9. Storage Architecture

Special Topics in Computer Systems:

Modern Storage Systems
(IC820-01)

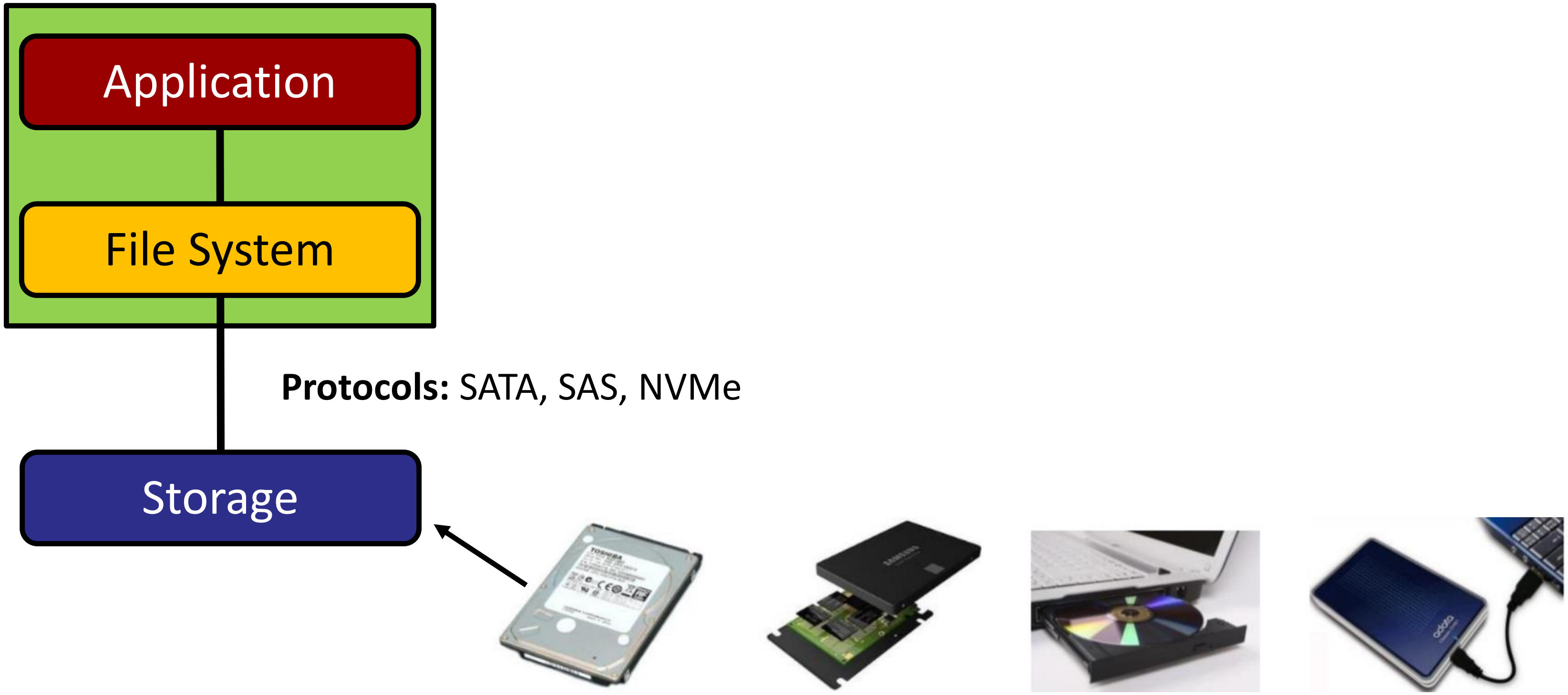
Instructor:

Prof. Sungjin Lee (sungjin.lee@dgist.ac.kr)

DAS, NAS, and SAN

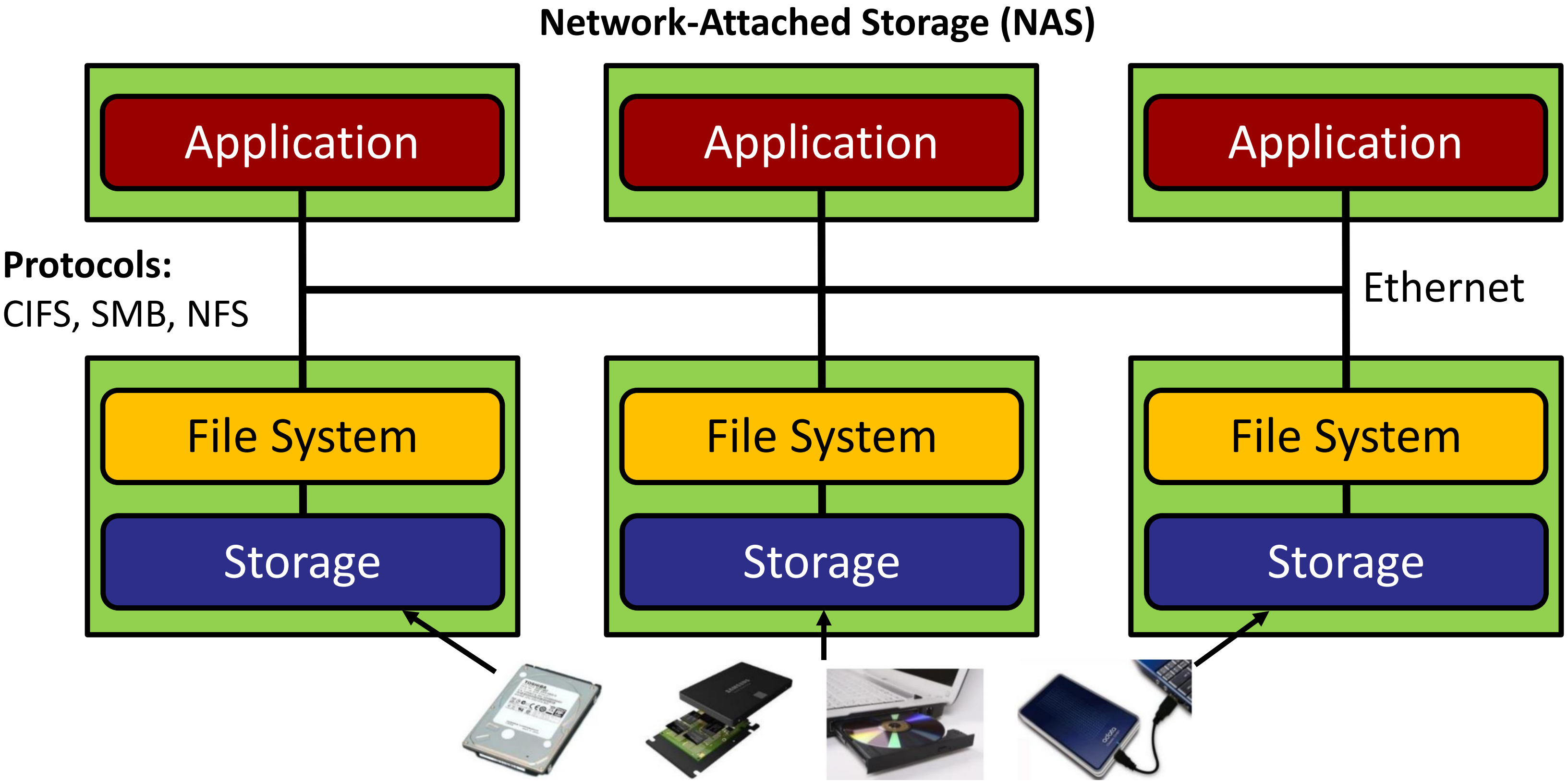
- **Directly-Attached Storage (DAS)**
  - Direct-attached storage device
  - Generally attached/dedicated to a specific server

Directly-Attached Storage



DAS, NAS, and SAN (Cont.)

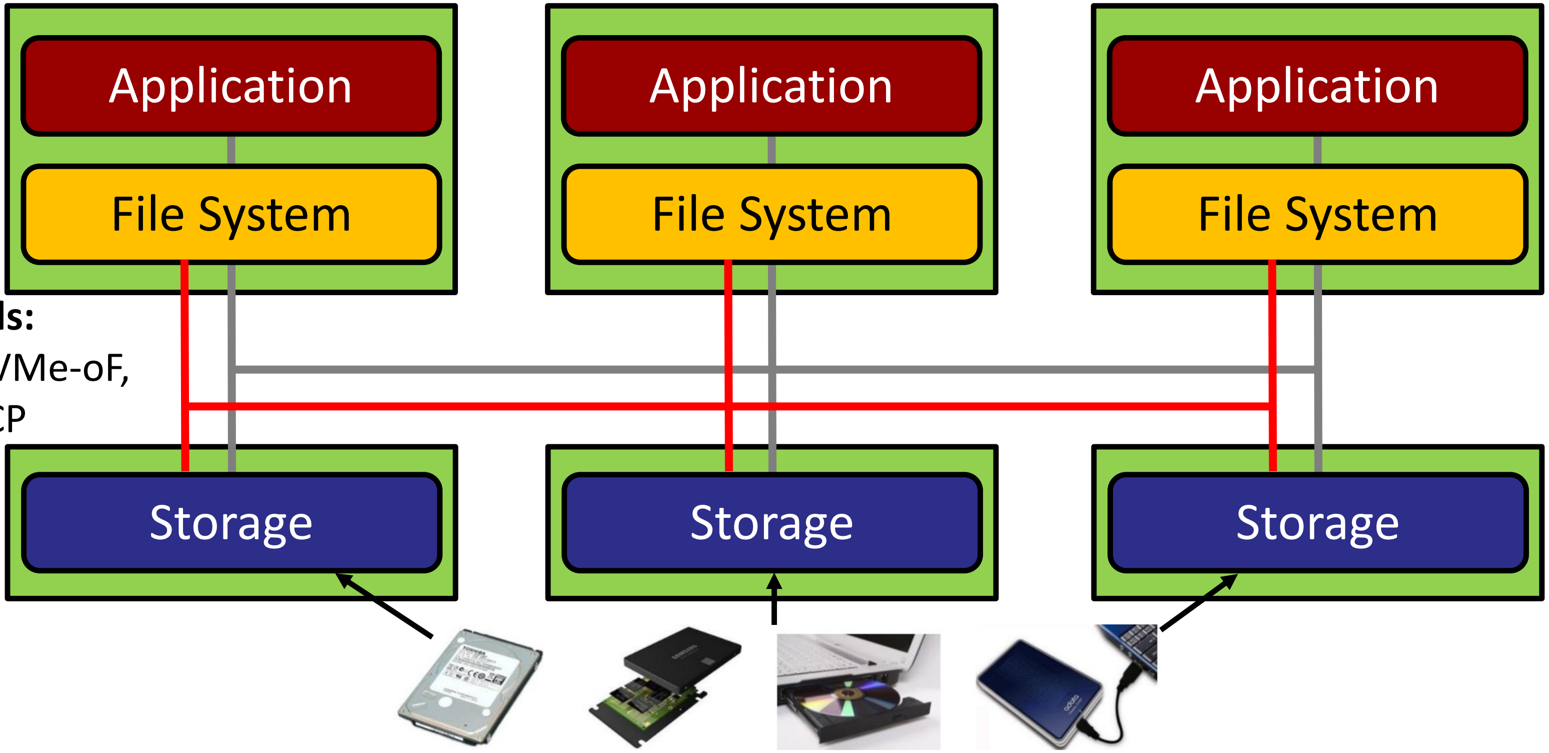
- **Network-Attached Storage (NAS)**
  - Connected to a server via a network
  - Can be shared or dedicated



DAS, NAS, and SAN (Cont.)

- **Storage-Area Network (SAN)**
  - Connected to a server via a storage network
  - Can be shared or dedicated

Storage-Area Network (SAN)



Protocols:
iSCSI, NVMe-oF,
FCoE, FCP

Outline

- Directly-Attached Storage (DAS)
- Network-Attached Storage (NAS)
- Storage-Area Network (SAN)

Storage Interface

- **Hard drives and SSDs use four major interfaces to communicate with the host system**
  - PATA: Parallel Advanced Technology Attachment
  - SATA: Serial Advanced Technology Attachment
  - SAS: Serial-Attached SCSI
  - NVMe: NVMe over peripheral component interconnect express
  - DIMM: The memory channel

Parallel ATA

16 - bit는 많으나 frequency가 적다.

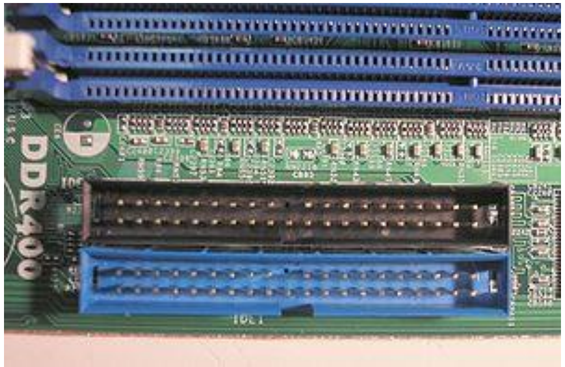
- **A direct connection to the 16-bit ISA bus introduced with the IBM PC/AT**
  - Use the Integrated Drive Electronics (IDE) protocol
  - With a 16-bit bus, two bytes are transmitted per bus transaction
  - Double-edge clocking mechanism for DMA transfers

can't increase frequency

100 MB/s =

25 MHz strobe x **2** (double data rate clocking) x **16 bits** per edge / 8 bits per byte

- **Provide up to the maximum throughput of 133 MB/s**
  - No further development



Motherboard sockets

IDE Cable

Serial ATA

- **The proactive evolution of the ATA interface from a parallel bus to a serial bus architecture**
  - Overcome the electrical constraints that are increasing the difficulty of continued speed enhancements of the parallel communication
  - Use either the IDE or Intel's Advanced Host Controller Interface (AHCI) protocol

150 MB/s =
1500 MHz clock x **1 bit** per clock x 8b/10b encoding / 8 bits per byte

빠른 HDD까지 충분하다.

- **SATA 3.0 provides up to 600 MB/s throughput**

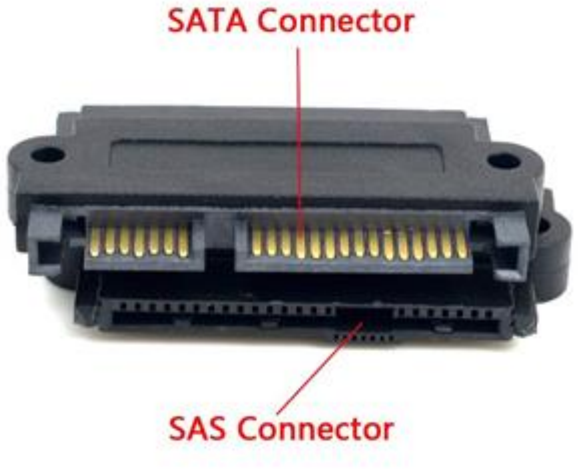
SATA Cable and Connector

Serial-Attached SCSI

- **Based on Small Computing System Interface (SCSI) by a floppy disk maker**
  - Improved to support a parallel bus later
- **Higher performance with full duplex**
  - The link can transfer data to and from the device simultaneously
- **High reliability and scalability**
  - High-Availability (HA): two ports for failover, error recovery, and error correction
  - A large number of disks: up to 255
  - But, need a special controller (e.g., HBA)

여러 device를 single system에 붙일 수 있다.

- **SAS 3.0 provides up to 1.2 GB/s throughput**



지금은 이것도 오래됨.

PCIe/NVMe

다 PCIe로 통합됨. 처음에는 GPU 용이였지만, Storage device의 성능이 늘어남에 따라 이전의 인터페이스들이 bottleneck이 되면서 PCIe를 쓰려고 하기 시작했고, 그래서 NVMe가 생김.

- **For nonvolatile memory attached to a computer over the high-speed PCIe bus (which is devised to support graphics)**
- **Provide much greater storage bandwidth than SATA and SAS**
  - Support multiple lanes (e.g., 1x, 2x, 4x, 8x, 16x): 1 GB/s per lane (PCIe 3.0)
  - Support multiple queues for better performance
    - 65,535 command queues (c.f., a single queue in AHCI)
    - 65,535 outstanding commands (c.f., 32 in AHCI)
  - Support full duplex

NVMe SSD
960 PRO
M.2
www.samsung.com/ssd
SAMSUNG ELECTRONICS CO., LTD.
1TB
P/N: MZVKP1T0HMJP
Model: MZ-V6P1T0 2016.08
PSID:1DZLL851J5A1S0FRNGTHQPFJW6EGR4G2
S/N: S3DSNY0H800004Y RATED: DC+3.3V ⎓ 2.7A
SAMSUNG
EUI64: 00253858610001D2 MSIP-REM-SEC-MZ-V6P1T0

NVMe SSD with M.2 form factor

DIMM

- **The fastest interface to the CPU, outperforming the NVMe/PCIe interface**
  - Storage media is seen as byte-addressable memory
  - No interrupt interrupts and deterministic latency

- **Products available in market**
  - **NVDIMM-N:**
    - Standard DRAM with the addition of NAND flash that stores DRAM’s data in event of power failure
  - **NVDIMM-F:**
    - Connect multiple SSDs to the DRAM bus
  - **Optane DIMM: (Introduced in 2019)**
    - Based on Intel’s 3D-Xpoint memory

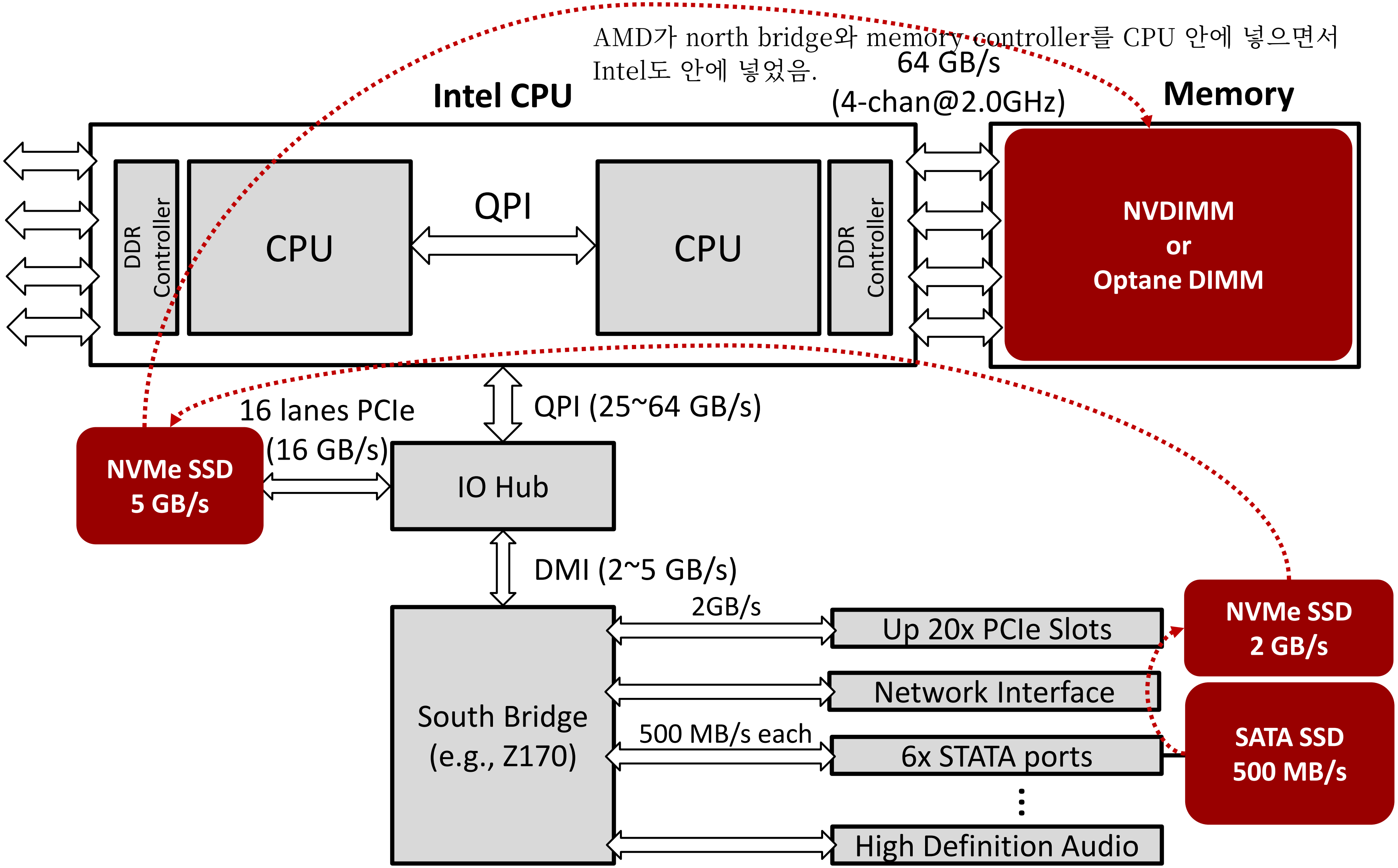
Summary

숫자가 정확하지는 않음

Interface	Mnemonic Meaning	Transfer Speed	Characteristics
SATA	Serial ATA	0.6 GB/s	Low cost
SAS	Serial Attached SCSI	1.2 GB/s	Supports multiple ports Error detection/correction
NVMe	Nonvolatile memory express over PCIe	1 GB/s per lane (3.0) 2 GB/s per lane (4.0)	Up to 16 lanes High command queue support
DIMM	Nonvolatile memory on memory channel	Up to 1 GB/s over 64-bit bus	Very low latency No interrupt Deterministic

- NVMe is becoming a standard interface both for desktop or server systems based on its high performance
- Optane DIMM will be alternative that will replace costly DRAM and slow SSD cache

Storage Bandwidth Hierarchy

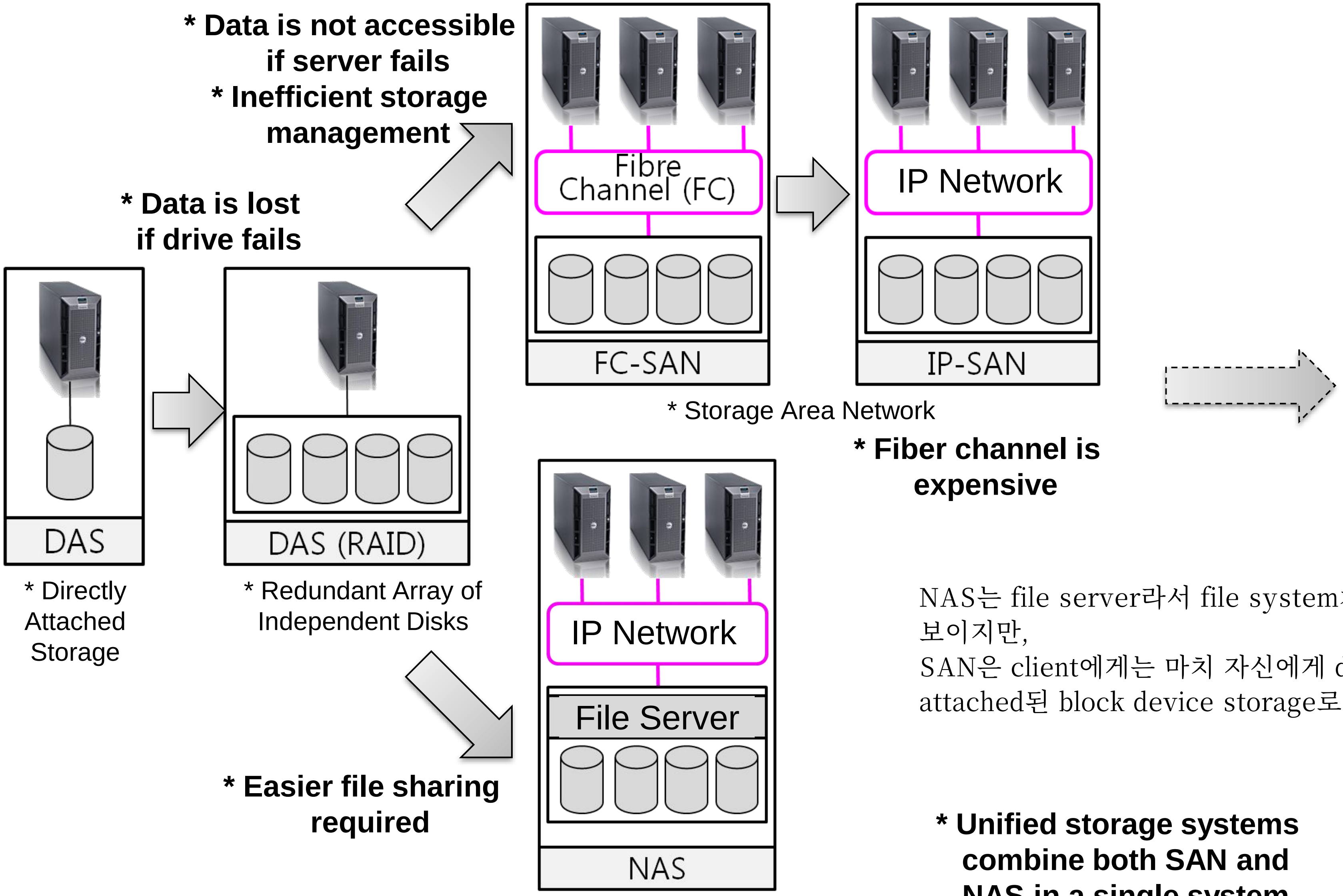


AMD가 north bridge와 memory controller를 CPU 안에 넣으면서 Intel도 안에 넣었음.

Outline

- Directly-Attached Storage (DAS)
- **Network-Attached Storage (NAS)**
- Storage-Area Network (SAN)

Evolution of Storage System



*** Data is lost if drive fails**

*** Data is not accessible if server fails**

*** Inefficient storage management**

* Directly Attached Storage

* Redundant Array of Independent Disks

* Storage Area Network

*** Fiber channel is expensive**

*** Easier file sharing required**

* Network-Attached Storage

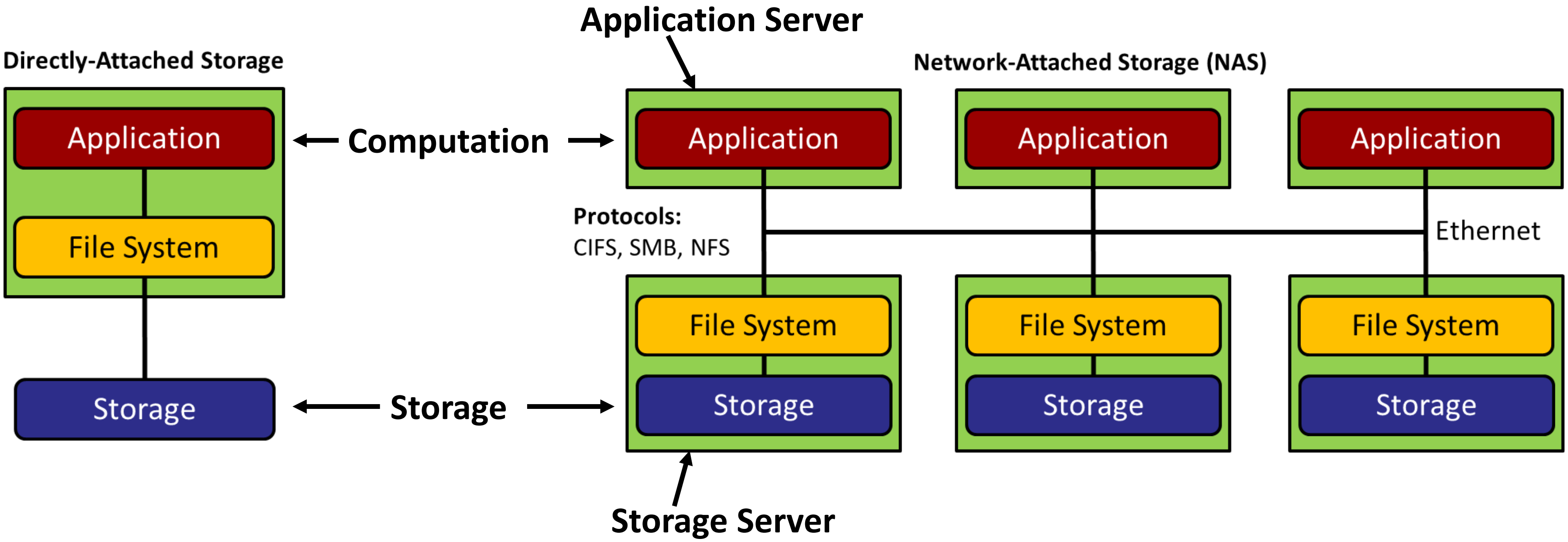
NAS는 file server라서 file system처럼 보이지만,
SAN은 client에게는 마치 자신에게 directly attached된 block device storage로 보인다.

*** Unified storage systems combine both SAN and NAS in a single system**

Computation & Storage Separation

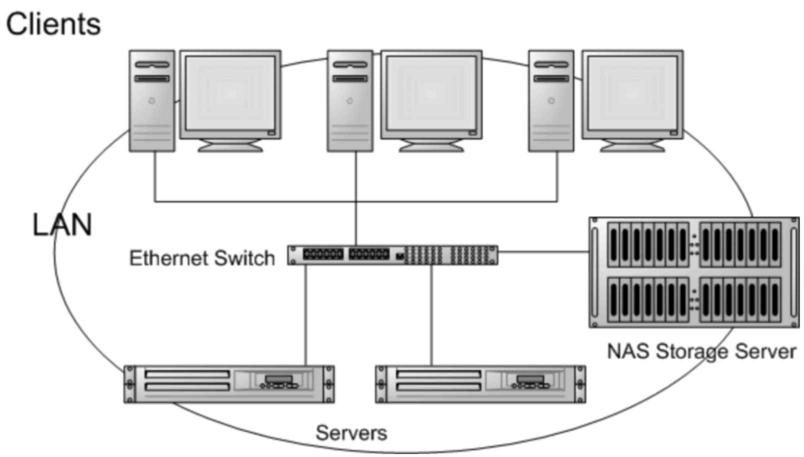
Scalability, Better management, Availability/Reliability, Sharing, Cost

- **Computation and storage are often separated in large scale systems**
  - ***Scalability***: add new application or storage servers depending on client's needs
  - ***Better management***: automatically back up user data
  - ***Availability*** / ***Reliability***: failure of a single server does not affect other servers
  - ***Sharing***: easy to share user contents
  - ***Cost***: thin provisioning, deduplication, and compression



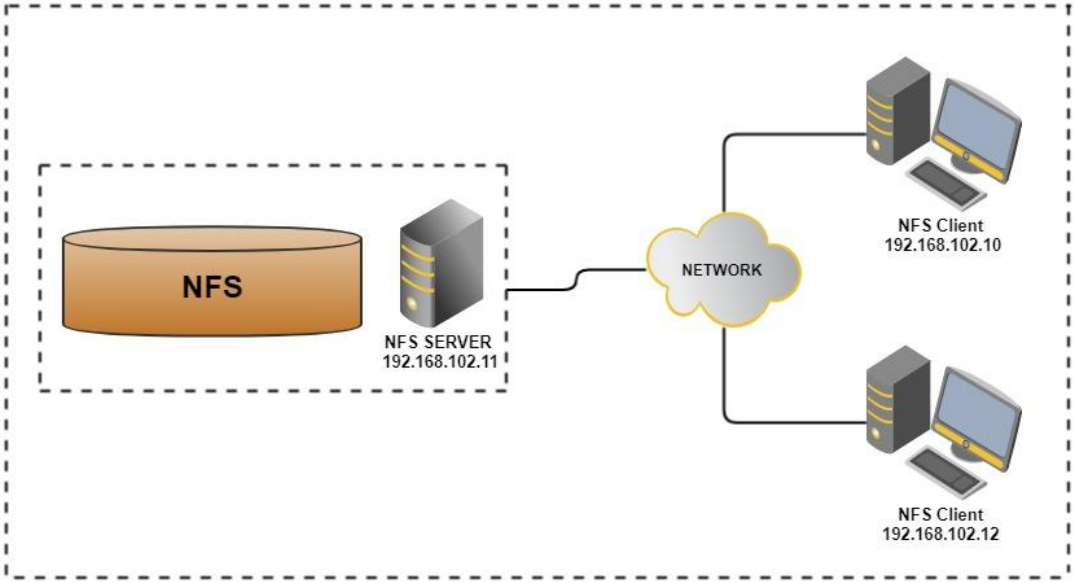
NAS Detail

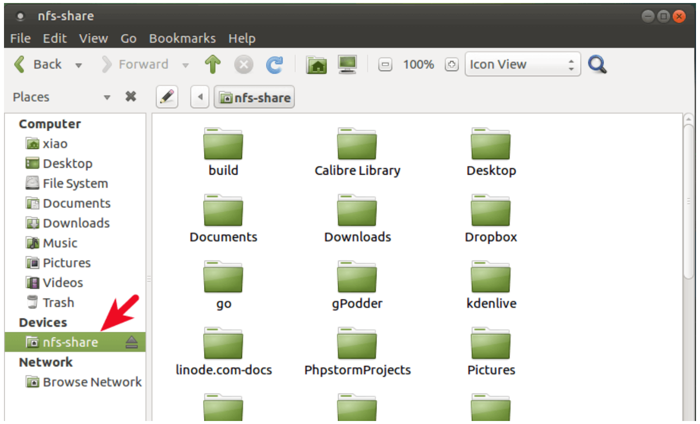
- **Provides file-level access to storage**
  - Ethernet connectivity through TCP/IP
  - NFS (Network File System)
  - CIFS (Common Internet File System)
  - SMB (Server Message Block)
- ***Networked file system*** **allows for concurrent access to data**
- **Several layers between data request and receipt**



NFS*

- A distributed file system protocol developed by Sun Microsystems in 1984
- Allow a user on a client computer to access files over a computer network much like local storage is accessed
- The NFS is an open standard defined in a Request for Comments (RFC), allowing anyone to implement the protocol



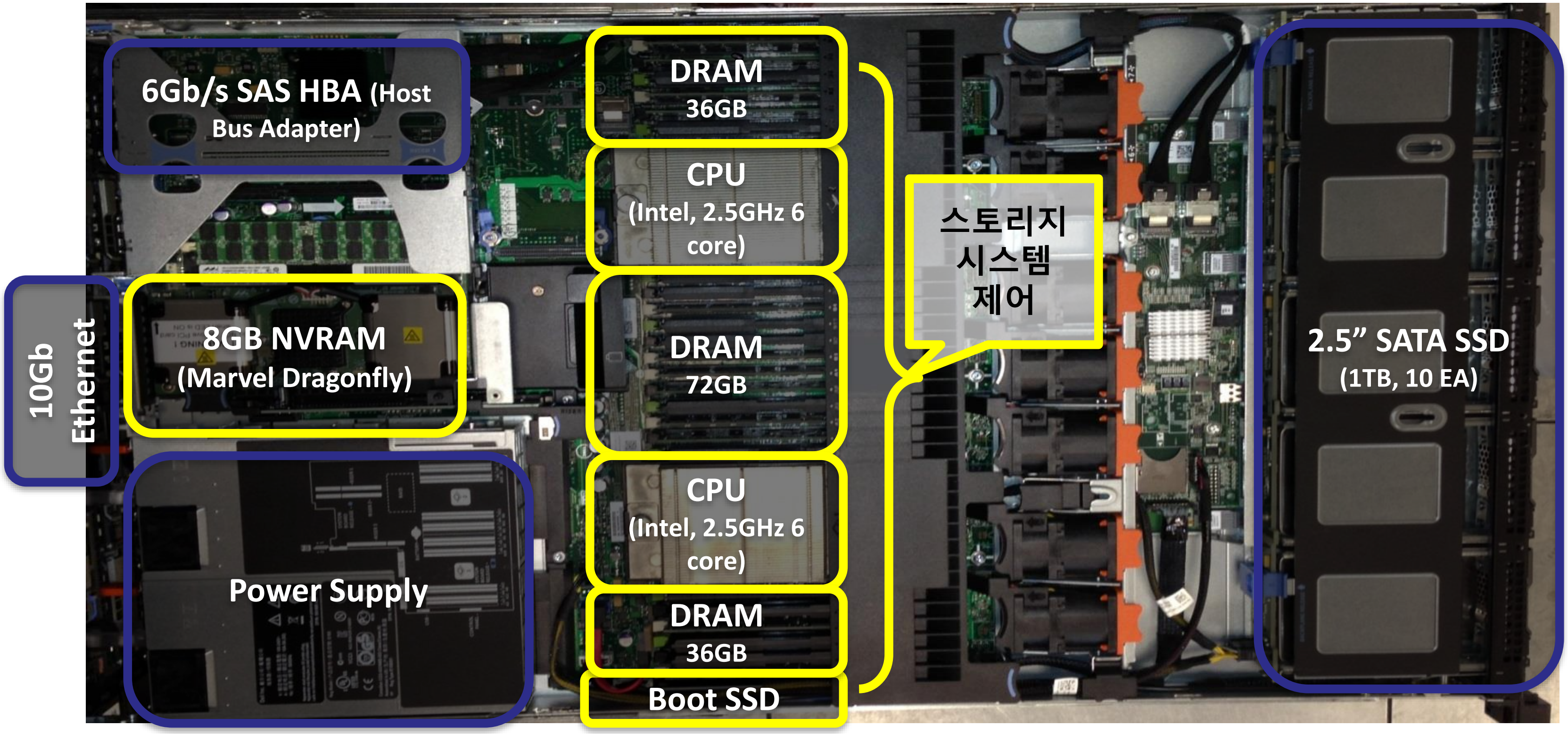


* Design and Implementation of the Sun Network Filesystem, USENIX ATC ‘85

Enterprise NAS Server

- Responsible for servicing user requests over the network
- Hardware specification is similar to high-end enterprise servers
- Perform *lots of jobs* internally

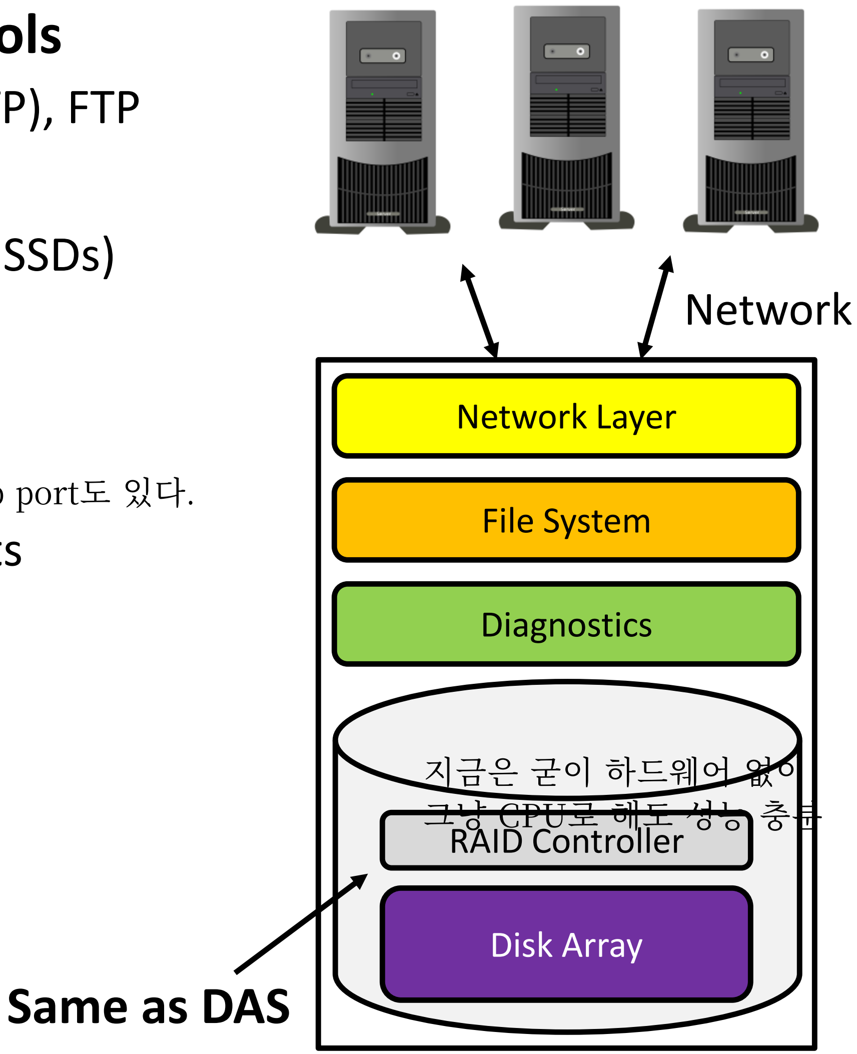
[Solidfire HW Architecture]


6Gb/s SAS HBA (Host Bus Adapter)
10Gb Ethernet
8GB NVRAM (Marvel Dragonfly)
Power Supply
DRAM 36GB
CPU (Intel, 2.5GHz 6 core)
DRAM 72GB
CPU (Intel, 2.5GHz 6 core)
DRAM 36GB
Boot SSD
스토리지 시스템 제어
2.5" SATA SSD (1TB, 10 EA)

Enterprise NAS Server Features

NAS의 기능들은?
다양한 file - sharing protocol 지원, 여러 disk 관리, Scalability, Fault tolerant, Data protection 등..

- **Support various file-sharing protocols**
  - Windows (CIFS), UNIX (NFS), Web (HTTP), FTP
- **Disk management**
  - Manage many disks (32 ~ 250 HDDs or SSDs)
- **Scales from GBs to TBs**
  - Scale up & scale out
- **Fault tolerant** NVRAM 뿐만이 아니라 backup port도 있다.
  - Dual, redundant, hot-swap components
- **Data protection**
  - RAID, Backup to disks & tape
- **Management software**
  - Manage & setup from remote location
- **Diagnostic software**
  - Predictive failure analysis and alters

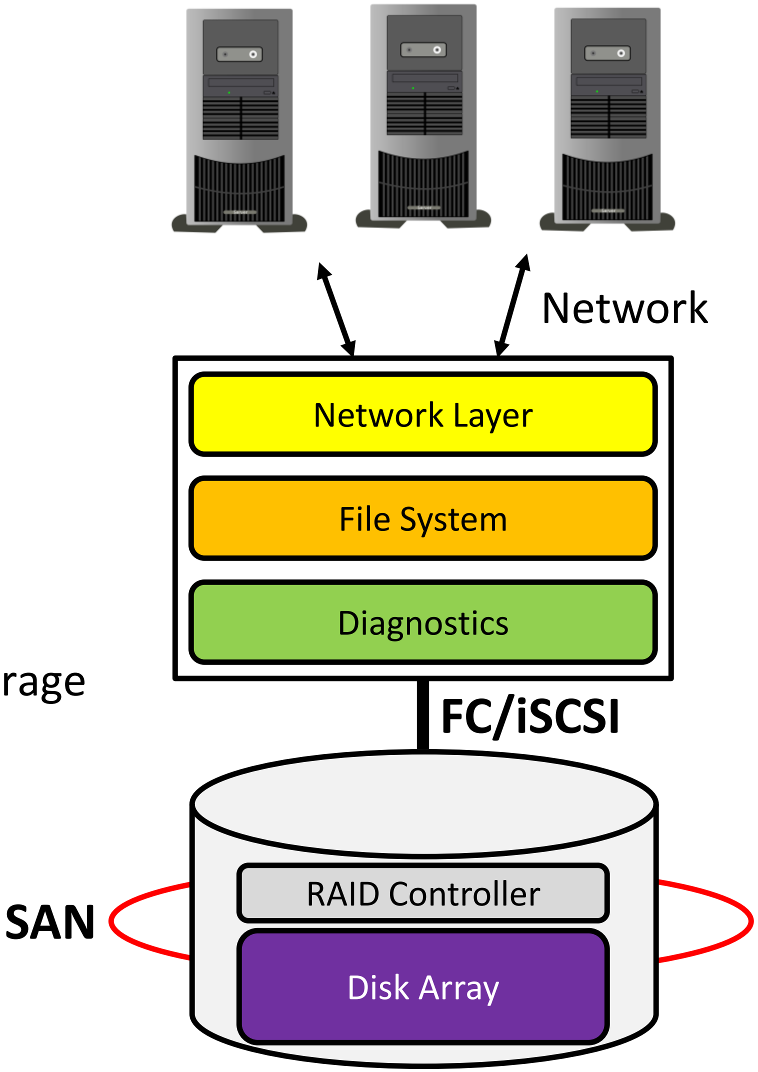


지금은 굳이 하드웨어 없이
그냥 CPU로 해도 성능 충분

NAS Gateway

- **Offers benefits and characteristics of NAS**
  - Connect to IP networks
  - Performs as a file server
  - Heterogeneous file sharing
  - Data protection
  - Clustering and failover features

- **NAS gateway is a NAS appliance with one exception**
  - Supports direct attachment to Fibre Channel storage or connection to a storage device across SAN
  - Do not have integrated disks for data storage



Outline

- Directly-Attached Storage (DAS)
- Network-Attached Storage (NAS)
- Storage-Area Network (SAN)

SAN Detail

SAN은 block - level stroage를 서버로 제공해준다.

- **SAN storage devices are connected over the network to servers**
- **Provides *block-level storage* that can be accessed by the applications running on any networked servers**

- **Differences between SAN and NAS**
  - While SANs provide block-level storage for servers, a NAS device provides file-level storage for end users
  - OS sees a SAN as a disk, while they see a NAS device as a file server

→ Latest storage boxes support ether NAS, SAN, or both, depending on configuration

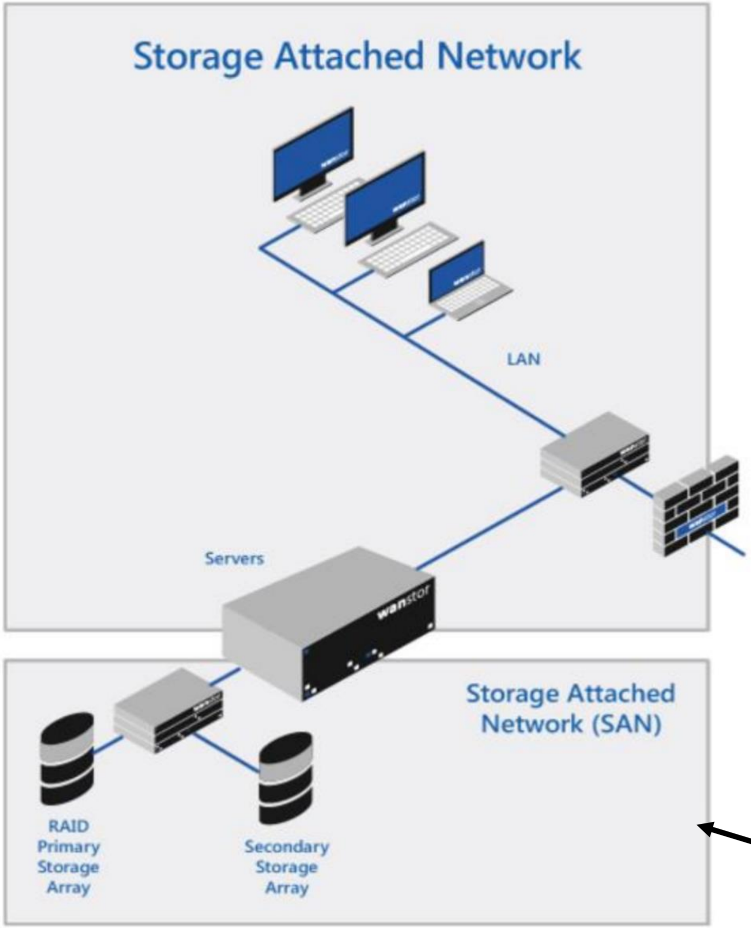
SAN Architecture

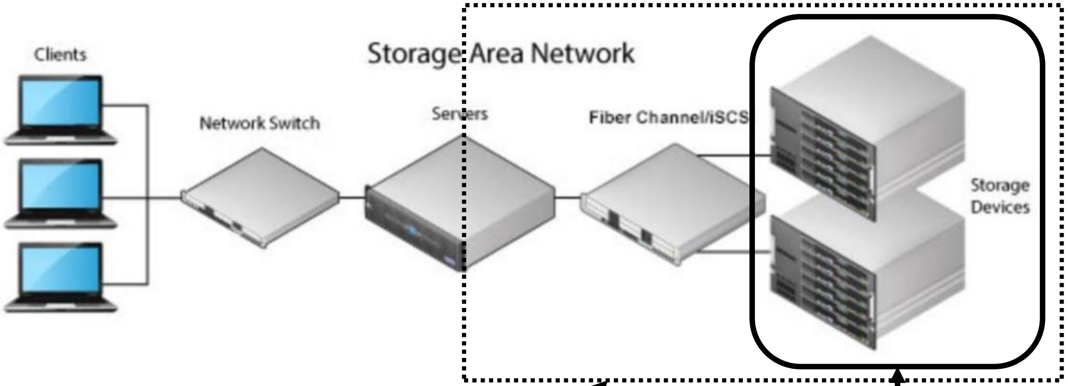
storage part와 application part의
network가 분리되어 있다.
(NAS는 분리되어있지 않음)
왜 separate network?
1. 성능 - network throughput
2. 보안
3. fiber channel 등의 special한 걸로
data access 부분과 server access 나누기

그러나 사실
이렇게 separate된
건 안 쓴다.

Provide services

Access data



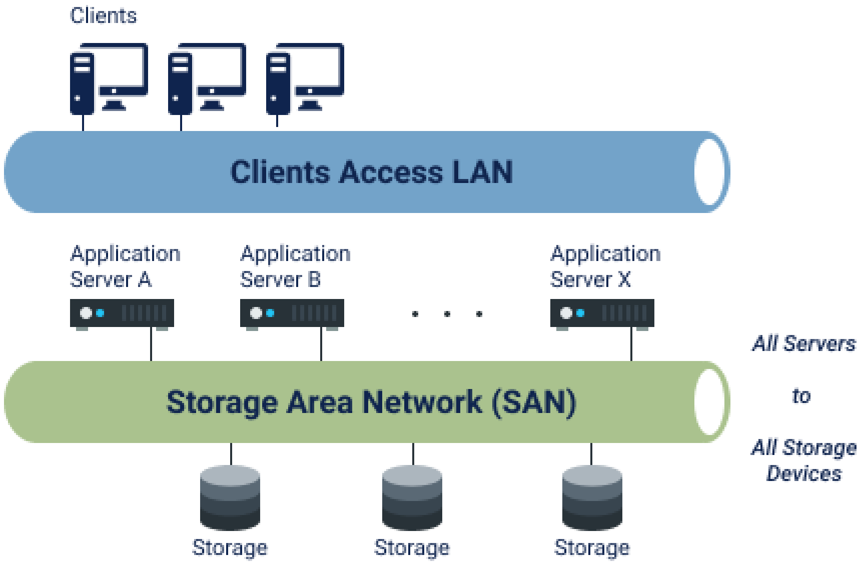


Separate
storage network

Almost the same
as storage servers
in NAS
(x86 CPU +
HDDs or SSDs)

Fibre Channel

- **Fibre Channel (FC) stands for a set of protocols, technologies and services used to build a "classic" SAN network**
  - Fibre Channel Protocol (FCP) - data transfer protocol that lets through SCSI commands.
  - Fibre optic infrastructure - used to transmit data to and/or from FC devices.
  - Name Service - acts as a database for connected devices. It is quite similar to a domain name system (DNS).
  - Set of flow control service



Fibre Channel (Cont.)

- **"FC SAN" implies a storage network built up of dedicated hardware adapters and switches, connected using fiber optics**
  - As the network is developed for high-loaded storage devices, it uses a strong cyclic redundancy check (CRC) – data is not corrupted when transmitted
  - Fewer retransmissions compared to TCP/IP and connection retries due to loss of data
  - More isolated compared to TCP/IP-based networks – lower security risks
  - Support 8Gbps, 16Gbps, and 32Gbps

- **Disadvantage**
  - Expensive – FC requires buying special network switches and storage adapters

iSCSI

IP 망을 통해 SCSI 명령을 전달함으로써 인트라넷을 거쳐 데이터 전송을 쉽게 하고 먼 거리에 걸쳐 스토리지를 관리하는 데 쓰인다.

SCSI와 거의 똑같지만 대신 TCP/IP channel을 사용한다.

- **The basic concept of iSCSI is simply putting SCSI commands inside of a typical TCP/IP channel**

  SCSI command를 TCP/IP channel에 넣는다.

  - Just install ***iSCSI Target/Initiator*** software onto your ***storage server*** and its ***clients***

- **Ethernet and TCP/IP are widely deployed and dominant**
  - Well understood technology; Low acquisition cost; Unlimited distance
  - A scalable technology with 10/100/1000/10000 Mbps
  - Allow the creation of a single physical network using familiar standards
    - VLAN may be used for separating storage traffic from intranet traffic
  - Bring interoperability & Ethernet economics to storage

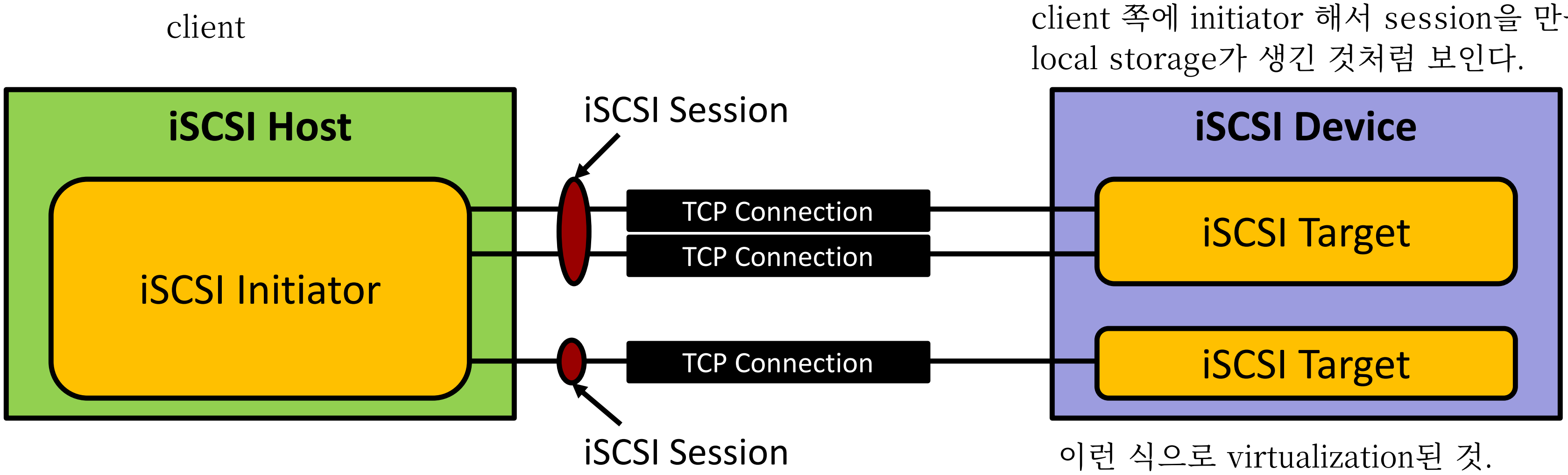
Fiber channel은 거리상의 한계가 있지만 tcp/ip 이므로 거리 괜찮음

iSCSI (Cont.)

- **TCP/IP over Ethernet are designed for common usage**
- **No strong data flow controls or built-in storage discovery services**
  - IP addresses of iSCSI storage and clients, frame sizes, LUN visibility, etc
  - Optimize the network for large data block transfers to get relatively high performance
  - Hardware-accelerated network adapters to offload iSCSI processing from a host server or client

iSCSI Connectivity

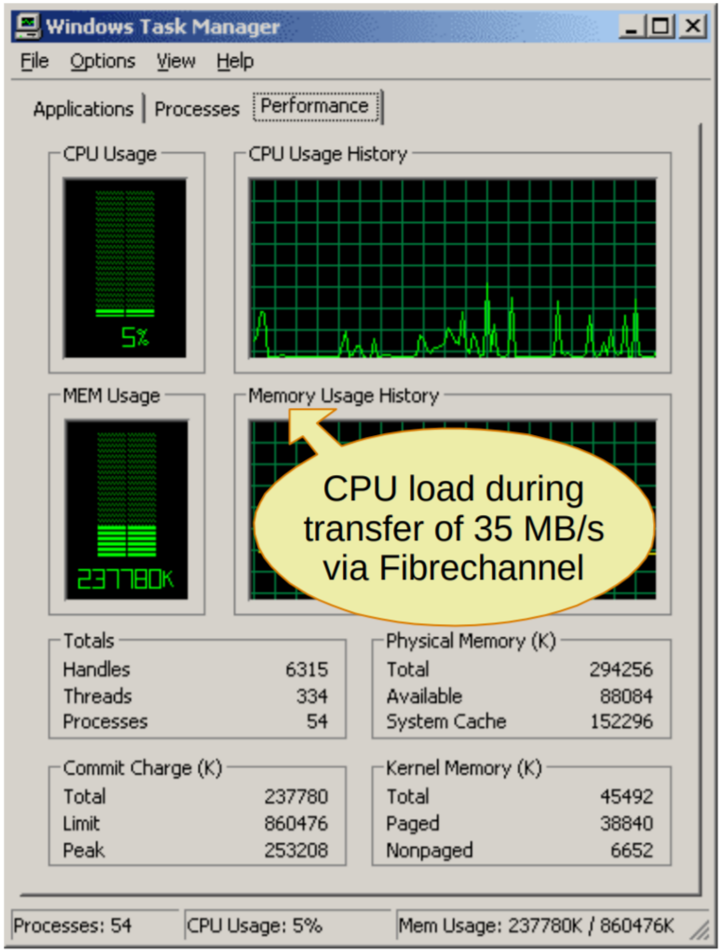
서버에 target을 만들어 두고
client 쪽에 initiator 해서 session을 만들면
local storage가 생긴 것처럼 보인다.



이런 식으로 virtualization된 것.
disk가 10개여도 virtualize해서 100개로
보이게 할 수 있다.

- Initiators and targets can be implemented in H/W or S/W
- Session between initiator and target
  - One or more TCP connections per session
  - Login phase begins each connection
- Services (e.g., authentication, security) negotiated during login
- TCP protocol provides
  - Delivery of SCSI commands in order
  - Recovery from lost connections

CPU Load



CPU load during transfer of 35 MB/s via Fibrechannel

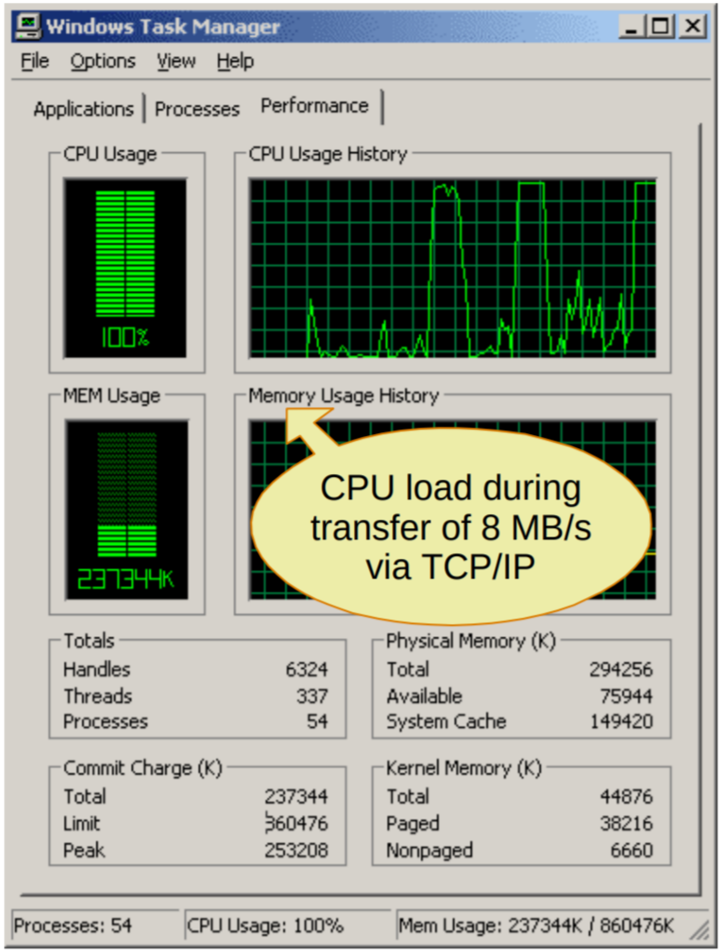
Totals	
Handles	6315
Threads	334
Processes	54

Physical Memory (K)	
Total	294256
Available	88084
System Cache	152296

Commit Charge (K)	
Total	237780
Limit	860476
Peak	253208

Kernel Memory (K)	
Total	45492
Paged	38840
Nonpaged	6652

Processes: 54 | CPU Usage: 5% | Mem Usage: 237780K / 860476K



CPU load during transfer of 8 MB/s via TCP/IP

Totals	
Handles	6324
Threads	337
Processes	54

Physical Memory (K)	
Total	294256
Available	75944
System Cache	149420

Commit Charge (K)	
Total	237344
Limit	860476
Peak	253208

Kernel Memory (K)	
Total	44876
Paged	38216
Nonpaged	6660

Processes: 54 | CPU Usage: 100% | Mem Usage: 237344K / 860476K

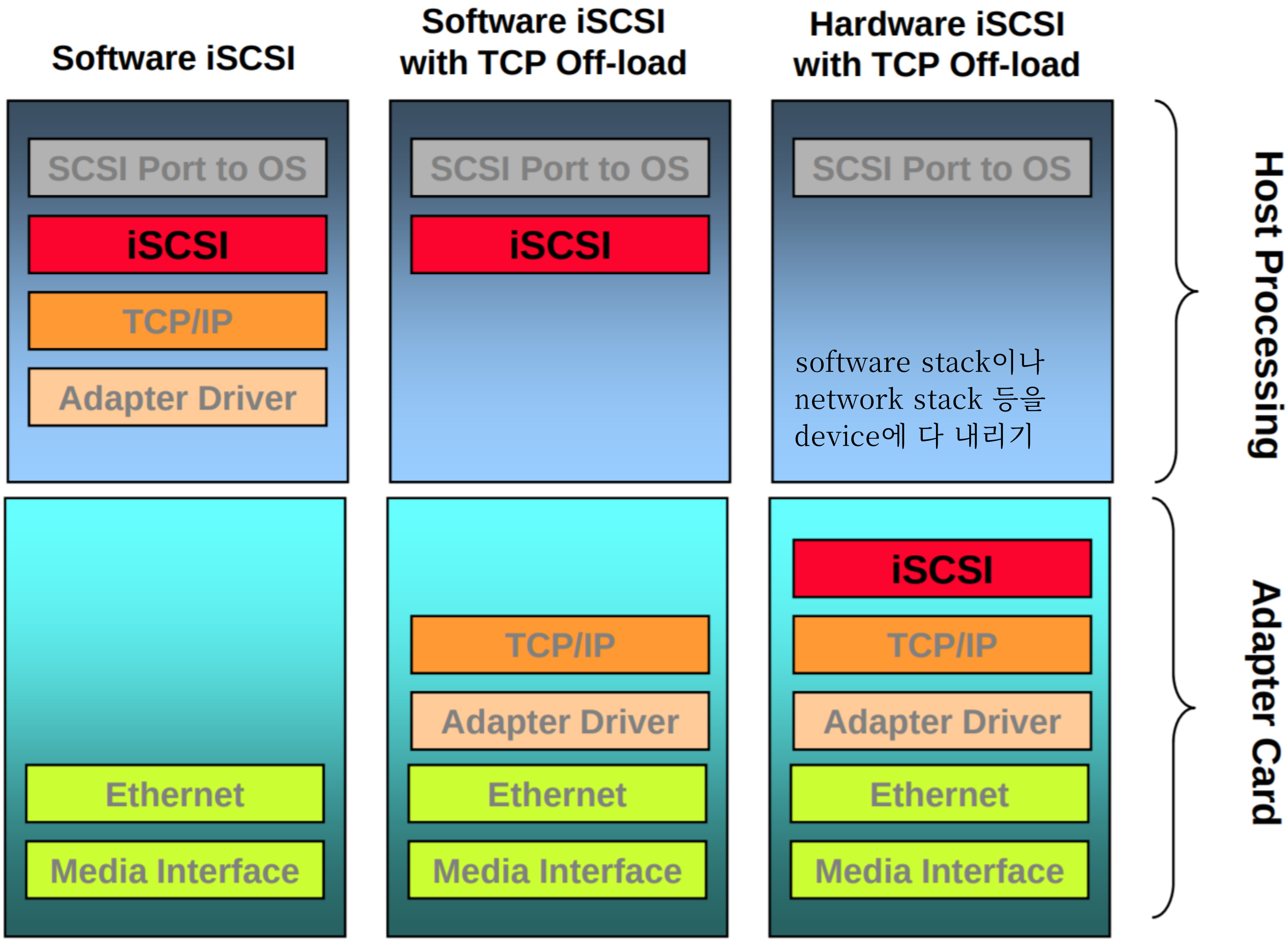
TCP/IP Overhead

- **Every TCP/IP connection that is part of an iSCSI session has processing overhead**
  - Connection setup / teardown
  - TCP state machine:
    - Acknowledge, timeout, and retransmission
    - Window management
    - Congestion control
  - Checksum calculation
  - TCP segmentation

iSCSI session에서 사용하는 TCP/IP connection은
CPU 사용량이 높아 Overhead가 된다.
그래서 TCP/IP Offload Engine (TOE)이 있기도 하다.

- ***TCP/IP Offload Engine*** **(TOE) helps at GbE NICs!**
  - 1 GbE links will not require full integrated TOEs
    - Increasing CPU performance might be sufficient
  - For higher than 10 GbE, TOE is necessary!

iSCSI & TOE Adapters



software stack이나
network stack 등을
device에 다 내리기

Outline

- Directly-Attached Storage (DAS)
- Network-Attached Storage (NAS)
- **Storage-Area Network (SAN)**
  - **NVMe-over-Fabric (NVMe-oF)**

NVMe-oF

NVMe - OF는
한 컴퓨터가 다른 컴퓨터에 연결된 NVMe device에 접근하기 위한 통신 프로토콜이다.

iSCSI 와 같다. 그러나 RDMA가 결합됐다.

- **NVMe-OF is a communication protocol that allows one computer to access NVMe devices attached to another computer**
  - Contrary to the standard NVMe protocol where NVMe devices are connected directly to PCIe bus

물론 concept 자체는 오래된 것

- **Combined with remote direct-memory access (RDMA)**
  - One computer can access another computer's memory as if that memory actually resided within the first computer
  - Don't need to go through the OS's I/O stack – run at speeds closer to the speed of memory

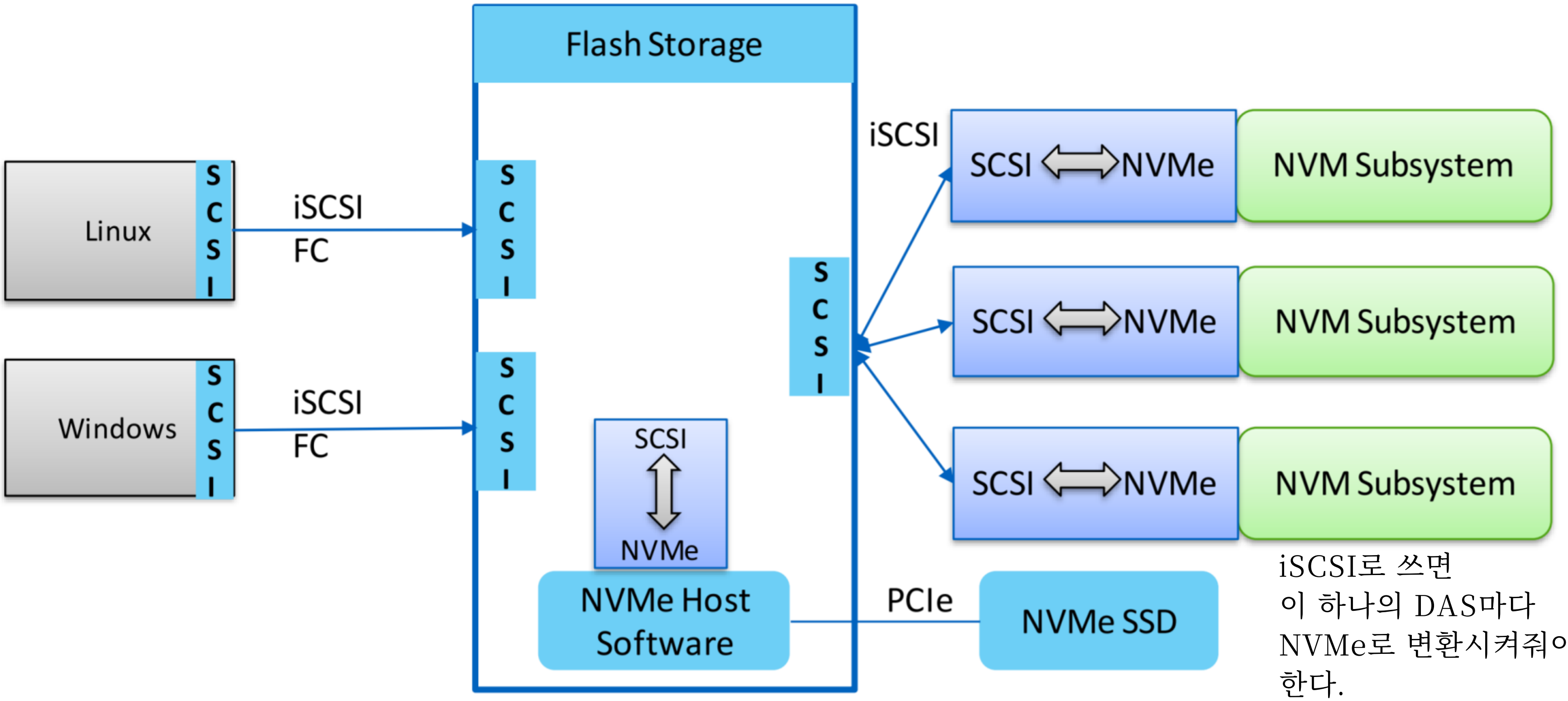
- **Implemented over Ethernet or InfiniBand**

- ***NVMe-oF will replace iSCSI in the future!***

지금은 많은 device가 NVMe interface이므로
이게 iSCSI를 대체할 것이다.

Why NVMe-oF?

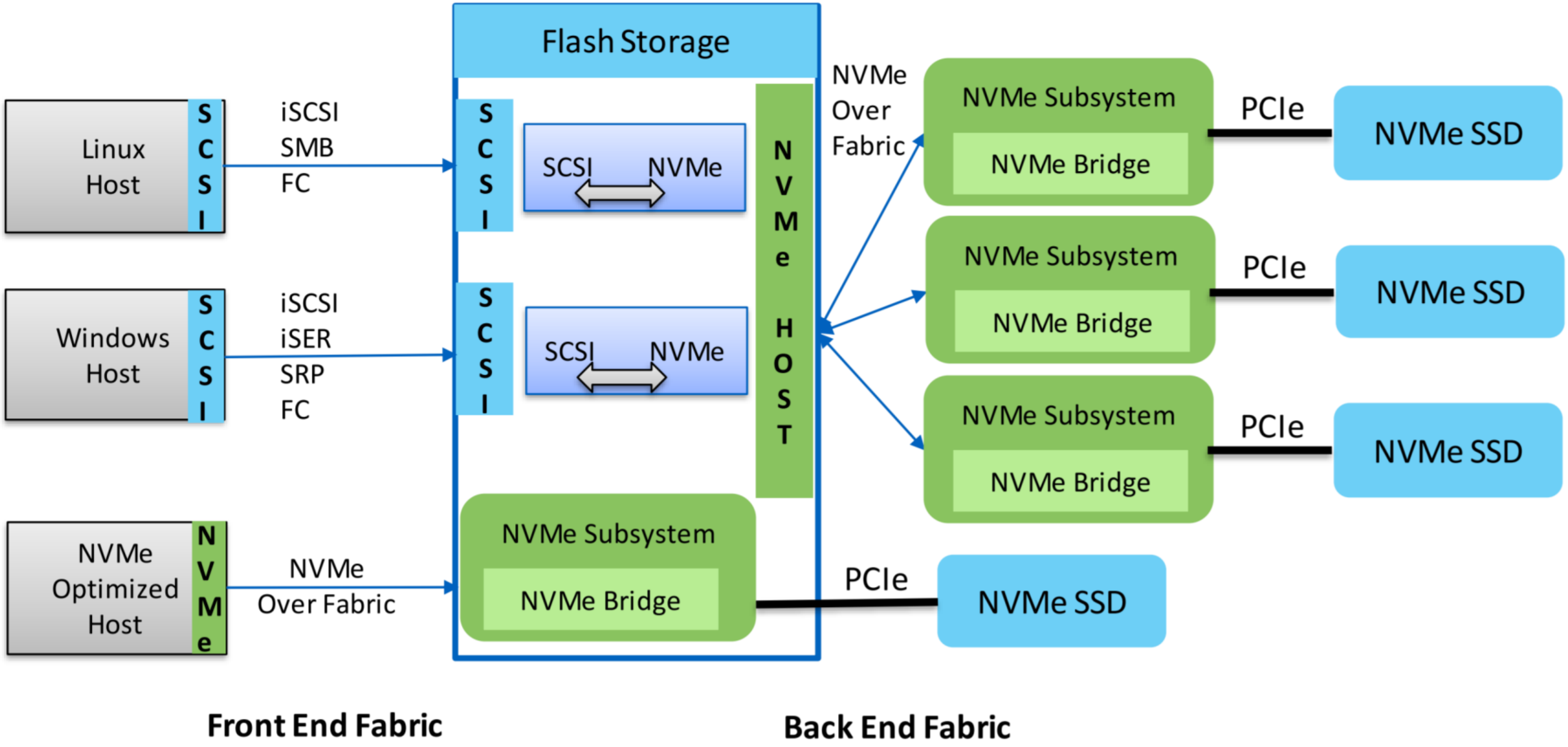
- **Protocol conversion bridge is required to access the data over network which increases I/O latency**



iSCSI로 쓰면
이 하나의 DAS마다
NVMe로 변환시켜줘야
한다.

Why NVMe-oF? (Cont.)

- NVMe-oF removes a burden on converting iSCSI comds to NVME cmds
- Enable us to take advantage of unique features of NVMe devices like multiple-queue architectures for fast storage



What is RDMA?

기본적으로 큰 관심을 받는 건
RDMA 때문이다.

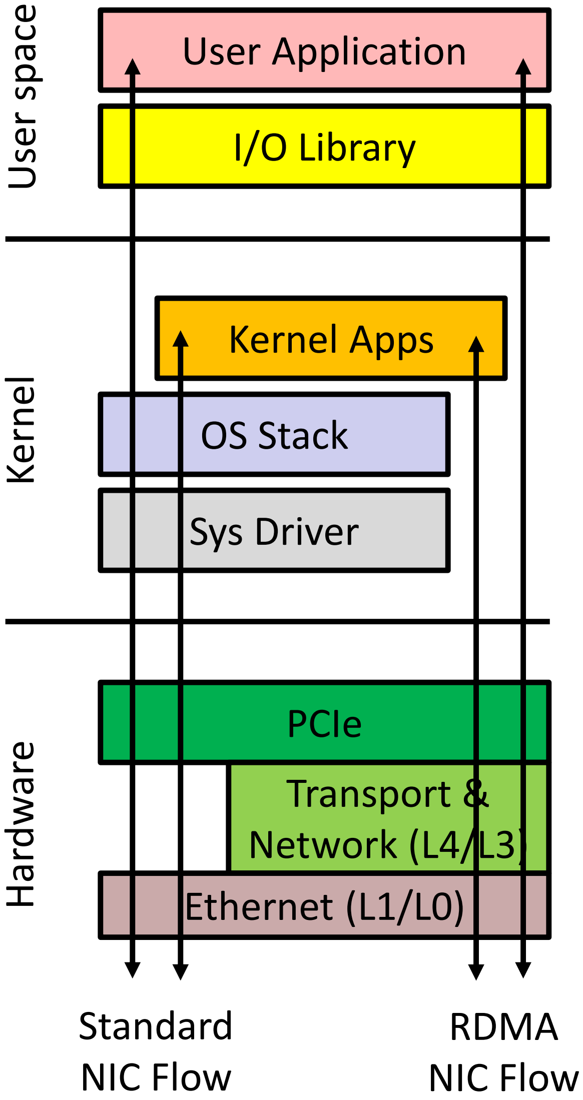
- **RDMA is a host-offload, host-bypass technology that allows an application (including storage) to make data transfers directly to/from another application's memory space**

- **The RDMA-capable Ethernet NICs (RNICs) – not the host – manage reliable connections between source and destination**

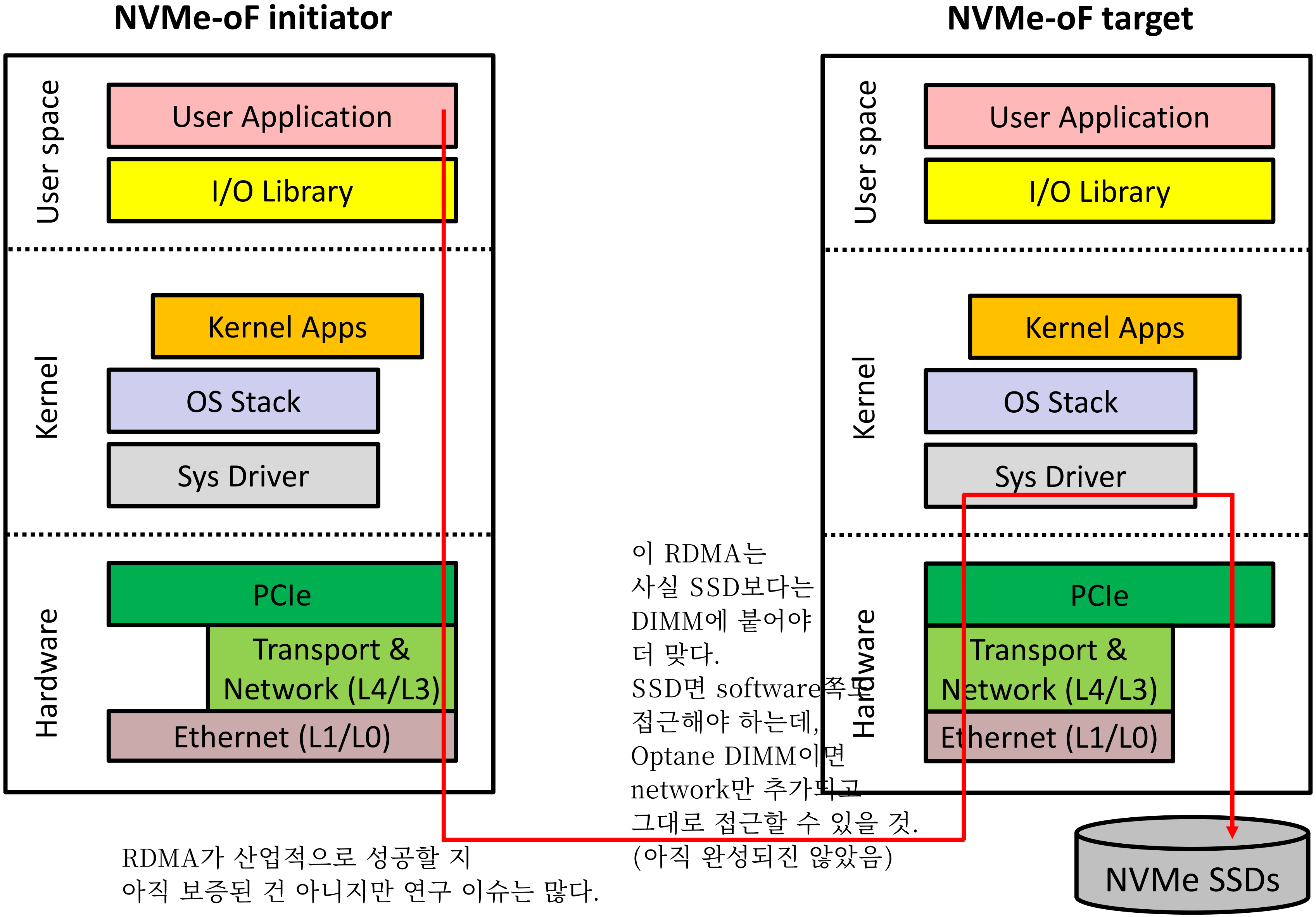
- **Applications communicate with the RDMA NIC using dedicated Queue Pairs (QPs) and Completion Queues (CQs)**
  - Suitable for the NVMe architecture

Benefits of RDMA

- **Bypass of system SW stack components that processes network traffic**
  - For user applications, RDMA bypasses the kernel altogether
  - For kernel applications, RDMA bypasses the OS stack and the system drivers
- **Direct data placement of data from one machine (real or virtual) to another machine – without copies**
- **Increased bandwidth while lowering latency, jitter, and CPU utilization**
- **Great for networked storage!**



How NVMe-oF w/ RDMA Works?



이 RDMA는
사실 SSD보다는
DIMM에 붙어야
더 맞다.
SSD면 software쪽도
접근해야 하는데,
Optane DIMM이면
network만 추가되고
그대로 접근할 수 있을 것.
(아직 완성되진 않았음)

RDMA가 산업적으로 성공할 지
아직 보증된 건 아니지만 연구 이슈는 많다.

End of Chapter 9