Storage Area Network

Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, iSCSI, InfiniBand and FCoE, Second Edition.
U. Troppens, W. Müller-Friedt, R. Wolafka, R. Erkens and N. Haustein
 ISBN: 978-0-470-74143-6

Joint master program Skoltech and CMC MSU

Prof. R. Smelyanskiy

Content

- Disk subsystems and their organizations
- JBOD
- RAID
- Intelligent Disc Subsystems (DS)
- CPU – DS tract
- SCSI
- Fibre Channel
- FC over TCP\IP
- File System and SAN
- Network File System (NFS)
- Network Attached Storage (NAS)
- Conclusion

Main trends in traffic growth in networks

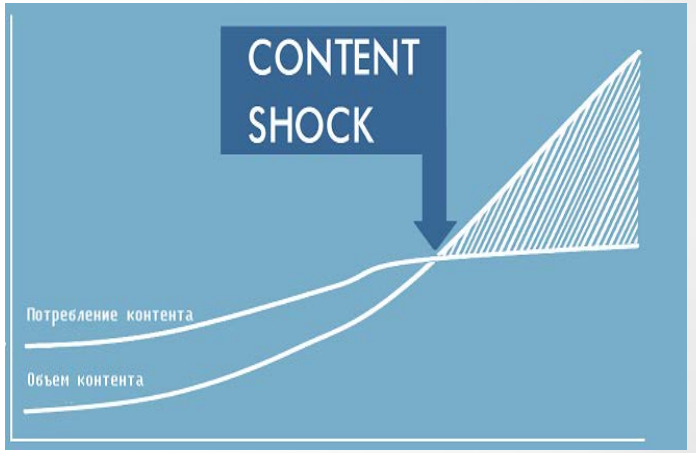
The main trends:

Global annual IP traffic: 2.3 ZB (zettabytes = 10 ^ 21) per year by 2020.

The amount of traffic from wireless and mobile devices will be two-thirds of the total IP traffic by 2020 and will exceed the one from fixed connected computers by 2020.

Traffic between the data center (DC) will dominate global traffic

Specifics of the growth of mobile traffic:



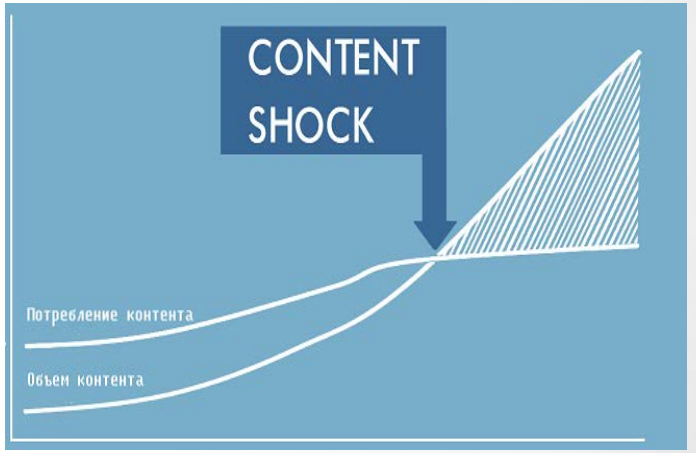
Expected that at the end of 2020, the volume of mobile traffic will increase by 8 times comparing 2015 and reach 30.6 EB / month (Exabyte = 10 ^ 18).

Mobile traffic during this period will grow three times faster than traffic in fixed networks.

The number of grains of sand on all the beaches of the Earth is **700,500,000,000,000,000,000** (or seven quintillion five quadrillion).
40 zettabytes is 57 times more than the number of grains of sand on all the beaches of the planet.
If you write 40 zettabytes of data to modern Blu-ray discs, the total weight of the discs (excluding paper and plastic packaging) is equal to the weight of **424 aircraft carriers**.

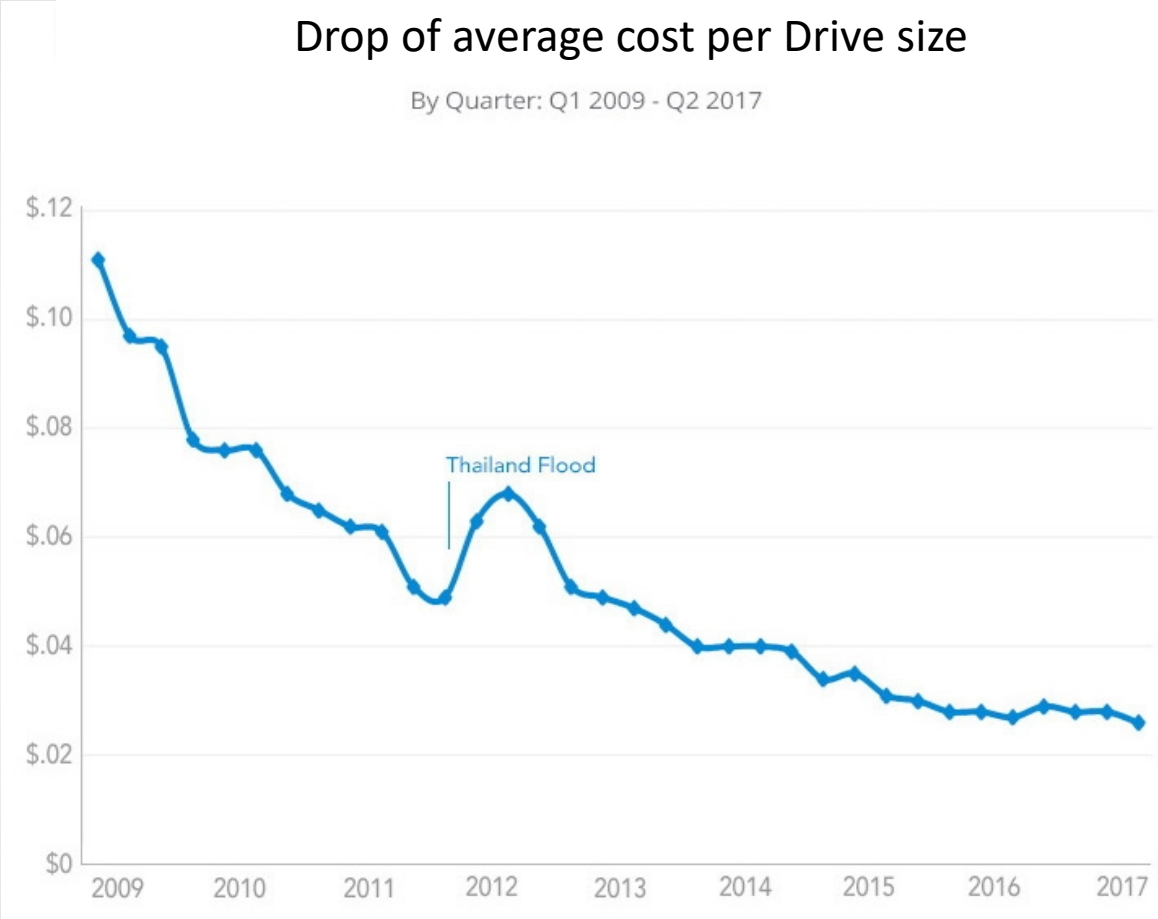
The main Features of the growth of gaming and video traffic

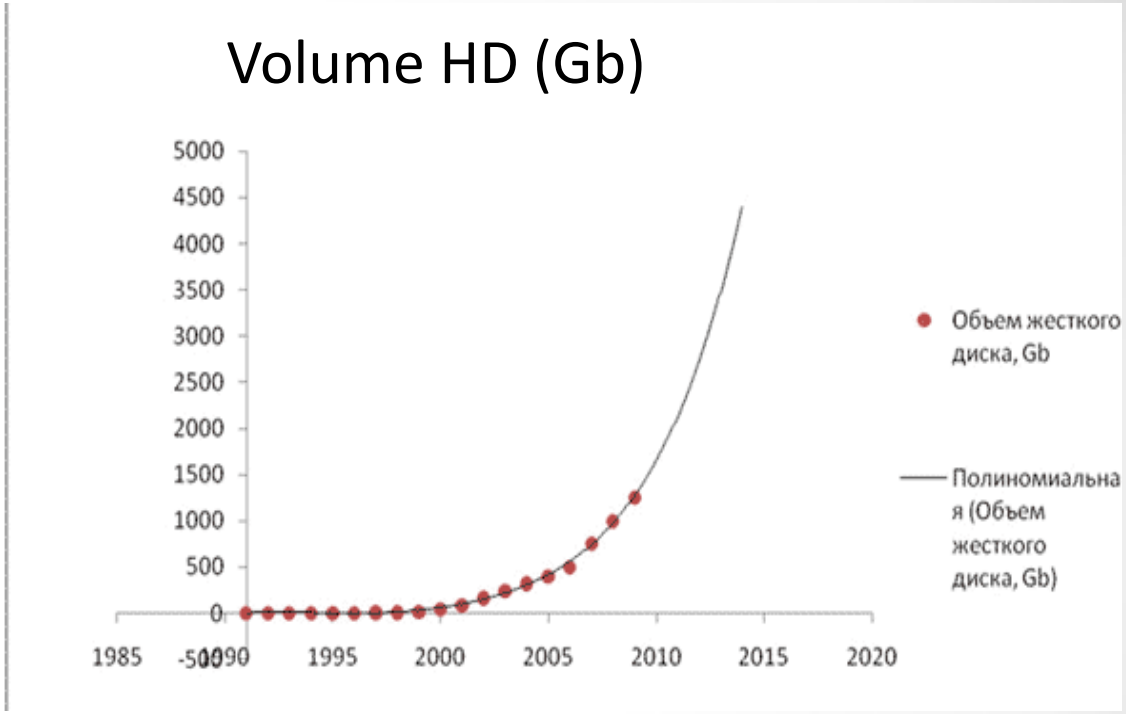
- In 2020, it will take more than 5 million years to view all the video content that will pass through the global IP networks every month.
- Virtual Reality traffic grew 4 times by 2015. By 2020, it grows another 61 times with an average annual growth rate of 127%.
- Over the past year, the volume of video surveillance traffic has almost doubled, and by 2020 it doubles.
- Gaming Internet traffic will grow by 7 times by 2020.
- The volume of consumer video traffic on demand by 2020 will almost double.
- IPTV traffic increased by 50 percent in 2015. By 2020, it will grow by 3.6 times.



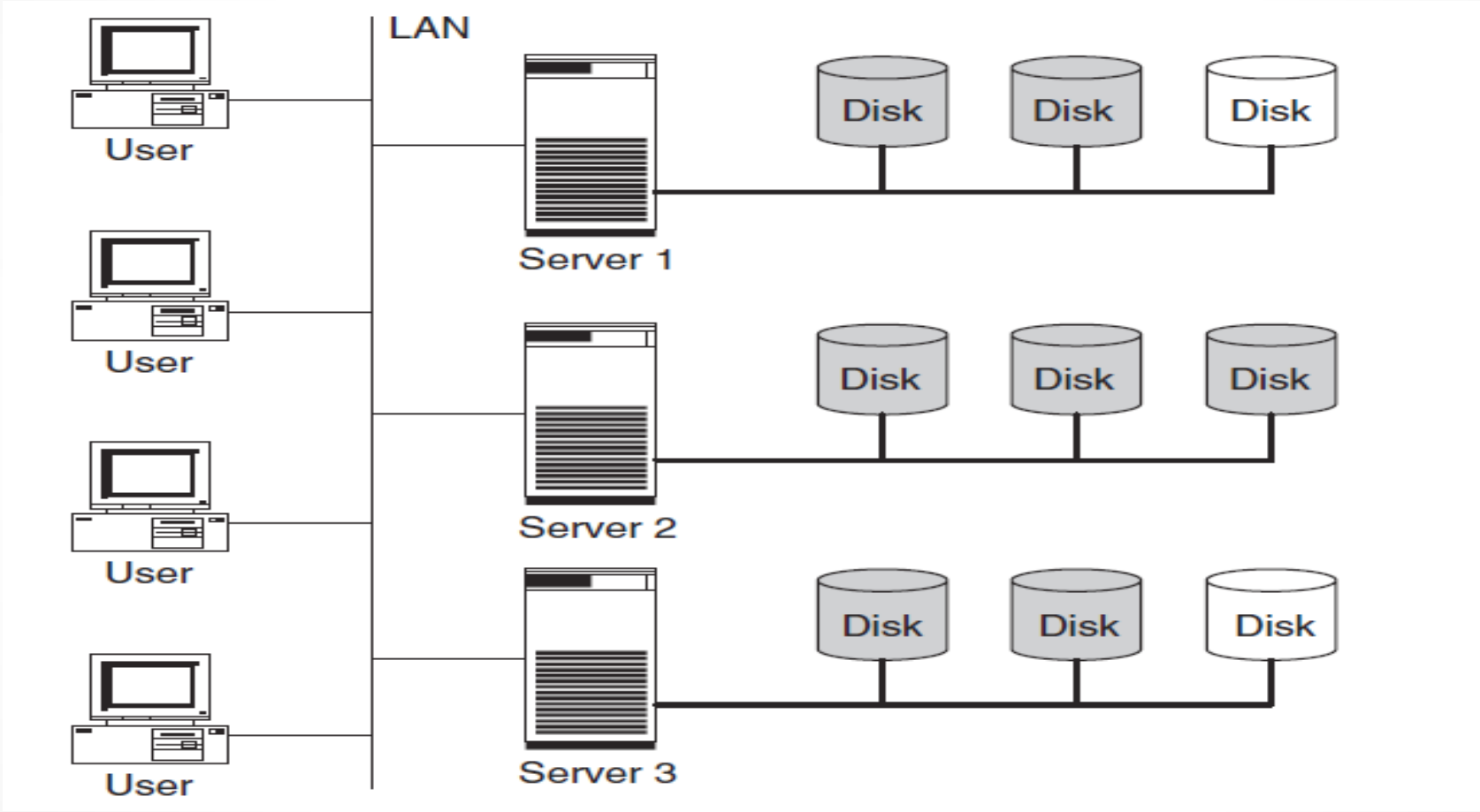
- **Data Center Growth**
- **The use of IT in the business processes of companies is growing**
- **The amount of data stored is growing.**
- **Increasing the volume of archives**

The main trends in Data storages

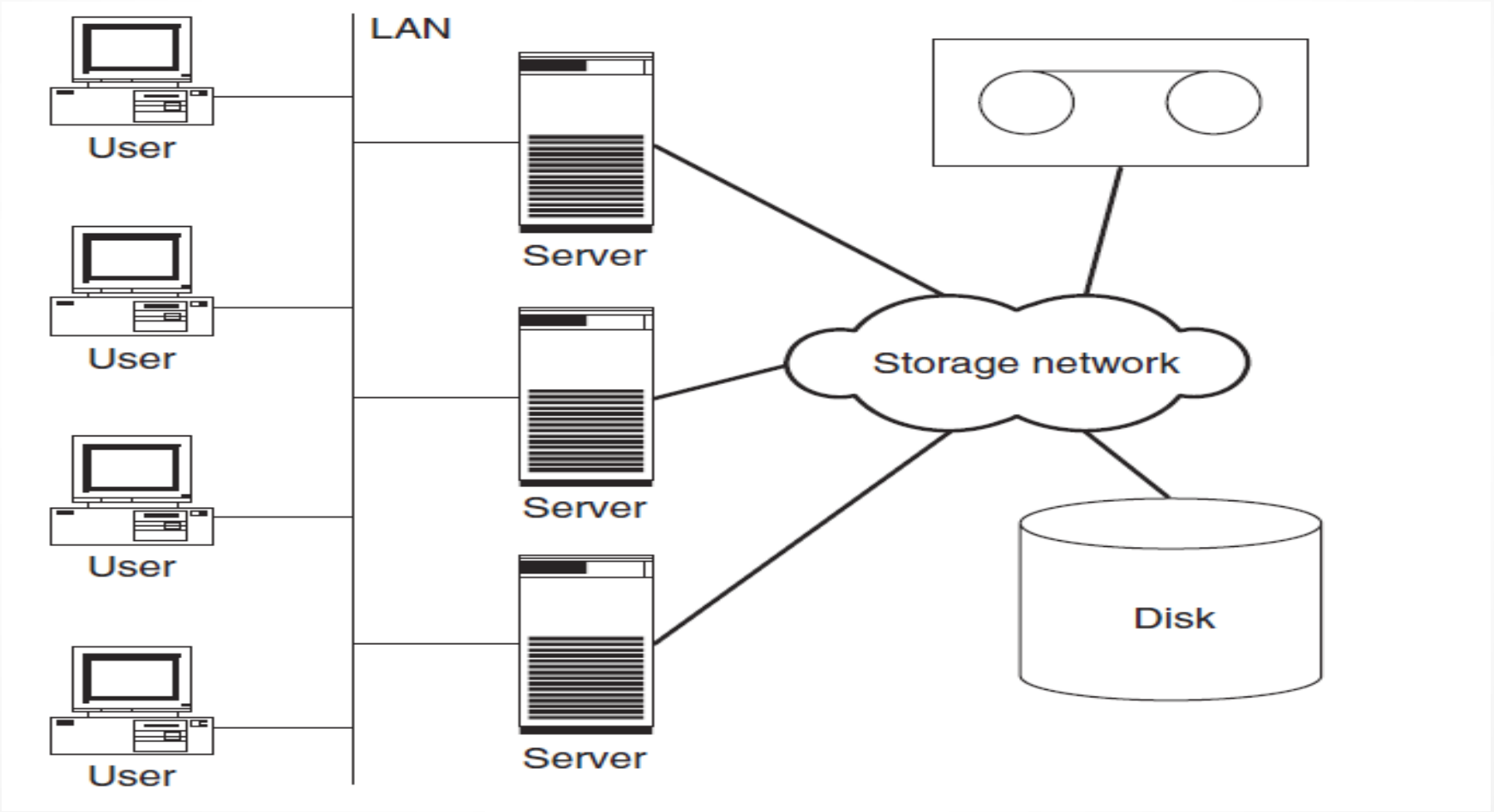




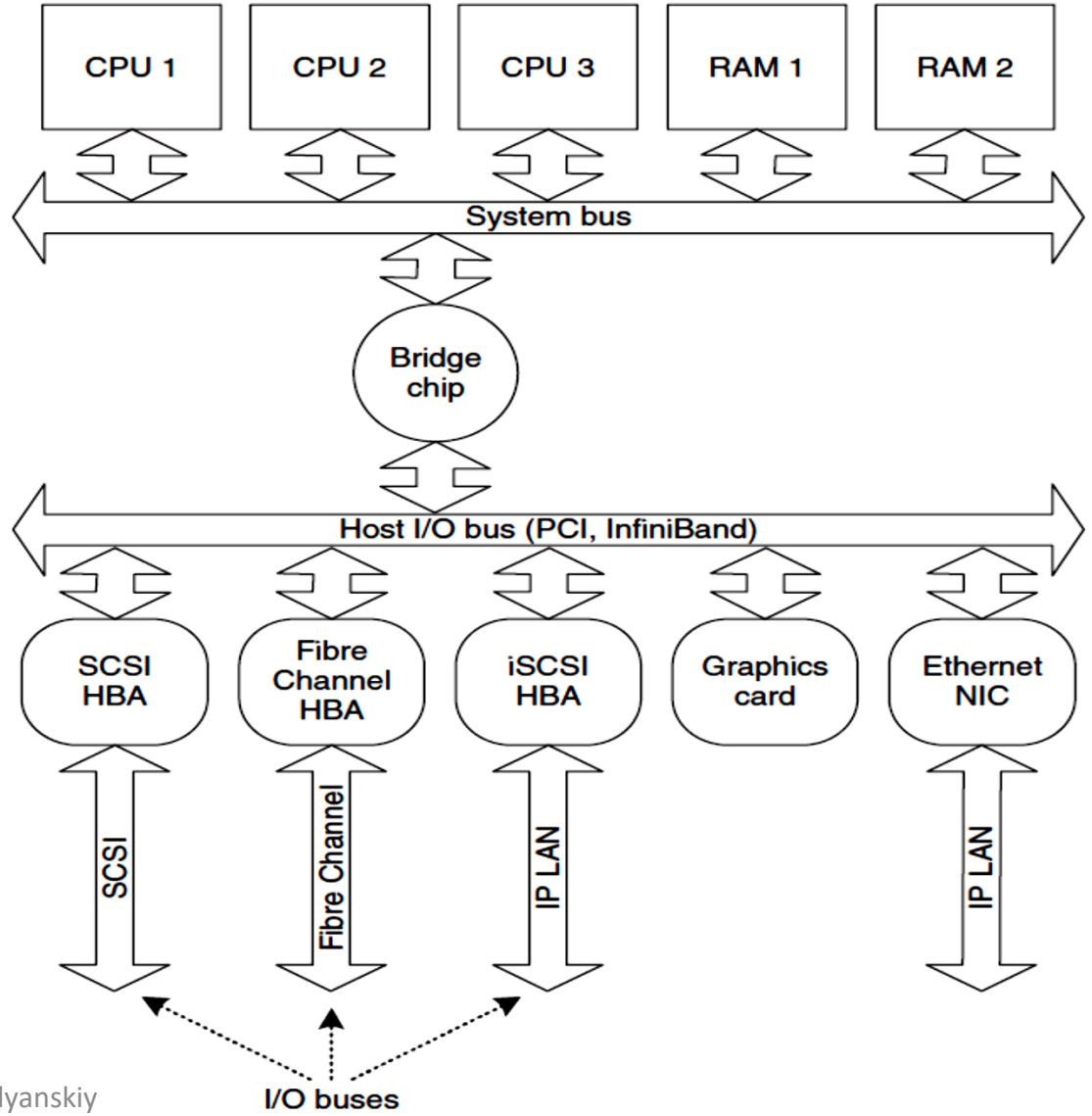
Server Centric Architecture



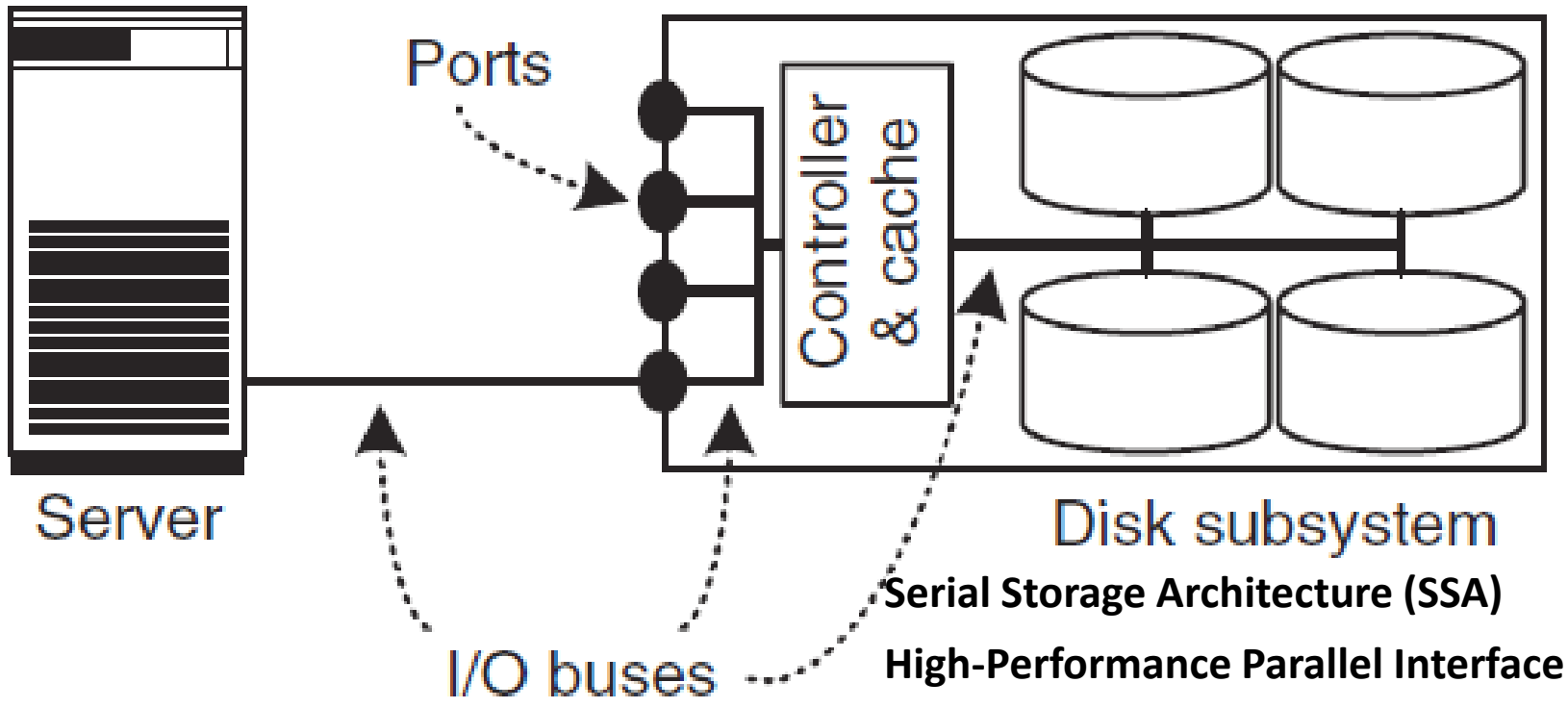
Storage Centric Architecture



THE PHYSICAL I/O PATH FROM THE SERVER CPU TO THE STORAGE SYSTEM



THE PHYSICAL I/O PATH FROM THE SERVER CPU TO THE DISK SUBSYSTEM



Serial Storage Architecture (SSA)

High-Performance Parallel Interface (HIPPI),

Advanced Technology Attachment (ATA),

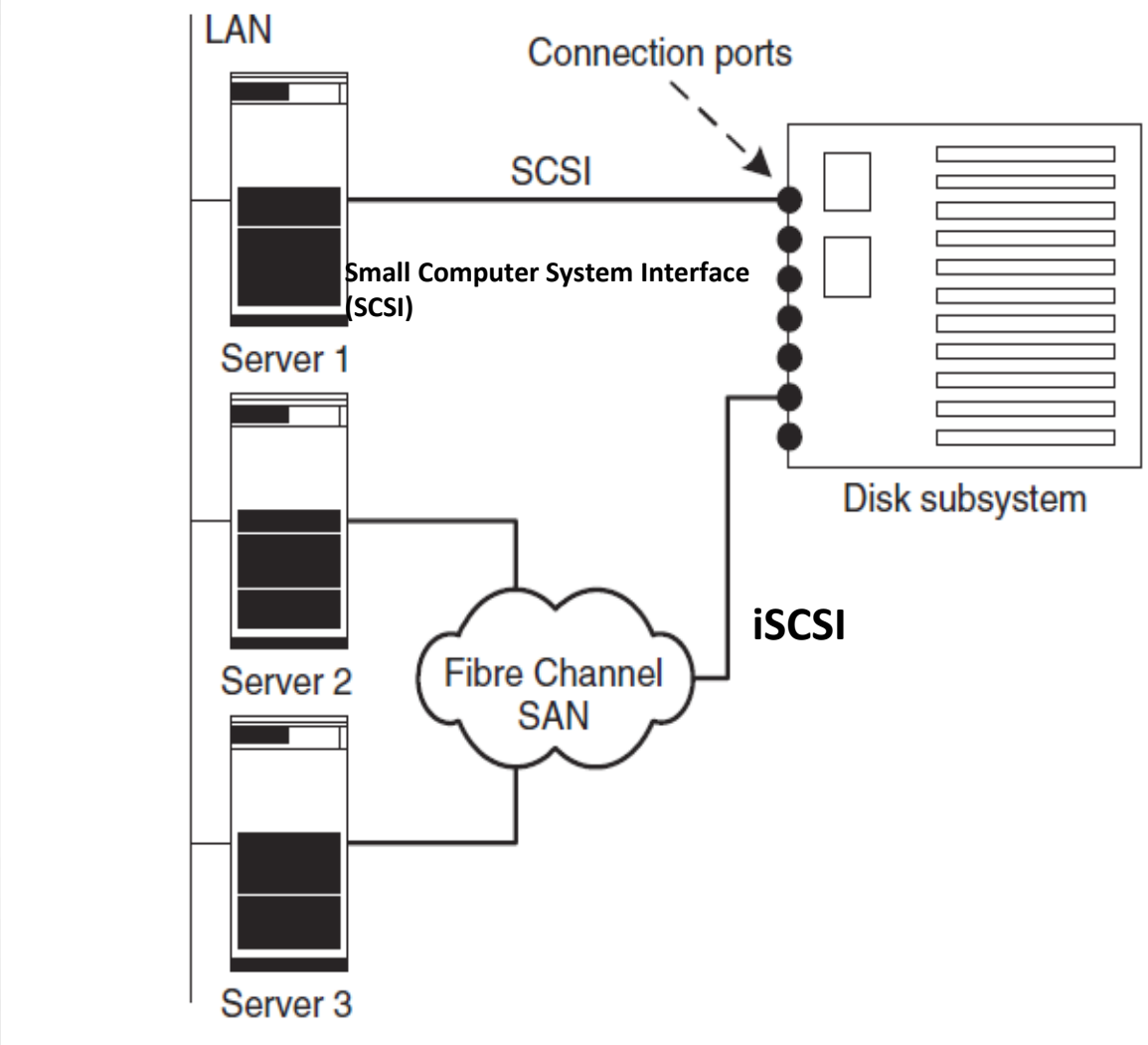
Integrated Drive Electronics (IDE),

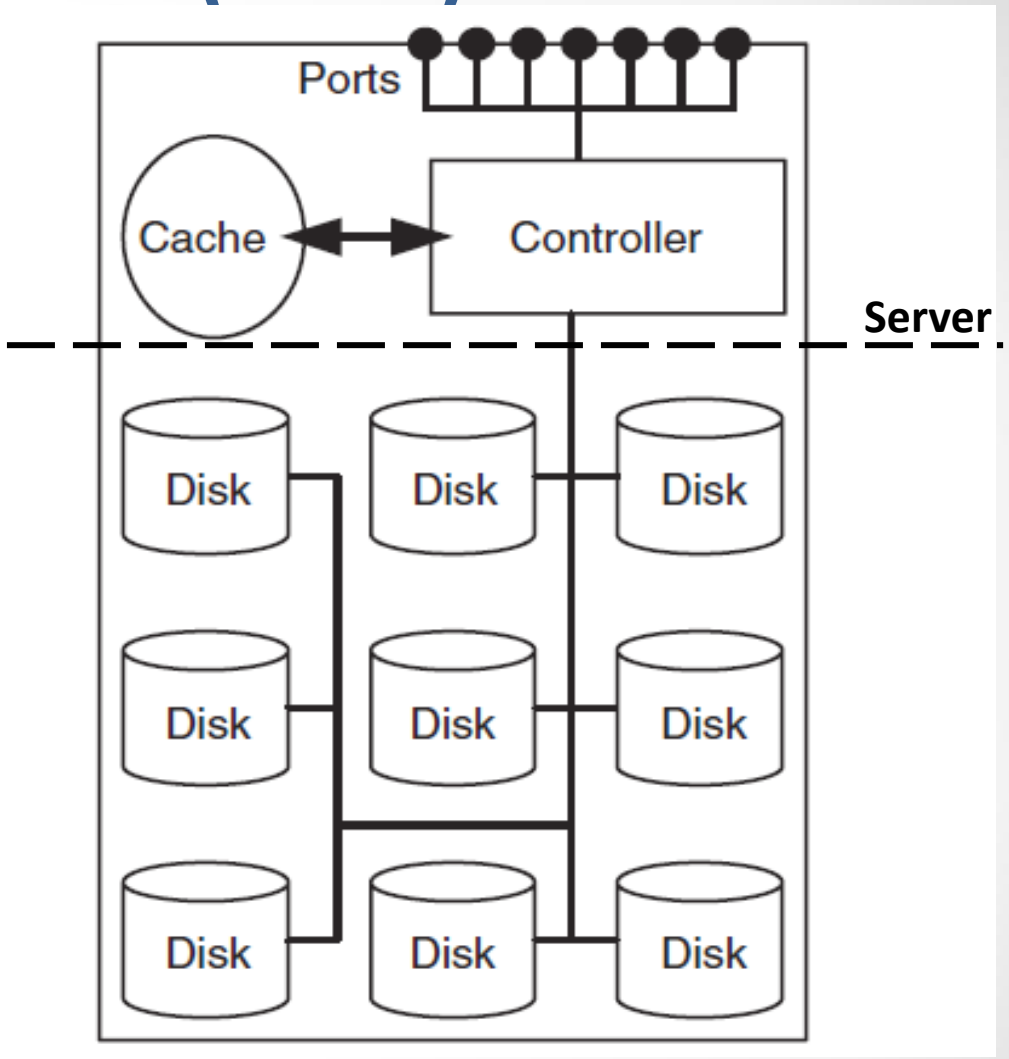
Serial ATA (SATA), Serial

Attached SCSI (SAS)

Universal Serial Bus (USB).

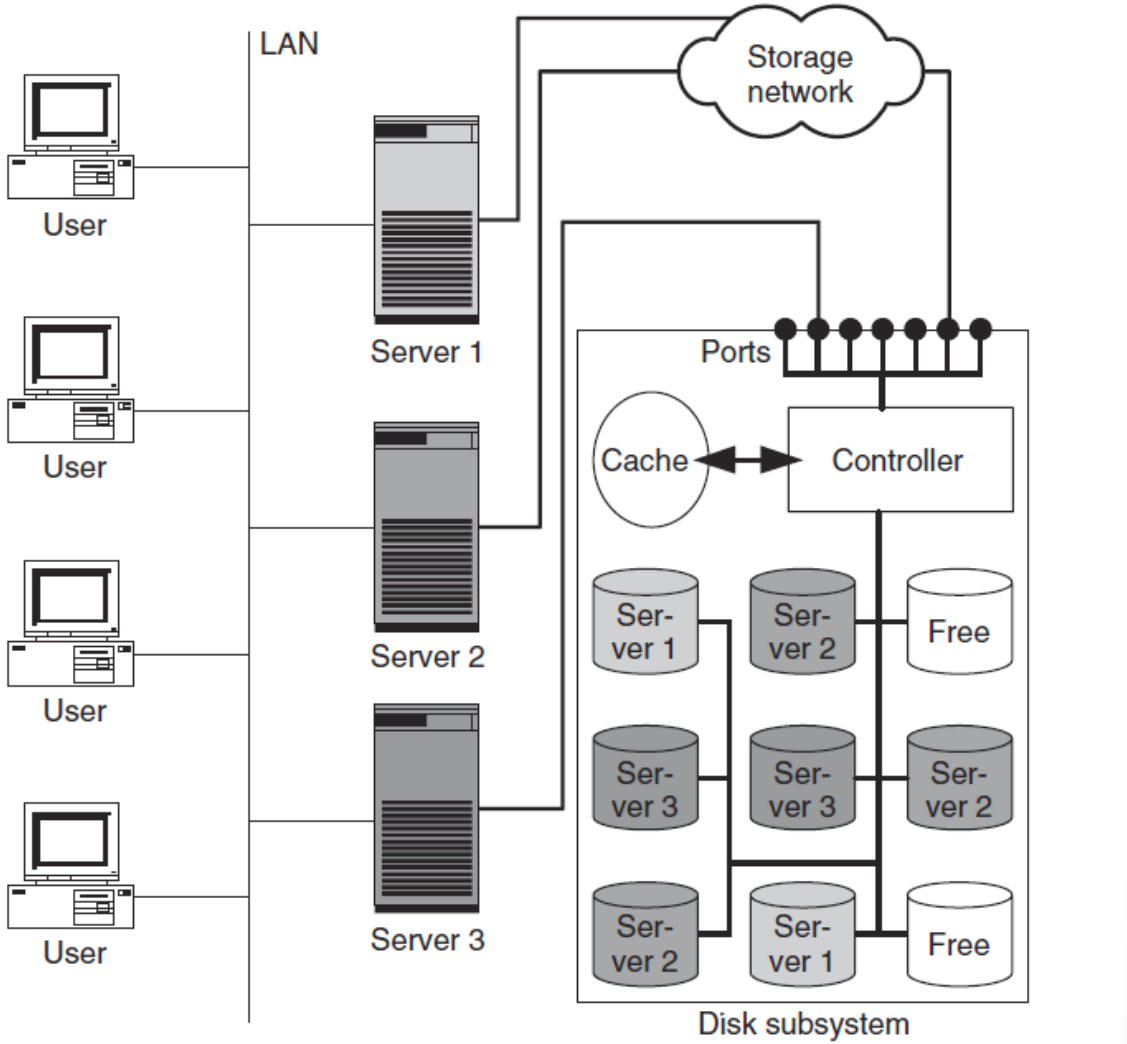
Disk Subsystem Architecture (JBOD)



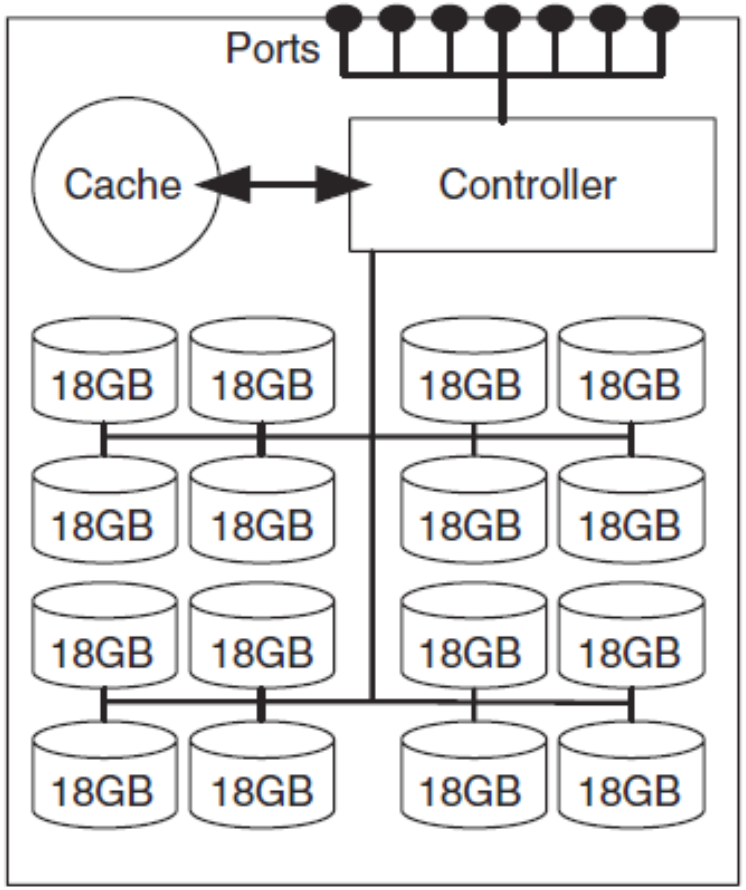


(1) no controller; (2) RAID controller; (3) intelligent controller with services like e.g. instant copy and remote mirroring .

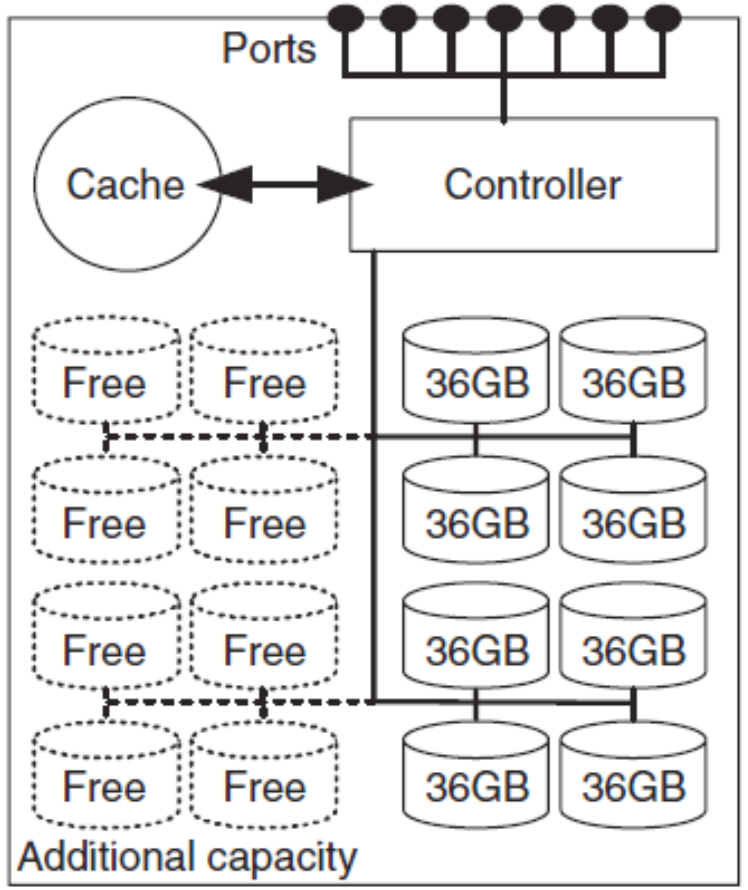
Disk Storage System - usage example



Disk Subsystem: internal organization and disk capacity

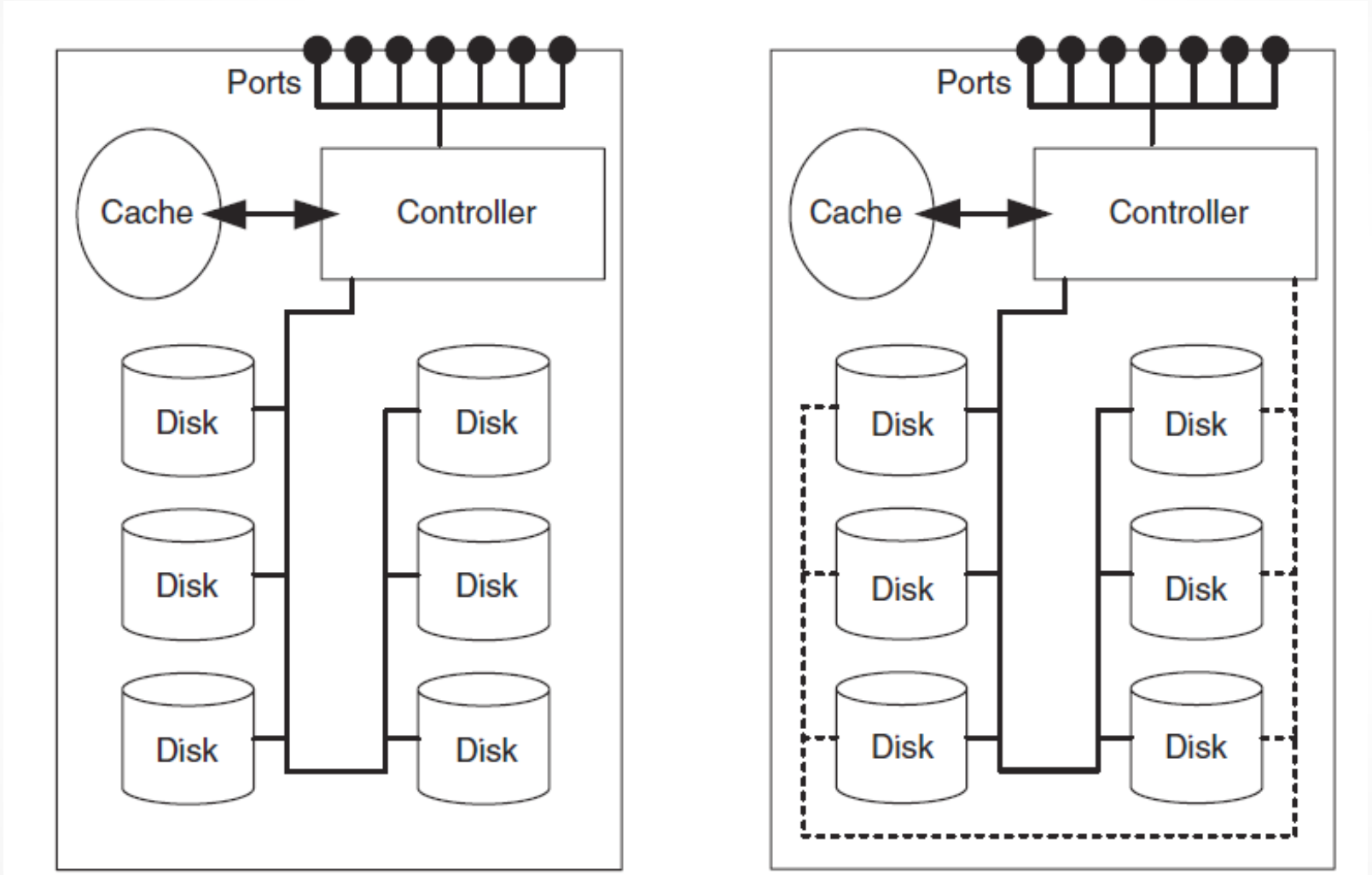


Small Drives



Large Drives

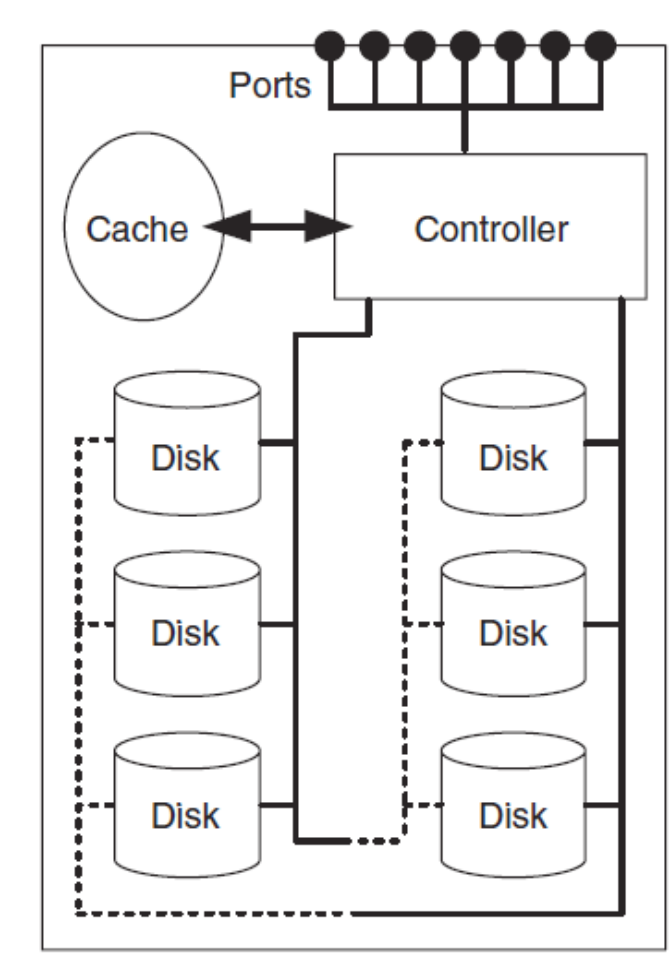
Disk Subsystem internal organization



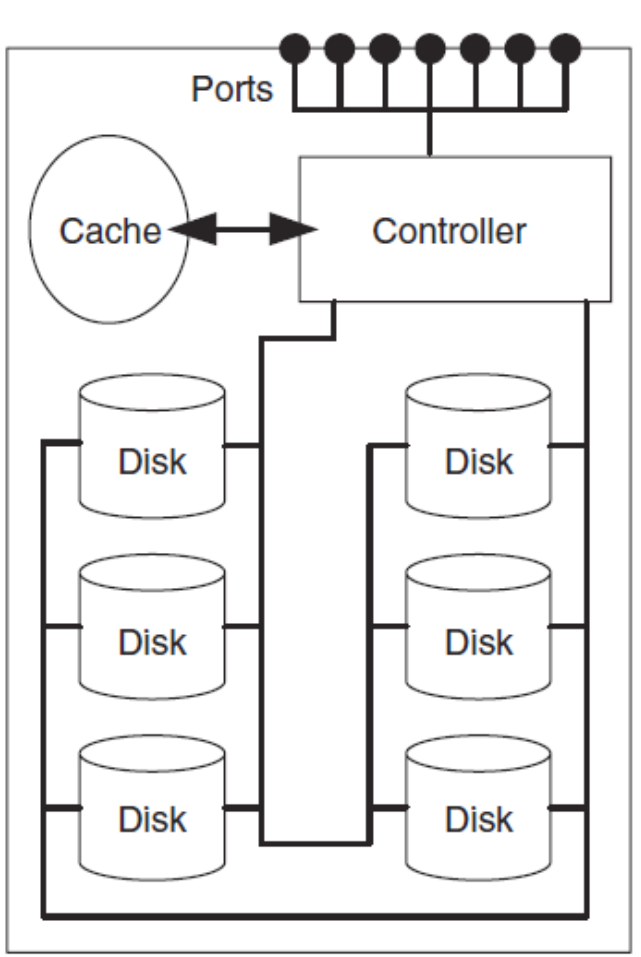
Active Disks connectivity

Active/Passive Disks connectivity

Active duplication



Active/Active with separation
No load sharing

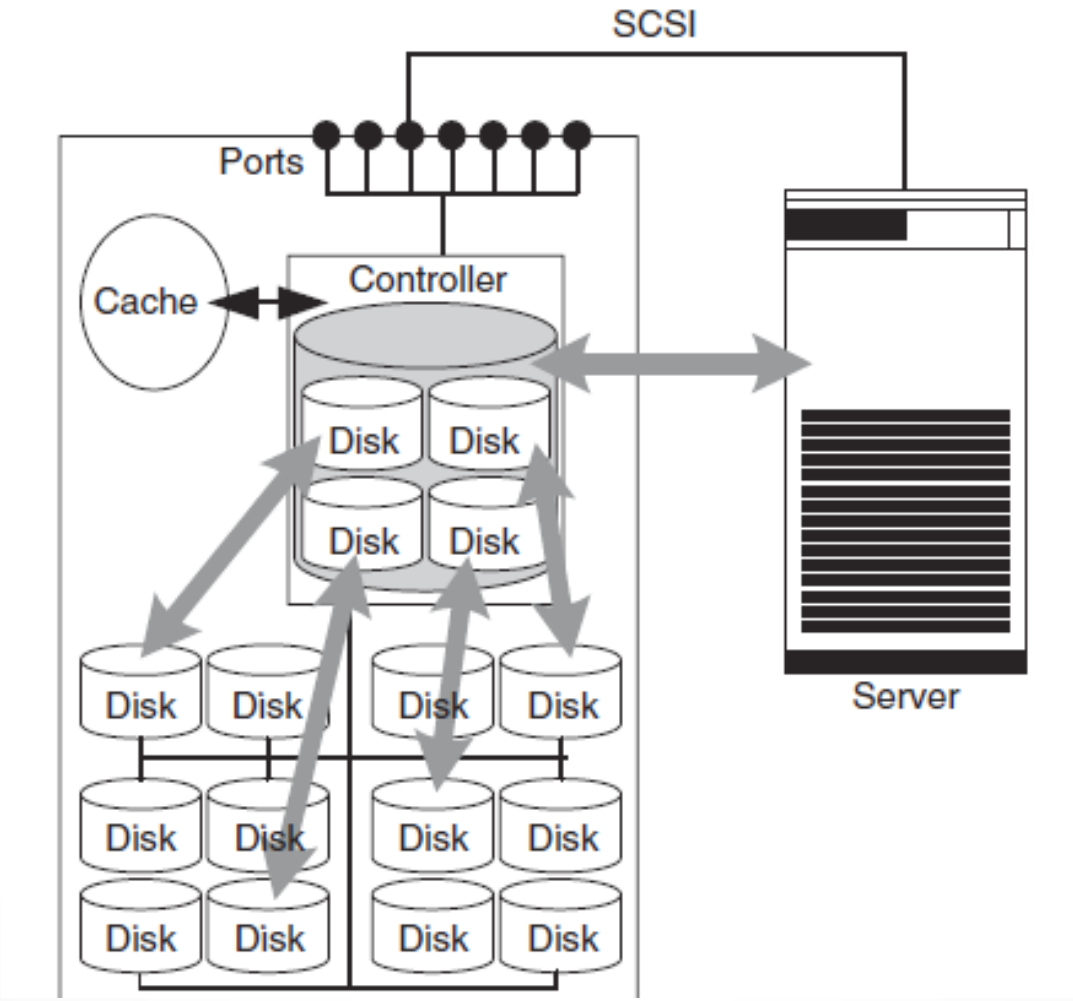


Active/Active without separation
Load sharing

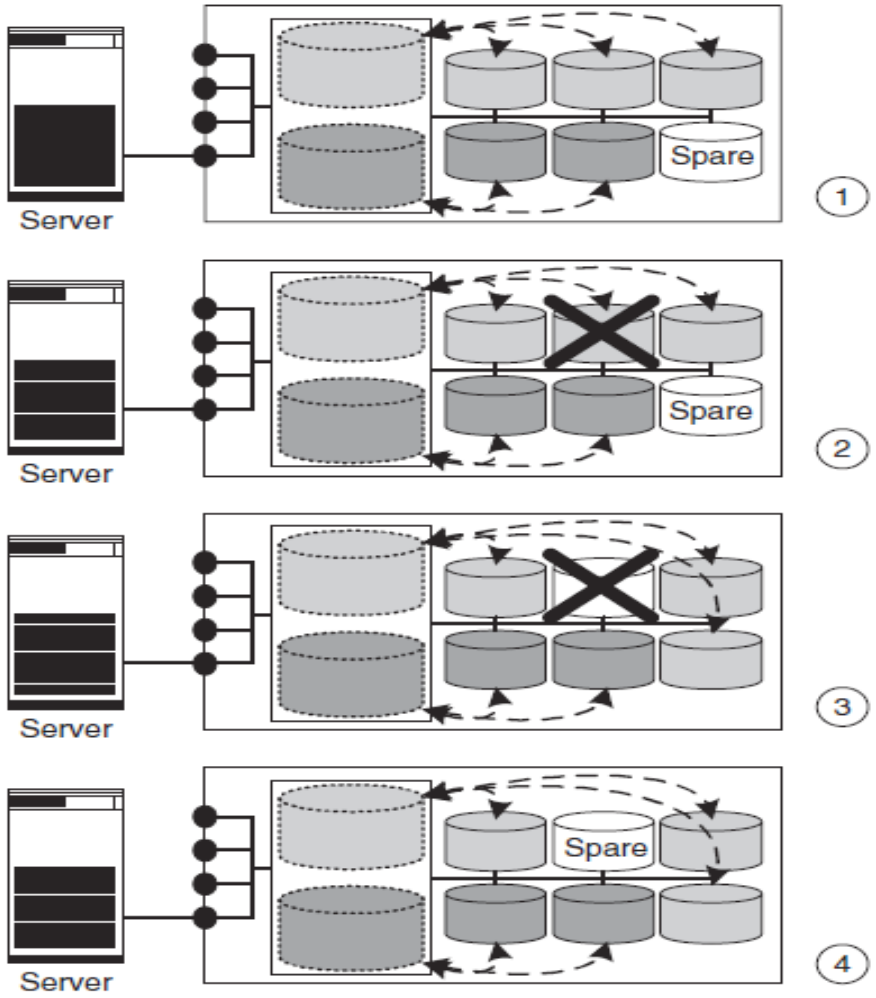
JBOD: JUST A BUNCH OF DISKS

- No internal controller
- The connections for I/O channels and power supply are taken outwards
- Small number of HDD
- Outside server can see JOBD as several independent disks
- Because of one point of connection JOBD it is the bottleneck of Disk Subsystem

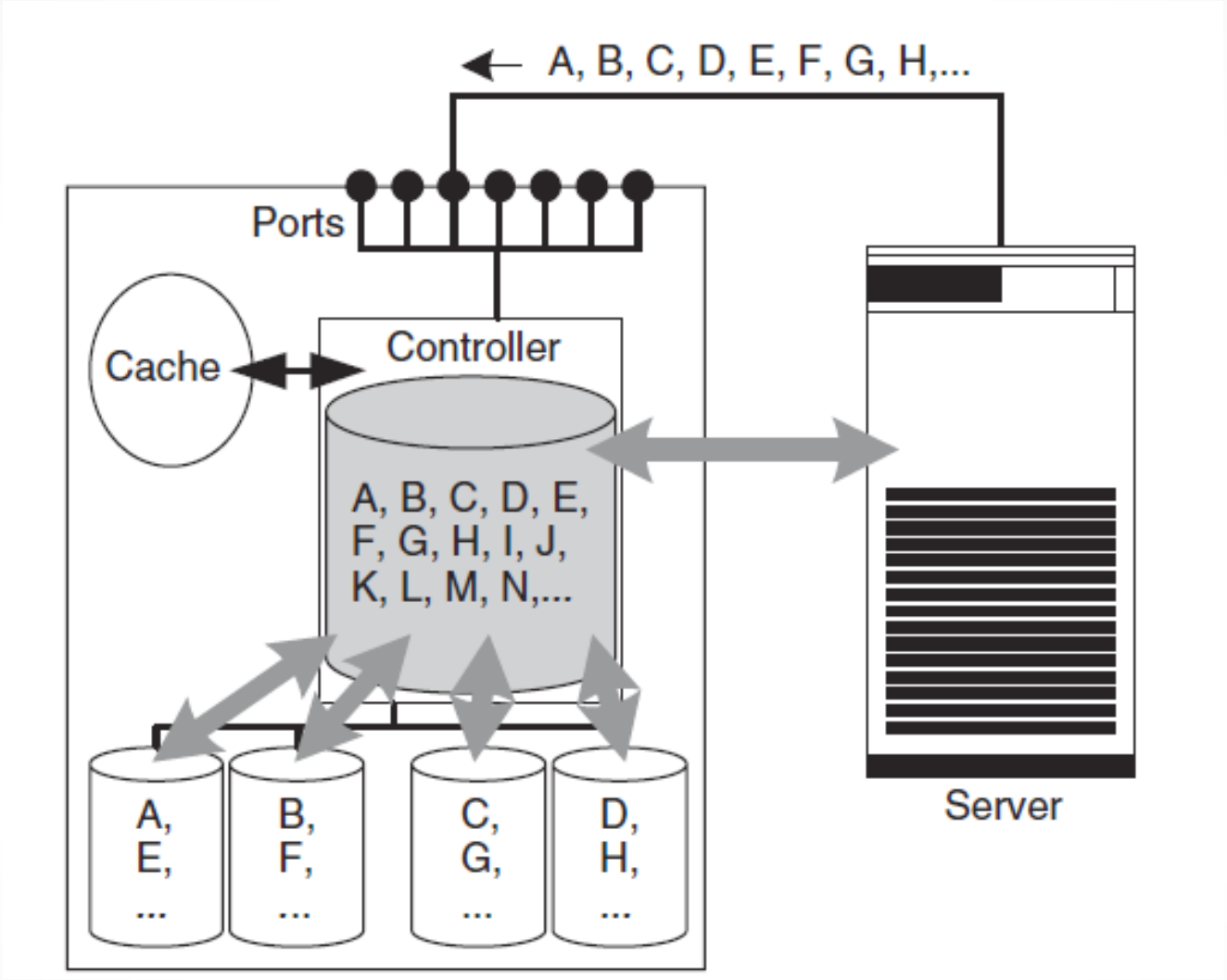
Storage Virtualization with RAID – Redundant Array of Independent Disks



Hot Spare Disk

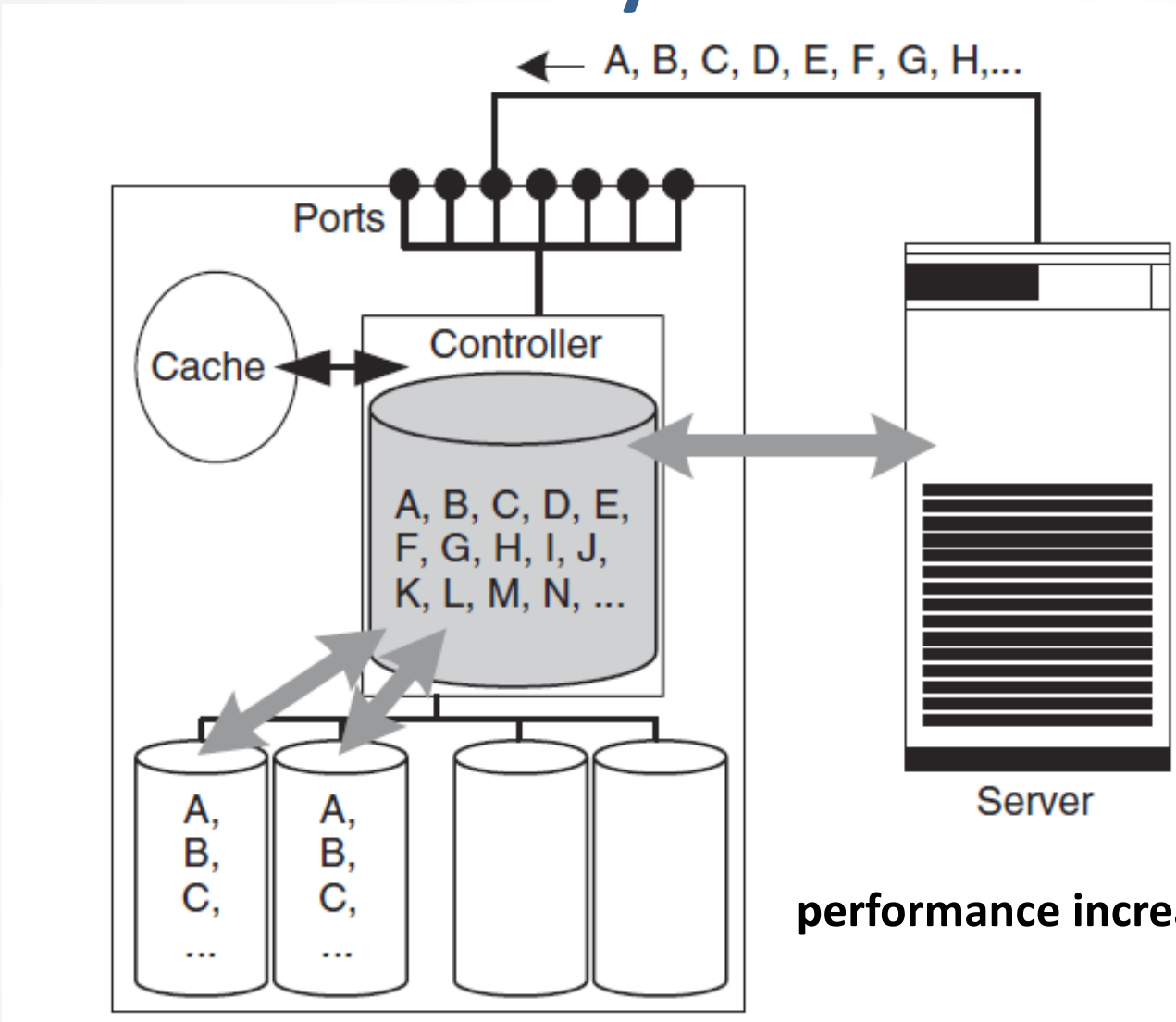


RAID 0: block-by-block striping



RAID 0 increases the performance of the virtual hard disk, but not its fault-tolerance.

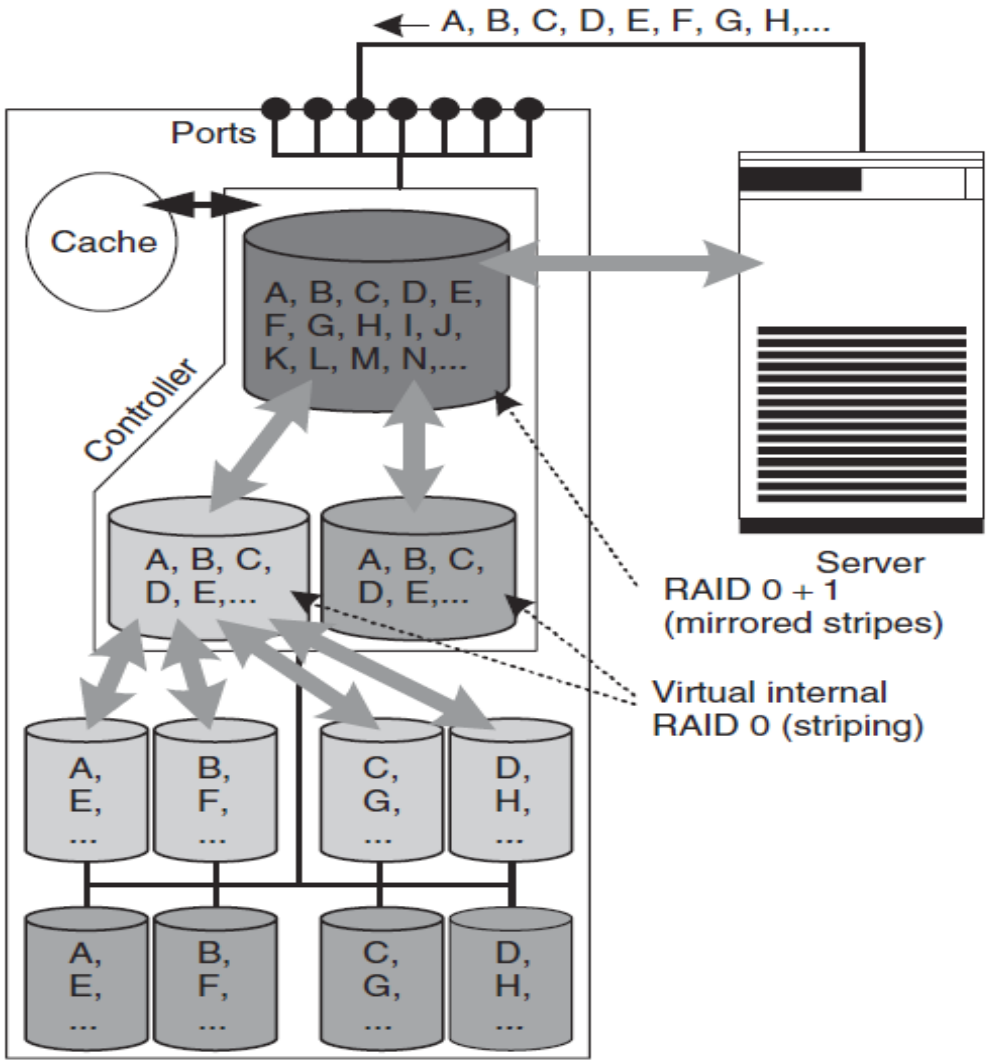
RAID 1: block-by-block mirroring



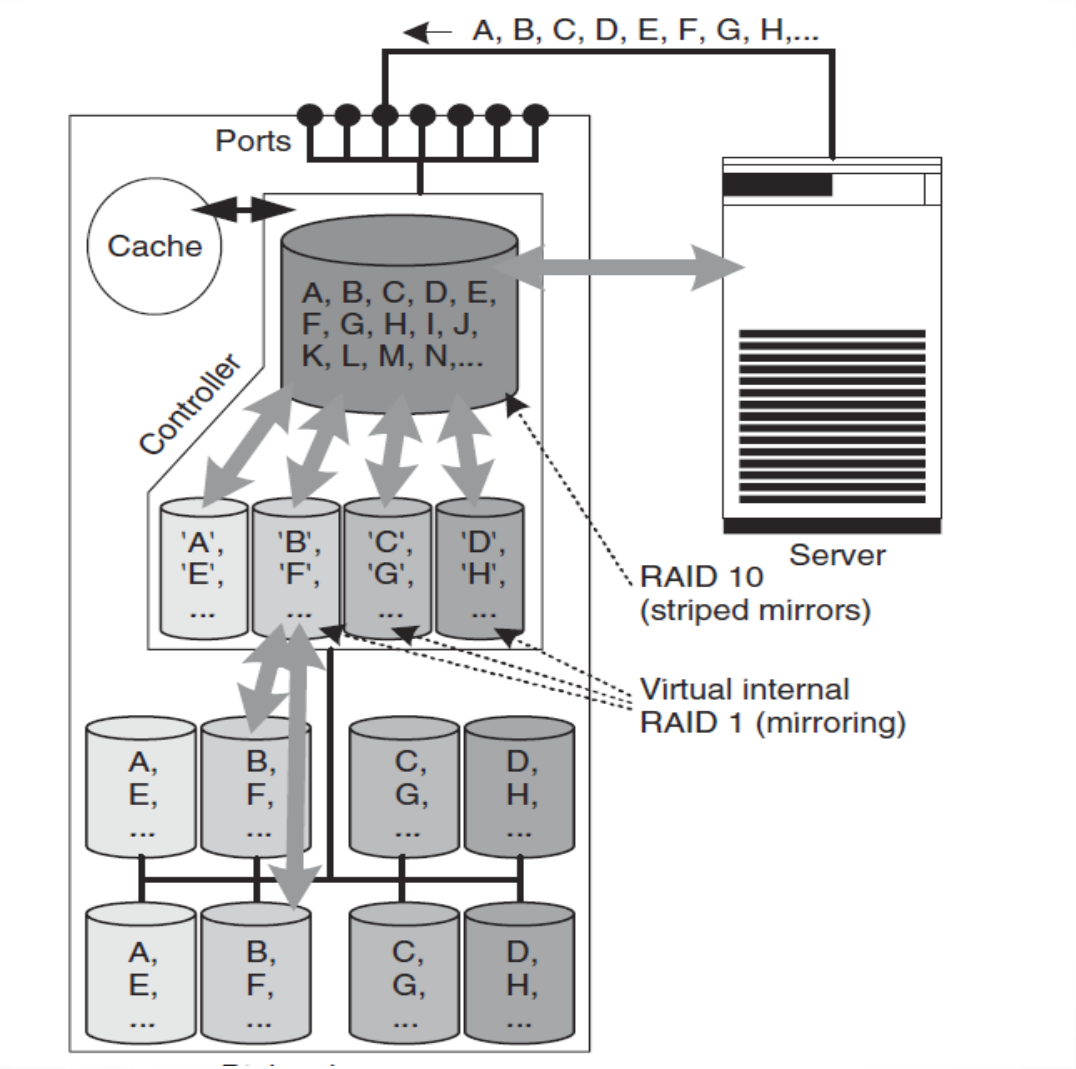
performance increases are only possible in read operations

fault-tolerance is of primary importance

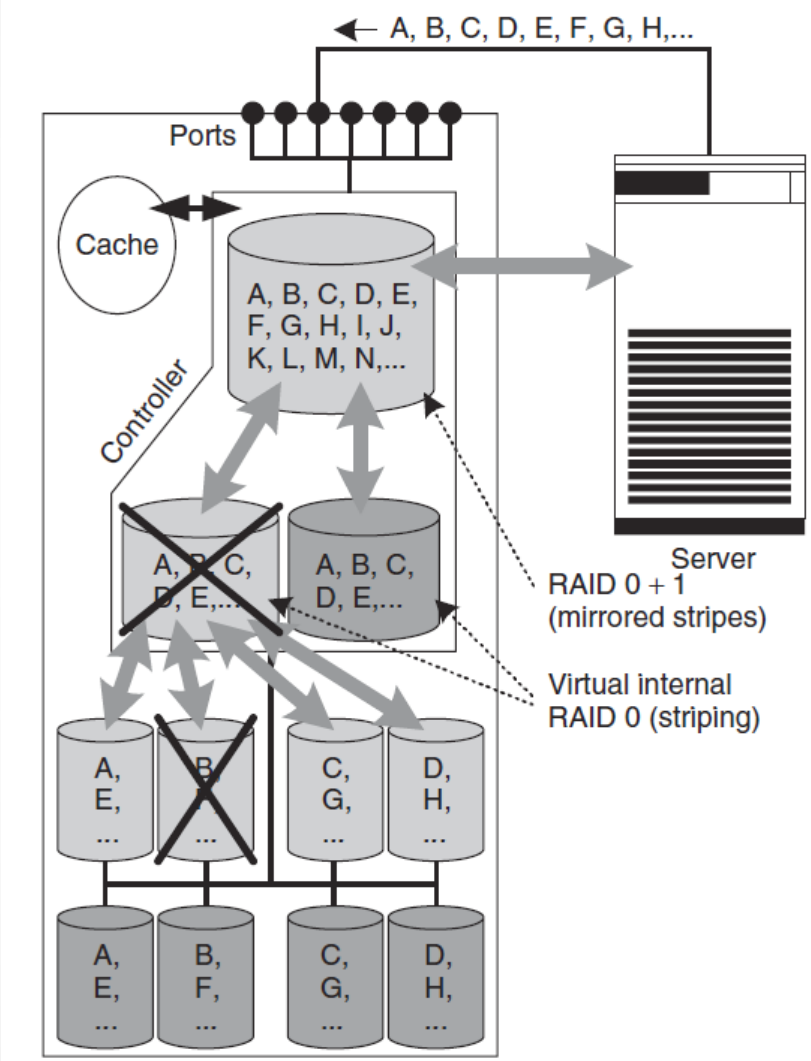
RAID = 0+1 (mirrored stripes)

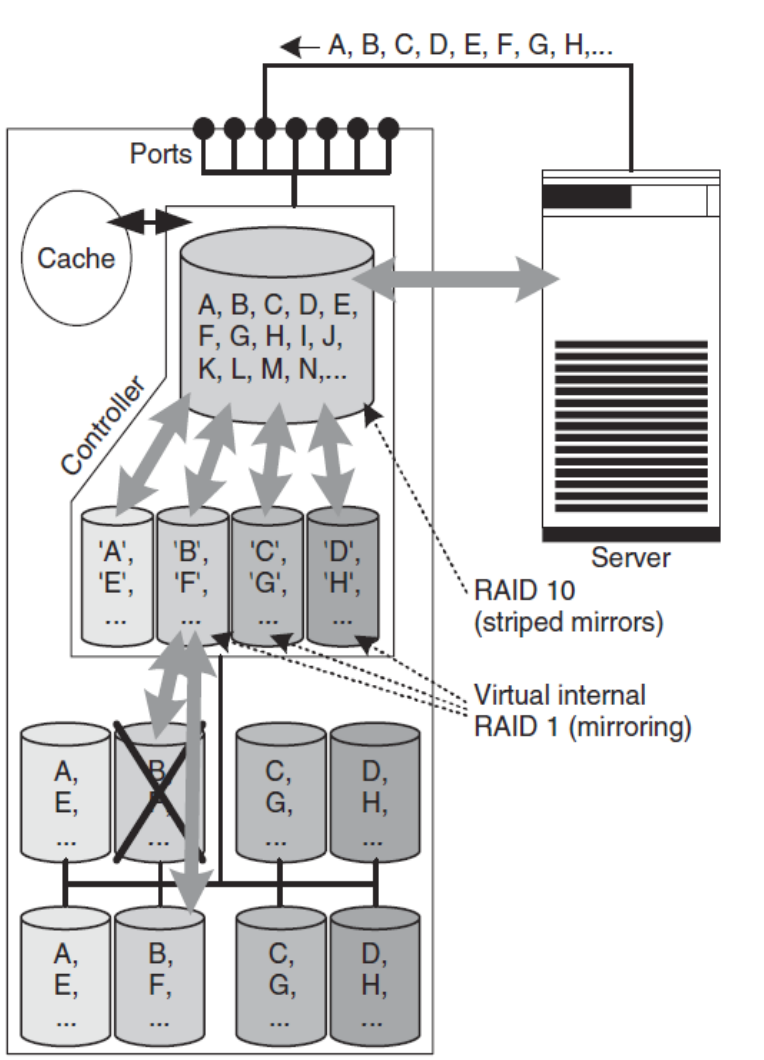


RAID = 1+0 (striped mirrors)

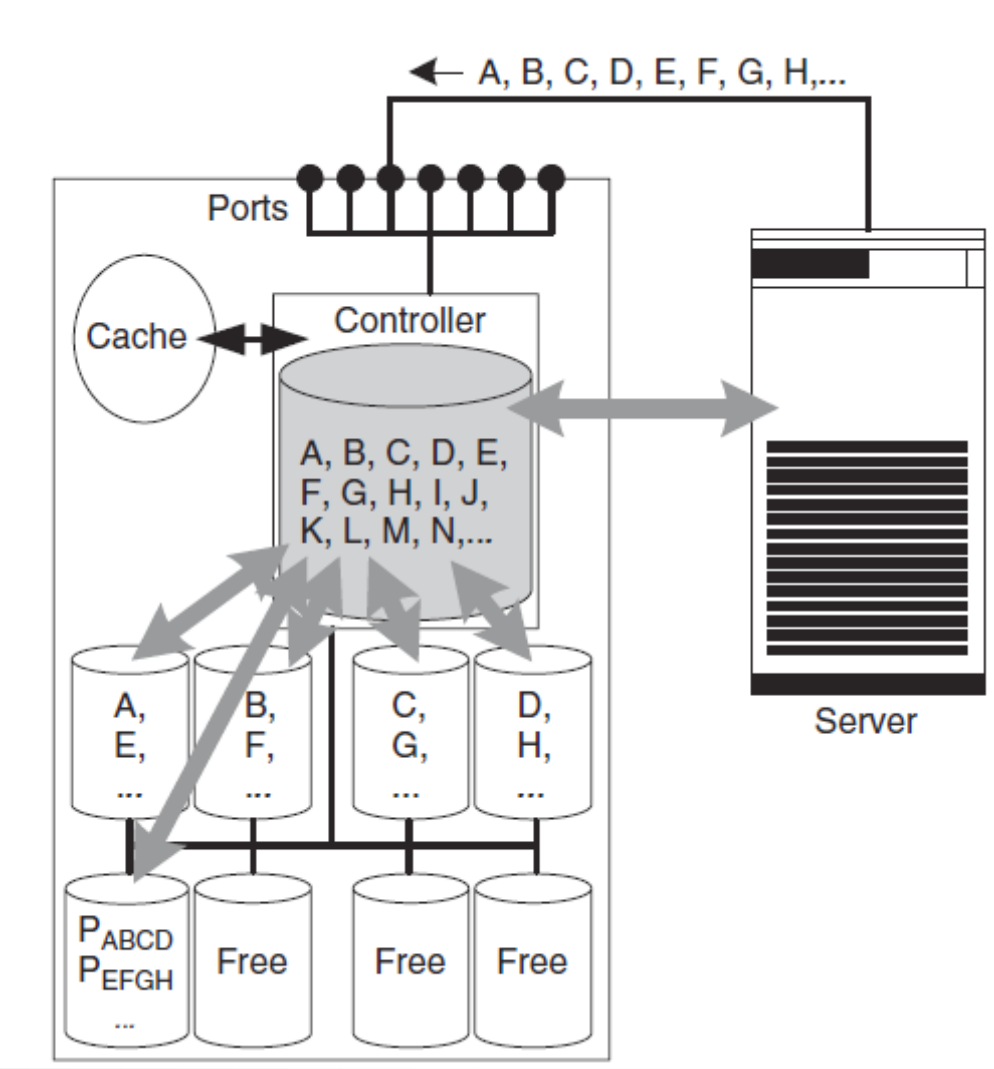


Comparison RAID 0+1 vs 1+0





RAID 4 and RAID 5: parity instead of mirroring



A was changed to **~A**

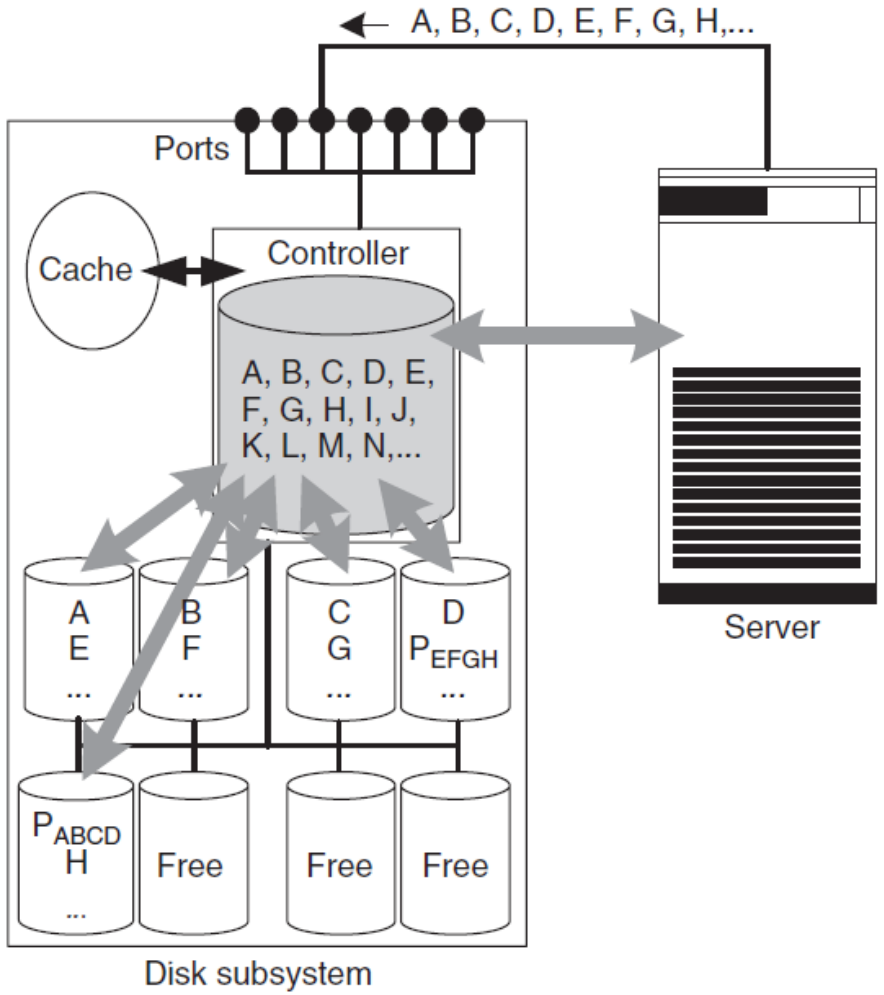
Δ = A xor ~A

~P = Δ xor P

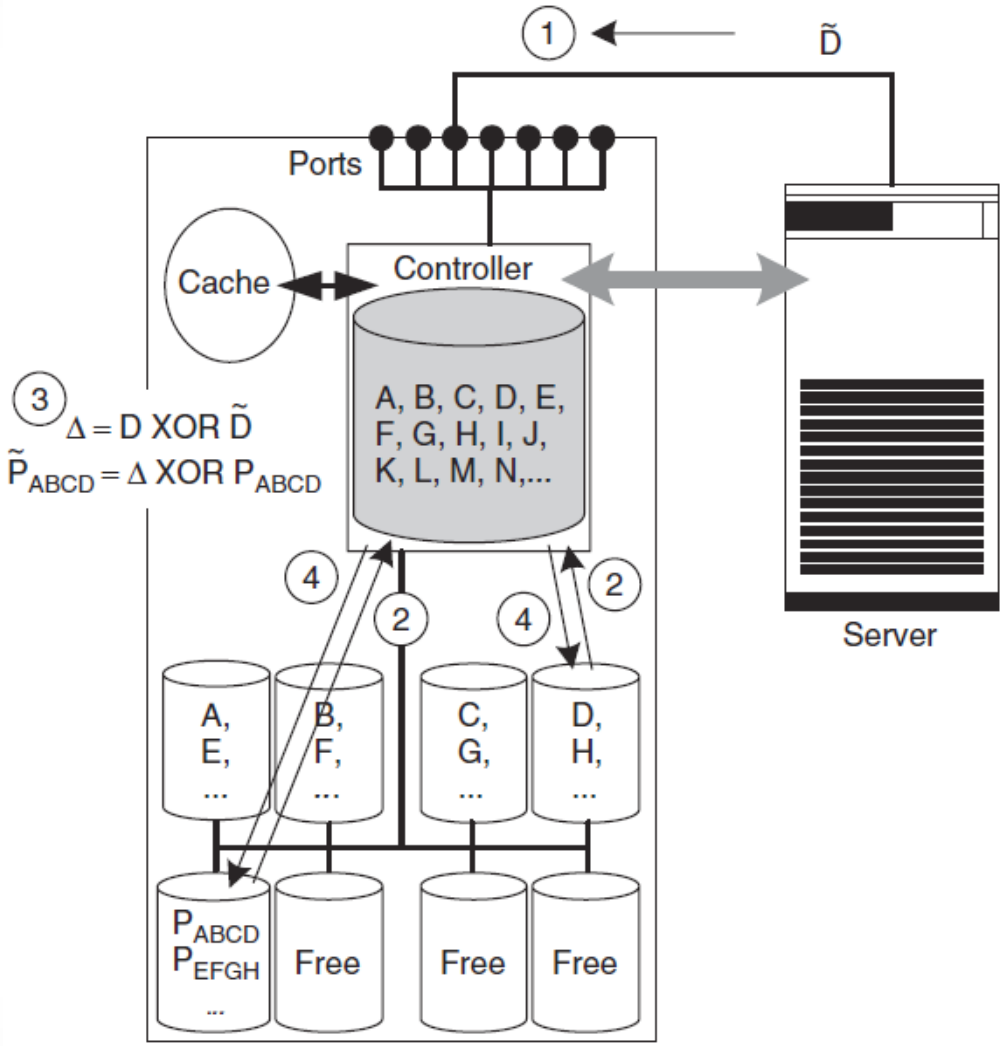
If block A was changed only,
that easy to recount P_{ABCD} , without knowing BCD.

But it need to count old block A,
that to count Δ

RAID 5 striped parity



Overheads RAID 4 and 5



RAID 6: double parity

- Recently 1TB HDD with BER 10 ^ -15 => one 100 TB sector is lost when reading
- 10 16X1T disk arrays will lose one array once a year +
- Operation mode is now 7X24
- RAID 6 uses an extra parity disk for group errors
- Cost increase
- Increase recording and correction operation time

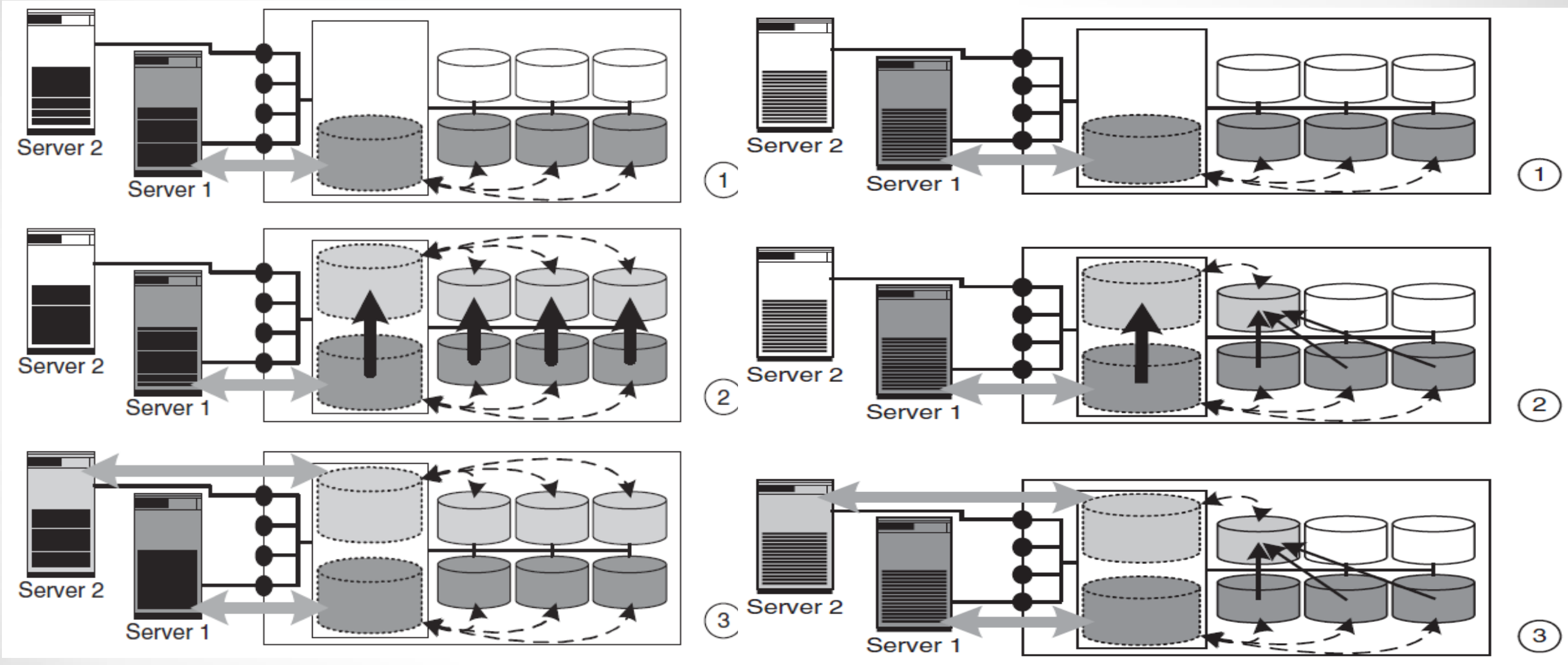
Comparison RAID scheems

RAID level	Fault-tolerance	Read performance	Write performance	Space requirement
RAID 0	None	Good	Very good	Minimal
RAID 1	High	Poor	Poor	High
RAID 10	Very high	Very good	Good	High
RAID 4	High	Good	Very very poor	Low
RAID 5	High	Good	Very poor	Low
RAID 6	Very high	Good	Very very poor	Low

CACHING: ACCELERATION OF HARD DISK ACCESS

- Cache on the hard disk
- Write cache in the disk subsystem controller
  - GB caches
  - Applications operates by blocks
  - The main point is to save the data in the cache even when power is off (UPS)
- Read cache in the disk subsystem controller

INTELLIGENT DISK SUBSYSTEMS (Instant copies)

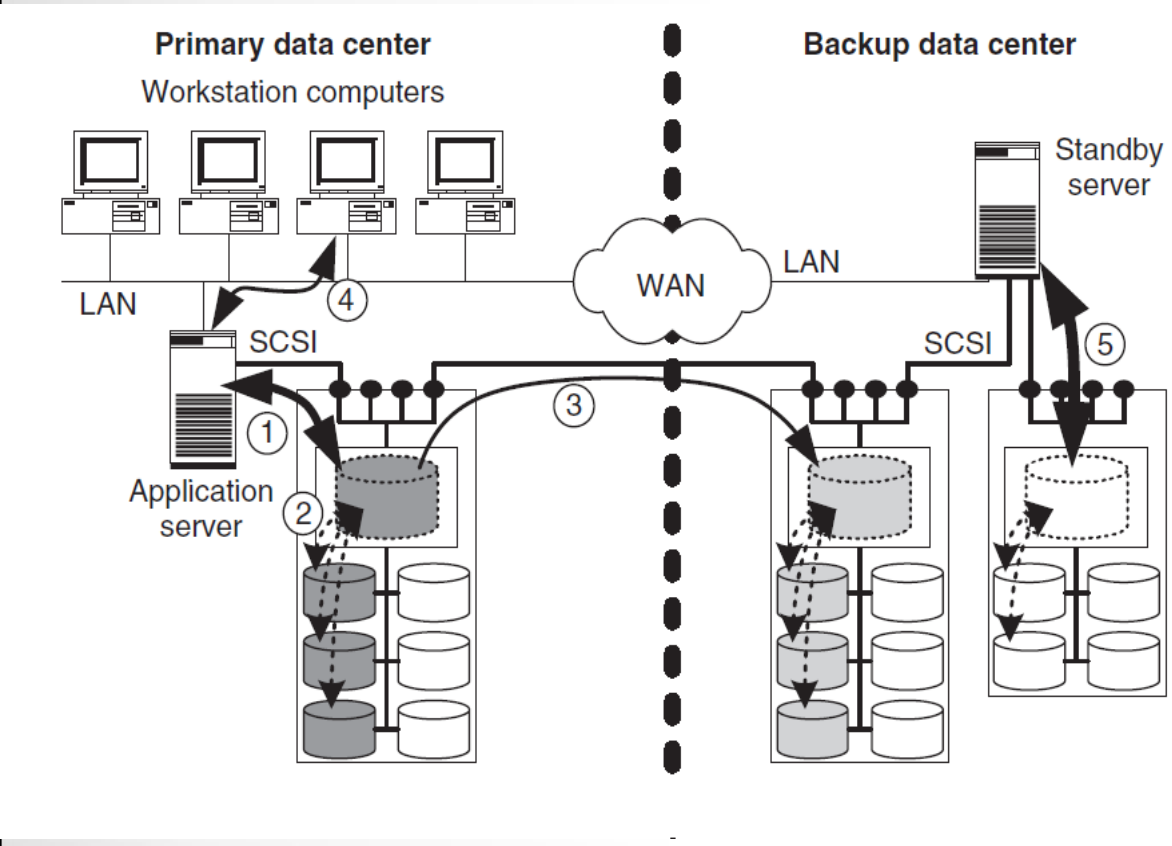


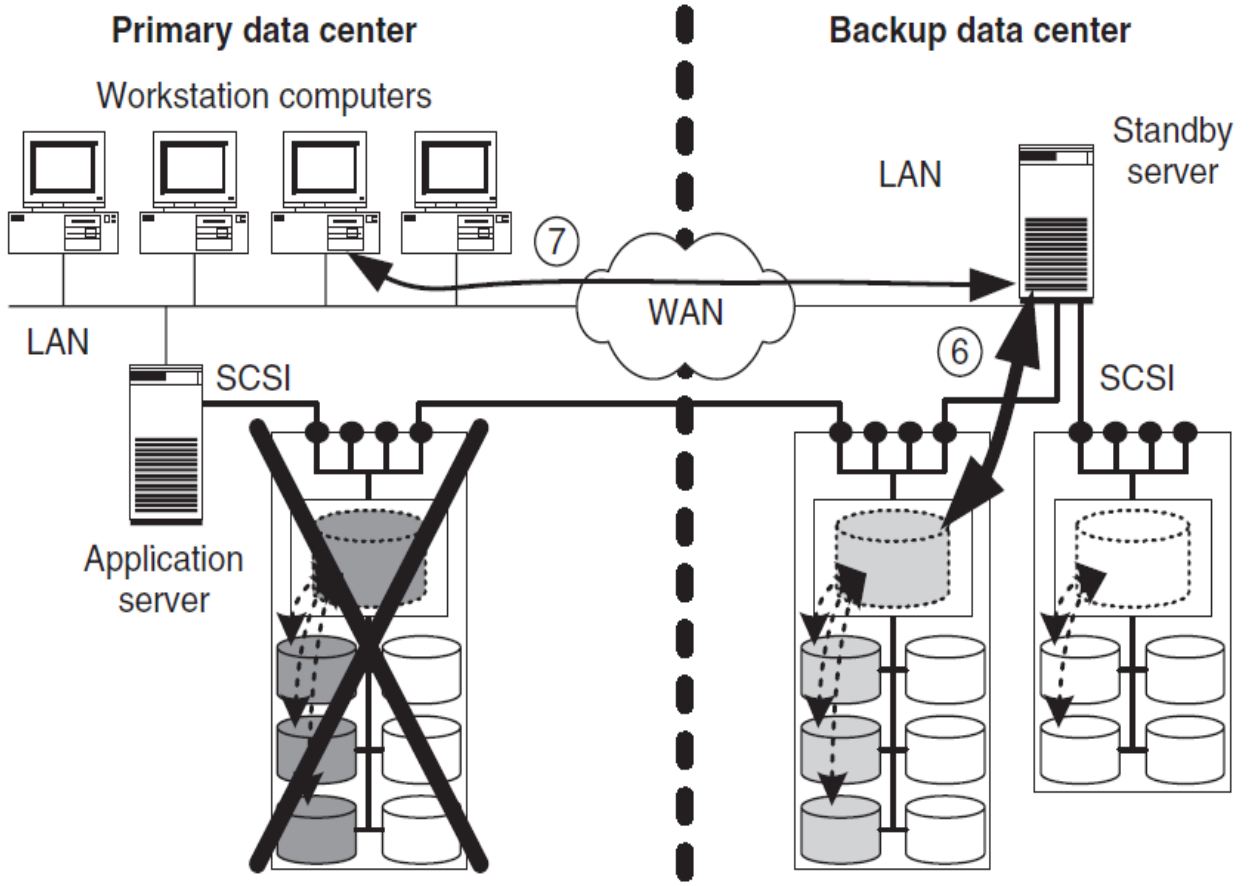
Space-efficient instant copy

Incremental instant copy

Reversal of instant copy

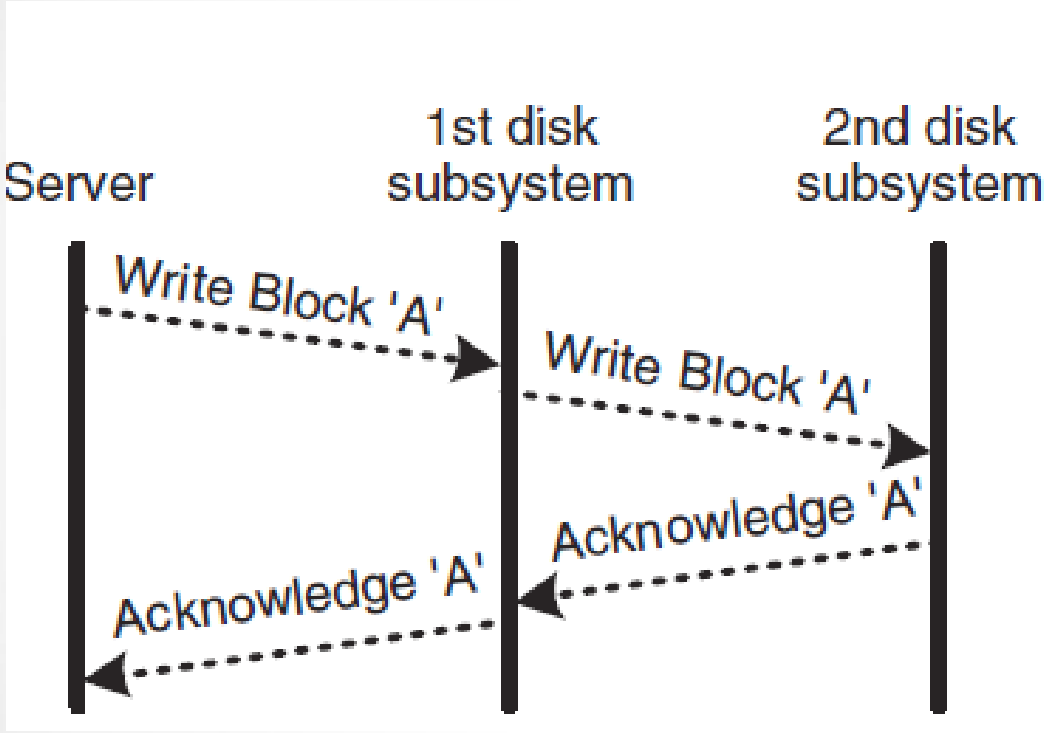
Remote mirroring (catastrophe resilience)



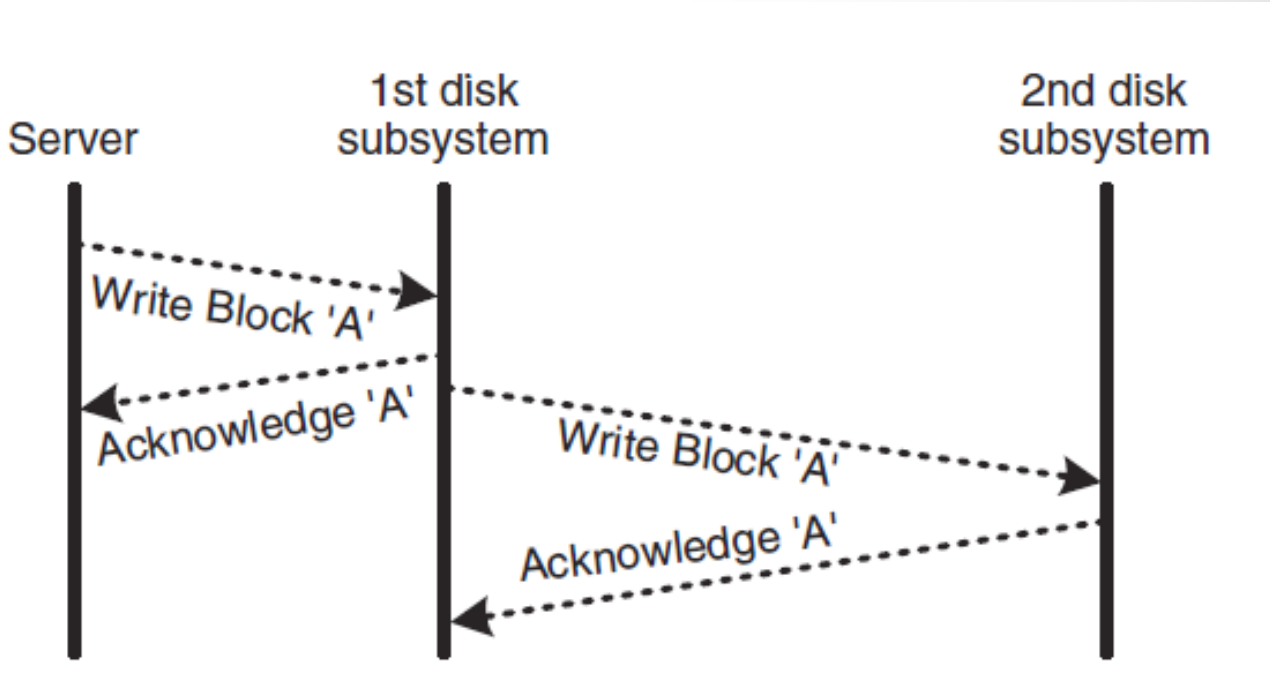


For data protection, the proximity of production data and data copies is fatal.

Synchronous vs Asynchronous remote mirroring

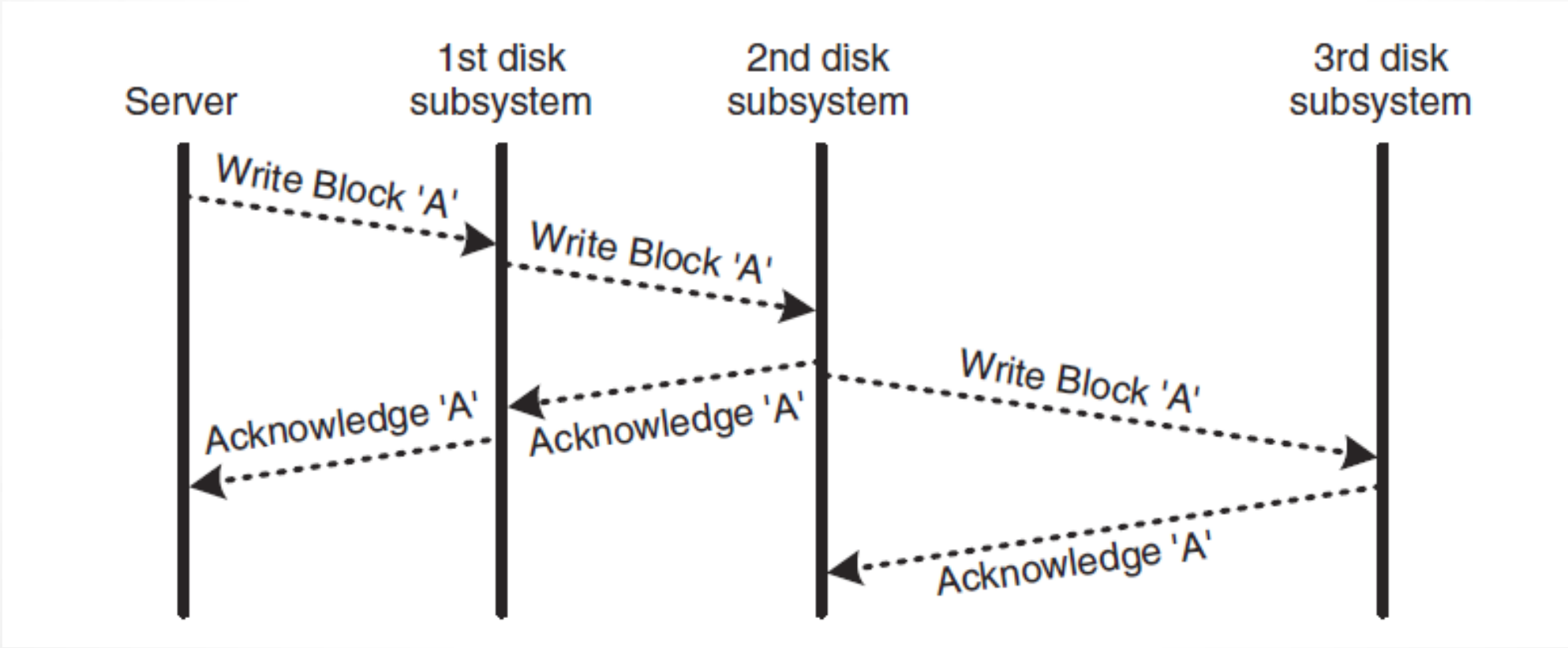


Synchronous

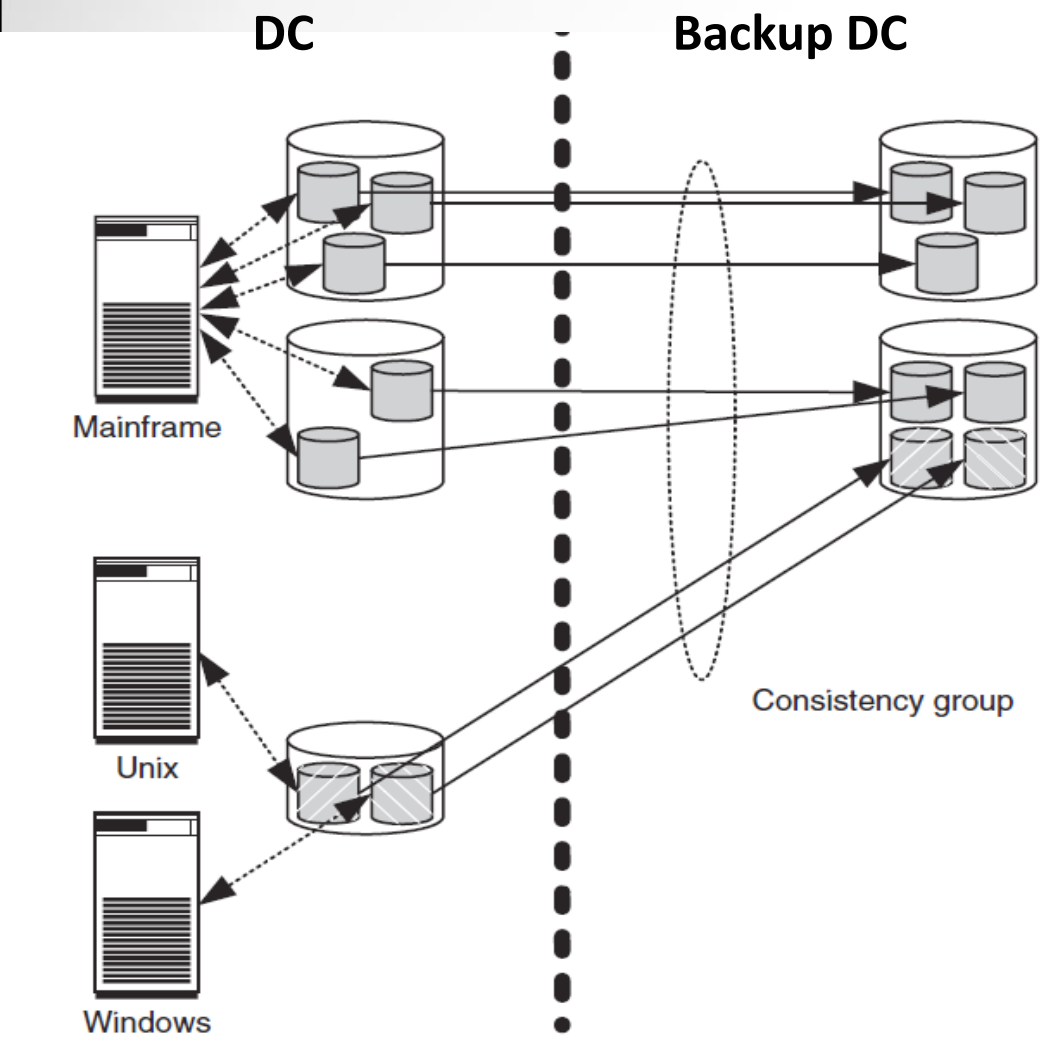


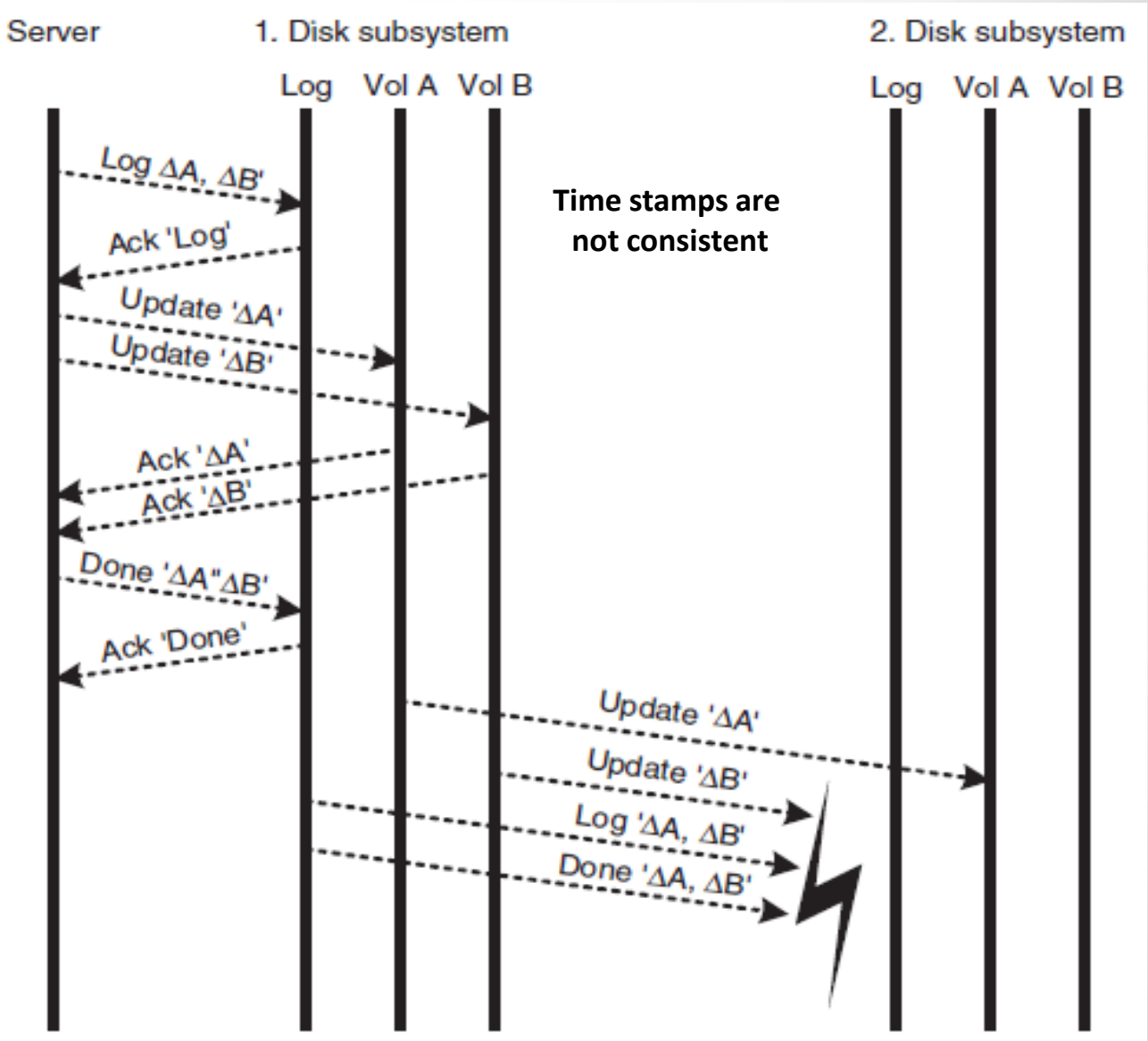
Asynchronous

Mirroring data over long distances

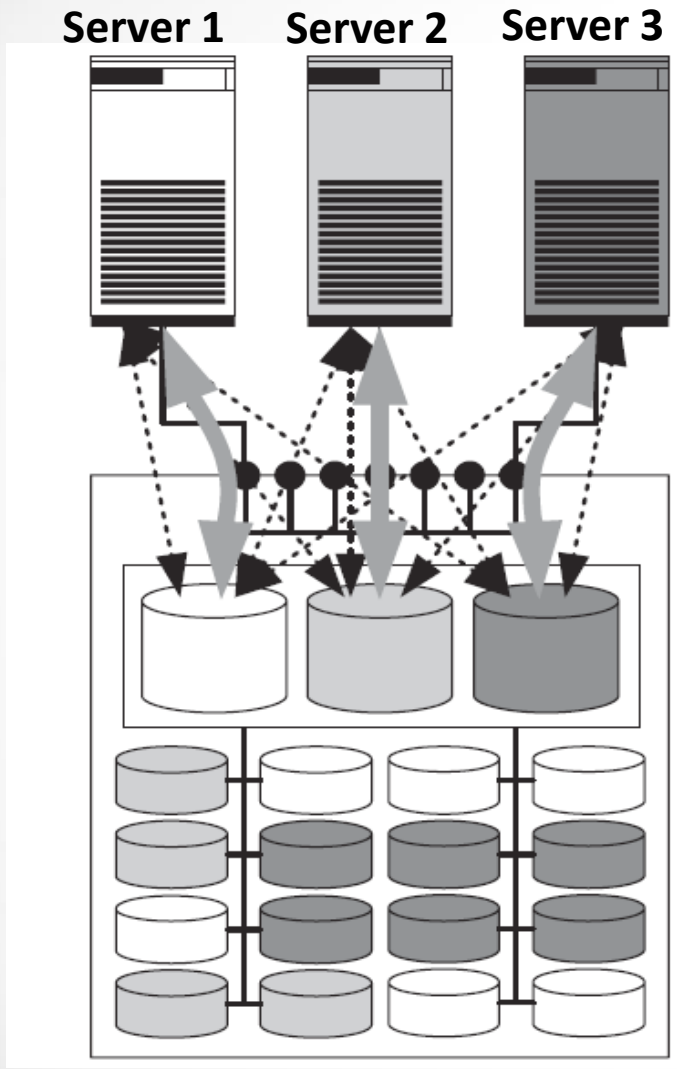


Consistency groups





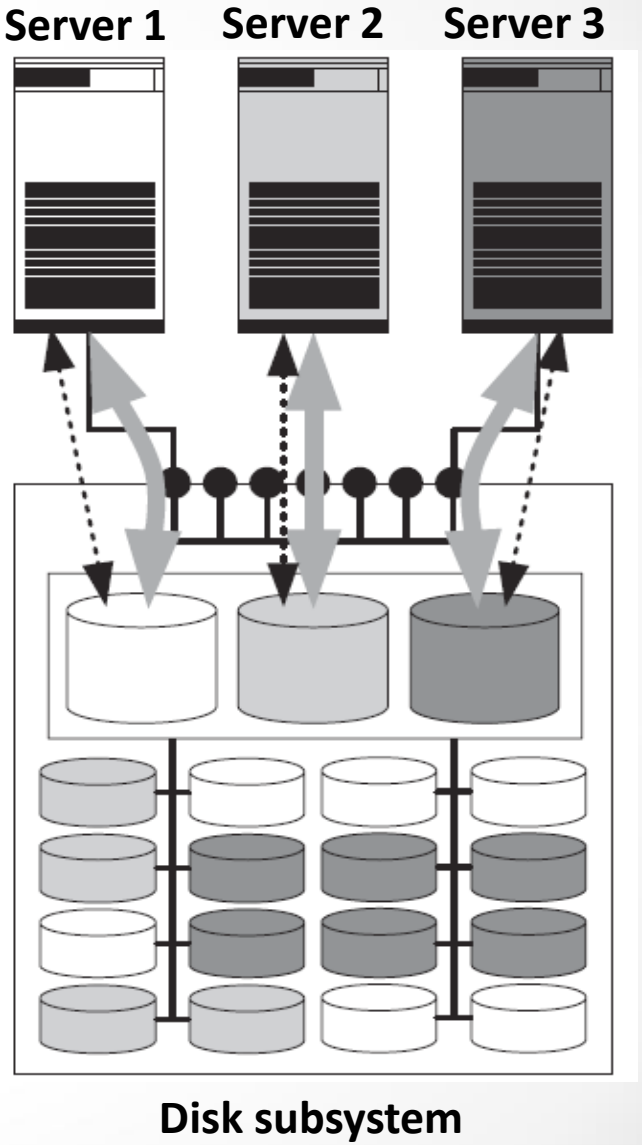
LUN masking



Logical
Unit
Number

Server uses LUN

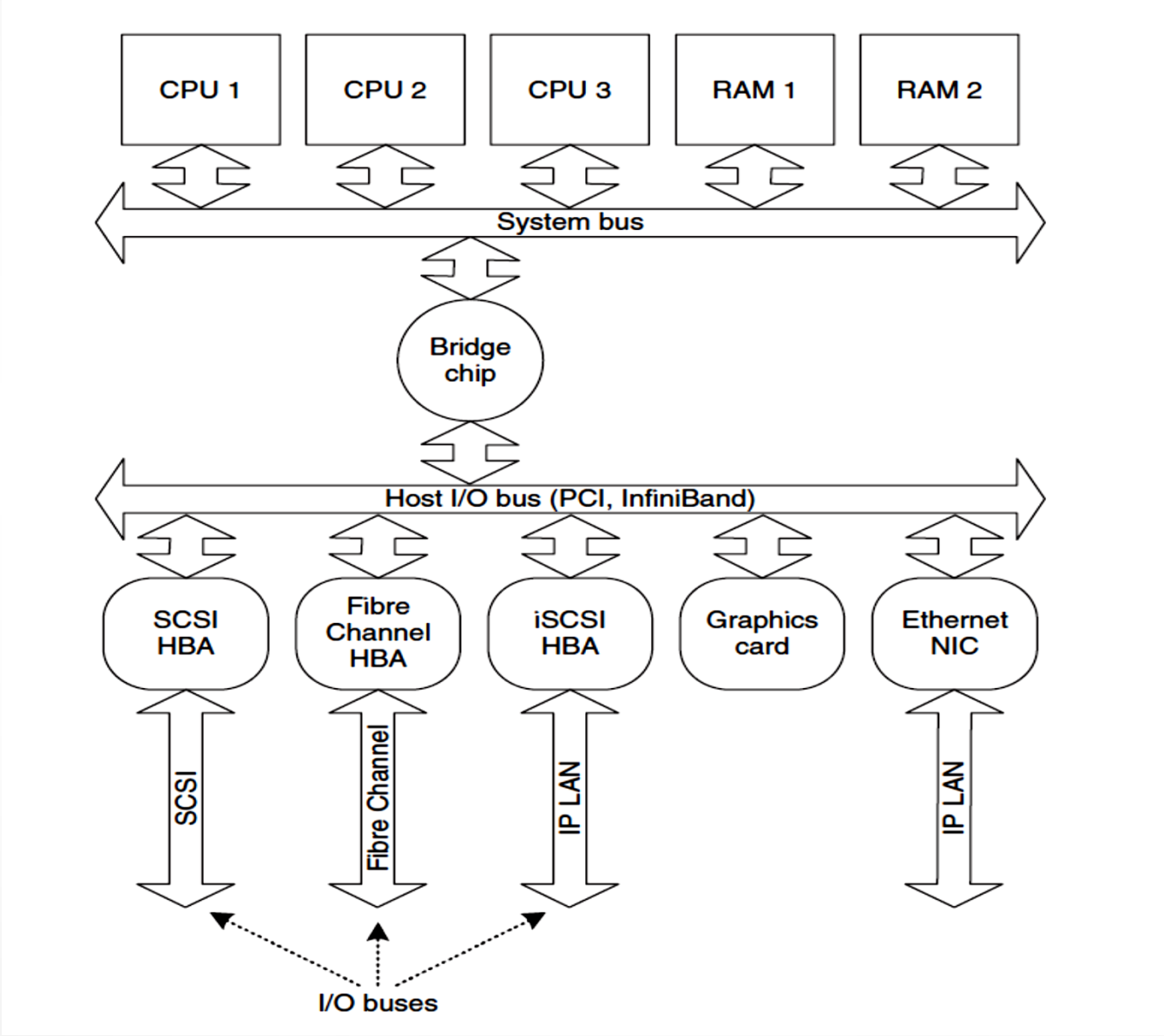
Server sees LUN



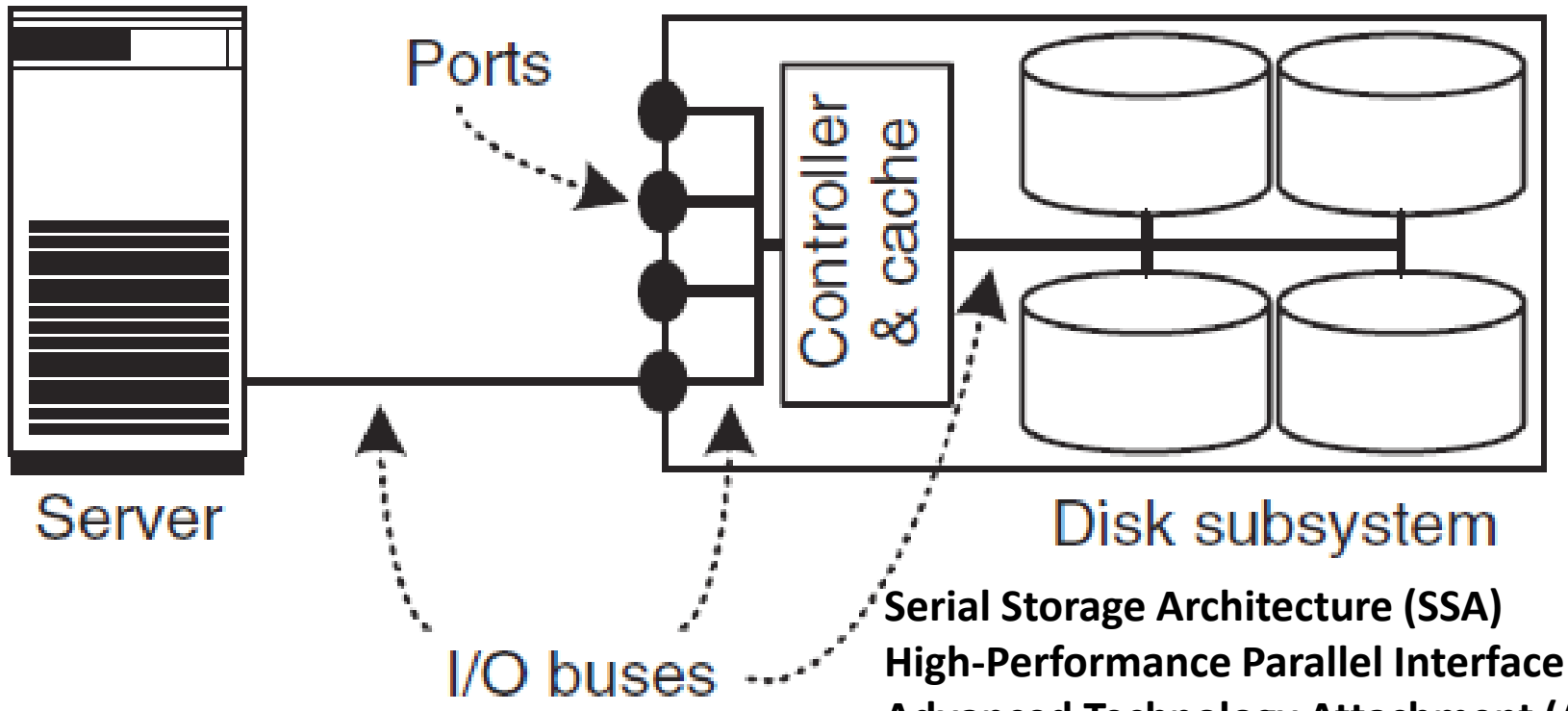
AVAILABILITY OF DISK SUBSYSTEMS

- The data is distributed across several disks using RAID mechanisms and provide redundant data (parity blocks).
- On each physical disk, the data is encoded with a Hamming code. In addition, the disk is equipped with a self-diagnostic subsystem, which controls the error rate, spindle vibration, etc. This allows you to proactively predict disk failures.
- Each disk is connected to the controller through at least two internal buses.
- The disk subsystem controller can be duplicated. The output of one instance will automatically activate the next instance. Active Standby
- Duplicated UPS cooling systems.
- DS connect to different electrical networks
- The server is connected to the DS via several lines.
- Use periodic instant copying to protect against logical errors. For example, creating an instant copy of data every hour. Then in case of failure and destruction of some table, it can be restored.
- Remote mirroring is used from physical destruction or damage to equipment (disaster tolerance). Combined with instant copying, these services guarantee data retention and consistency even for multiple virtual disks or disk subsystems.
- LUN masking protects against unauthorized access, simplifies the work of the system administrator, and protects against accidental failures in the operation of server applications and their equipment.

THE PHYSICAL I/O PATH FROM THE SERVER CPU TO THE STORAGE SYSTEM

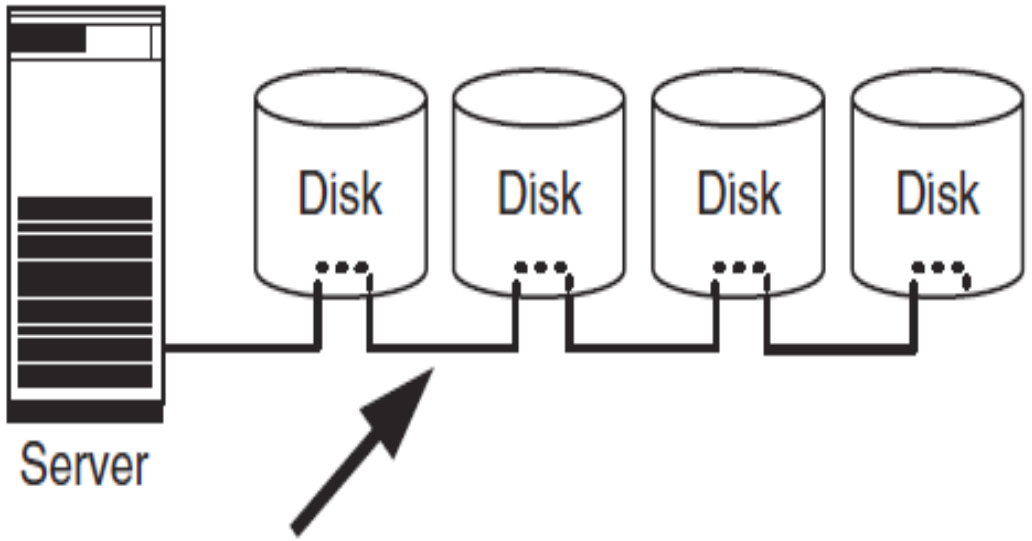


THE PHYSICAL I/O PATH FROM THE SERVER CPU TO THE DISK SUBSYSTEM



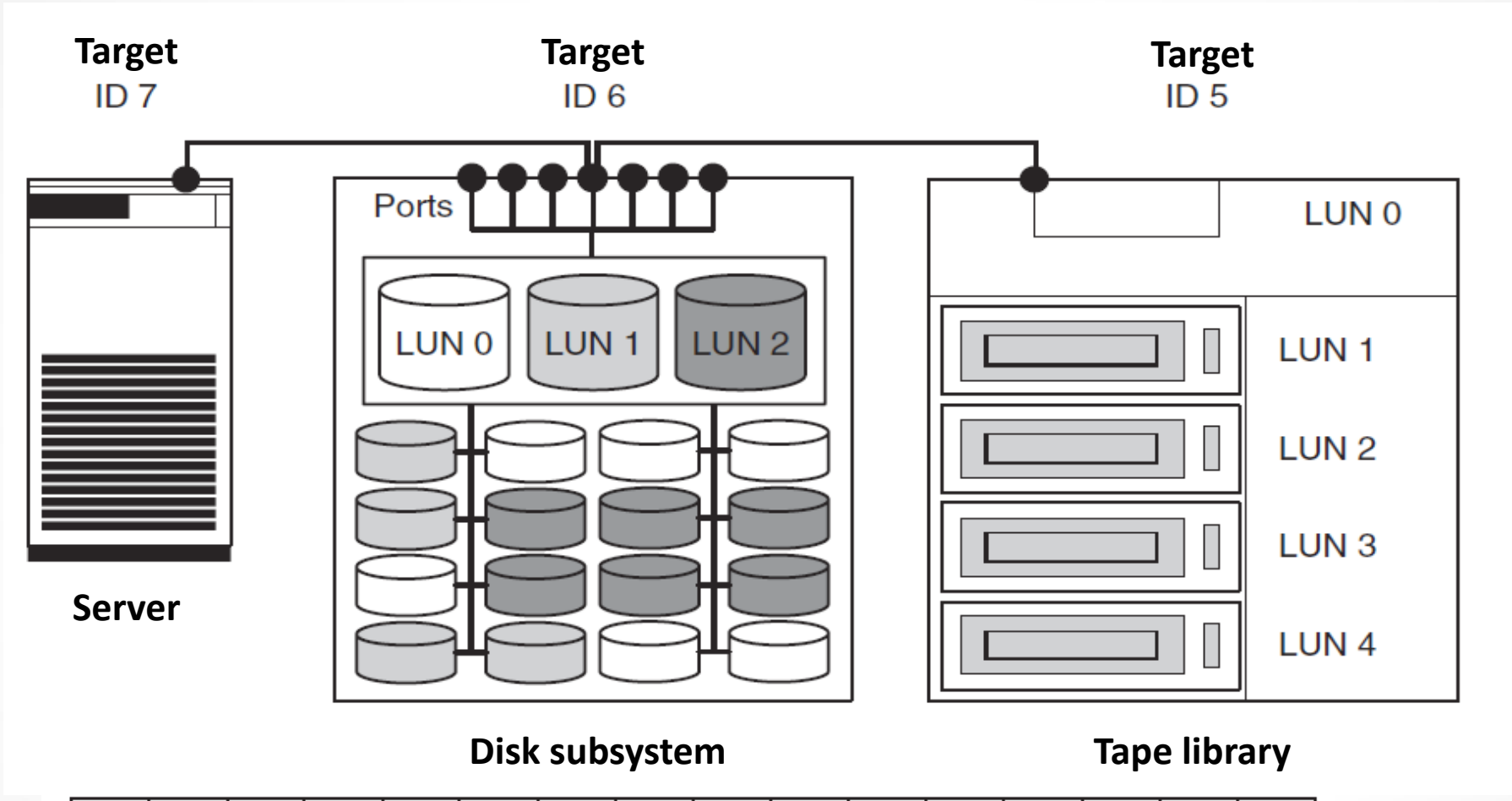
Serial Storage Architecture (SSA)
High-Performance Parallel Interface (HIPPI),
Advanced Technology Attachment (ATA),
Integrated Drive Electronics (IDE),
Serial ATA (SATA), Serial
Attached SCSI (SAS)
Universal Serial Bus (USB).

Small Computer System Interface (SCSI)

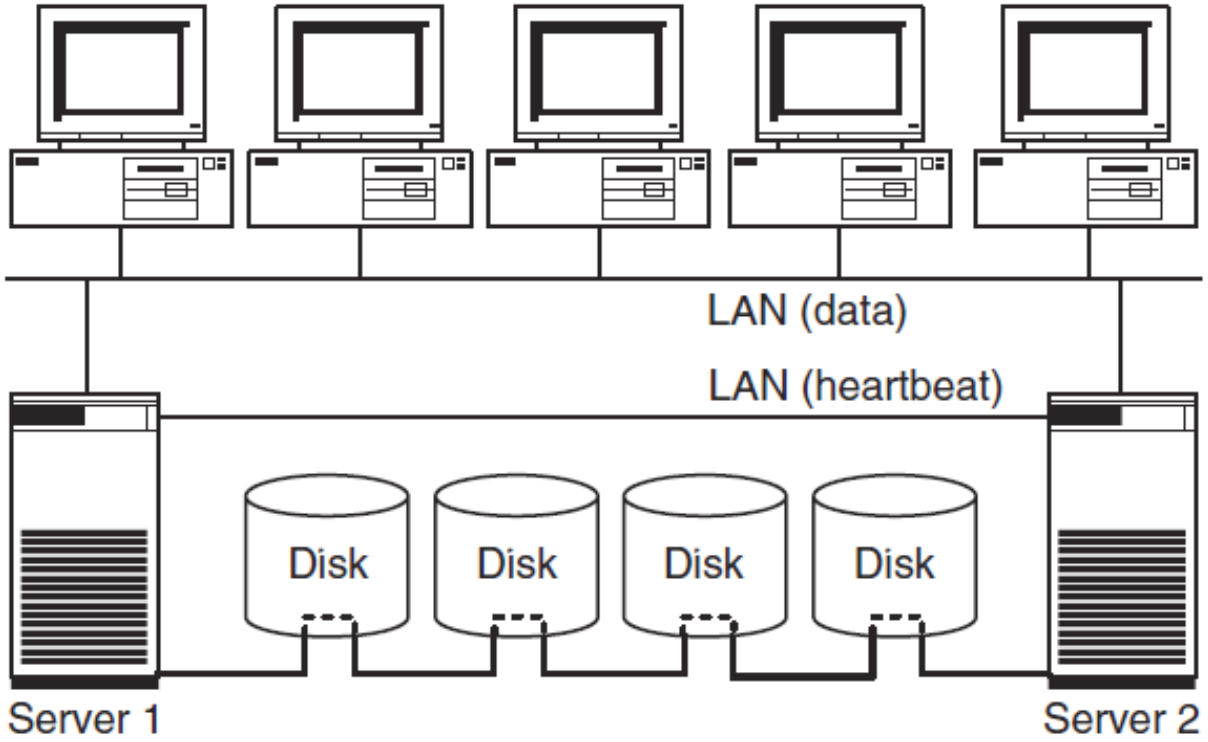


SCSI version	MByte/s MByte/s	Bus width	Max. no. of devices
SCSI-2	5	8	8
Wide Ultra SCSI	40	16	16
Wide Ultra SCSI	40	16	8
Wide Ultra SCSI	40	16	4
Ultra2 SCSI	40	8	8
Wide Ultra2 SCSI	80	16	16
Ultra3 SCSI	160	16	16
Ultra320 SCSI	320	16	16

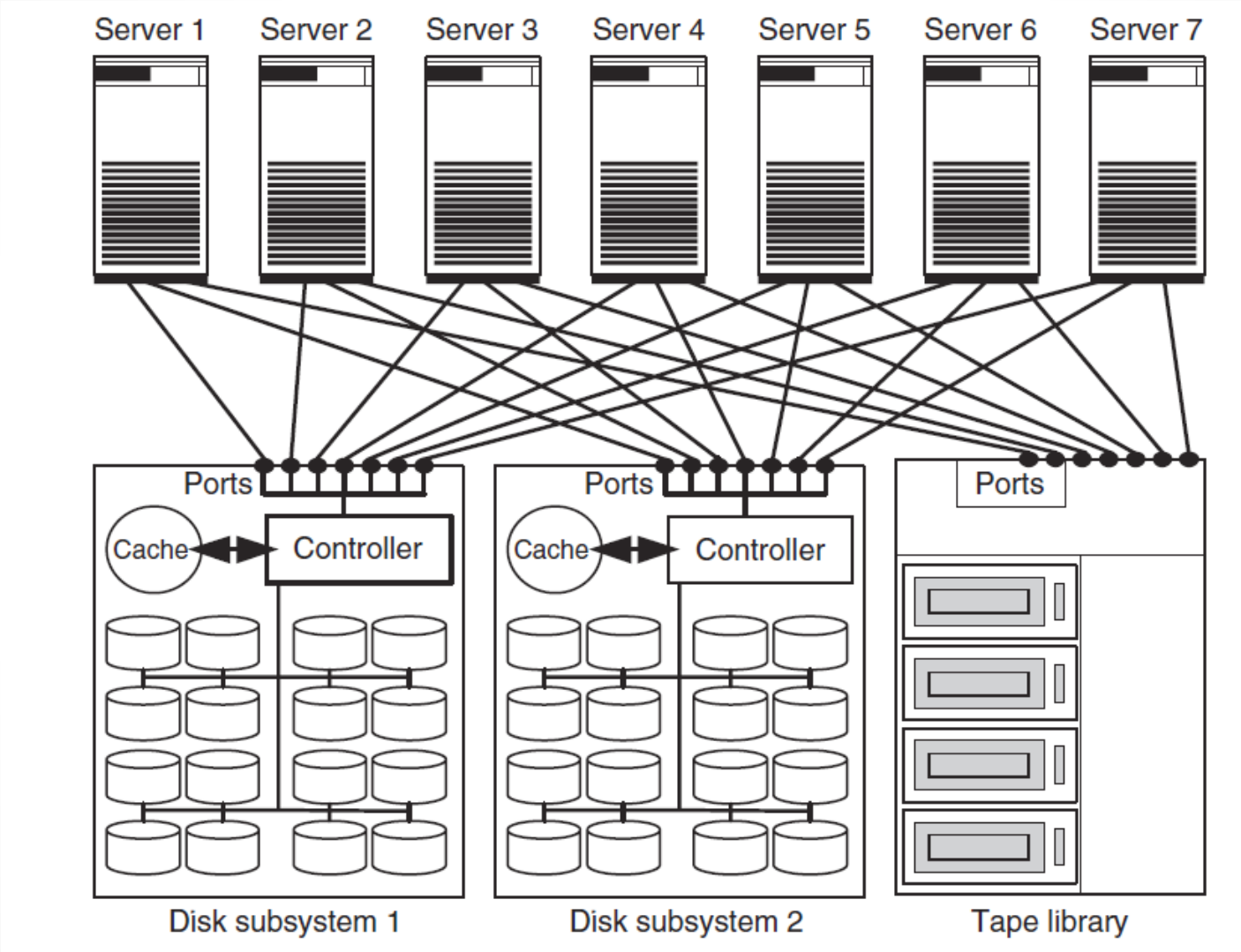
Device addressing on SCSI



SCSI storage networks



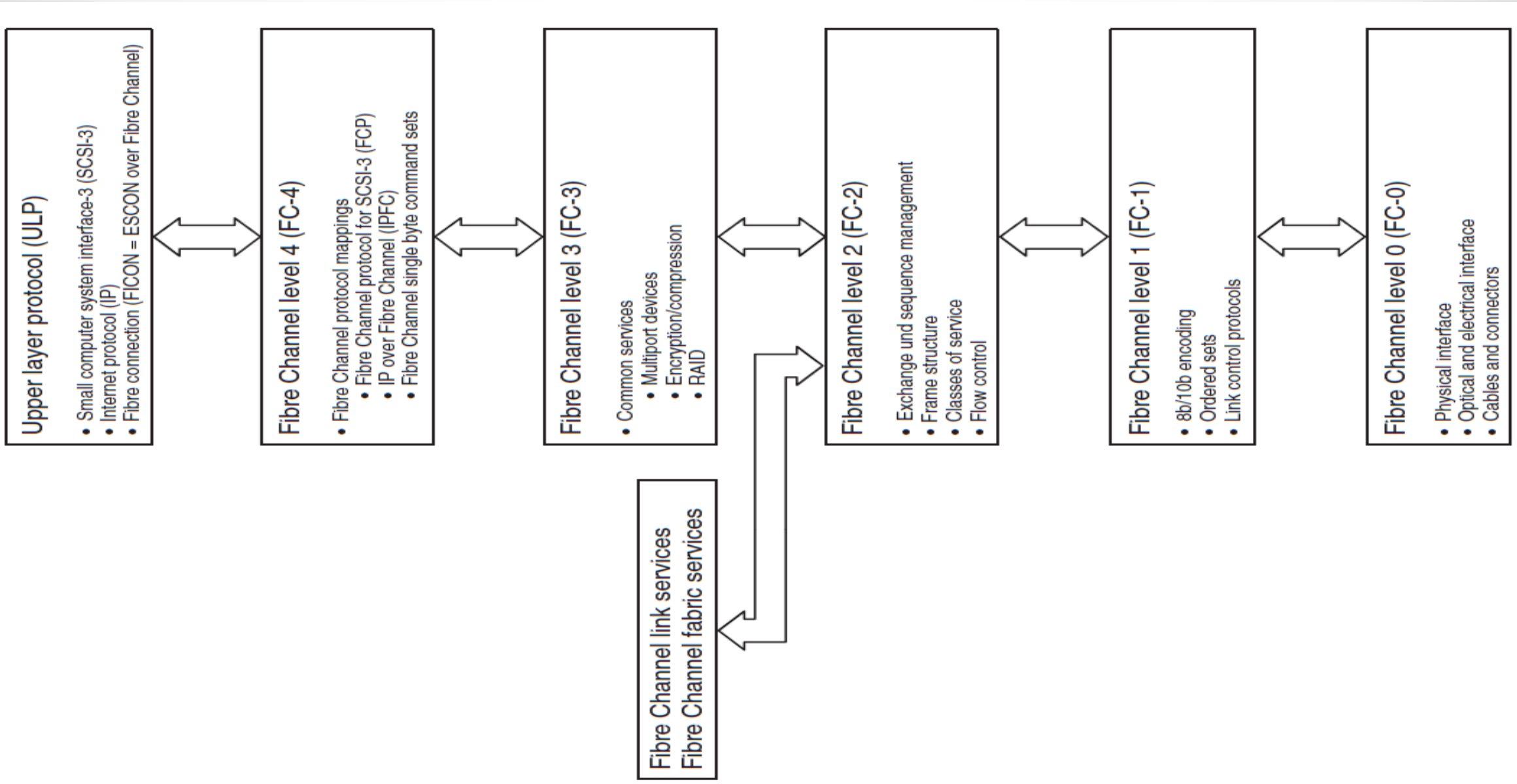
SCSI SAN with multiport storage system



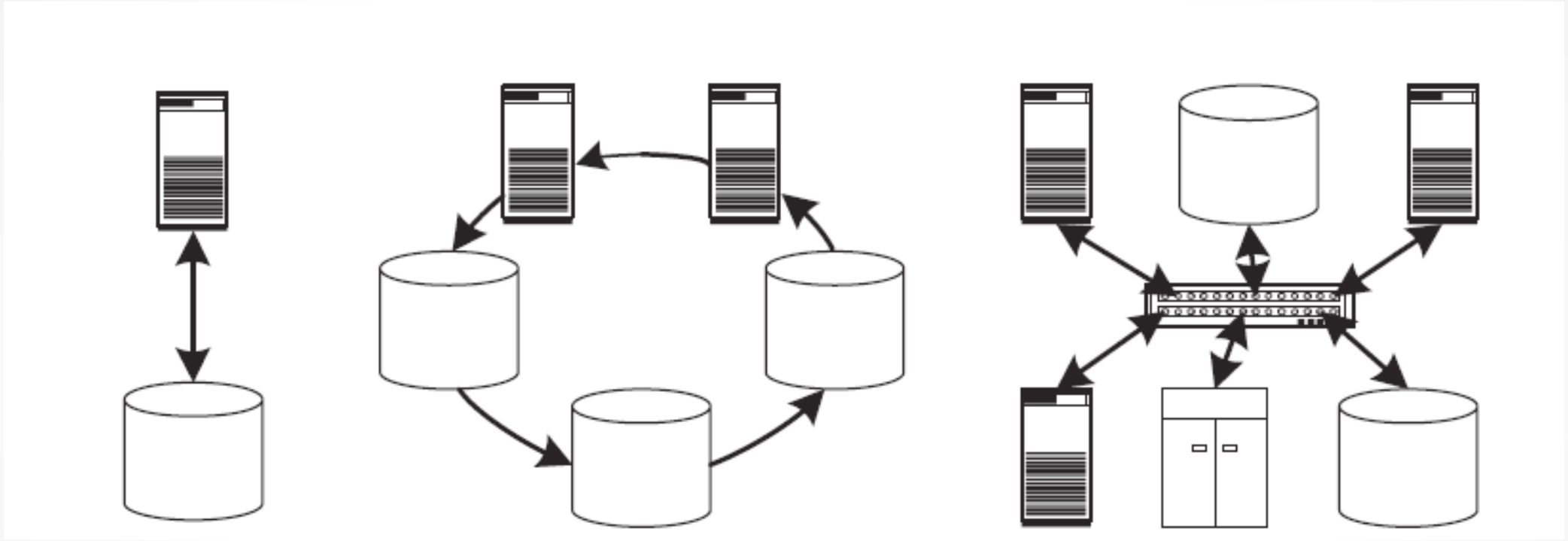
Fibre Channel (2009)

- Originally developed as a backbone technology for the connection of LANs
  - Serial transmission for high speed and long distances;
  - Low rate of transmission errors;
  - Low delay (latency) of the transmitted data;
  - Implementation of the Fibre Channel Protocol (FCP) in hardware on HBA cards to free up the server CPUs
- Fiber Channel IPI (Intelligent Peripheral Interface), SCSI, HIPPI (High Performance Parallel Interface), ATM, IP и 802.2 (Ethernet).
- Fibre Channel - n x100 Mbps over 10 km,
  where n – number of channels. Bandwidth limit - 4,25 Gbps.
- Physical environment
  - Fiber
  - Copper cable as coax as twisted pair (less 200 MBps)

Fiber Channel protocol stack



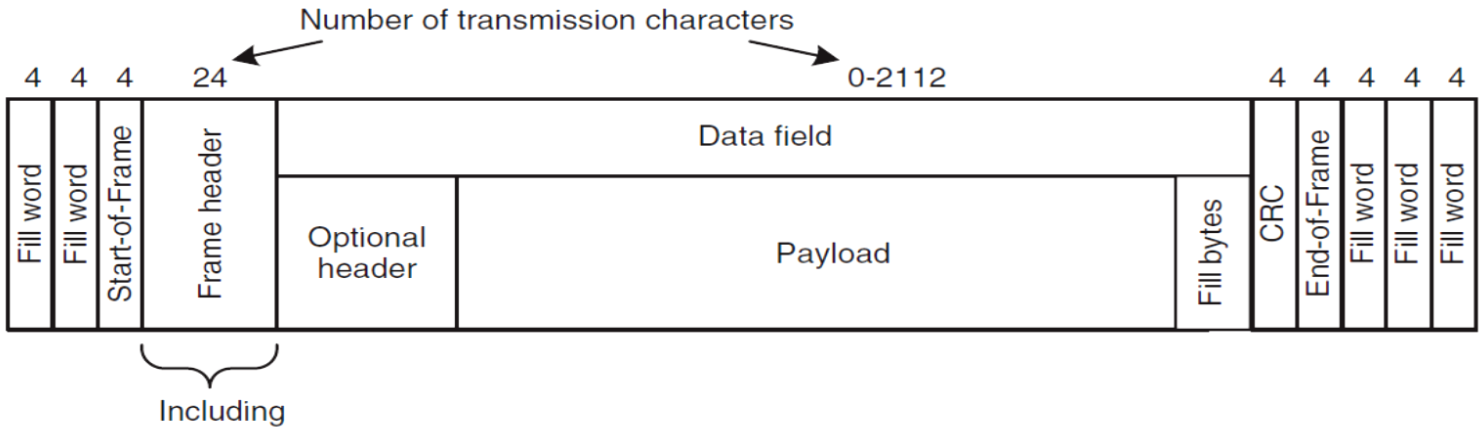
Links, ports and topologies



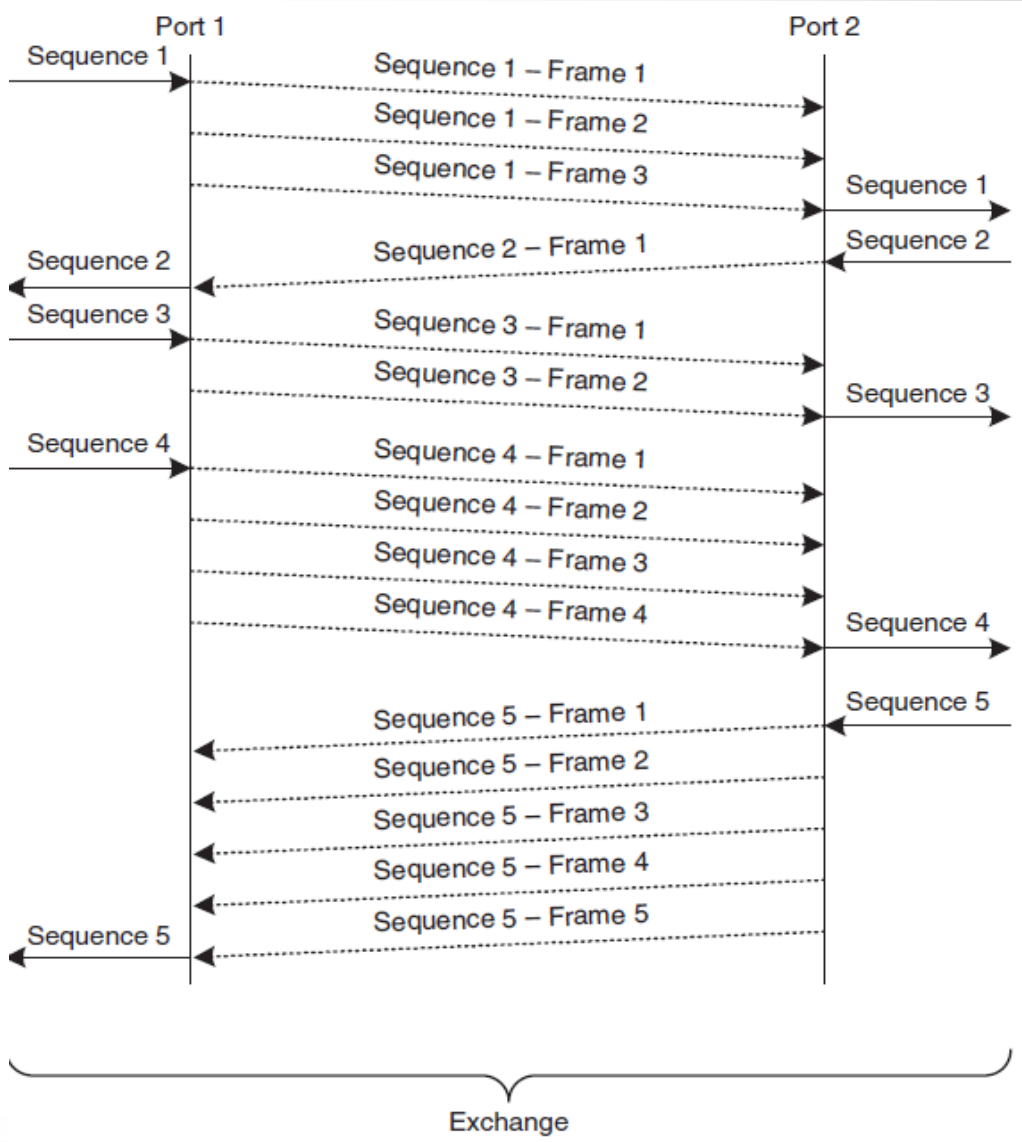
Fibre Channel: port types

- N-Port – describes the capability of a port as an end device (server, storage device), also called node, to participate in the fabric topology or to participate in the point-to-point topology as a partner.
- F-Port – F-Ports are the counterpart to N-Ports in the Fibre Channel switch.
- L-Port – describes the capability of a port to participate in the arbitrated loop topology as an end device (server, storage device).
- NL-Port – capable operate as an N-Port as an L-Port.
- FL-Port – allows a fabric to connect to a loop.
- E-Port – transmit the data from end devices that are connected to two different Fibre Channel switches. Two Fibre Channel switches are connected together by E-Ports.
- G-Port – can operate as E as FL depend on port configuration.
- B-Port – for connecting two FC switches via ATM, SDH, Ethernet or IP, e.g. two FC SAN could be connected through WAN (Bridge port).

FC-2: data transfer

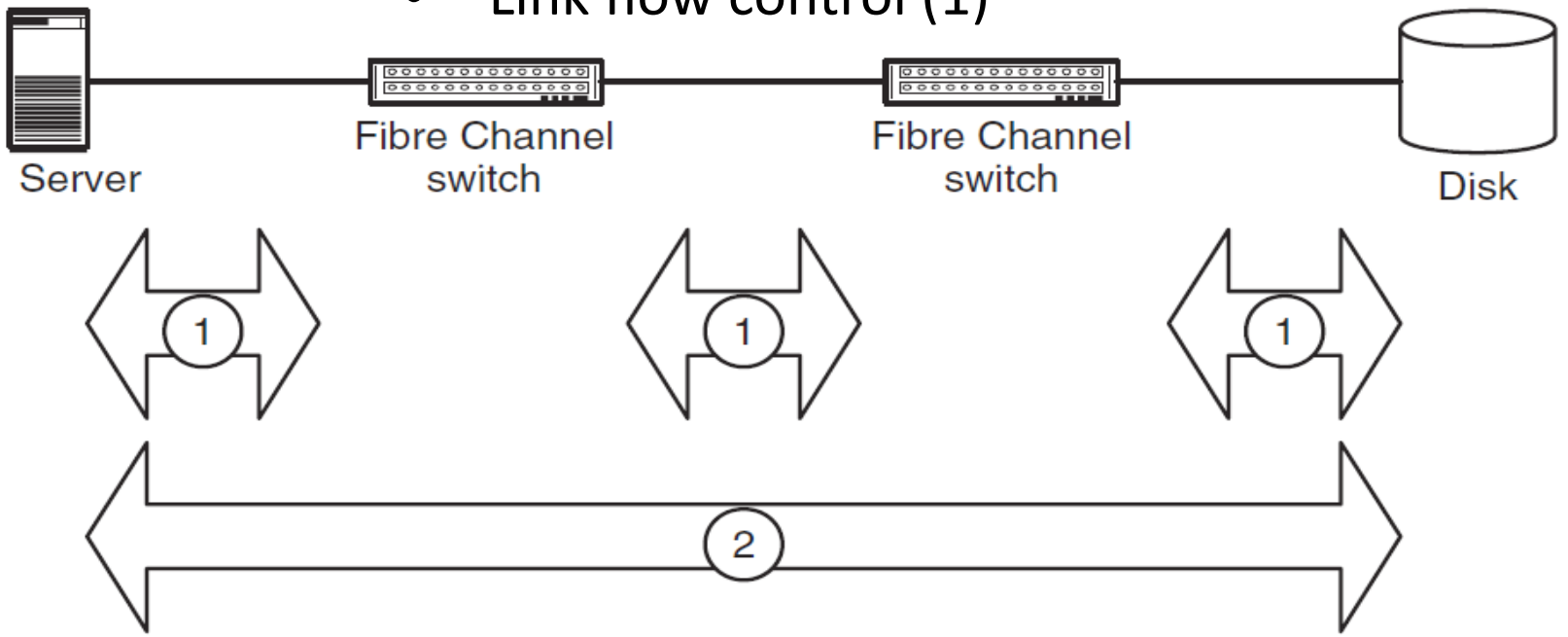


- Frame Destination Address (D_ID)
- Frame Source Address (S_ID)
- Sequence ID
- Number of the frame within the sequence
- Exchange ID



FC-2: Flow control

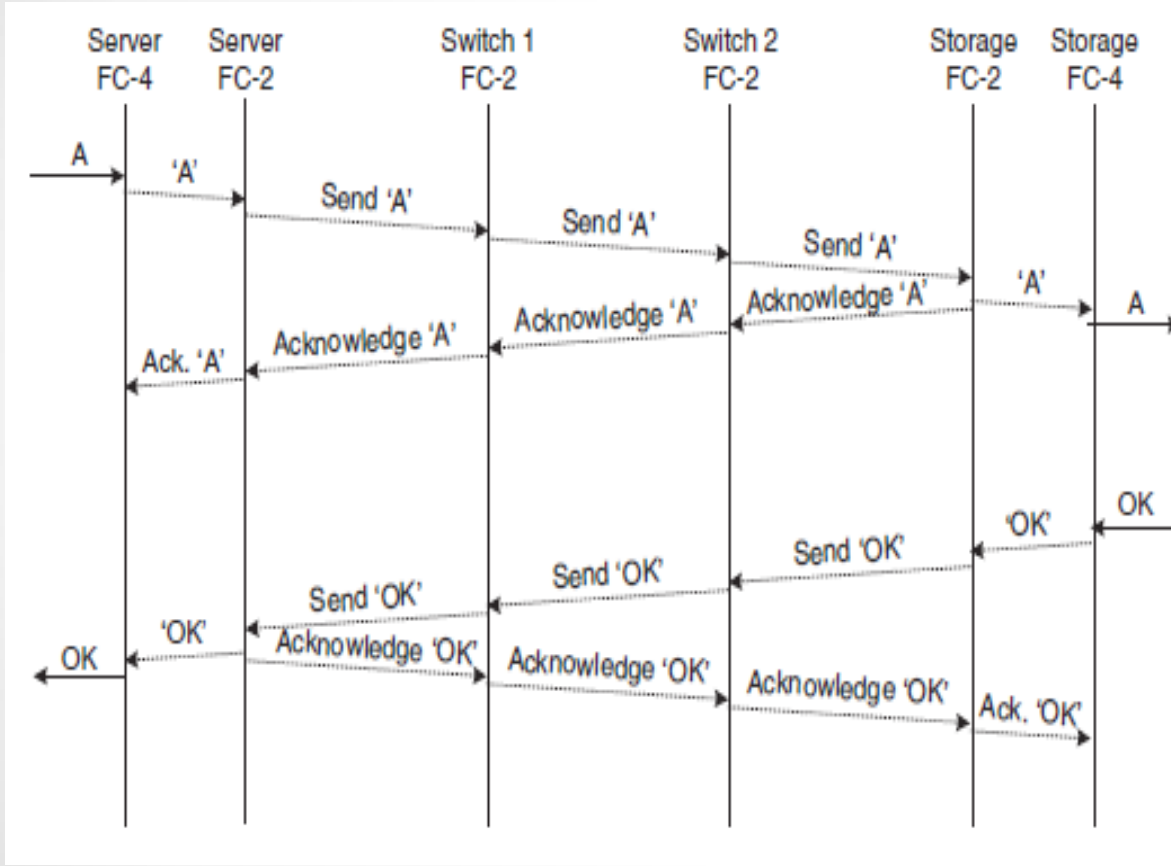
- Credential scheem
- E2E flow control (2)
- Link flow control (1)

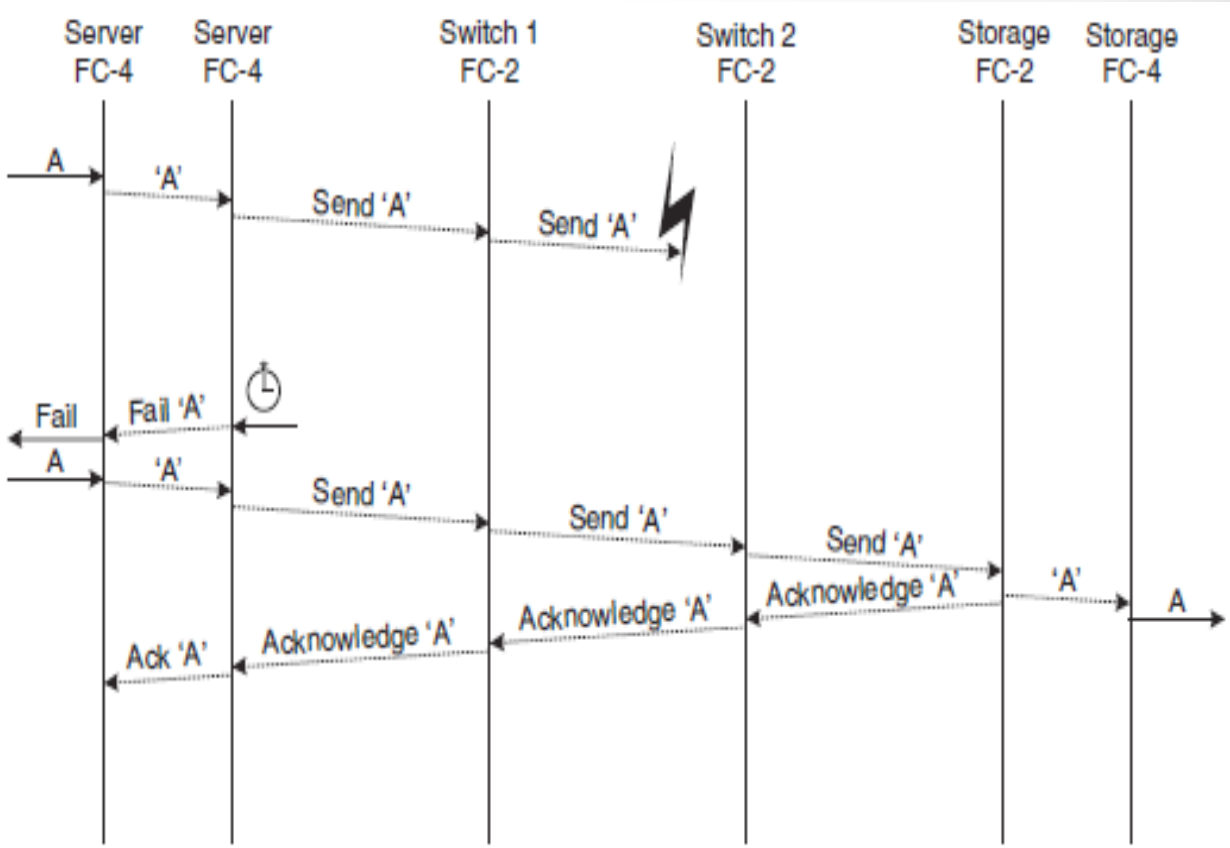


FC-2: Service classes

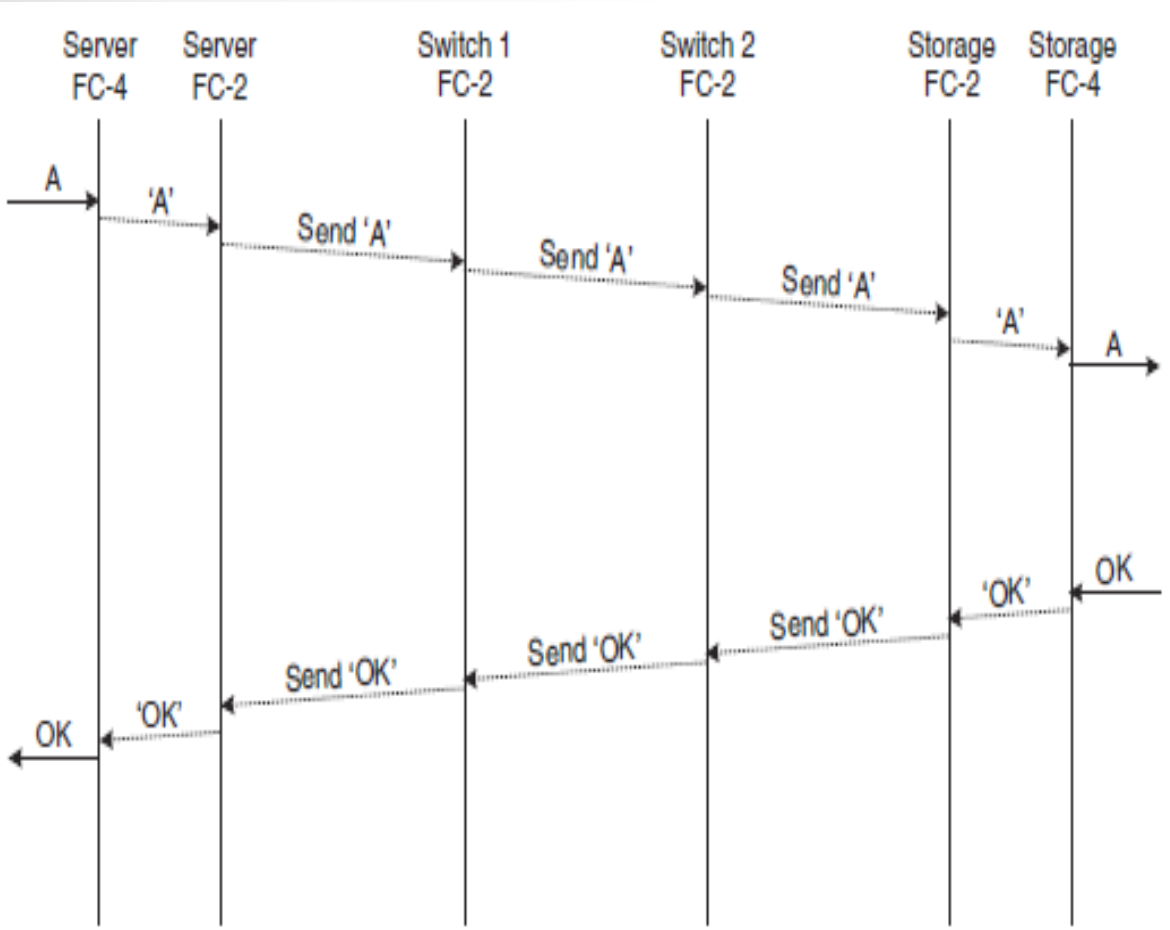
Class 1	A point-to-point connection (end-to-end) between ports of type n_port through circuit switching.
Class 2	Connectionless packet switched, which guarantees delivery of data. A port can communicate simultaneously with any number of ports of type n_port in duplex mode. The order of frame delivery is not guaranteed, except for P2P or XA connection. There is flow control. This class is typical for local area networks, where the data delivery time is not critical.
Class 3	Exchange of datagrams without a connection and without a delivery guarantee. There is flow control. Applies to SCSI channels.
Class 4	Provides the allocation of a certain fraction of the channel capacity with a given quality of service (QoS). Topology only matrix with n_port. The order of delivery of frames is guaranteed.
Class 5	Regulatory documents are in preparation.
Class 6	Provides group-service with switching.

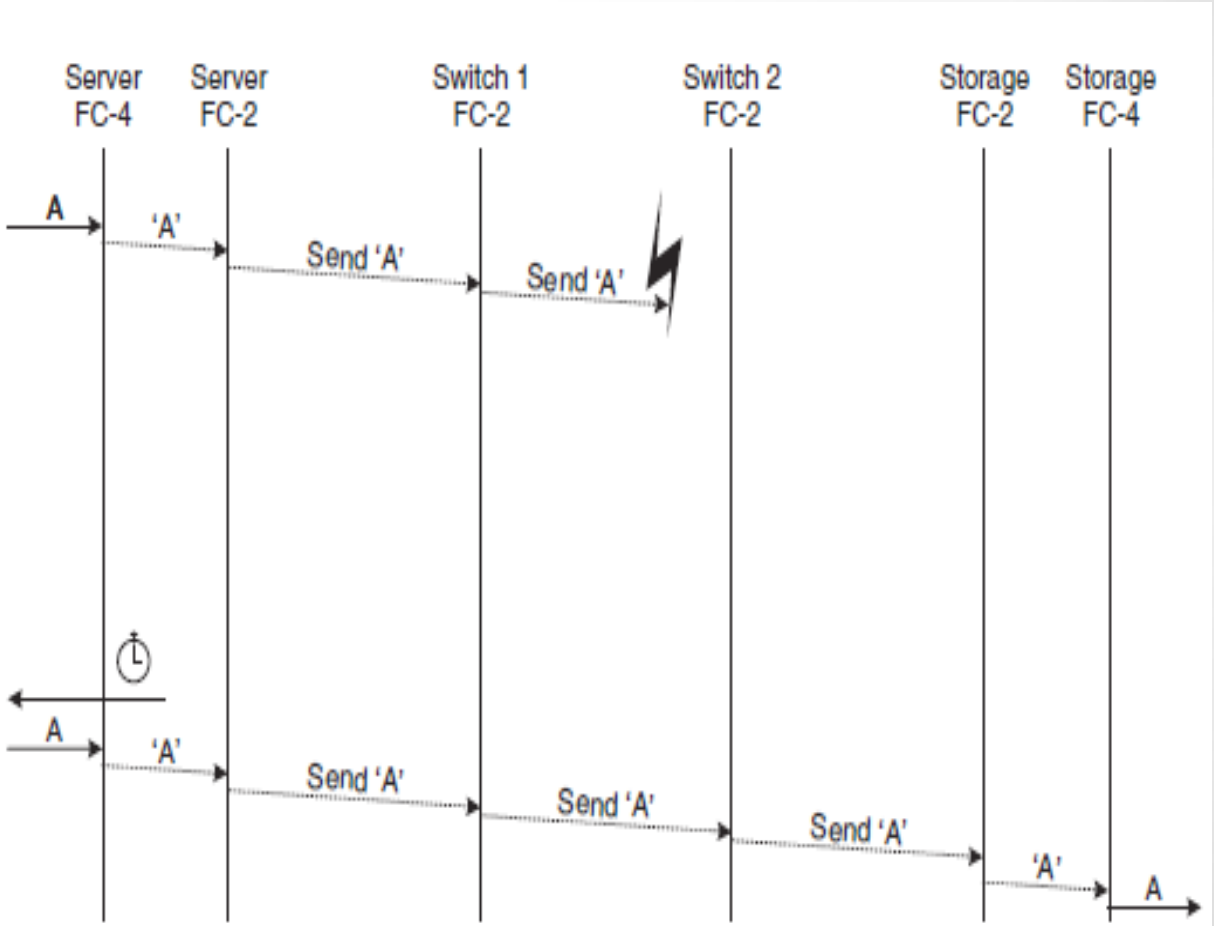
FC-2 class 2


Server FC-4
Server FC-2
Switch 1 FC-2
Switch 2 FC-2
Storage FC-2
Storage FC-4
A
'A'
Send 'A'
Send 'A'
Send 'A'
'A'
A
Acknowledge 'A'
Acknowledge 'A'
Acknowledge 'A'
Ack. 'A'
OK
'OK'
Send 'OK'
Send 'OK'
Send 'OK'
'OK'
OK
Acknowledge 'OK'
Acknowledge 'OK'
Acknowledge 'OK'
Ack. 'OK'


Server FC-4
Server FC-4
Switch 1 FC-2
Switch 2 FC-2
Storage FC-2
Storage FC-4
A
'A'
Send 'A'
Send 'A'
Fail
Fail 'A'
A
'A'
Send 'A'
Send 'A'
Send 'A'
'A'
A
Acknowledge 'A'
Acknowledge 'A'
Ack 'A'
Acknowledge 'A'

FC-2: class 3

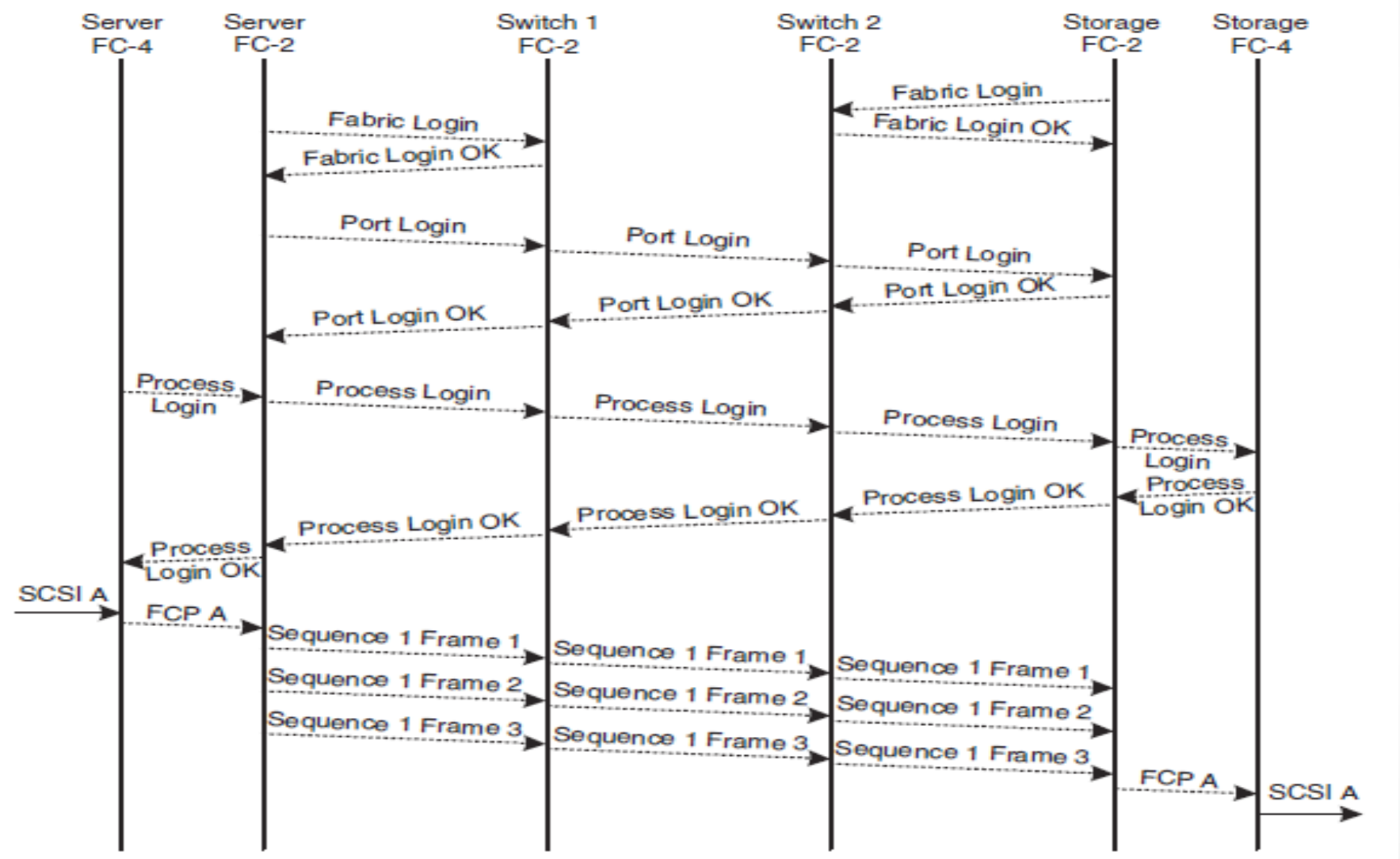

Server FC-4
Server FC-2
Switch 1 FC-2
Switch 2 FC-2
Storage FC-2
Storage FC-4
A
'A'
Send 'A'
Send 'A'
Send 'A'
'A'
A
OK
'OK'
Send 'OK'
Send 'OK'
Send 'OK'
'OK'
OK


Server FC-4
Server FC-2
Switch 1 FC-2
Switch 2 FC-2
Storage FC-2
Storage FC-4
A
'A'
Send 'A'
Send 'A'
A
'A'
Send 'A'
Send 'A'
Send 'A'
'A'
A

FC-3: services

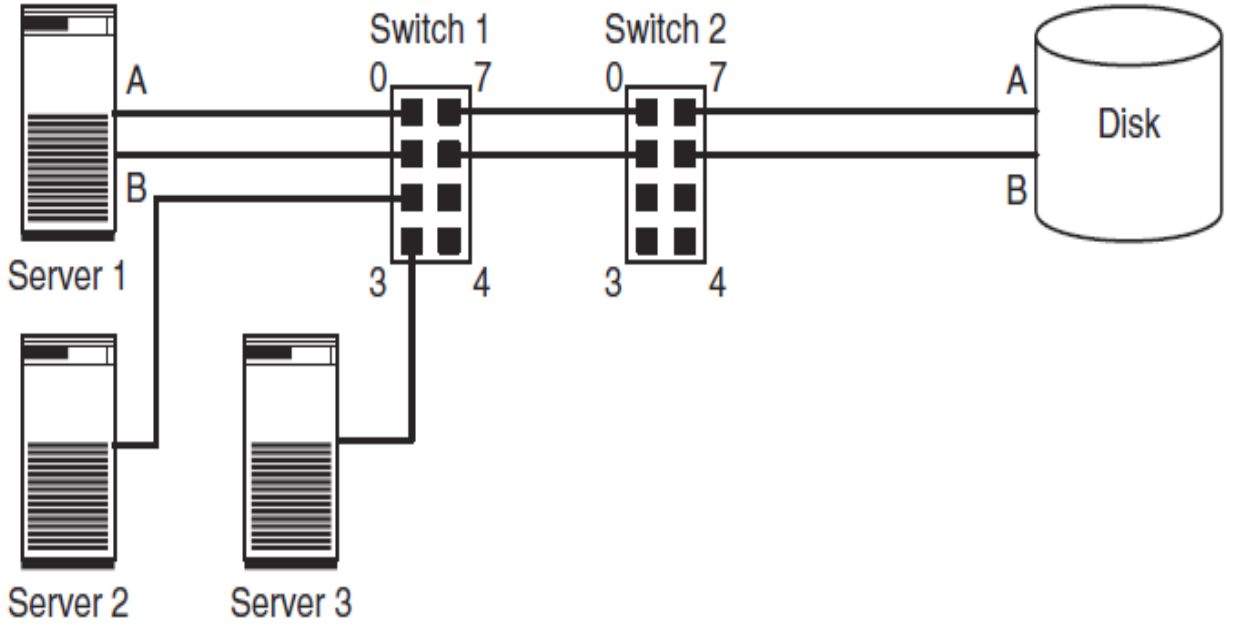
- Striping manages several paths between multiport end devices. Striping could distribute the frames of an exchange over several ports and thus increase the throughput between the two devices.
- Multipathing combines several paths between two multiport end devices to form a logical path group. Failure or overloading of a path can be hidden from the higher protocol layers.
- Compressing the data to be transmitted, preferably realized in the hardware on the HBA.
- Encryption of the data to be transmitted, preferably realized in the hardware on the HBA.
- Mirroring and other RAID levels are the last example that are mentioned in the Fibre Channel standard as possible functions of FC-3.

Line services: Identifying and Addressing



Addressing

- Names and addresses in FC
- All devices FC network have 64 bit. names
- WWN vs FCN
- WWN: WWPN. WWNN (Node Name)_
- FLOG – 24 bit port address
- S_ID vs D_ID
- AL - 8 bit AL_PA (Arbitrated Loop Physical Address)



Port_ID	WWPN	WWNN	Device
010000	20000003 EAFE2C31	2100000C EAFE2C31	Server 1, Port A
010100	20000003 C10E8CC2	2100000C EAFE2C31	Server 1, Port B
010200	10000007 FE667122	10000007 FE667122	Server 2
010300	20000003 3CCD4431	2100000A EA331231	Server 3
020600	20000003 EAFE4C31	50000003 214CC4EF	Disk, Port B
020700	20000003 EAFE8C31	50000003 214CC4EF	Disk, Port A

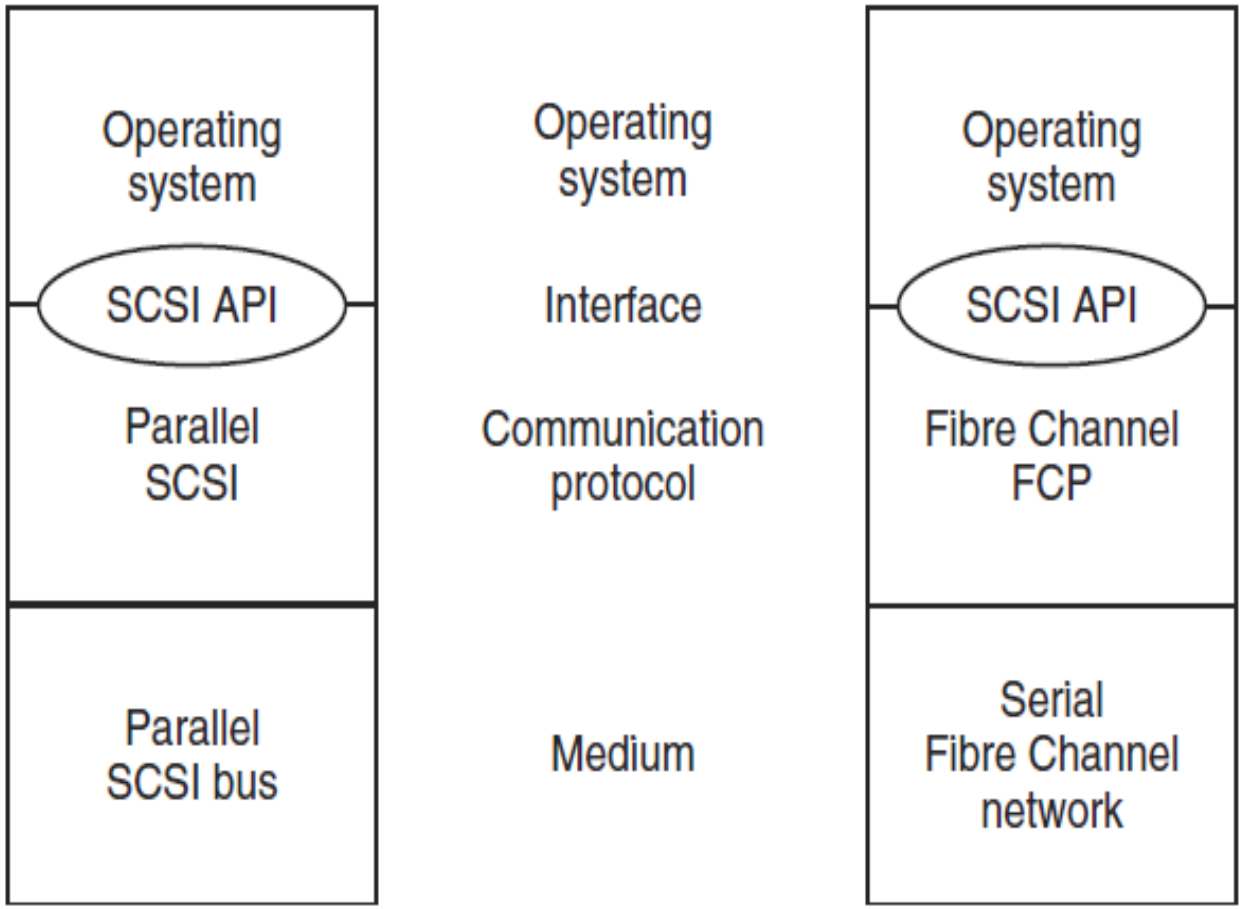
Switching Environment Services

- KCC are needed for managing infrastructure and flows in the FC
- All services implement defined servers, witch have strictly defined addresses.
- FLOG server responsible for processing all incoming fabric login request
- Fabric controller monitors for all changes in FC network
- Name server – responsible for DB all names N_Port's

Address	Description
0×FF FF FF	Broadcast addresses
0×FF FF FE	Fabric Login Server
0×FF FF FD	Fabric Controller
0×FF FF FC	Name Server
0×FF FF FB	Time Server
0×FF FF FA	Management Server
0×FF FF F9	Quality of Service Facilitator
0×FF FF F8	Alias Server
0×FF FF F7	Security Key Distribution Server
0×FF FF F6	Clock Synchronisation Server
0×FF FF F5	Multicast Server
0×FF FF F4	Reserved
0×FF FF F3	Reserved
0×FF FF F2	Reserved
0×FF FF F1	Reserved
0×FF FF F0	Reserved

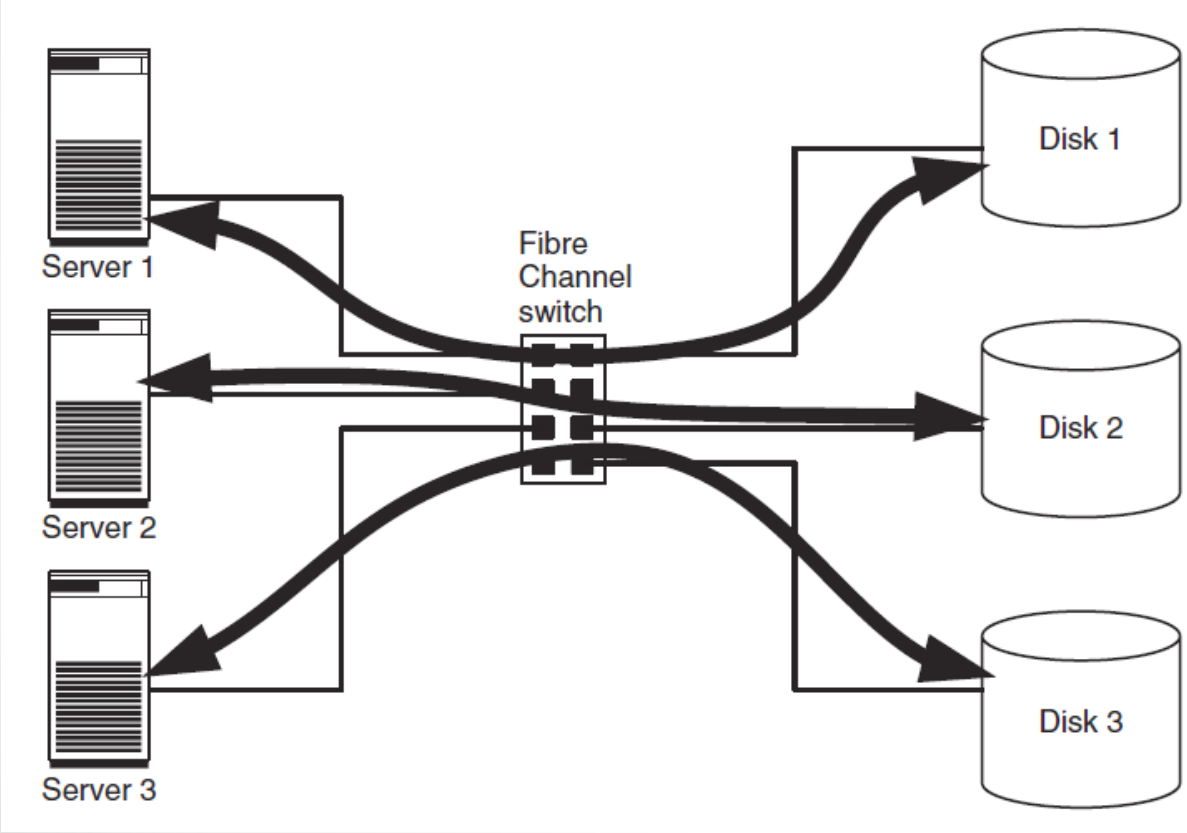
The Fibre Channel standard specifies the addresses at which the auxiliary services for the administration and configuration of the Fibre Channel network can be addressed.

FC-4: ULP

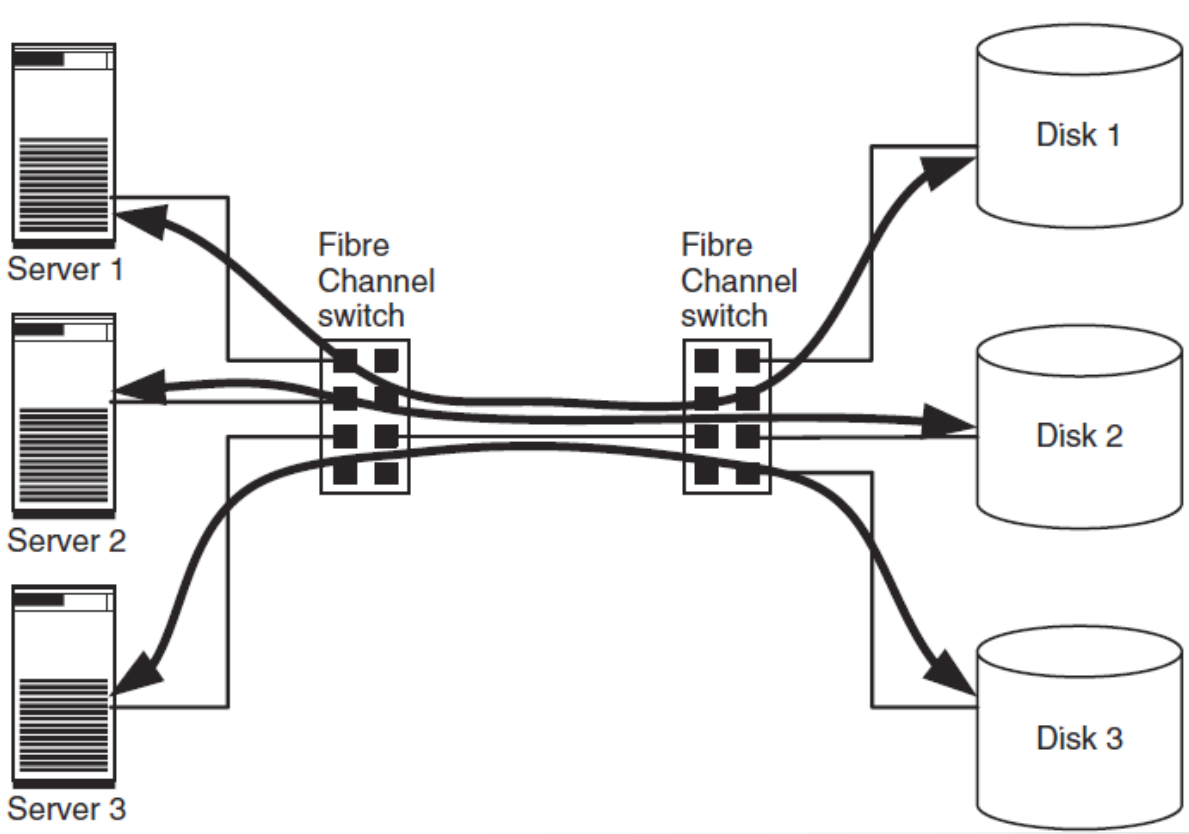


Fibre Channel FCP makes its services available to the operating system via the SCSI API. The purpose of this is to ease the transition from SCSI to Fibre Channel SAN.

Fabric topologes

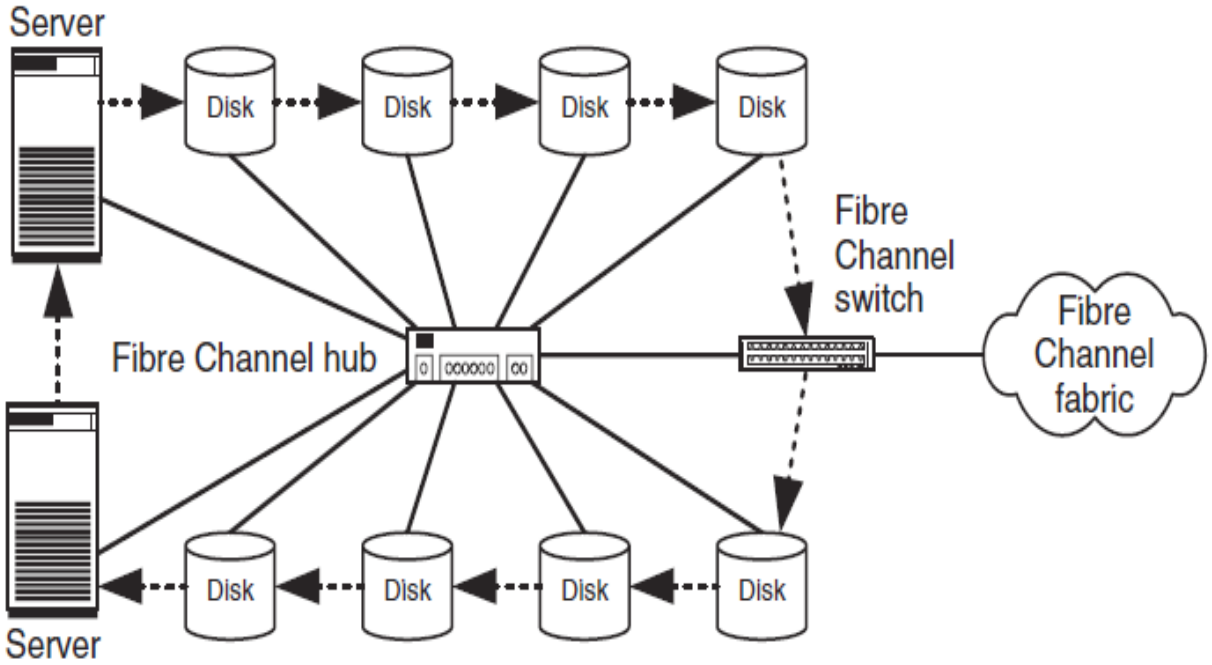


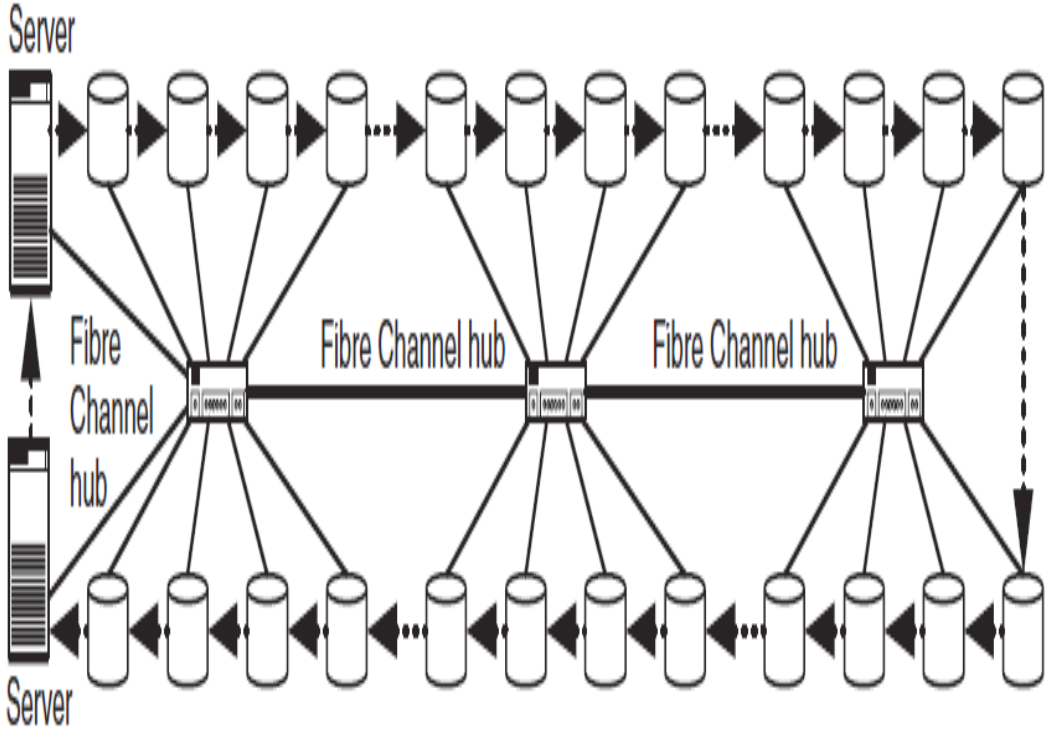
several connections at full bandwidth.



Inter-switch links (ISLs).

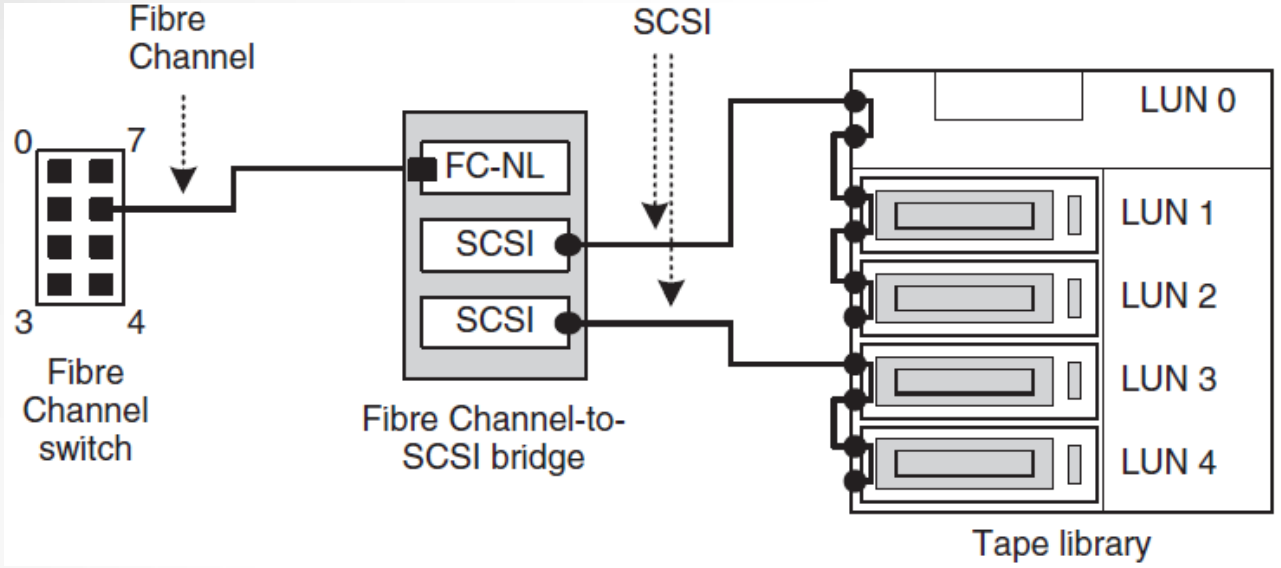
Arbitrated loop topologes



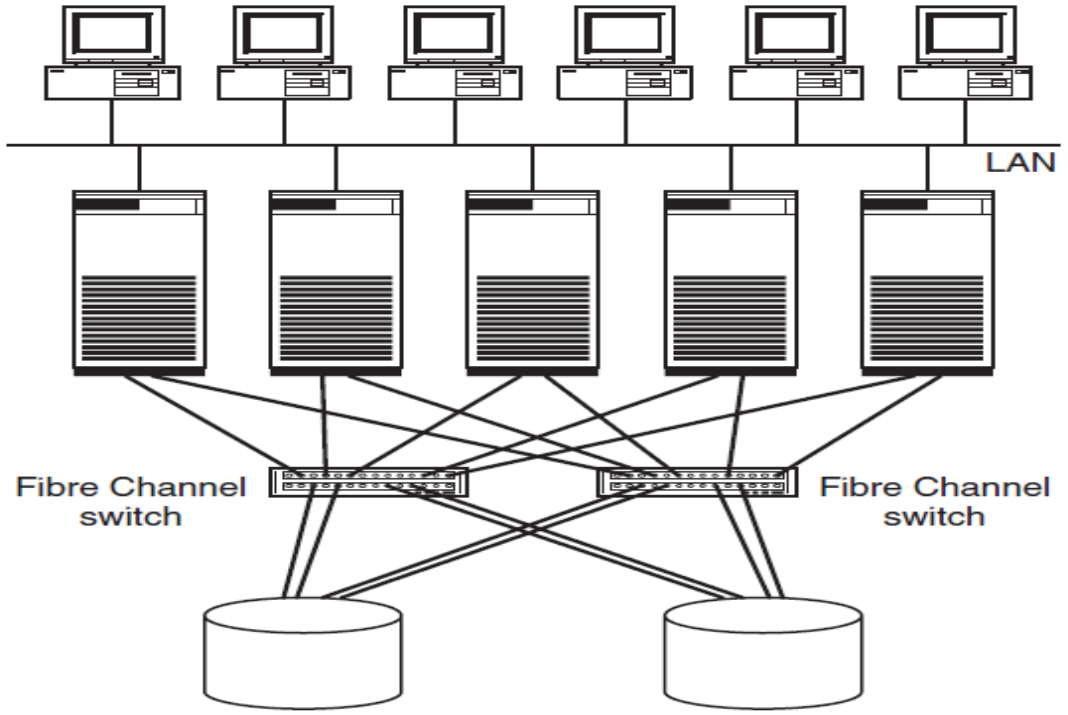


arbitrated loops are significantly cheaper than the components for a fabric

Fibre Channel SAN

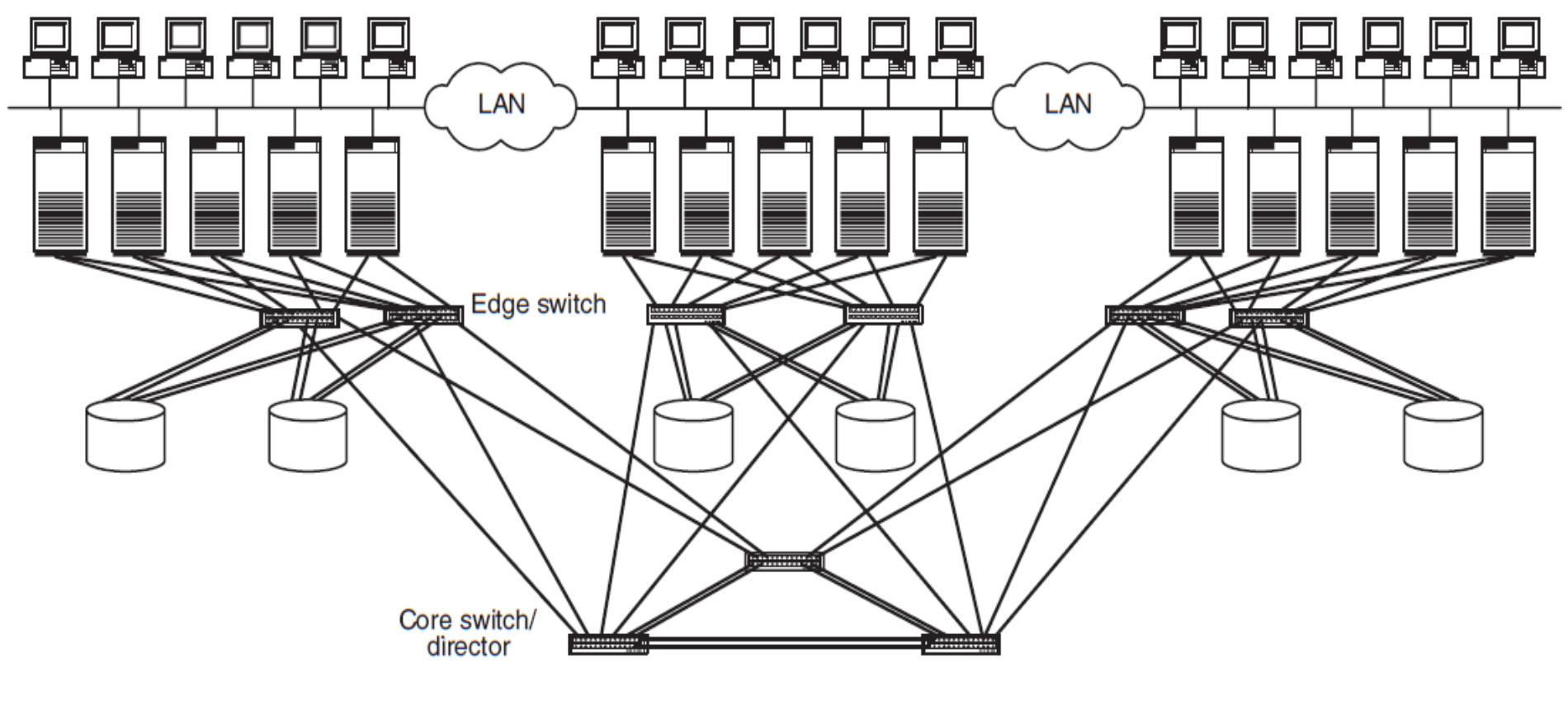


Fibre Channel-to-SCSI bridges translate between Fibre Channel FCP and SCSI.

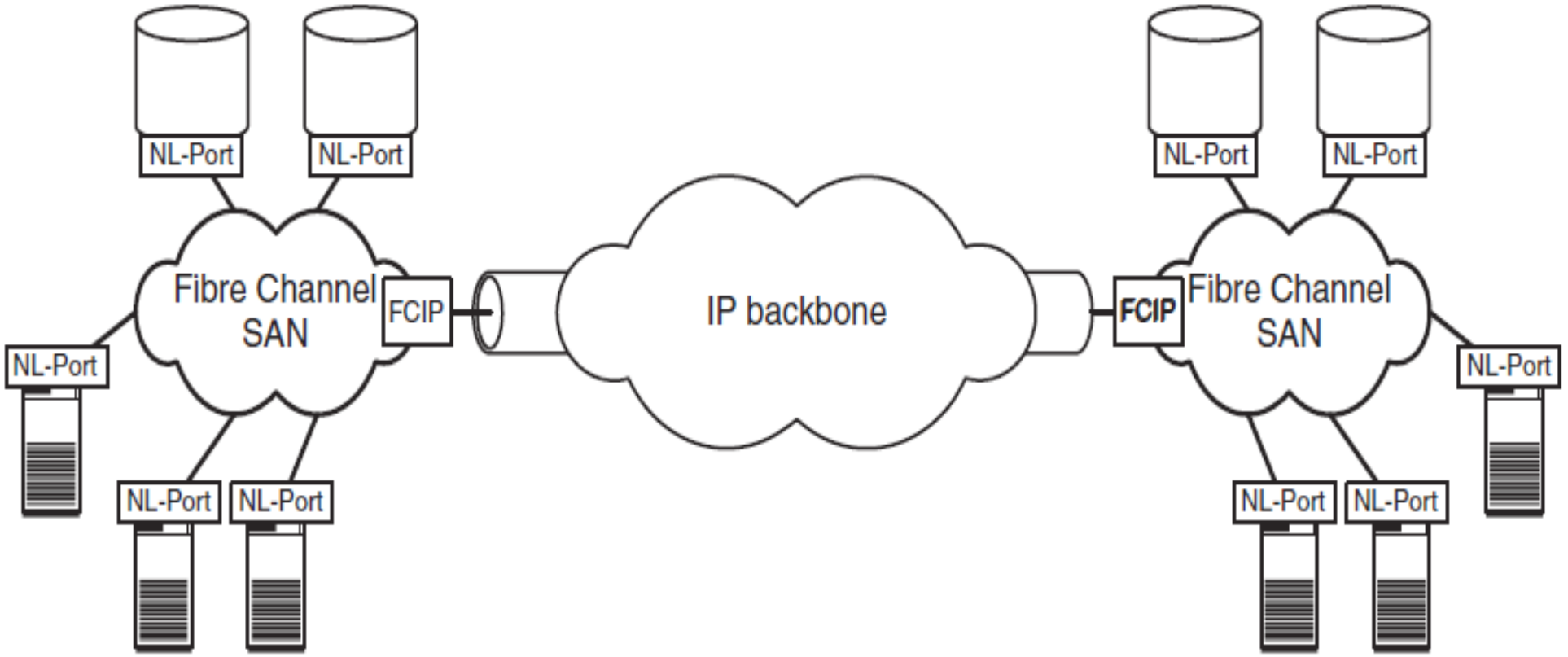


High available FC SAN configuration on dual FC fabric

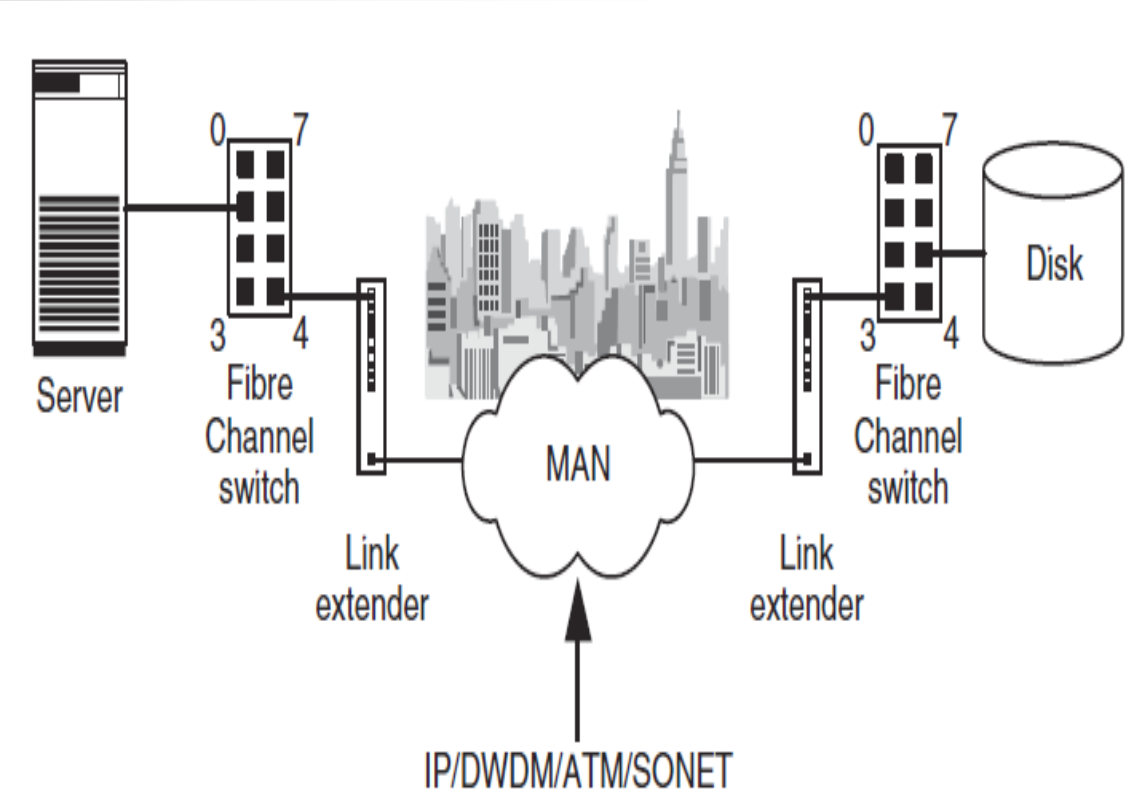
Build-in redundency

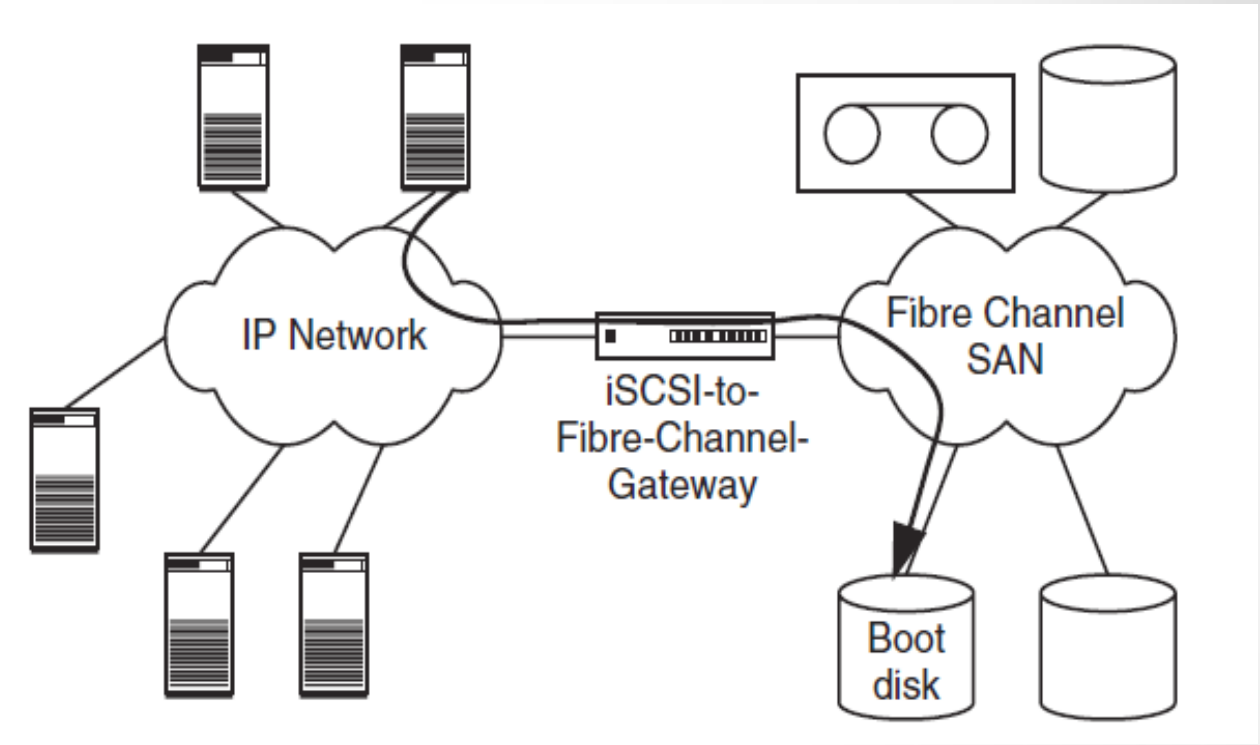


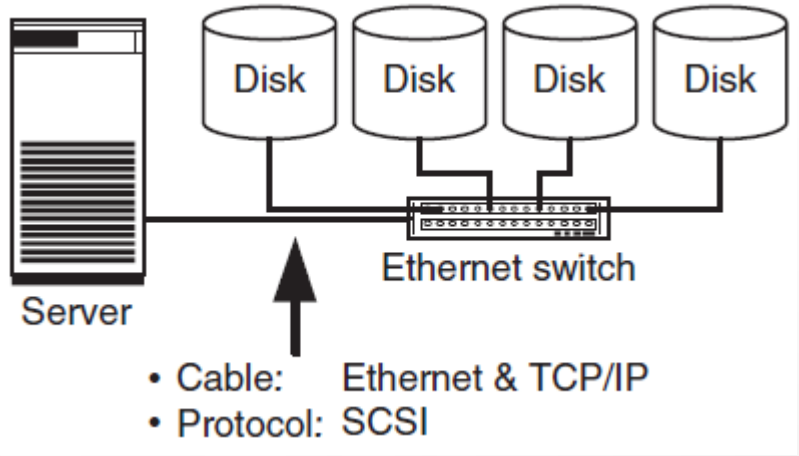
Connecting FC_SAN through TCP/IP backbone



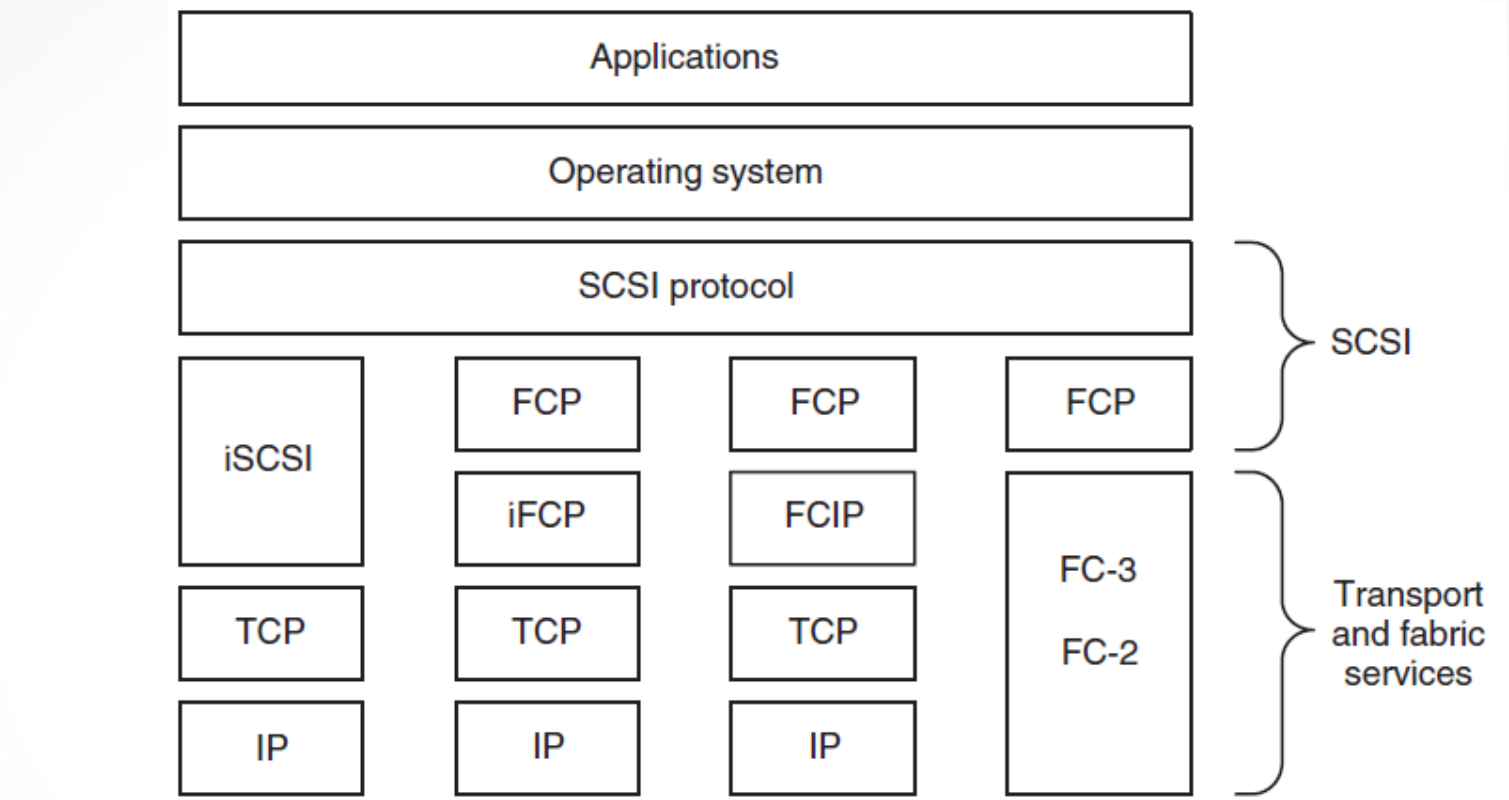
Metro SAN and IP_FC SAN

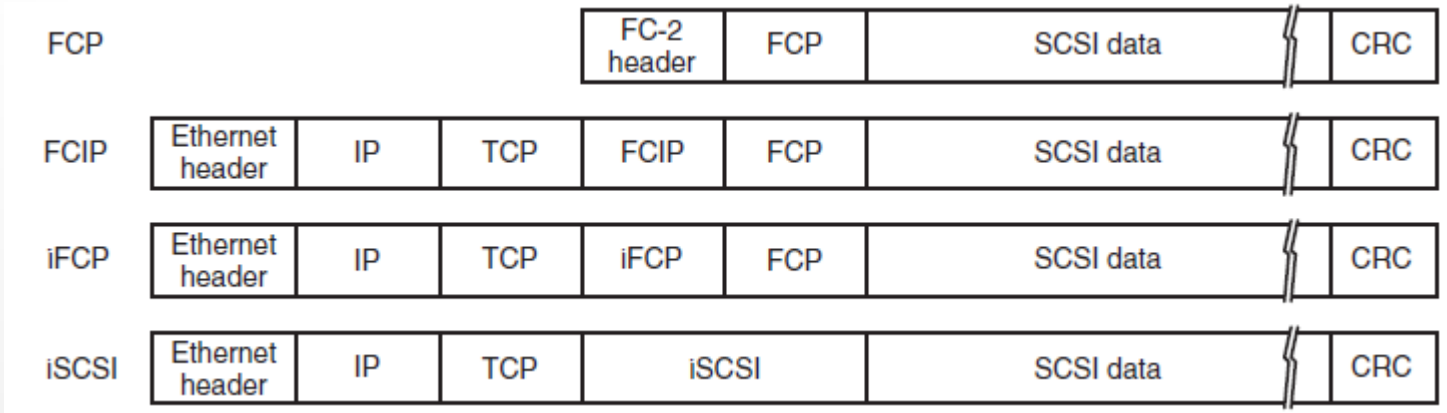




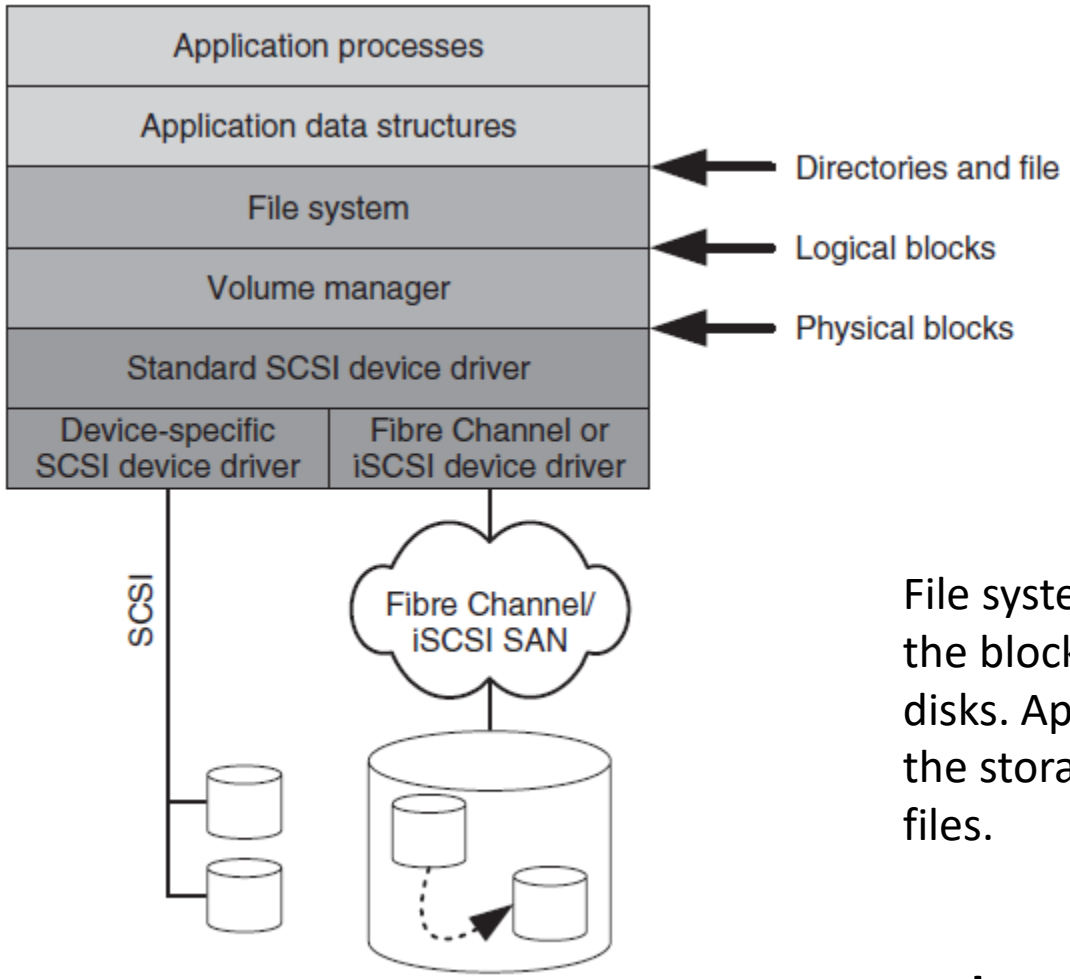


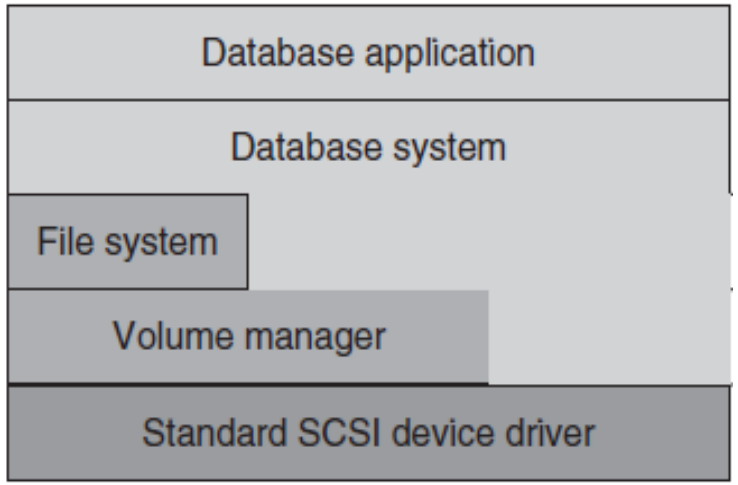
PDU encapsulation scheme

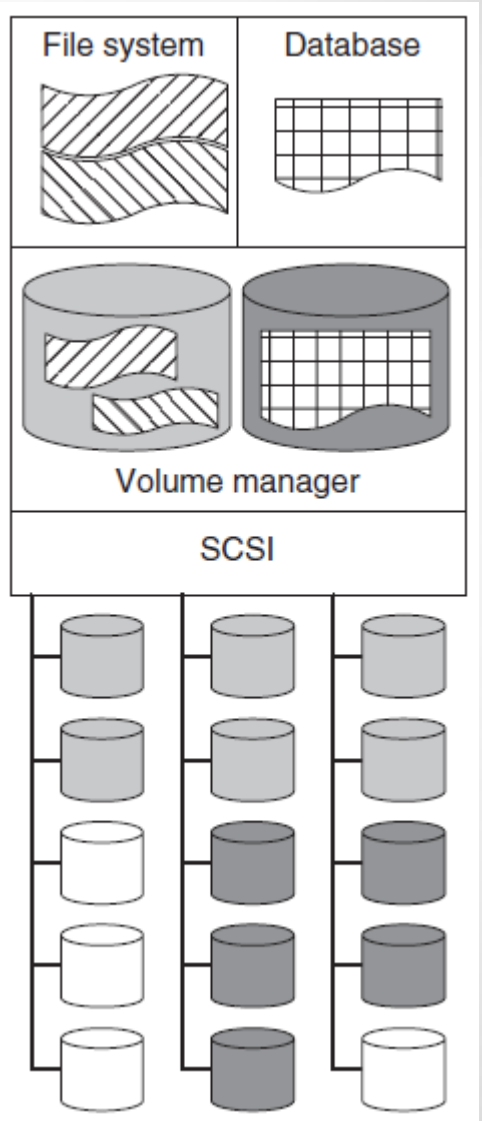




FILE SYSTEM



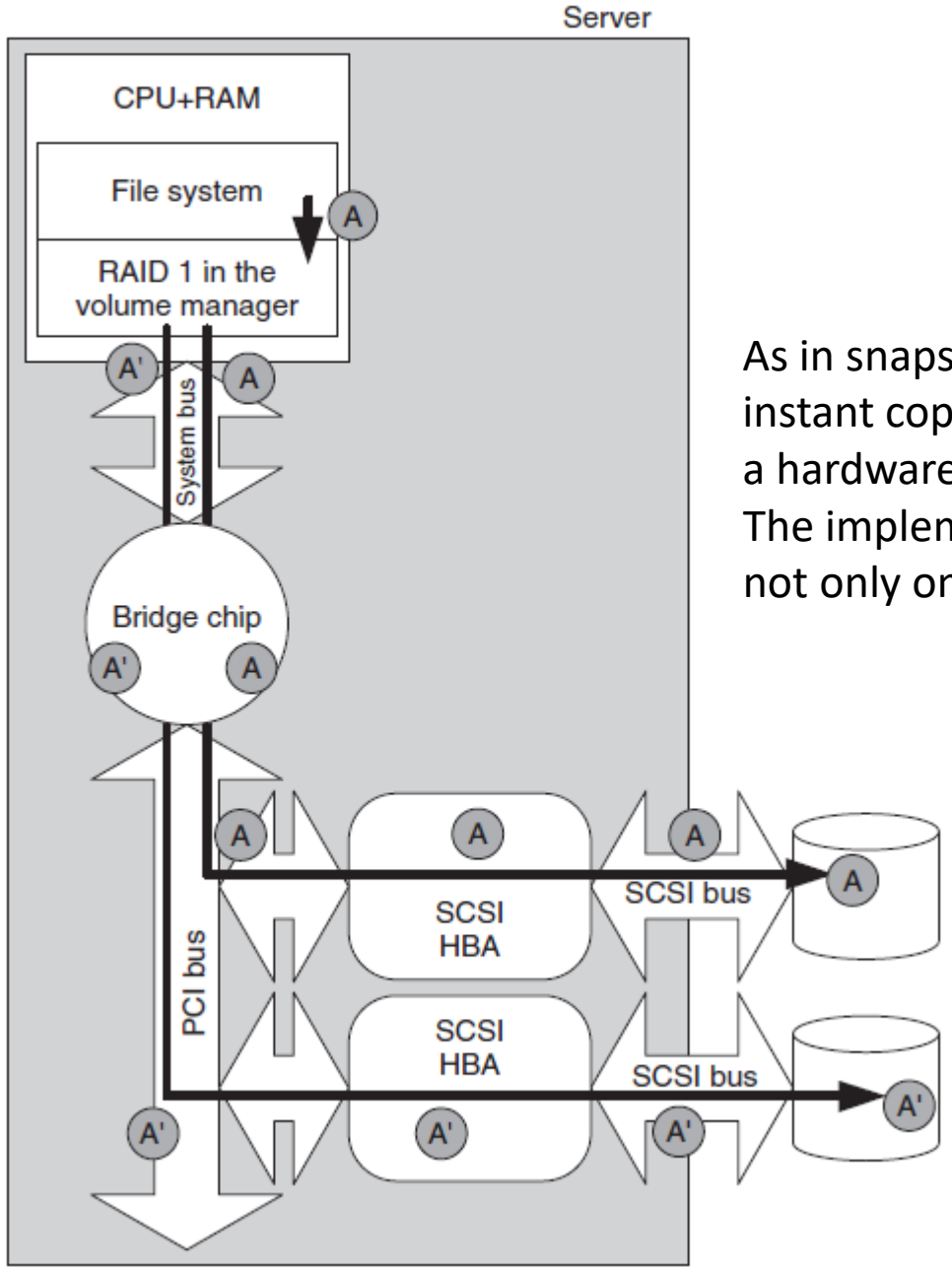




File system and volume manager manage the blocks of the block-oriented hard disks. Applications and users thus use the storage capacity of the disks via directories and files.

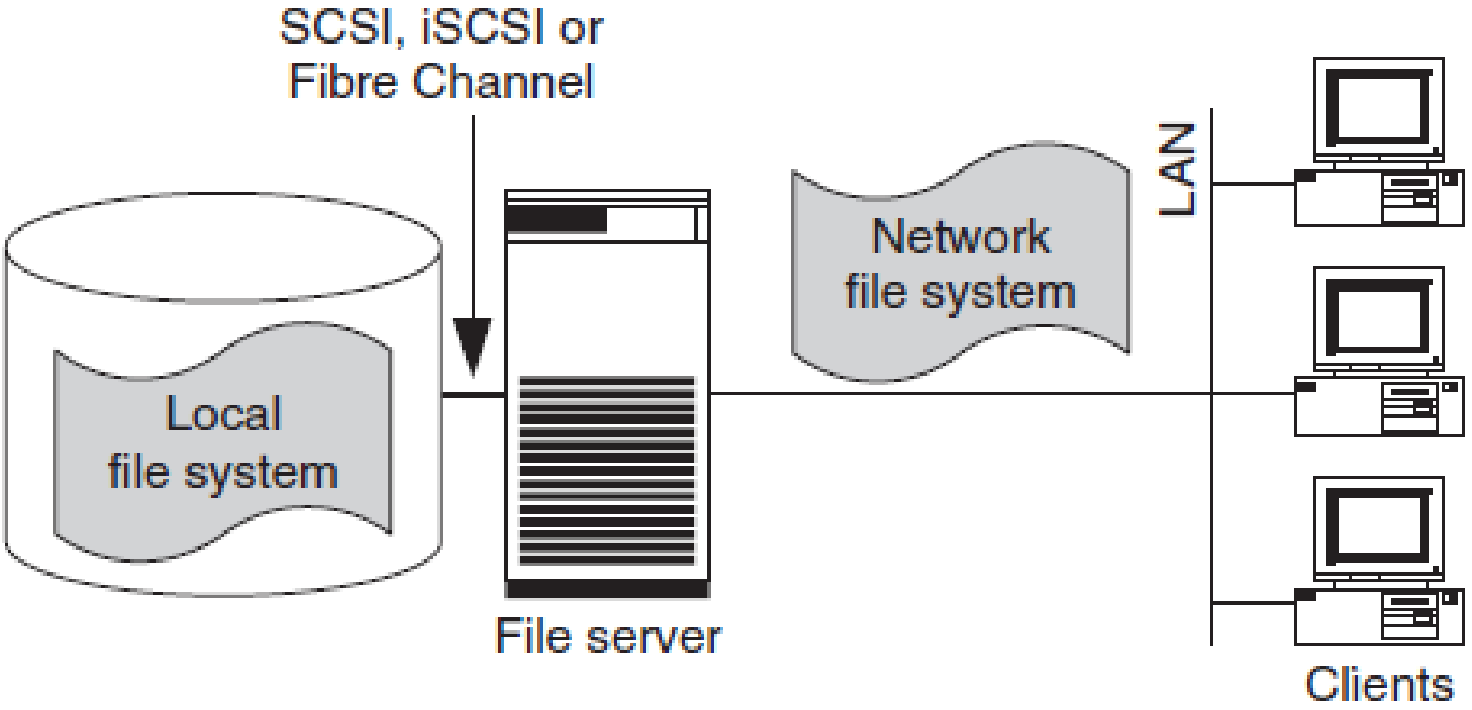
Journaling, Snapshots, Volume Manager

RAID in Volume Manager



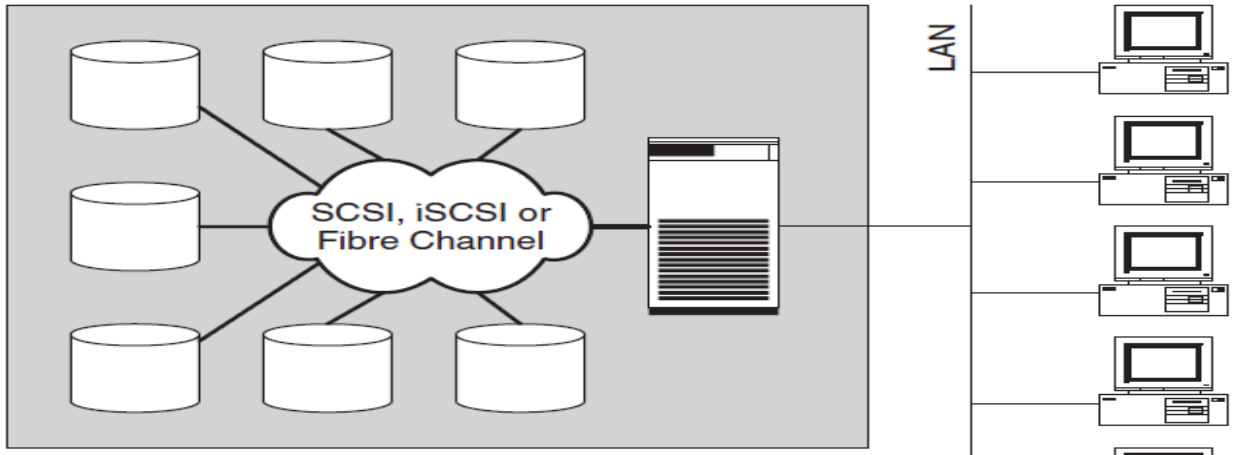
As in snapshots, here too functions such as RAID, instant copies and remote mirroring are implemented in a hardware-independent manner in the volume manager. The implementation of RAID in the volume manager loads not only on the server's CPU, but also on its buses

Network File System

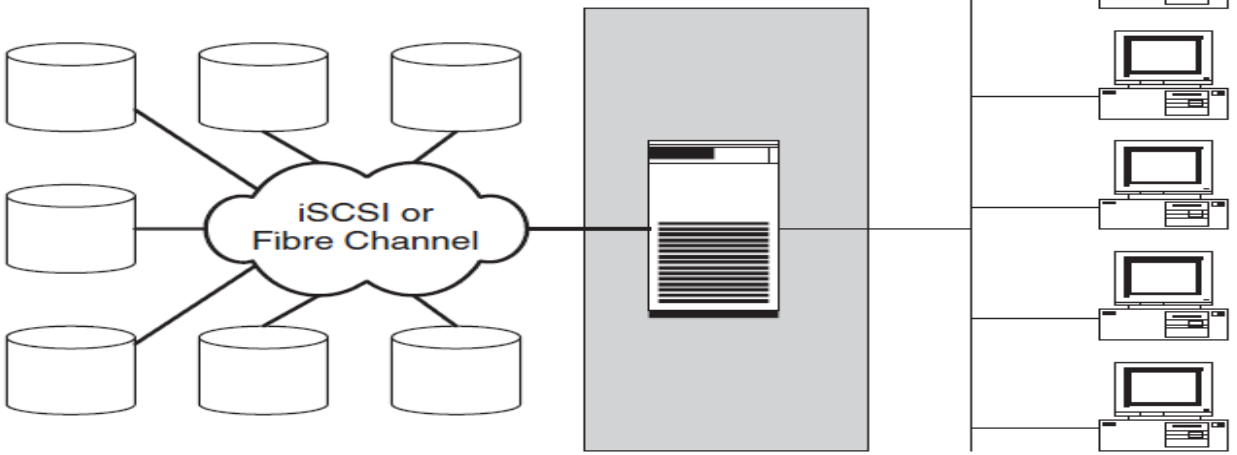


Network file systems make local files and directories available over the LAN.
Several end users can thus work on common files (for example, project data, source code).

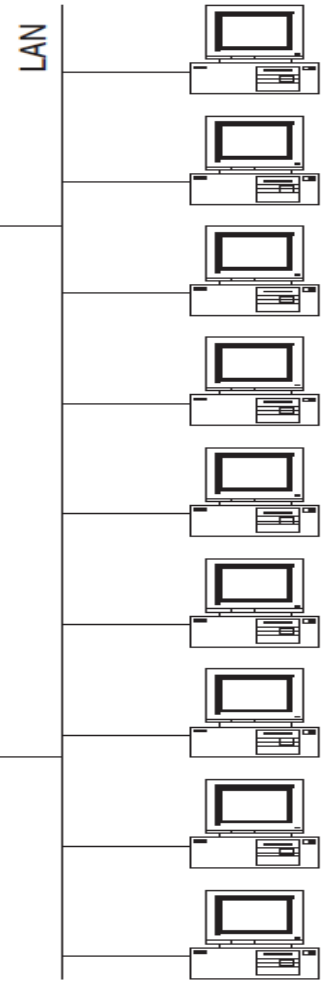
Network Attached Storage



NAS server



NAS gateway



Clients

Conclusion

- Hard disks provide their storage in the form of blocks that are addressed via cylinders, sectors and tracks.

- Currently, interoperability is provided between the interfaces of SPD SCSI, FC, Ethernet, Infiniband, which allows using the traditional TCP \ IP protocol stack to build a SAN

- File systems manage the blocks of the hard disks and make their storage capacity available to users in the form of directories and files.
- Network file systems and shared disk file systems make it possible to access to the common data set via LAN from various computers with additional functions: Journaling, Snapshots; Volume manager.

- The performance of network file systems is limited by two factors:
  (1) all data accesses to network file systems have to pass through a single file server;
  (2) current network file systems such as NFS and CIFS and the underlying network protocols are not suitable for a high throughput.