Storage Area Network

чл.-корр. РАН Смелянский Р.Л. Доп. главы компьютерных сетей

Сетевые Хранилища Данных

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 © 2009 John Wiley & Sons Ltd. ISBN: 978-0-470-74143-6

Содержание

- Архитектура информационной инфраструктуры
- Дисковые подсистемы и их организация
- JBOD
- RAID
- Интеллектуальные ДПС
- От CPU до ДПС
- SCSI
- Fibre Channel
- Заключение

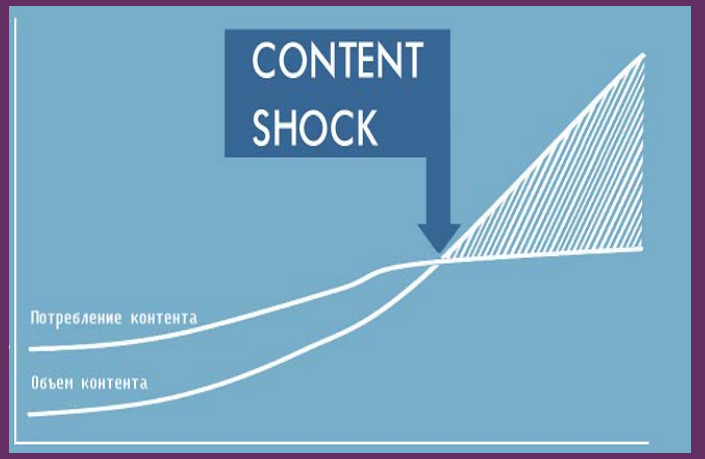
Основные тренды роста трафика в сетях

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 will dominate traffic between the data center (DC)

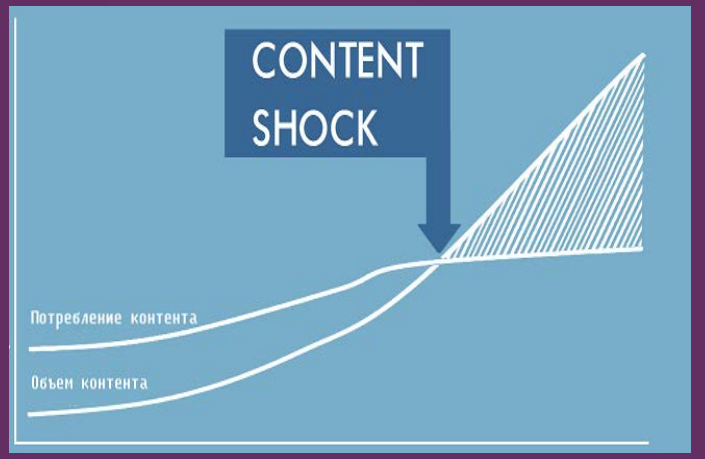


Specifics of the growth of mobile traffic:

From 2015 to 2020, the volume of mobile traffic will increase by 8 times and reach in 2020 the figure of 30.6 EB / month (Exabyte = 10 ^ 18). Mobile traffic during this period will grow three times faster than traffic in fixed networks.

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 will grow 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 will increase tenfold.
- 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**

Тренды

1. Объем данных, созданных в течение следующих трех лет, будет больше, чем объем данных, созданных за последние 30 лет. А в течение следующих пяти лет мир будет создавать в три раза больше данных, чем в предыдущие пять.

2. К 2024 году развлекательные данные будут составлять 40% от мирового объема данных. Растет число датчиков интернета вещей, собирающих информацию от различных устройств.
Кол
вск

3. Д на 4%
в те

4. К
гос
эфф
из г

Более 175 зеттабайт данных будет создано, скопировано и использовано в мире к 2025.

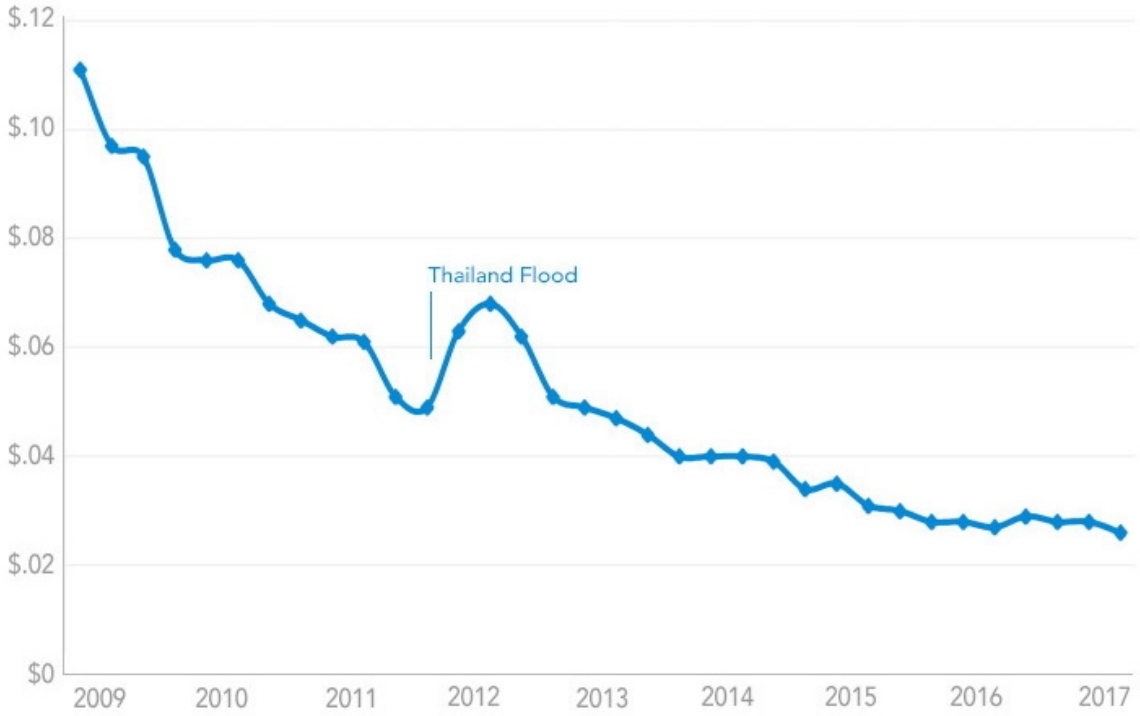
https://www.idc.com/getdoc.jsp?containerId=IDC_P38353

Объём информации в мире **возрастает ежегодно на 30 %**. В среднем на человека в год в мире производится $2{,}5 \cdot 10^8$ байт.

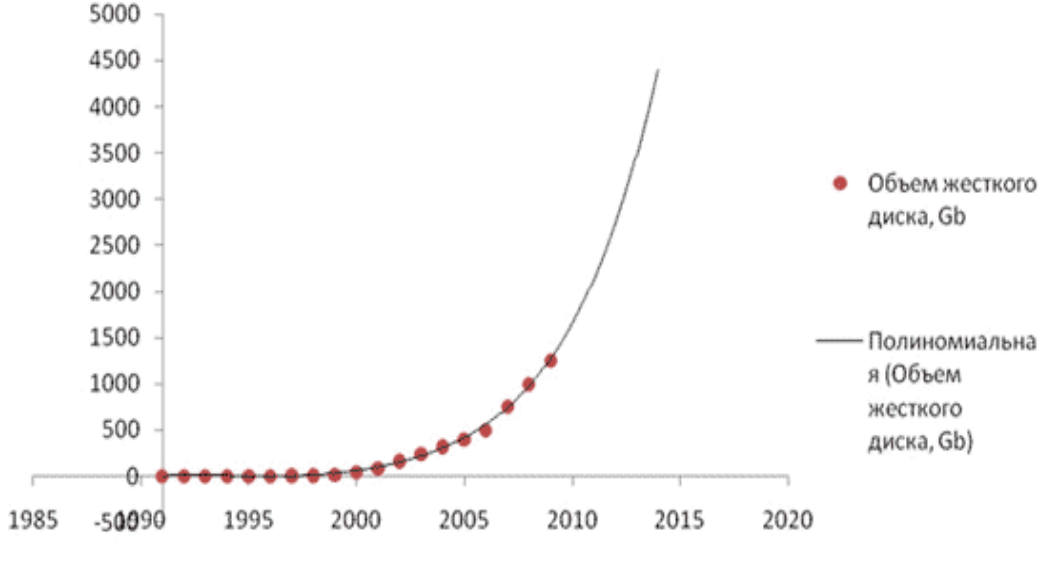
Соо
реплицированным данным (скопированным и использованным) составляет примерно 1:9.
При этом наблюдается дальнейшее смещение в сторону реплицированных данных:
по прогнозам, к 2024 году это соотношение будет 1:10.

The main trends in Data storages

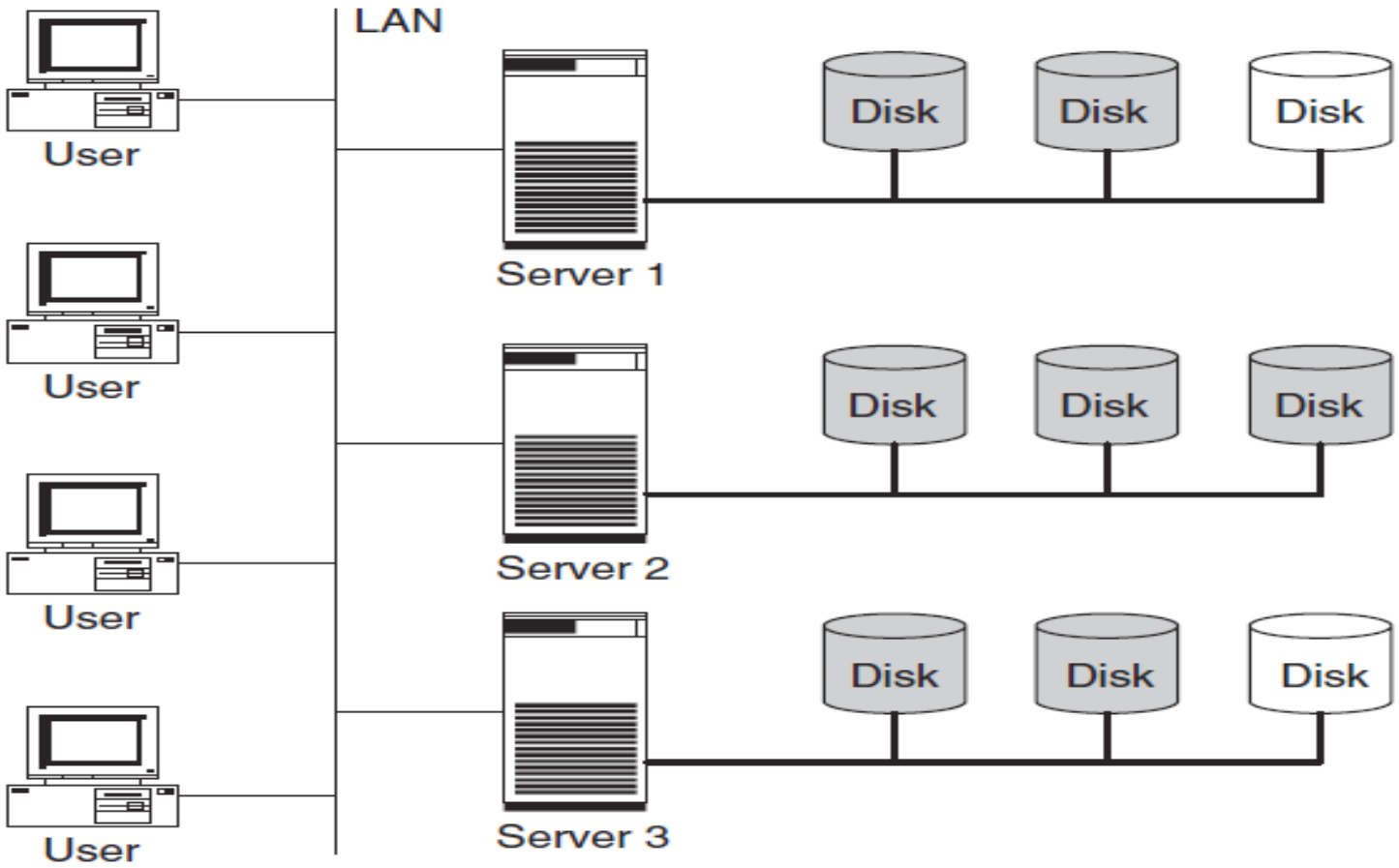
Drop of average cost per Drive size



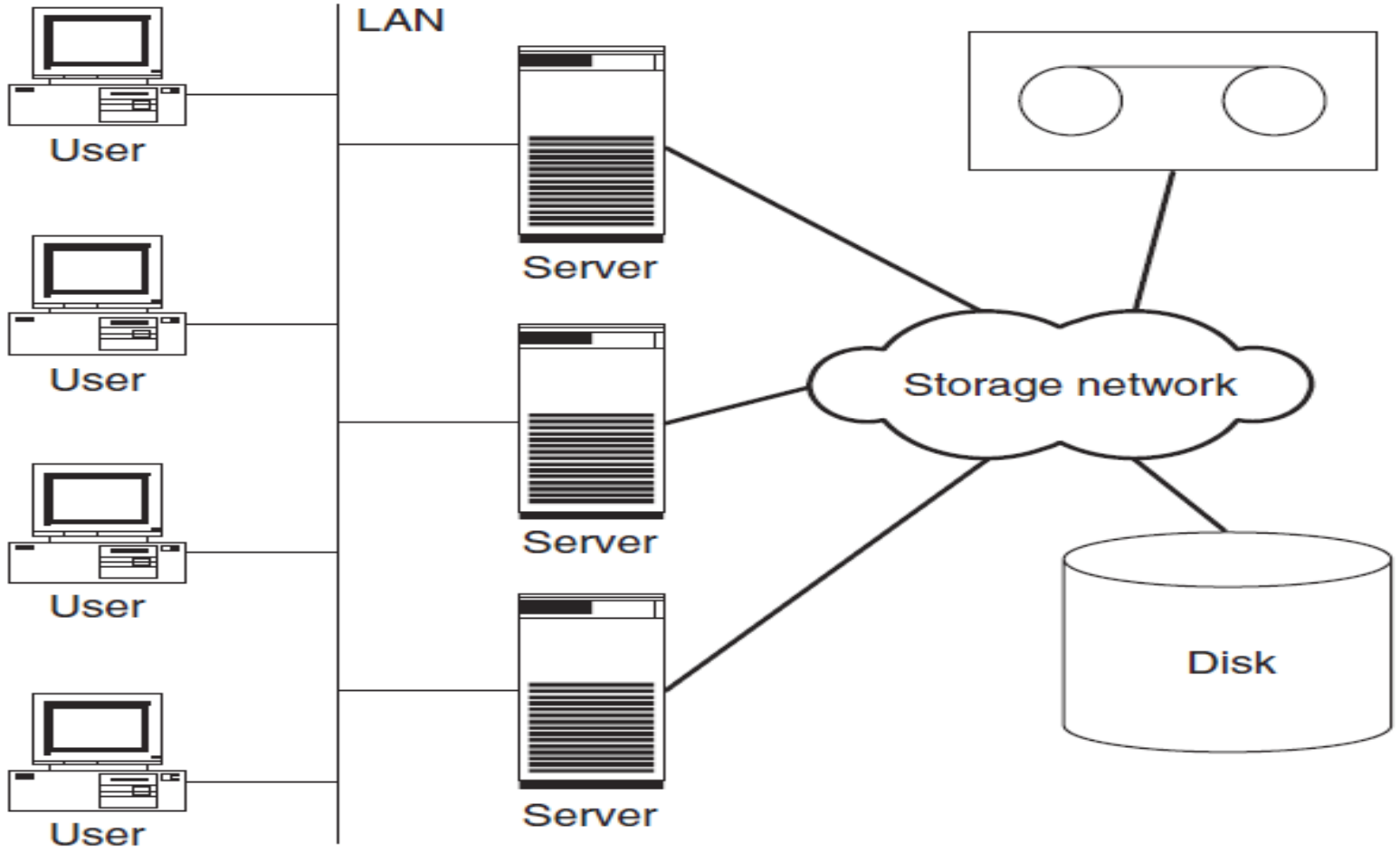
Volume HD (Gb)



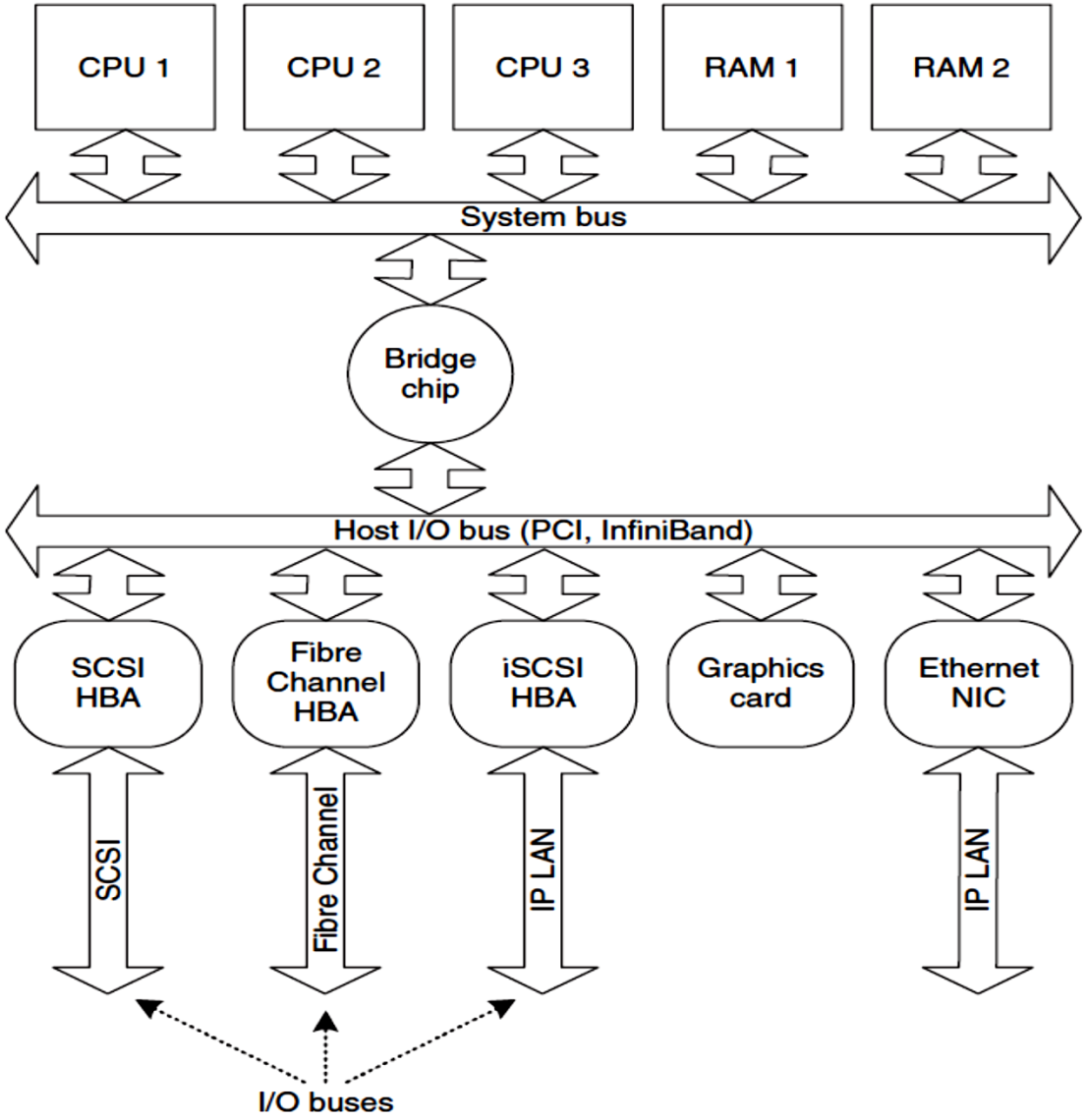
Server Centric Architecture



Storage Centric Architecture



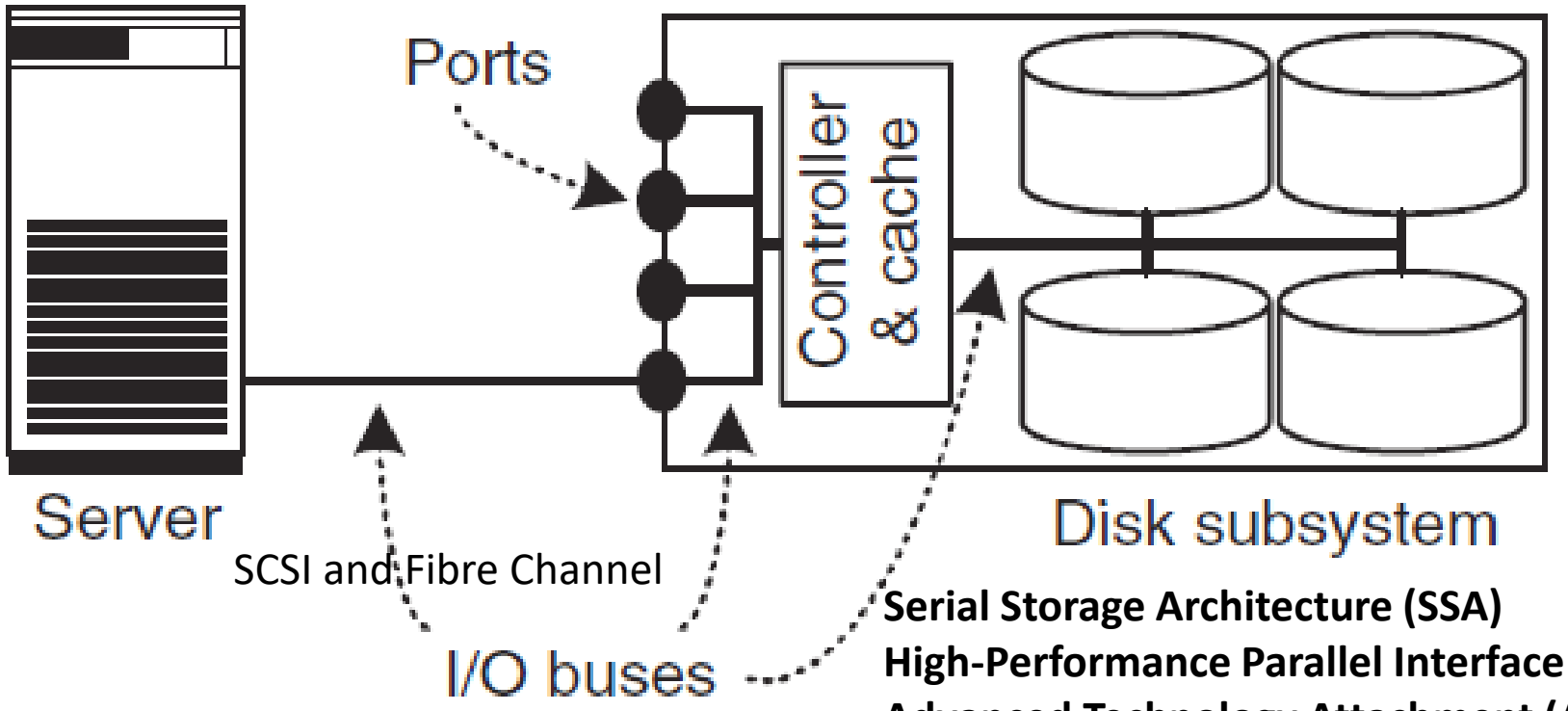
THE PHYSICAL I/O PATH FROM THE SERVER CPU TO THE STORAGE SYSTEM



PCI – Peripheral Component Interconnection
PCI gen.4 – 512 GBps

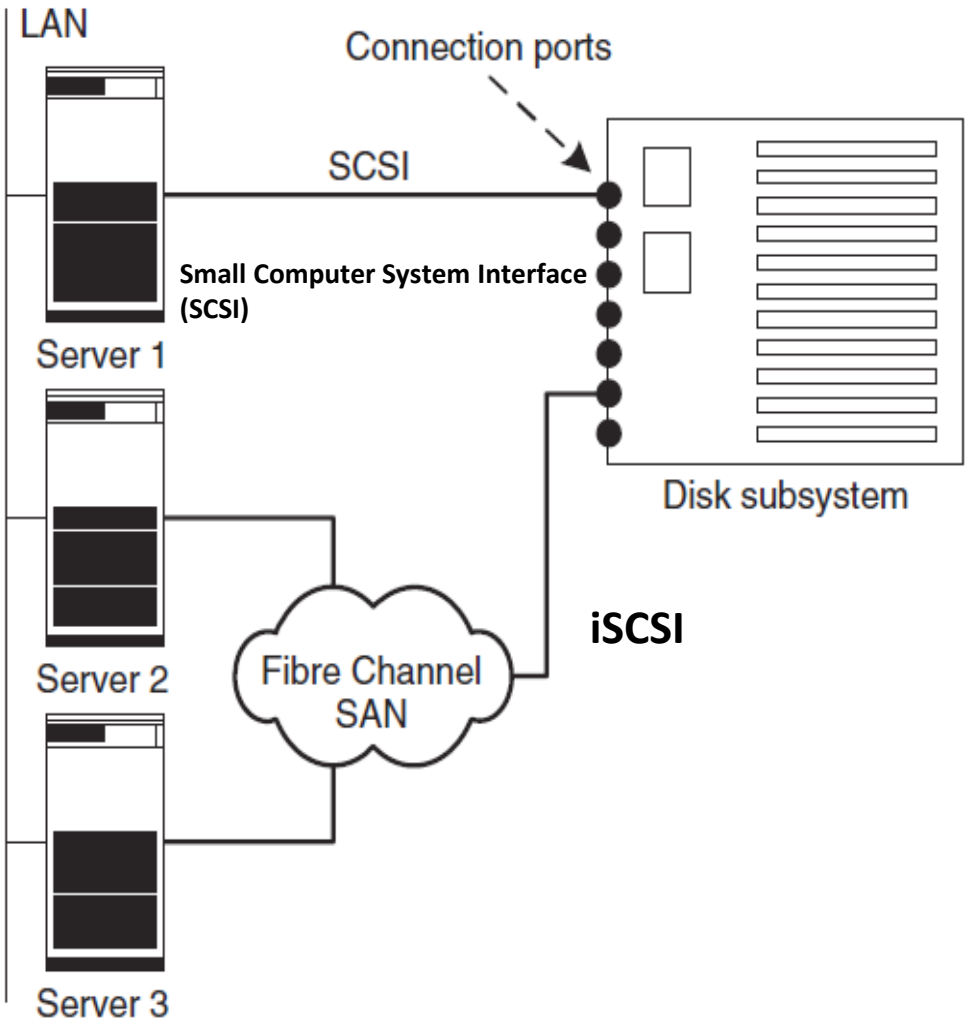
NIC – ASIC drivers

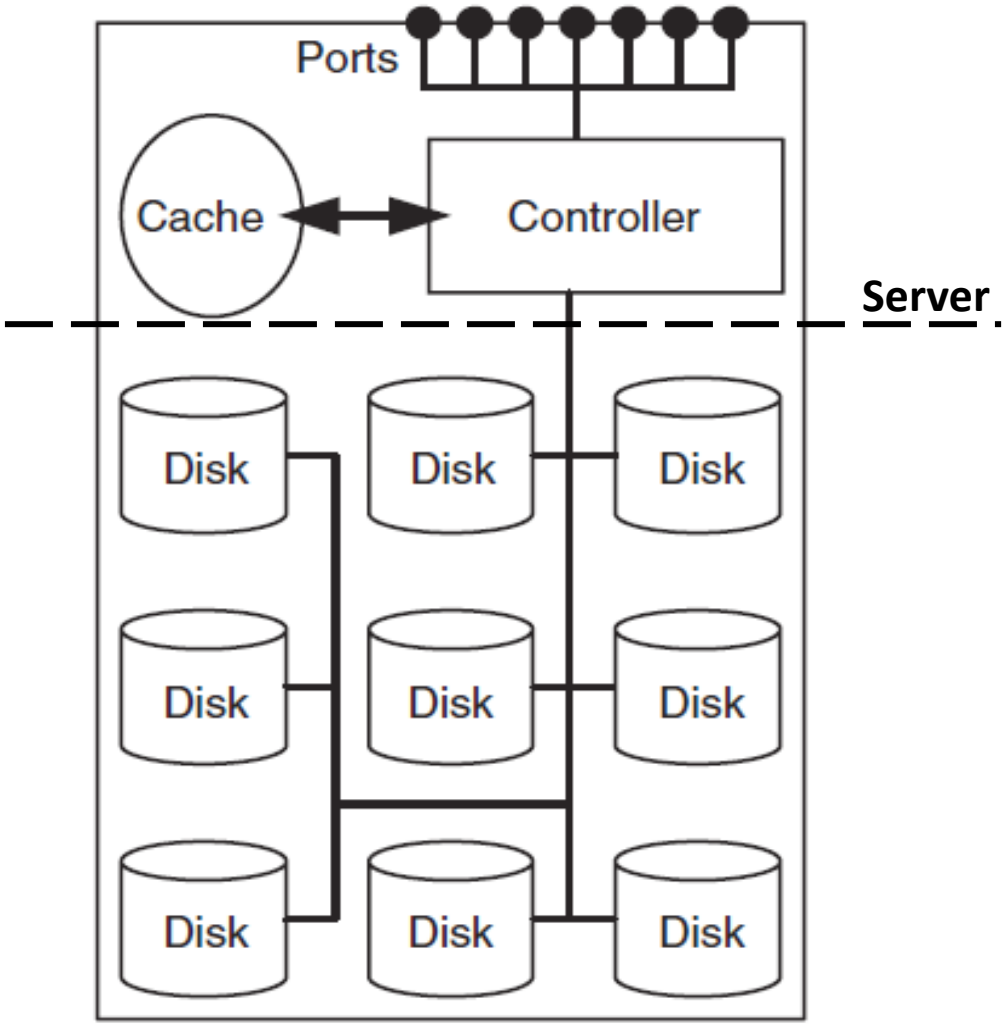
THE PHYSICAL I/O PATH FROM THE SERVER CPU TO THE DISK SUBSYSTEM



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).

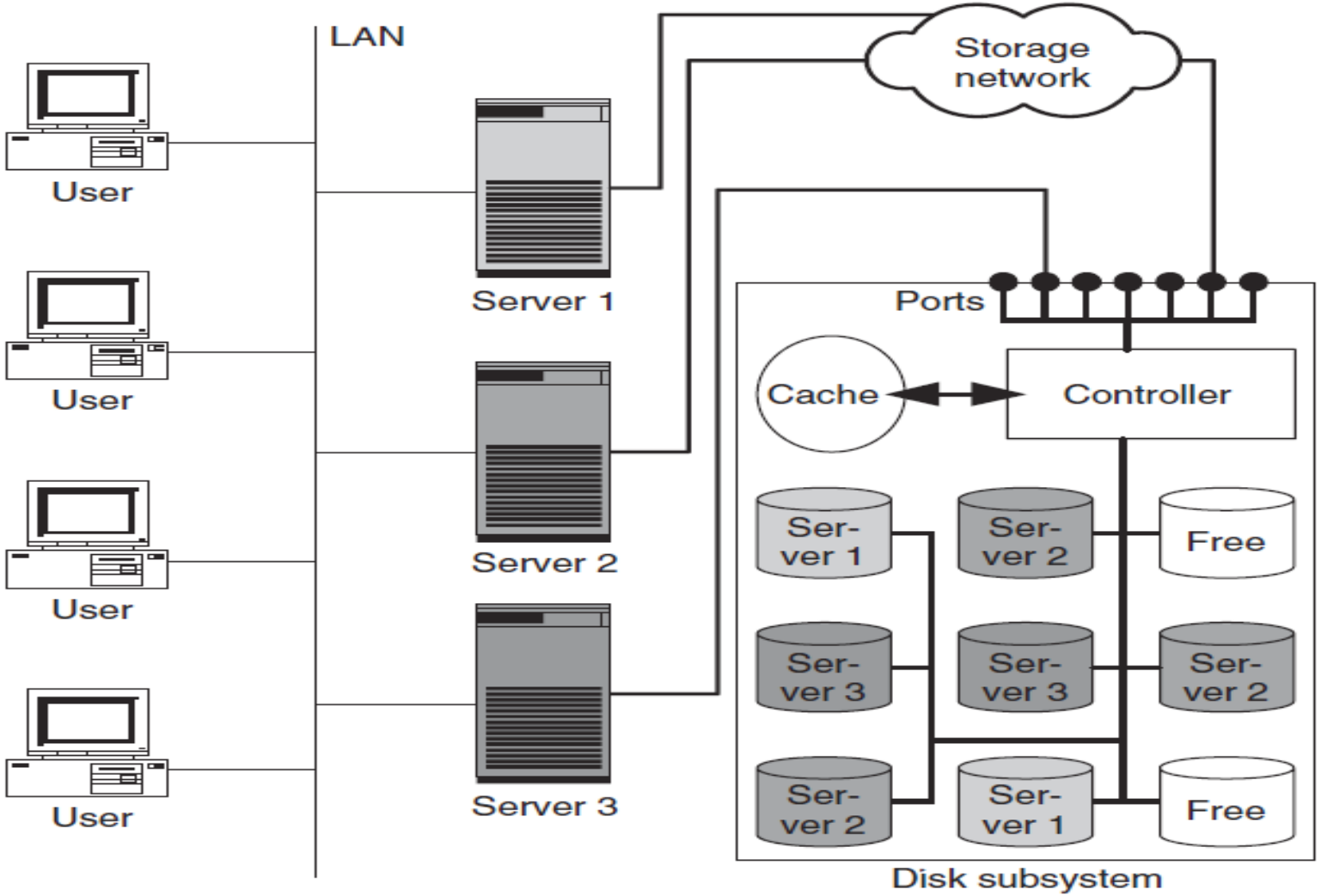
Disk Subsystem Architecture (JBOD)





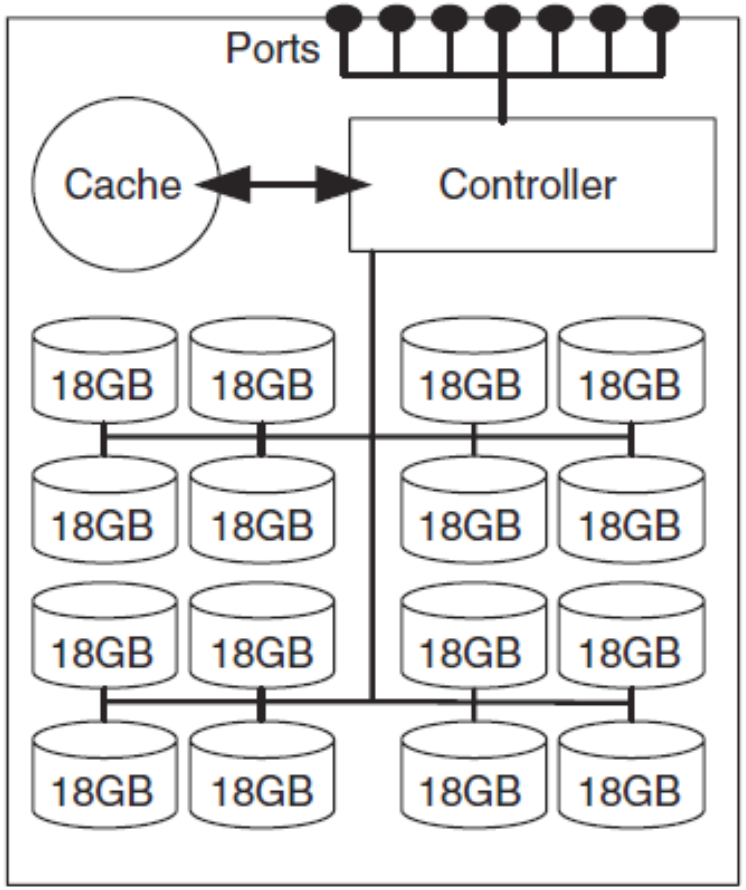
(1) no controller; (2) RAID controller; (3) intelligent controller with services like e.g. instant copy and remote mirroring .

Disk Storage System - usage example

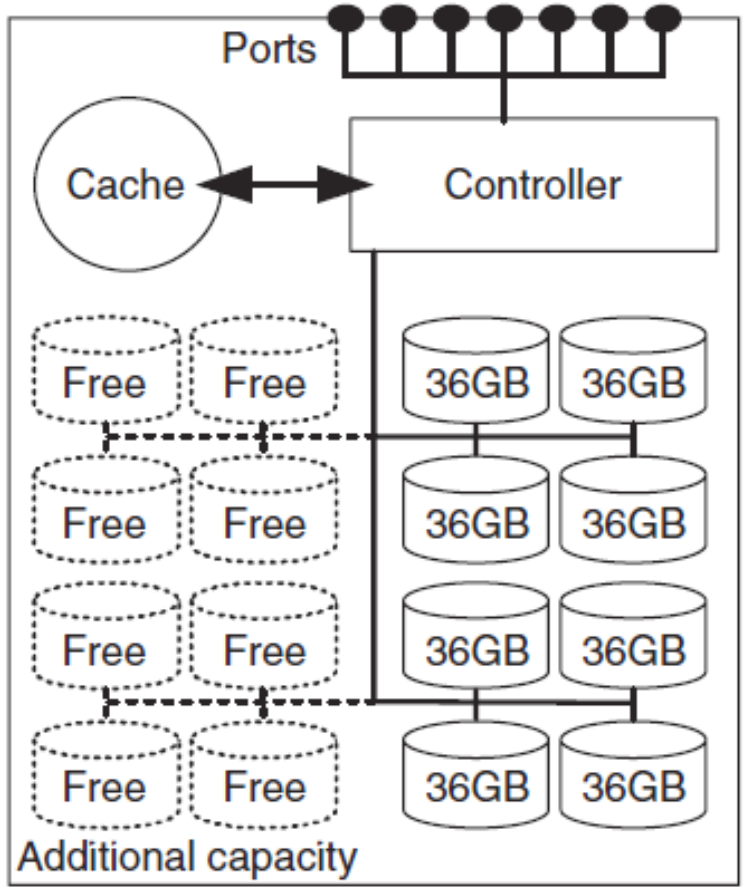


Max amount of disk capacity is limited
Max number of HDD or SDD in the same JOBD is limited

Disk Subsystem: internal organization and disk capacity



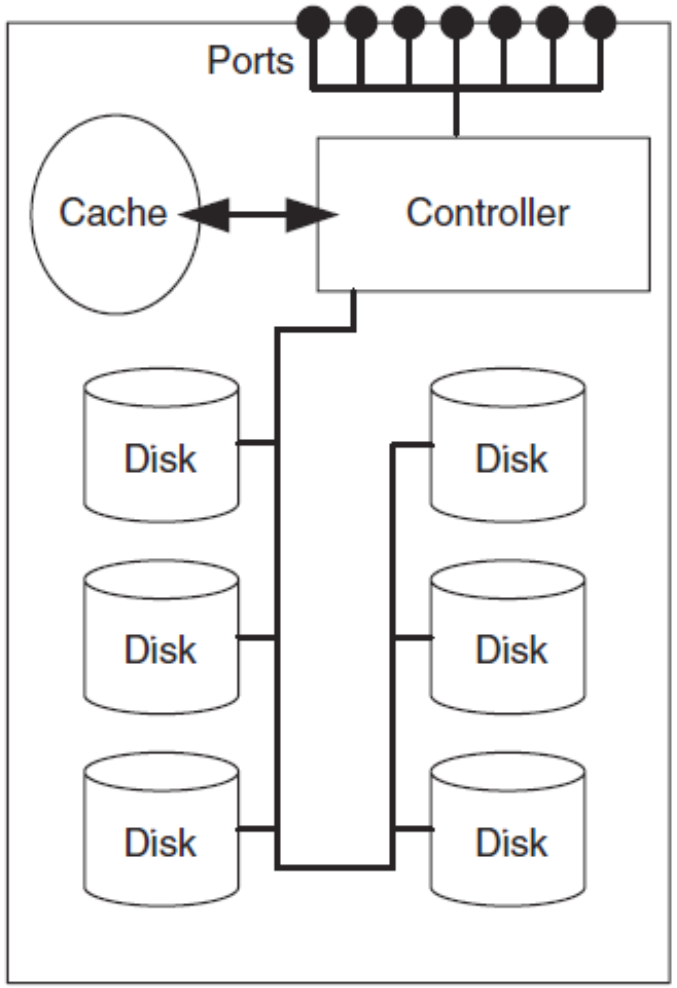
Маленькие диски



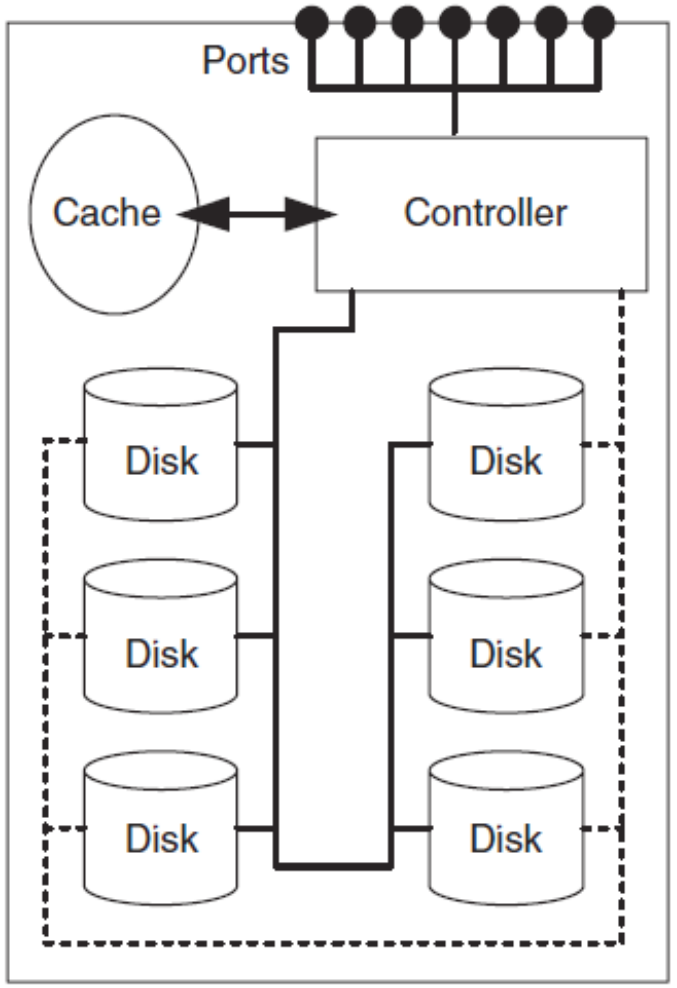
Большие диски

What files are deals with?

Disk Subsystem internal organization

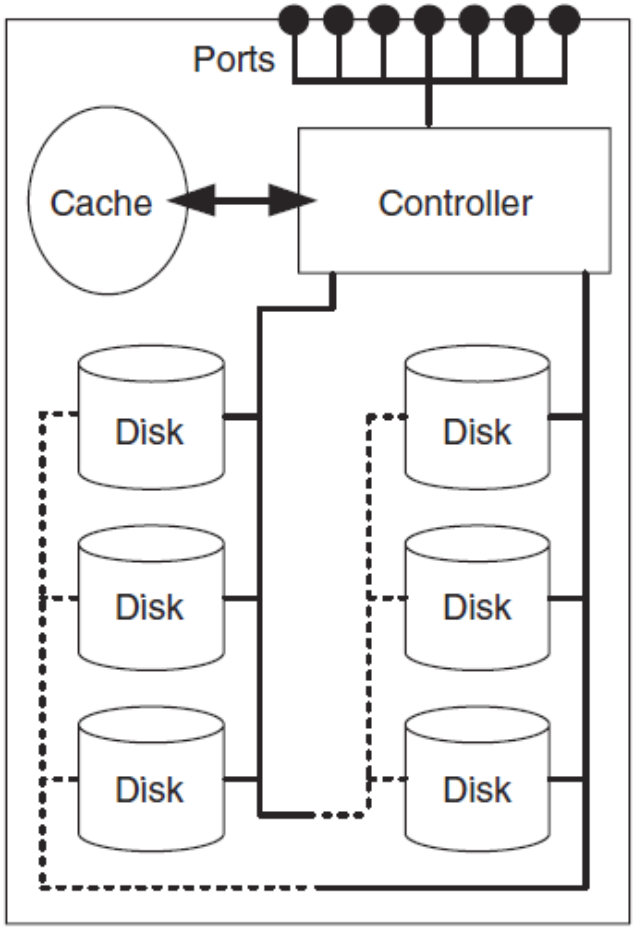


Active Disks connectivity

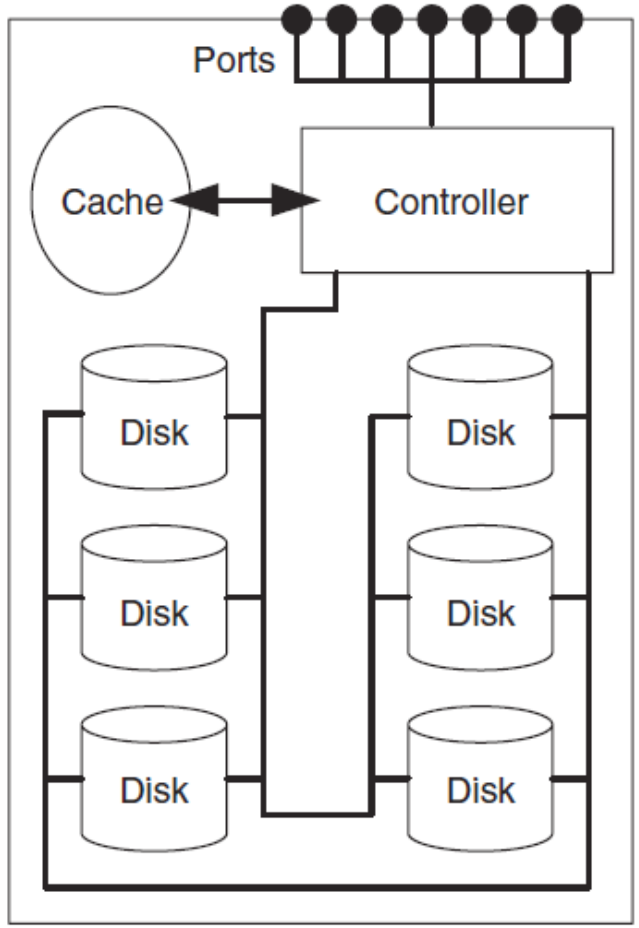


Active/Passive **Disks connectivity**

Active duplication



Active/Active with separation
No load sharing

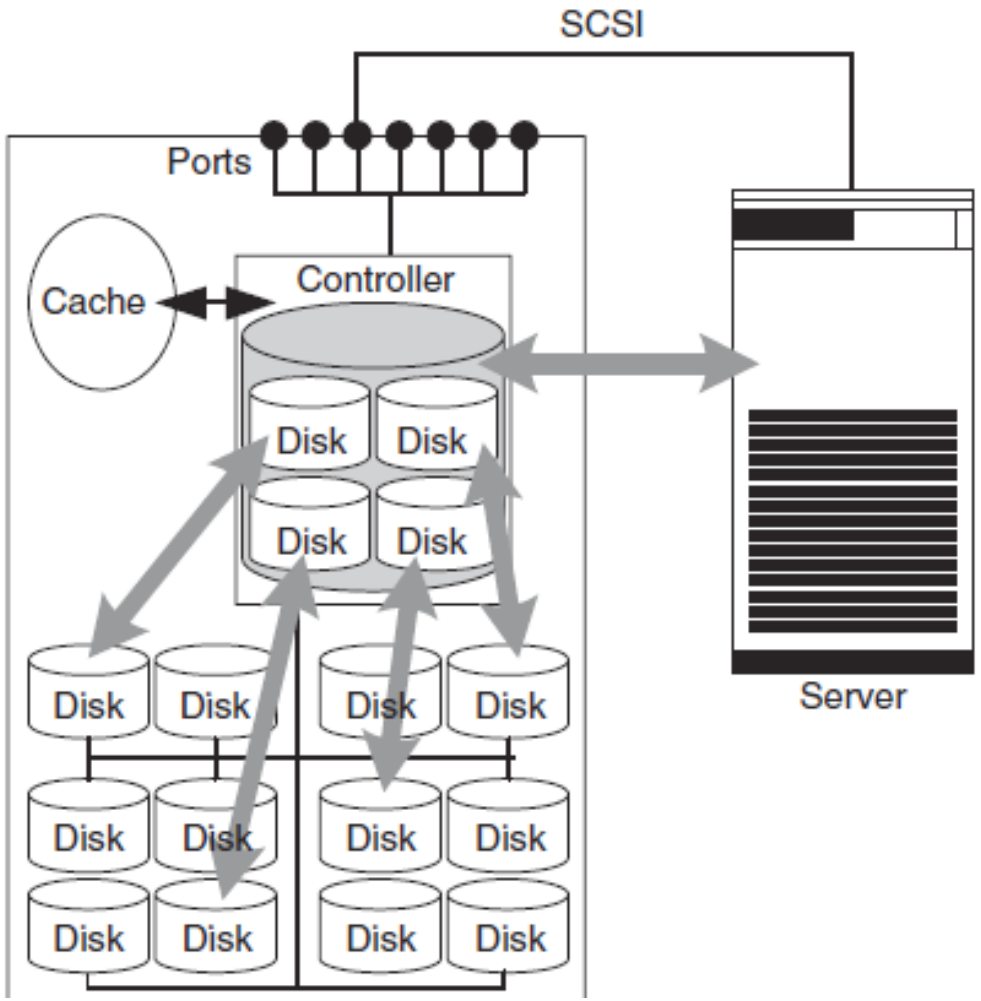


Active/Active without separation
Load sharing

JBOD: JUST A BUNCH OF DISKS

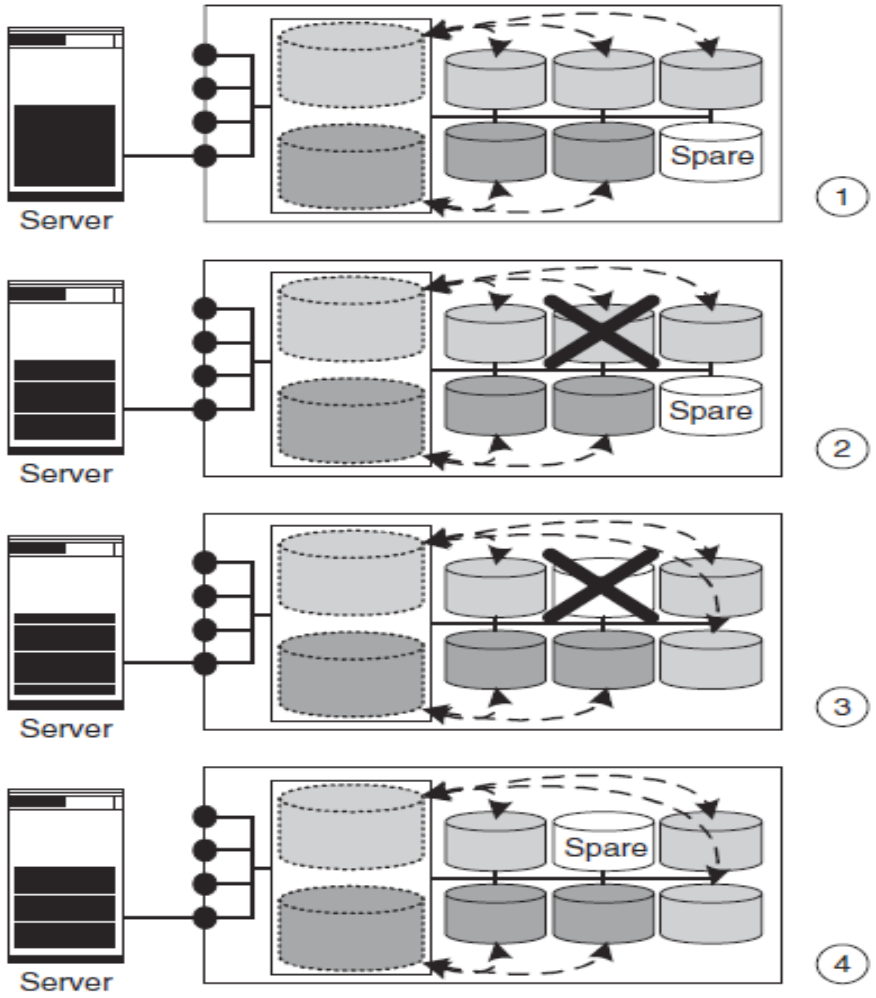
- No internal controller
- The connections for I/O channels and power supply are taken outwards
- Small number of HDD
- Outside server can see JOBD as several independent disks
- Because of one point of connection JOBD it is the bottleneck of Disk Subsystem

Storage Virtualization with RAID – Redundant Array of Independent Disks

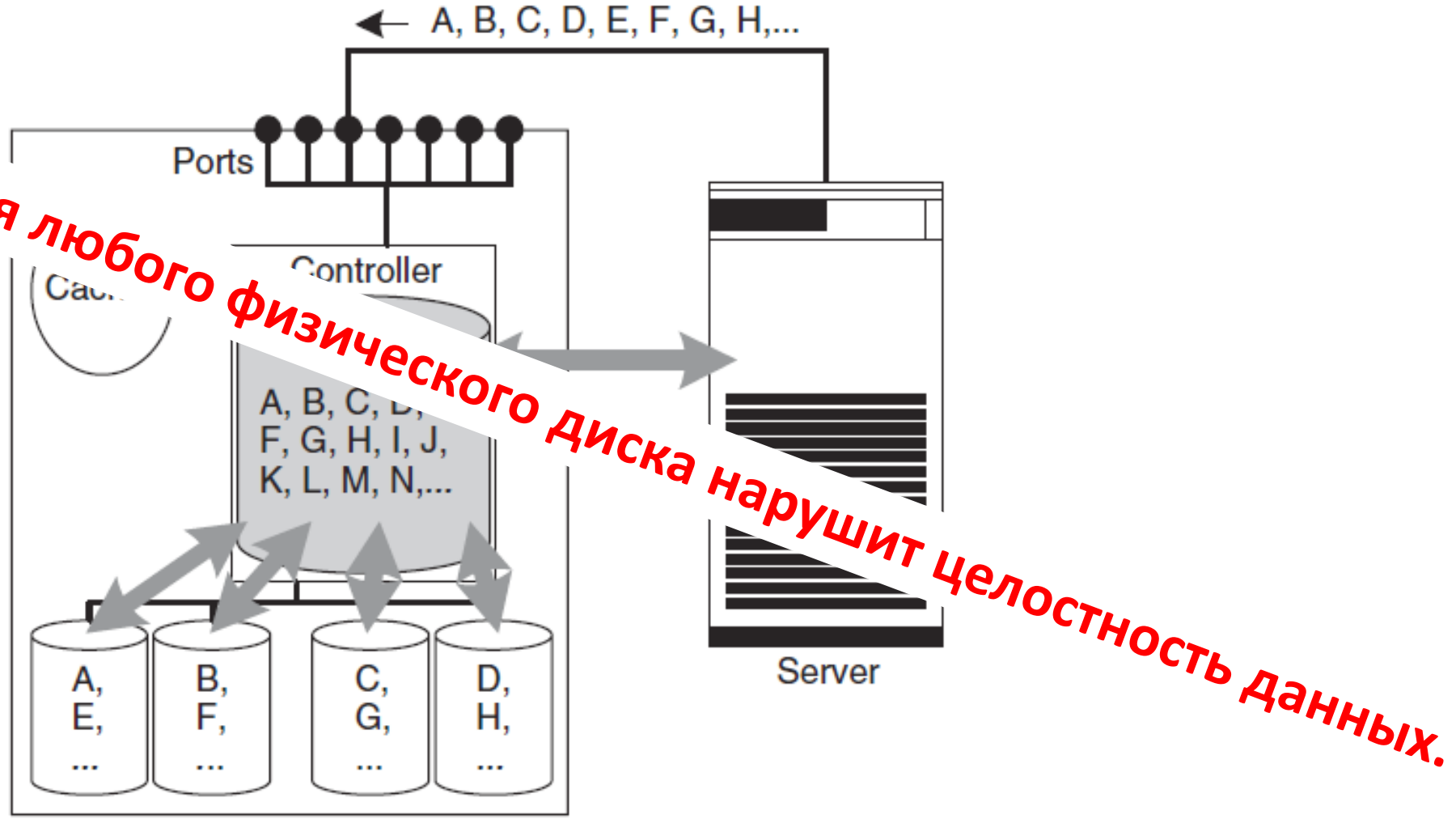


- **Increase performance by striping**
- **Increase fault-tolerance by redundancy**

Hot Spare Disk



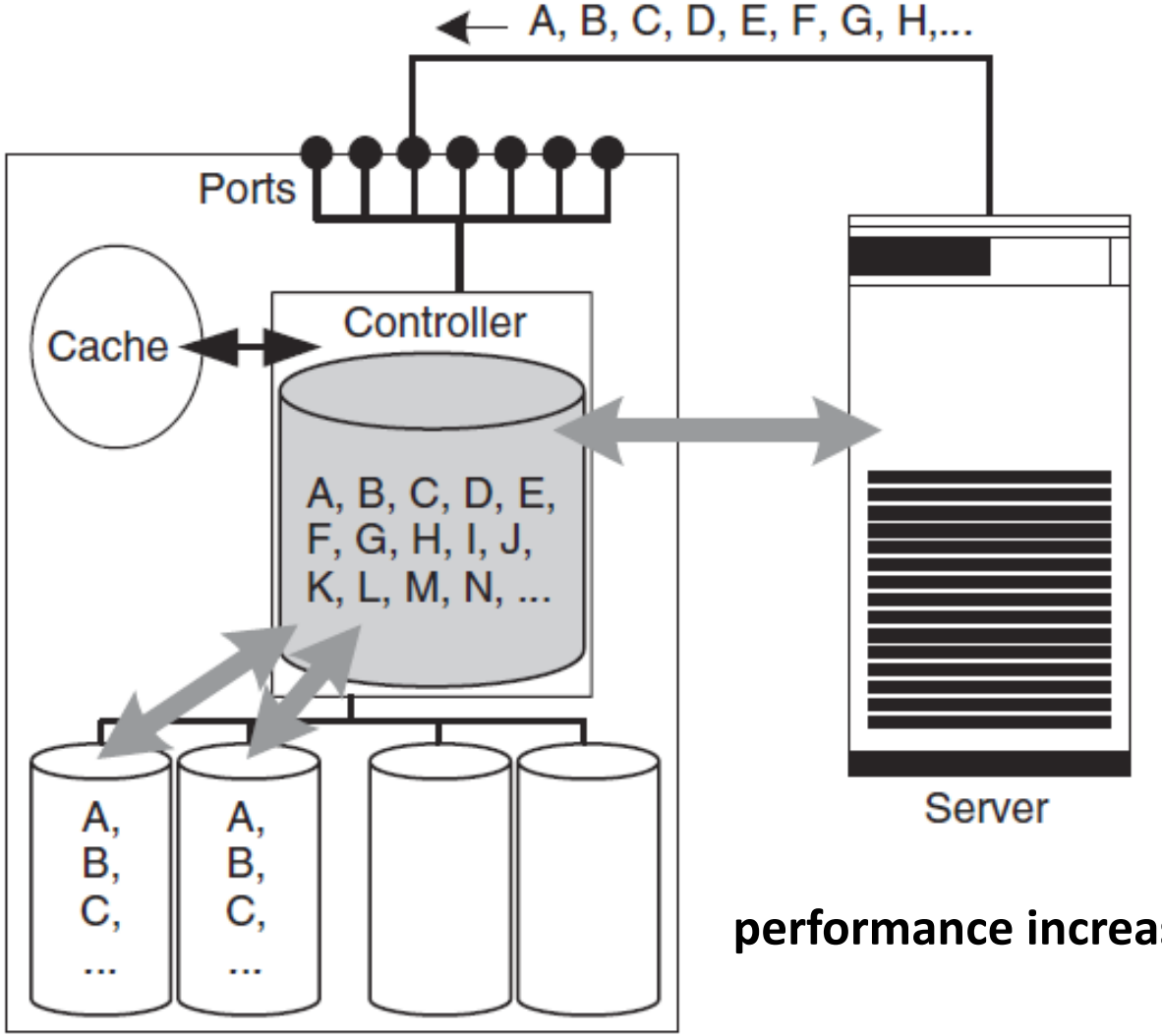
RAID 0: block-by-block striping



Выход из строя любого физического диска нарушит целостность данных.

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

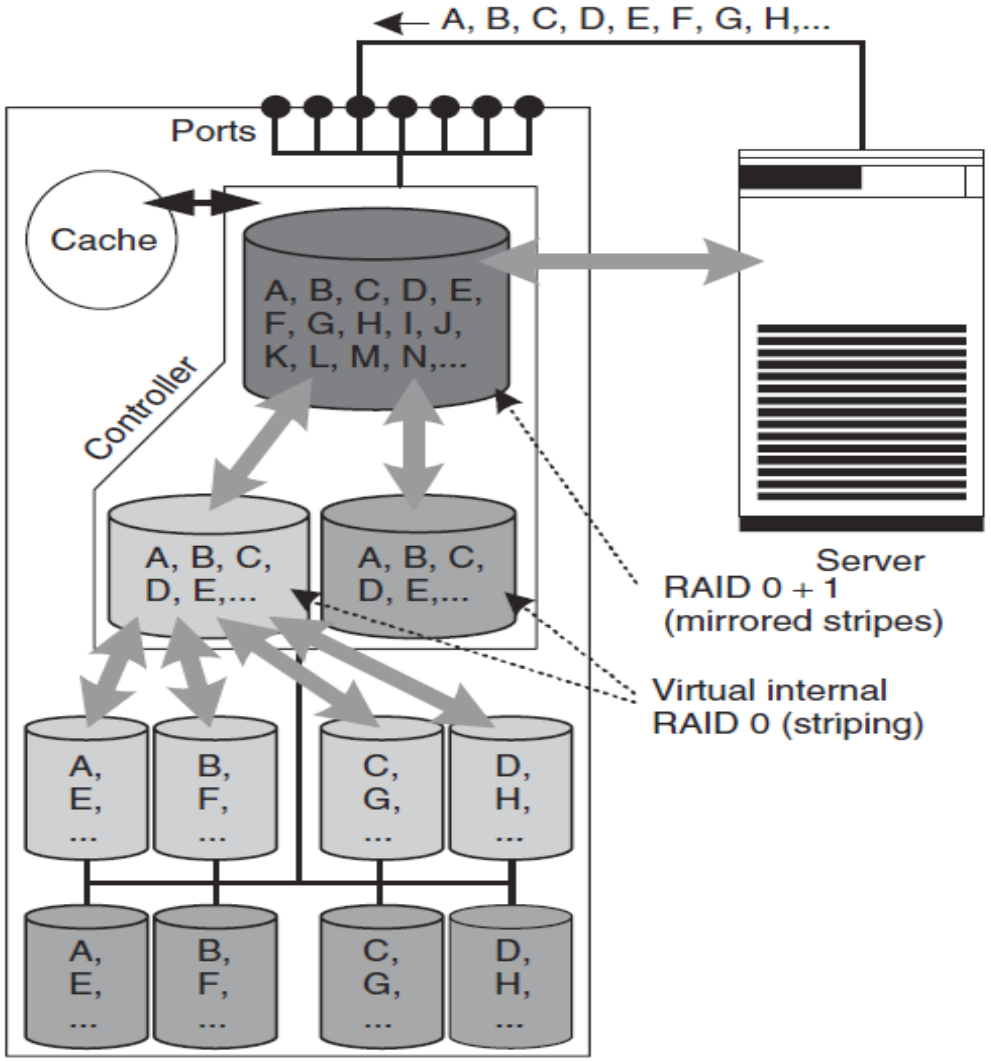
RAID 1: block-by-block mirroring



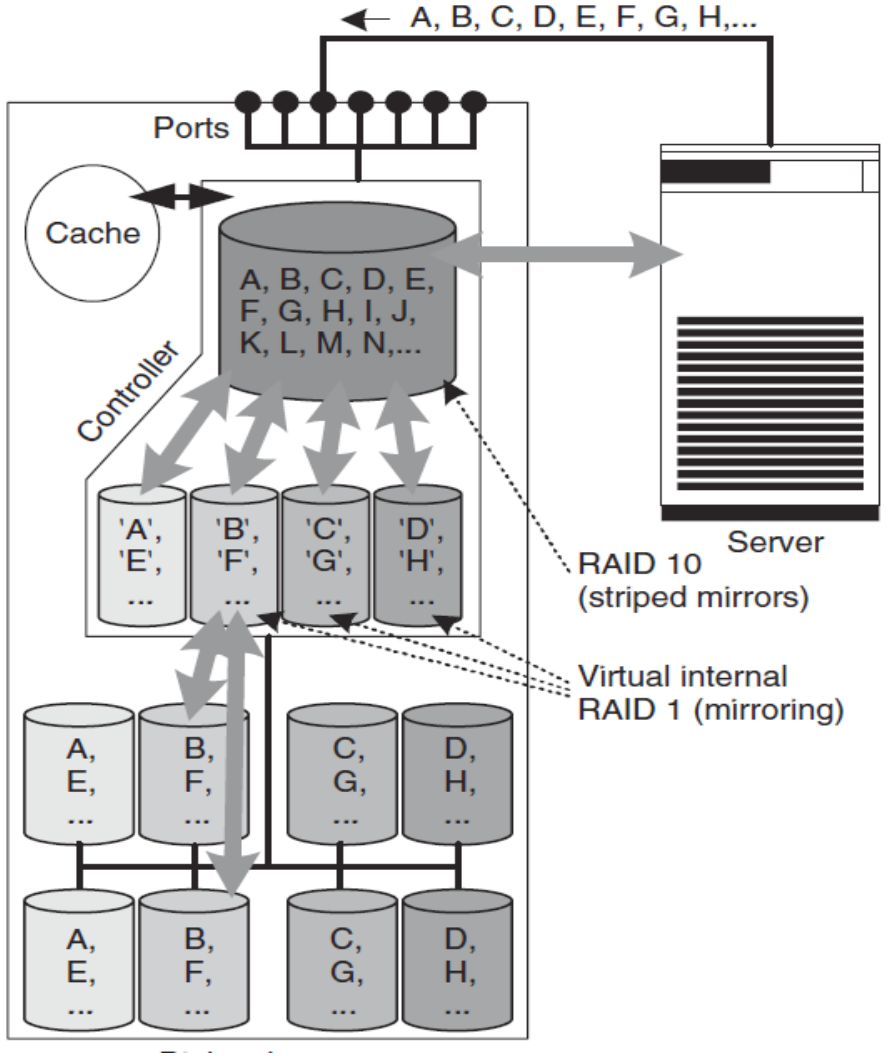
performance increases are only possible in read operations

fault-tolerance is of primary importance

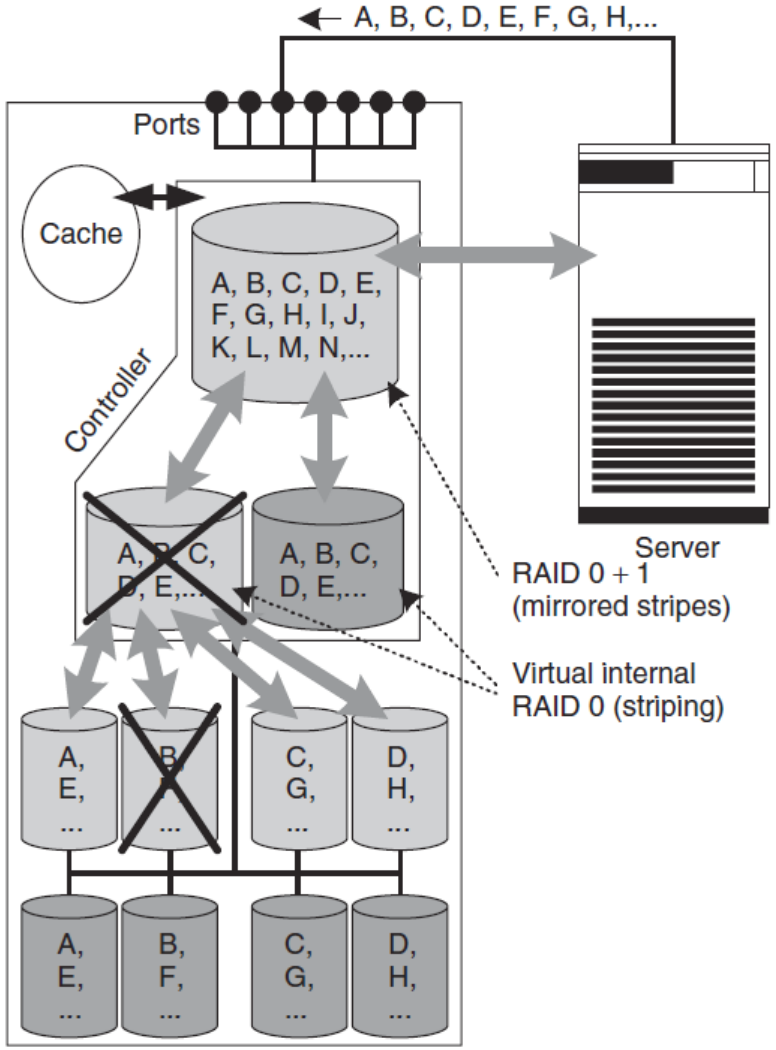
RAID = 0+1 (mirrored stripes)

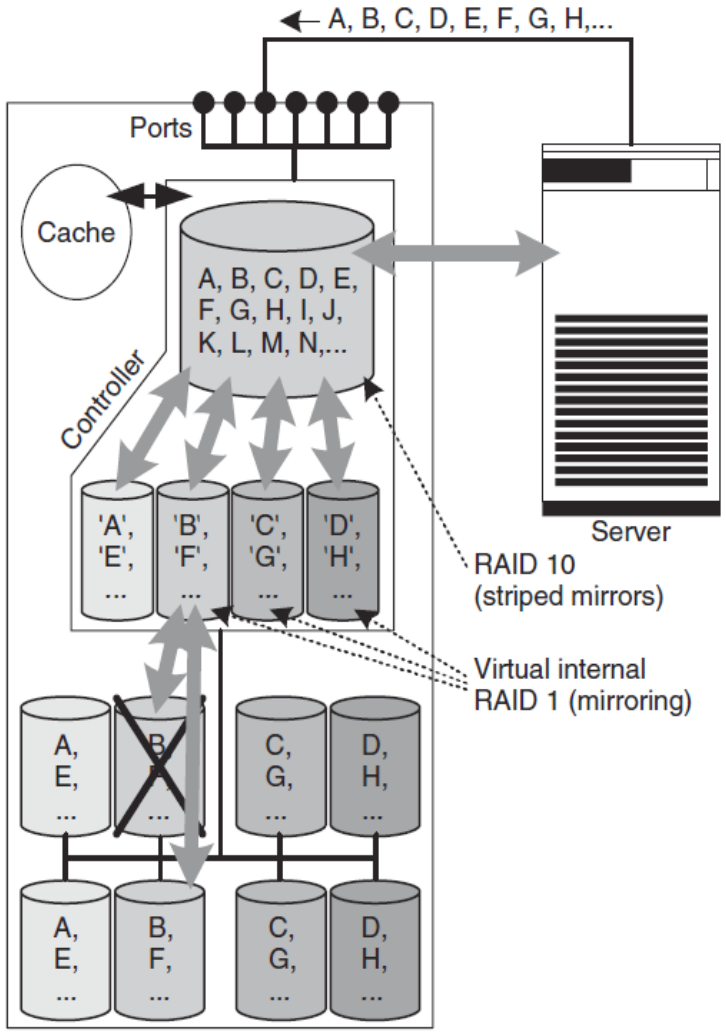


RAID = 1+0 (striped mirrors)

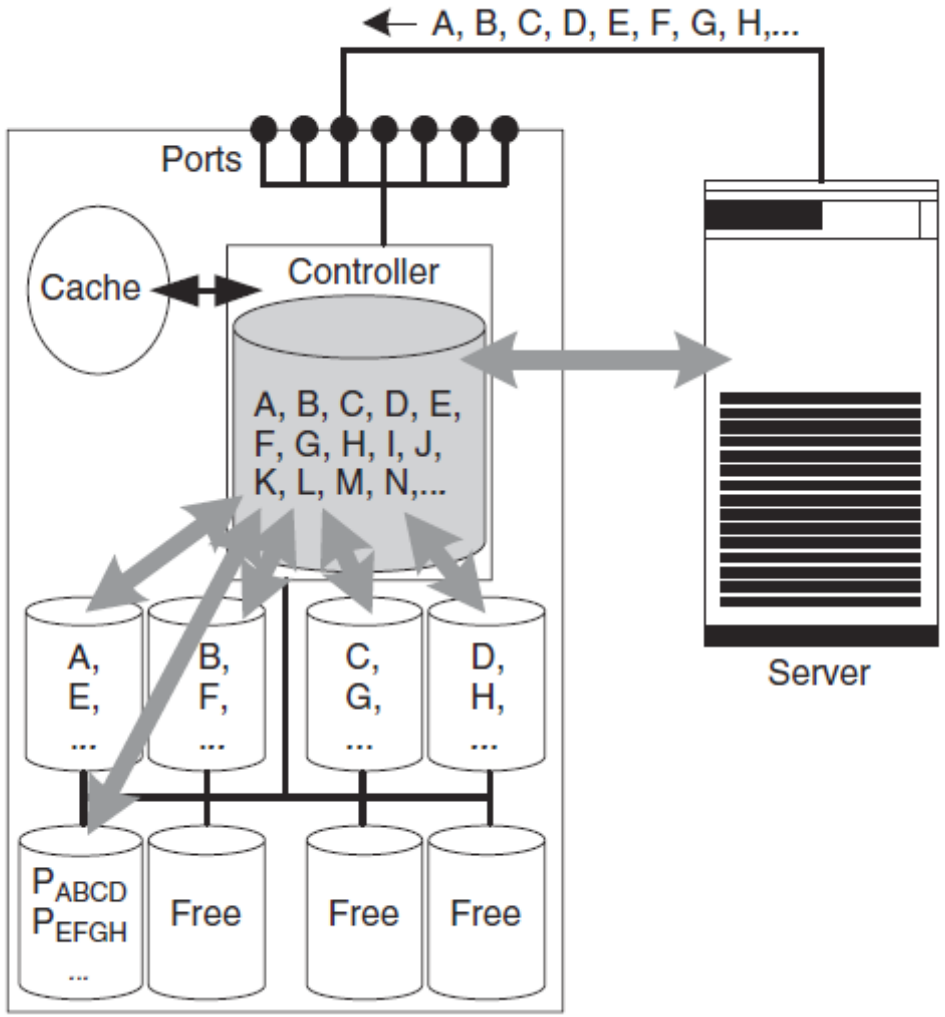


Comparison RAID 0+1 vs 1+0





RAID 4 and RAID 5: parity instead of mirroring



A изменили на ~A

Δ = A xor ~A

~P = Δ xor P

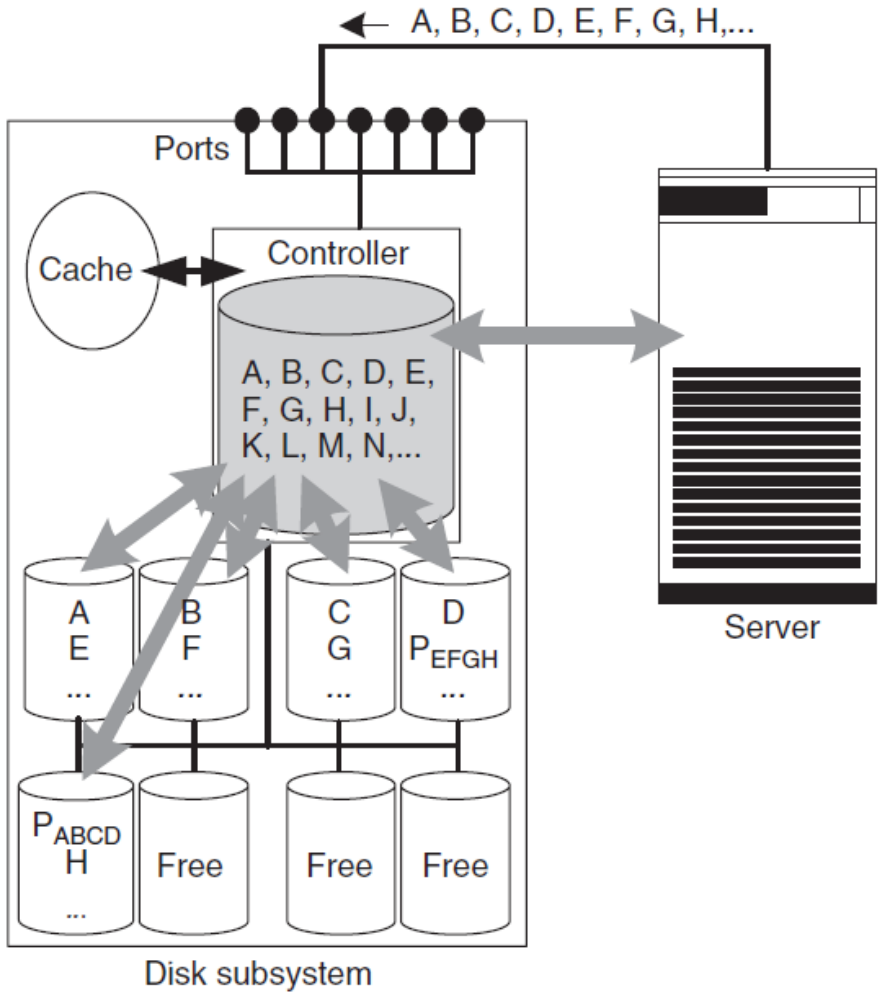
Если изменился только блок A,

То легко пересчитать P_{ABCD}, не зная BCD.

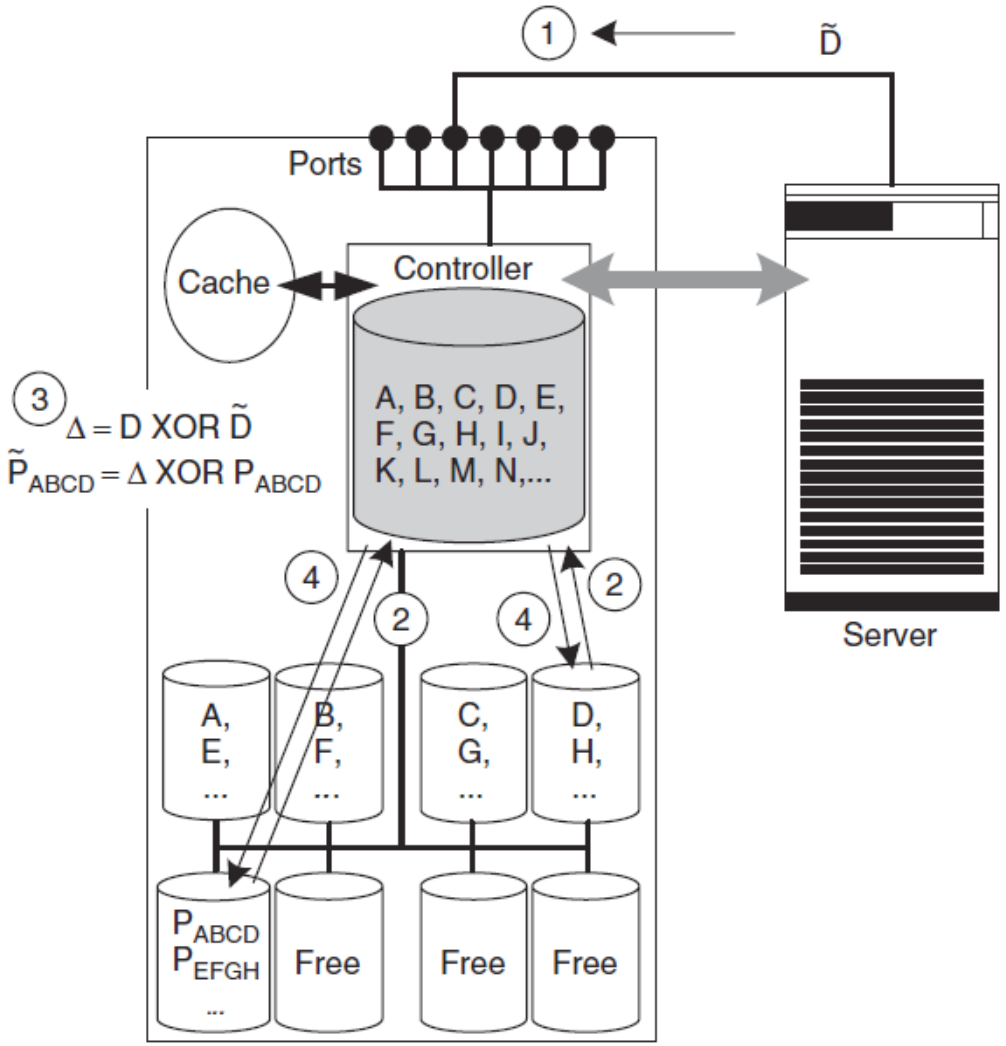
Однако надо считать старый блок A,

чтобы рассчитать Δ

RAID 5 striped parity



Overheads RAID 4 and 5



RAID 6: double parity

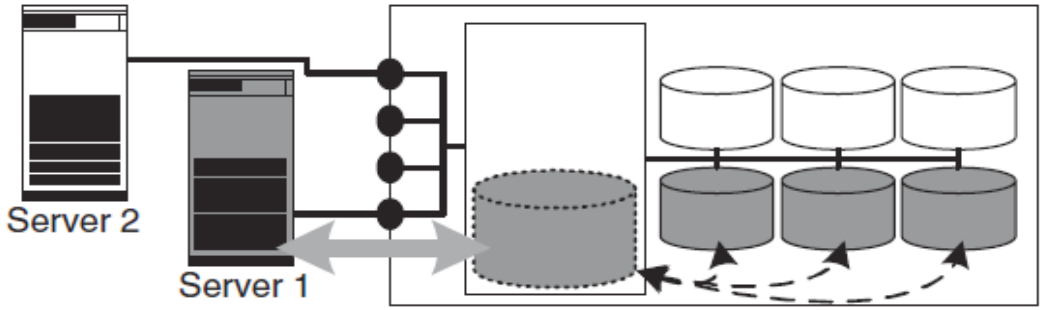
Comparison RAID schemes

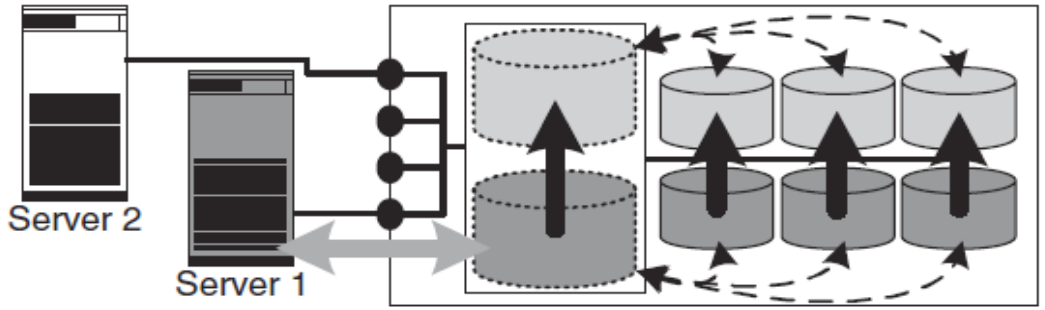
RAID level	Fault-tolerance	Read performance	Write performance	Space requirement
RAID 0	None	Good	Very good	Minimal
RAID 1	High	Poor	Poor	High
RAID 10	Very high	Very good	Good	High
RAID 4	High	Good	Very very poor	Low
RAID 5	High	Good	Very poor	Low
RAID 6	Very high	Good	Very very poor	Low

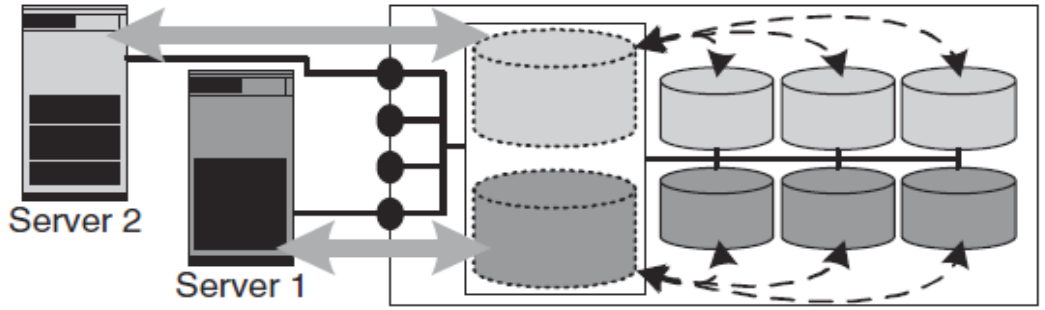
- Recently 1TV HDD with BER 10 ^ -15 => one 100 TB sector is lost when reading
- 10 16X1TV disk arrays will lose one array once a year +
- Operation mode is now 7X24
- RAID 6 uses an extra parity disk for group errors
- Cost increase
- Increase recording and correction operation time

- **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**

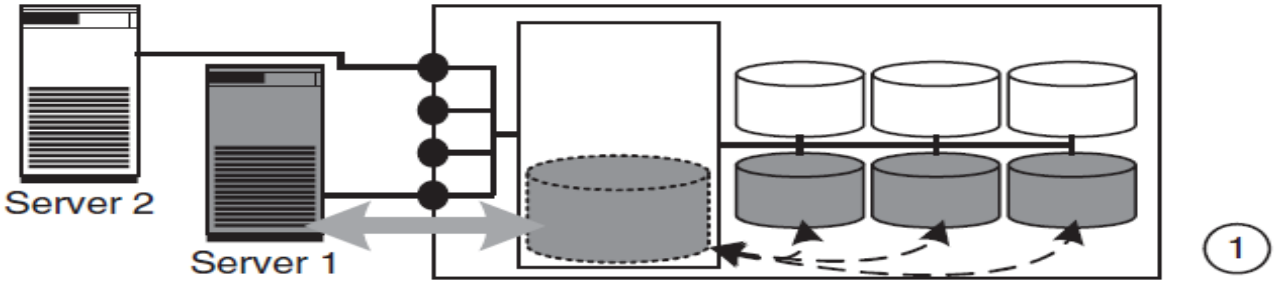
INTELLIGENT DISK SUBSYSTEMS (Instant copies)

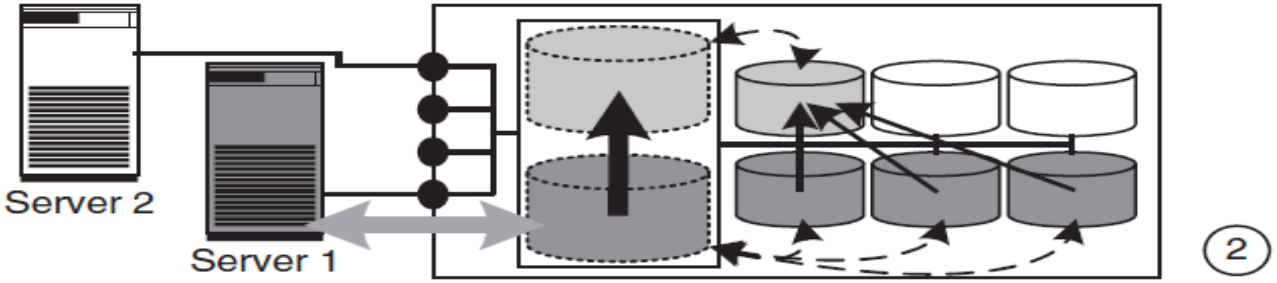


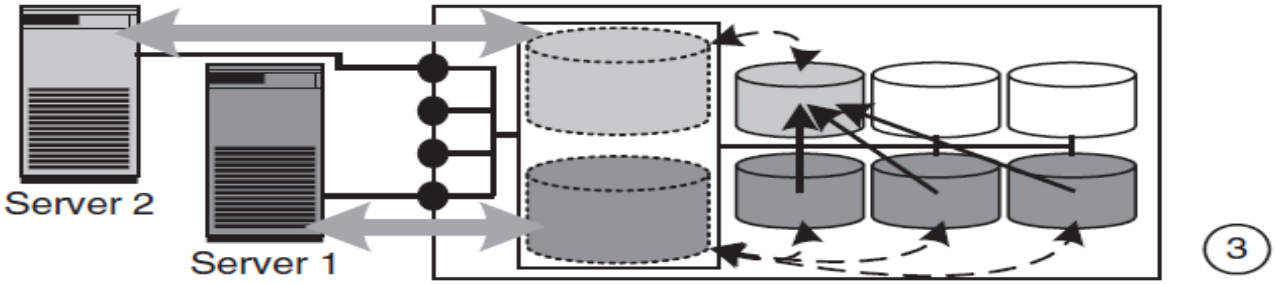




Space-efficient instant copy



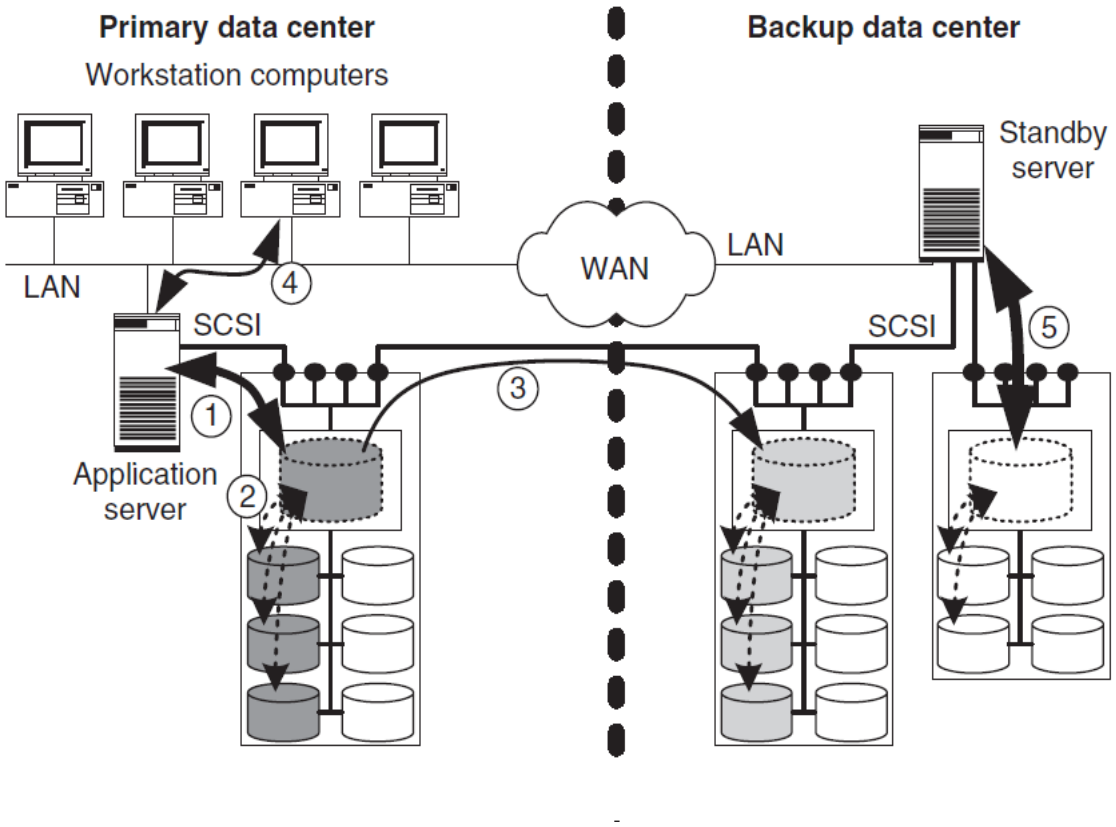


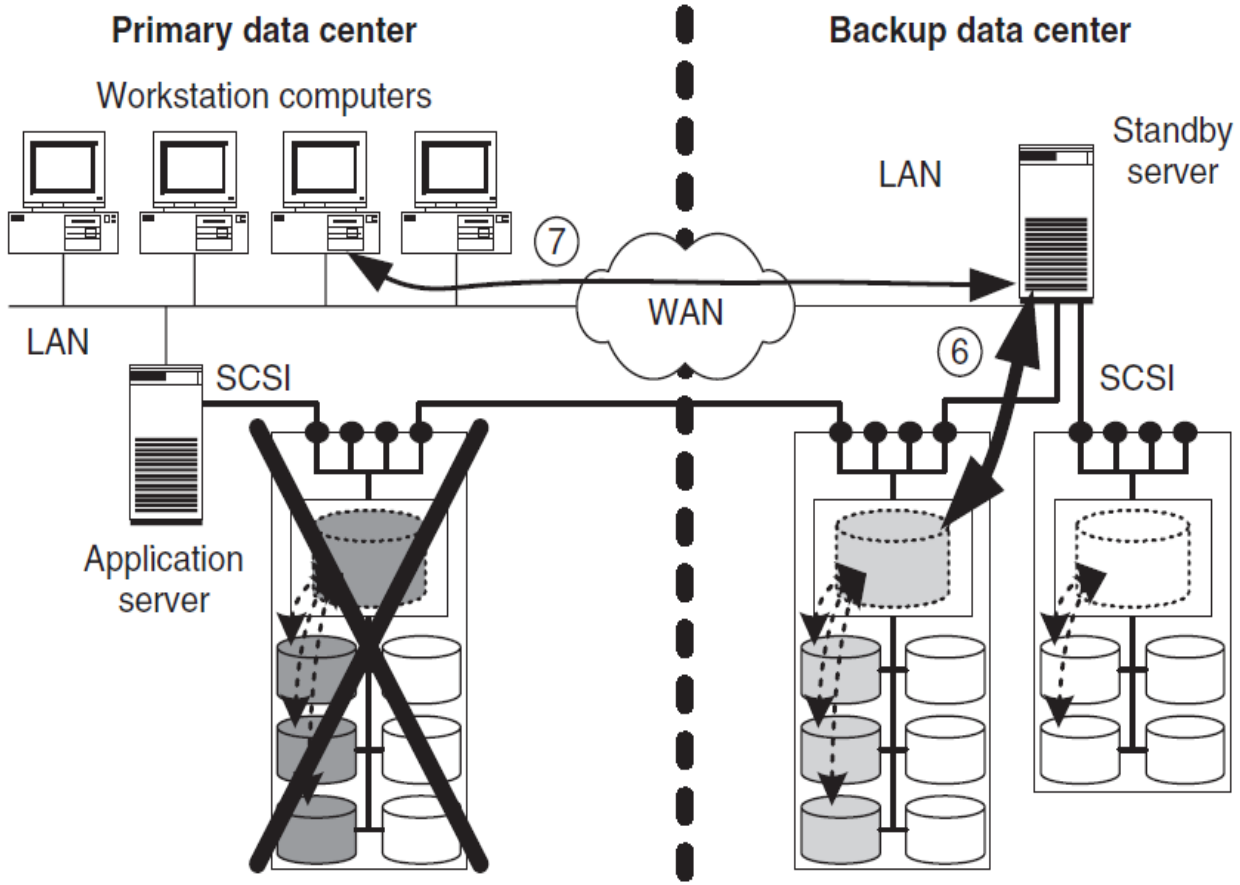


Incremental instant copy

Reversal of instant copy

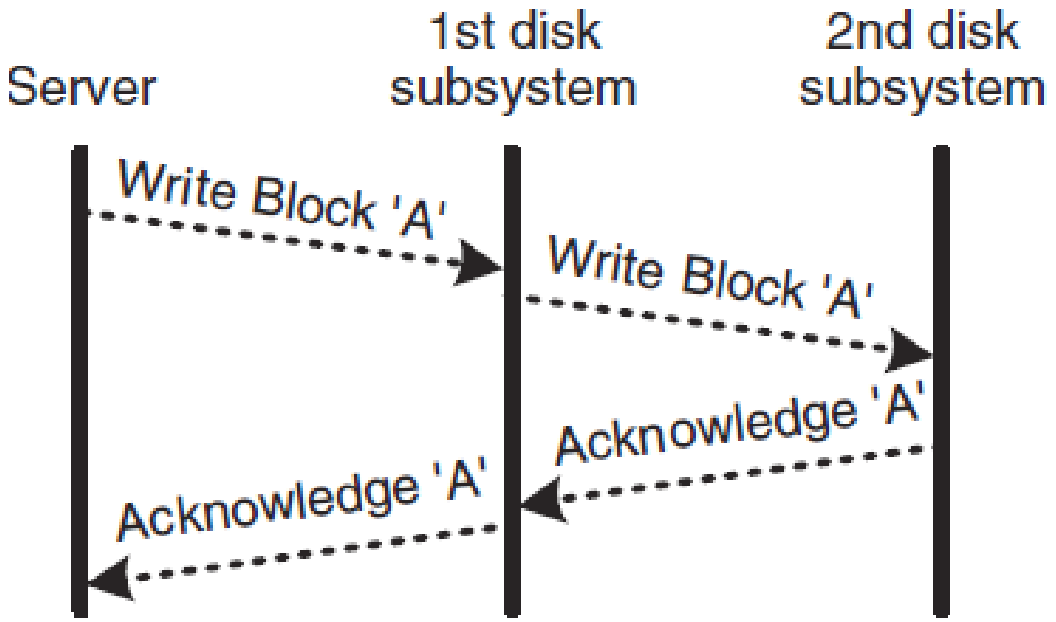
Remote mirroring (catastrophe resilience)



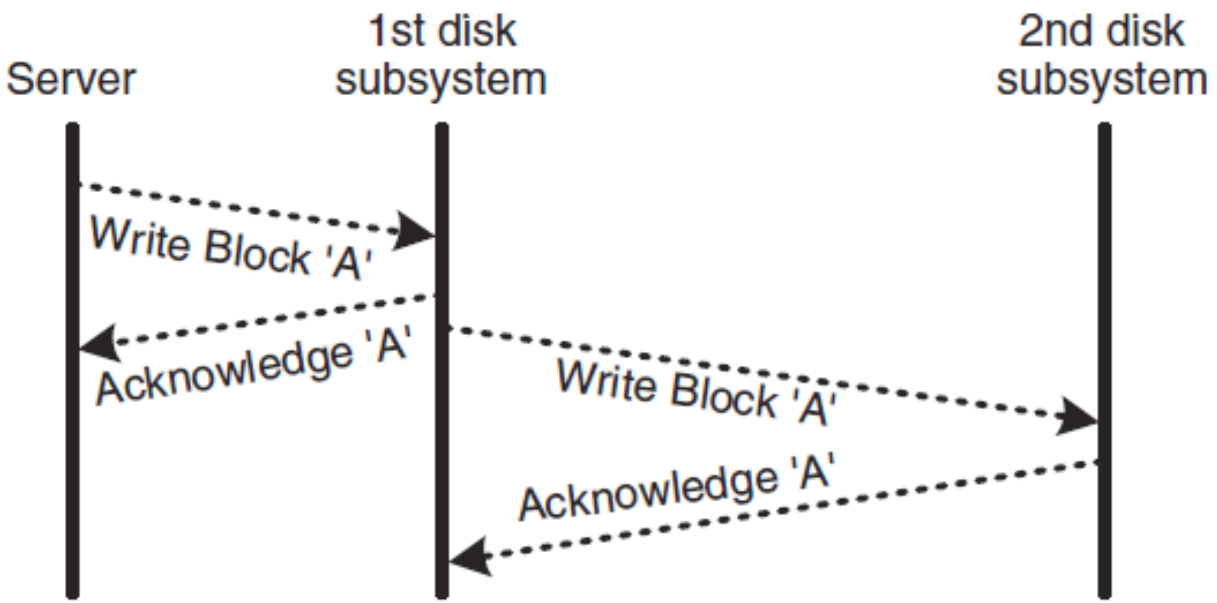


For data protection, the proximity of production data and data copies is fatal.

Synchronous vs Asynchronous remote mirroring

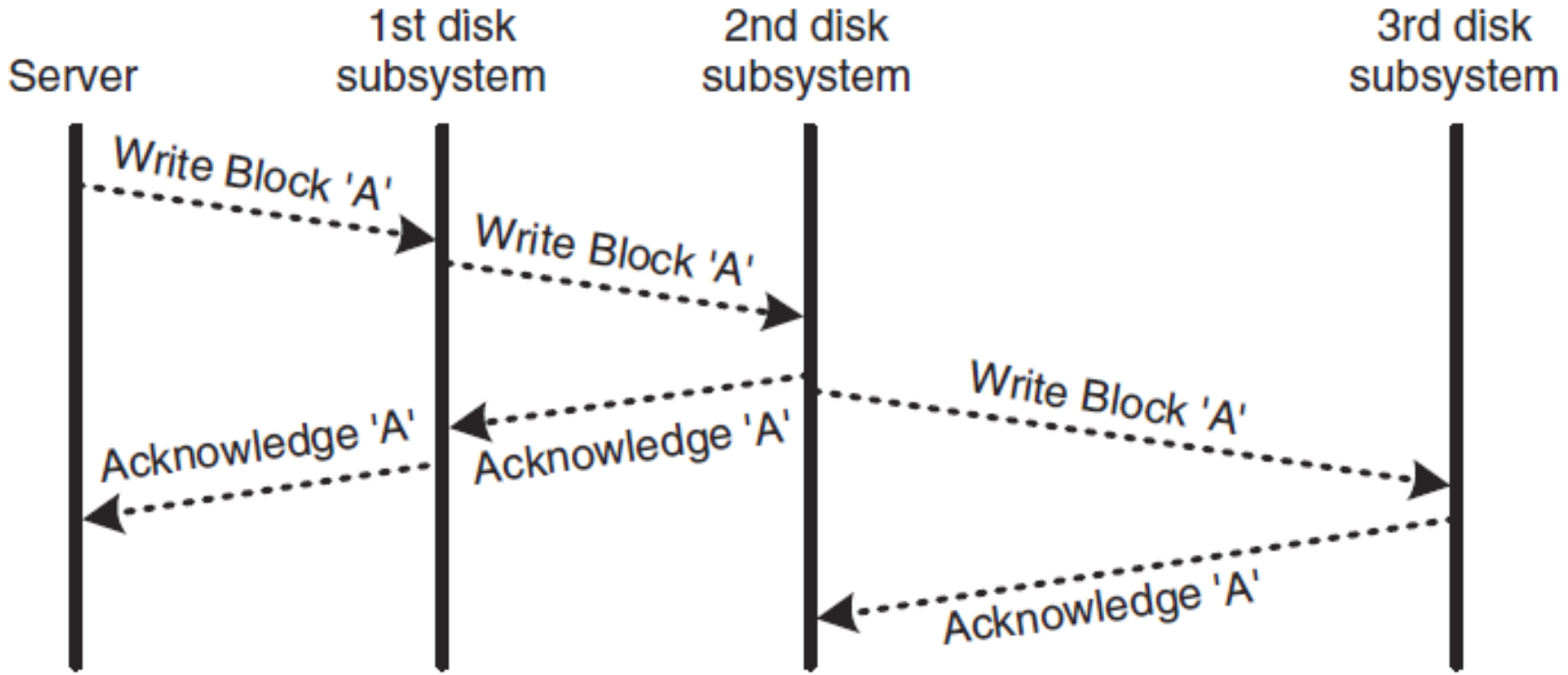


Synchronous

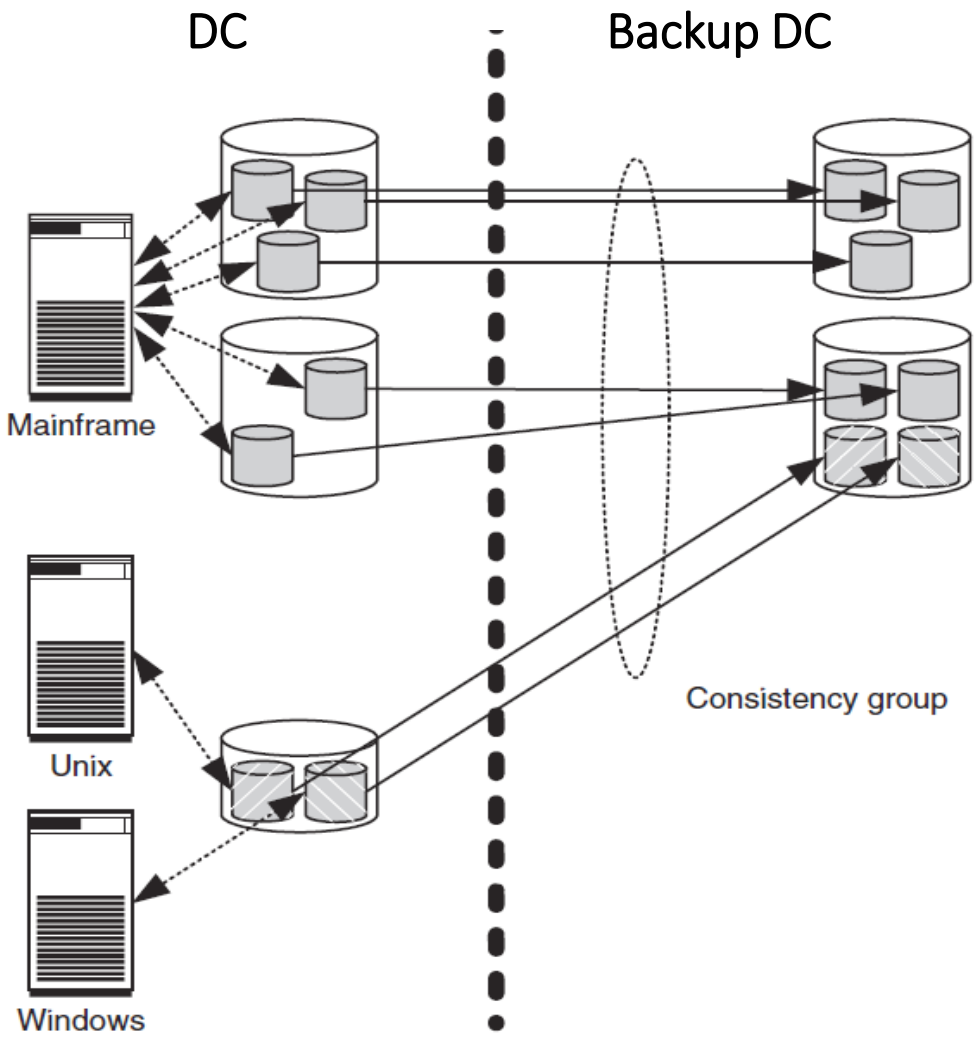


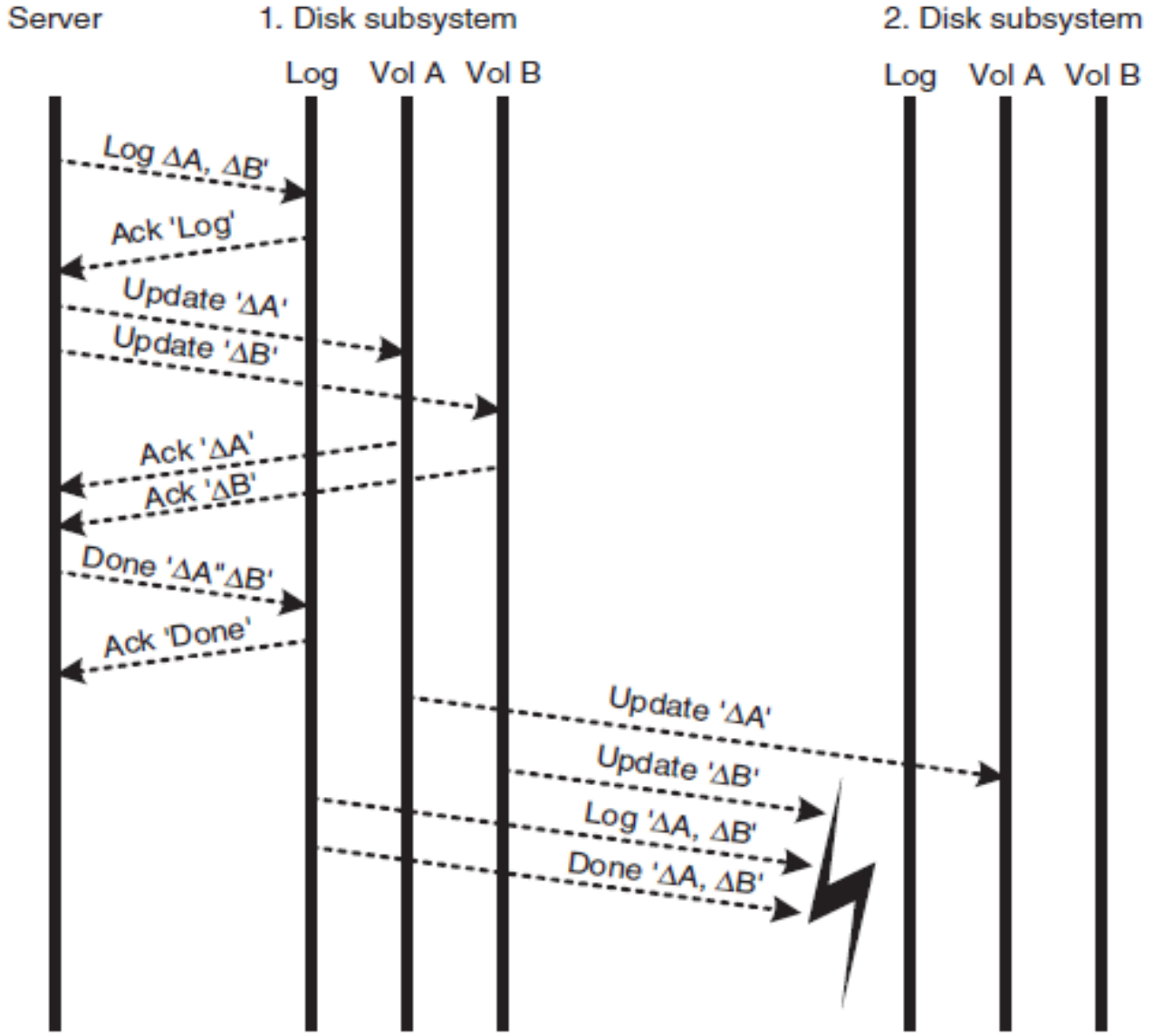
Asynchronous

Mirroring data over long distances

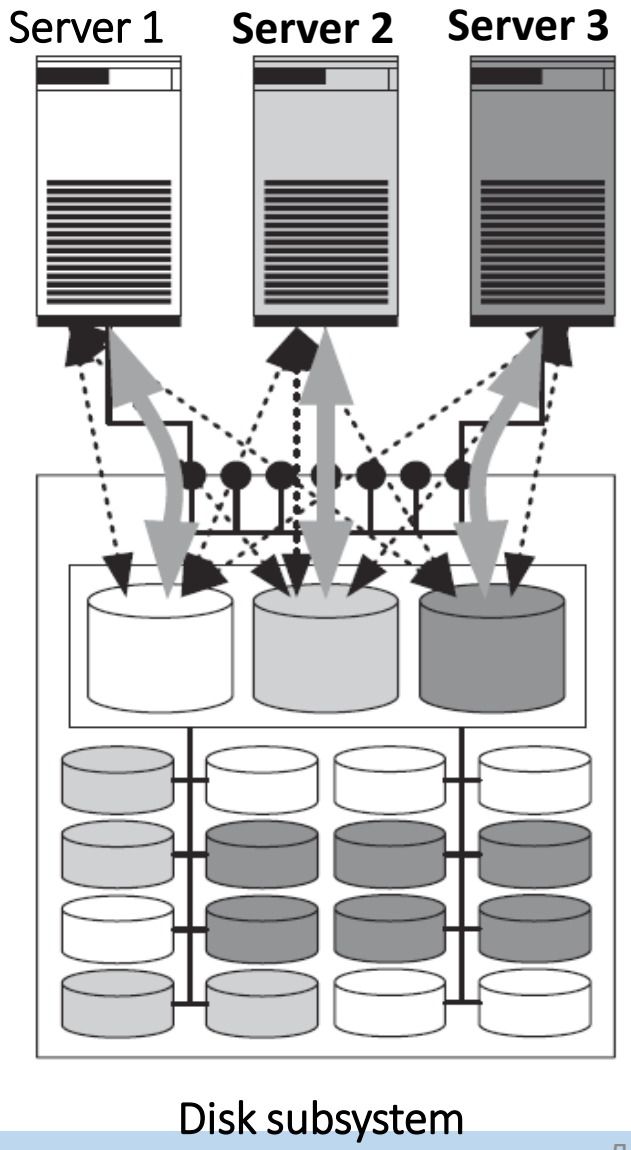


Consistency groups





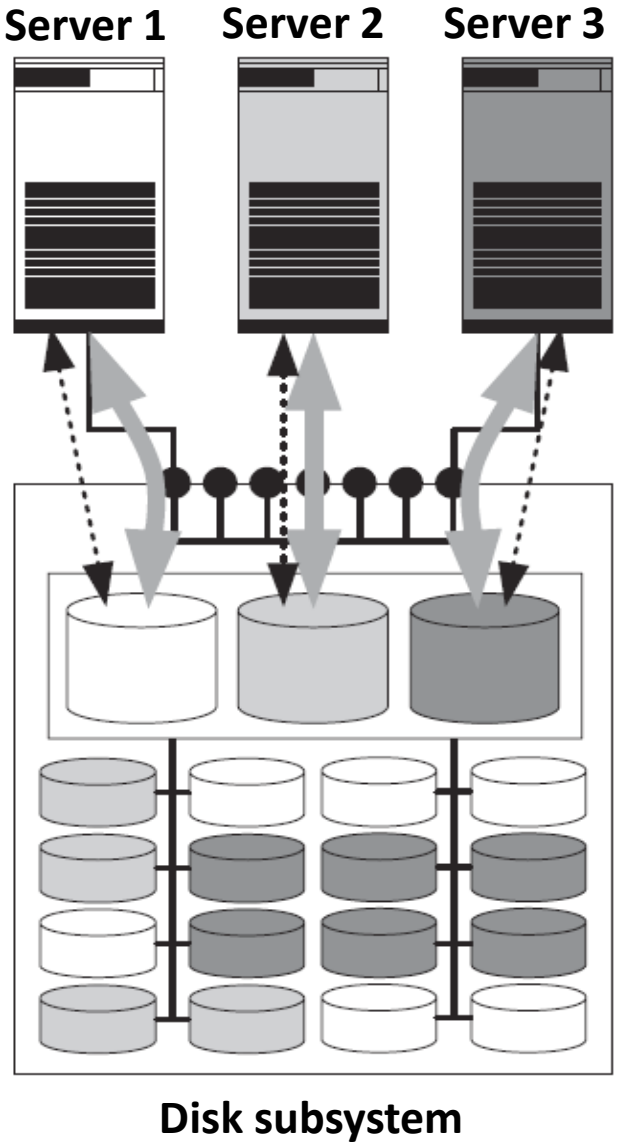
LUN masking



Logical
Unit
Number

Server uses LUN

Server sees LUN



AVAILABILITY OF DISK SUBSYSTEMS

- Данные распределяются по нескольким дискам с использованием механизмов RAID и предоставляют избыточные данные (блоки четности).
- На каждом физическом диске данные закодированы кодом Хэмминга. Кроме того, диск оснащен подсистемой самодиагностики, контролирующей частоту ошибок, вибрацию шпинделя и т. д. Это позволяет заблаговременно прогнозировать отказы диска.
- Каждый диск подключается к контроллеру как минимум через две внутренние шины.

• Контроллер дисковой подсистемы можно дублировать. Выход одного экземпляра автоматически активирует следующий экземпляр. Активный режим ожидания

• Дублированные системы охлаждения ИБП. DSподключаются к разным электрическим сетям

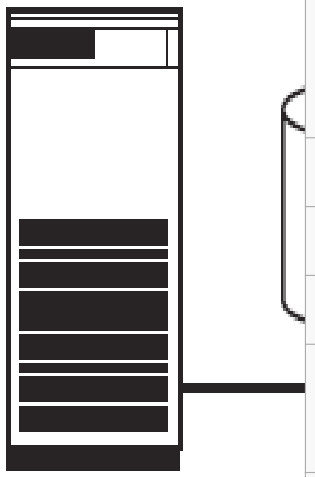
• Сервер подключен к DS несколькими линиями.

• Используйте периодическое мгновенное копирование для защиты от логических ошибок. Например, создание мгновенной копии данных каждый час. Тогда в случае сбоя и уничтожения какой-то таблицы ее можно будет восстановить.

• Удаленное зеркалирование используется при физическом уничтожении или повреждении оборудования (аварийная устойчивость). В сочетании с мгновенным копированием эти услуги гарантируют сохранение и согласованность данных даже для нескольких виртуальных дисков или дисковых подсистем.

• Маскировка LUN защищает от несанкционированного доступа, упрощает работу системного администратора и защищает от случайных сбоев в работе серверных приложений и их оборудования.

Small Computer System Interface (SCSI)

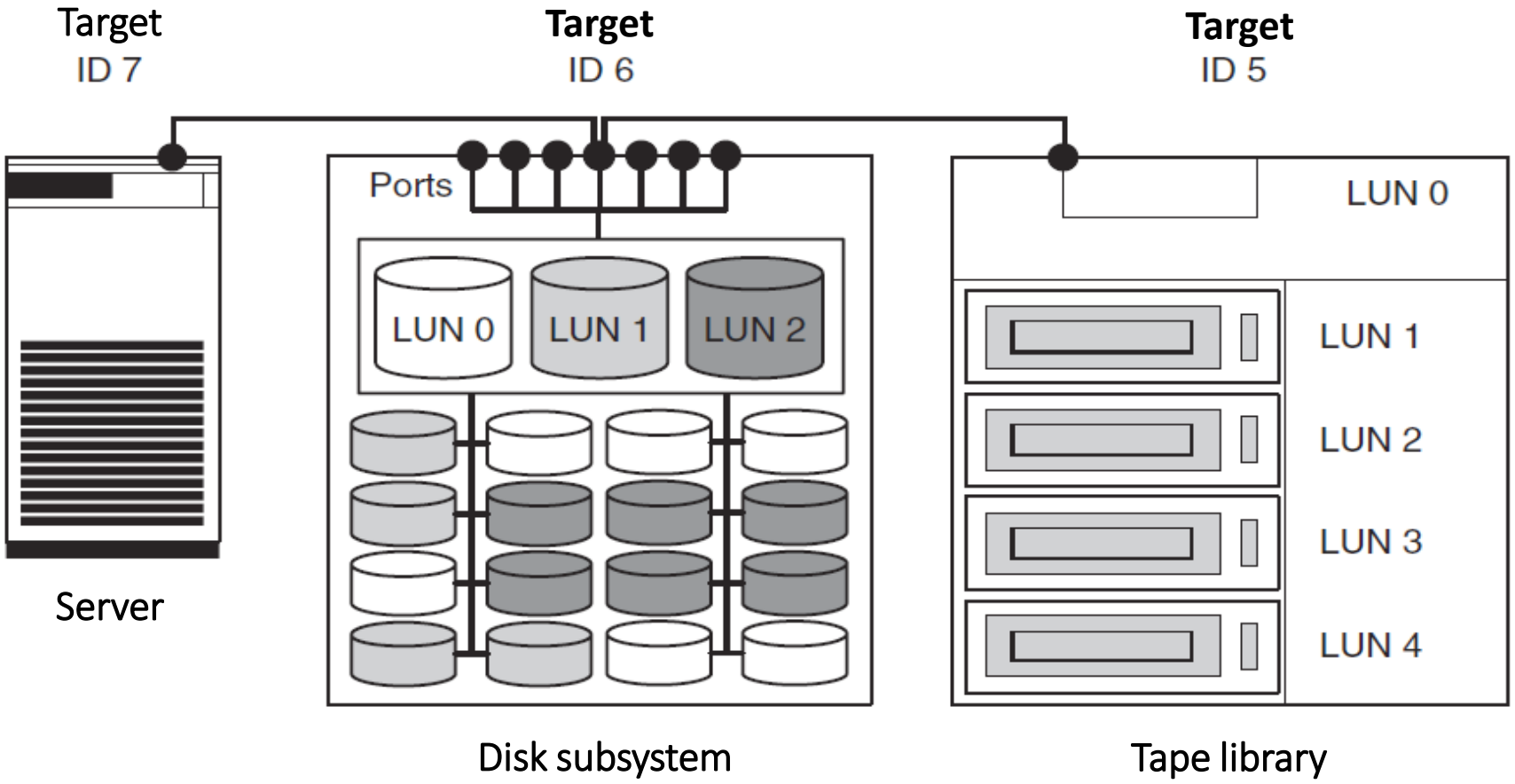


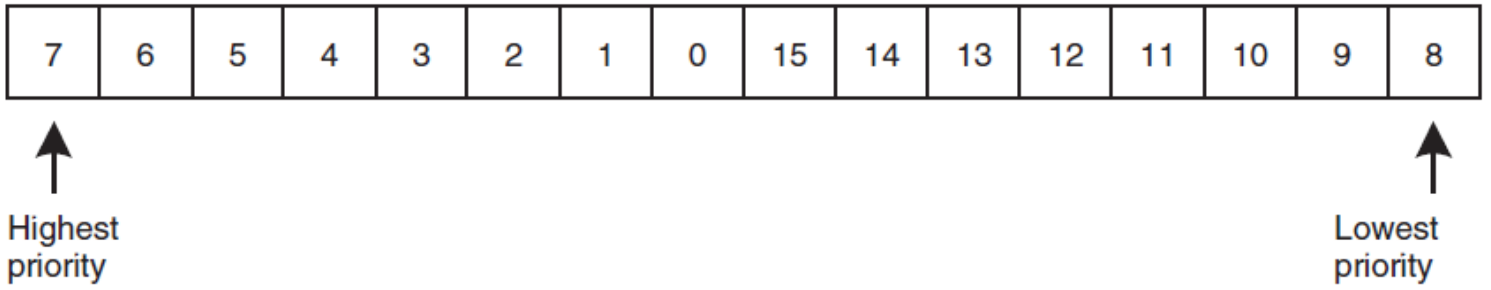
Server

- Cable:
- Protoc

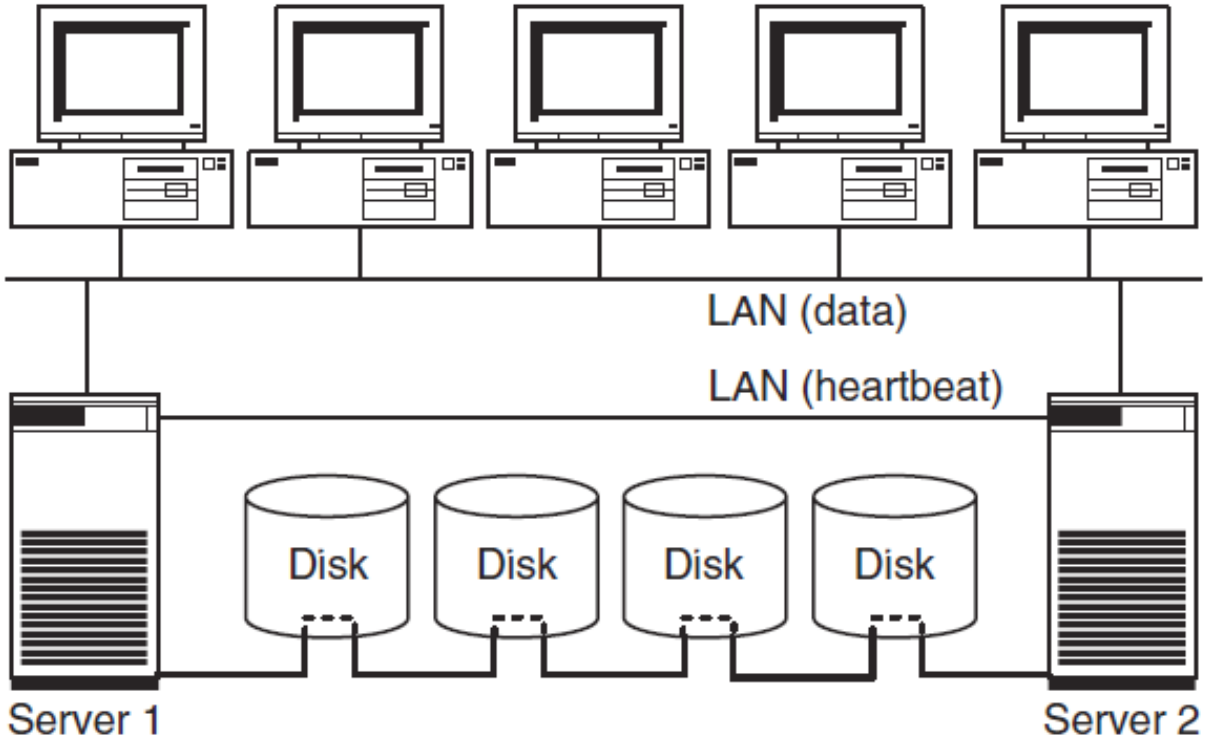
Обзор интерфейсов SCSI					
Наименование	Разрядность шины	Частота шины	Пропускная способность	Максимальная длина кабеля	Максимальное количество устройств
SCSI	8 бит	5 МГц	5 Мб/с	6 м (25 м с HVD)	8
Fast SCSI	8 бит	10 МГц	10 Мб/с	3 м (25 м с HVD)	8
Wide SCSI	16 бит	10 МГц	20 Мб/с	3 м (25 м с HVD)	16
Ultra SCSI	8 бит	20 МГц	20 Мб/с	1,5—3 м (25 м с HVD)	4—8
Ultra Wide SCSI	16 бит	20 МГц	40 Мб/с	1,5—3 м (25 м с HVD)	4—16
Ultra2 SCSI	8 бит	40 МГц	40 Мб/с	12 м (25 м с HVD)	8
Ultra2 Wide SCSI	16 бит	40 МГц	80 Мбайт/сек	12 м (25 м с HVD)	16
Ultra3 SCSI	16 бит	40 МГц DDR	160 Мб/с	12 м	16
Ultra-320 SCSI	16 бит	80 МГц DDR	320 Мб/с	12 м	16
Ultra-640 SCSI	16 бит	160 МГц DDR	640 Мб/с	10 м	16

Device addressing on SCSI

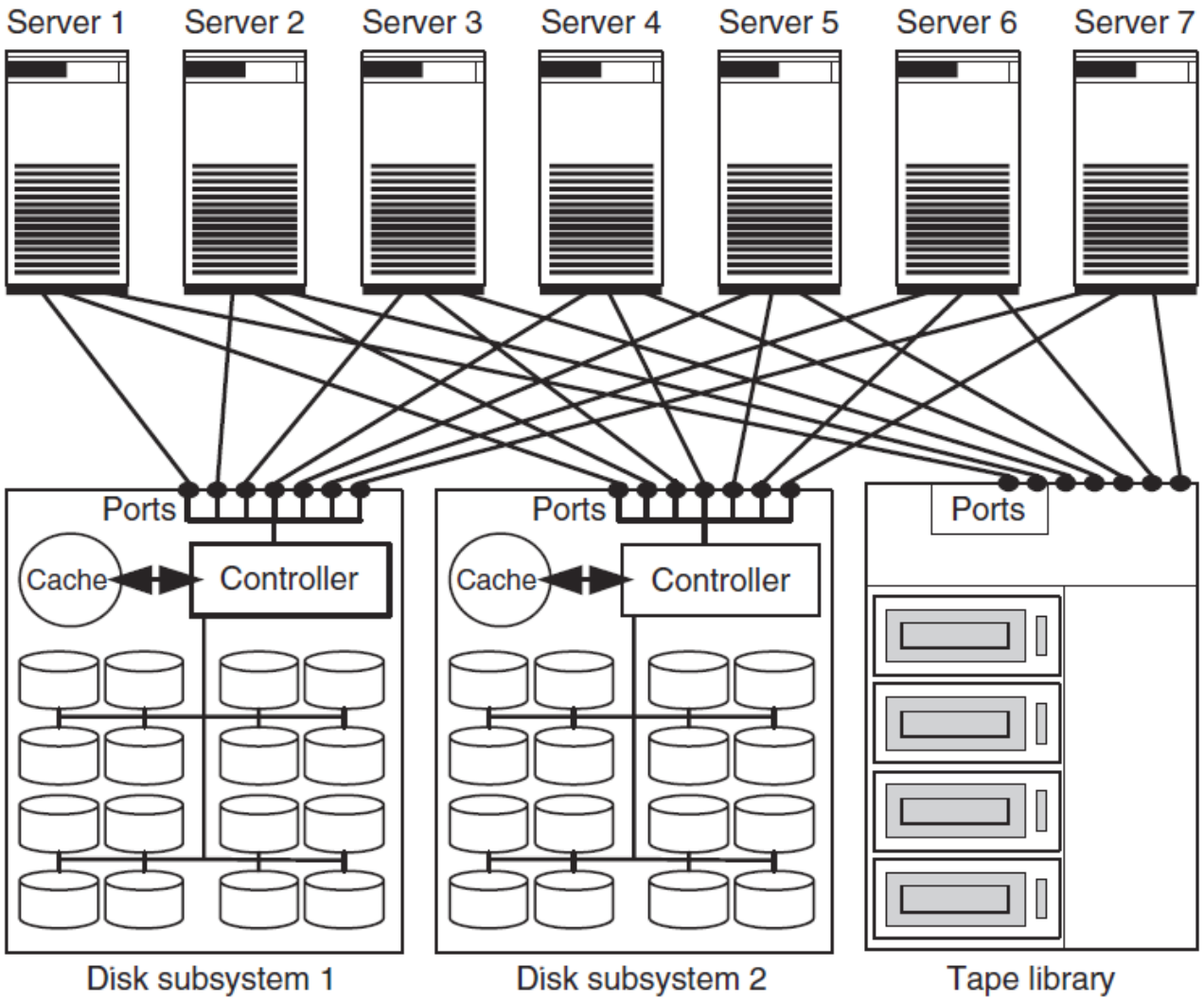




Faultolerant SCSI storage networks



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

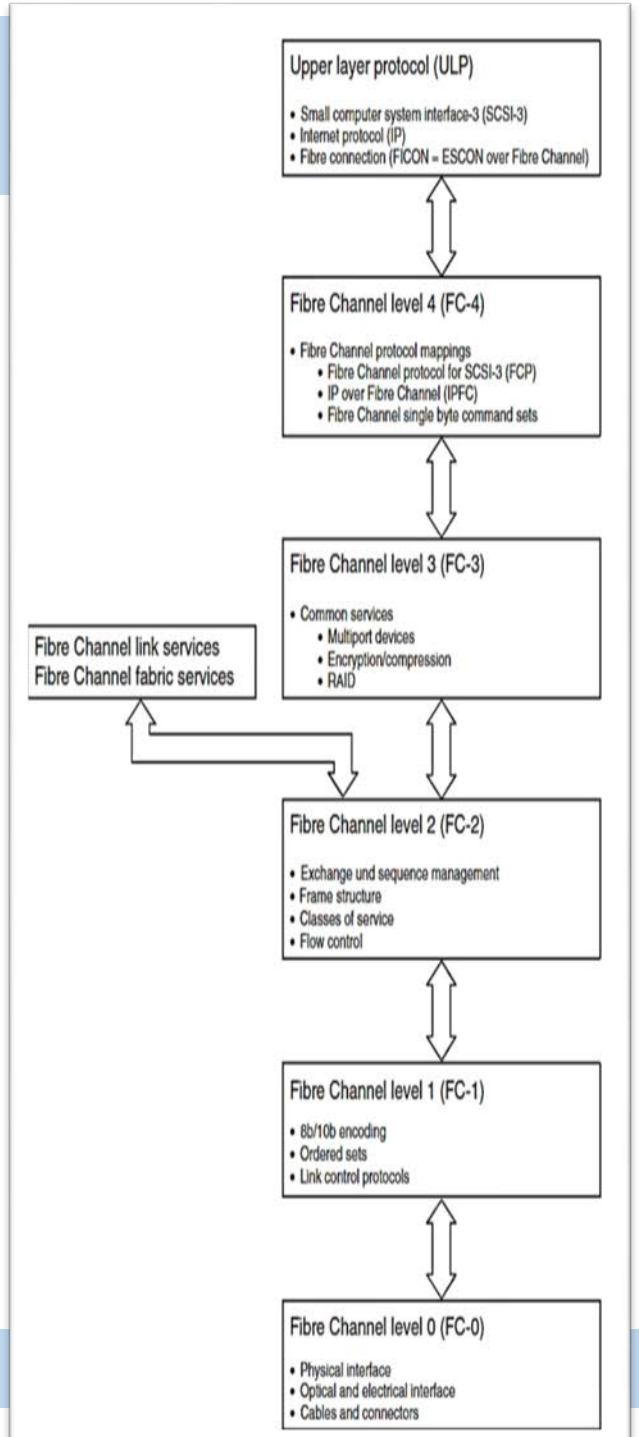
FC - 0 определяет физические характеристики интерфейса и среды, включая кабели, разъемы, драйверы (ECL, LED, лазеры), передатчики и приемники. Вместе с FC-1 этот уровень образует физический слой.

FC - 1 определяет метод кодирования/декодирования (8B/10B) и протокол передачи, где объединяется пересылка данных и синхронизирующей информации.

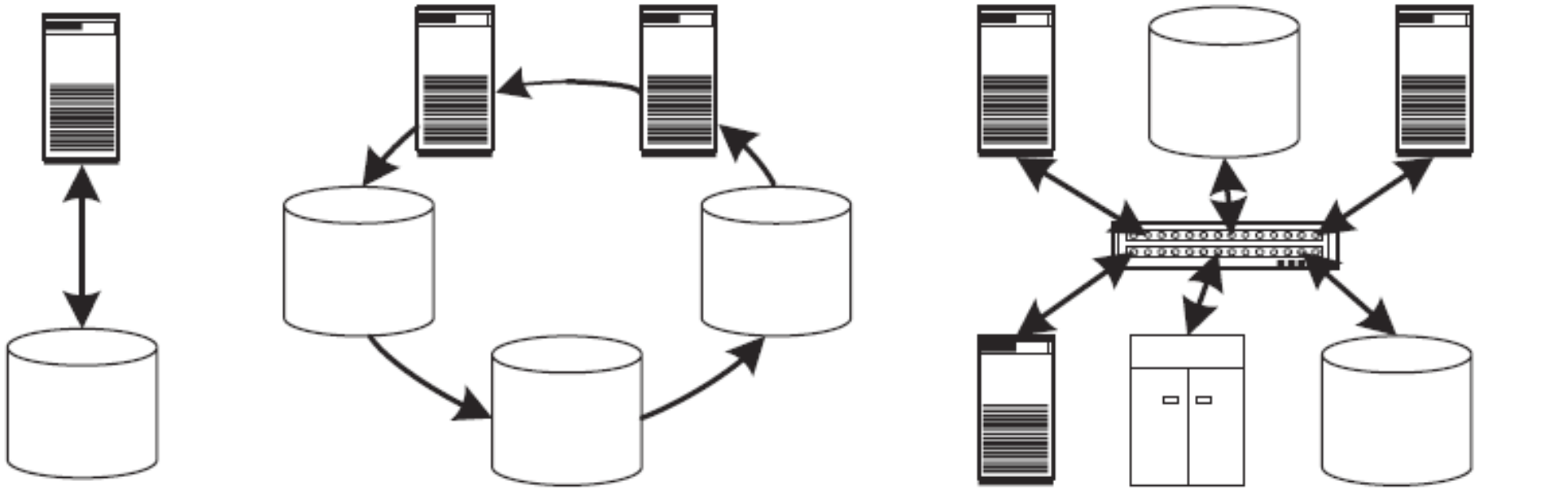
FC - 2 (управление передачей) определяет правила сигнального протокола (управление передачей), классы услуг, топологию, методику сегментации, задает формат кадра и описывает передачу информационных кадров.

FC - 3 (адресация) определяет работу нескольких портов на одном узле и обеспечивает общие виды сервиса.

FC - 4 обеспечивает реализацию набора прикладных команд и протоколов вышележащего уровня (например, для SCSI, IPI, IEEE 802, SBCCS, HIPPI, IP, ATM и т.д.)



Links, ports and topologies



Point-to-point Arbitrated loop Fabric **(матрица)**

Fibre Channel: port types

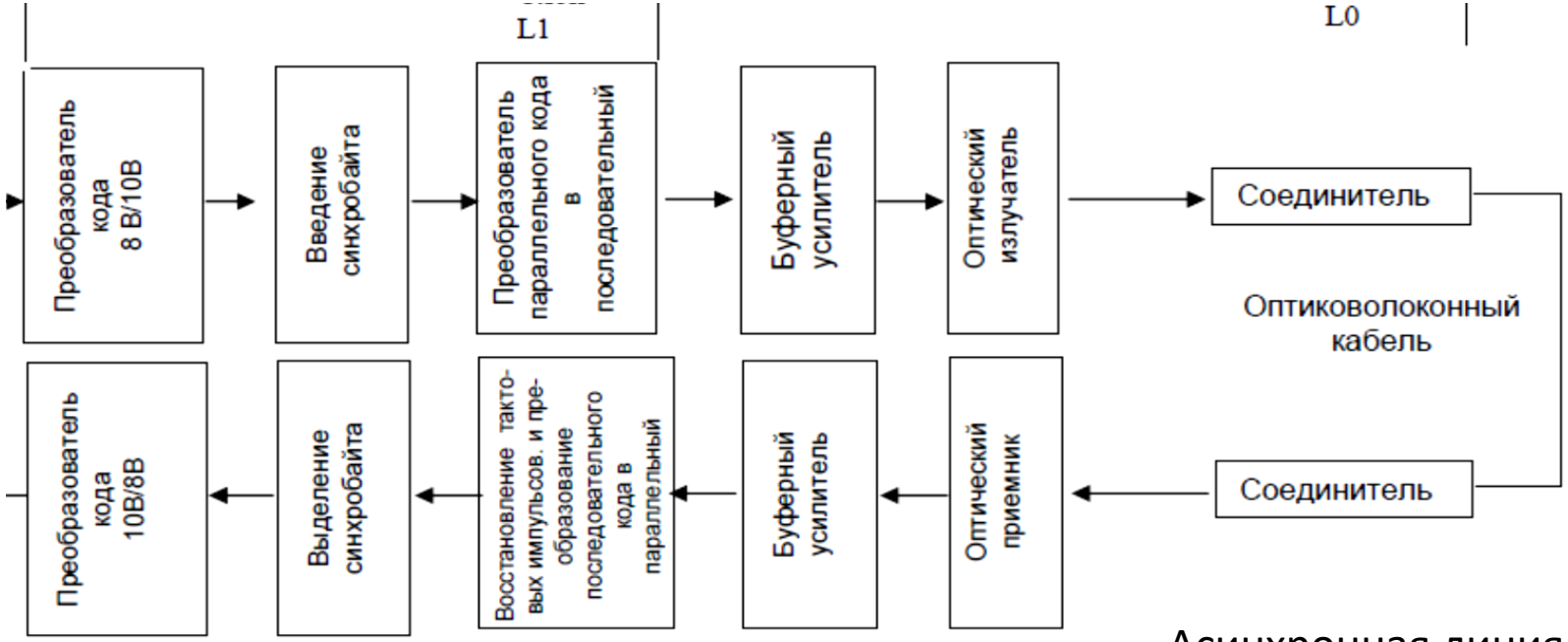
- **N-Port** – describes the capability of a port as an end device (server, storage device), also called node, to participate in the fabric topology or to participate in the point-to-point topology as a partner.
- **F-Port** – knows how it can pass a frame that an N-Port sends to it through the Fibre Channel network on to the desired end device.
- **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.
- **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.

FC-0: разъемы, кабели и кодировка

- Bandwidth 100 Mbps upto 20 Gbps in one direction
- Fibre Channel therefore transmits the bits serially.
- Single bit error BER = 10^12 (for a 100 Mbit/s connection under full load a bit error may occur only every 16.6 minutes)
- Fiber-optic cables are more expensive than copper cables, but they have some advantages:
  - Greater distances possible than with copper cable;
  - Insensitivity to electromagnetic interference;
  - No electromagnetic radiation;
  - No electrical connection between the devices;
  - No danger of ‘cross-talking’;
  - Greater transmission rates possible than with copper cable

FC-1: кодировка, упорядоченные наборы, управление линией

- 8b/10b кодирование
- Transmission words
  - Data word: SOF, 4 bytes, EOF
  - Ordered set: EOF, K28.5, SOF
- Управление линией

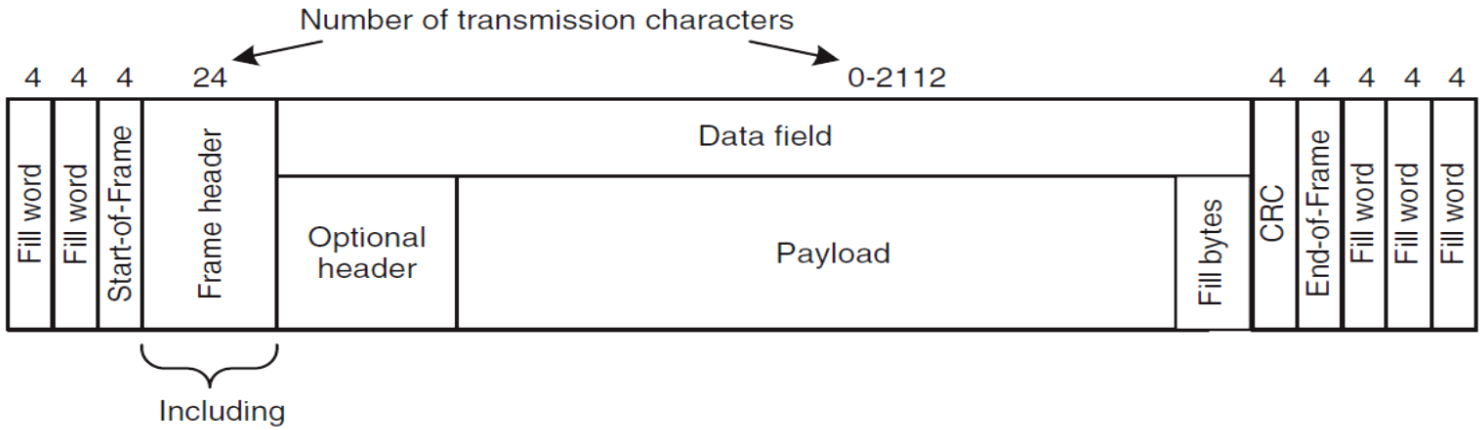


Асинхронная линия последовательной передачи

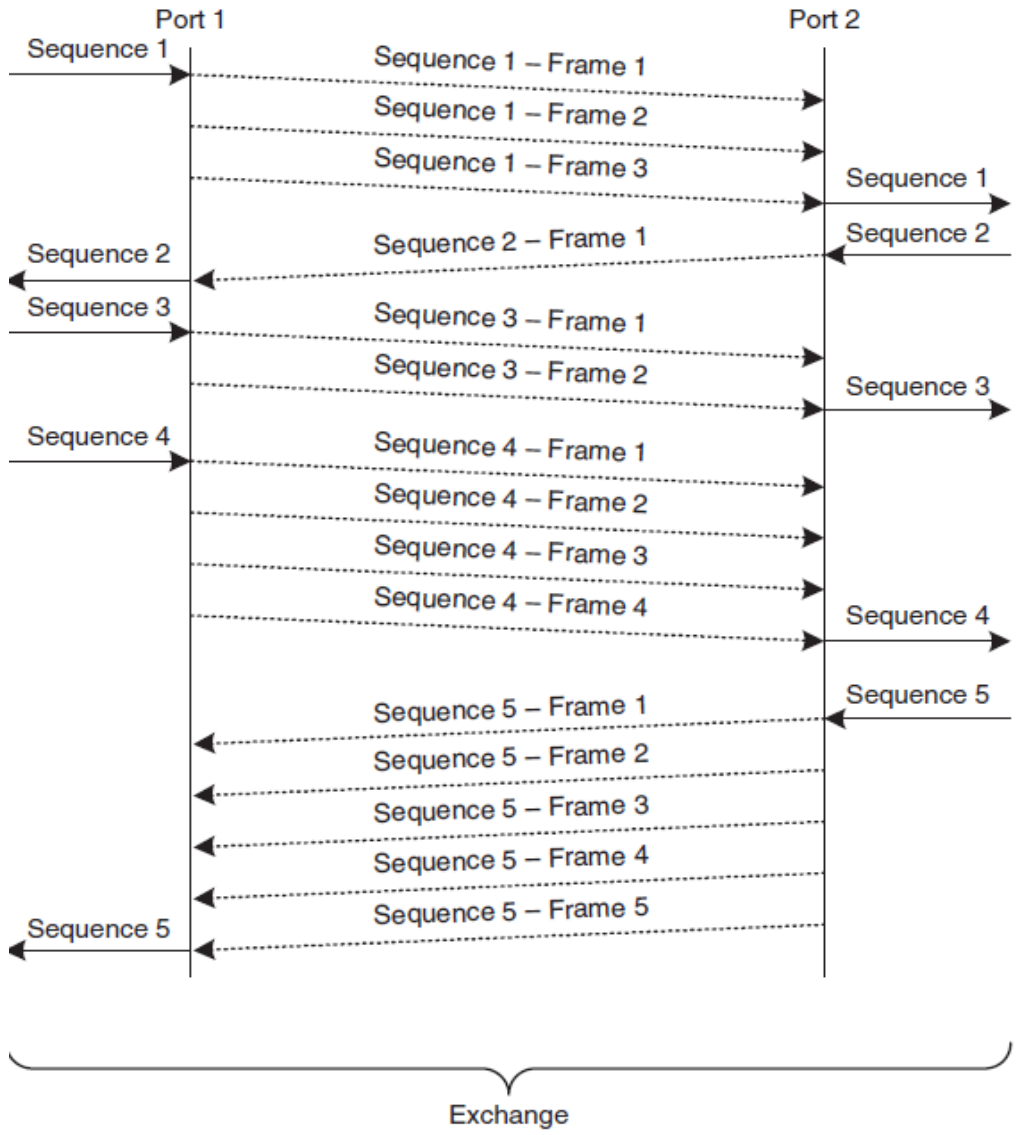
Т а б л и ц а А.1 –Специальные коды

Обозначения специальных кодов	TP –	TP+
	abcdei fghj	abcdei fghj
K28.0	001111 0100	110000 1011
K28.1	001111 1001	110000 0110
K28.2	001111 0101	110000 1010
K28.3	001111 0011	110000 1100
K28.4	001111 0010	110000 1101
K28.5	001111 1010	110000 0101
K28.6	001111 1000	110000 1001
K23.7	111010 1000	000101 0111
K27.7	110110 1000	001001 0111
K29.7	101110 1000	010001 0111
K30.7	011110 1000	100001 0111

FC-2: data transfer

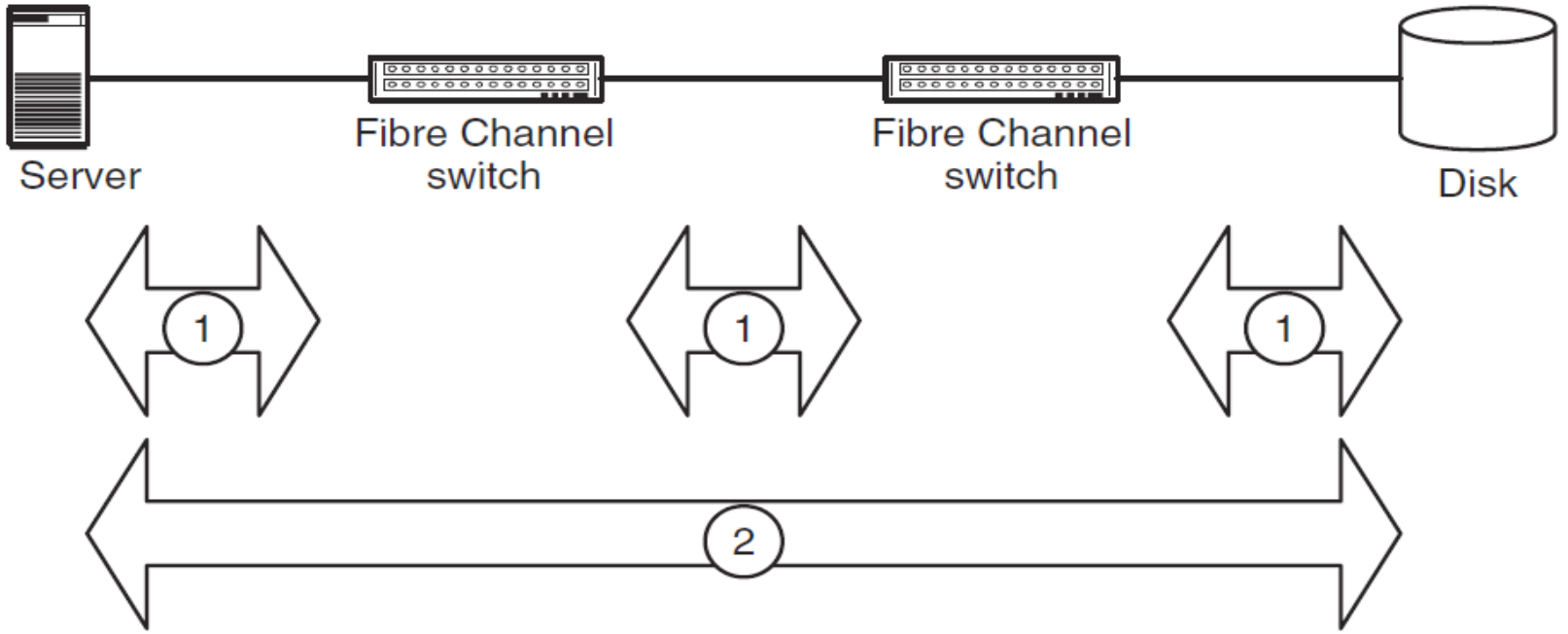


- Frame Destination Address (D_ID)
- Frame Source Address (S_ID)
- Sequence ID
- Number of the frame within the sequence
- Exchange ID



FC-2: Flow control

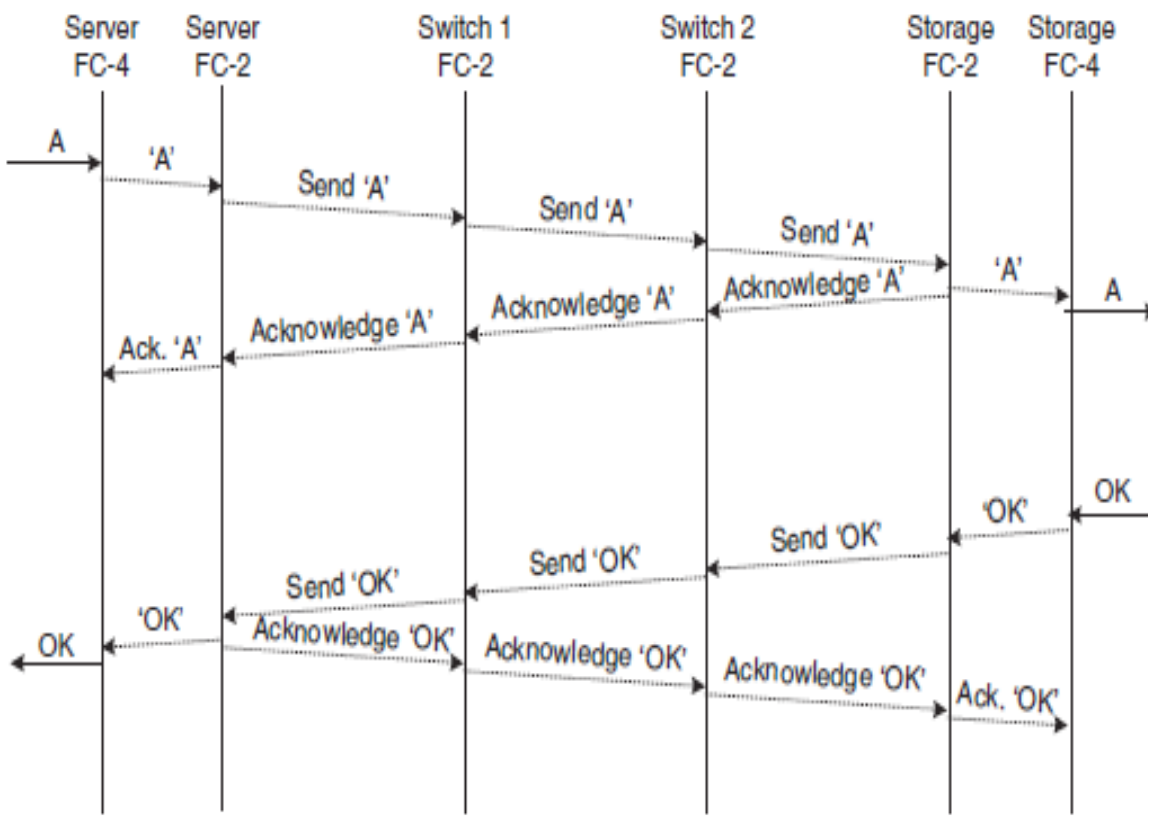
- Credential scheem
- E2E flow control (2)
- Link flow control (1)

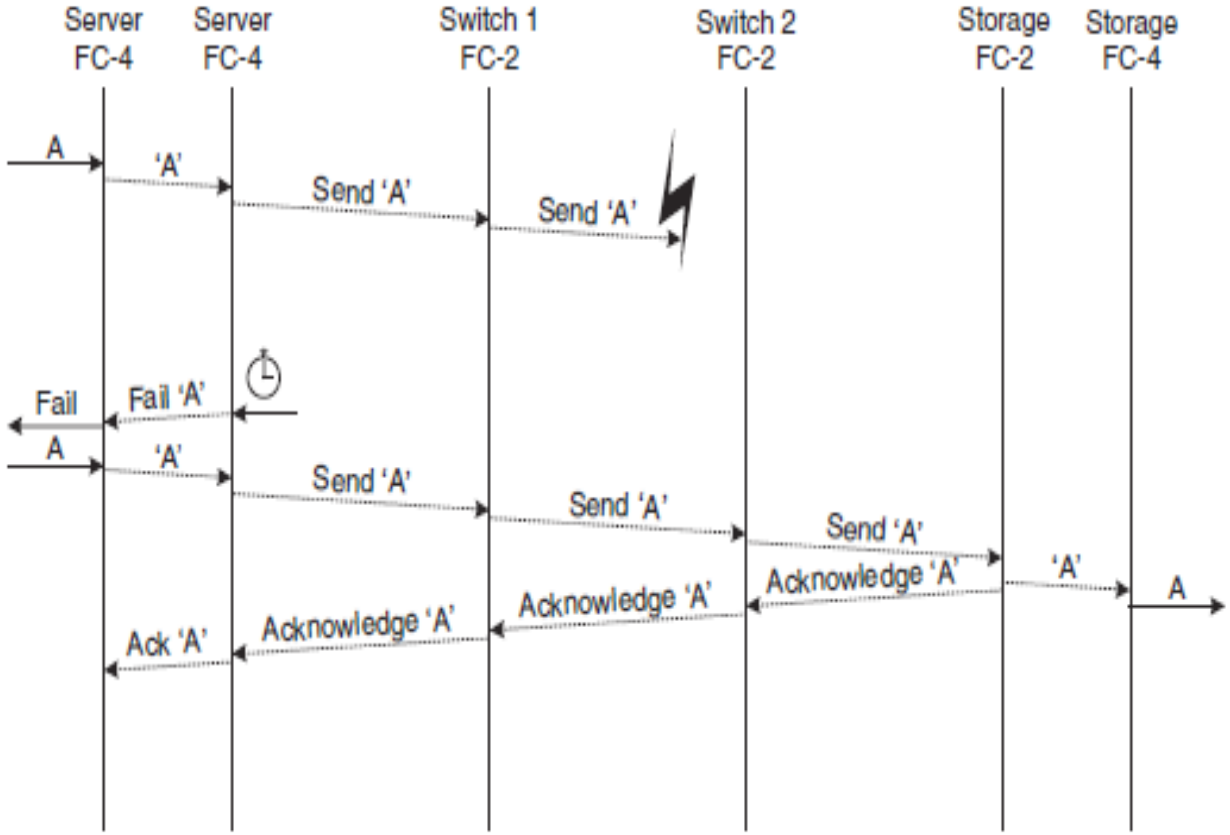


FC-2: Service classes

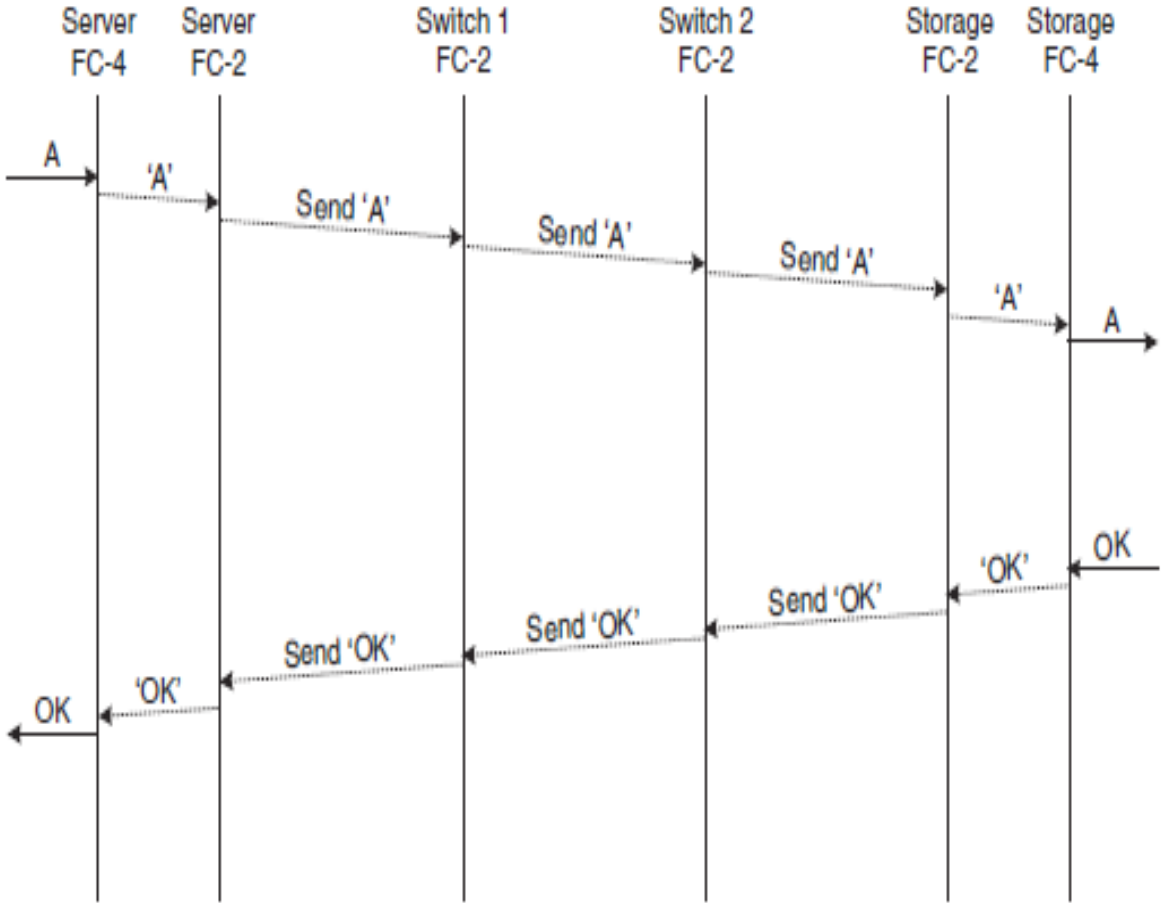
Class 1	A point-to-point connection (end-to-end) between ports of type n_port through circuit switching.
Class 2	Connectionless packet switched, which guarantees delivery of data. A port can communicate simultaneously with any number of ports of type n_port in duplex mode. The order of frame delivery 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.

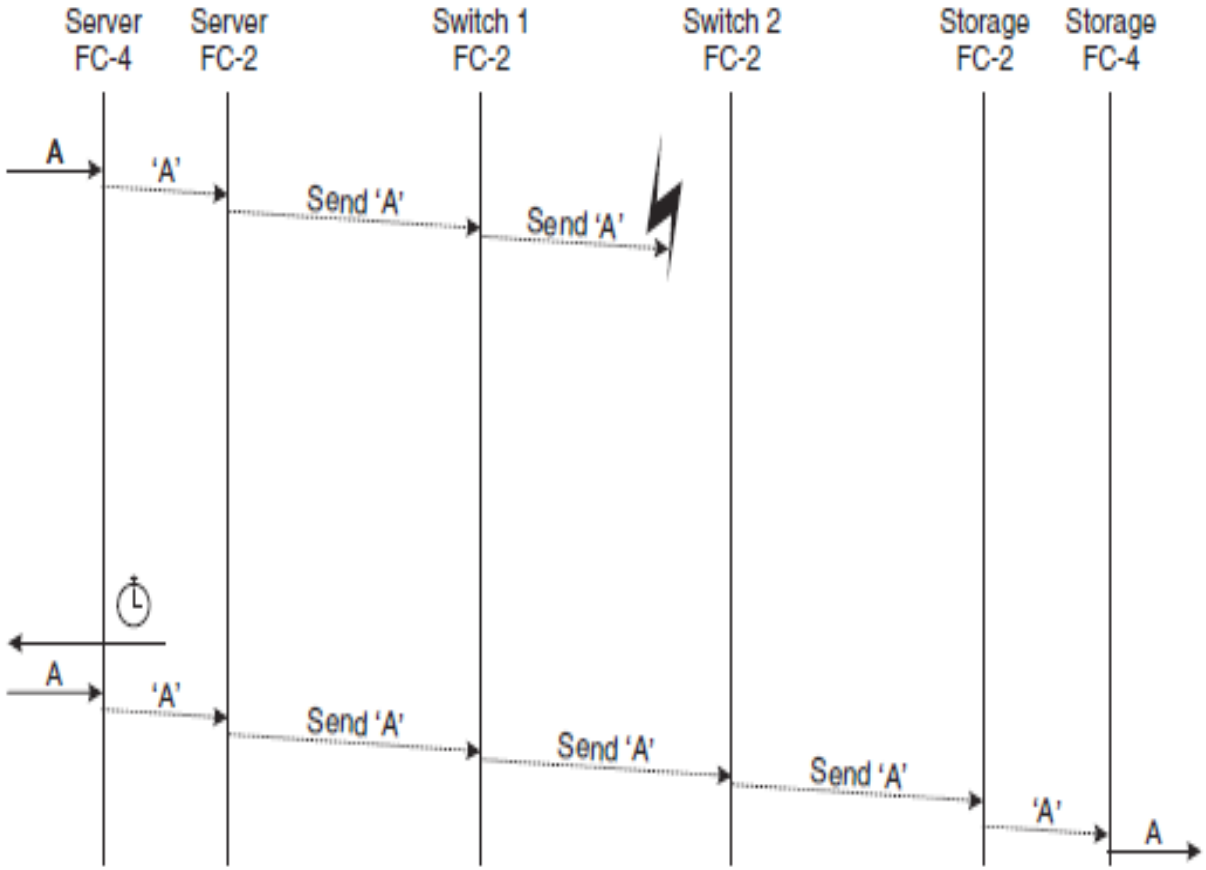
FC-2 class 2


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

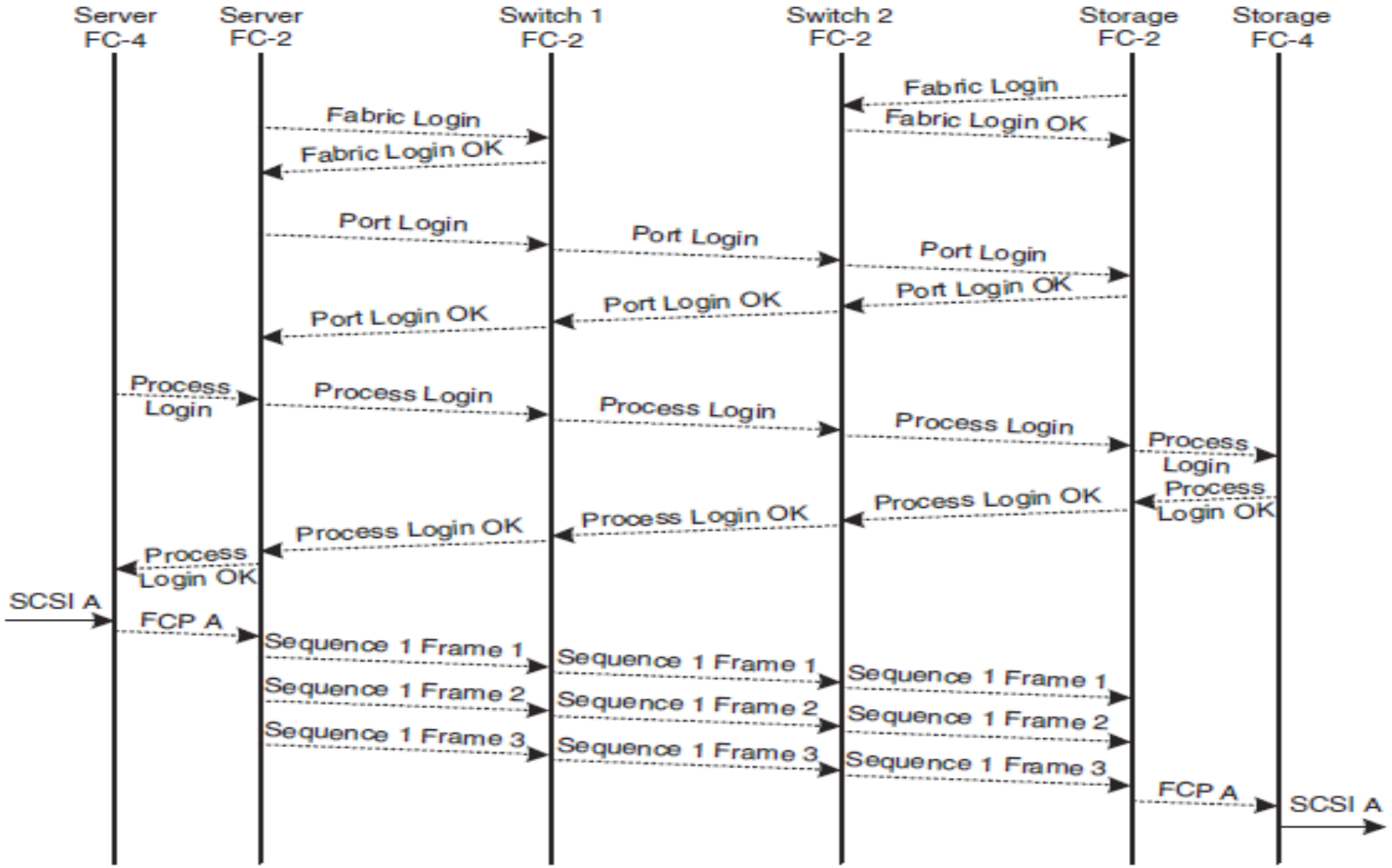

Server FC-4
Server FC-2
Switch 1 FC-2
Switch 2 FC-2
Storage FC-2
Storage FC-4
A
'A'
Send 'A'
Send 'A'
Send 'A'
'A'
A
OK
'OK'
Send 'OK'
Send 'OK'
Send 'OK'
'OK'
OK


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

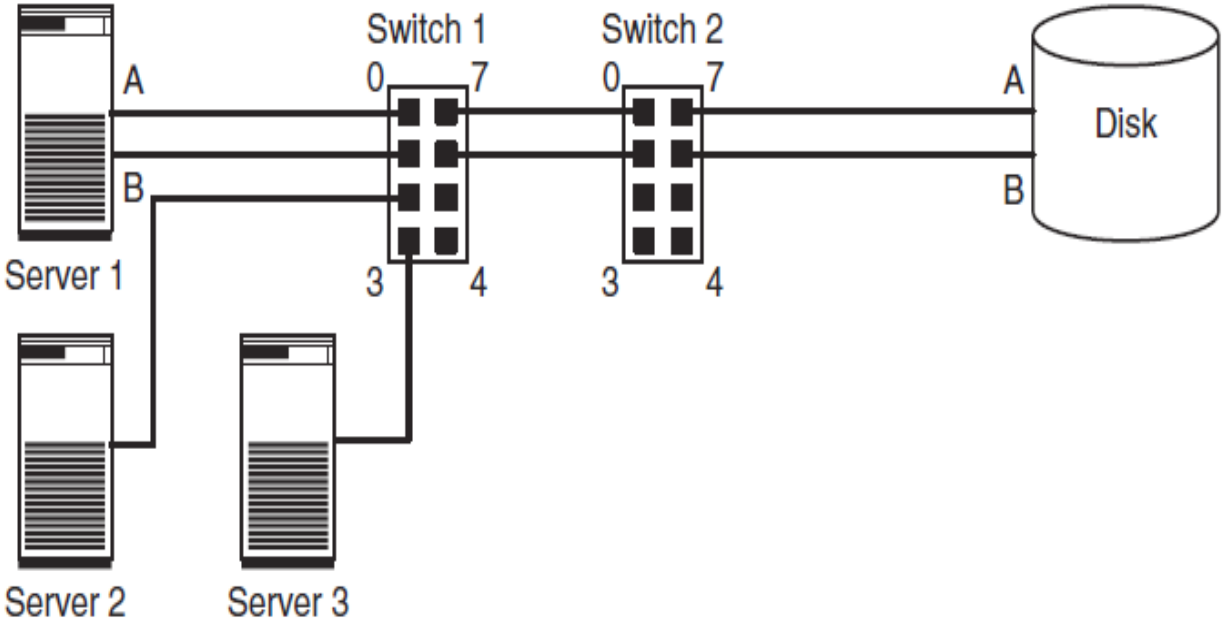
- Striping manages several paths between multiport end devices. Striping could distribute the frames of an exchange over several ports and thus increase the throughput between the two devices.
- Multipathing combines several paths between two multiport end devices to form a logical path group. Failure 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.

Службы линии: идентификация и адресация



Адресация

- Имена и адреса в FC
- У всех устройств FC сети есть 64 бит. имена
- WWN vs FCN
- WWN: WWPN. WWNN
- FLOG – 24 bit port address
- S_ID vs D_ID
- КсА 8 bit AL_PA (Arbitrated Loop Physical Address)



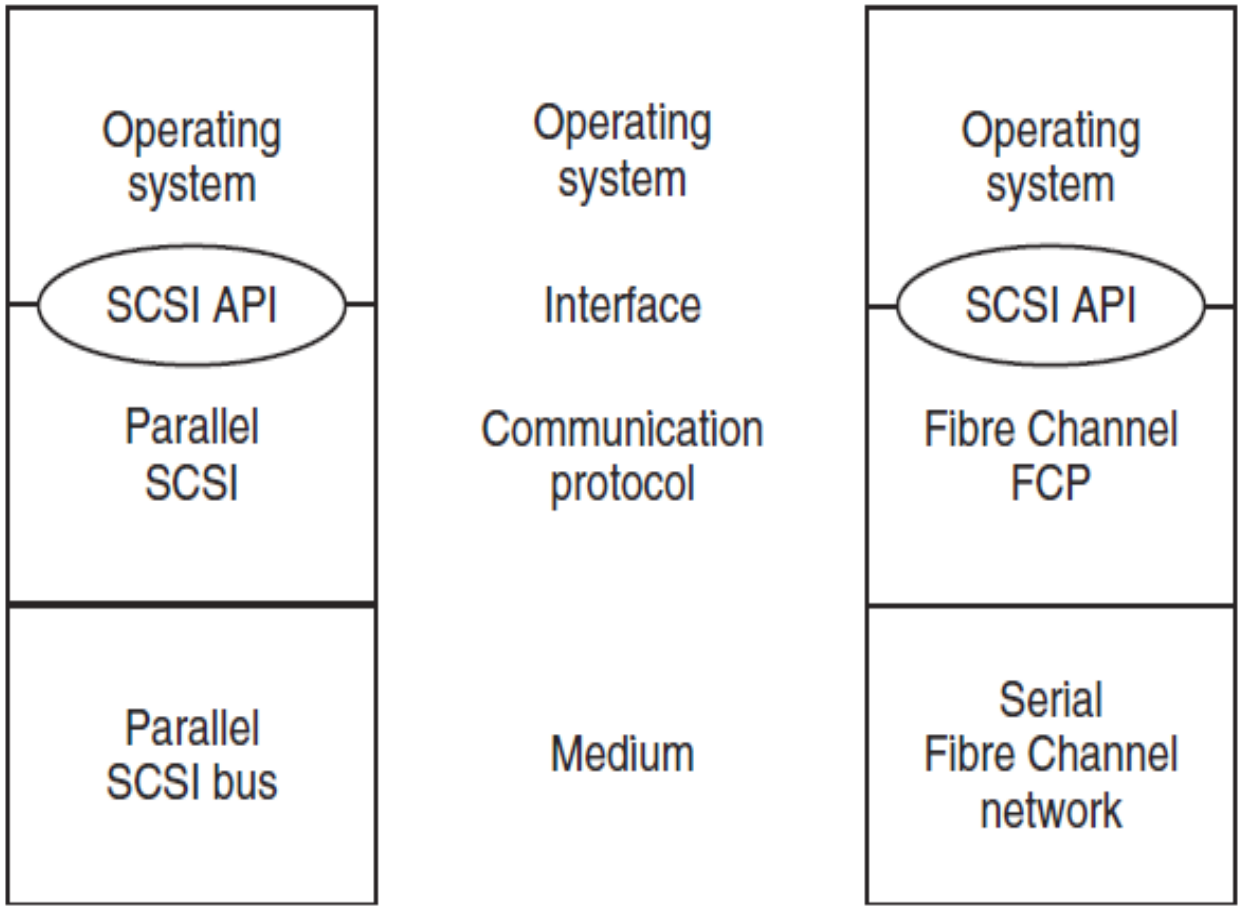
Port_ID	WWPN	WWNN	Device
010000	20000003 EAFE2C31	2100000C EAFE2C31	Server 1, Port A
010100	20000003 C10E8CC2	2100000C EAFE2C31	Server 1, Port B
010200	10000007 FE667122	10000007 FE667122	Server 2
010300	20000003 3CCD4431	2100000A EA331231	Server 3
020600	20000003 EAFE4C31	50000003 214CC4EF	Disk, Port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

Сервисы коммутационной среды

- Сервисы коммутационной среды нужны для управления инфраструктурой и потоками в FC
- Все сервисы реализуют определённые сервера, которые имеют строго определенные адреса.
- FLOG сервер отвечает за обработку всех входящих fabric login request
- За всеми изменениями в FC сети следит fabric controller
- Name server – отвечает за БД все имен N_Port’ов

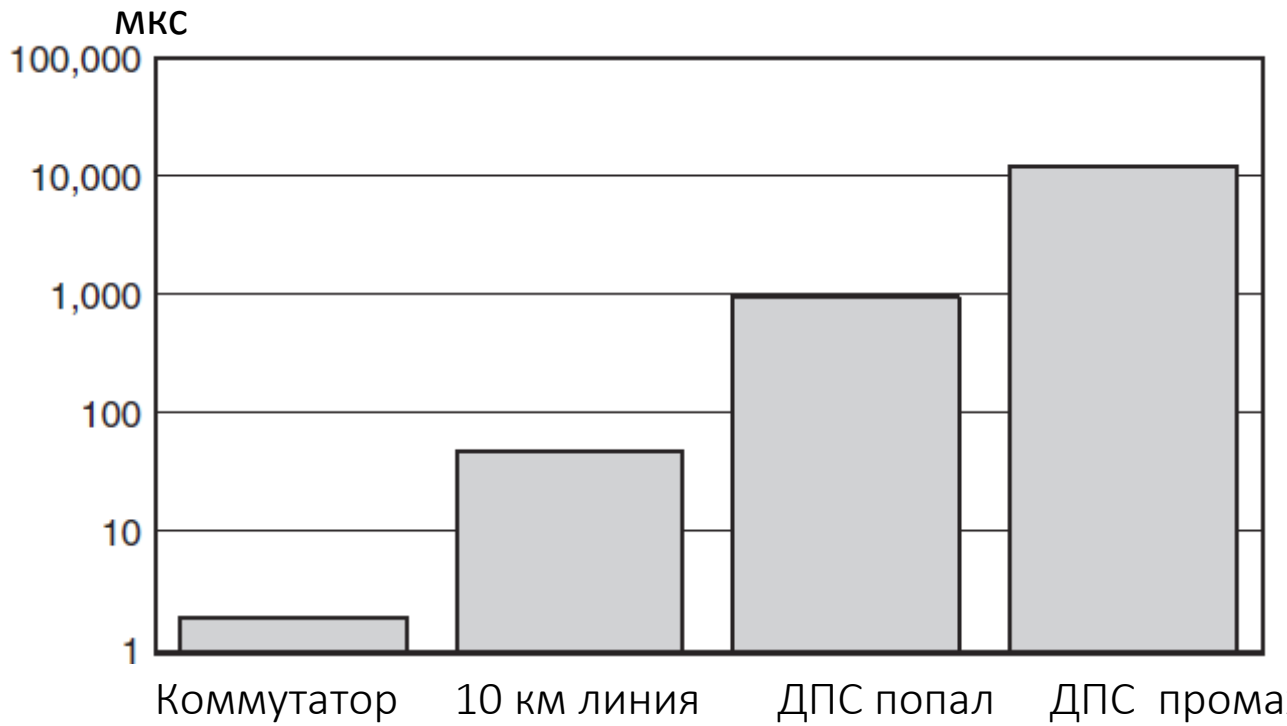
Address	Description
0×FF FF FF	Broadcast addresses
0×FF FF FE	Fabric Login Server
0×FF FF FD	Fabric Controller
0×FF FF FC	Name Server
0×FF FF FB	Time Server
0×FF FF FA	Management Server
0×FF FF F9	Quality of Service Facilitator
0×FF FF F8	Alias Server
0×FF FF F7	Security Key Distribution Server
0×FF 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

FC-4: ULP


Operating system
SCSI API
Parallel SCSI
Parallel SCSI bus
Operating system
Interface
Communication protocol
Medium
Operating system
SCSI API
Fibre Channel FCP
Serial Fibre Channel network

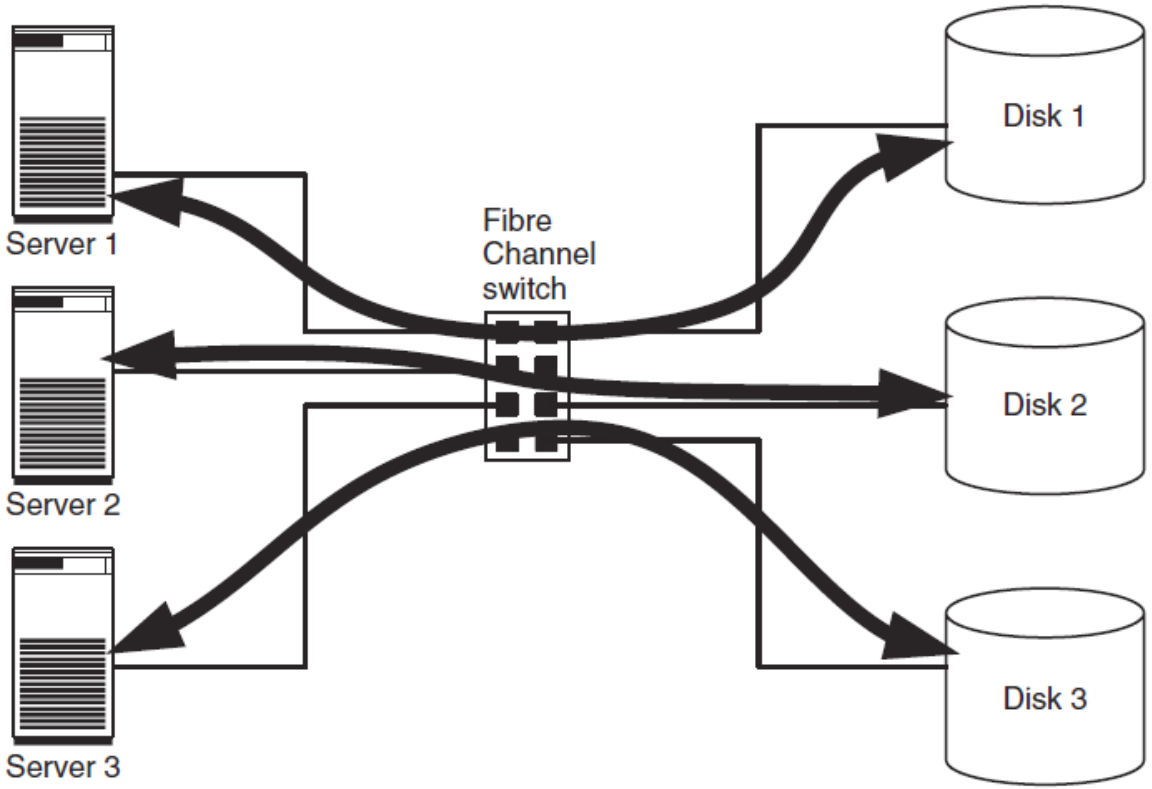
Fibre Channel SAN

- P2P – FC до 10 км, SCSI не более 25 метров
- SCSI – медь, FC - оптика

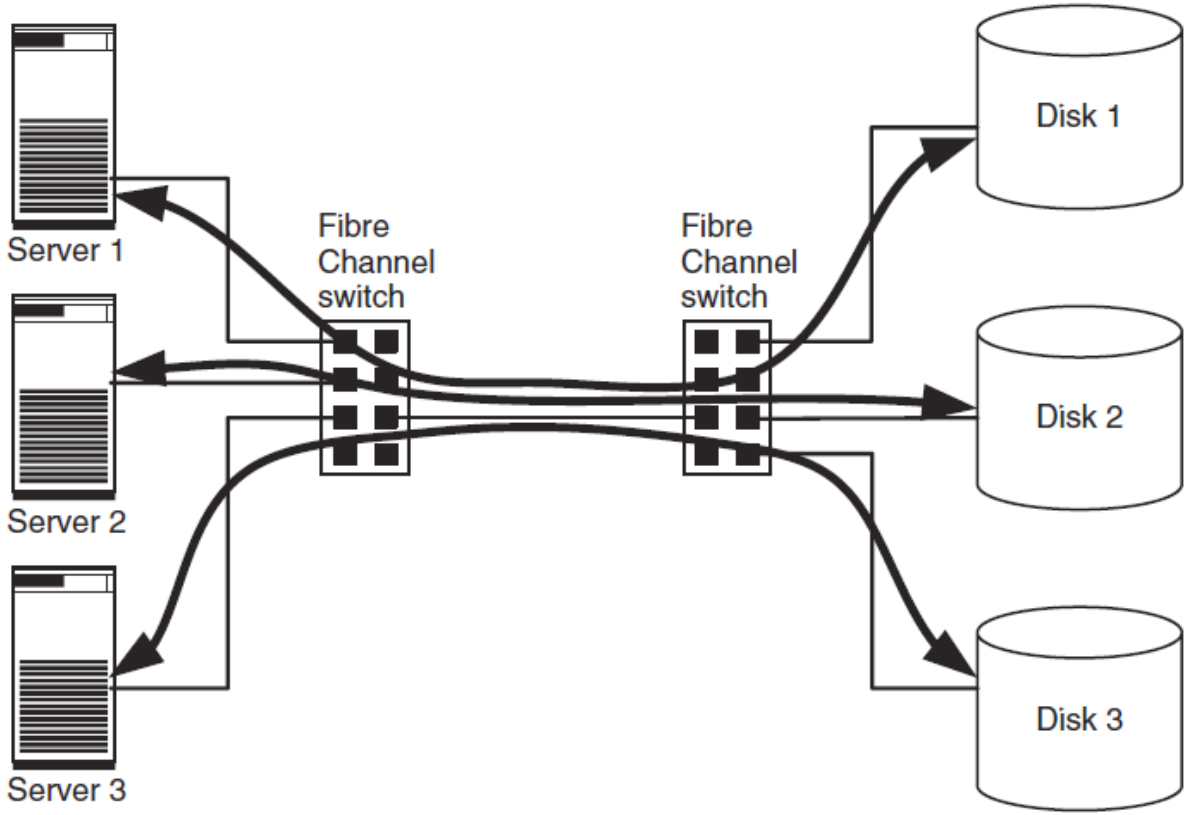


Задержки на различных компонентах FC сети

Fabric topolides

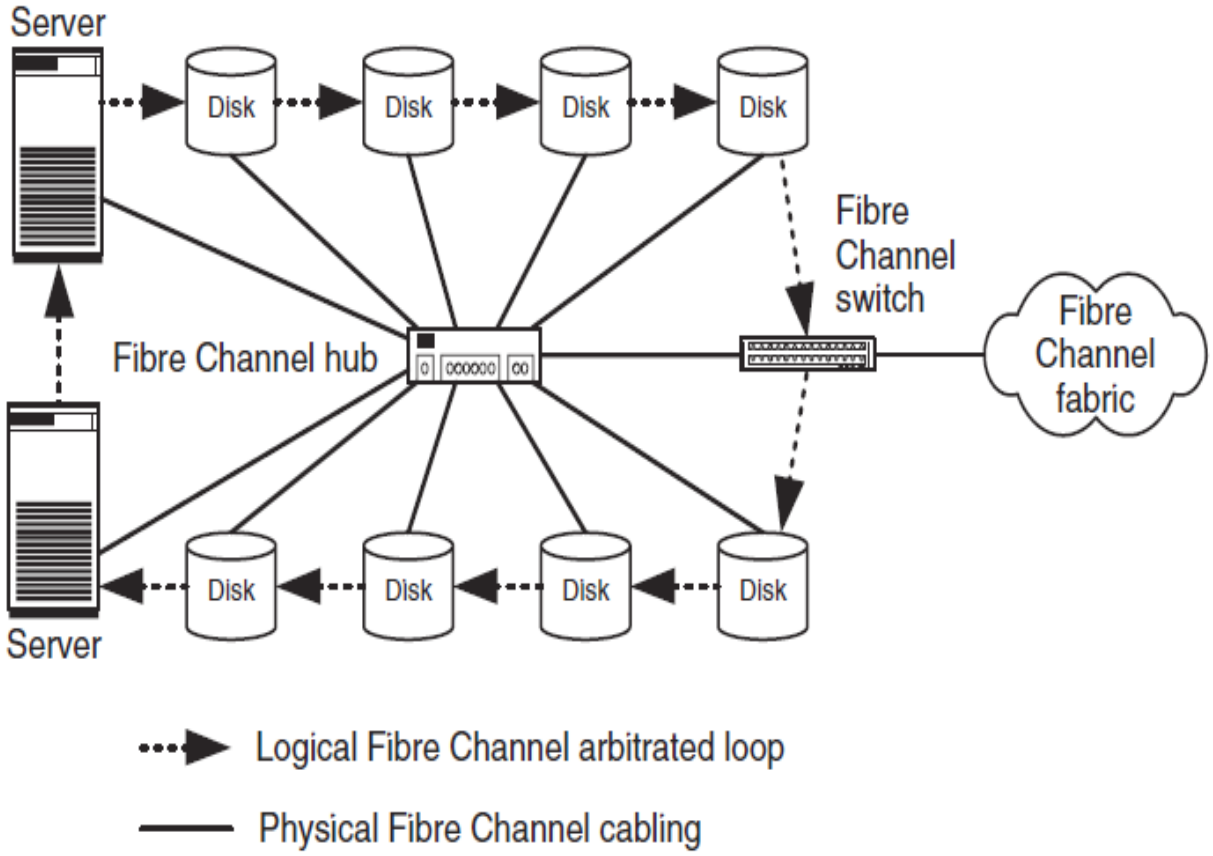


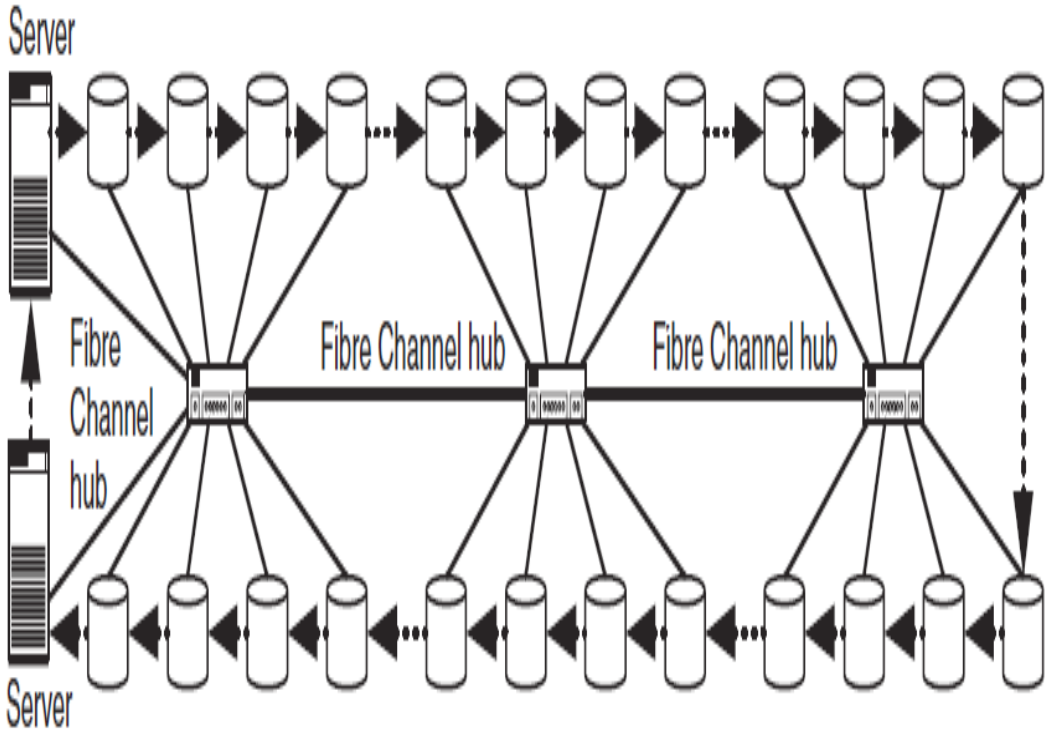
several connections at full bandwidth.



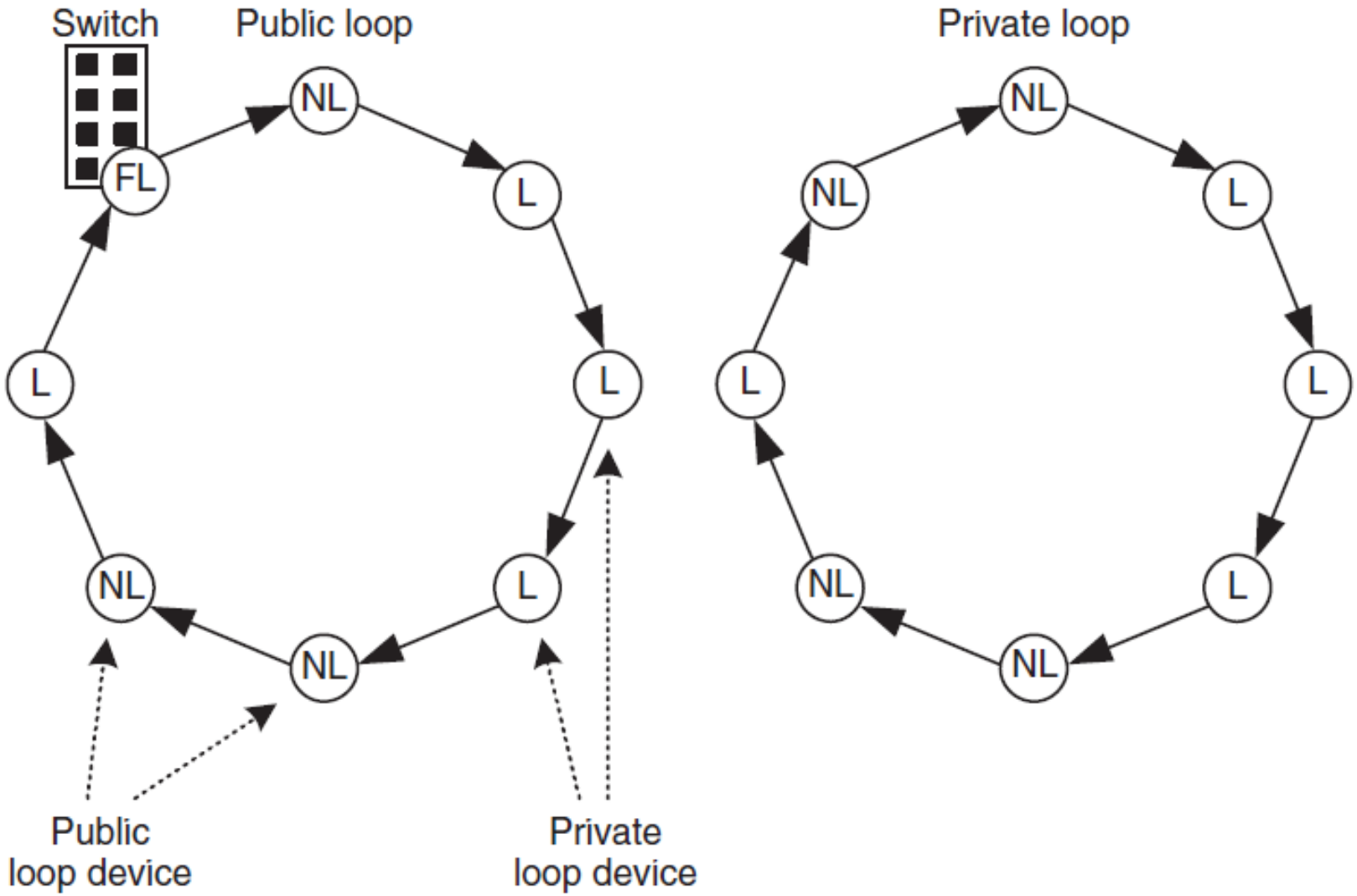
Inter-switch links (ISLs).

Arbitrated loop topologes

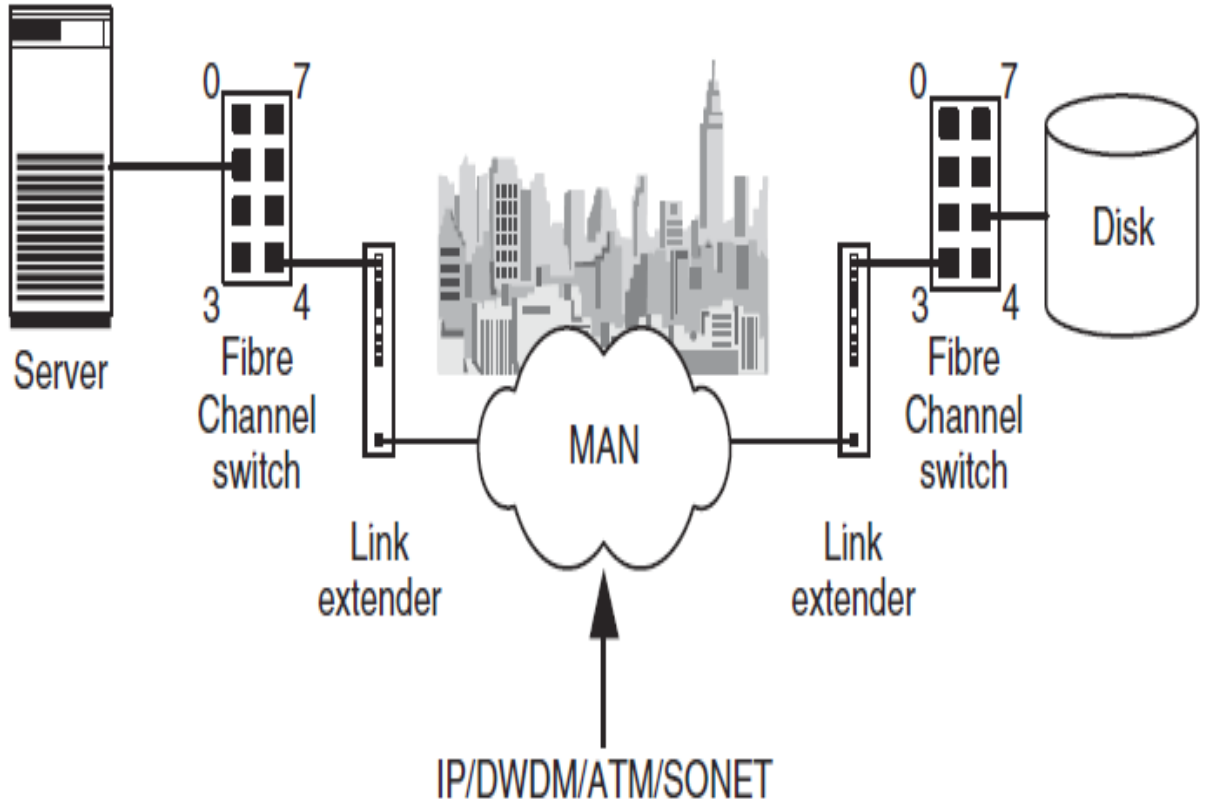


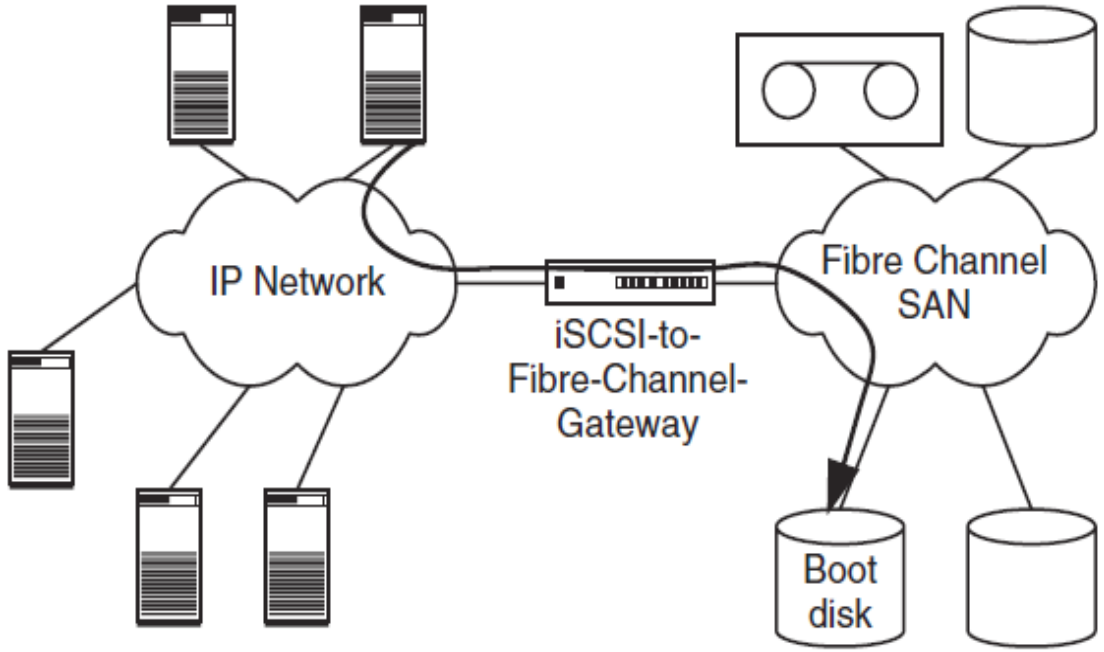


Private vs Public loops

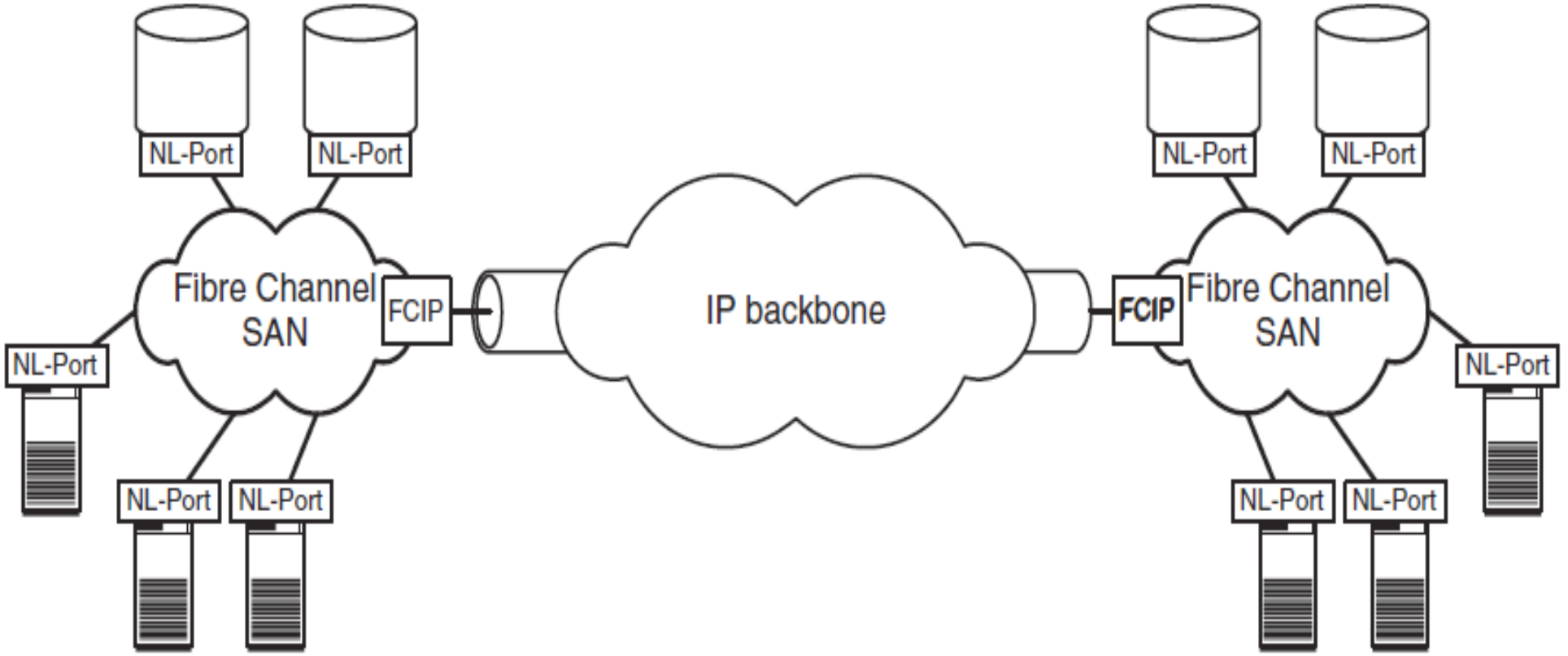


Metro SAN and IP_FC SAN





Соединение FC_SAN через TCP/IP магистраль



ВОРОСЫ?

http://arccn.ru/

smel@arccn.ru

+7 (495) 240-50-63

@ArccnNews