Dr. Stefan Fütterling – The Open Group Distinguished Architect – stefan.fuetterling@capgemini.com
Martin Bachmaier – IT Versatilist, Lenovo R&D – mbachmaier@lenovo.com
Daniel Amor – The Open Group Master Architect – danny@dxc.com

Duale Hochschule Baden-Württemberg Stuttgart
IT Architekturen, 2025-11

Teil 3

Central storage

The big picture

From
… laptop with local storage
… to laptop with external storage (USB drives)
… to laptop or desktop with Network Storage (Synology, QNAP, …)

… to server with local storage
… to server(s) providing such **Network Storage** (which is inside this server)
… to server(s) using **external, central storage** (physically separated, connection via dedicated storage network)

Zentralisierter Storage

3.1 Architectures and basics

Intro
Network Attached Storage (NAS)
FiberChannel SAN (FC-SAN)
Ethernet-based SAN (IP-SAN and FCoE-SAN)

3.2 RAID levels

RAID 0, RAID 1, RAID 5, RAID 6, RAID 10

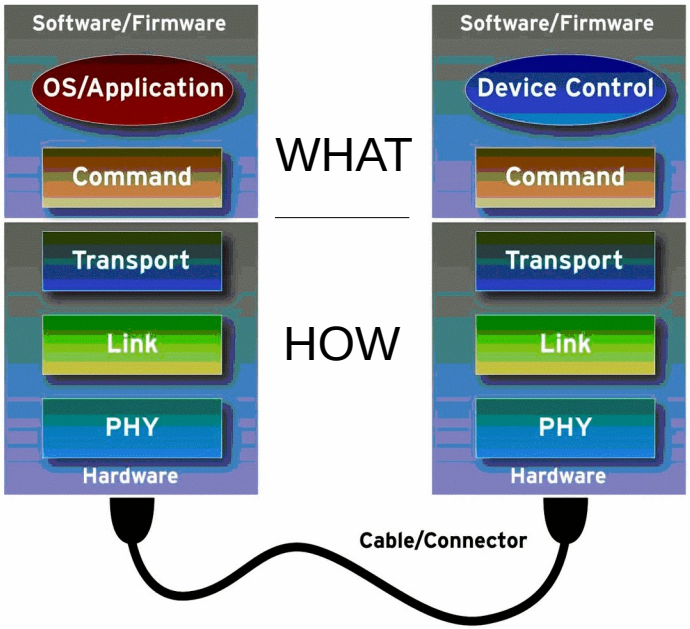
3.3 Enterprise storage subsystems

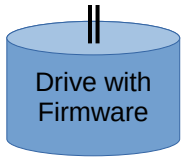
Example Storage Appliance (NetApp)
Enterprise Storage Features
Software-Defined Storage (SDS) → Foundation for Hyperconverged Infrastructure (HCI)

Intro: our approach today

- **Laptop local storage**
  - Used to be SATA
  - Nowadays NVMe/PCIe
- **Generally speaking: always two aspects**
  - Protocol (the WHAT)
  - Bus/interface/medium (the HOW)

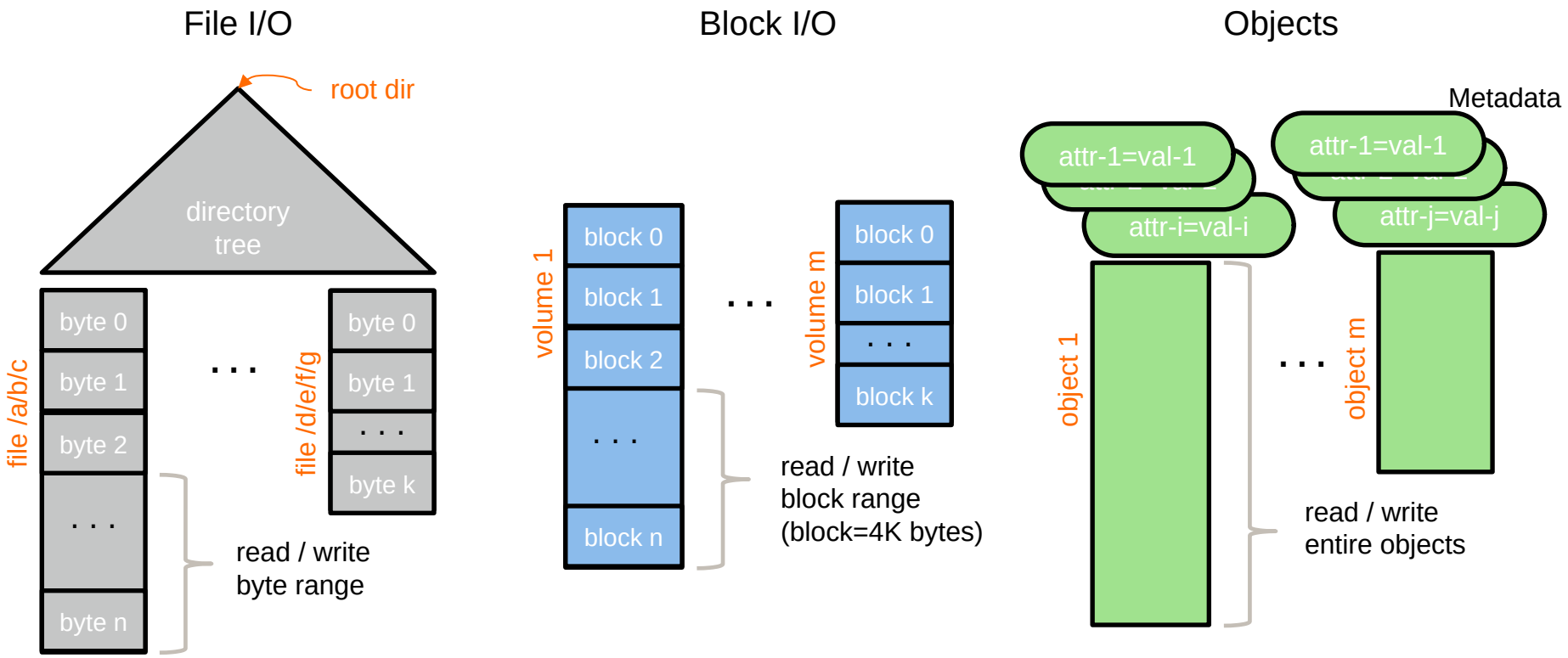
  Examples (protocol / interface):
  SATA / SATA
  SATA / USB
  NVMe / PCIe





Today's storage device protocols:
- SATA
- SCSI
- NVMe

Intro: three Storage types



Local file system
(Your laptop)

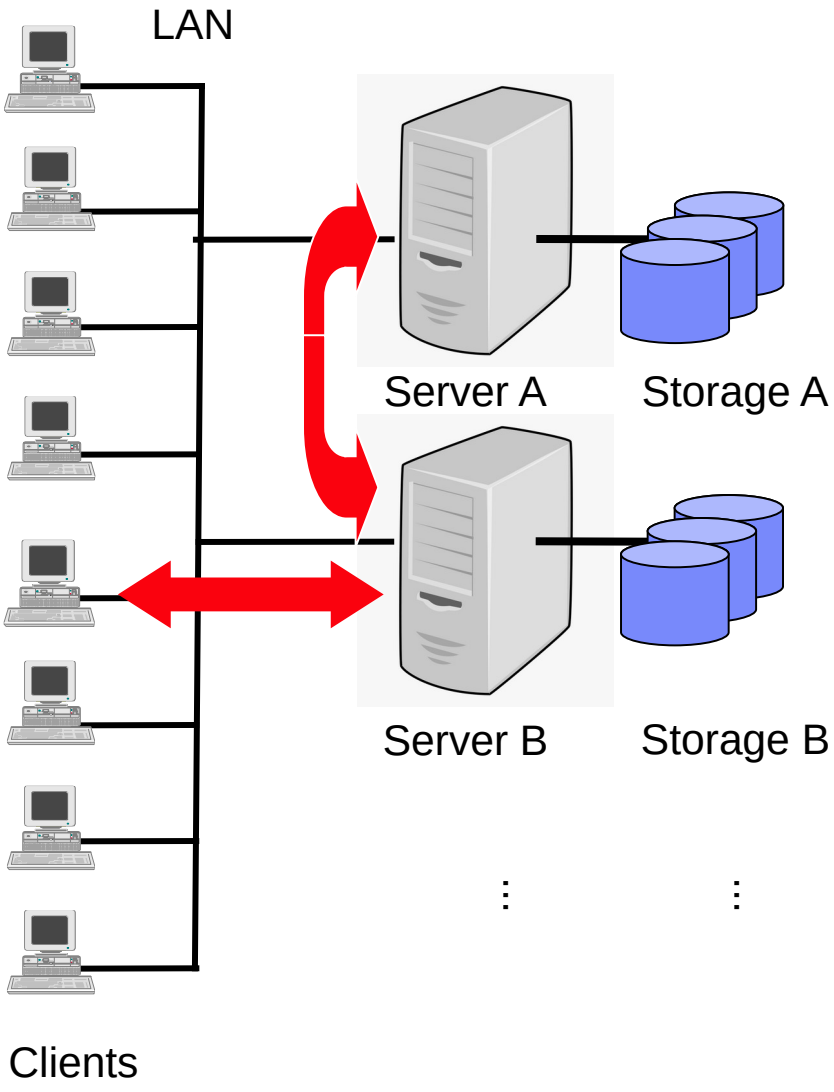
Local drive,
USB drive (via SCSI commands)

AWS S3
Azure Blob Storage

Intro: two design approaches

- Network Attached Storage (NAS)
  - Storage … accessible .. over the network
- Storage Area Network (SAN)
  - A dedicated storage network

Simple Client/Server using file servers (NAS)



Access to file server via common network (LAN/WLAN).
File server properties:

- Data access from multiple clients (Sharing)
- Dedicated storage space (usually big)
- Growth leads to information islands (by adding file servers)
- Difficult to implement backup and disaster recovery
- Performance bottlenecks
  - Network becomes the bottleneck because it competes for bandwidth with general LAN traffic
  - File server itself become the bottleneck as more and more clients access it
- Single point of failure

Network Attached Storage (NAS)

- **Network Attached Storage is central storage that can be accessed over the LAN**
- Conventional file server, modern NAS filer appliances
- Pro:
  - No need for a dedicated storage network
  - Cheap and easy to set up
- Contra:
  - Additional traffic on the LAN
  - TCP/IP overhead causes additional load on CPU for server and client
- Protocols:
  - **TCP/IP using NFS** for UNIX and Linux environments (Network file system)
  - **TCP/IP using CIFS** for Windows environments (Common Internet file system)
- Access (from consumer perspective) on file level: **File I/O**

Storage Area Network (SAN)

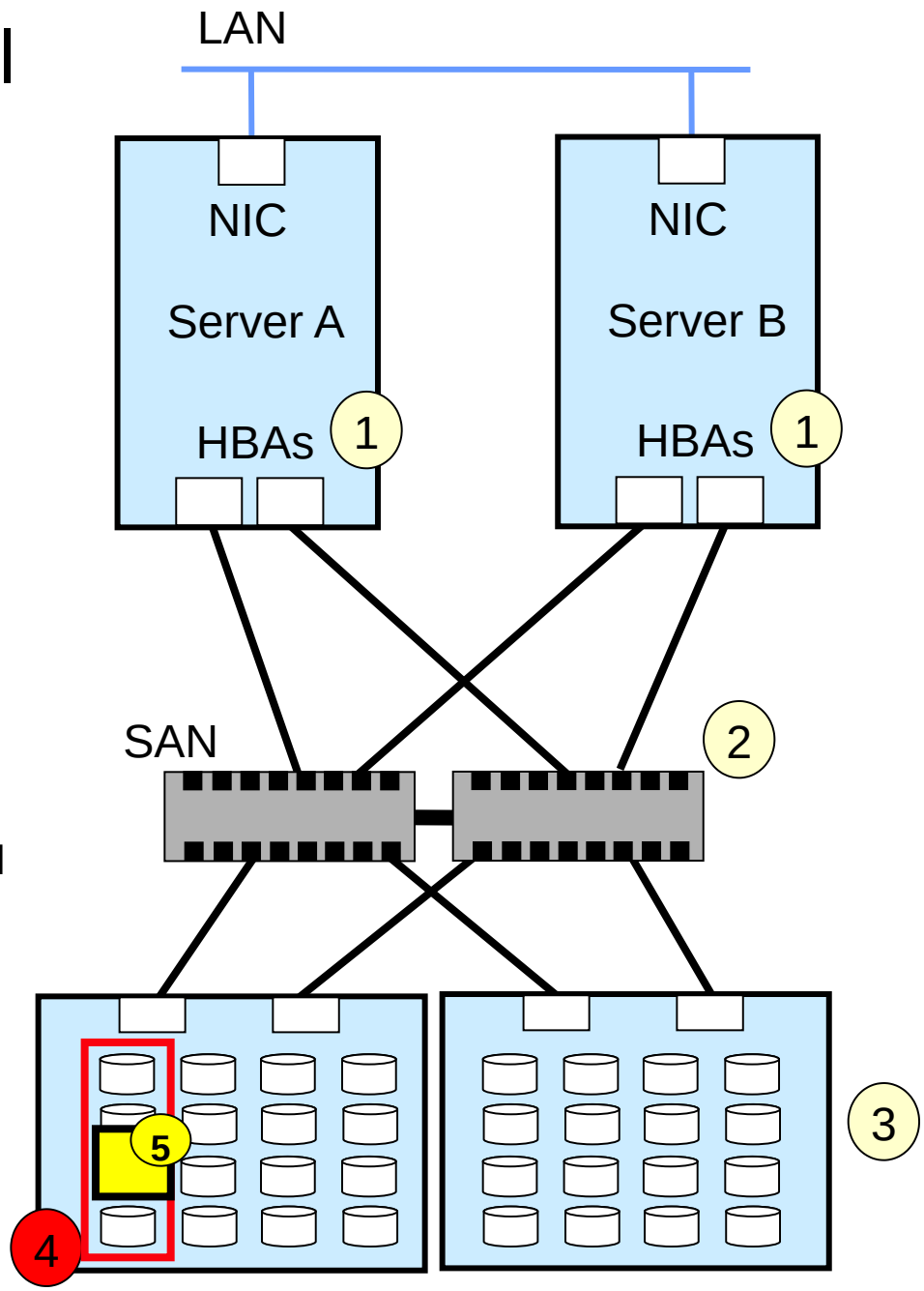
- **Definition of the “Storage Networking Industry Association” (SNIA)**
  - A storage area network is any high-performance network whose primary purpose is to enable storage devices to communicate with computer systems and with each other.
  - The type of interconnect is intentionally not specified
- **Dedicated storage network** (different from LAN)
- Different types of Storage Area Network:
  - **FC-SAN**: FibreChannel transport network using FC protocol
  - **IP-SAN**: Ethernet transport network using TCP/IP + iSCSI protocol
  - **FCoE-SAN:** Ethernet transport network using FibreChannel over Ethernet (FCoE) prot.
- Access (from consumer perspective) on block level: **Block I/O**

Storage Area Network in detail

- Components:
  - **Host Bus Adapter (HBA)** (1)
    (FibreChannel or Ethernet HBA)
  - **Cabling and SAN switches** (2)
    (FibreChannel or Ethernet)
  - **Disk subsystem** (3)
    - holds disks
    - provides access to storage (via abstraction layer)
    - implements features in the abstraction layer (RAID, Security, Encryption, …)

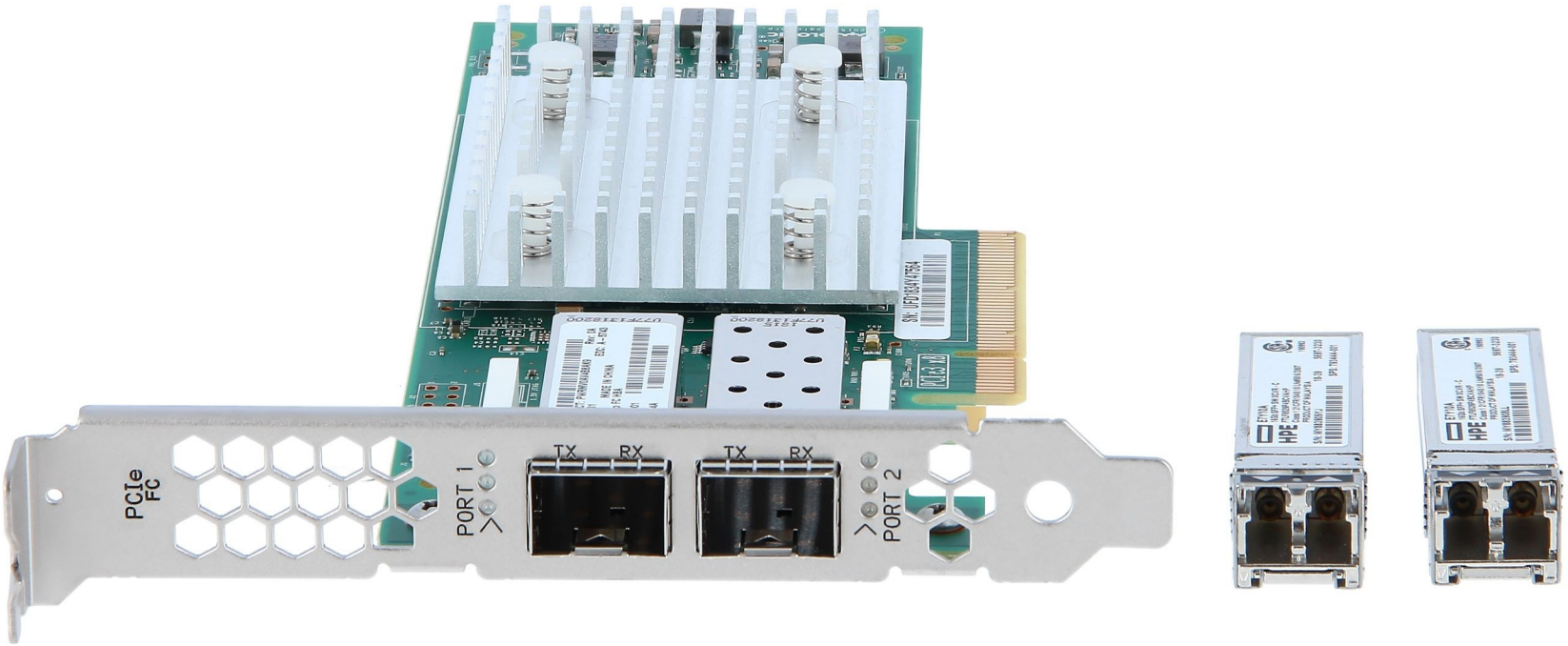
- Disk subsystem abstraction layer:
  - Disk drives are grouped into **Disk Arrays** (4)
  - Arrays are assigned **RAID Levels**
  - Definition of logical storage space within those disk arrays as **virtual Disk (vDisk)** or **Logical Unit (LU)** using numbers (LUN) (5)

- Mapping of LUNs to server („zoning“):
  - LUN is mapped to one or more server using unique IDs
  - Content of LUNs is usually a file system (ext4, NTFS, ...)



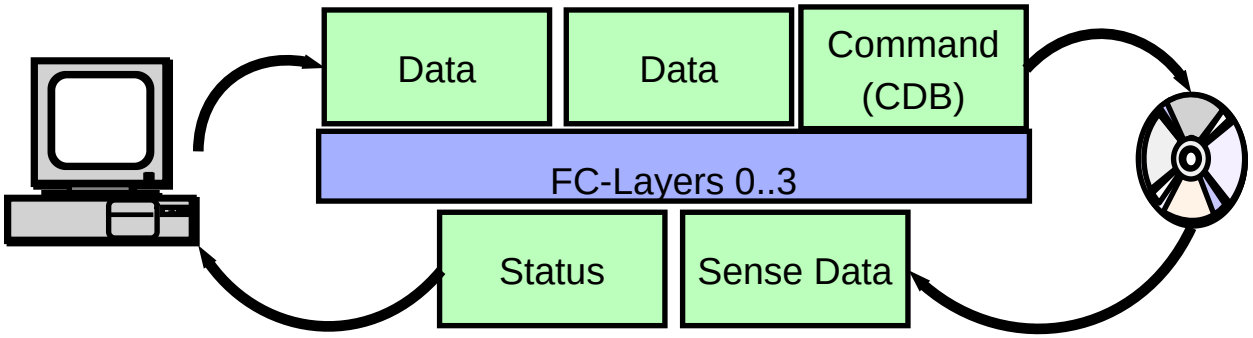
FibreChannel standard

- **ANSI standard since 1995**
- **Properties**
  - Low CPU consumption on the server (processed on the HBA)
  - Connection-oriented
  - Addressing via 64-bit identifier (World Wide Port Name, WWPN)
  - Definition of SAN switches (allows for a fabric topology)
  - Block I/O
  - Access control via zoning (mapping of LUNs to server HBAs)
- **FC protocol (FCP)**
  - FC protocol is SCSI compatible
  - easy integration into existing SCSI protocol stack
- **Performance**
  - Currently used speeds: mostly 32 Gbit/s bidirectional
    (32 Gbit/s products on the market since 2016, first 64 Gbit/s HBA 2020)
  - Achievable net throughput ~ 1600 MByte/s (for 16 Gb/s fabric)

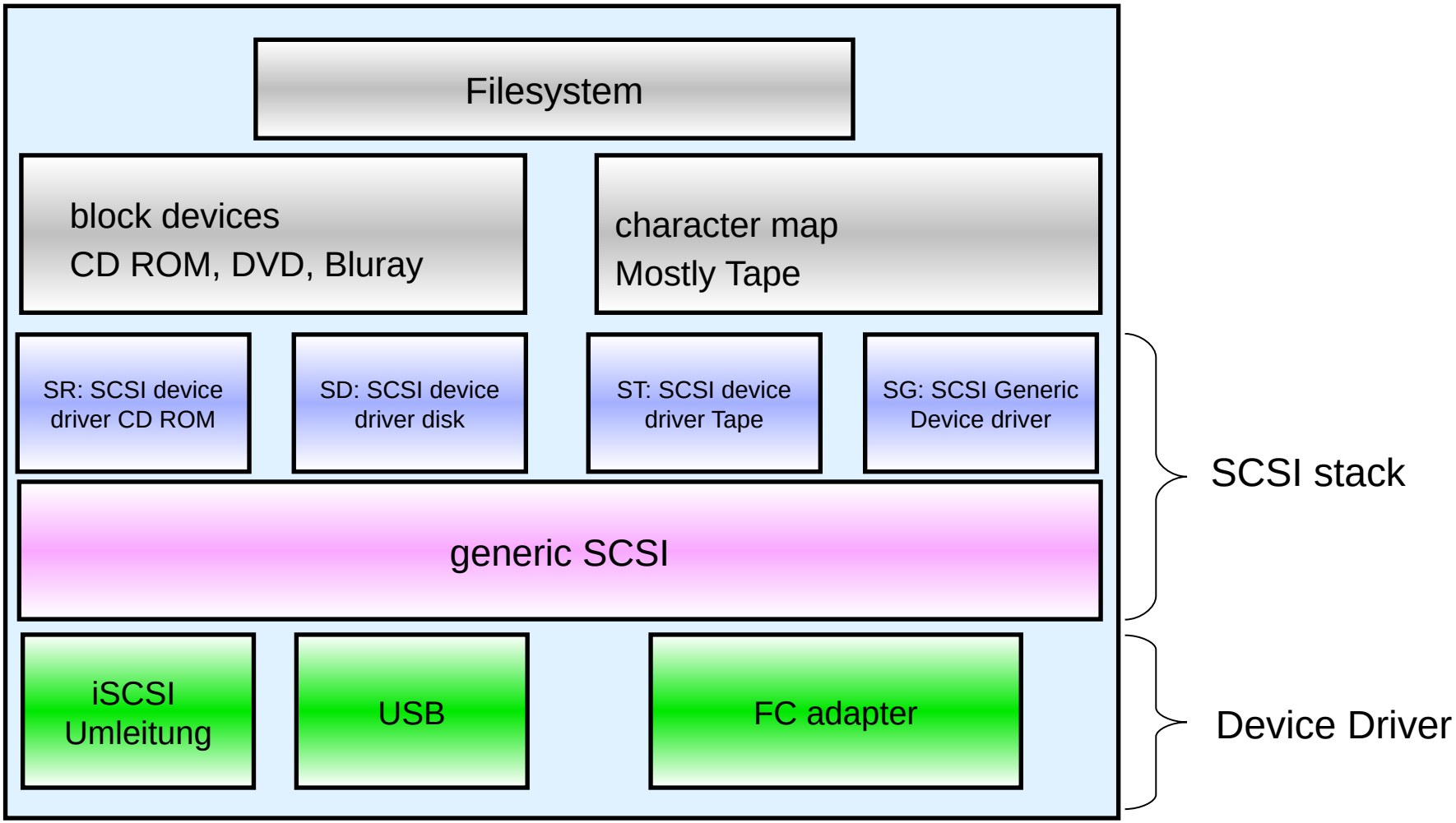

PCIe
FC
PORT 1
TX
RX
TX
RX
PORT 2

FibreChannel Protocol (FCP) layers

FC-4 Mapping	**Interface to Upper Level Protocol**	FCP (SCSI) / FICON (SB-2) / IPFC
FC-3	**Common Services – Not implemented!**	Multicast, Striping, Compression...
FC-2 Signaling Protocol	**Framing Protocol, Flow Control**	Packet Segmentation Flow Control Classes of Service
FC-1 Transmission Protocol	**Encode / Decode, Link Control**	Data Transport
FC-0 Interface / Media	**Physical Characteristics** Single Mode Fibre / Multimode Fibre / Copper	Cable, plug, signal encoding

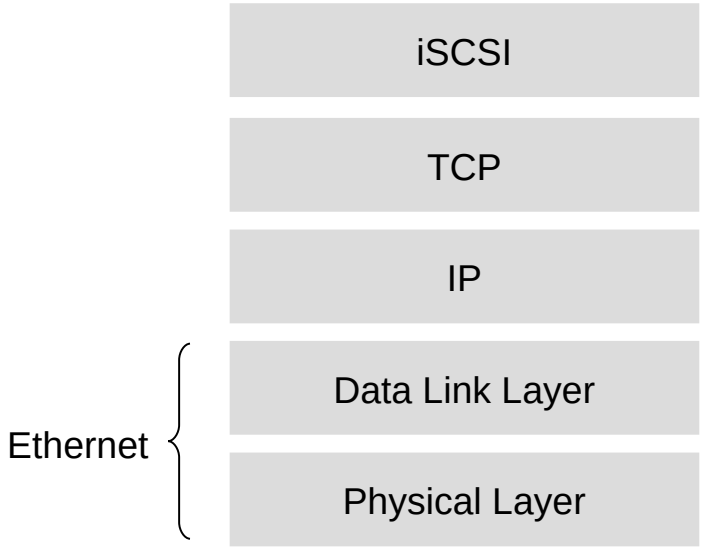


Example: how FC integrates with the Linux stack



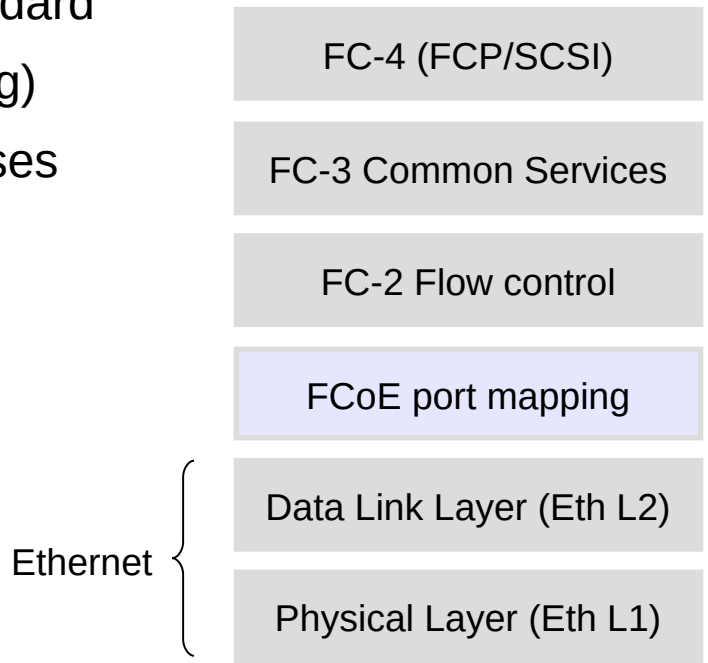
IP-SAN using the iSCSI protocol

- iSCSI = “Internet SCSI”
- iSCSI sits on top of TCP/IP like an application layer protocol
- SCSI I/O requests are encapsulated in TCP/IP packets and sent to the iSCSI target
- iSCSI packets are received from the target TCP/IP stack and decapsulated. It is then recognized as an iSCSI packet and handed over to the iSCSI stack. Finally, the packet is handed over to the generic Linux SCSI driver to handle the IO
- iSCSI packets can be routed (and firewalled..)



FCoE-SAN

- FCoE = FibreChannel over Ethernet
- Switches FCP protocol via Ethernet frames (not TCP/IP frames)
- Ultimate goal is convergence of existing data center networks – LAN and SAN – to one single physical infrastructure
- FC-packets are encapsulated in Ethernet frames → higher throughput than TCP/IP+iSCSI but not routable
- FCoE requires extension to conventional Ethernet standard
  - Lossless Ethernet fabric (Data Center Bridging)
  - Mapping between FC-ports and MAC addresses
  - Defined only for 10 Gbit Ethernet and above

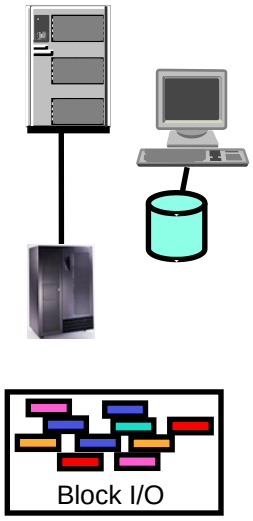


DAS, NAS, SAN

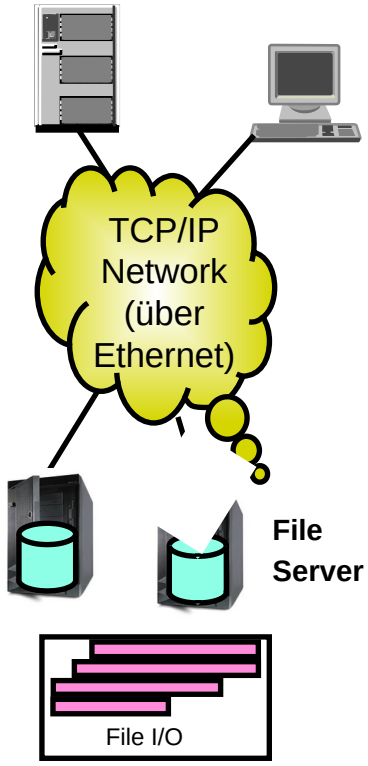
Storage dedicated, non-shared

Storage shared

Direct Attached Storage (DAS)

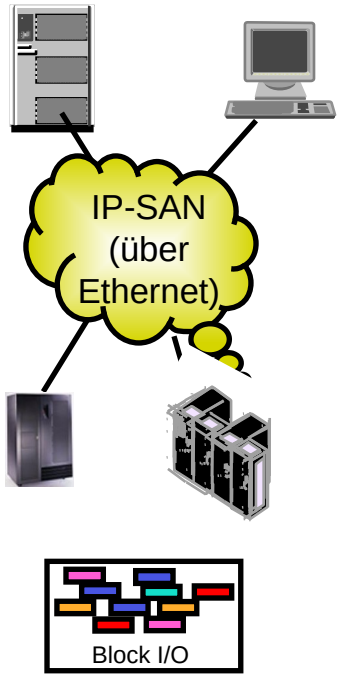


Network-Attached Storage (NAS)

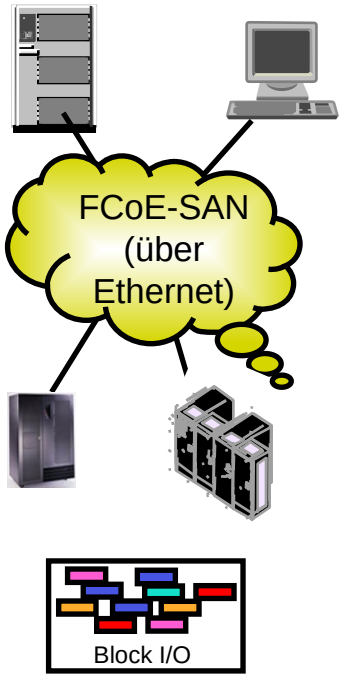


Based on distributed file systems (NFS, CIFS, ..)

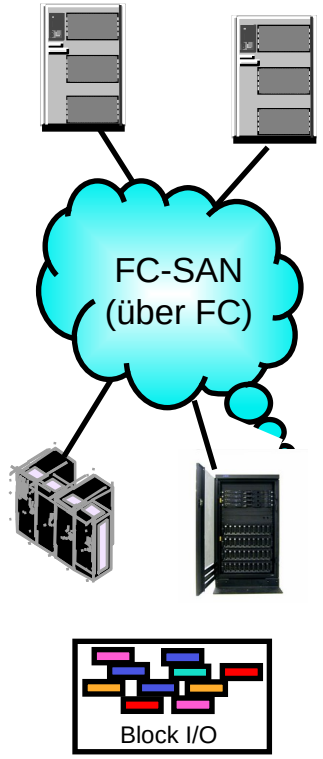
IP-SAN (iSCSI Protocol)



FCoE-SAN (FiberChannel over Ethernet)



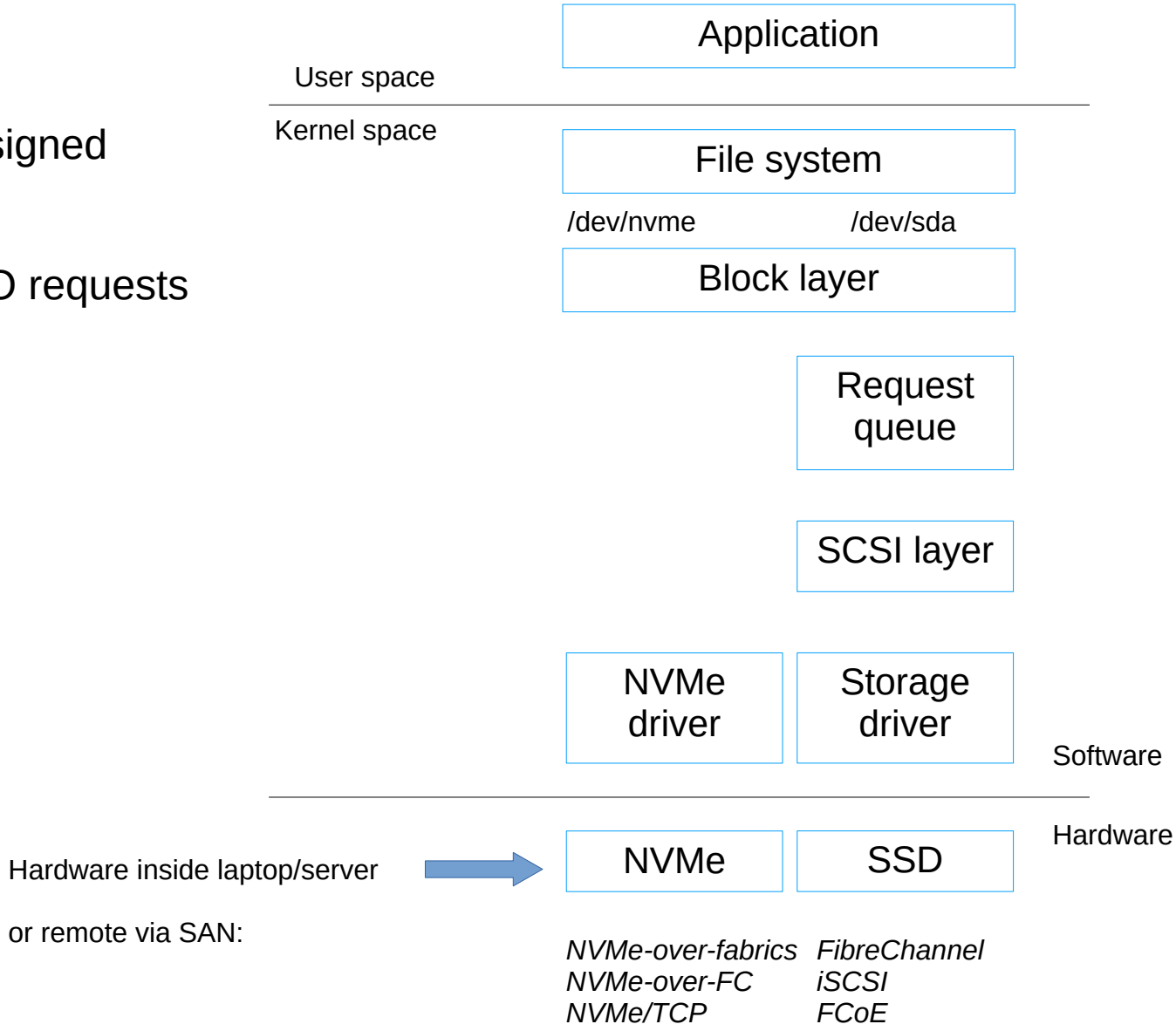
FC-SAN (FC Protocol)

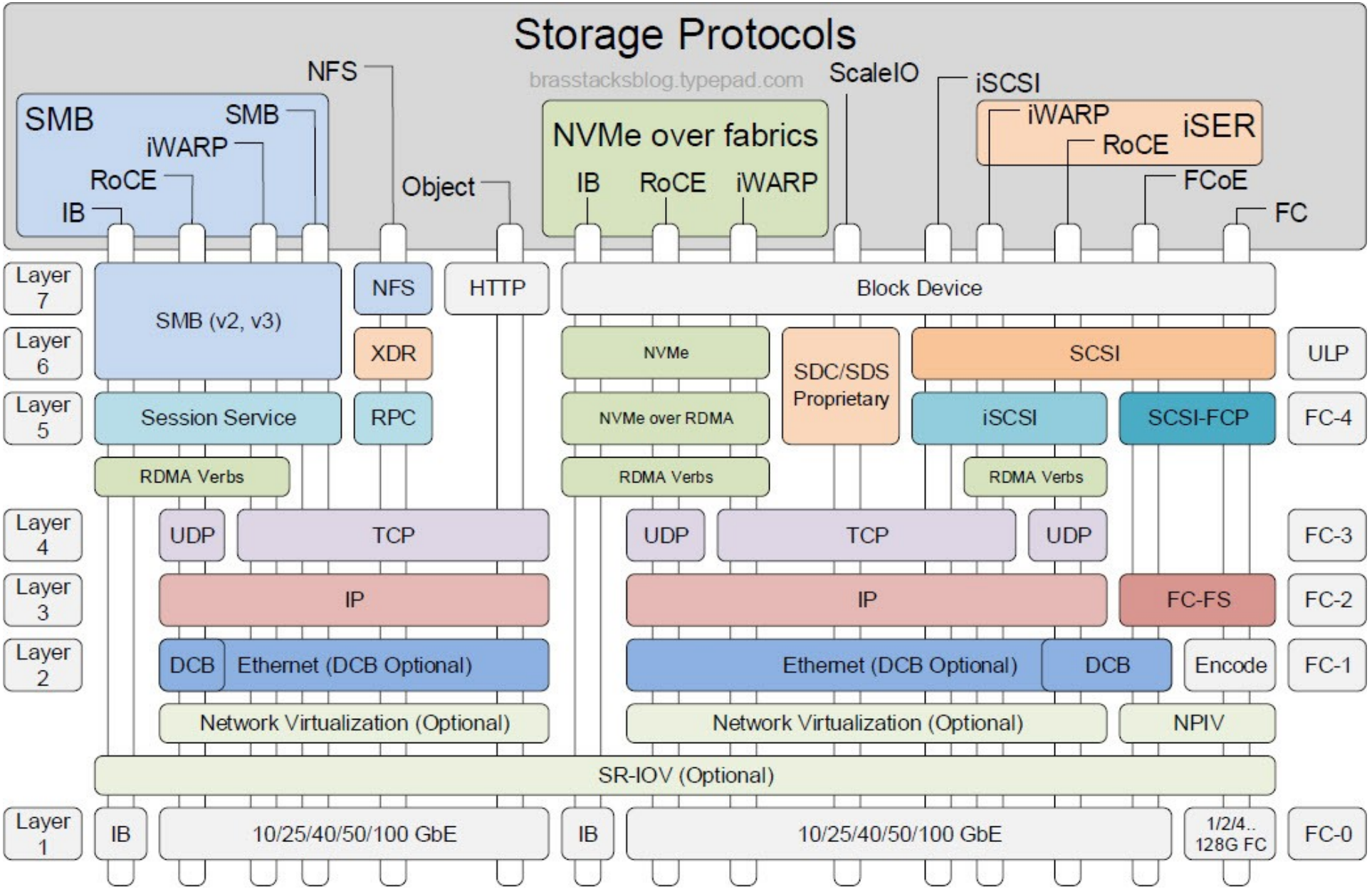


NVMe and why it is great

Storage protocol that is designed from the ground up for

- high throughput
- high number of parallel IO requests
- low latency





Source 2016 Dell-EMC presentation at https://brasstacksblog.typepad.com/.a/6a013487ed263f970c01b7c877d4db970b-pi

Zusammenfassung DAS, NAS, SAN

Storage Typ	Technologie	Zugriff	Vor- und Nachteile
Direct Attached Storage (DAS)	Lokale Festplatten z.B. SAS oder S-ATA	Block I/O	☑ geringe Investitionskosten ☑ gute Performance ☒ geringe Skalierbarkeit, oft geringe Auslastung ☒ hohe Administrationskosten
Network Attached Storage (NAS)	Kommunikations-netz TCP/IP mit NFS / CIFS	File I/O	☑ flexible Einsatzmöglichkeiten, gute Skalierbarkeit ☑ zentrales Speichermanagement, z.B. Backup ☑ vorhandenes Kommunikationsnetzwerk ☒ geringe Performance, z.B. für Datenbankanwendungen
IP-basiertes Storage Area Network (IP-SAN)	Speichernetzwerk auf Ethernet Basis TCP/IP mit iSCSI Protokoll	Block I/O	☑ gute Skalierbarkeit und Ausfallsicherheit ☑ zentrales Speichermanagement, z.B. Backup ☒ mäßige Performance ☒ zusätzliche Investitionskosten (Ethernet HW)
FCoE Storage Area Network (FCoE-SAN)	Speichernetzwerk auf Ethernet Basis FCoE Protokoll	Block I/O	☑ gute Skalierbarkeit und Ausfallsicherheit ☑ zentrales Speichermanagement, z.B. Backup ☑ gute Performance (höhere Transportleistung als IP-SAN) ☒ zusätzliche Investitionskosten (Ethernet HW)
FiberChannel Storage Area Network (FC-SAN)	Speichernetzwerk auf FiberChannel Basis FCP Protokoll	Block I/O	☑ gute Skalierbarkeit und Ausfallsicherheit ☑ zentrales Speichermanagement, z.B. Backup ☑ sehr hohe Performance ☒ sehr hohe Investitionskosten

3.1 Architectures and basics

- Intro
- Network Attached Storage (NAS)
- FiberChannel SAN (FC-SAN)
- Ethernet-based SAN (IP-SAN and FCoE-SAN)

3.2 RAID levels

- RAID 0, RAID 1, RAID 5, RAID 6, RAID 10

3.3 Enterprise storage subsystems

- Example Storage Appliance (NetApp)
- Enterprise Storage Features
- Software-Defined Storage (SDS)

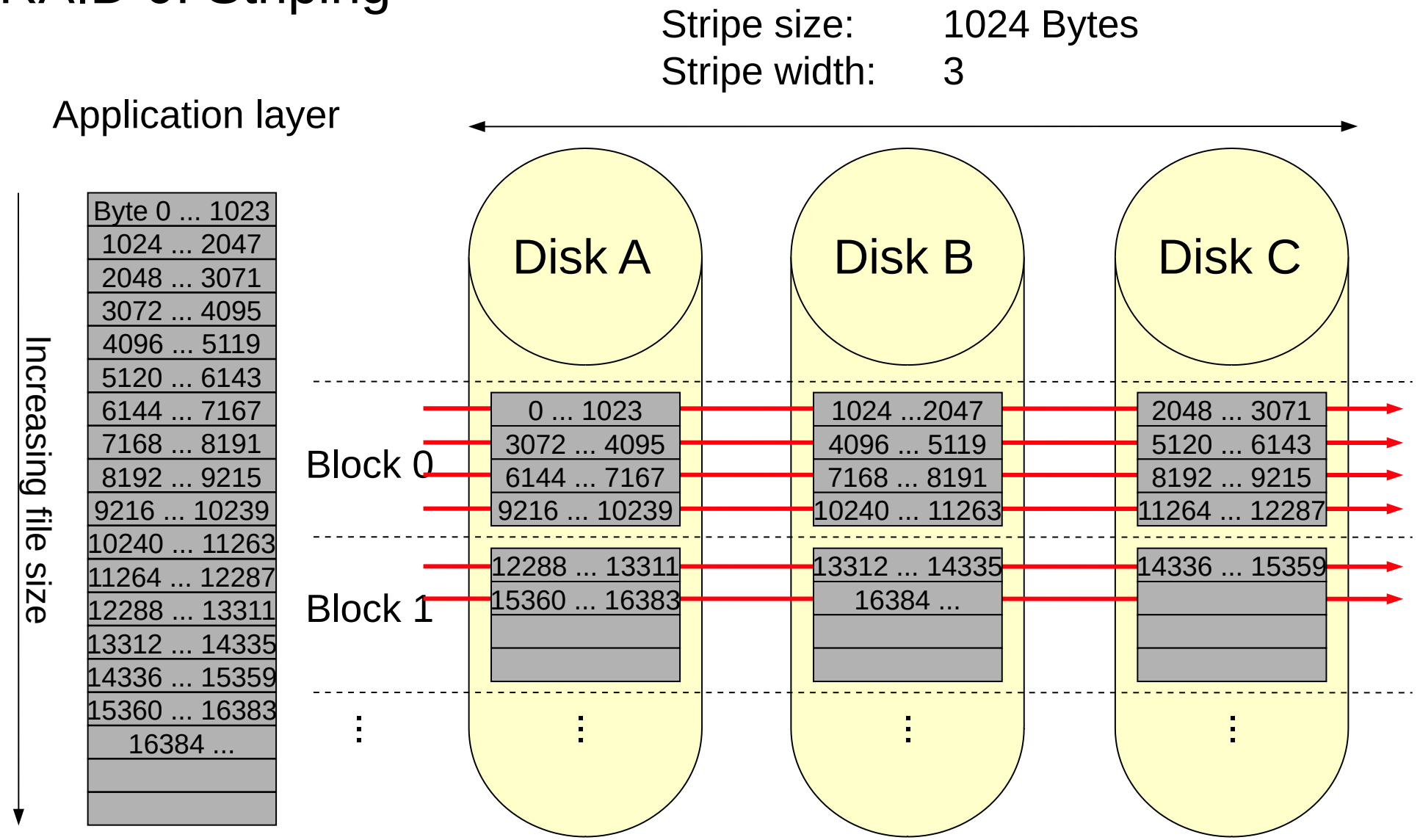
Storage Systems - RAID

- Enterprise storage systems contain conventional hard disk drives or flash devices
- Hard disk drives have a specified mean time to failure (MTTF)
- Introduction of **RAID (Redundant Array of Independent Disks)** concepts with the following goals:
  - Allow for redundancy to increase availability
  - Reduce file access time via parallel access to multiple disks
  - Allow for different combinations of these goals
  - RAID can be implemented in software (virtual disk arrays) or in hardware (physical disk arrays that appear as one logical disk arrays towards the operating system)
  - Hardware RAID is transparent to operating systems and provides certain benefits (calculation offloading, easy to replace, ..)

RAID 0: Striping

- Distribute data over multiple disks (ie. multiple disk access arms)
- No redundancy but parallelization
- Can achieve great read and write troughput (depending on stripe width and stripe size)
- Minimum if two disks required (stripe width=2), maximum number of disks (stripe width=N) depends on HW RAID controller
- Stripe size varies from 1 KB up to 1 MB and is user selectable. Granularity depends on controller or operating system support
- Use disks of identical size only – otherwise you waste disk space
- Capacity = (size of smallest disk in the array) * (number of disks)

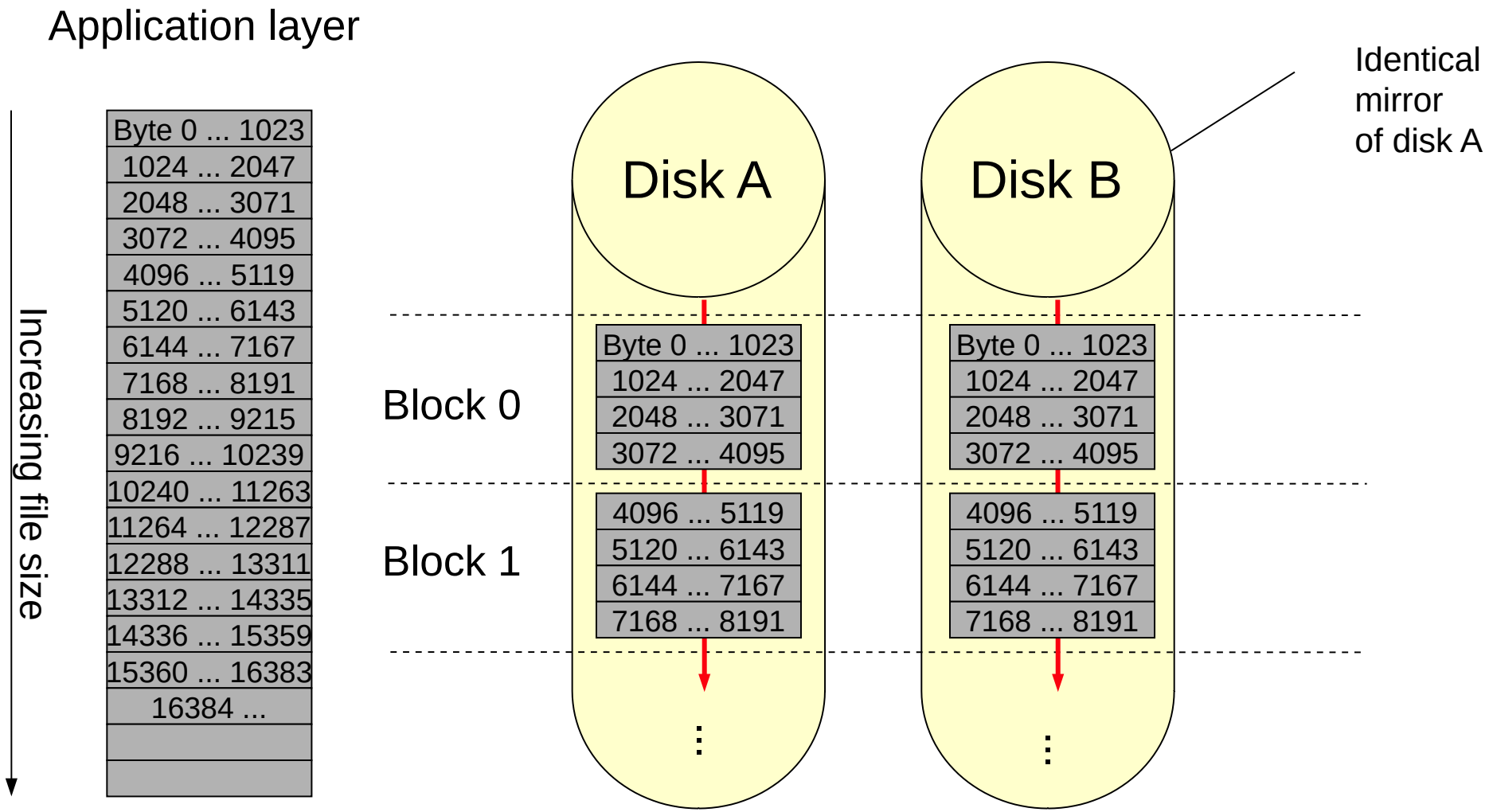
RAID 0: Striping



RAID 1: Mirroring

- Data is written to two disks in parallel
  - A rare variant of mirroring is duplexing when you also use two RAID controller
- Full redundancy with very little impact in case of a disk failure
  - If a disks is dead the disk arrays runs in degraded mode – ie as a single disk
  - Most RAID controller support a hot spare disk that automatically replaces a broken disk and trigger an automatic rebuild of the array
  - Very fast rebuild, just copy the data. No parity calculation required
  - Read performance can be better than a single drive (chances are one disk delivers data a bit quicker then the other one) but not as good as other RAID levels
  - Write performance is good but less than a single drive (both disks must be written to). Faster than other RAID levels because no parity needs to be calculated
- Requires pairs of disk drives
  - Efficiency 50%
  - Capacity = size of the smaller disk

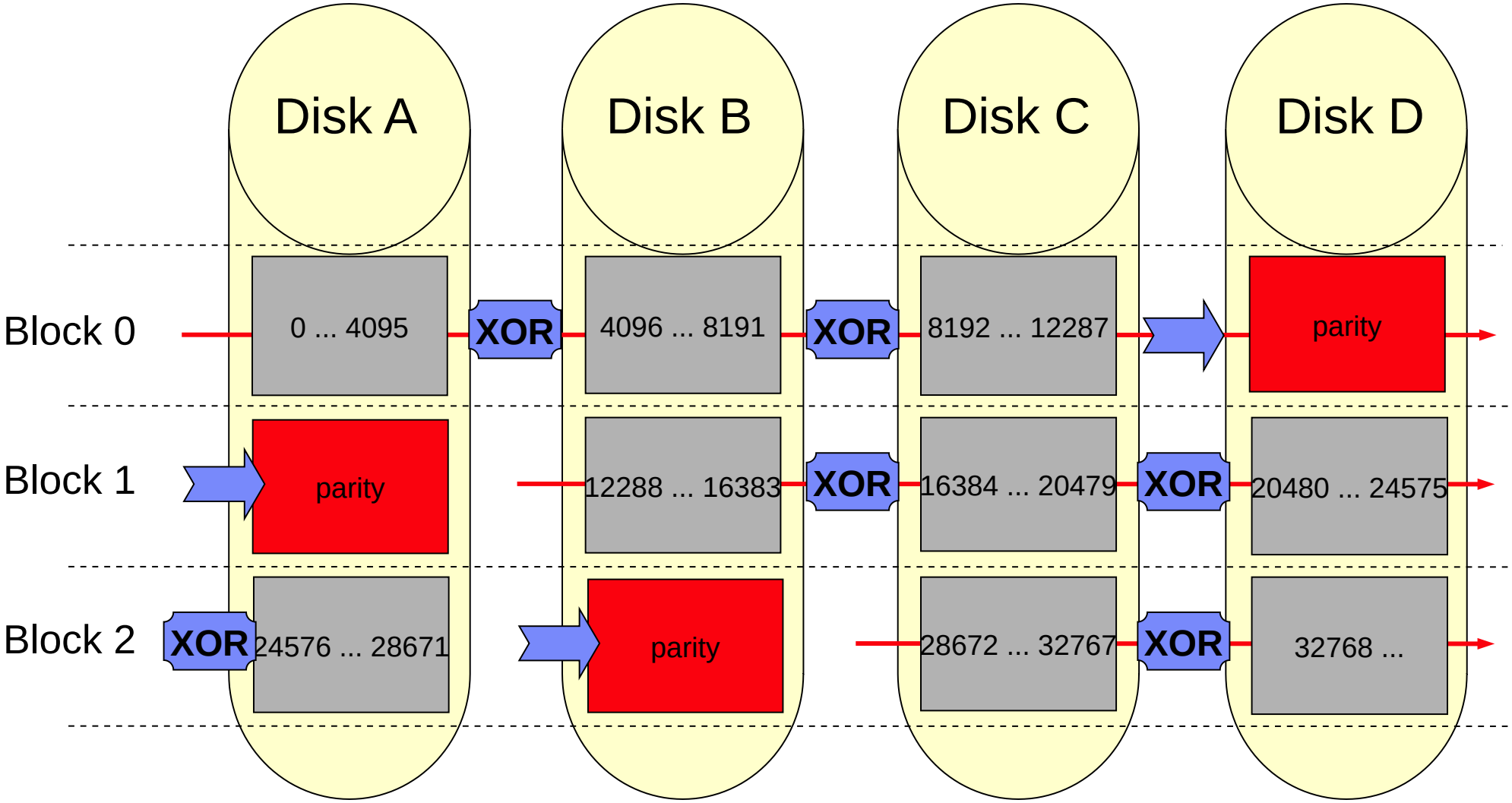
RAID 1: Mirroring



RAID 5: Striping with distributed parity

- Very popoluar and very often used in enterprise environments
- Stripe size in the range of 1KB to 1 MB. Parity is calculated with an XOR operation over the disk stripes
- Parity information is distributed over all participating disks
- In case of a disk outage the data can be recovered from the remaining data plus parity information (quite time consuming)
- Full redundancy but with a performance penalty
  - Parity calculation for every write IO operation required
  - Big sequential writes: minor impact because all disks utilized anyhow
  - Small random writes: big overhead because read from all disks and then a write to all disks plus parity is necessary (read is required to be able to calculate the new parity)
- At least three disks required
- Capacity = (size of smallest disk in the array) * (number of disks - 1)

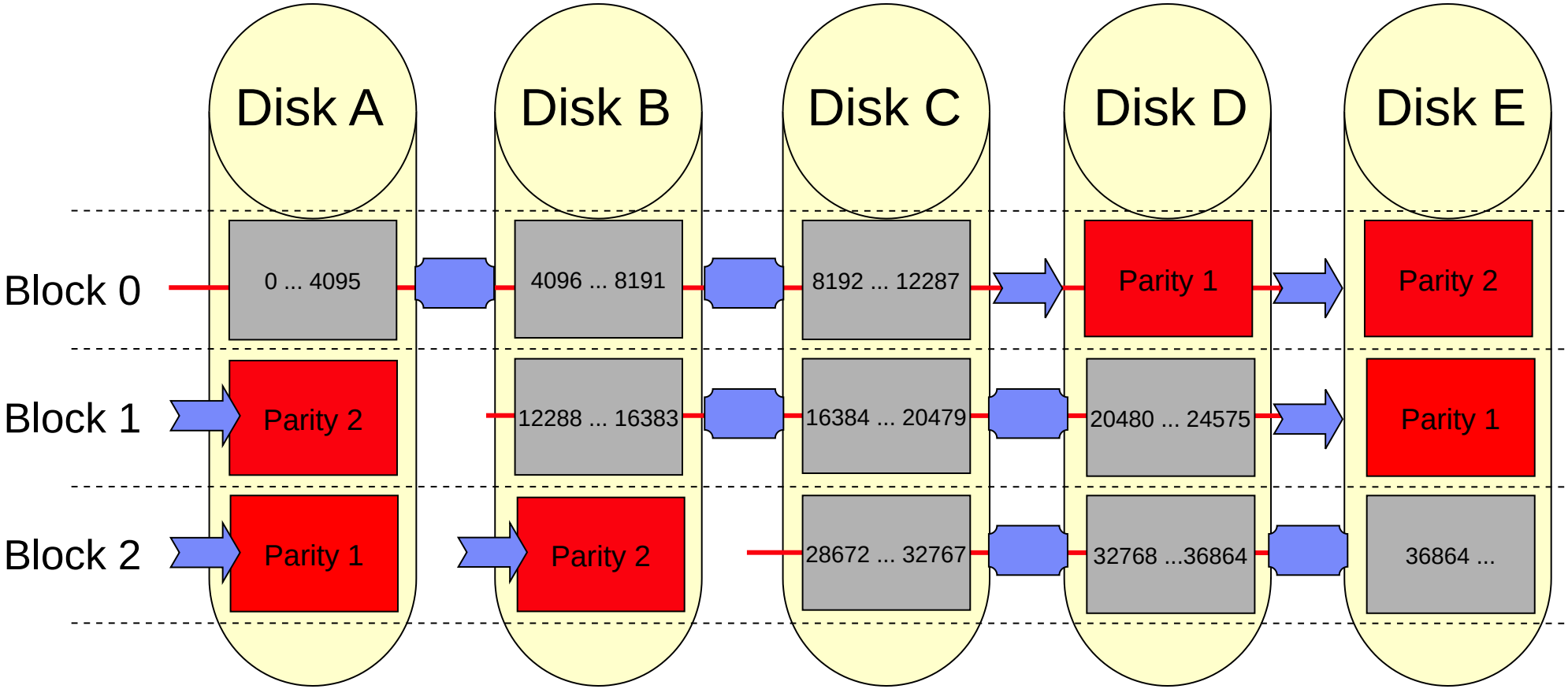
RAID 5: Striping with distributed parity



RAID 6: Striping with two distributed parities

- Gets used more often nowadays
- Striping like RAID5 – no dedicated parity drives (distributed parity)
- Double parity allows for two disk drives to fail
- Two different parity algorithms required! (usually XOR and Galois field algebra)
- Second parity calculation quite expensive
- Block level striping
- Because today's disk drives are multi TB in size the time to restore after a disk failure is getting longer. The risk for data loss increases substantially because no additional disk drive is allowed to fail with RAID 5 (and RAID restore activity means additional load on disk drives)
- Requires at least four disk drives
- Capacity = (size of smallest disk in the array) * (number of disks - 2)

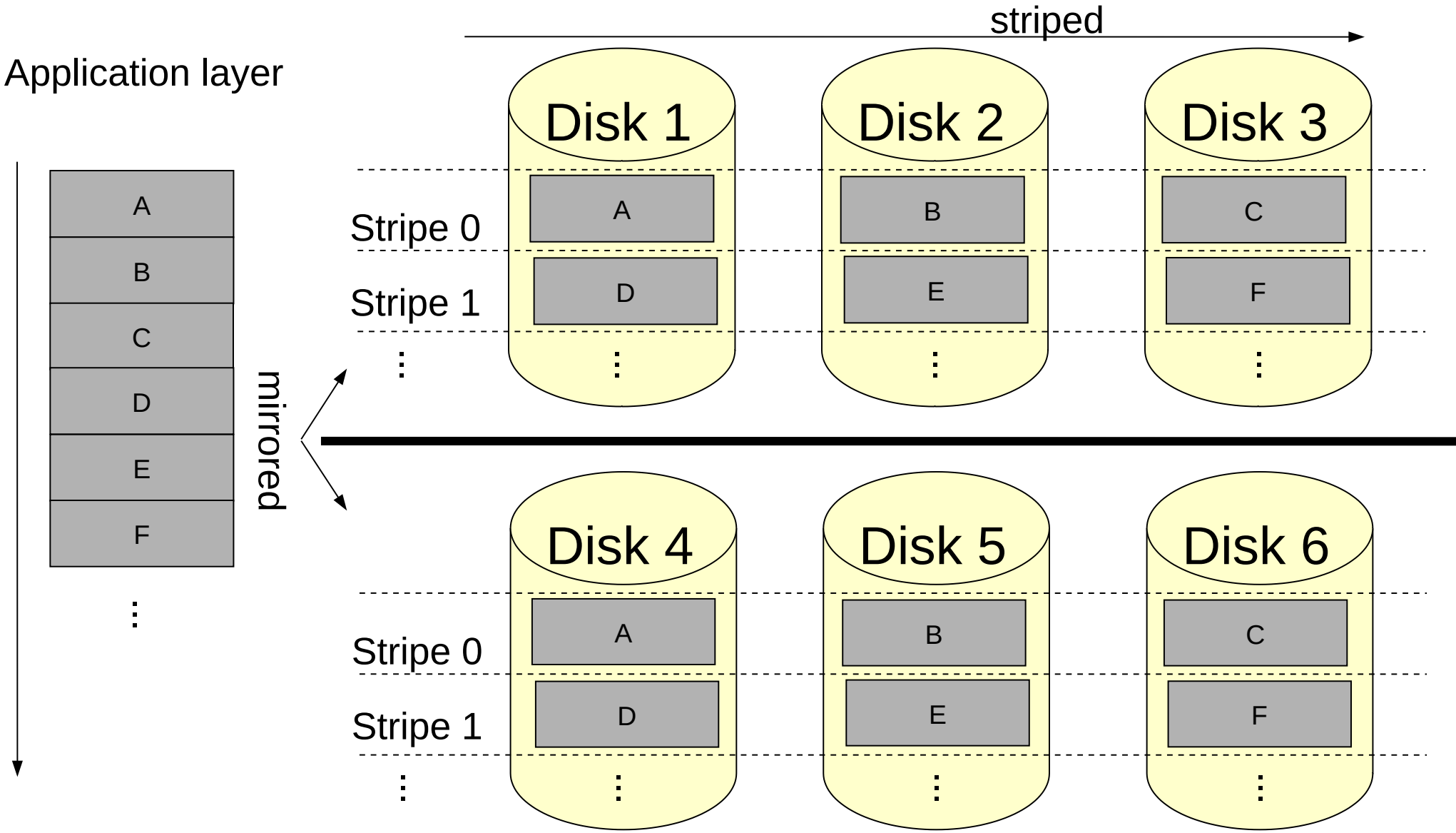
RAID 6: Striping with two distributed parities



RAID 10 (or RAID 0+1)

- Combination of RAID 0 (striping) and RAID 1 (mirroring)
  - High performance (striping) and high availability (mirroring) without burden of parity calculation
  - Cost intensive: requires additional disk drives (minimum of 4)
- Capacity = (size of smallest disk in the array) * (number of disks) / 2
  - Efficiency: 50%, like RAID 1 (if all disks are of the same size)
  - Very good performance and availability

RAID 10 (or RAID 0+1)



RAID Konzepte im Vergleich

RAID 1	▪ Spiegelung ▪ Hohe Ausfallsicherheit ▪ Gute Leseperformance (Zugriff auf Daten über mehrere Platten möglich) ▪ Etwas schlechtere Schreibperformance (Schreiben auf mehrere Platten) ▪ Nutzkapazität 50% ▪ Doppelter Speicherbedarf, hohe Kosten
RAID 5	▪ Striping mit Parity (distributed Parity über alle Disks) ▪ Gute Ausfallsicherheit ▪ Verschlechterung des Laufzeitverhaltens für die Dauer des Ausfalls einer Disk (durch die Rückberechnung von fehlenden Datenblöcken aus der Parity) ▪ Nutzkapazität 87% (bei 8 Disks im Array)
RAID 6	▪ Striping mit doppelter Parity (distributed double Parity über alle Disks) ▪ Sehr gute Ausfallsicherheit (2 Disks eines Arrays können ausfallen) ▪ Nutzkapazität 75% (bei 8 Disks im Array)
RAID 10	▪ Striping und Spiegelung der Daten (Kombination aus RAID 1 und RAID 0) ▪ Kombiniert Eigenschaften von Striping und Spiegelung, wie hohe Performance (keine Parity Berechnung) und Ausfallsicherheit ▪ Nutzkapazität 50% ▪ Doppelter Speicherbedarf, hohe Kosten