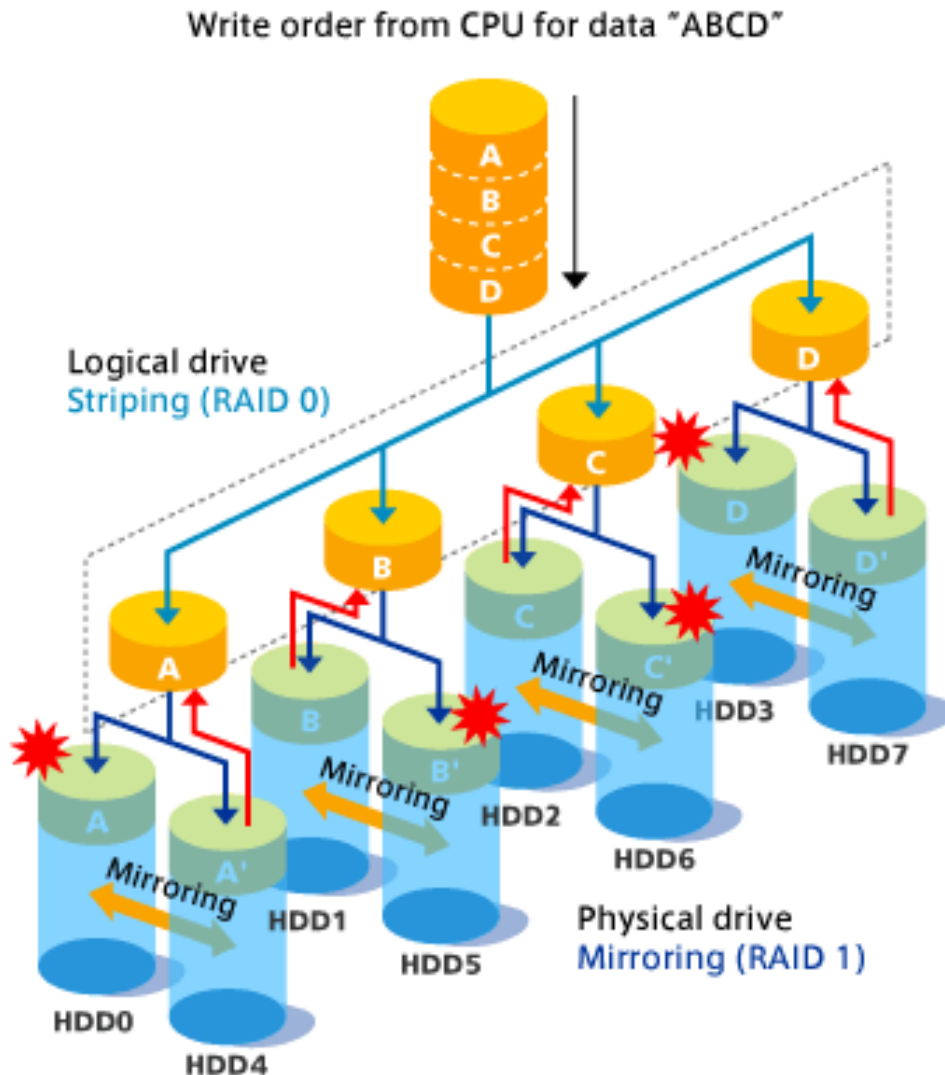
(e) RAID 6: P + Q redundancy.

P + Q redundancy scheme

Allow multiple disk failures

RAID controller ใด ๆ จะทำ HOT SWAP และ HOT SPARE ได้

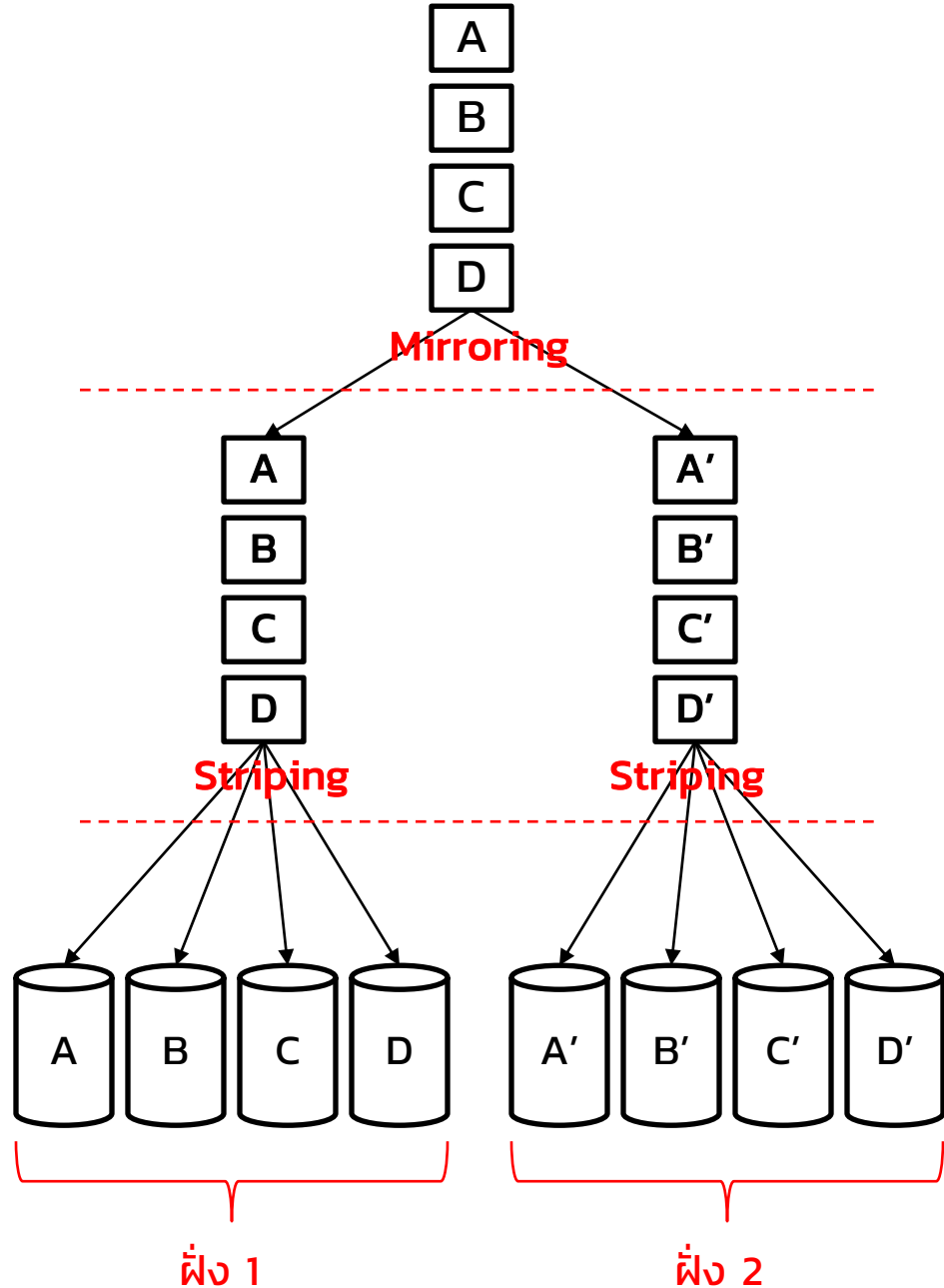
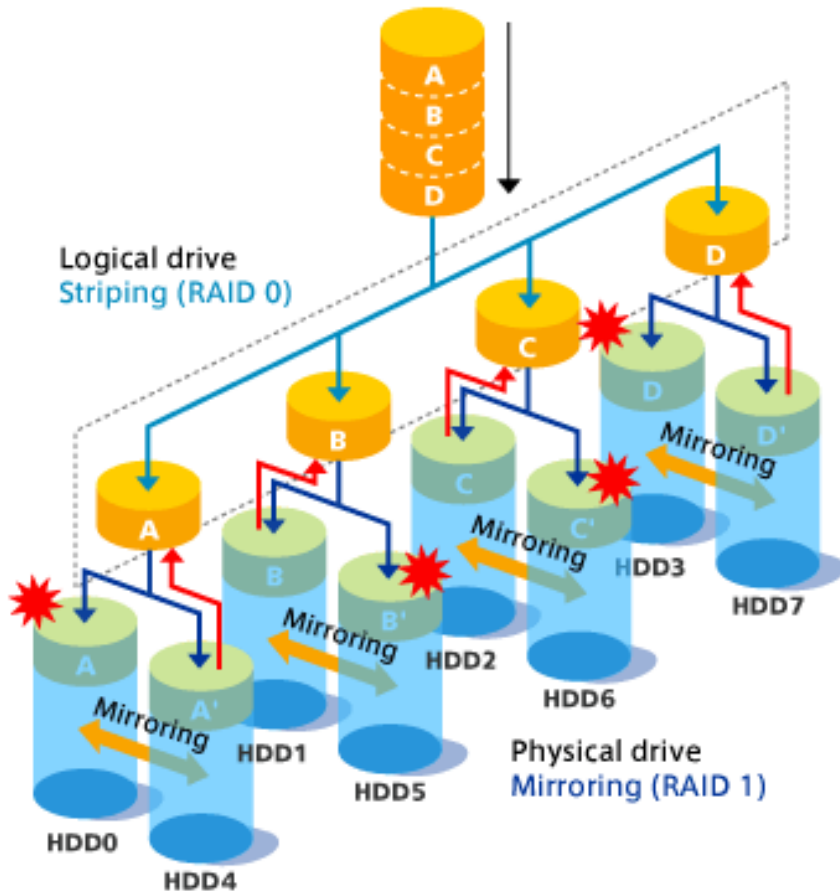
RAID 1+0 (more tolerant than RAID 0+1)



ทน disk เสียได้มากที่สุด 4 ตัว
และรับได้หลาย combination ด้วย
ขออย่าให้เป็น disk ตัวที่คู่กัน

RAID 1+0 vs. 0+1

Write order from CPU for data "ABCD"



ทน disk เสียได้มากที่สุด 4 ตัว และต้องอยู่พียงเดียวกันหมด
ถ้ามี disk เสียทั้งสองพียง แม้เพียงพียงละตัว ก็จะไม่ทำงานต่อไม่ได้

กำหนดให้มี disk ทั้งหมด 8 ตัว และความน่าจะเป็นที่ disk ตัวหนึ่งจะเสียหลังจากใช้งานไประยะเวลาหนึ่งเท่ากับ 0.001 จงแสดงการคำนวณความน่าจะเป็นที่ RAID 1+0 จะเสียจนทำงานต่อไม่ได้ และแสดงการคำนวณของ RAID 0+1 ด้วย

RAID 1+0	HD1 HD2		Prob <u>ที่ disk 1 คู่ จะเสียคือ</u> 0.001^2 ต้องเสียทั้ง 2 ตัว
	HD3 HD4		Prob <u>ที่ disk 1 คู่ จะใช้งานได้คือ</u> $1 - 0.001^2$ มีตัวใดตัวหนึ่งใช้ได้
	HD5 HD6		Prob <u>ที่ disk ทั้ง 4 คู่จะใช้งานได้คือ</u> $(1 - 0.001^2)^4$
	HD7 HD8		RAID1+0 ใช้งานไม่ได้ (มี disk อย่างน้อย 1 คู่ที่ mirror กันเสีย) Prob <u>ที่ disk ทั้ง 4 คู่จะเสียคือ</u> $1 - (1 - 0.001^2)^4 = 0.000004$
RAID 0+1	HD1	HD5	Prob <u>ที่ฝั่งหนึ่งจะทำงานได้เท่ากับ</u> $(1 - 0.001)^4$ ต้องดีทั้ง 4 ตัว
	HD2	HD6	Prob <u>ที่ฝั่งหนึ่งจะเสียเท่ากับ</u> $1 - (1 - 0.001)^4$ มีตัวใดตัวหนึ่งเสีย
	HD3	HD7	Probe <u>ที่จะเสียทั้งสองฝั่ง</u> $(1 - (1 - 0.001)^4)^2 = 0.000016$
	HD4	HD8	(มีโอกาสเสียมากกว่า RAID 1+0 ถึง 4 เท่า)

Array Devices

Device	Identification	Temp.	Reads	Writes	Errors	FS	Size	Used	Free
▼ Parity	ST8000VN0022-2EL112_ZA15Q627 - 8 TB (sdq)	90 F	45	124	0				
▼ Parity 2	ST8000VN0022-2EL112_ZA15RG65 - 8 TB (sdd)	97 F	45	124	0				
▼ Disk 1	ST3000DM001-1CH166_W1F3X376 - 3 TB (sdo)	81 F	6487	14	0	xfs	3 TB	2.82 TB	184 GB
▼ Disk 2	TOSHIBA_DT01ACA300_Y44TLOS6S - 3 TB (sdn)	84 F	1926	12	0	xfs	3 TB	2.74 TB	260 GB
▼ Disk 3	TOSHIBA_DT01ACA300_Y44TLORG6S - 3 TB (sdp)	86 F	601	12	0	xfs	3 TB	2.63 TB	373 GB
▼ Disk 4	ST3000DM001-1CH166_W1F3WXV0 - 3 TB (sdm)	82 F	245	5	0	xfs	3 TB	2.68 TB	317 GB
▼ Disk 5	TOSHIBA_DT01ACA300_54G0E13GS - 3 TB (sdl)	88 F	424	10	0	xfs	3 TB	2.23 TB	768 GB
▼ Disk 6	WDC_WD30EZR6-00D8PB0_WD-WCC4N1229543 - 3 TB (sdk)	84 F	334	11	0	xfs	3 TB	2.63 TB	370 GB
▼ Disk 7	ST8000AS0002-1NA17Z_Z840MM2Q - 8 TB (sdj)	88 F	30,962	11	0	xfs	8 TB	4.53 TB	3.47 TB
▼ Disk 8	TOSHIBA_DT01ACA300_45AS963GS - 3 TB (sdi)	86 F	437	11	0	xfs	3 TB	2.60 TB	400 GB
▼ Disk 9	TOSHIBA_DT01ACA300_Y37AV61GS - 3 TB (sdh)	90 F	340	11	0	xfs	3 TB	2.62 TB	375 GB
▼ Disk 10	TOSHIBA_DT01ACA300_54G0MNTGS - 3 TB (sdf)	88 F	291	5	0	xfs	3 TB	2.50 TB	498 GB
▼ Disk 11	ST8000AS0002-1NA17Z_Z840PXVB - 8 TB (sdg)	93 F	430	12	0	xfs	8 TB	4.01 TB	3.99 TB
▼ Disk 12	ST8000AS0002-1NA17Z_Z840PRJW - 8 TB (sdb)	90 F	255	5	0	xfs	8 TB	6.93 TB	1.07 TB
▼ Disk 13	ST8000AS0002-1NA17Z_Z840PJ9W - 8 TB (sde)	93 F	253	5	0	xfs	8 TB	5.93 TB	2.07 TB
Total	Array of fifteen devices	88 F	43,075	372	0		59 TB	44.8 TB	14.2 TB

Cache Devices

Linux OS (NAS) ใช้ disk ขนาดใดพอสมควรก็ได้ ขยายความจุได้ง่าย ซื้อ disk มาเพิ่ม

Device	Identification	Temp.	Reads	Writes	Errors	FS	Size	Used	Free
▼ Cache	Samsung_SSD_850_EVO_500GB_S21HNSAG142765B - 500 GB (sdr)	81 F	21,163	10,447	0	btrfs	500 GB	206 GB	294 GB

Array Started

Chapter 11 Mass-Storage Structure

11.1 Overview of Mass-Storage

Structure 449

11.2 HDD Scheduling 457

11.3 NVM Scheduling 461

11.4 Error Detection and Correction 462

11.5 Storage Device Management 463

11.6 Swap-Space Management 467

11.7 Storage Attachment 469

11.8 RAID Structure 473

11.9 Summary 485

Practice Exercises 486

Further Reading 487

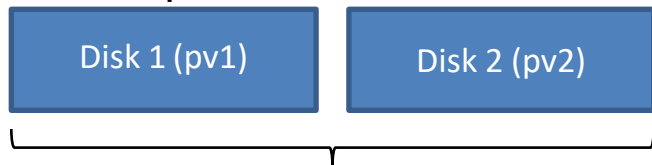
Fiber Switch

Server

SAN

Controller 2 ตัว
ทำงานแทนกันได้

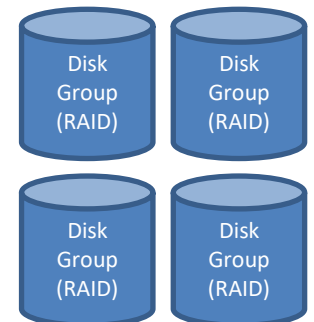
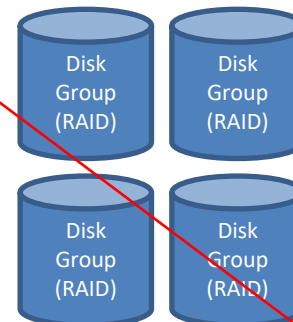
OS จะเห็น pool A, B เป็น disk 1, 2
ถ้าไม่ใช่ multipath จะเห็น disk 1, 2 อย่างละ 4 ตัว



ให้ disk1 และ disk2 เป็น physical volume (pv) ชื่อ pv1 และ pv2
pv ก็คือ storage blocks เรากำหนดให้ disk ตัวใดเป็น pv ก็ได้
สร้าง volume group (vg) ชื่อ san-vg = pv1 + pv2
ตัดแบ่ง san-vg ให้เป็น logical volume (lv) หลาย ๆ ตัวได้
OS/VM จะมองเห็น lv เป็น disk สามารถ partition และ format ได้

Controller A

Controller B



Pool A

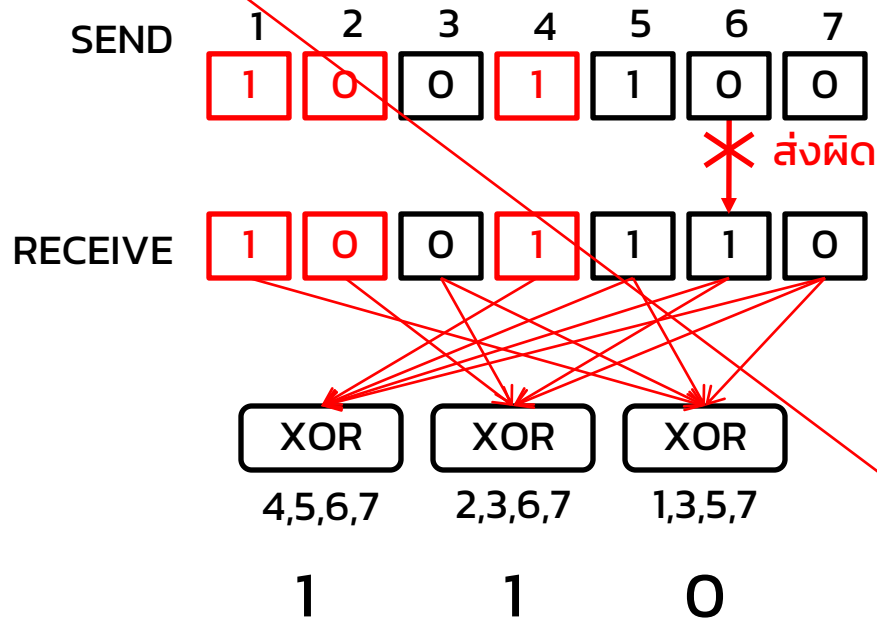
Pool B

Pool เป็น unit ที่ย้ายได้
ระหว่าง SAN

ถ้าไม่ใช่ SAN จะใช้ Disk1 และ Disk2 ที่เป็น hard disks ก็ได้ แต่จะไม่ทนต่อความเสียหาย ได้ data rate น้อยกว่า SAN และไม่มี multipath

RAID (Redundant Array of Inexpensive Disks)

Error correction (Hamming code)



บิตที่ 110 = 6 ผิด
ถ้าได้ 000 คือไม่มีที่ผิด

ใส่ redundant 3 บิต ต่อข้อมูล 4 บิต

Information	1 p_0	2 p_1	3 u_3	4 p_2	5 u_2	6 u_1	7 u_0
0	0	0	0	0	0	0	0
1	1	1	0	1	0	0	1
2	0	1	0	1	0	1	0
3	1	0	0	0	0	1	1
4	1	0	0	1	1	0	0
5	0	1	0	0	1	0	1
6	1	1	0	0	1	1	0
7	0	0	0	1	1	1	1
8	1	1	1	0	0	0	0
9	0	0	1	1	0	0	1
10	1	0	1	1	0	1	0
11	0	1	1	0	0	1	1
12	0	1	1	1	1	0	0
13	1	0	1	0	1	0	1
14	0	0	1	0	1	1	0
15	1	1	1	1	1	1	1