Storage

Jin-Soo Kim
(*jinsoo.kim@snu.ac.kr*)

Systems Software &
Architecture Lab.

Seoul National University

Fall 2022

Storage: A Logical View

- Abstraction given by block device drivers ("*block interface*")

- Operations
  - Identify(): returns N
  - Read (start sector #, # of sectors, buffer address)
  - Write (start sector #, # of sectors, buffer address)

Source: Sang Lyul Min (Seoul National University)

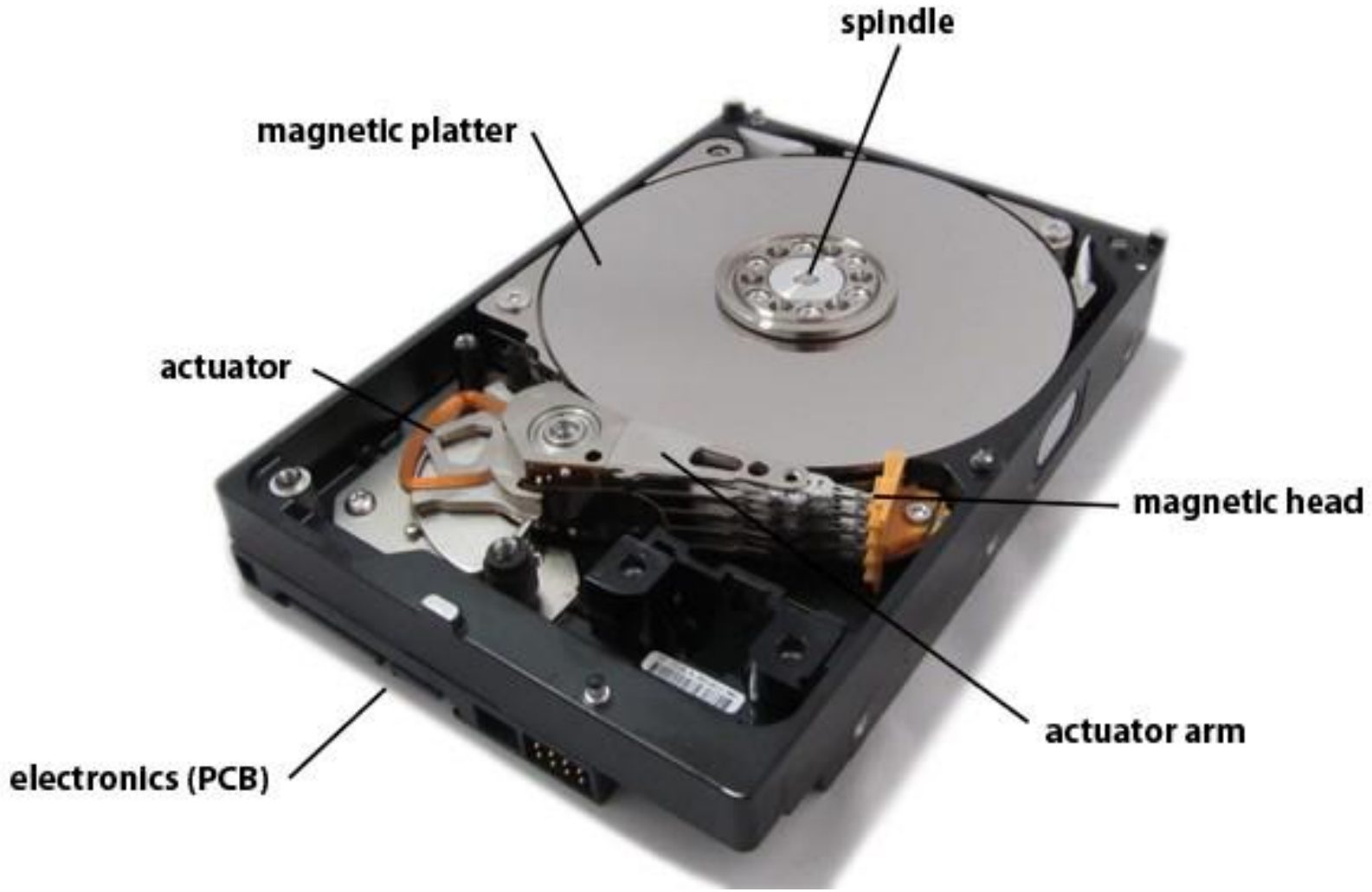
Hard Disk Drives (HDDs)

The First HDD

- IBM 305 RAMAC (1956)
  - First commercially produced hard disk drive
  - 5 MB capacity, 50 platters each 24” in diameter, $10,000/MB

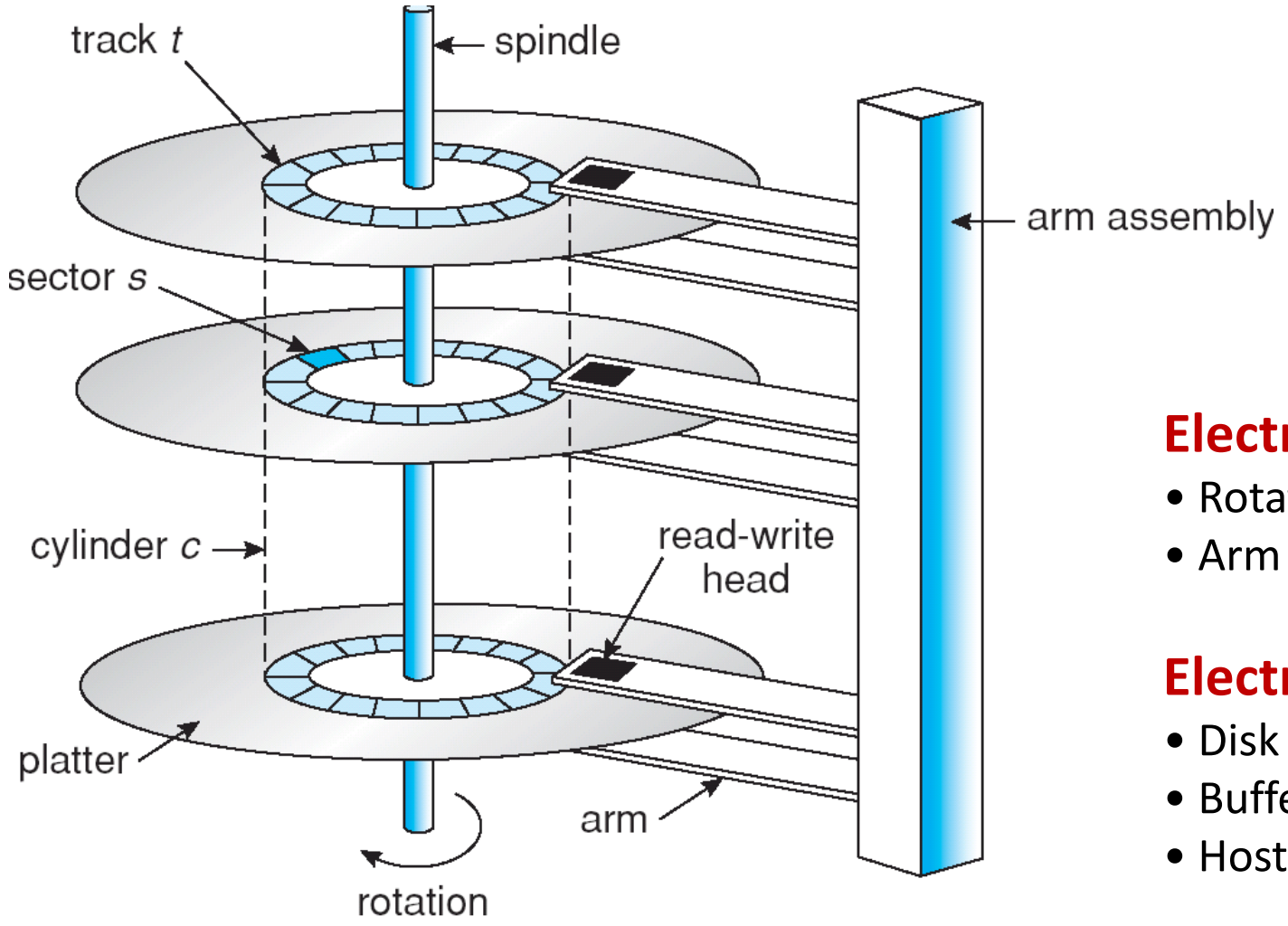
https://medium.com/@tonyjlum/future-of-data-storage-a65dc0cf3e40, https://courses.engr.illinois.edu/cs241/sp2012/lectures/38-disk.pptx

Anatomy of a HDD



http://www.helpdisc.rs/eng/data-recovery/hard-disks

Physical Drive Geometry



Electromechanical

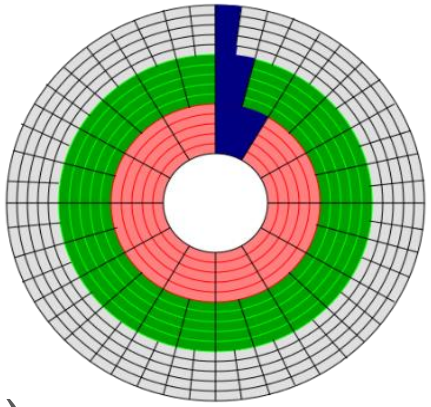
- Rotating disks
- Arm assembly

Electronics

- Disk controller
- Buffer
- Host interface

Interfacing with HDDs

- CHS (Cylinder-Head-Sector) scheme
  - The OS needs to know all disk "geometry" parameters
  - Modern disks are more complicated:
    Sector remapping, ZBR (Zone Bit Recording)
  - Can't be generalized to other devices (e.g., tapes, networked storage)



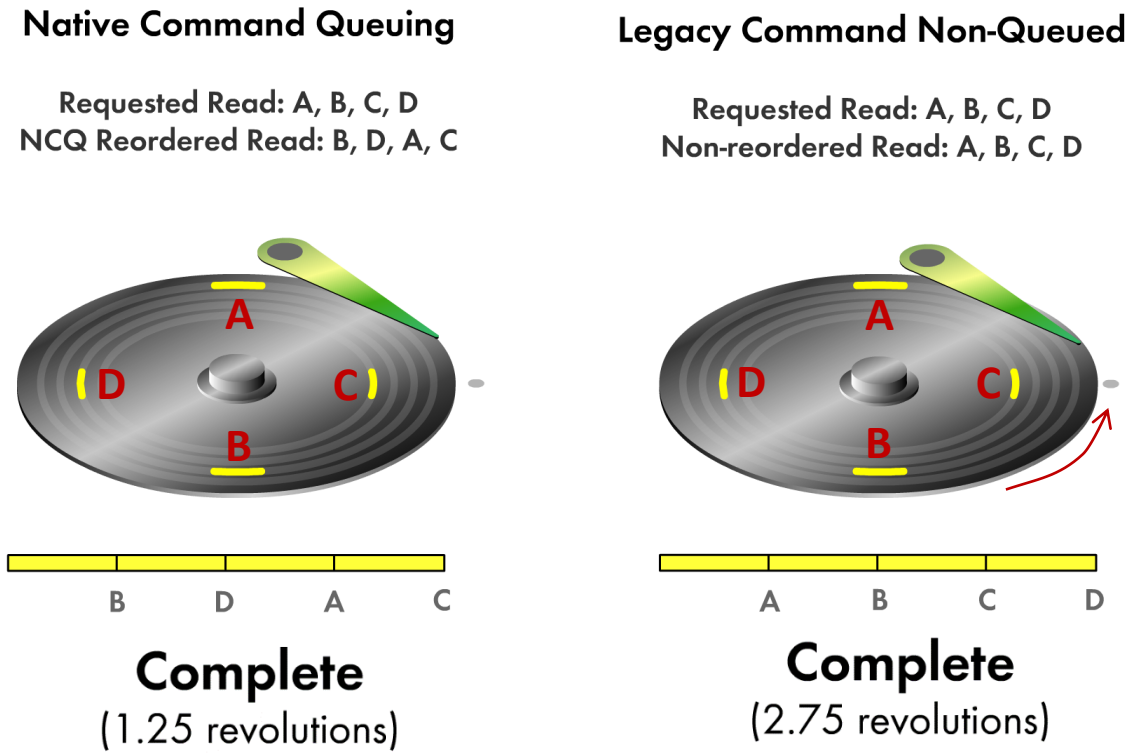
- Logical block addressing (LBA) scheme
  - First introduced in SCSI
  - Disk is abstracted as a logical array of blocks [0, …, N-1]
  - Disk maps an LBA to its physical location
  - Physical parameters of a disk are hidden from OS
  - 48-bit address with a release of ATA-6 in 2003

HDD Performance Factors

- Seek time (T_{seek})
  - Moving the disk arm to the correct cylinder
  - Depends on the cylinder distance (not purely linear cost)
  - Average seek time is roughly one-third of the full seek time
- Rotational delay ($T_{rotation}$)
  - Waiting for the sector to rotate under head
  - Depends on rotations per minute (RPM)
  - 5400, 7200 RPM common, 10K or 15K RPM for servers
- Transfer time ($T_{transfer}$)
  - Transferring data from surface into disk controller, sending it back to the host

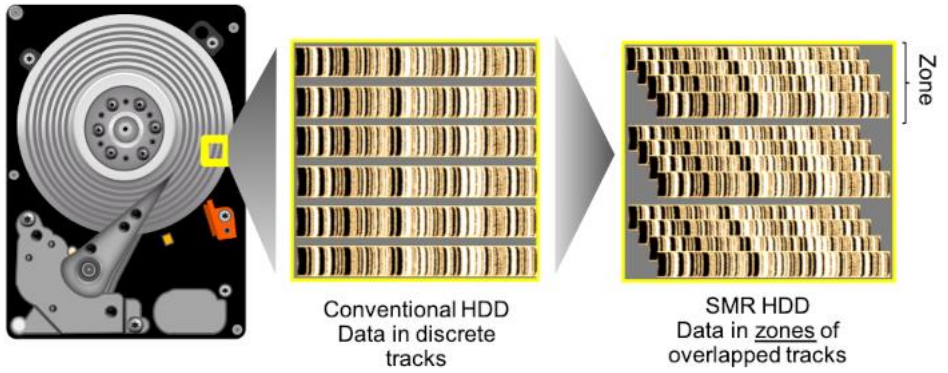
SATA NCQ

- Enqueue up to 32 commands in the drive
- Process them in an out-of-order fashion



SMR Disks

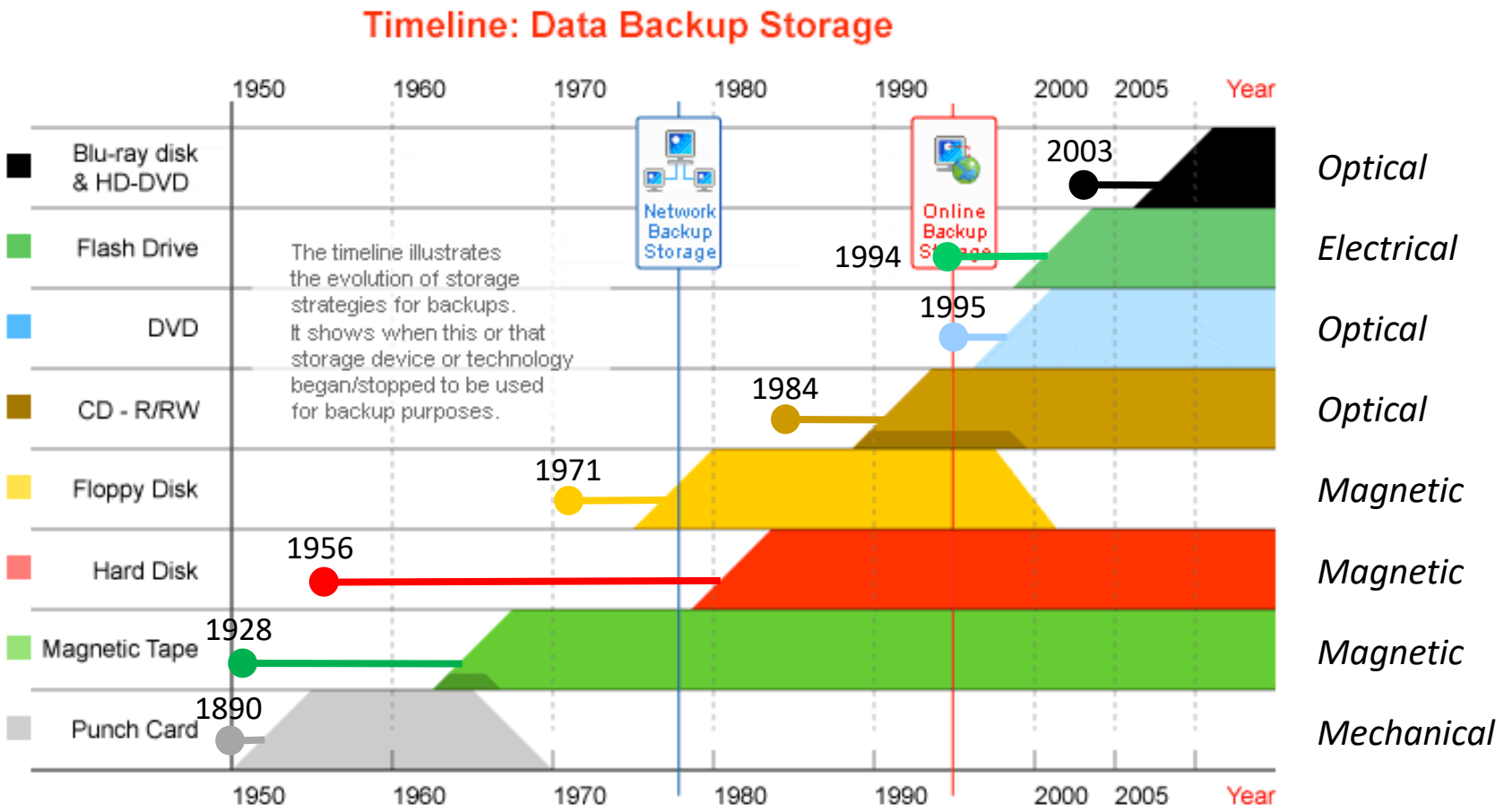
- Shingled Magnetic Recording
  - Recording heads are wider than reading heads
  - Write new tracks that overlap part of the previously written magnetic track
  - Higher storage capacity compared to CMR
- Command interface
  - Sequential zones (+ conventional zones)
  - SCSI ZBC (Zoned Block Commands)
  - ATA ZAC (Zoned Device ATA Command Set)
  - Report Zones, Reset Zone Write Pointer, Open Zone, Close Zone, Finish Zone
- Device-managed vs. Host-managed vs. Host-aware



https://zonedstorage.io/introduction/smr/

Solid-State Drives (SSDs)

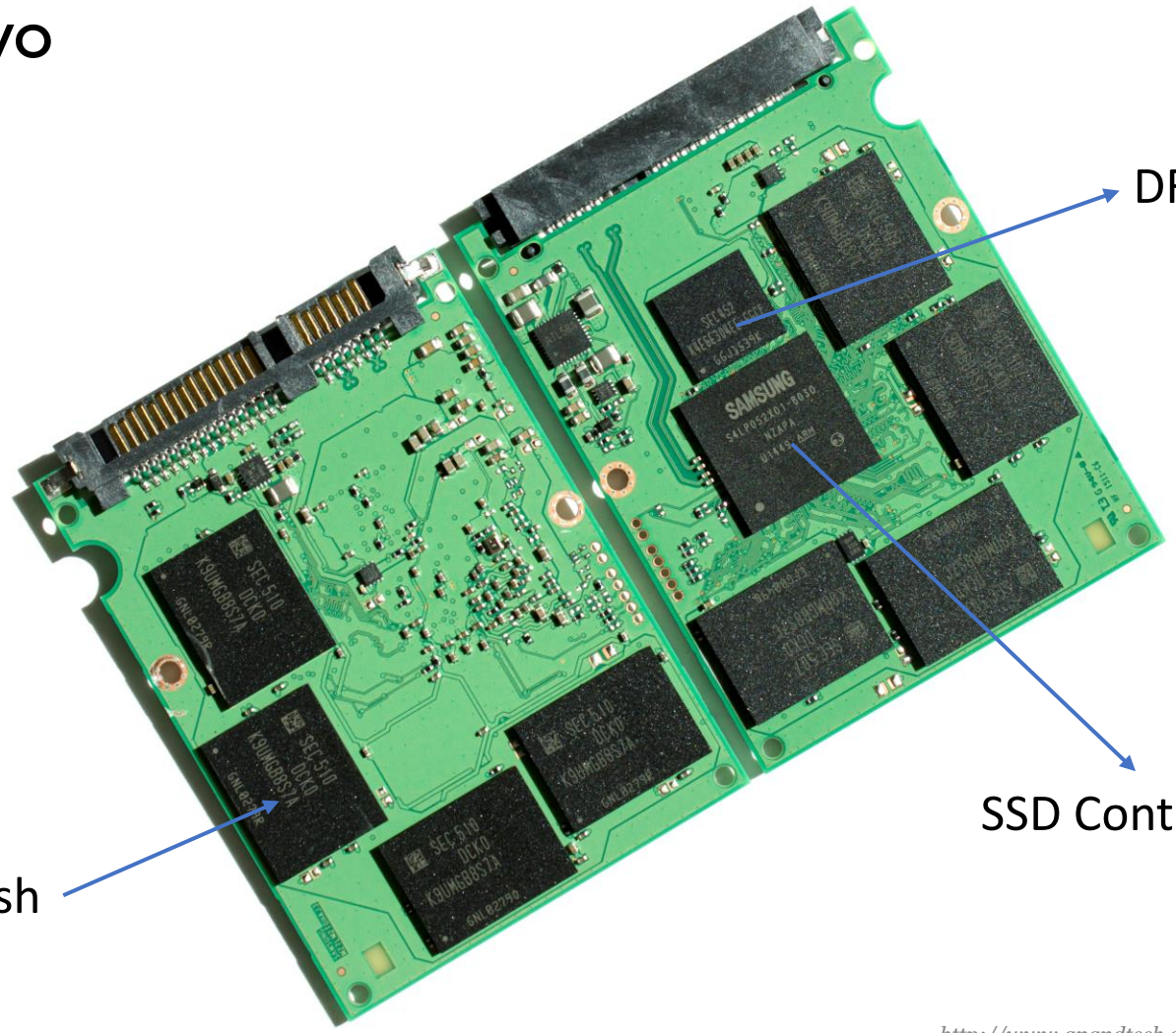
A Quest for Non-Volatile Storage



http://www.backuphistory.com

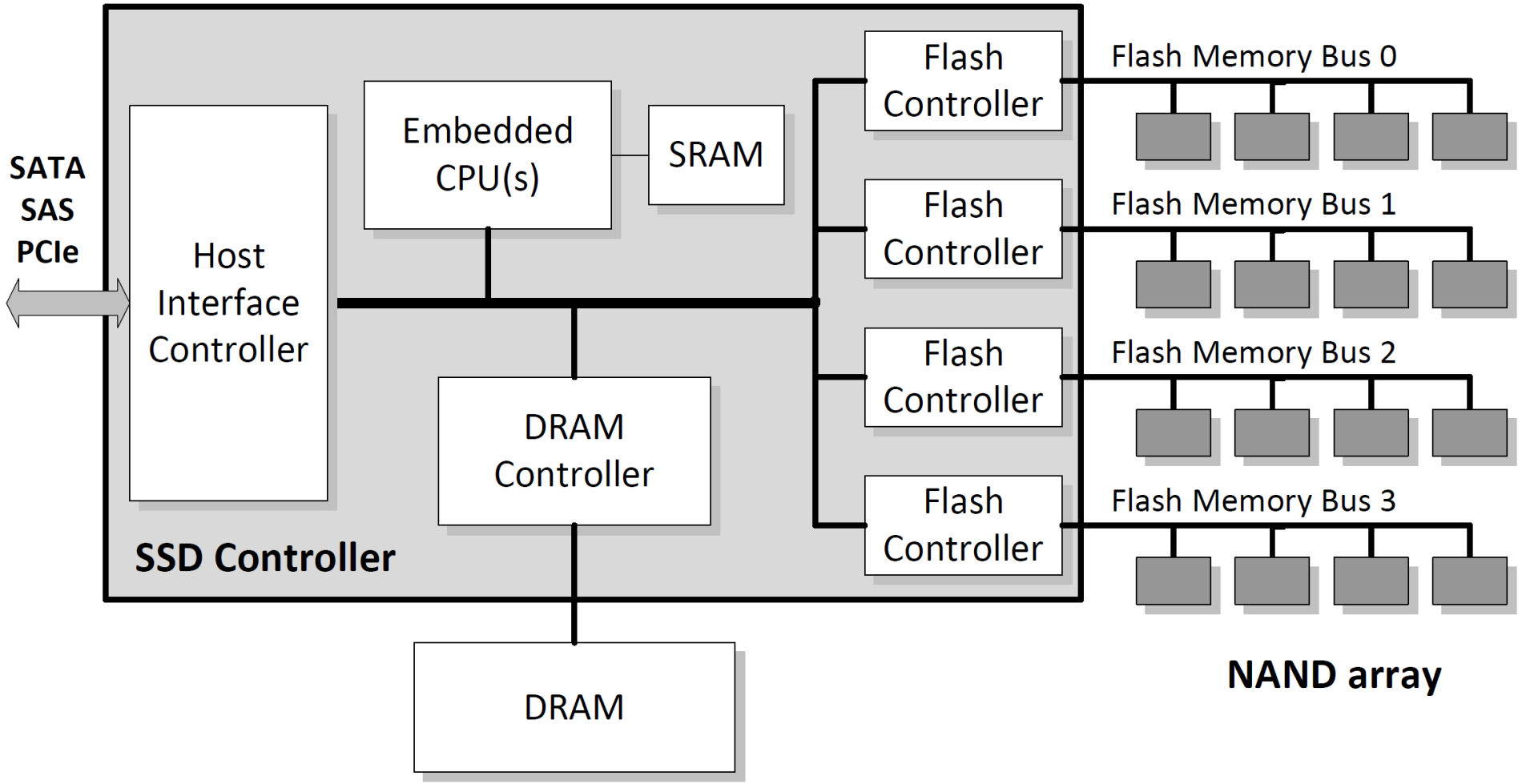
Anatomy of an SSD

- Samsung 850 Evo

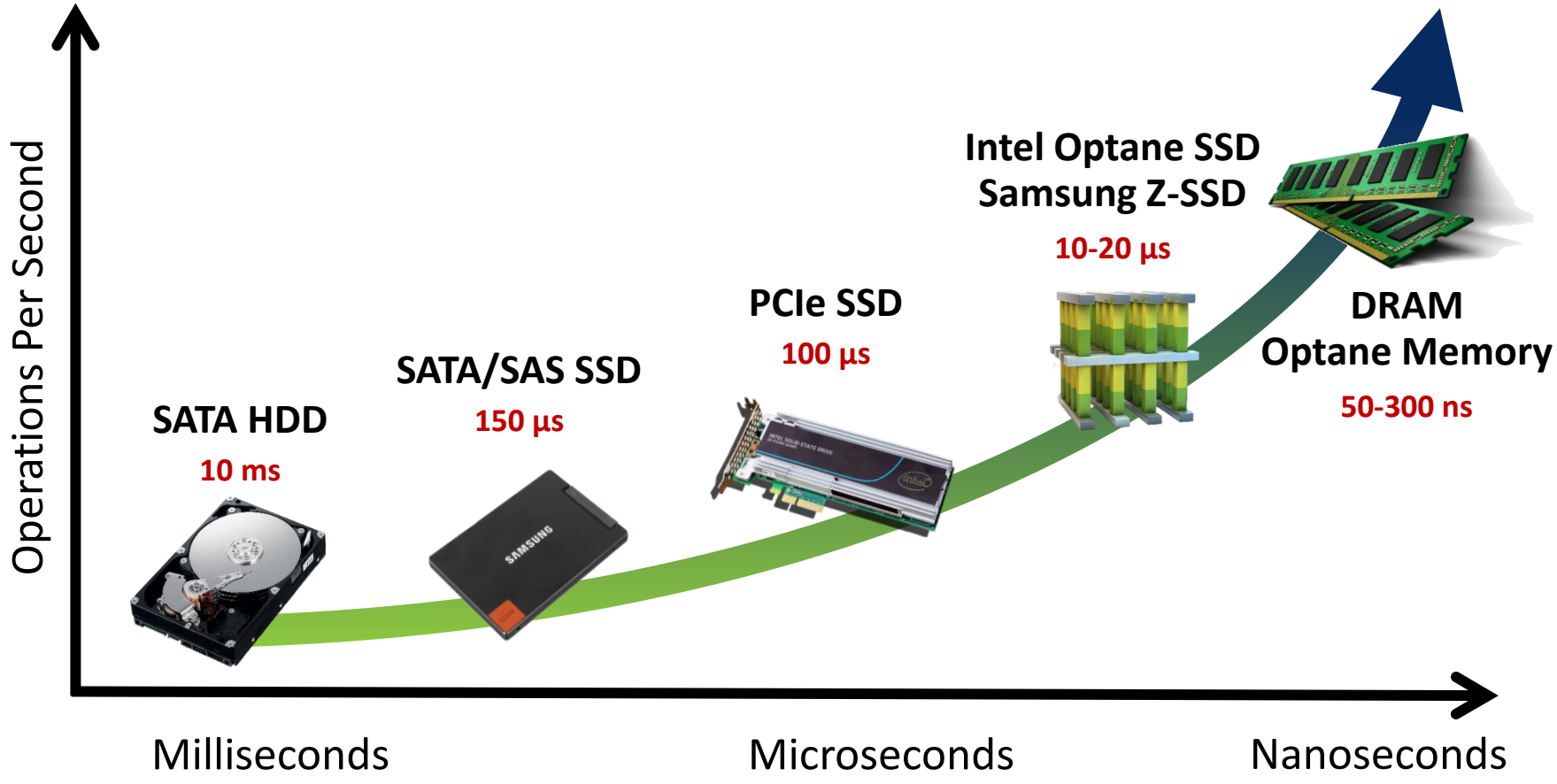


http://www.anandtech.com/show/9451/the-2tb-samsung-850-pro-evo-ssd-review

SSD Internals

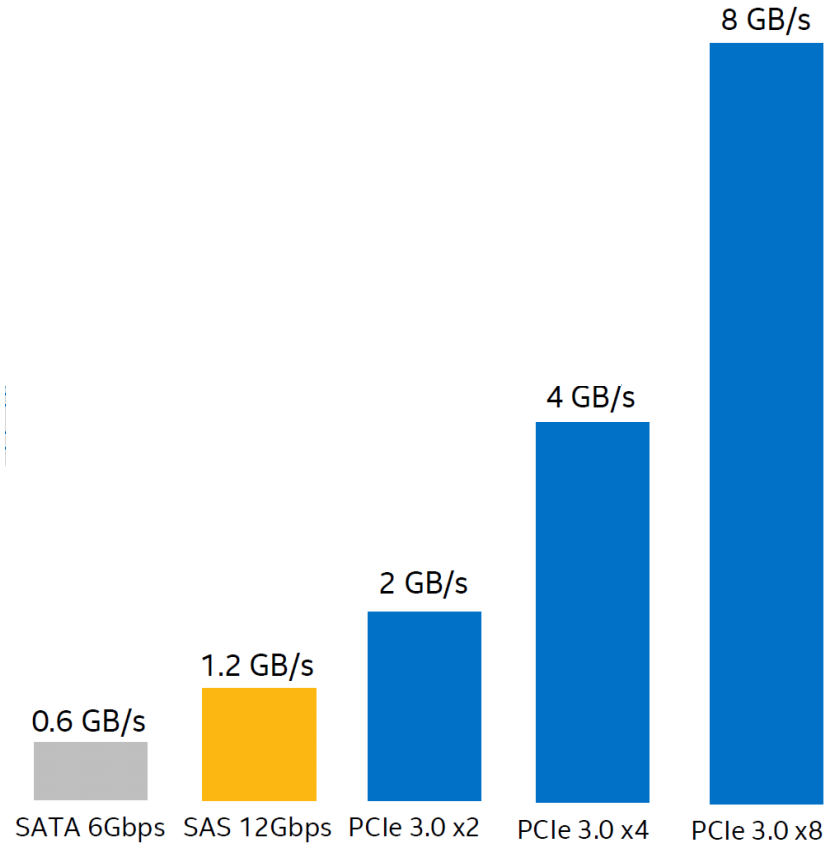


Moving Closer to the Processor

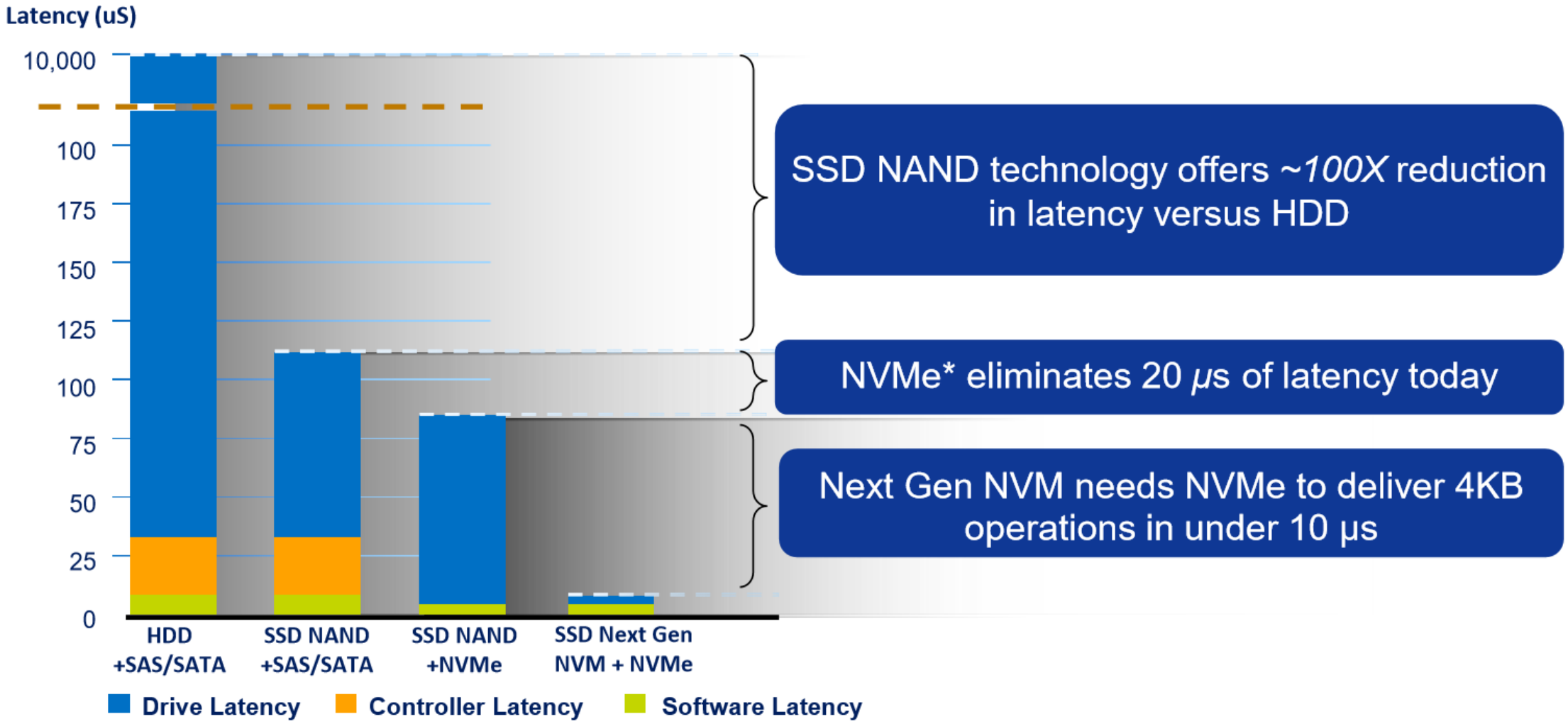


NVMe (NVM Express)

- The industry standard interface for high-performance NVM storage
  - NVMe 1.0 in 2011 by NVM Express Workgroup
  - NVMe 2.0 in 2021
- PCIe-based
- Lower latency
  - Direct connection to CPU
  - No HBA (Host Bus Adapter) required: reduced power and cost
- Scalable bandwidth
  - 1GB/s per lane (PCIe Gen3)
  - Up to 32 lanes



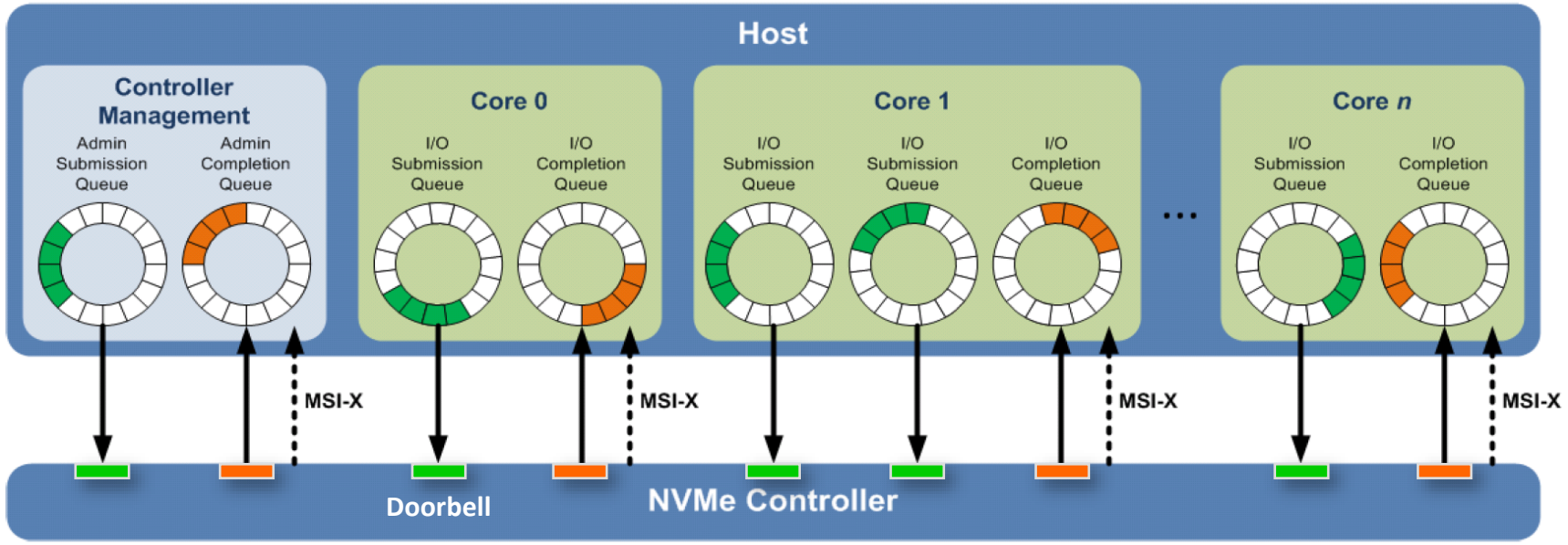
NVMe Benefits



A. Huffman, NVM Express Past, Present, and Future, FMS 2016.

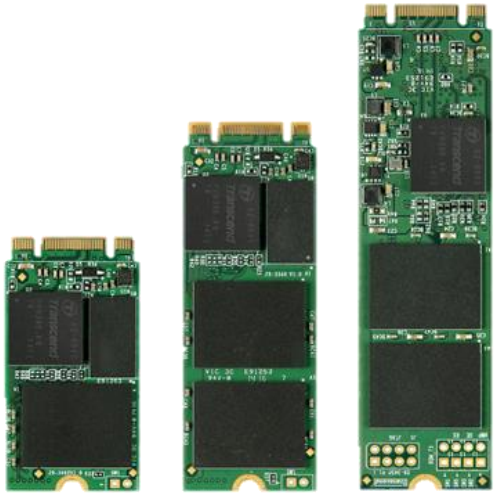
NVMe Overview

- Deep queue: 64K commands per queue, up to 64K queues
- Streamlined command set: only 13 required commands
- One register write to issue a command ("doorbell")
- Support for MSI-X and interrupt aggregation



NVMe SSD Form Factors

Add-in-card (AIC)

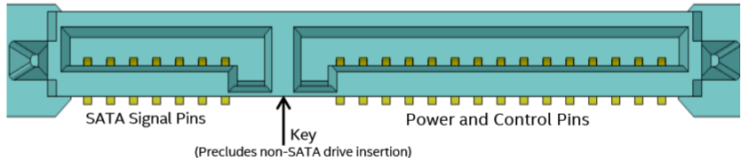


2242 2260 2280

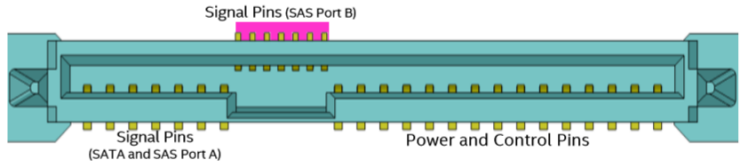
M.2 (PCIe: Up to x4)

U.2
(SFF-8639:
Up to x4)

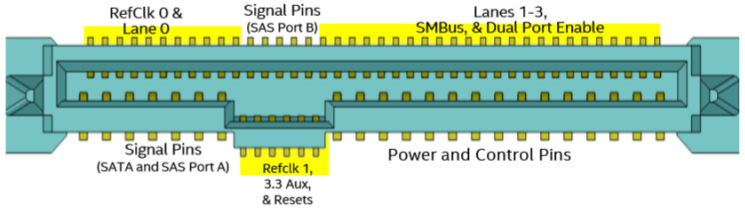
SATA



SAS



U.2



eMMC

- Embedded MultiMediaCard
  - Embedded storage solution with an MMC interface
  - JEDEC standard JESD84
  - Parallel, half-duplex interface with 1/4/8-bit data bus width
- eMMC evolution
  - eMMC 4.5: 1.6Gbps, 200MB/s, 2010 (Used in Galaxy S4)
  - eMMC 5.0: 3.2Gbps, 400MB/s, 2013 (Used in Galaxy S5)
  - eMMC 5.1: 3.2Gbps, 400MB/s, 2015
- Synchronous operation
  - One command at a time
  - Packed command (4.5+), Command queuing (5.1, up to 32)

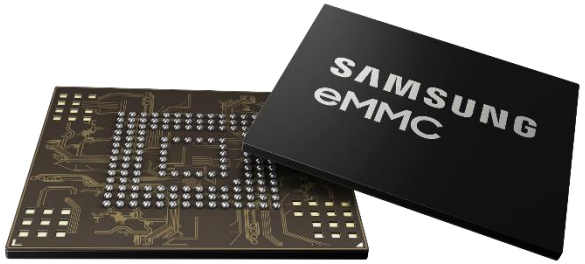
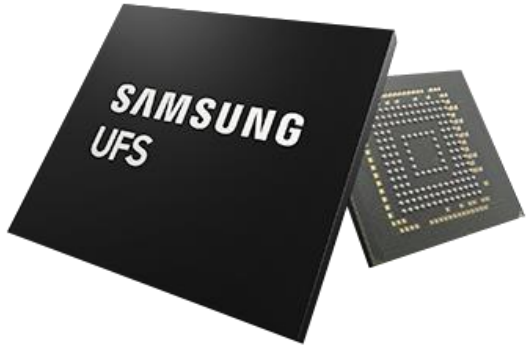


Image from https://news.samsung.com/

UFS

- Universal Flash Storage
  - Next generation flash storage for mobile devices
  - JEDEC standard JESD220
  - High-speed, full-duplex, serial interface
  - Based on SCSI command set
- UFS evolution
  - UFS 1.0: 150MB/s, single lane, 2011
  - UFS 2.0: 600MB/s, x2 lanes, 2013 (Used in Galaxy S6)
  - UFS 3.1: 1450MB/s, x2 lanes, 2020 (Used in Galaxy S21)
- Asynchronous operation
  - Higher random IOPS due to command queuing (up to 256)



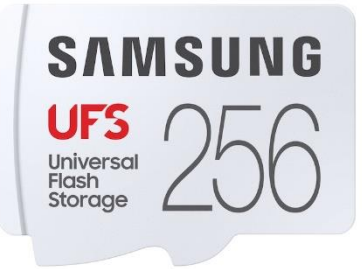


Image from https://www.samsung.com/

Summary

Transactions, eCommerce

Enterprise

Cloud

Data Center

Desktop PCs, Laptops

Consumer

Smartphones, Tablets

Mobile

?

NVM

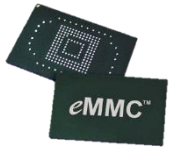
NVMe

SAS

SATA

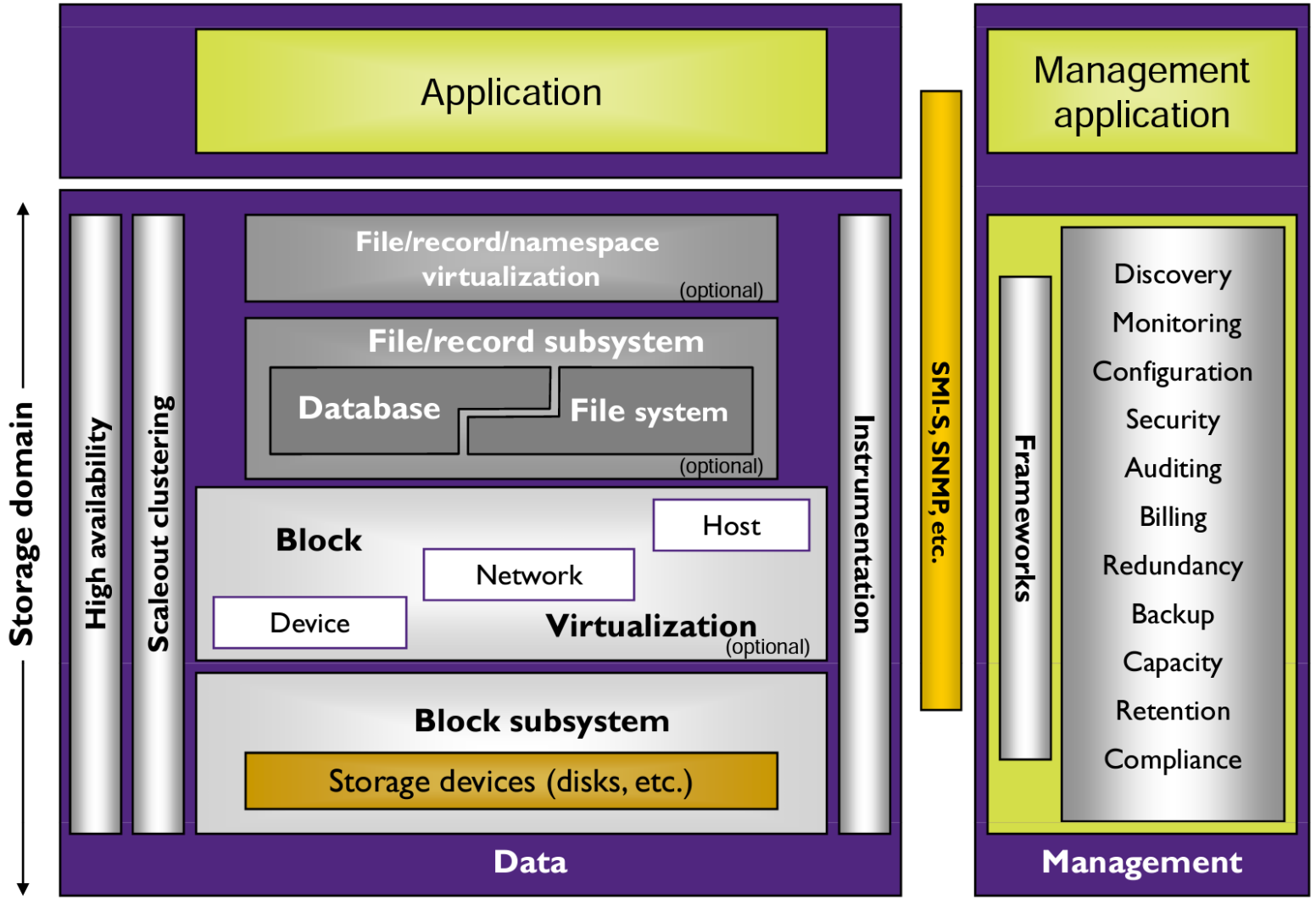
UFS

eMMC



SNIA Shared Storage Model

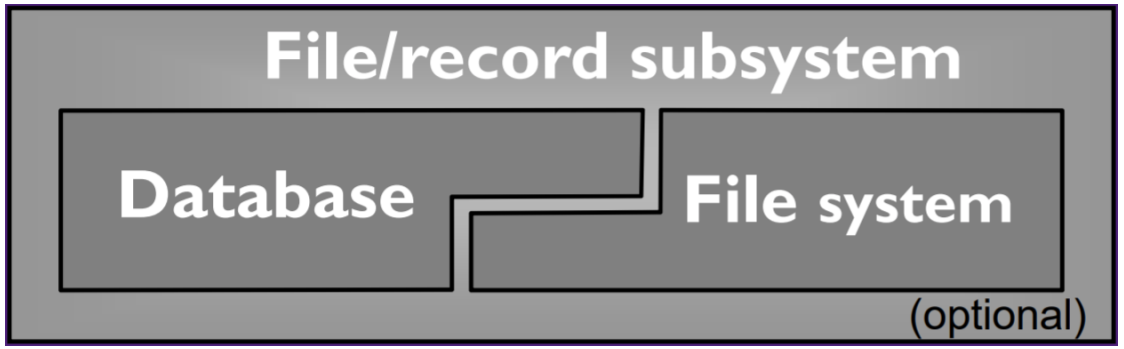
SNIA Shared Storage Model (SSM)



Source: https://www.snia.org/education/storage_networking_primer/shared_storage_model

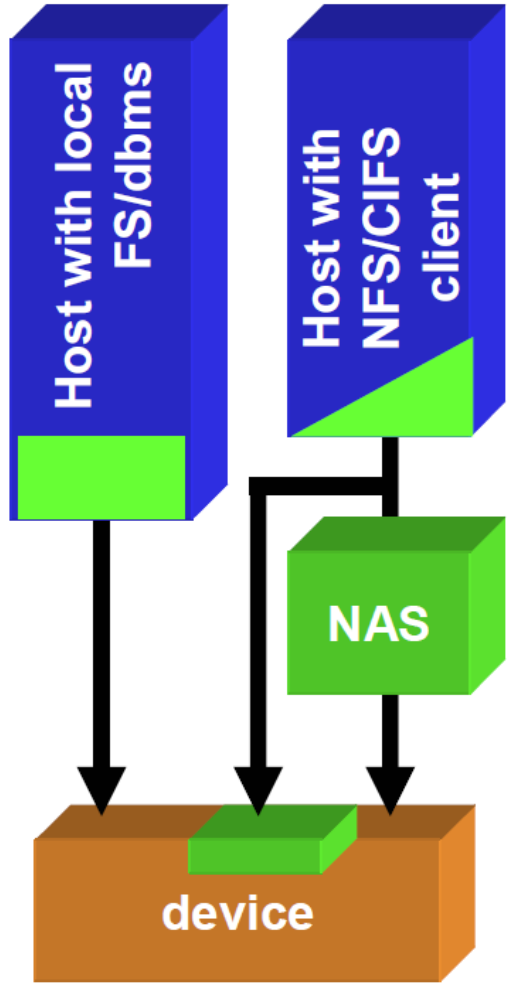
File/Record Layer

- "Access methods"
  - File system, DBMSes
- Primary responsibility
  - Fine-grain naming & indexing
  - Space allocation and clustering
  - Protection, etc.
- Secondary responsibility
  - Caching for performance
  - Coherency in distributed systems



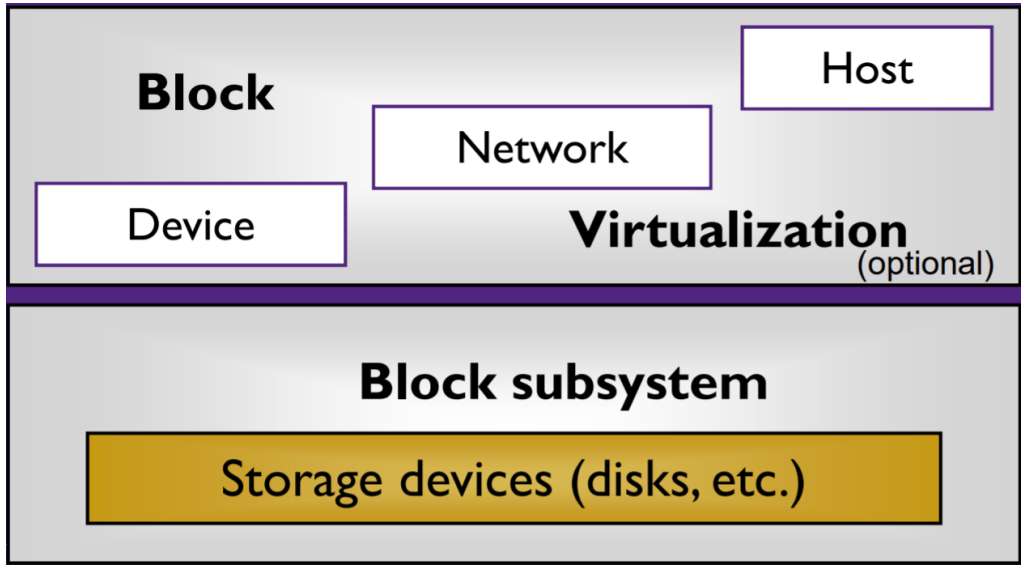
File/Record Layer: Where?

- Solely in the host
  - Traditional host-based file systems and DBMSes
- In both client and server
  - Split their functions between the client (host) and the server system (e.g., network file systems: NFS, CIFS, etc.)
  - File (database) server: a host with locally attached block storage device
  - NAS head: a dedicated-function computer acting as a file server and relying on external block storage devices
  - Storage device: disk array or "smart disk"



Block Layer

- Primary responsibility
  - Providing low-level storage to higher layers with an access interface that supports one or more linear vectors of fixed-size blocks (e.g. SCSI Logical Units (LUs))
- Secondary responsibility
  - Caching
  - Tiering
- "Native" storage devices
  - Disk drives, SSDs, tape drives, ...
- Block aggregation
  - Aggregation or "virtualization"



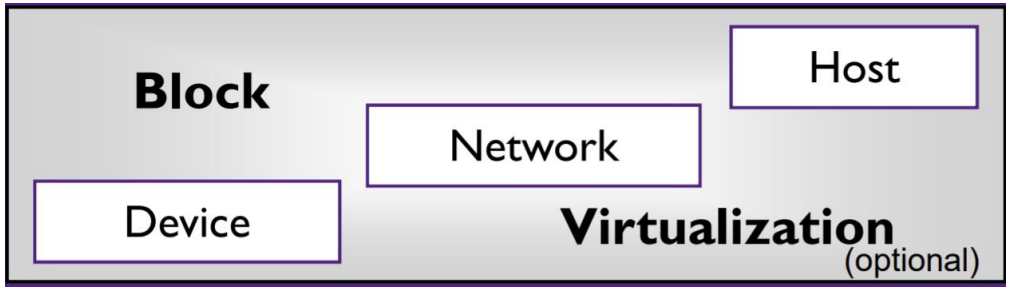
Block Aggregation

- Space management
  - Making a large block vector from many smaller ones
  - Packing many small block vectors into a large one
- Striping
  - For performance (load balancing, throughput)
- Redundancy
  - Full: local & remote mirroring, RAID-1/10, ...
  - Partial: RAID-3/4/5, ...
  - Snapshots

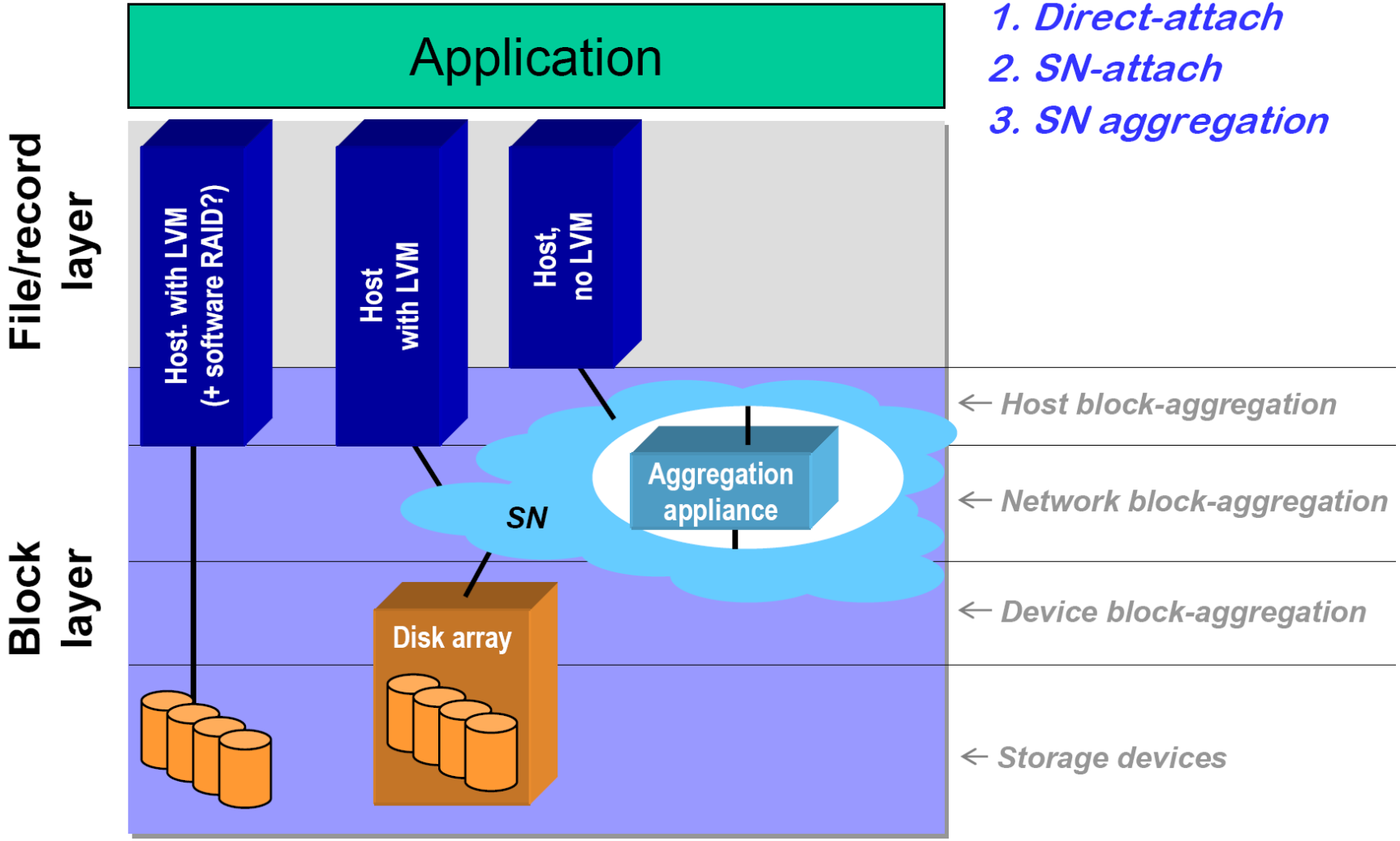
Block Aggregation: Where?

- Host-side
  - Logical Volume Managers (LVMs)
    - Mapping between logical volume and physical volume (linear, striped, ...)
    - Resizing a logical volume, snapshot support, ...
  - Software RAIDs, device drivers, HBAs, ...
- Storage Network (SN)-based
  - Specialized SN appliances
- Device-based
  - Disk arrays or Flash arrays
  - RAID controllers
  - Disk controllers

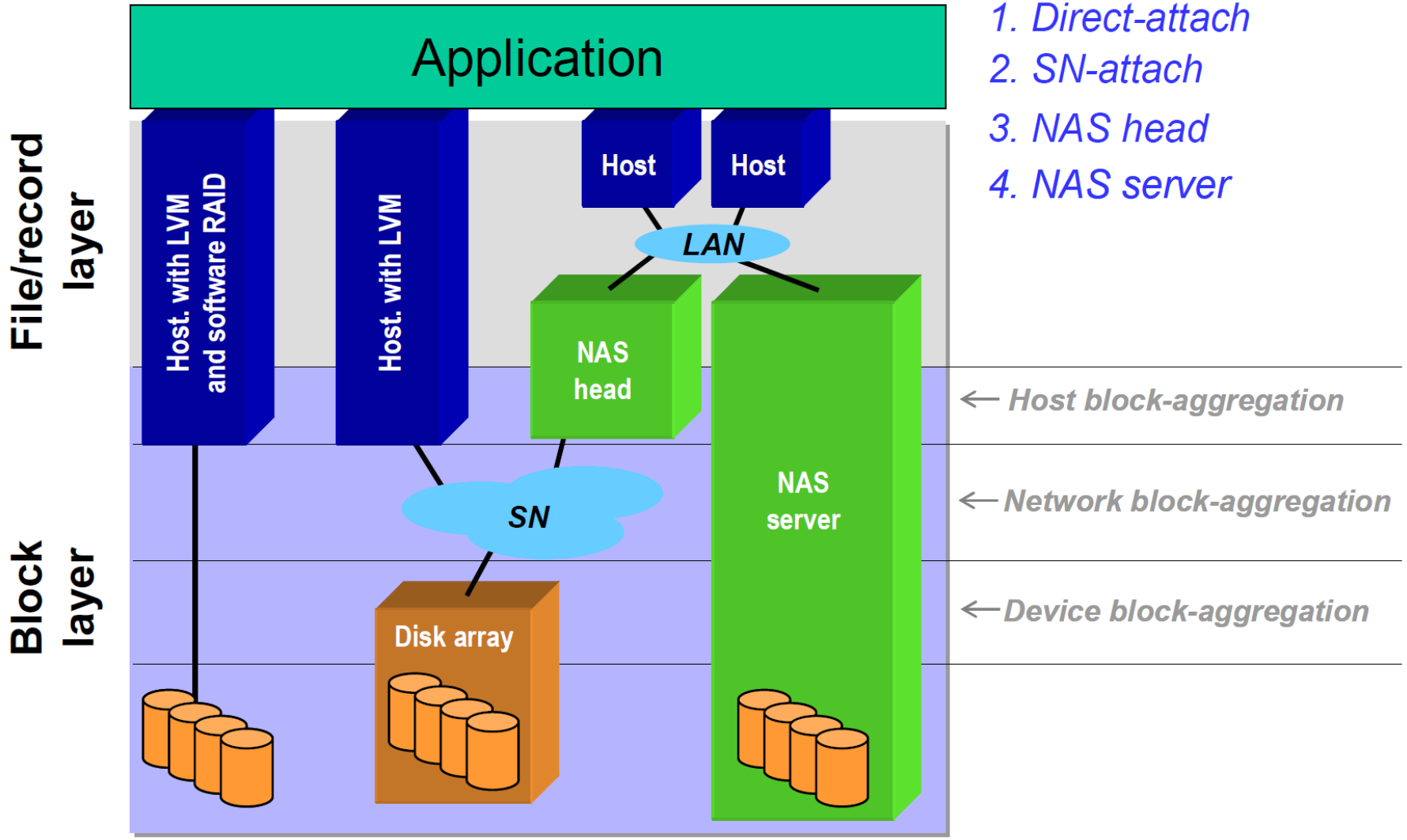
Storage network(SN):
any dedicated network installed for storage traffic (Fibre channel, Ethernet, etc.)



Block-based Architecture



File-based + Block-based Architecture



Summary: Access Paths