Сетевые Хранилища Данных
–
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

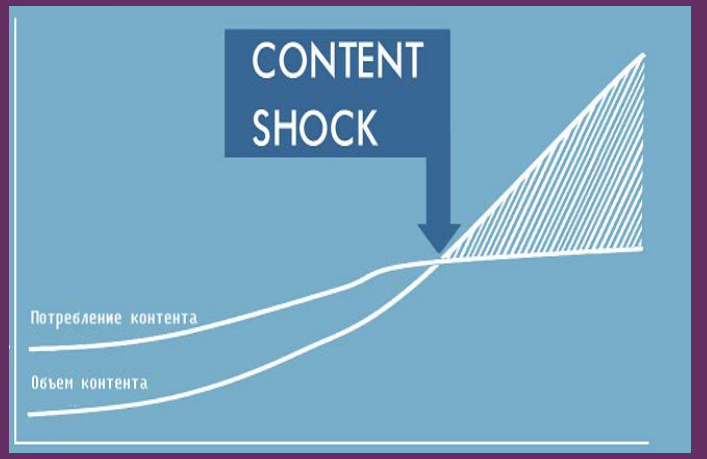
Содержание

- 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

Основные тренды:

- Глобальный годовой IP трафик: 2.3 ZB (зеттабайт = 10^21) в год к 2020 году.
- Объем трафика с беспроводных и мобильных устройств составит **две трети общего IP трафика к 2020 году и** превысит объем трафика со стационарных компьютеров к 2020 году.
- Трафик **Доминировать будет трафик между ЦОД**



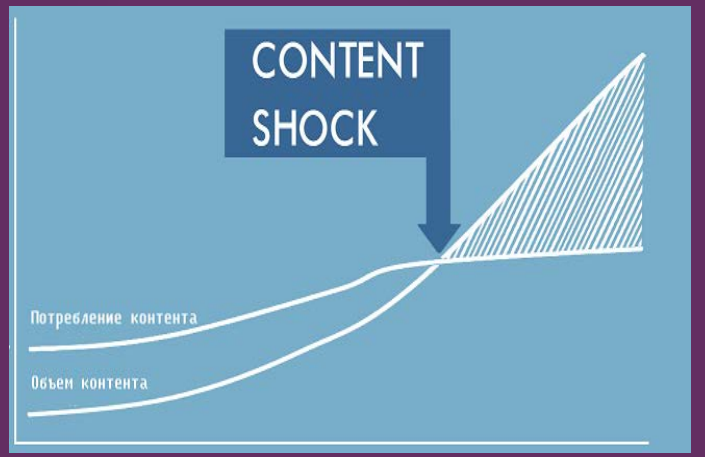
Особенности роста мобильного трафика:

- С 2015 по 2020 годы объем мобильного трафика возрастет в 8 раз и достигнет в 2020 г. показателя 30,6 ЭБ/мес (Эксабайт = 10^18).
- Мобильный трафик в этот период будет расти в три раза быстрее, чем трафик в фиксированных сетях.

Основные тренды роста трафика в сетях

Особенности роста игрового и видеотрафика:

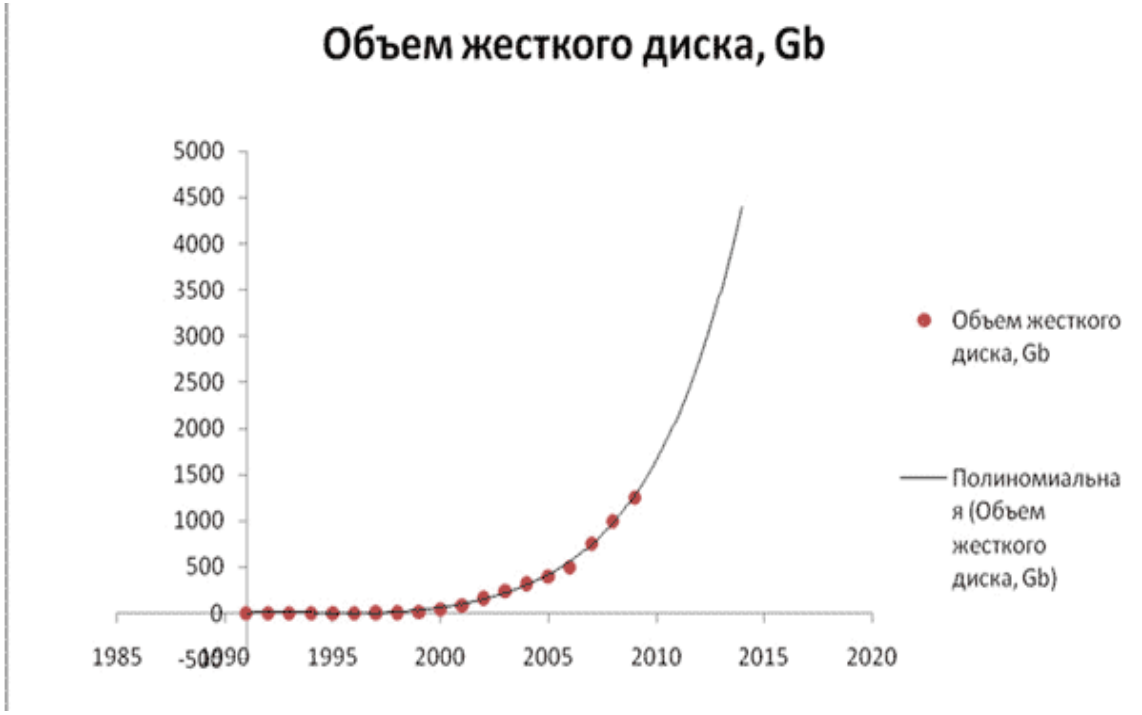
- В 2020 году для просмотр всего видеоконтента, который будет проходить через глобальные IP сети каждый месяц, потребуется более 5 миллионов лет.
- Трафик виртуальной реальности вырос к 2015 году в 4 раза. К 2020 году он вырастет еще в 61 раз при среднегодовом темпе роста в 127%.
- За последний год объем трафика видеонаблюдения практически удвоился, а к 2020 г. вырастет десятикратно.
- Игровой интернет-трафик вырастет к 2020 году в 7 раз.
- Объем потребительского трафика видео по требованию к 2020 году вырастет почти в два раза.
- Трафик IPTV увеличился в 2015 году на 50 процентов. К 2020 году вырастет в 3,6 раза.



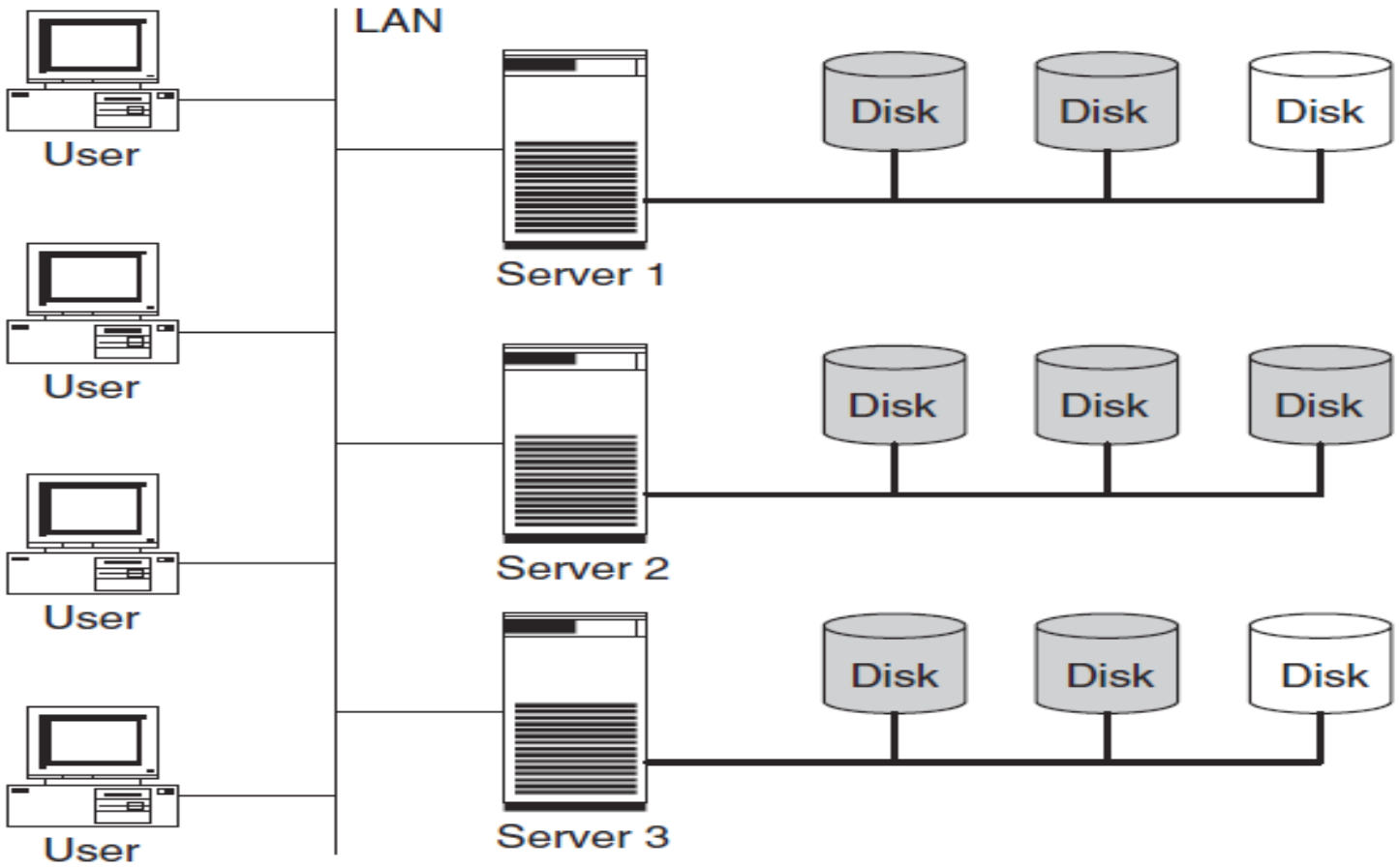
source: blogs.cisco.com

Тенденции систем хранения

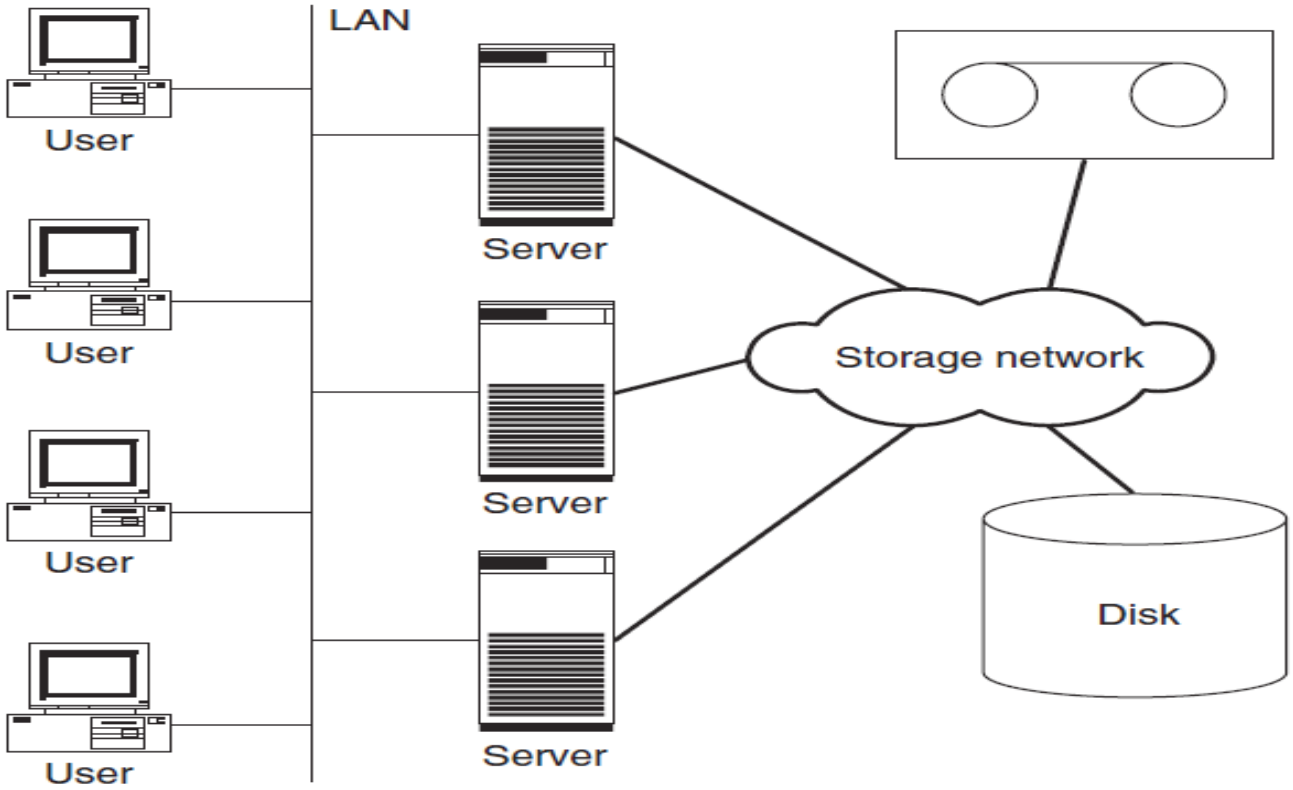
- Средняя стоимость ГБ для HDD



Server Centric Architecture

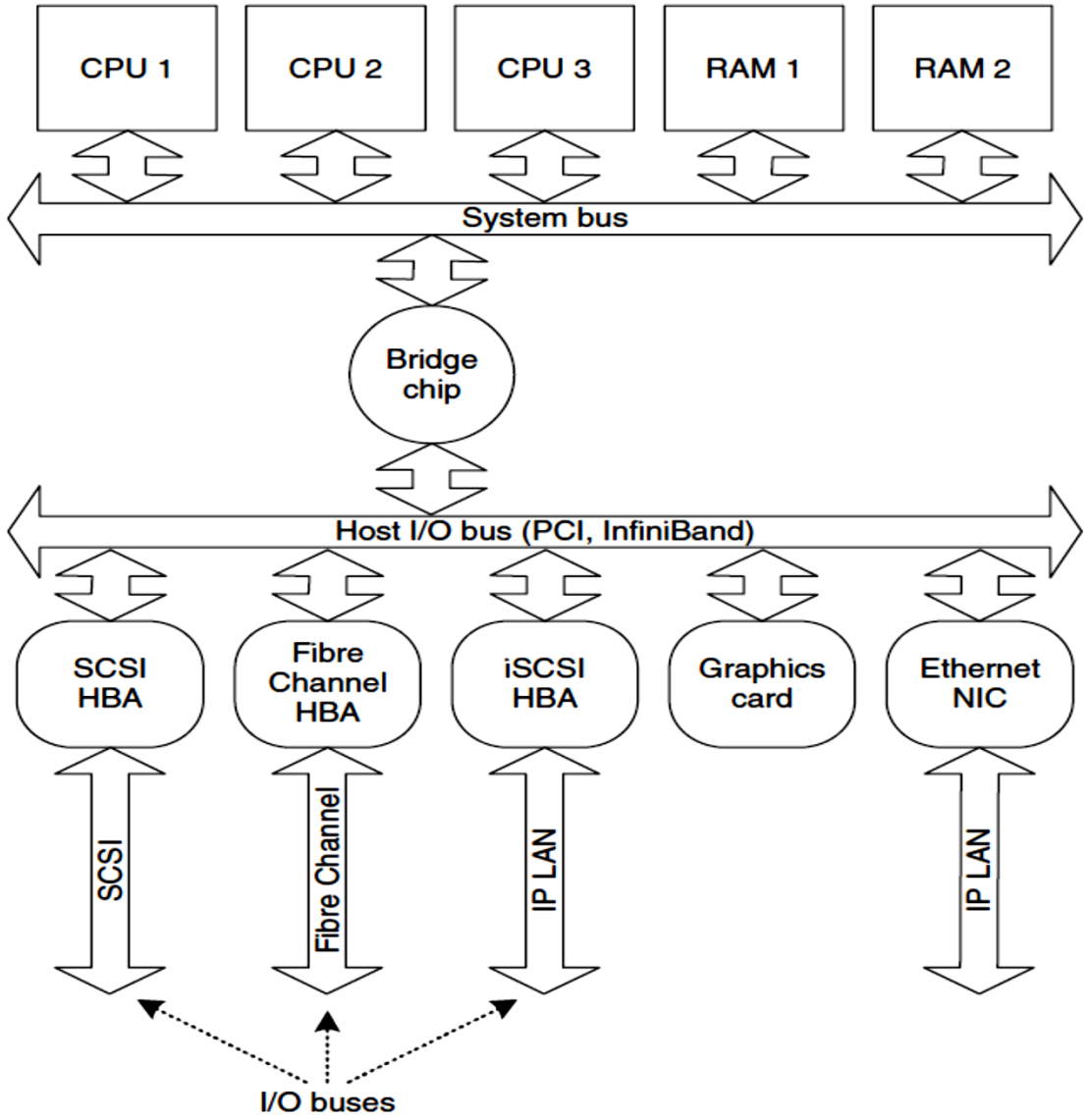


Storage Centric Architecture

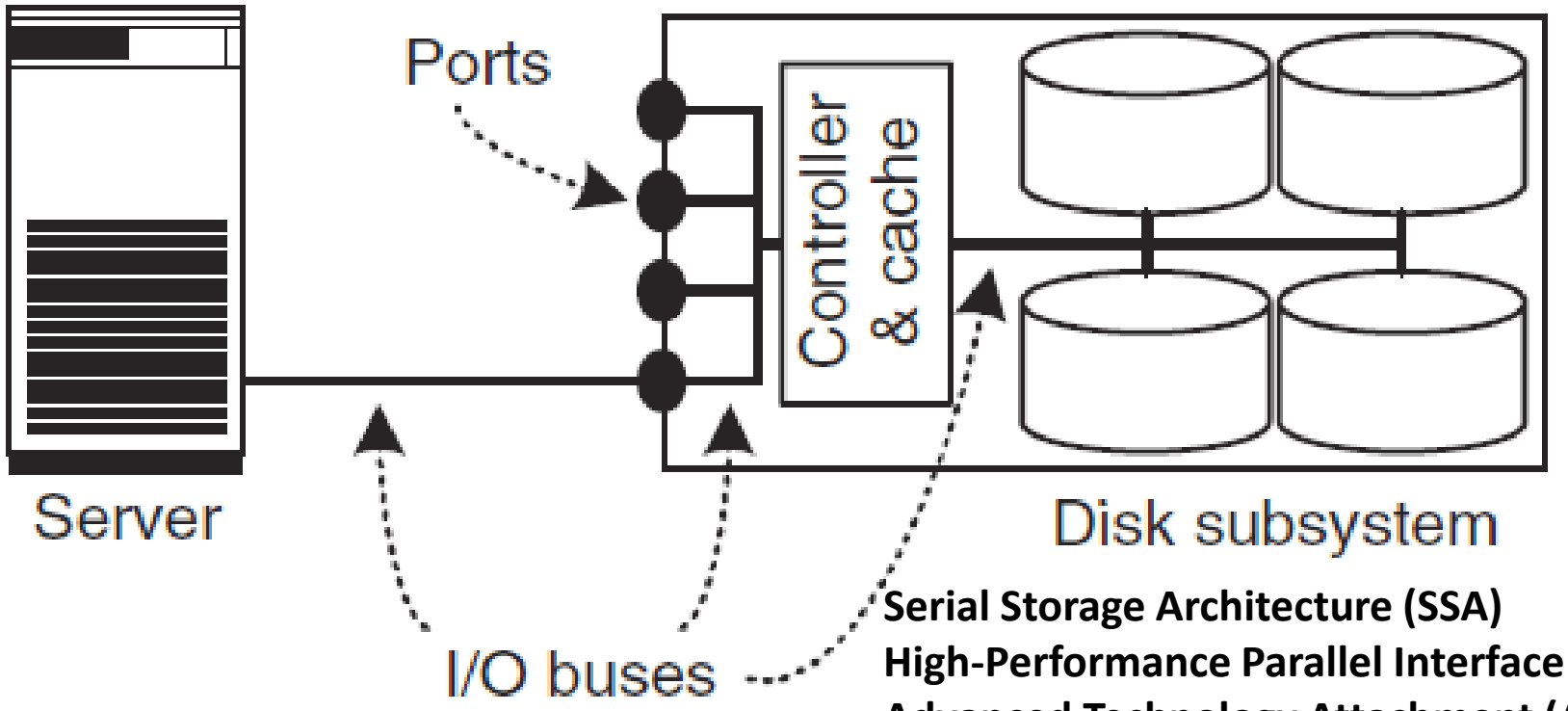


1. One server – one **Disk under loaded**
2. **Software upgrade**
3. **Data safety and consistency**
4. Automation **data balancing** between HD system in Data Storage Network

Тракт от CPU до СХД

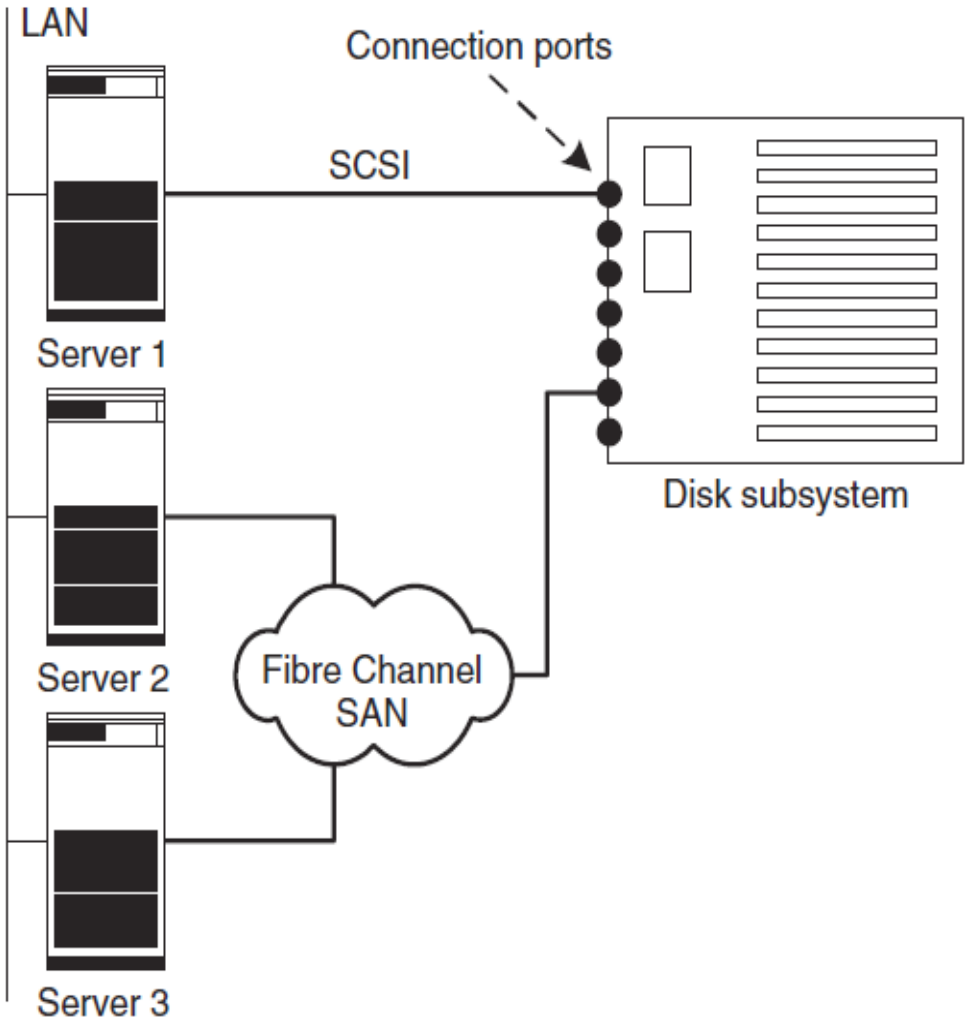


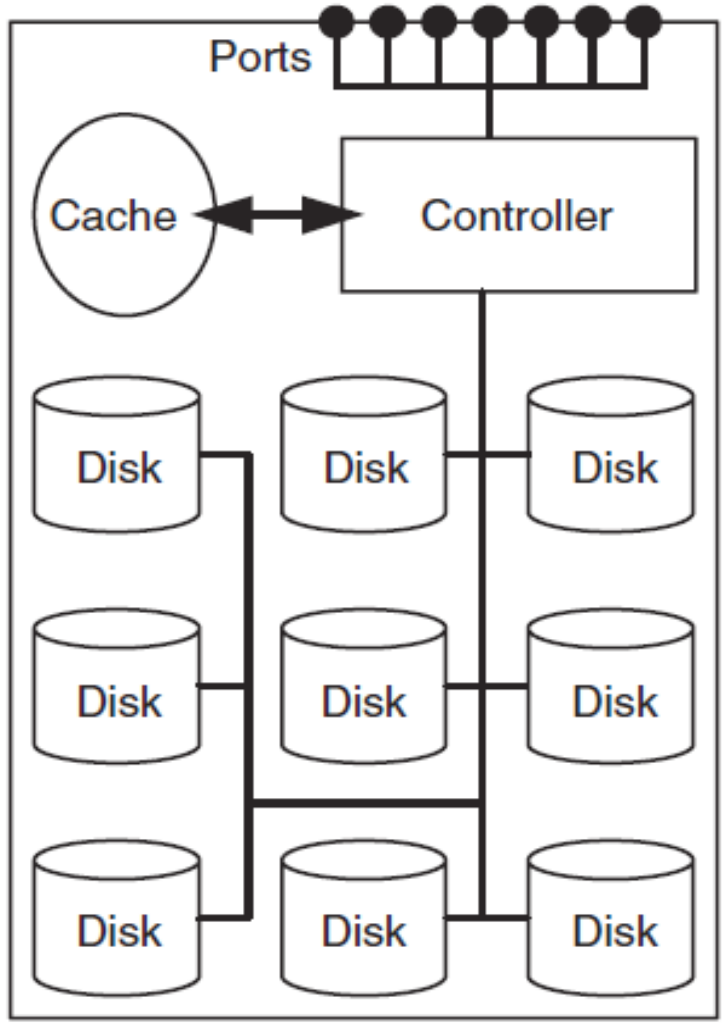
Тракт от CPU до СХД (2)



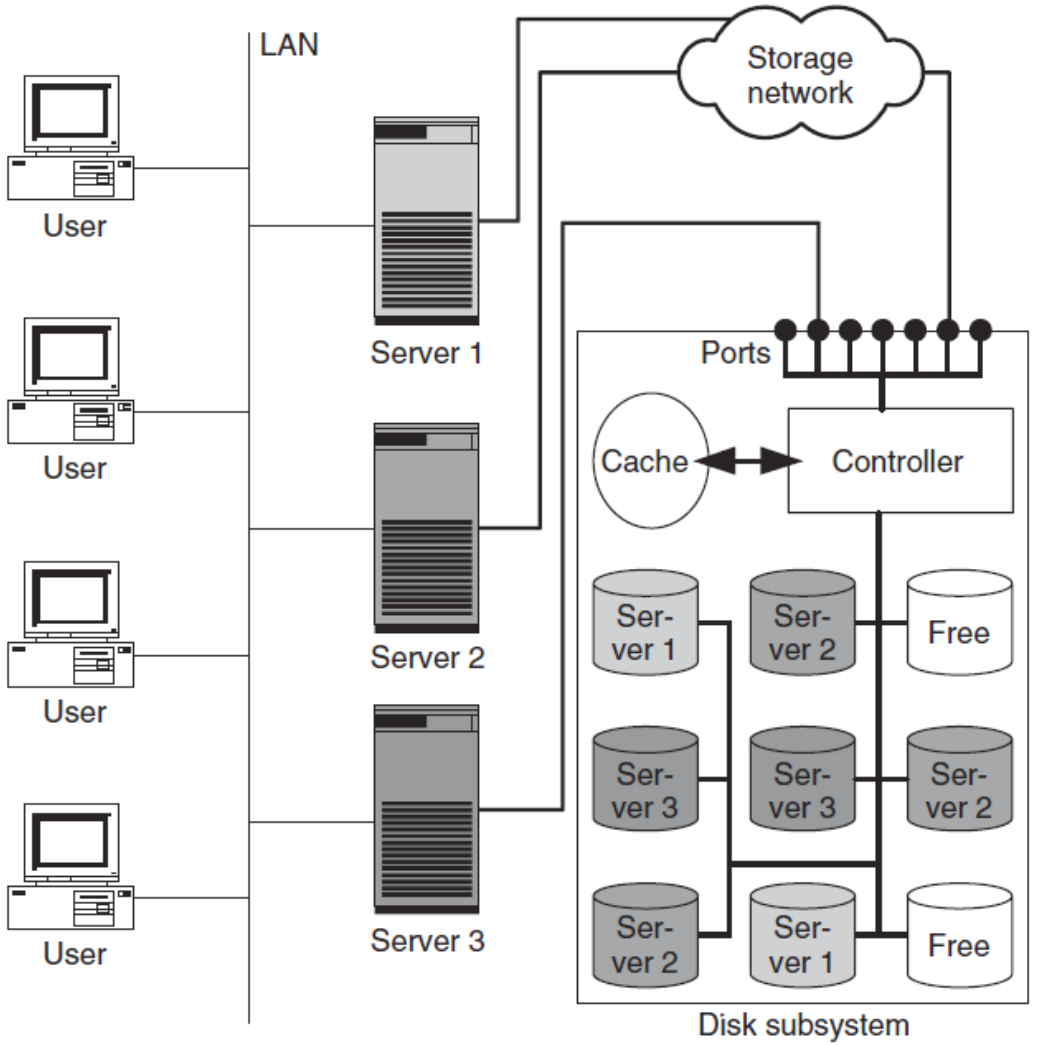
**Serial Storage Architecture (SSA)
High-Performance Parallel Interface (HIPPI),
Advanced Technology Attachment (ATA),
Integrated Drive Electronics (IDE),
Serial ATA (SATA), Serial
Attached SCSI (SAS)
Universal Serial Bus (USB).**

Архитектура Дисковой подсистемы (ДПС)

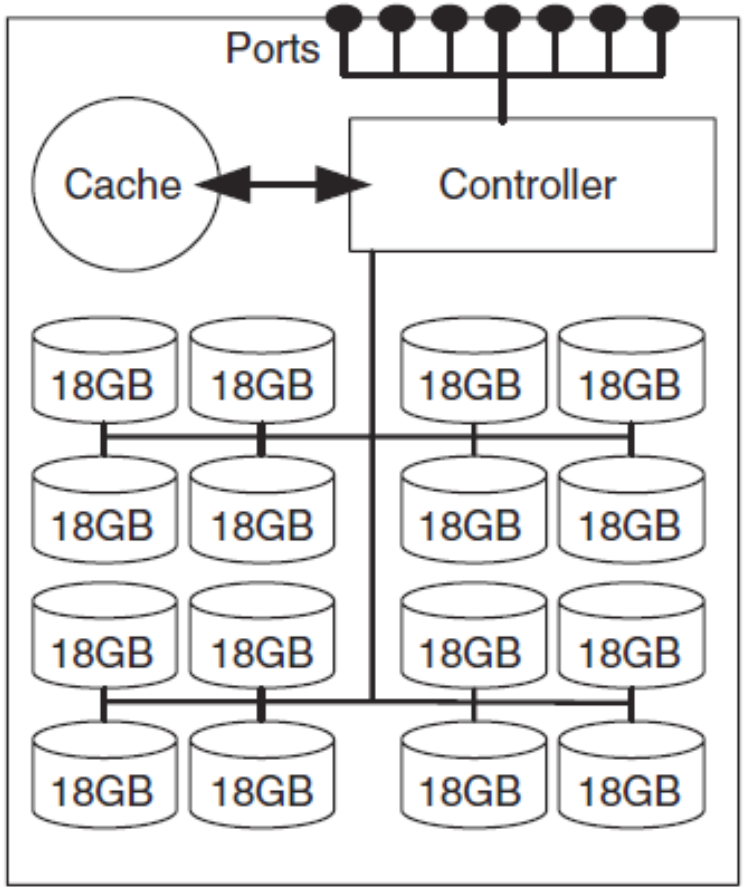




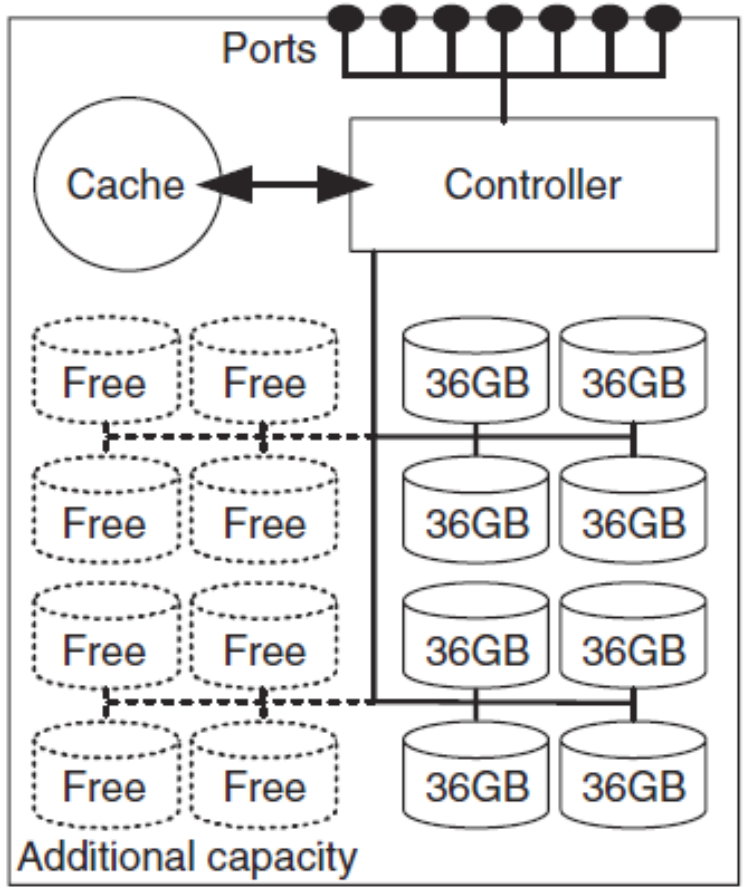
Пример использования ДПС



Внутренняя организация и емкость дисков

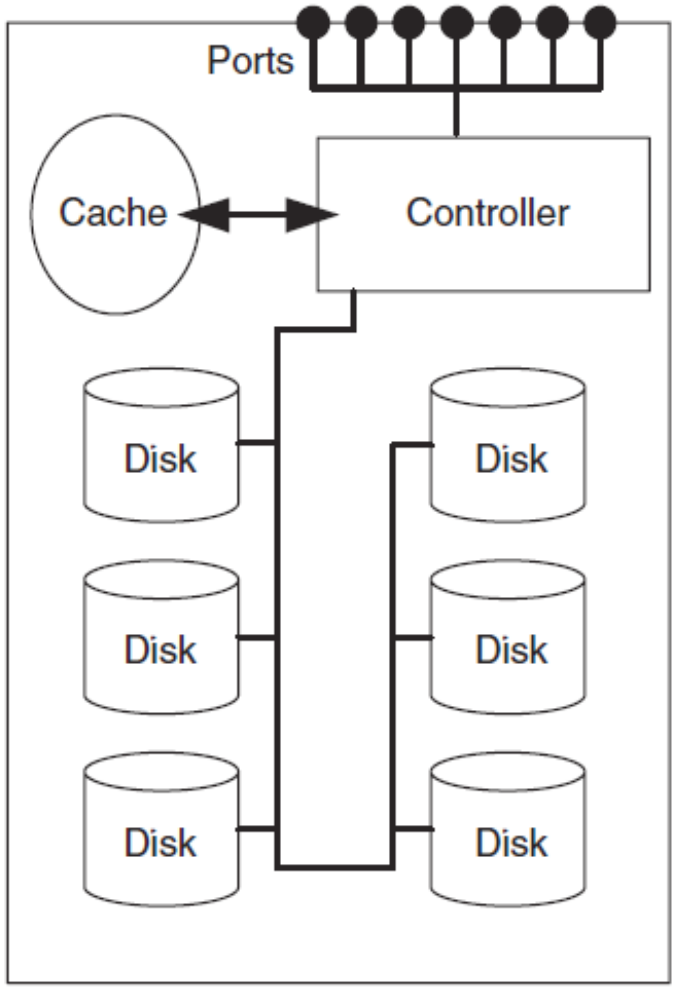


Маленькие диски

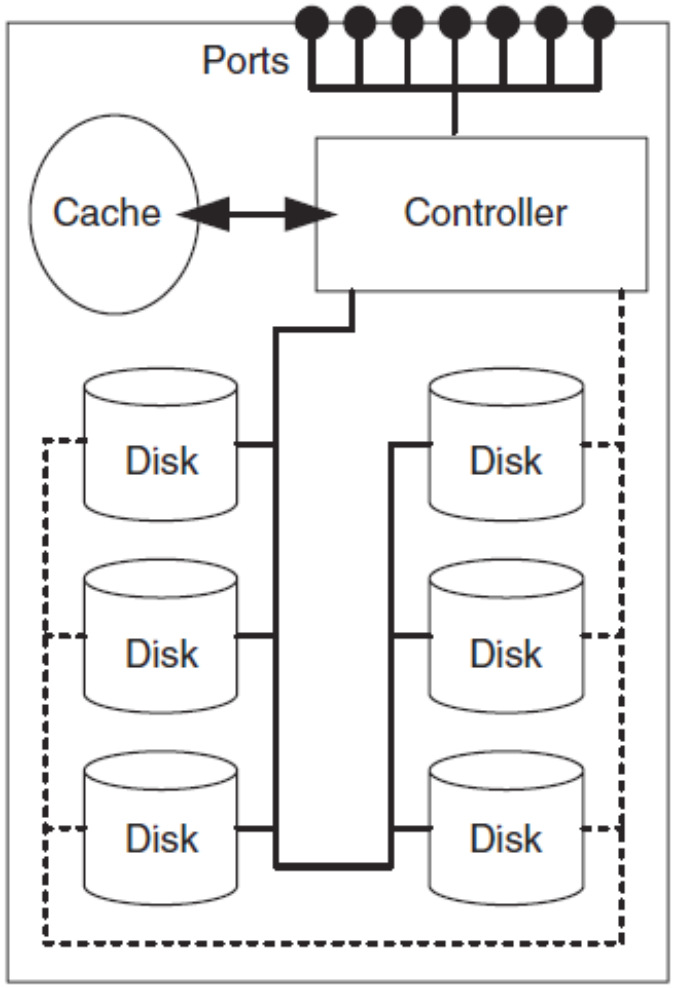


Большие диски

Внутренняя организация ДПС

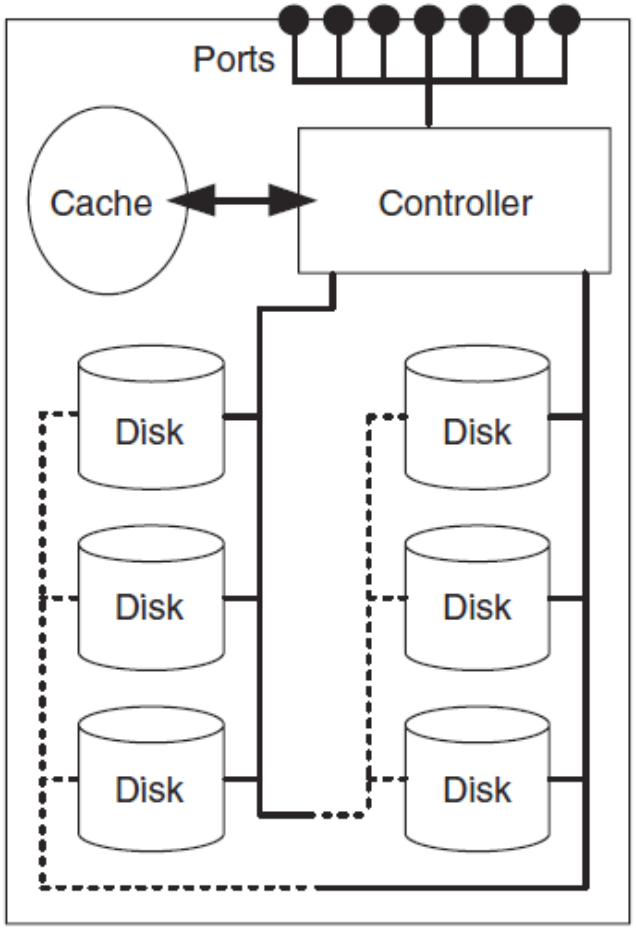


Одинарная

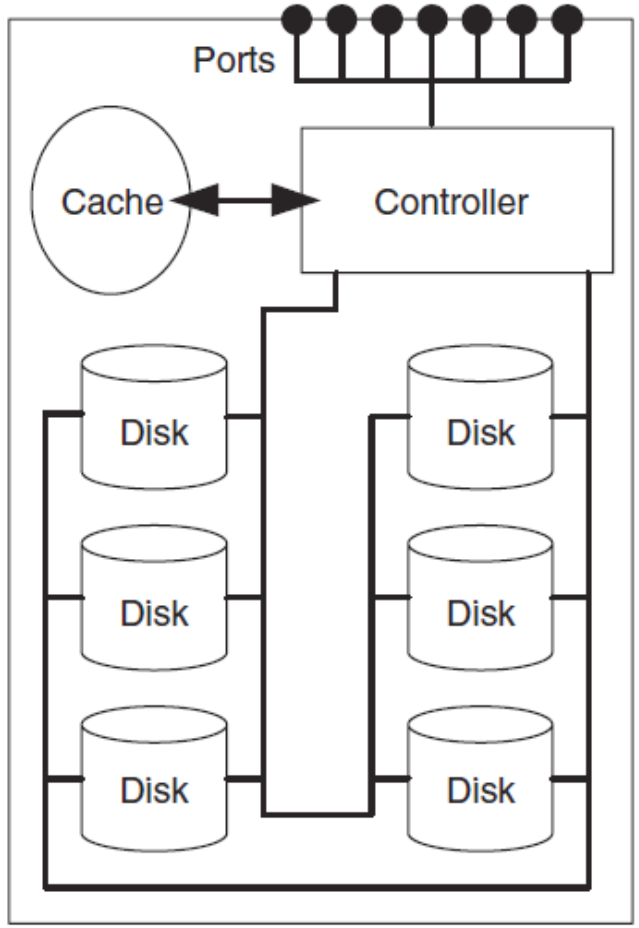


Дублированная

Активное дублирование

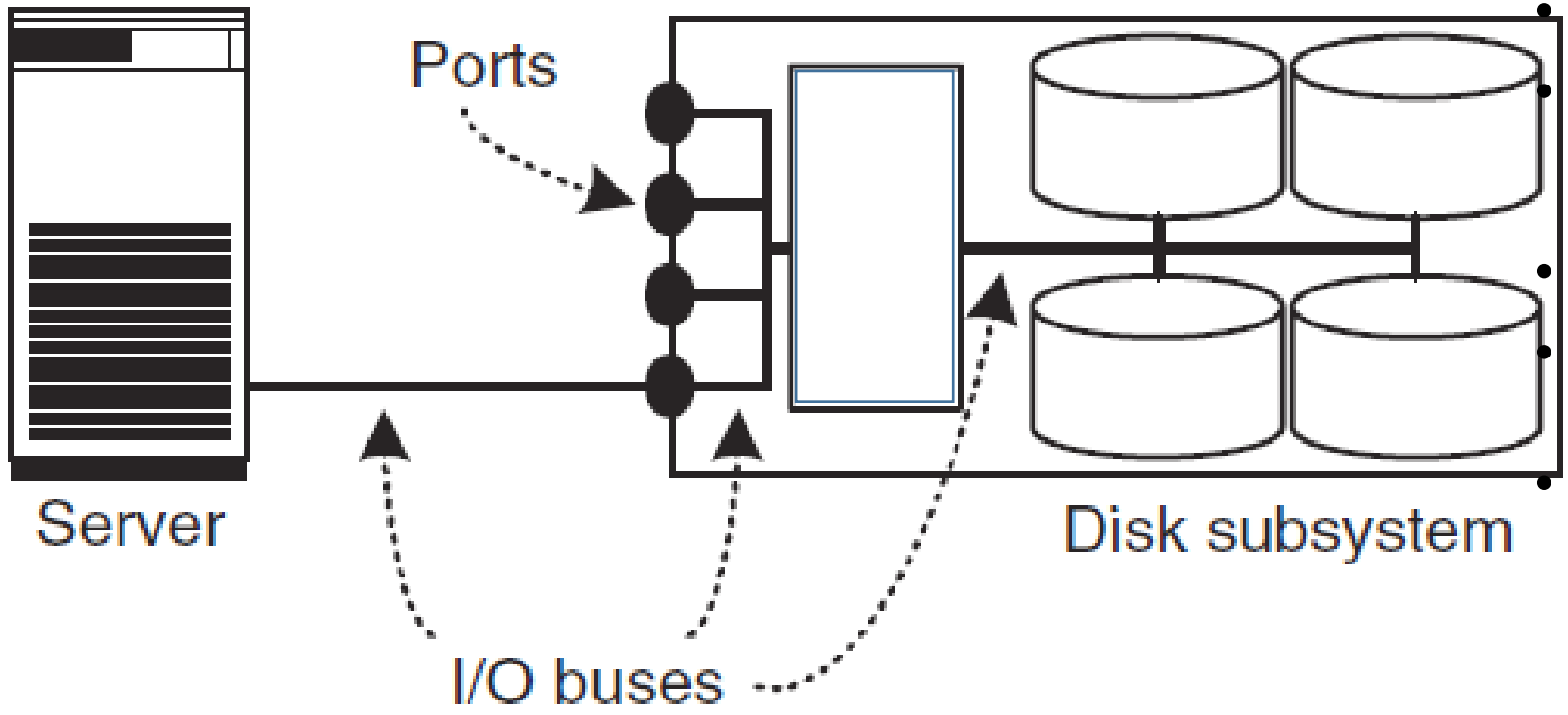


Active/Active без разделения
С фиксированным распределением дисков



Active/Active без разделения
С динамическим распределением дисков

JBOD: JUST A BUNCH OF DISKS



No internal controller

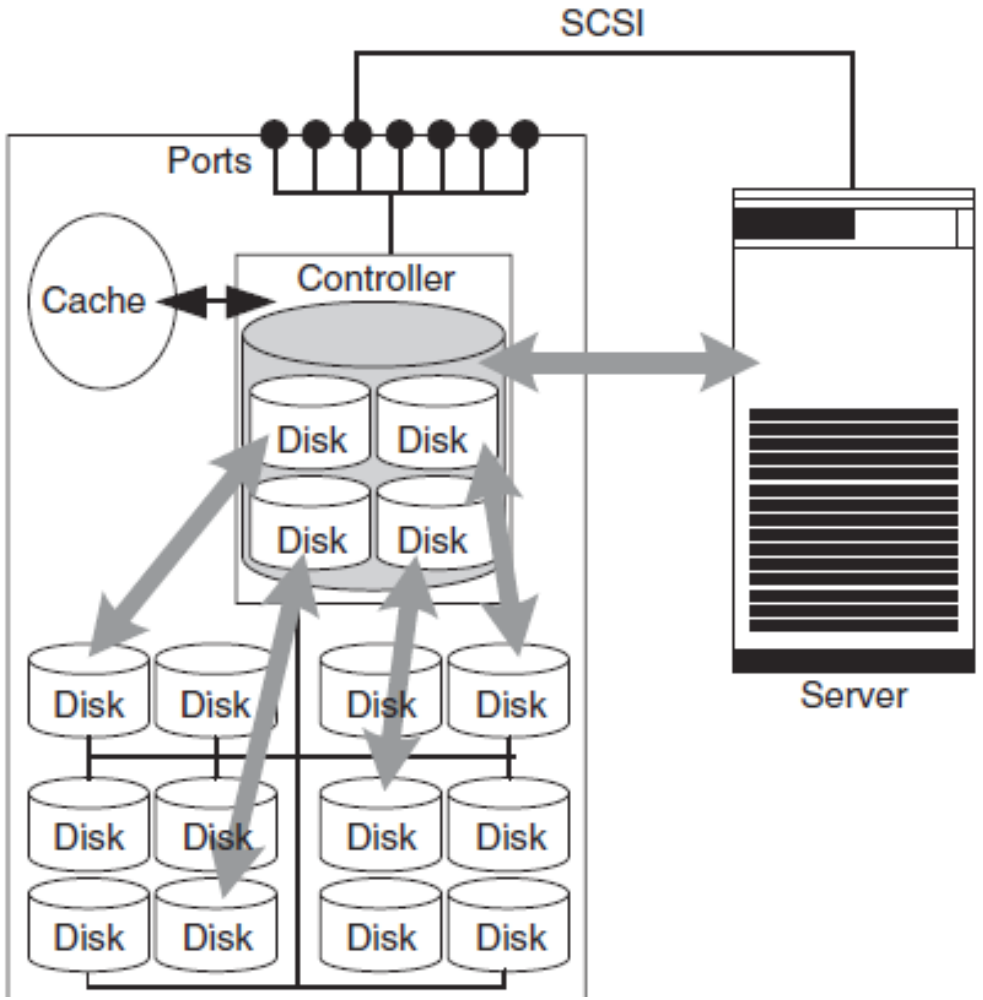
The connections for I/O channels and power supply are taken outwards

Small number of HDD

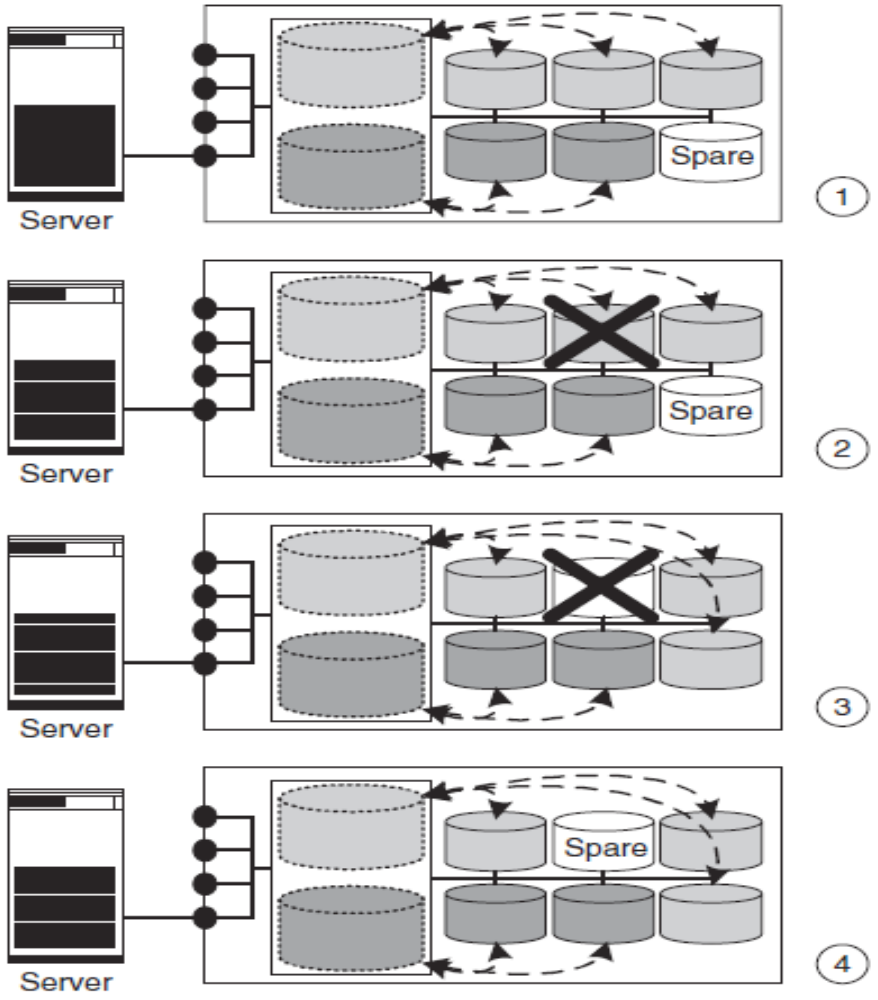
Outside server can see JOBD as **several independent disks**

Because of one point of connection JOBD it is the **bottleneck of Disk Subsystem**

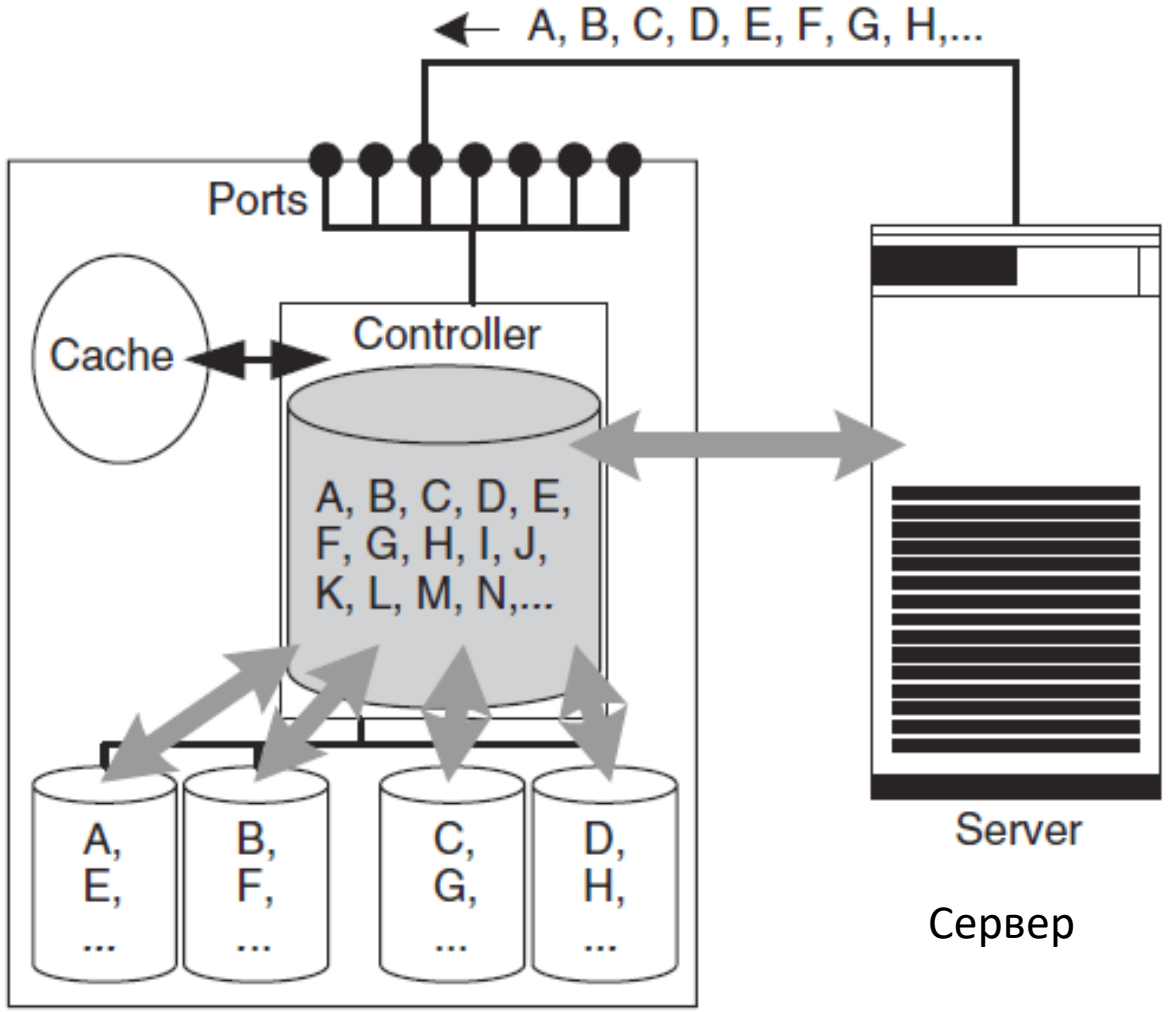
RAID – Избыточный массив независимых дисков



Горячий дисковый резерв

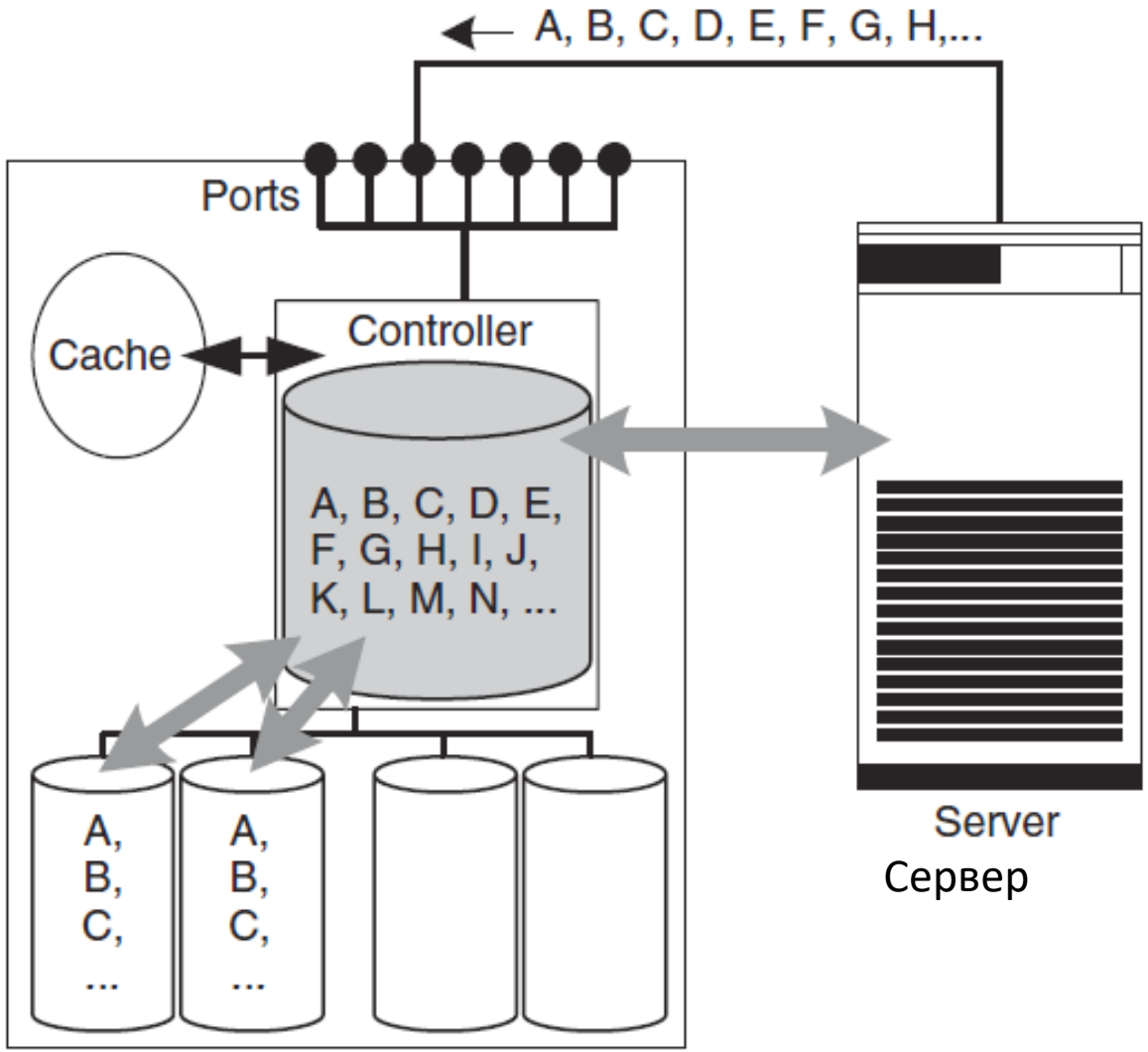


Уровни RAID = 0 (allocation)

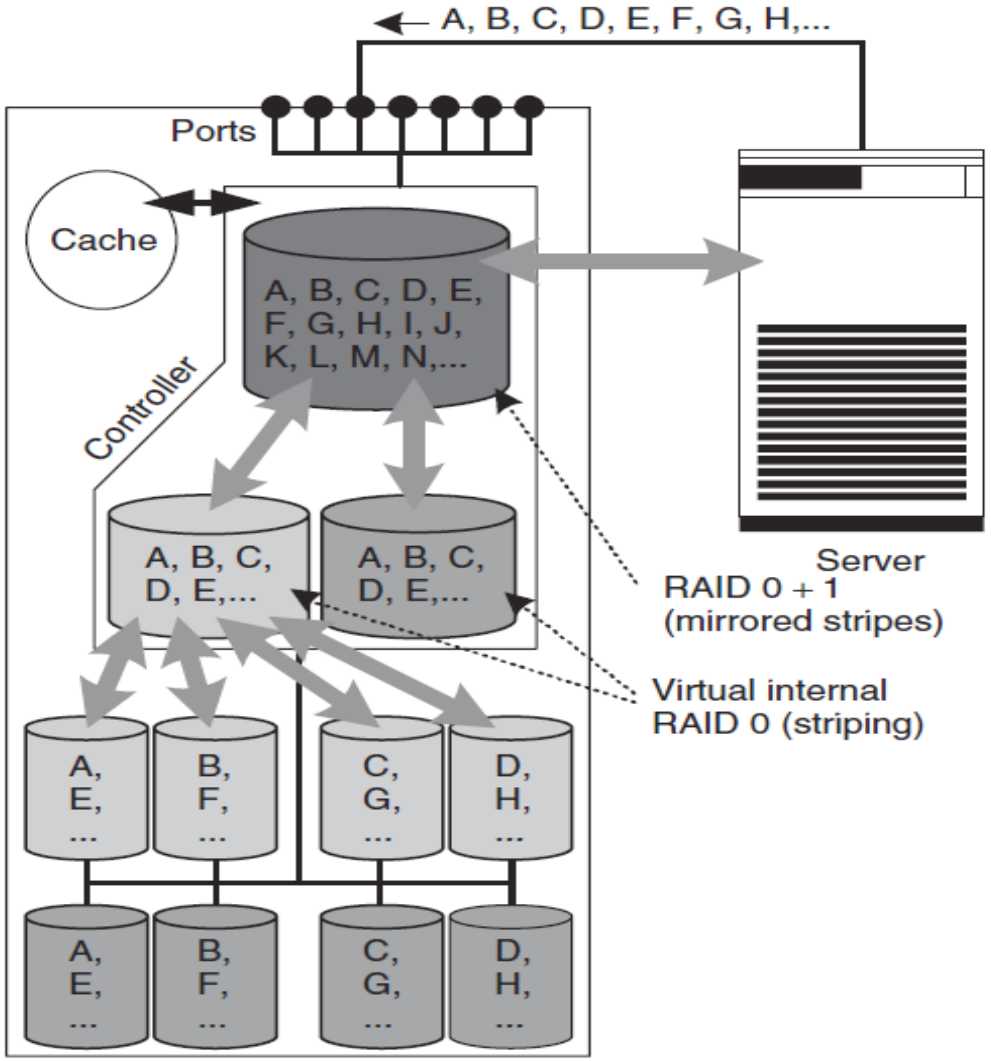


Выход из строя любого физического диска нарушит целостность данных.

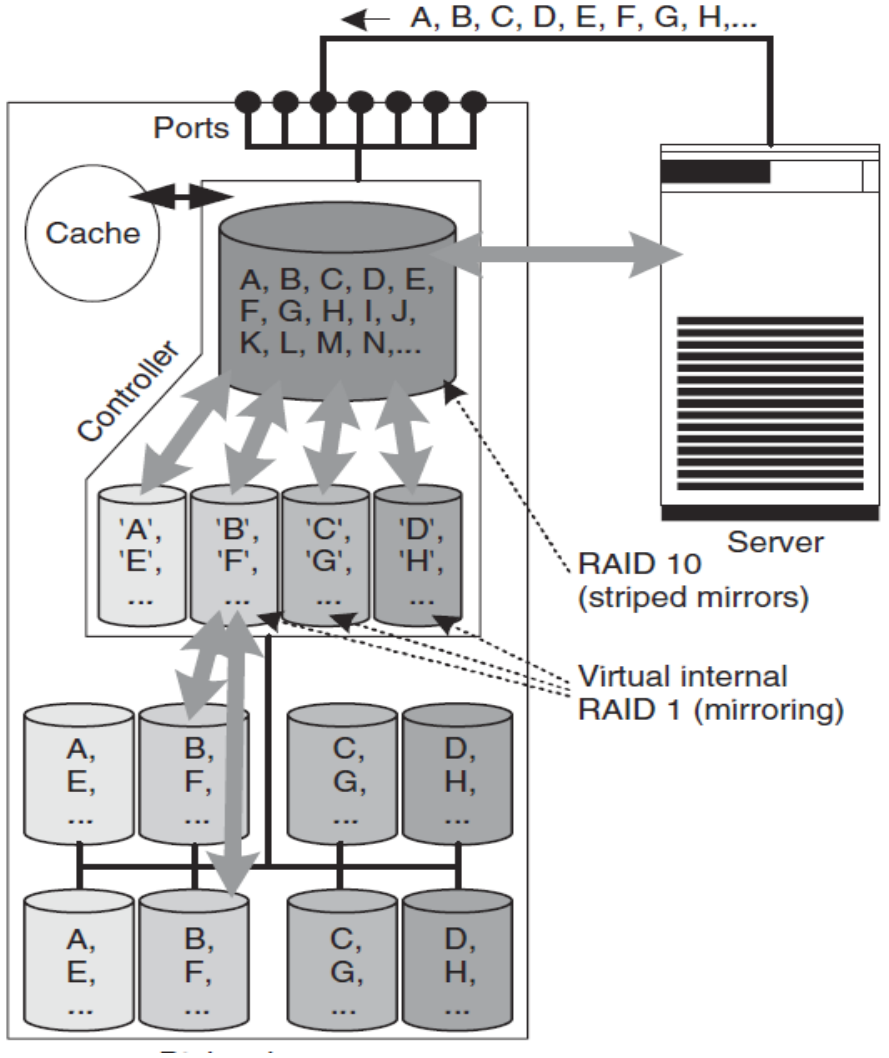
RAID = 1 (зеркалирование)



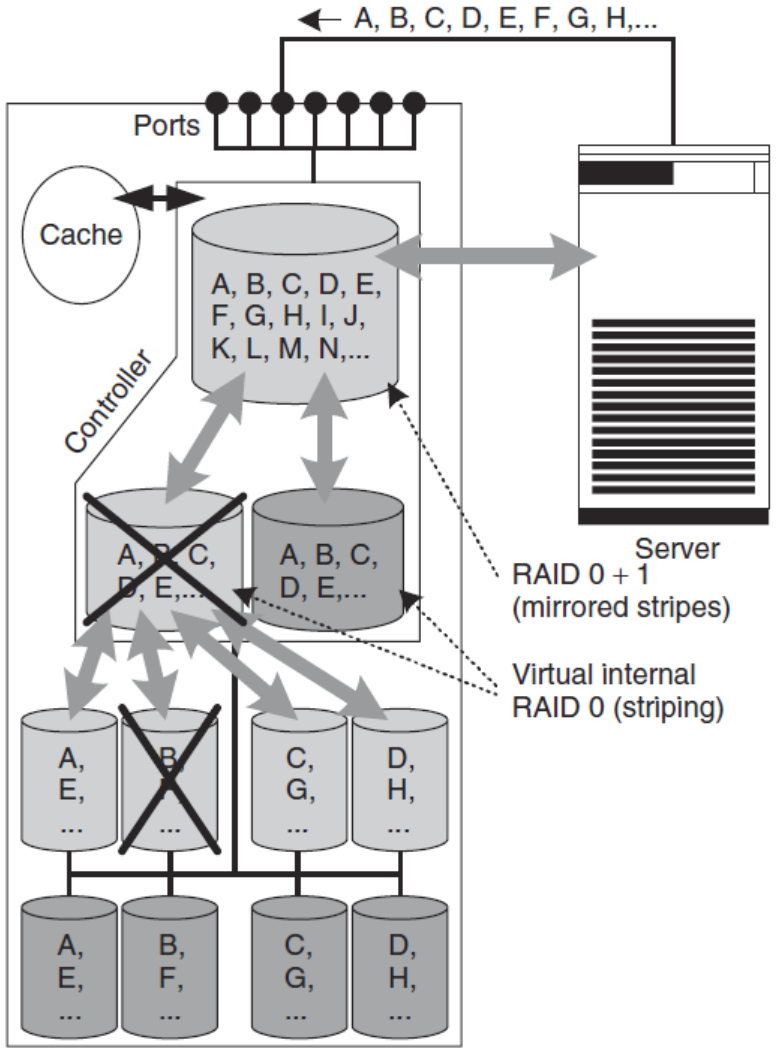
RAID = 0+1

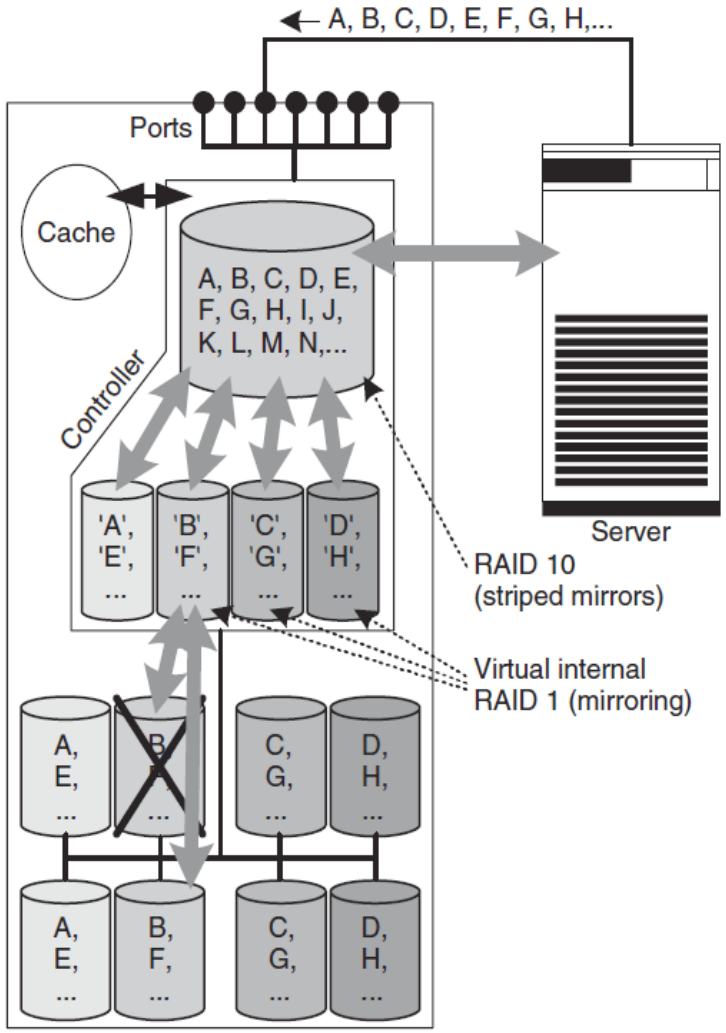

A, B, C, D, E, F, G, H,...
Ports
Cache
A, B, C, D, E, F, G, H, I, J, K, L, M, N,...
Controller
A, B, C, D, E,...
A, B, C, D, E,...
Server
RAID 0 + 1 (mirrored stripes)
Virtual internal RAID 0 (striping)
A, E, ...
B, F, ...
C, G, ...
D, H, ...
A, E, ...
B, F, ...
C, G, ...
D, H, ...

RAID = 1+0

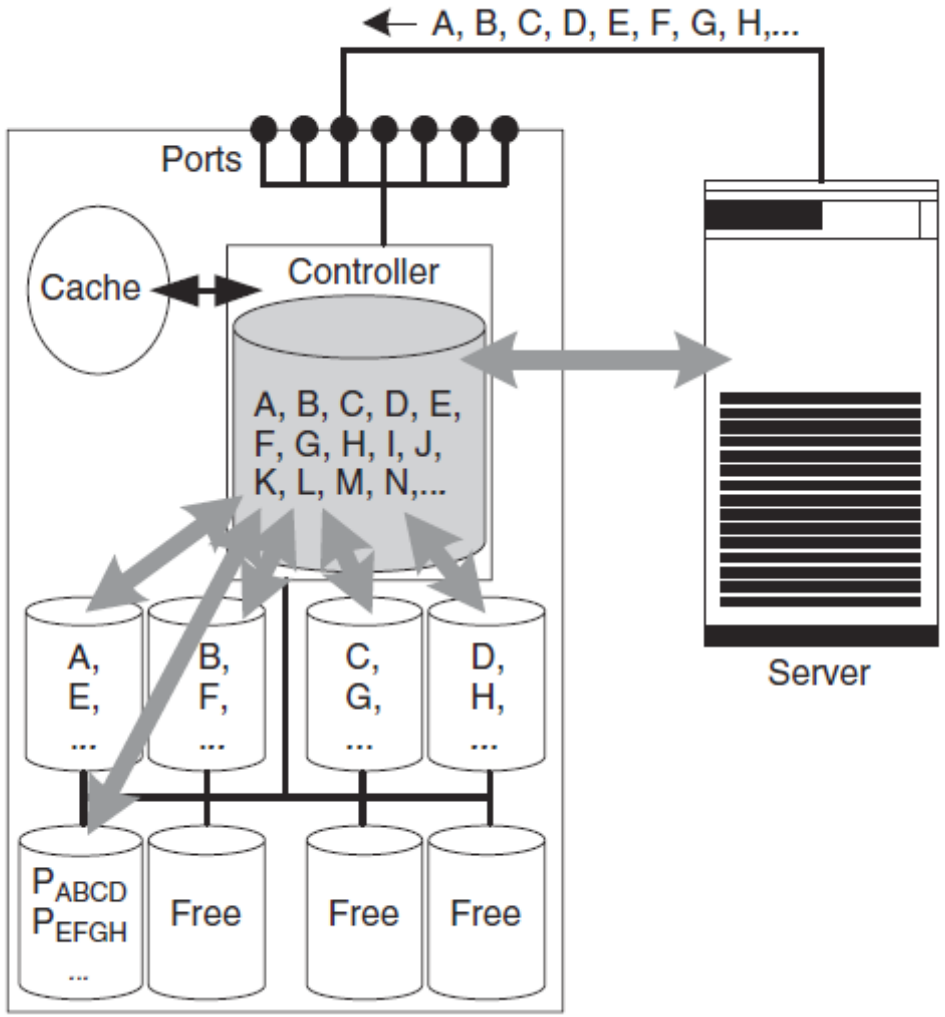

A, B, C, D, E, F, G, H,...
Ports
Cache
A, B, C, D, E,
F, G, H, I, J,
K, L, M, N,...
Controller
'A',
'E',
...
'B',
'F',
...
'C',
'G',
...
'D',
'H',
...
Server
RAID 10
(striped mirrors)
Virtual internal
RAID 1 (mirroring)
A,
E,
...
B,
F,
...
C,
G,
...
D,
H,
...
A,
E,
...
B,
F,
...
C,
G,
...
D,
H,
...

Сравнение RAID 0+1 и 1+0





RAID = 4



A изменили на ~A

Δ = A xor ~A

~P = Δ xor P

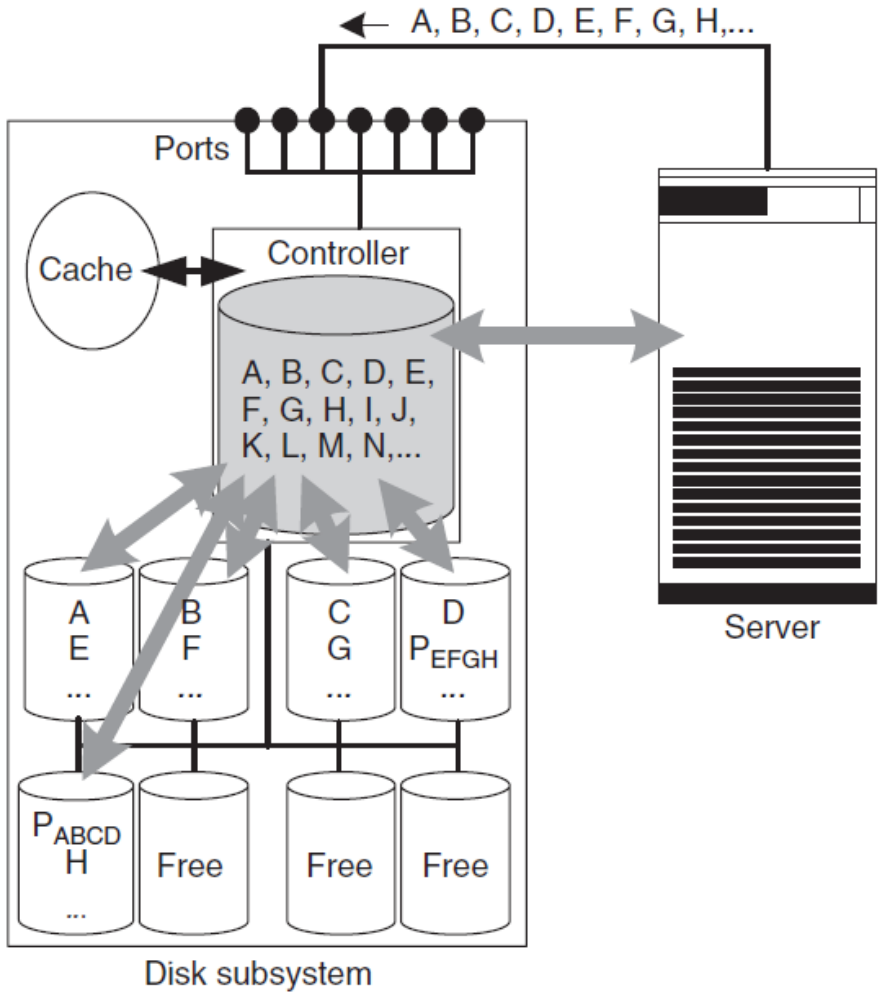
Если изменился только блок A,

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

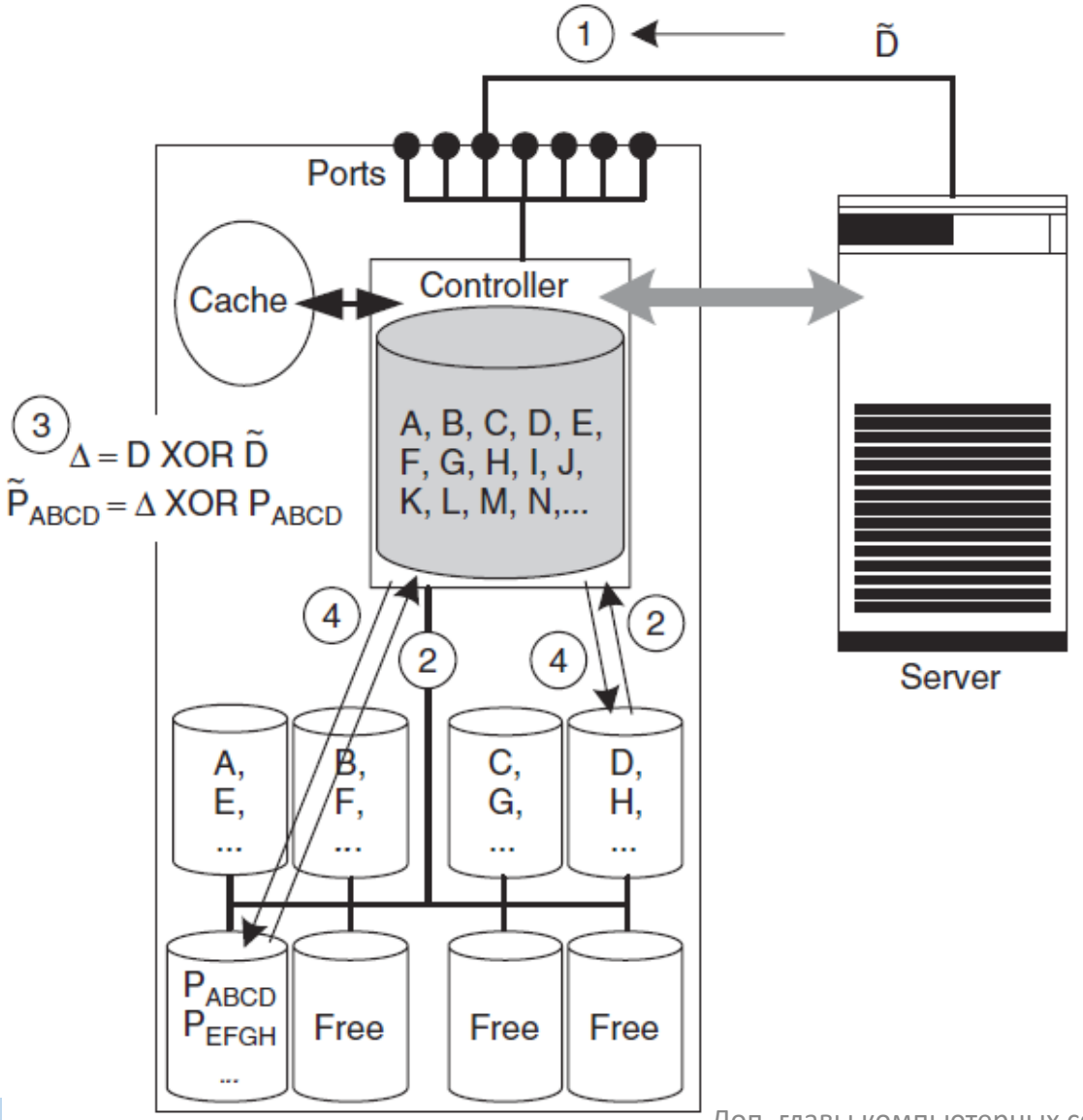
Однако надо считать старый блок A,

чтобы рассчитать Δ

RAID = 5


A, B, C, D, E, F, G, H,...
Ports
Cache
Controller
A, B, C, D, E,
F, G, H, I, J,
K, L, M, N,...
A
E
...
B
F
...
C
G
...
D
PEFGH
...
PABCD
H
...
Free
Free
Free
Server
Disk subsystem

Накладные расходы на запись в RAID = 4 и 5



- Сервер записывает измененный блок данных (1).
- RAID контроллер считывает старый блок данных и соответствующий ему блок четности (2)
- И рассчитывает новый блок четности (3).
- Наконец записывает новый блок данных и соответствующий ему блок четности (4).

RAID = 6

- Современные диски 1TB с BER 10^-15 => 100 TB одним сектором без ошибок не считать
- 10 дисковых массивов по 16X1TB будут терять один массив 1 раз в год+
- Режим эксплуатации теперь 7X24
- RAID 6 использует дополнительный диск четности для групповых ошибок
  - Увеличение затрат
  - Увеличение времени операции записи и коррекции

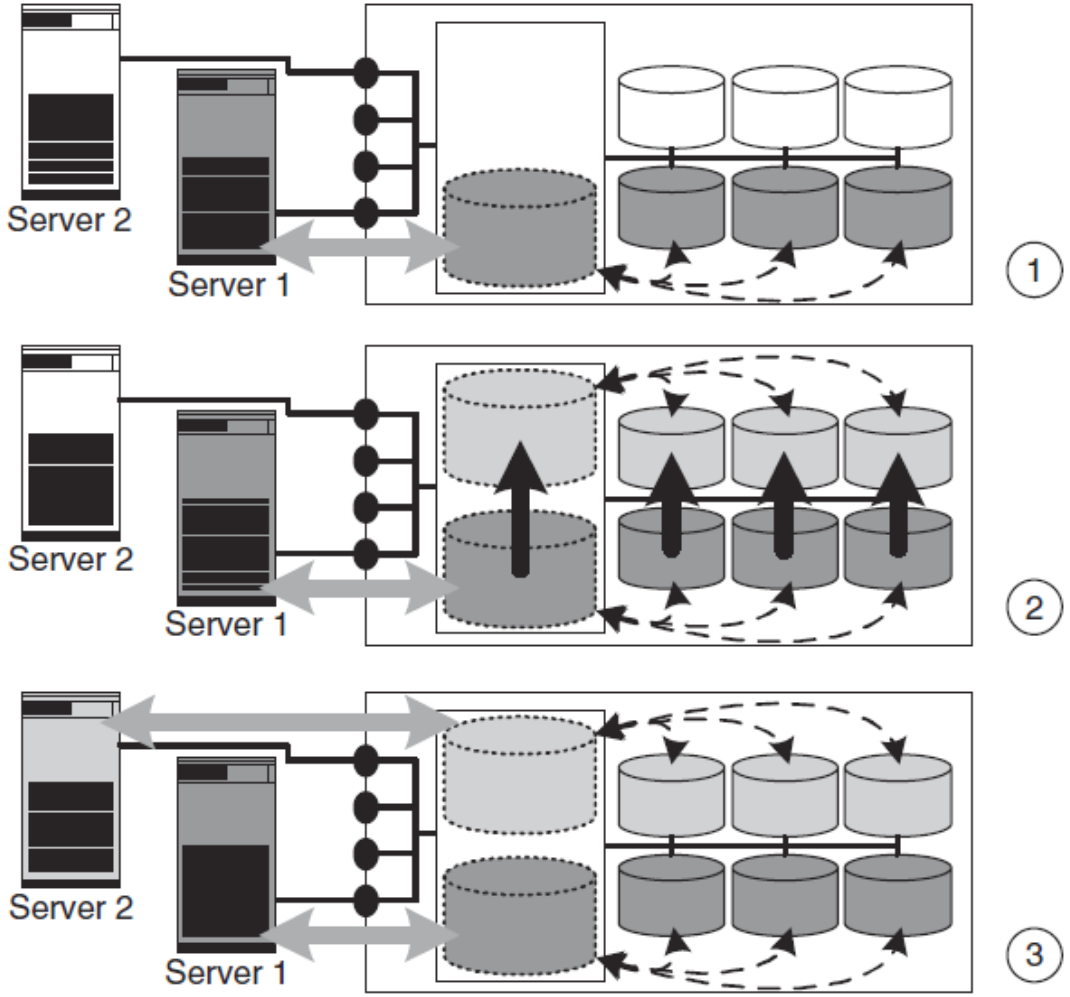
Сравнение схем RAID массивов

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

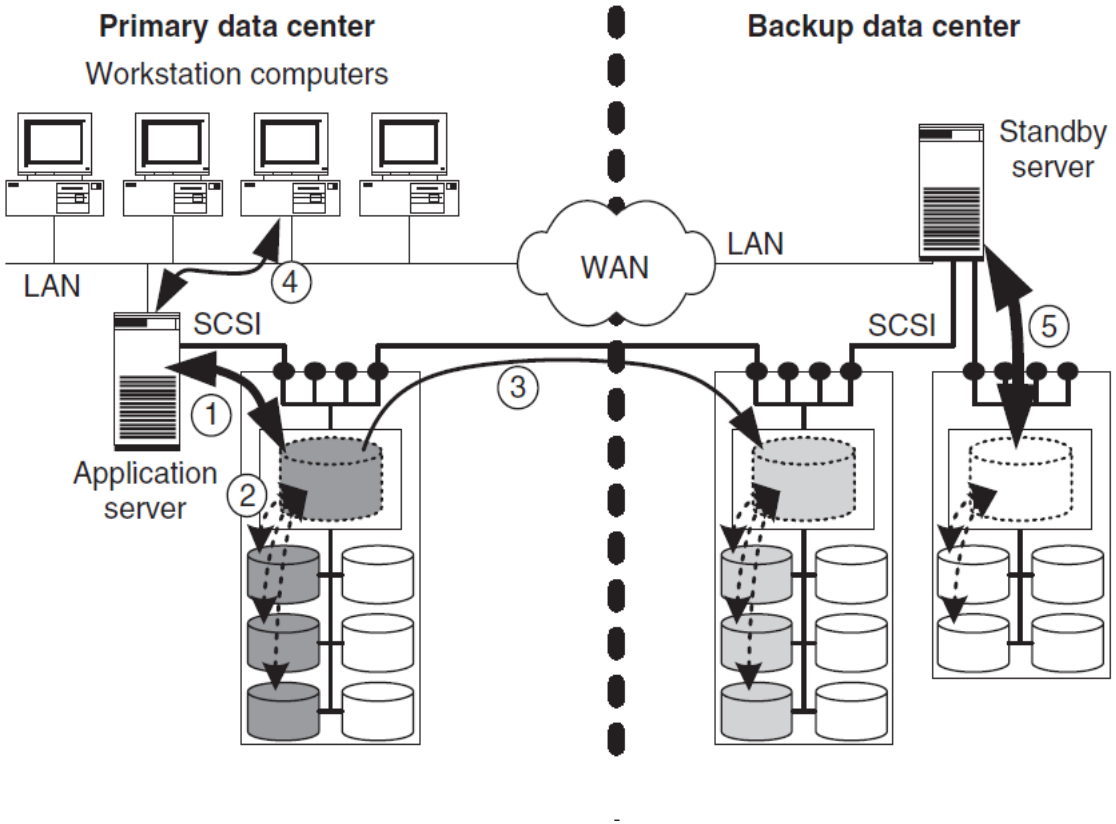
Caching – ускорение дисковых хранилищ

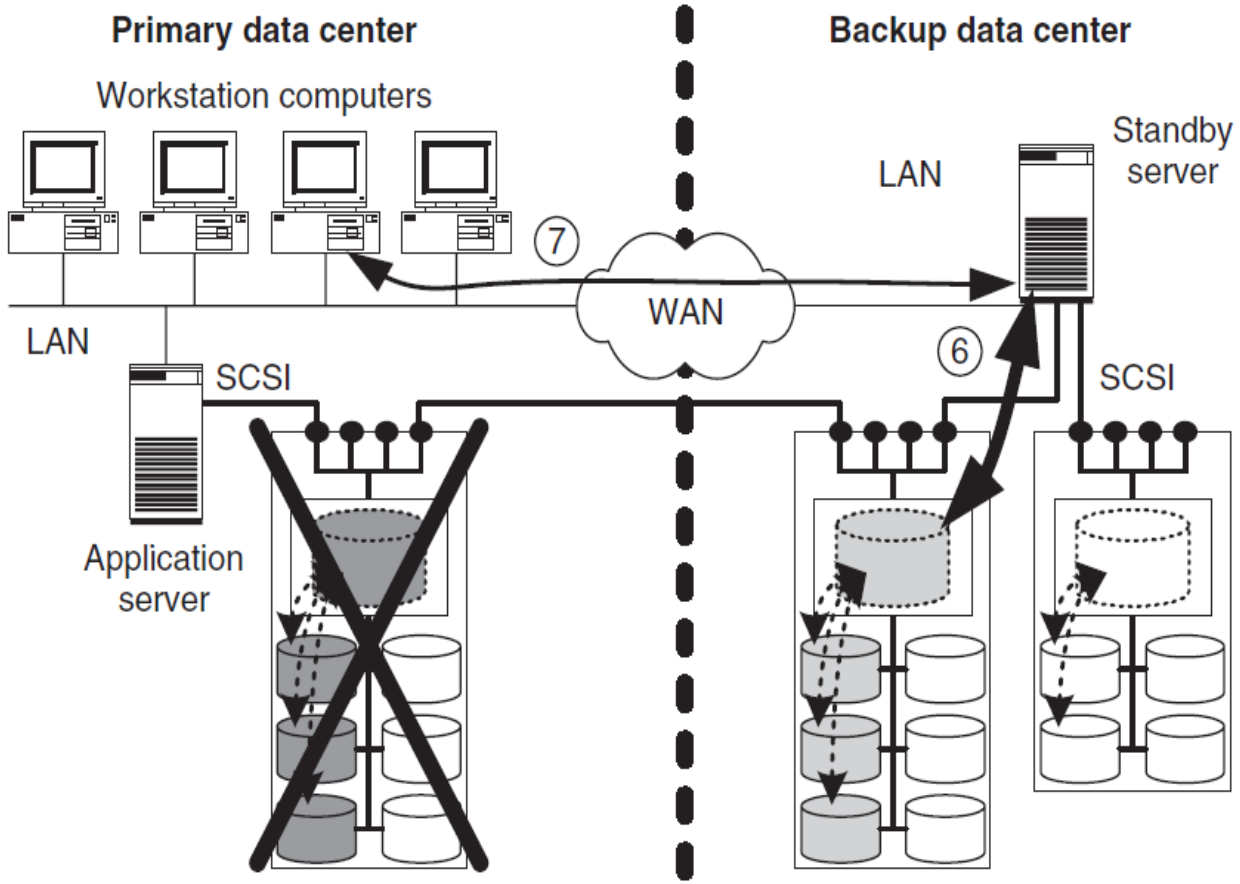
- Кэш на уровне HD
- Кэш на уровне контроллера ДС при записи
  - ГБ кэш
  - Главное сохранить данные в кэш даже при отключении питания (UPS)
  - Важно для блочных приложений
- Кэш для ускорения чтения контроллером ДС

Интеллектуальная дисковая подсистема (мгновенное копирование)

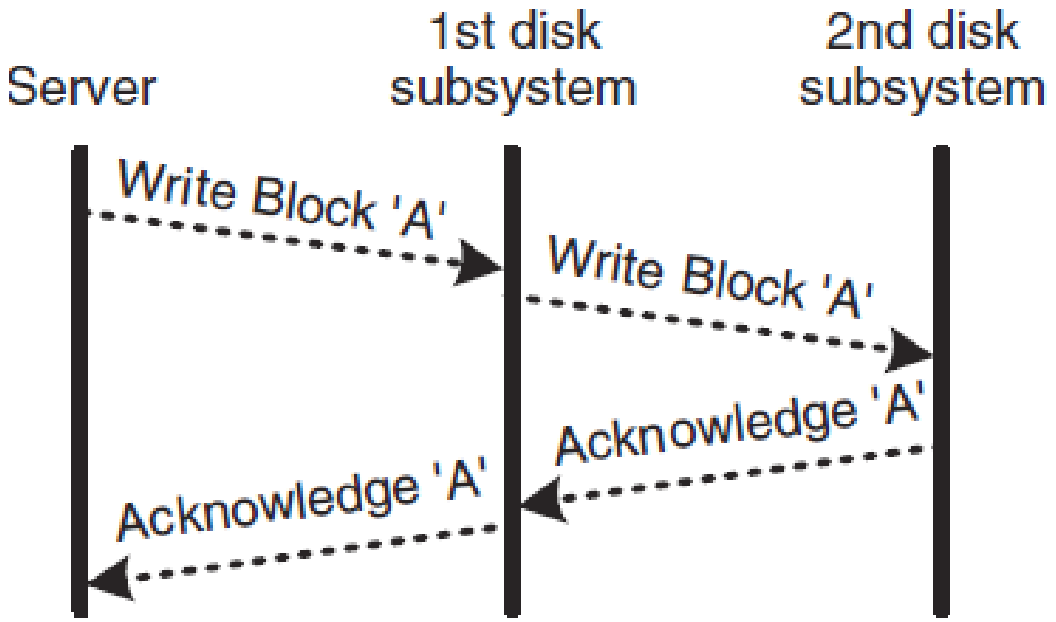


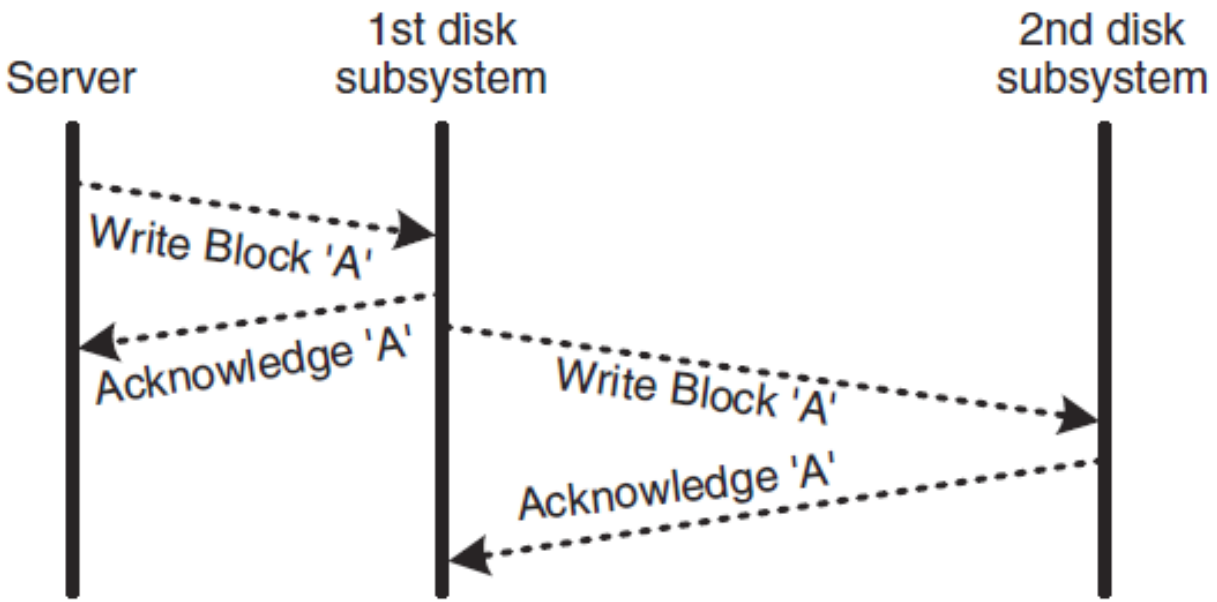
Удалённое зеркалирование



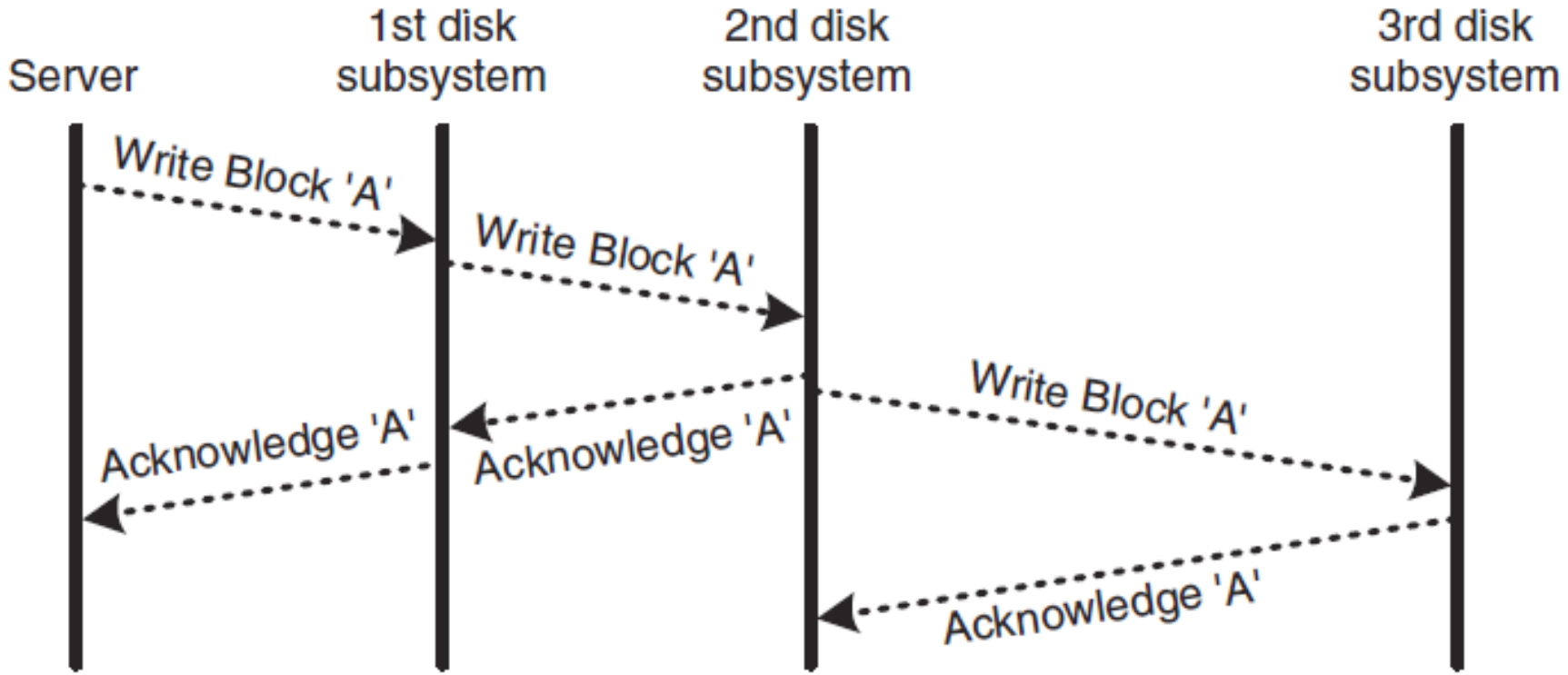


Синхронное и асинхронное удаленное зеркалирование

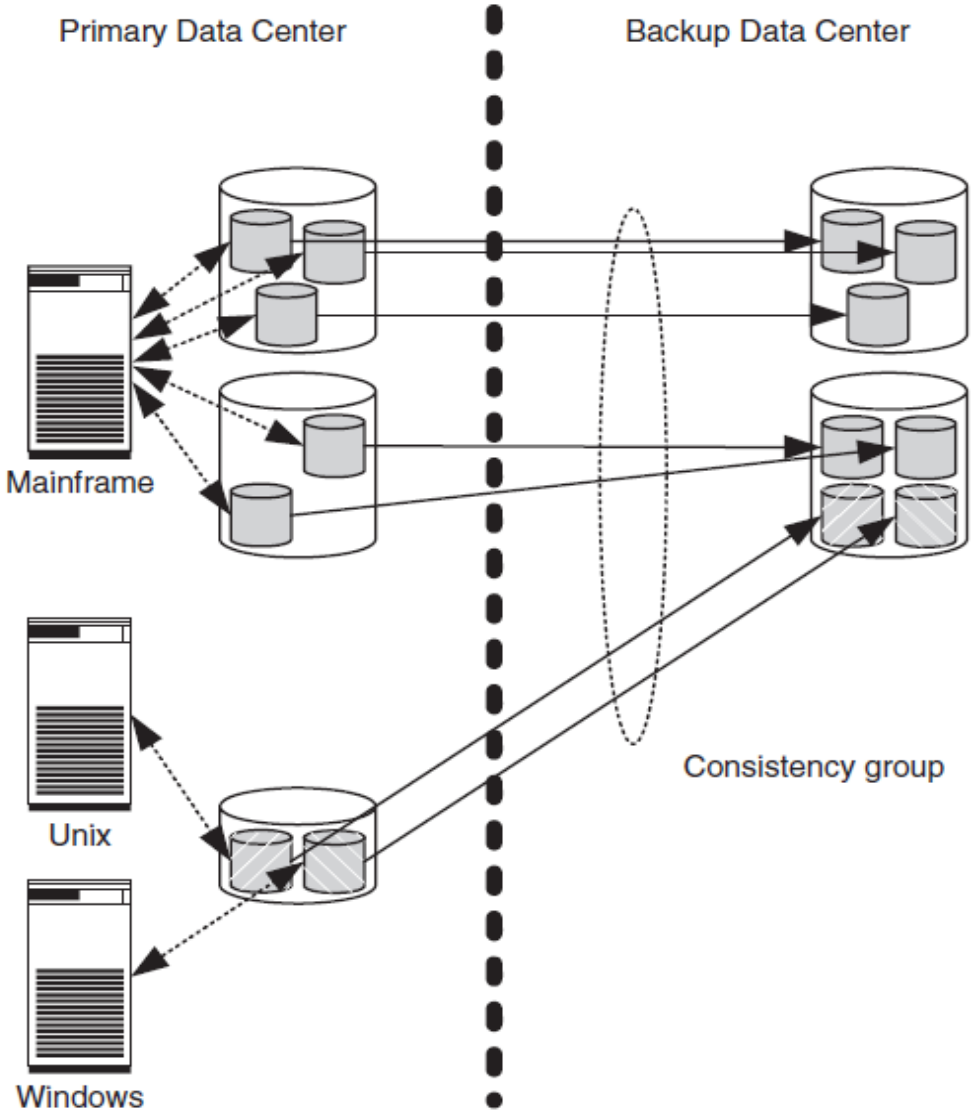


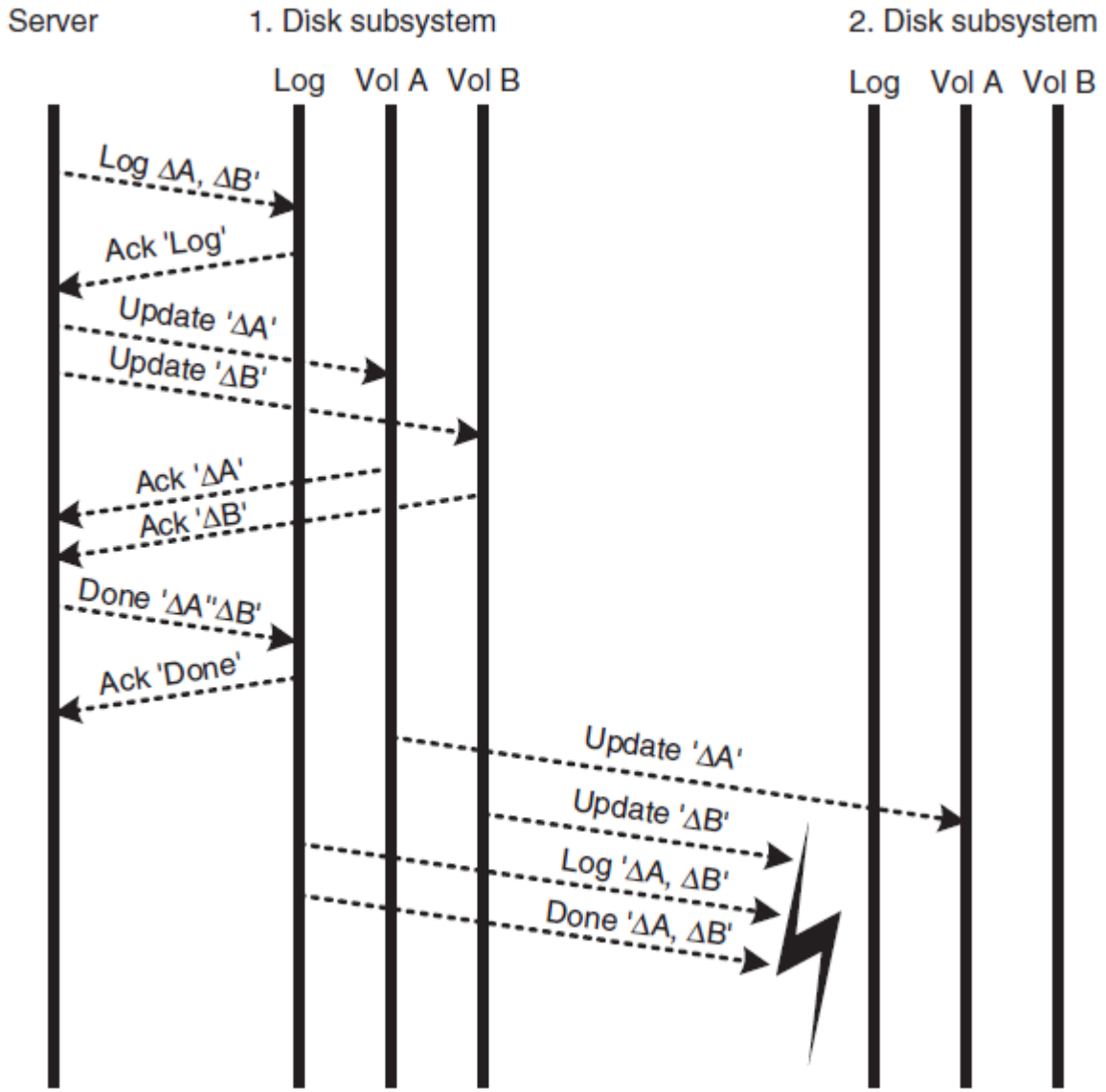


Комбинированная схема удалённого зеркалирования

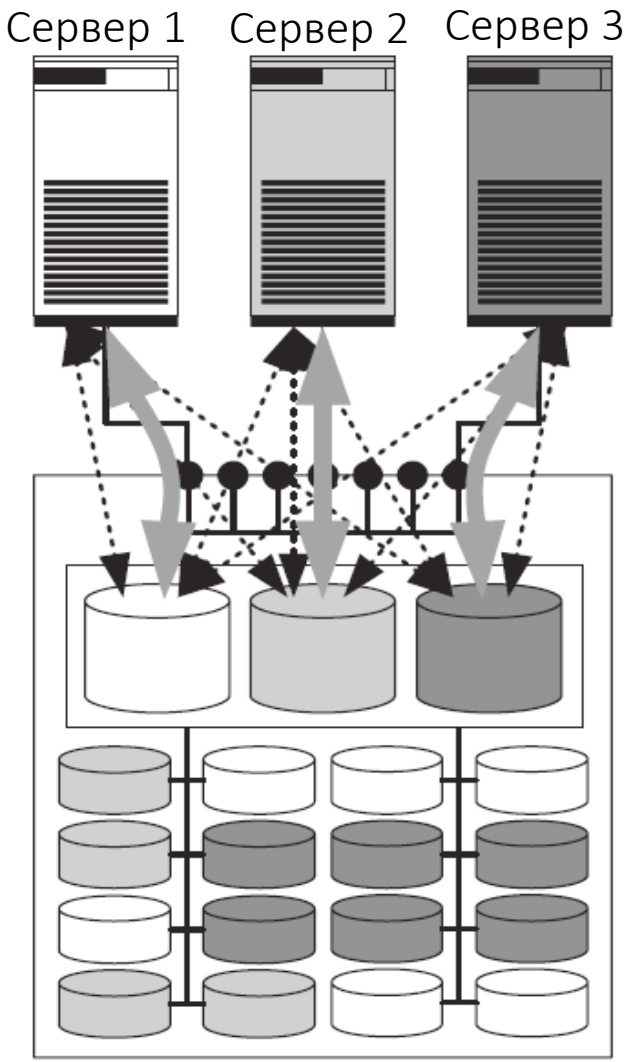


Группы консистентности





LUN маскирование

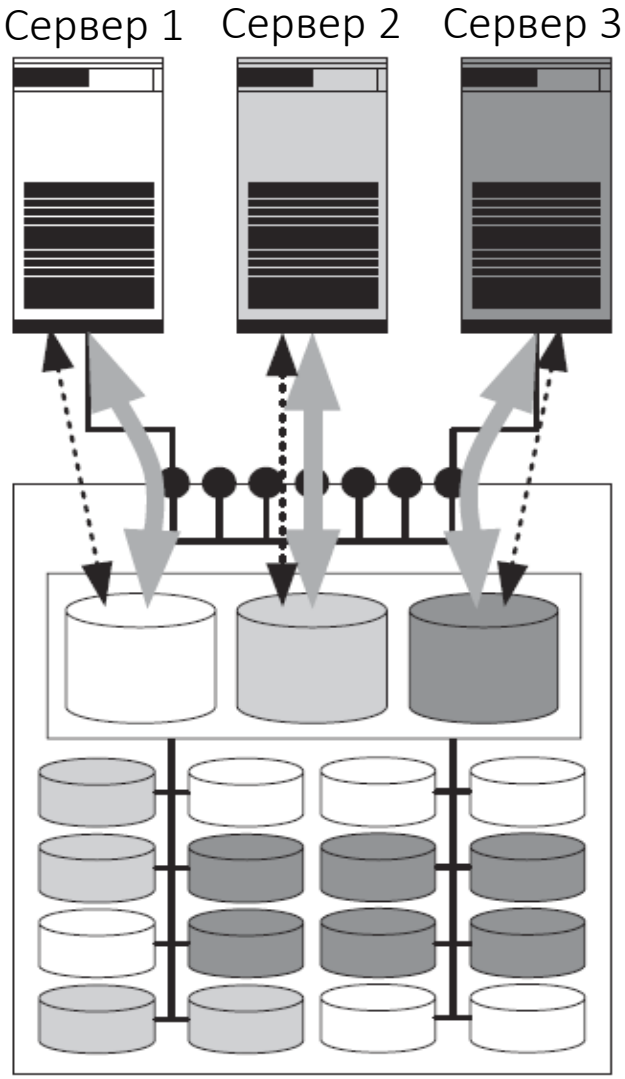


Дисковая подсистема

Logical
Unit
Number

Сервер использует LUN

Сервер видит LUN



Дисковая подсистема

Устойчивость работоспособности ДПС

• Данные распределяют по нескольким дискам с помощью механизмов RAID и снабжают избыточными данными (блоки четности).

• На каждом физическом диске данные закодированы кодом Хемминга. Кроме этого диск оснащен подсистемой самодиагностики, которая контролирует частоту ошибок, вибрацию шпинделя и т.д. Это позволяет проактивно прогнозировать отказы диска.

• Каждый диск подсоединен к контроллеру хотя бы через две внутренние шины.

• Контроллер дисковой подсистемы может быть продублирован. Выход одного экземпляра, автоматически будет активизировать следующий экземпляр. Схема Active-Standby.

• Дублируются UPS, системы охлаждения.

• ДС подключают к разным электрическим сетям

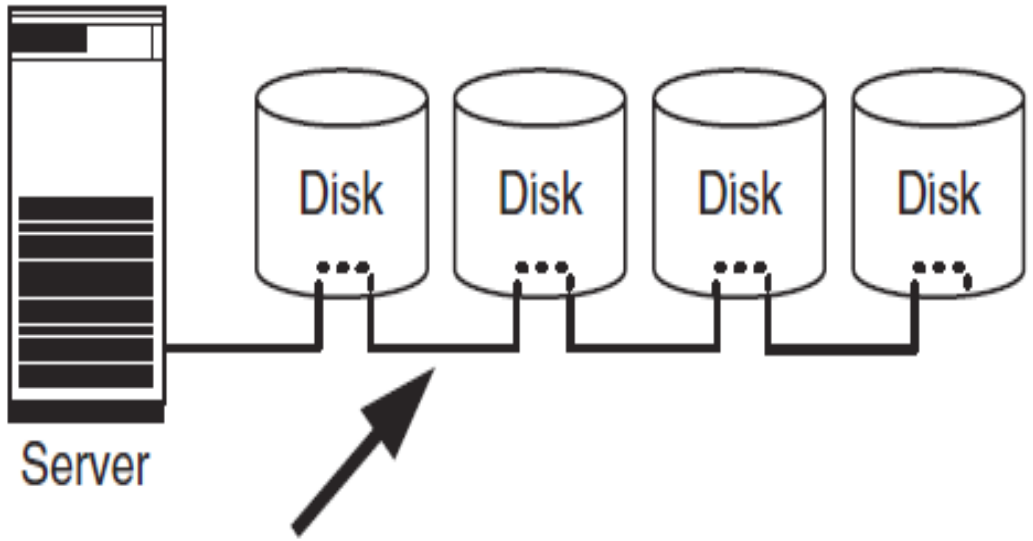
• Сервер соединяют с ДС через несколько линий.

• Используют периодическое мгновенное копирование для защиты от логических ошибок. Например, создание мгновенной копии данных через каждый час. Тогда в случае сбоя и уничтожения какой-то таблицы, она может быть восстановлена.

• Удаленное зеркалирование используют от физического уничтожения или повреждения оборудования (катастрофоустойчивость). В сочетании с мгновенным копирование эти сервисы гарантируют сохранение и консистентность данных даже для нескольких виртуальных дисков или дисковых подсистем.

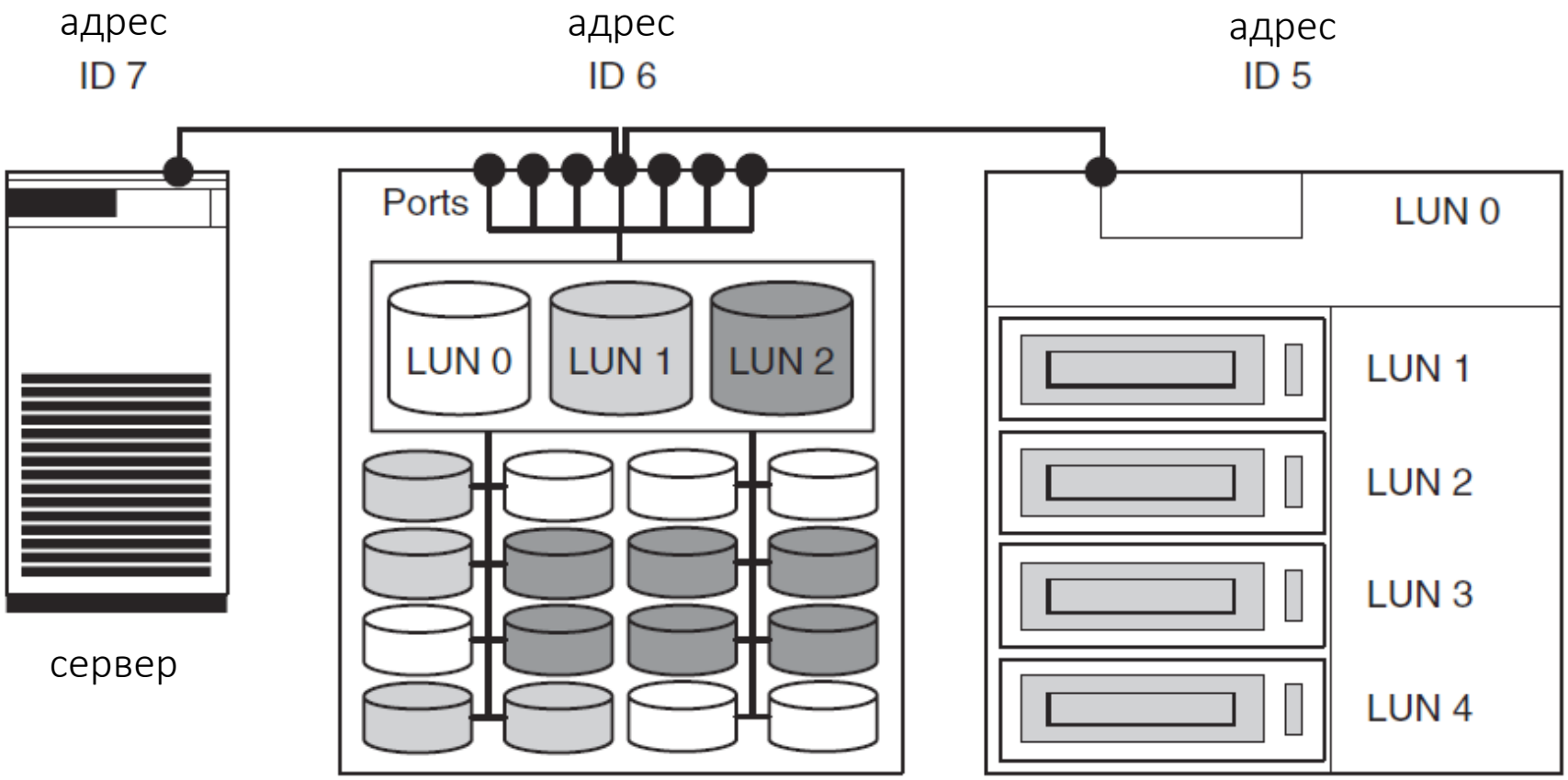
• LUN маскирование защищает от несанкционированного доступа, упрощает работу системного администратора, защищает от случайных сбоев в работе приложений серверов и их оборудования.

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

Адресация устройств на шине SCSI

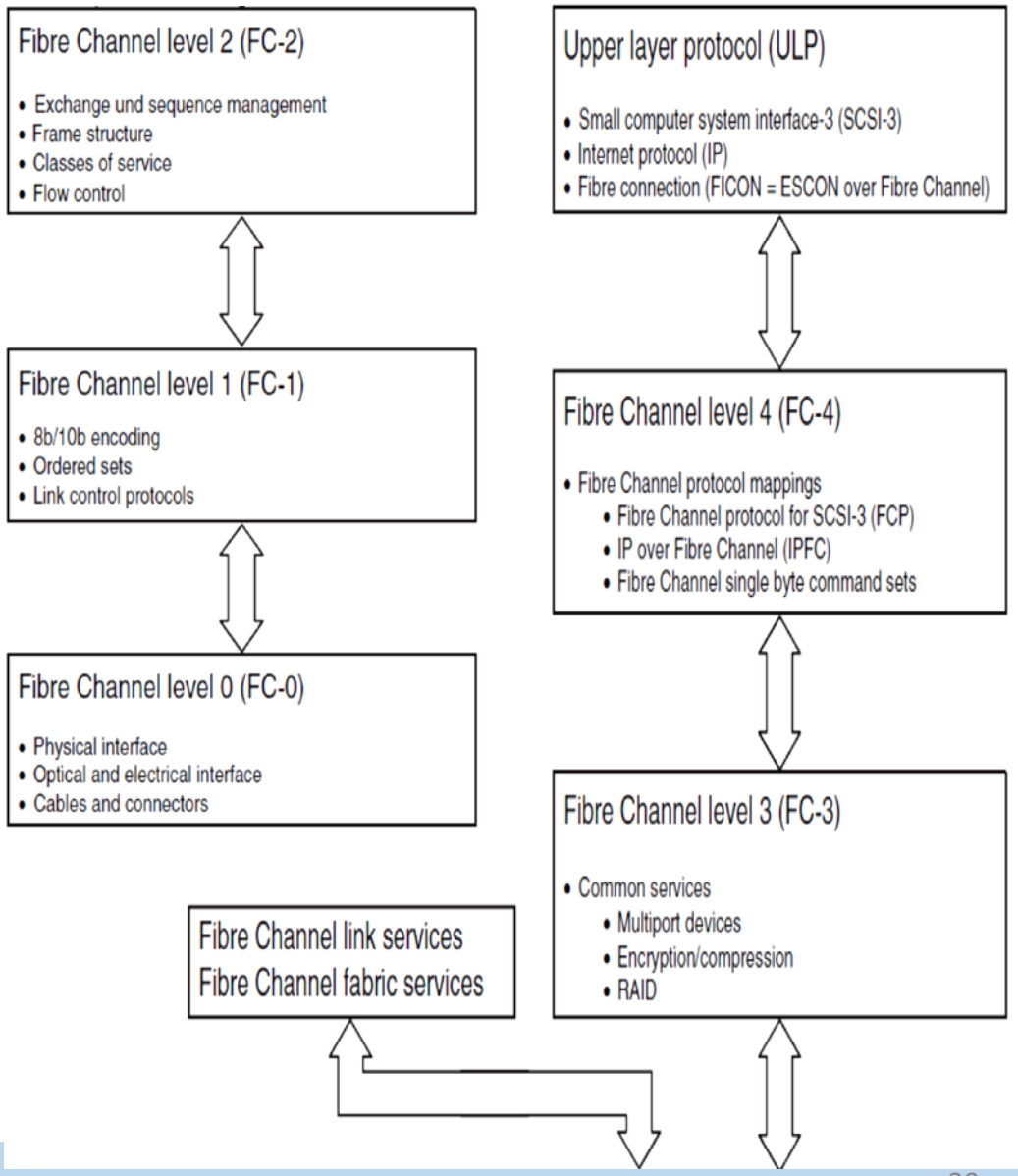


Fibre Channel

- I/O канал для LAN
  - Последовательной передачи на большие расстояния и высокой скорости.
  - Низким уровнем ошибок передачи
  - Малой задержкой передачи
  - Дешевизной реализации FC на картах HBA на серверах
- Fiber Channel сопрягаем с IPI (Intelligent Peripheral Interface), SCSI, HIPPI (High Performance Parallel Interface), ATM, IP и 802.2 (Ethernet).
- Fibre Channel - n x100МБ/с при длинах канала 10 км и более,
  где n – число каналов. Предельная скорость передачи - 4,25 Гбс
- В качестве физической среды может использоваться одномодовое или многомодовое оптическое волокно, коаксиал, витая пара при скоростях до 200 МБ/с.

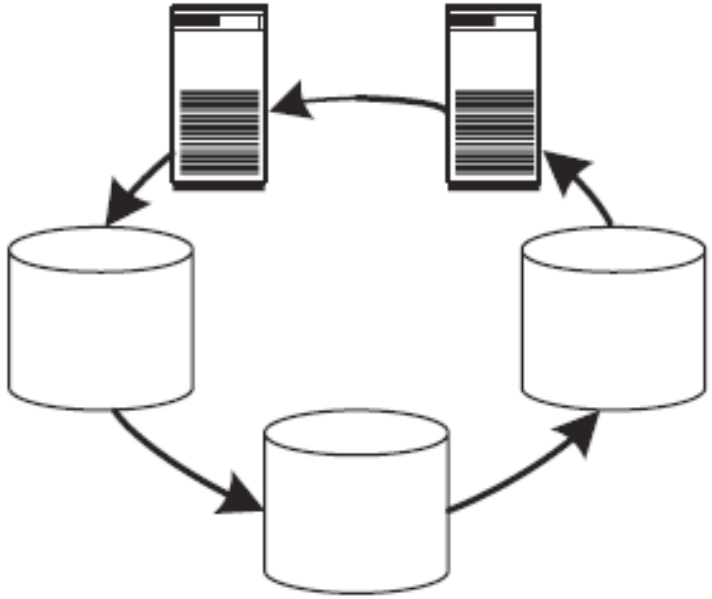
Стек Fiber Channel

- 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 протокола управления, 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.)

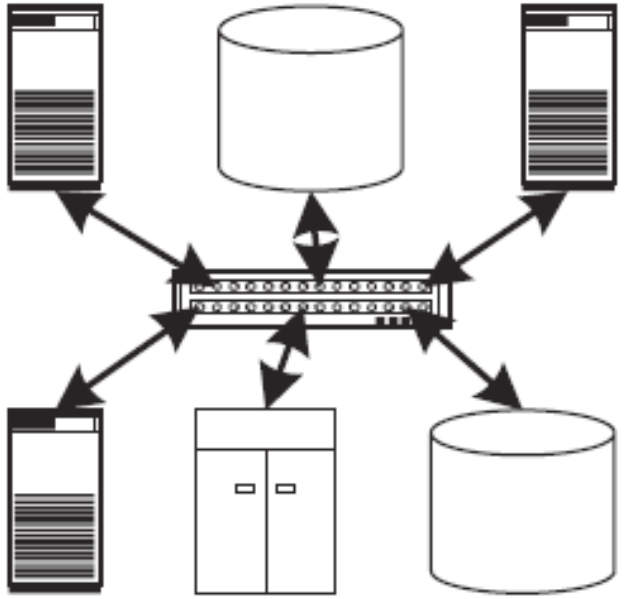


Топологии FC

Точка-Точка

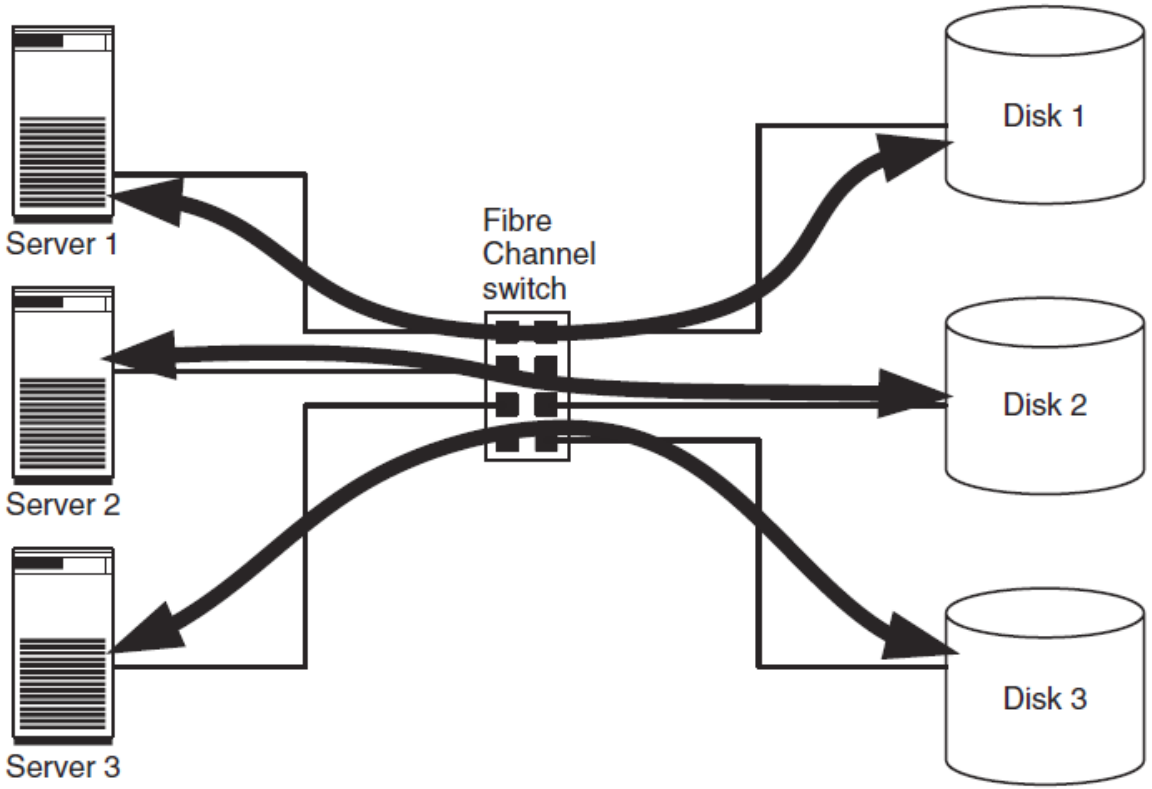


Кольцо с арбитражем

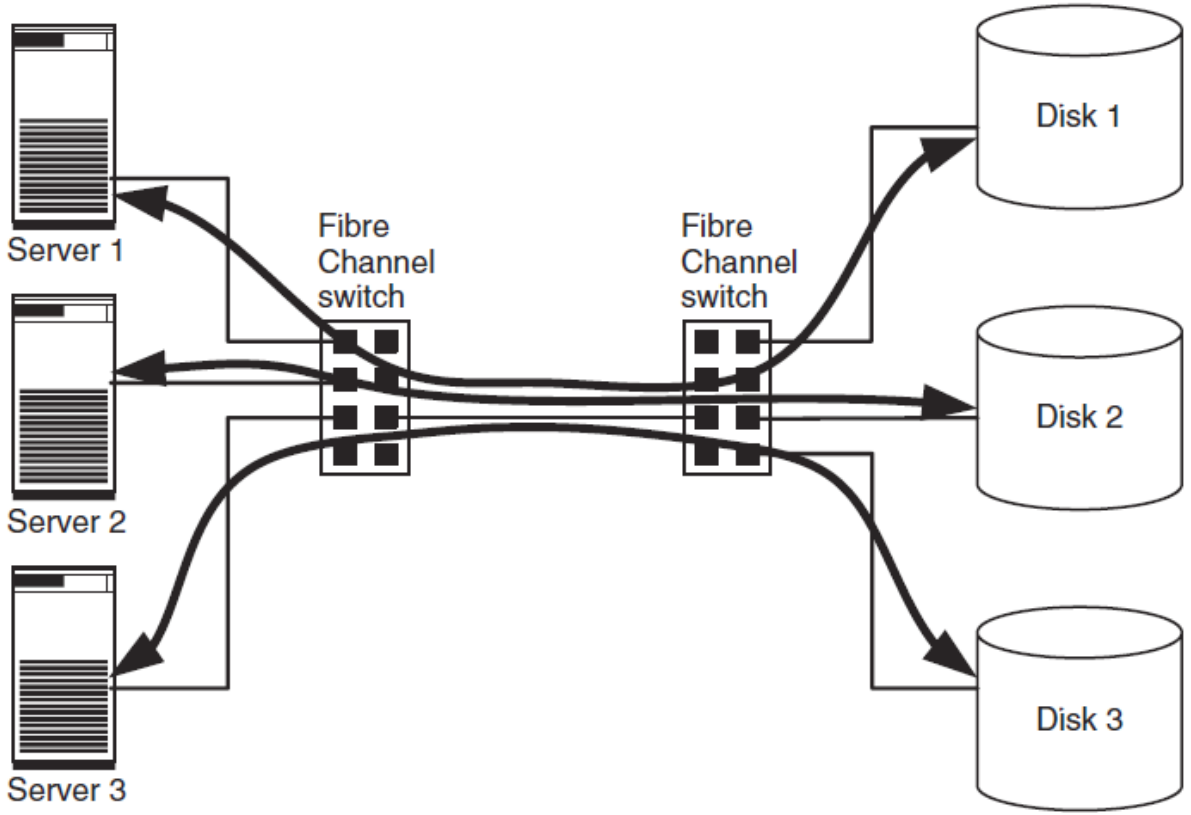


Коммутатор

Топология коммутационной среды (КМ)

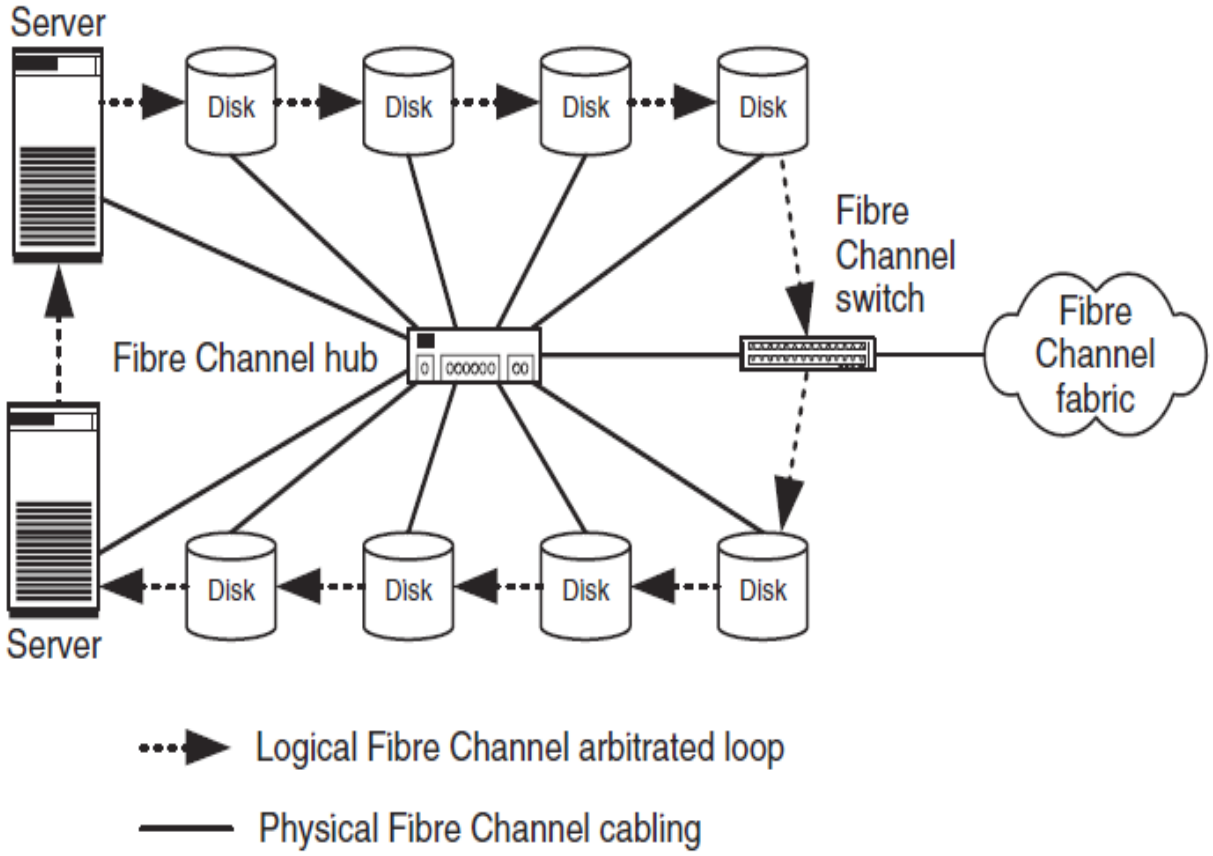


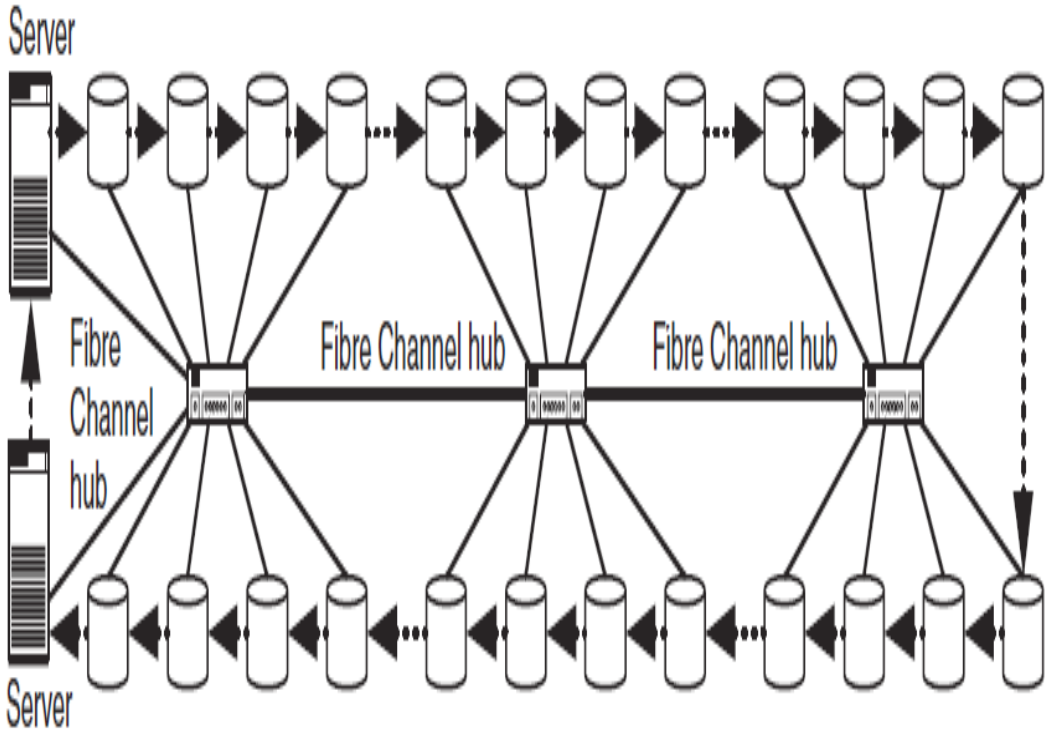
Несколько соединений на полной скорости.



Несколько соединений через один ISL.

КсА топология



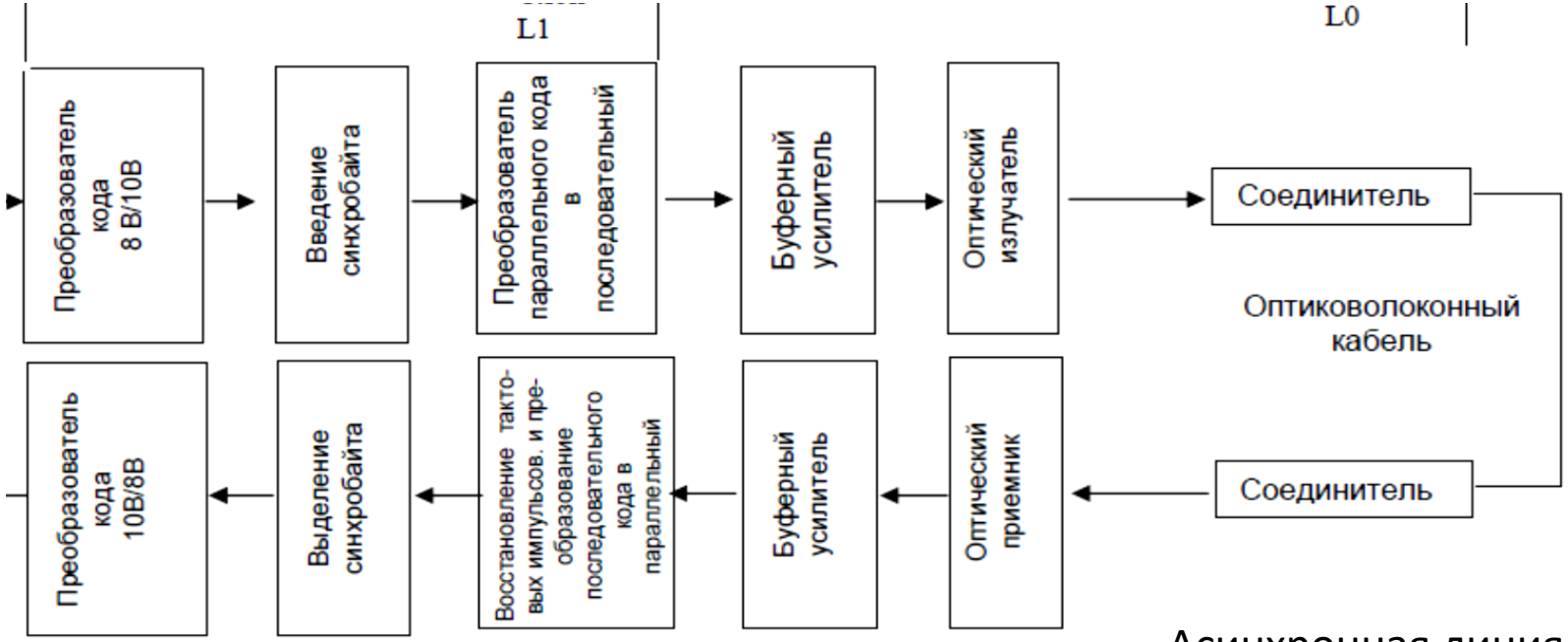


Fibre Channel: типы портов

- N-Port – порт для соединения устройства с коммутатором или другим устройством
- F-Port – порт для подключения к коммутатору
- L-Port – порт для КсА
- NL-Port – работает как N Port или как L-Port. Можно подключать порт через коммутатор или в КсА.
- FL-Port – для подключения коммутатора в КсА
- E-Port – для соединения двух FC коммутаторов
- G-Port – может настраиваться как E или FL в зависимости от подключения.
- B-Port – для соединения двух FC коммутаторов через ATM, SDH, Ethernet или IP. Например две FC SAN могут быть соединены через WAN.

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



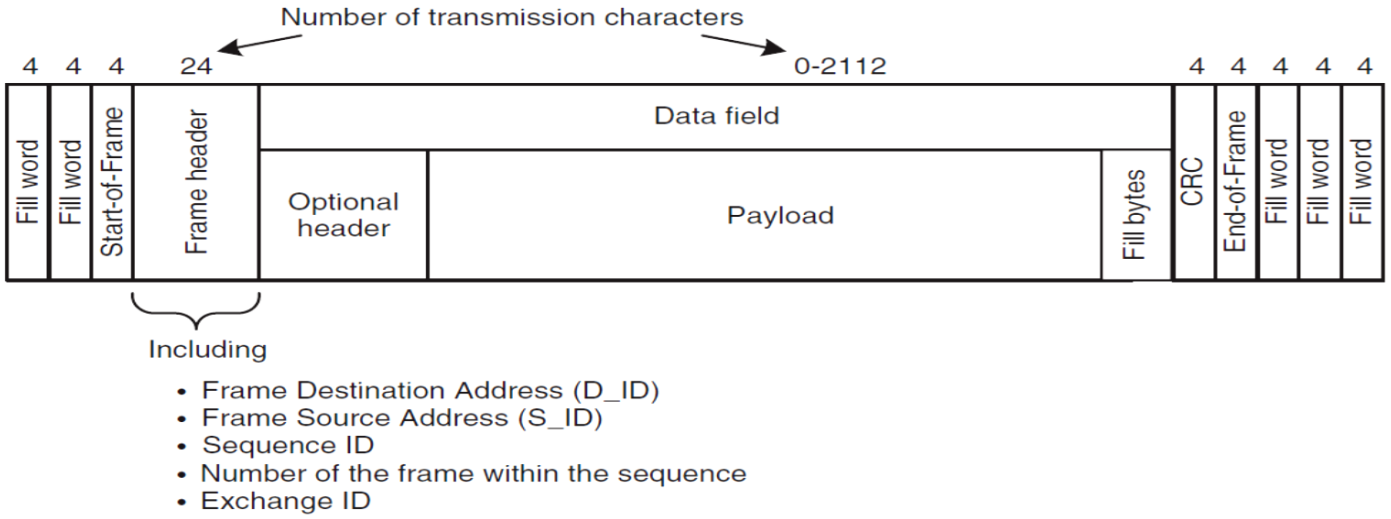
Асинхронная линия последовательной передачи

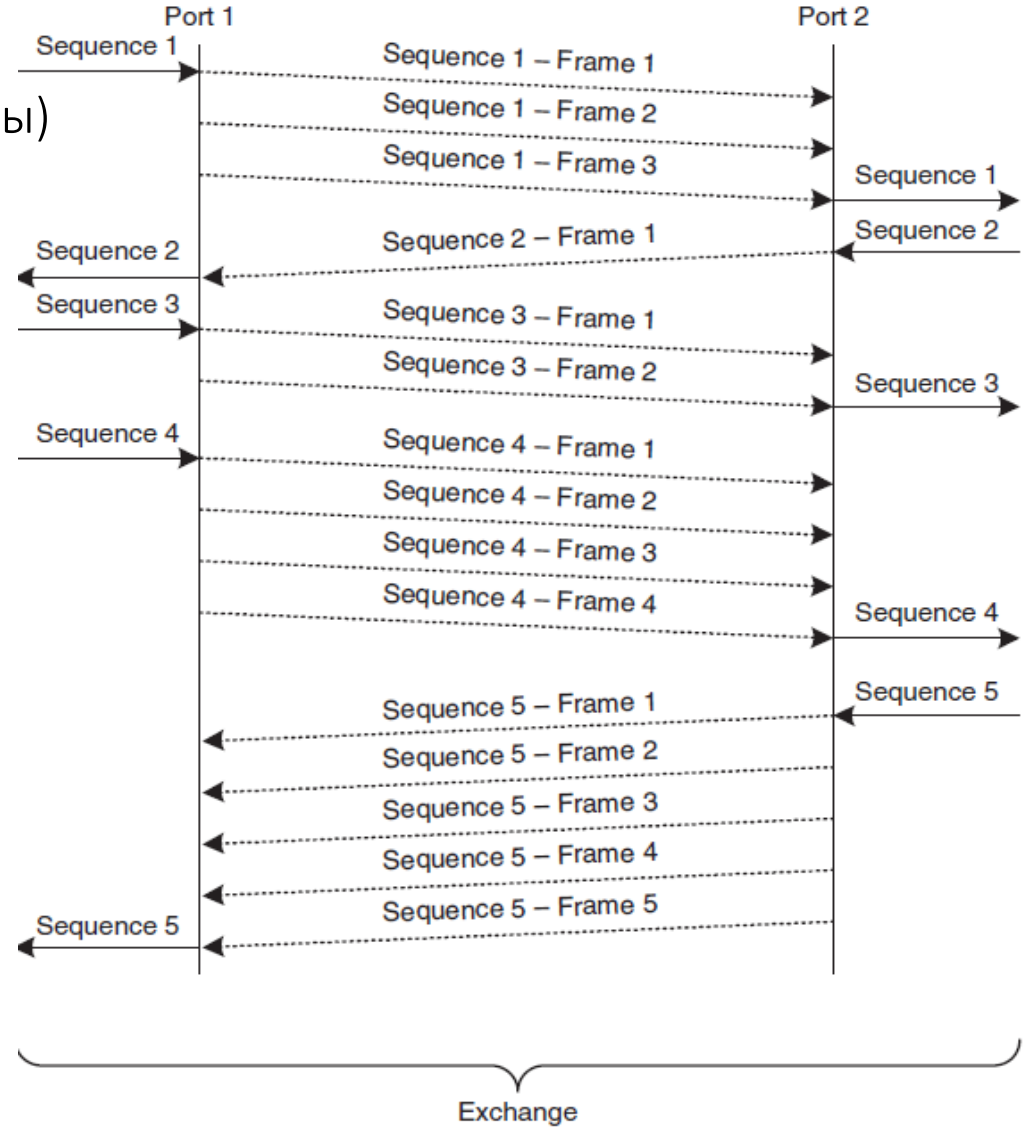
Т а б л и ц а А.1 –Специальные коды

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: передача данных

- Структура передачи данных данных
  - Exchange – сессия между логическими сущностями (процессы)
  - Sequence – последовательность кадров
  - Frame – управления и данных (2 112 Б)
- Управление потоком
- Классы сервиса





FC-2: Классы обслуживания

Класс 1 Соединение точка-точка (end-to-end) между портами типа n_port через коммутацию каналов. Класс удобен для аудио и видео приложений, например, видеоконференций. После установления соединения используется вся доступная полоса пропускания канала. При этом гарантируется, что кадры будут получены в том же порядке, в каком они были посланы. Есть управление потоком.

Класс 2 Без установления соединения с коммутацией пакетов, гарантирующий доставку данных. Порт может взаимодействовать одновременно с любым числом портов типа n_port в режиме дуплекс. Не гарантируется порядок доставки кадров, кроме соединения P2P или KcA. Есть управление потоком. Этот класс характерен для локальных сетей, где время доставки данных не является критическим.

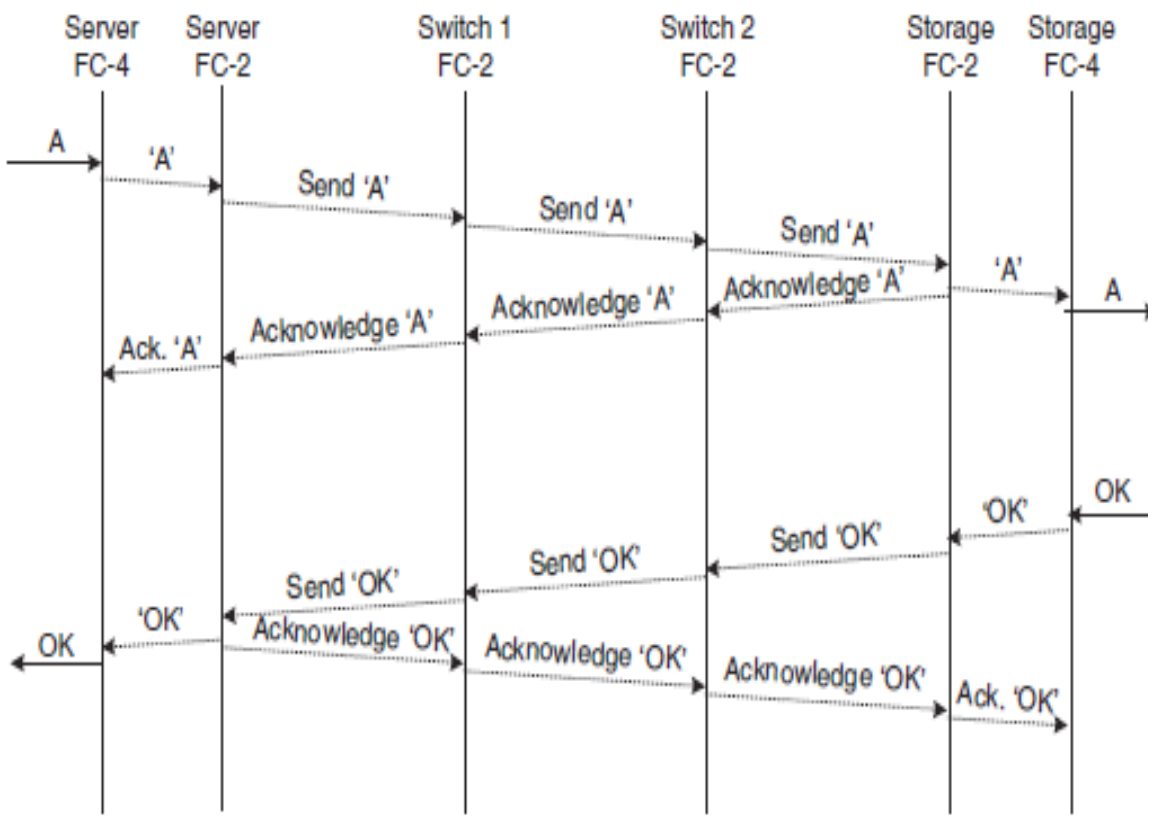
Класс 3 Обмен дейтограммами без установления соединения и без гарантии доставки. Есть управление потоком. Применяется для каналов SCSI.

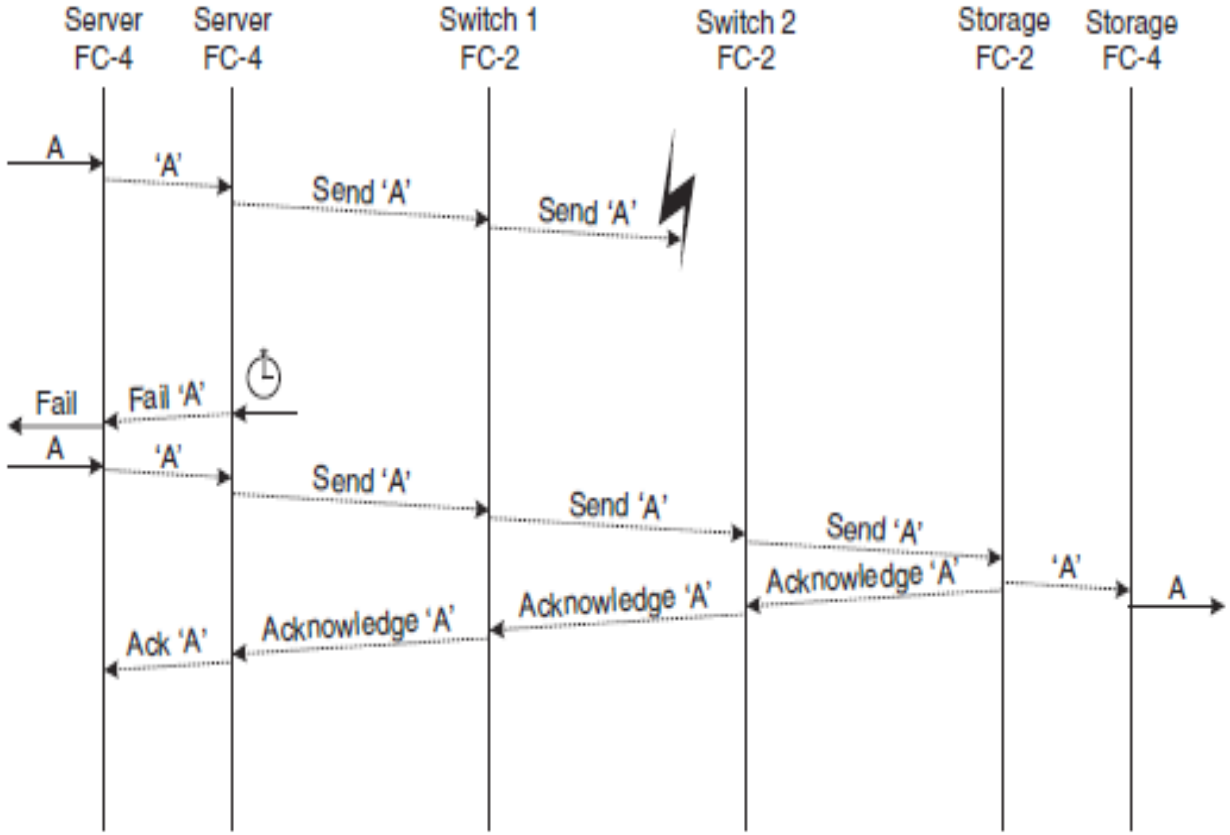
Класс 4 Обеспечивает выделение определенной доли пропускной способности канала с заданным качеством обслуживания (QoS). Только для топологии матрица с n_port. Гарантируется порядок доставки кадров.

Класс 5 Регламентирующие документы находятся в процессе подготовки.

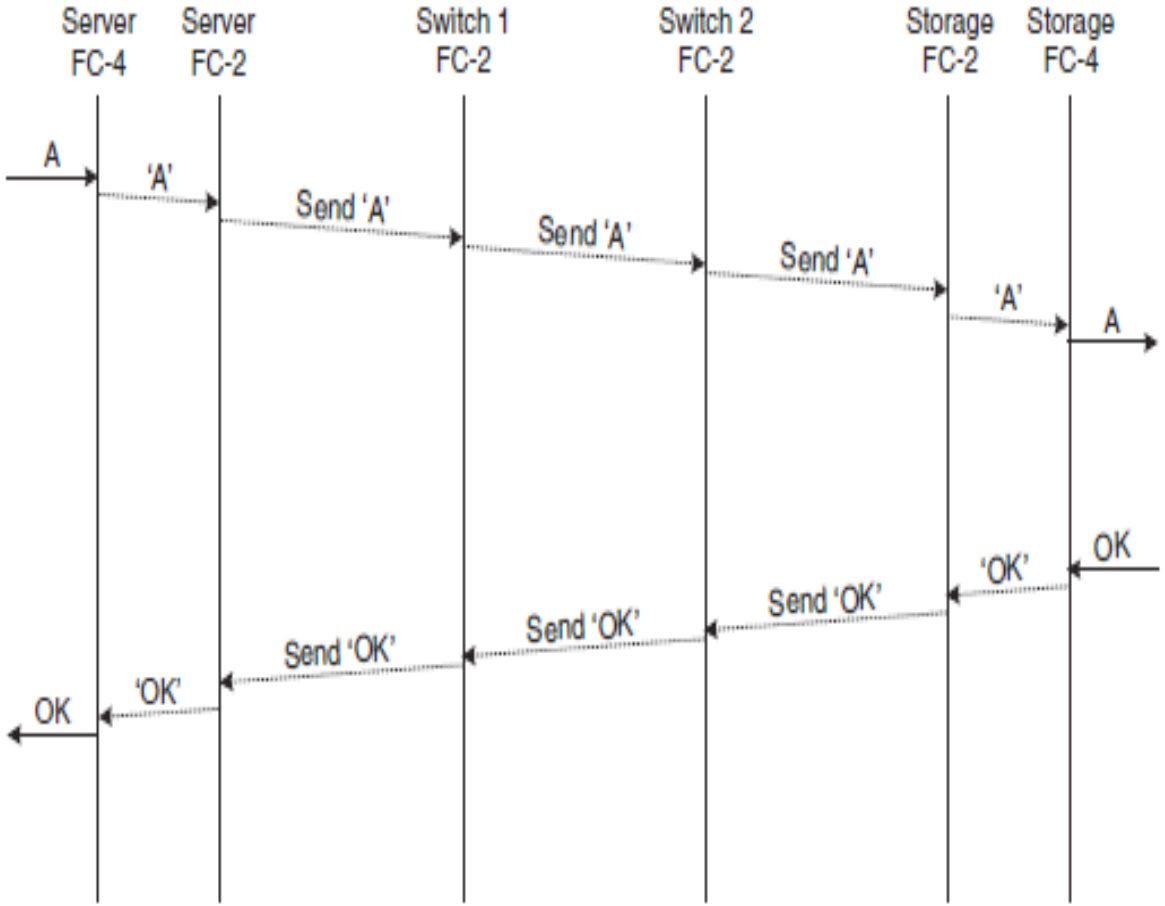
Класс 6 Предусматривает групповое-обслуживание с коммутацией.

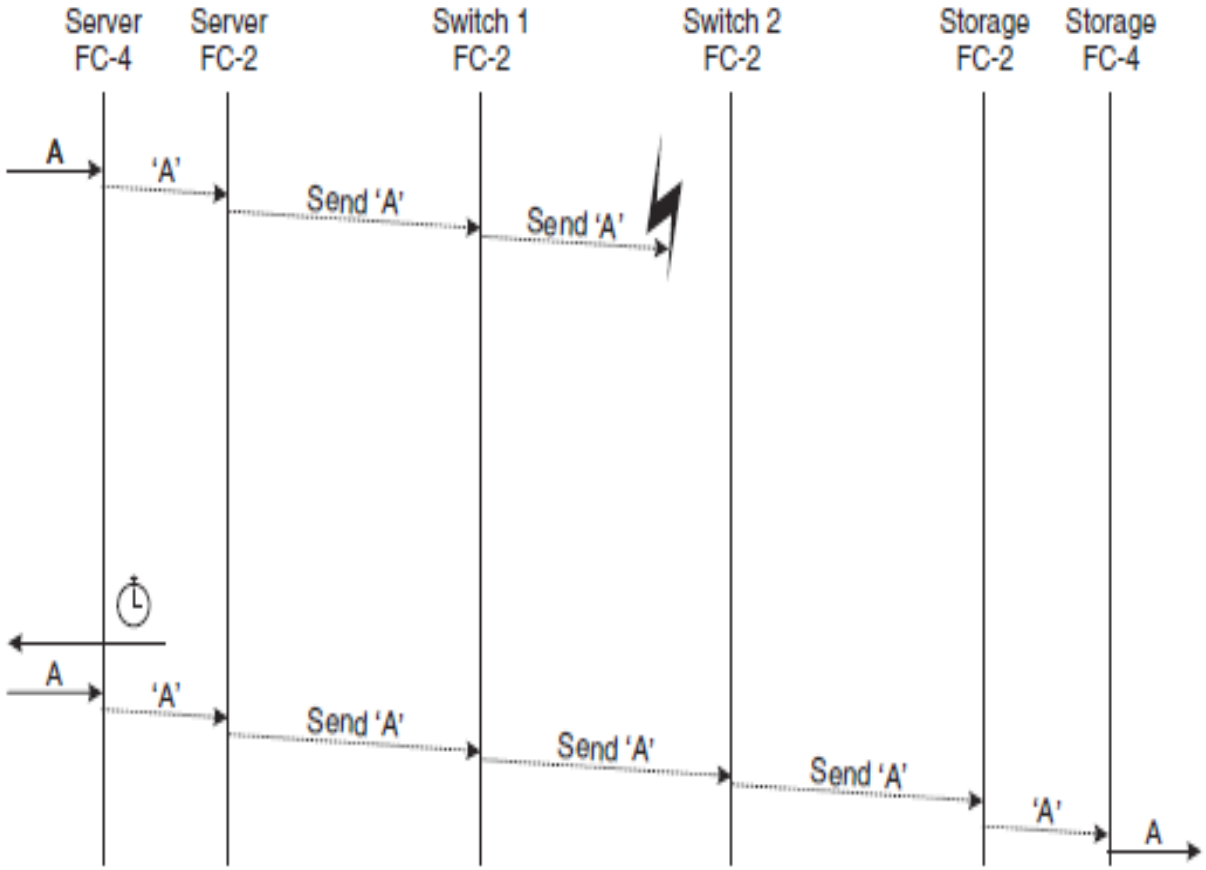
FC-2 класс 2


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: класс 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

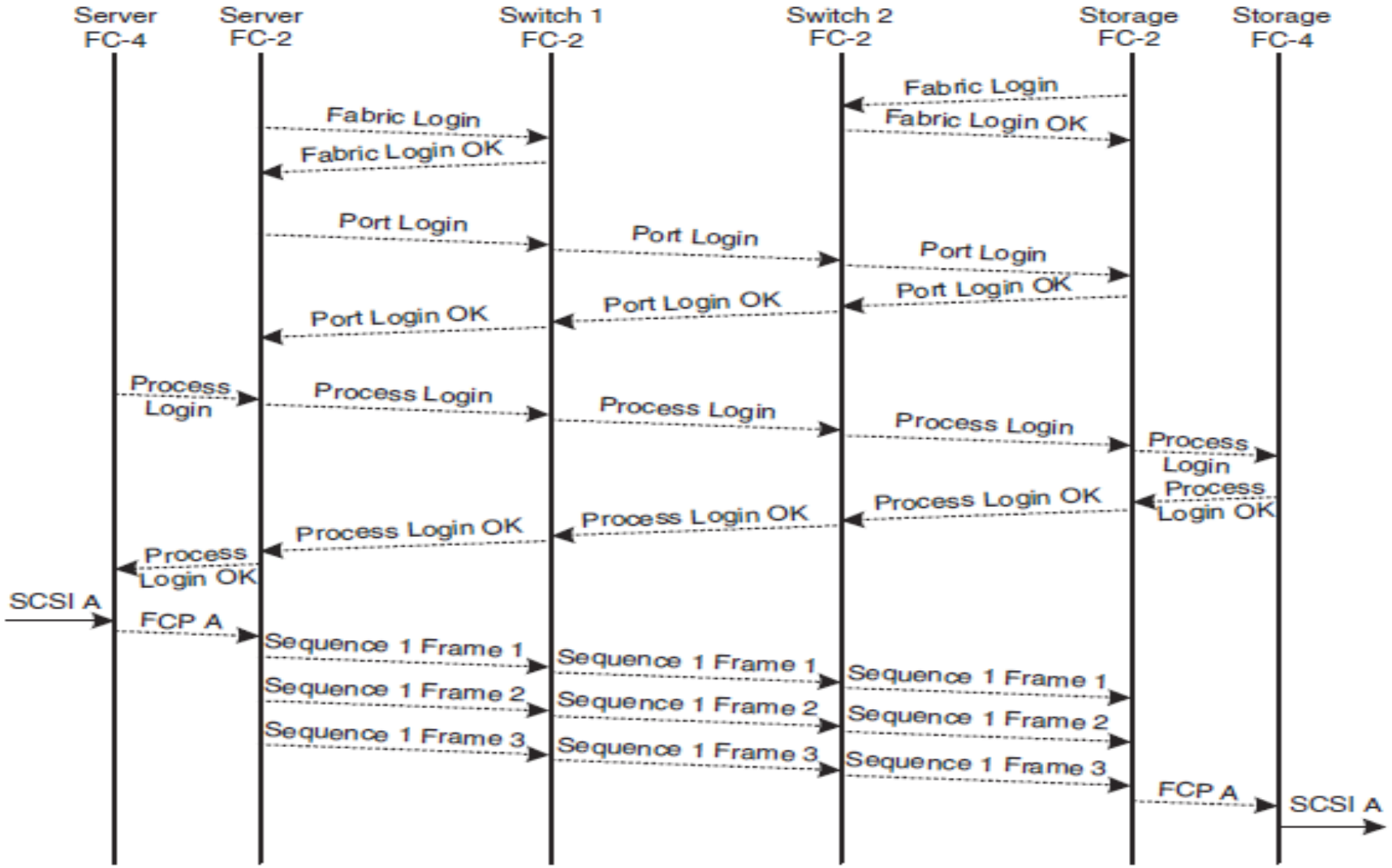

Server FC-4
Server FC-2
Switch 1 FC-2
Switch 2 FC-2
Storage FC-2
Storage FC-4
A
'A'
Send 'A'
Send 'A'
A
'A'
Send 'A'
Send 'A'
Send 'A'
'A'
A

FC-3: сервисы

- Распределение кадров по маршрутам между много портовыми устройствами для увеличения пропускной способности
- Формирование логических групп маршрутов для управления переполнением на маршруте или сбоя, чтобы не загружать верхние уровни стека.
- Компрессия передаваемых данных (на HBA)
- Шифрование данных (на HBA)
- Зеркалирование

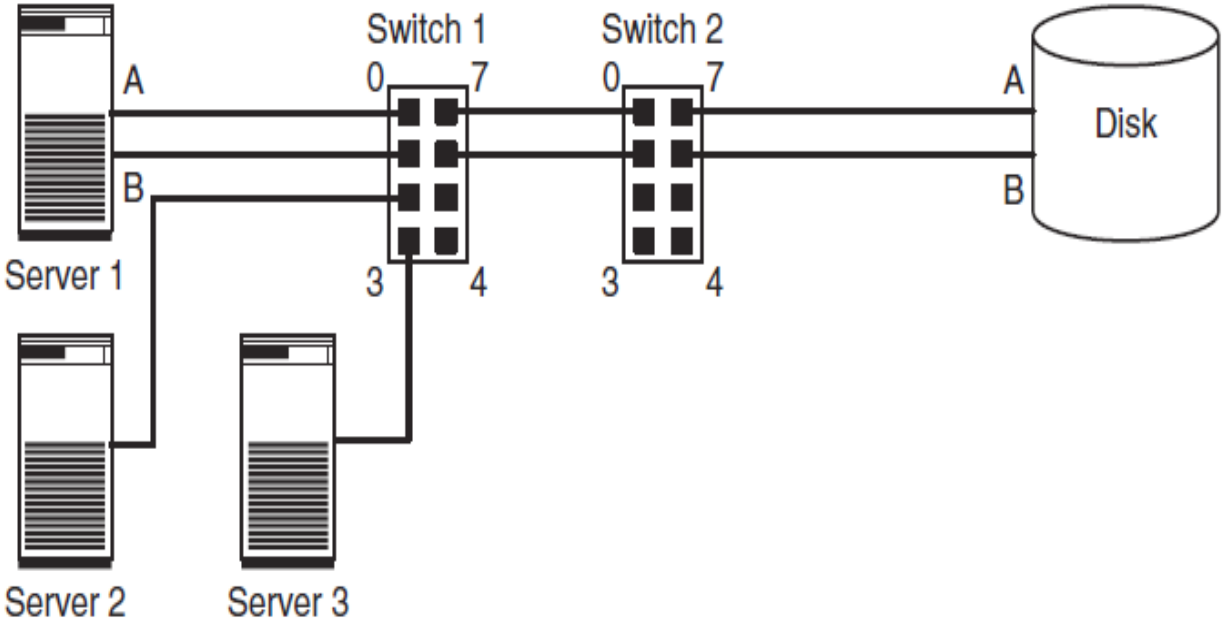
Пока это в планах

Службы линии: идентификация и адресация



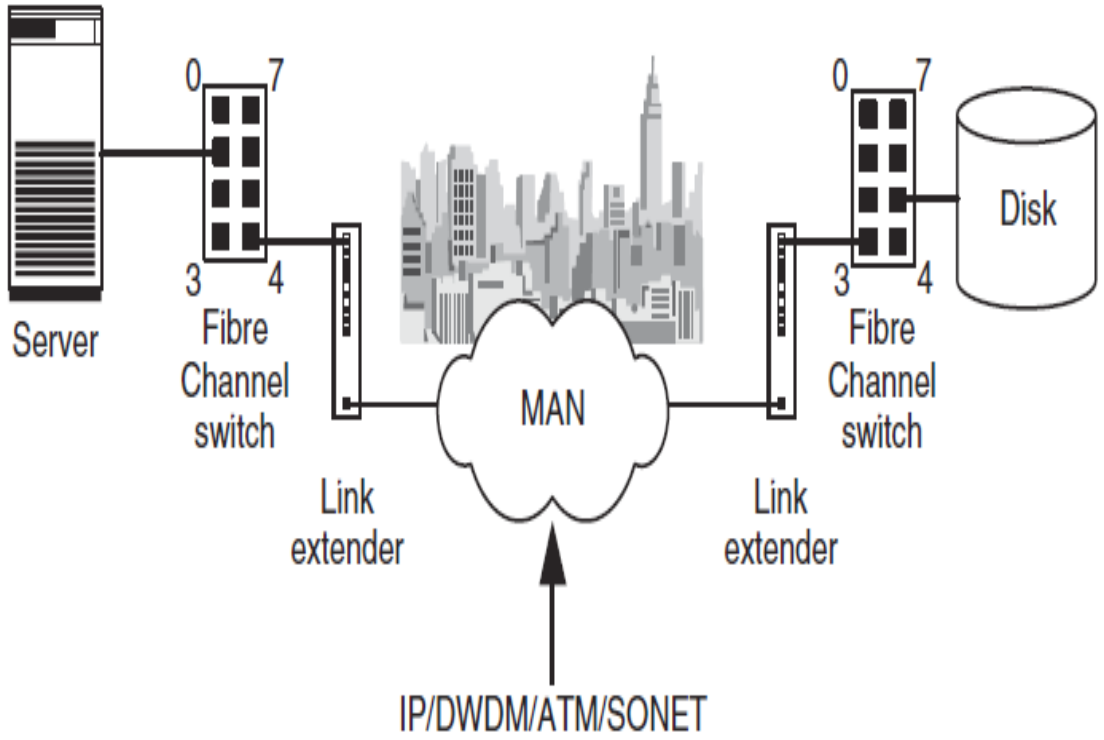
Адресация

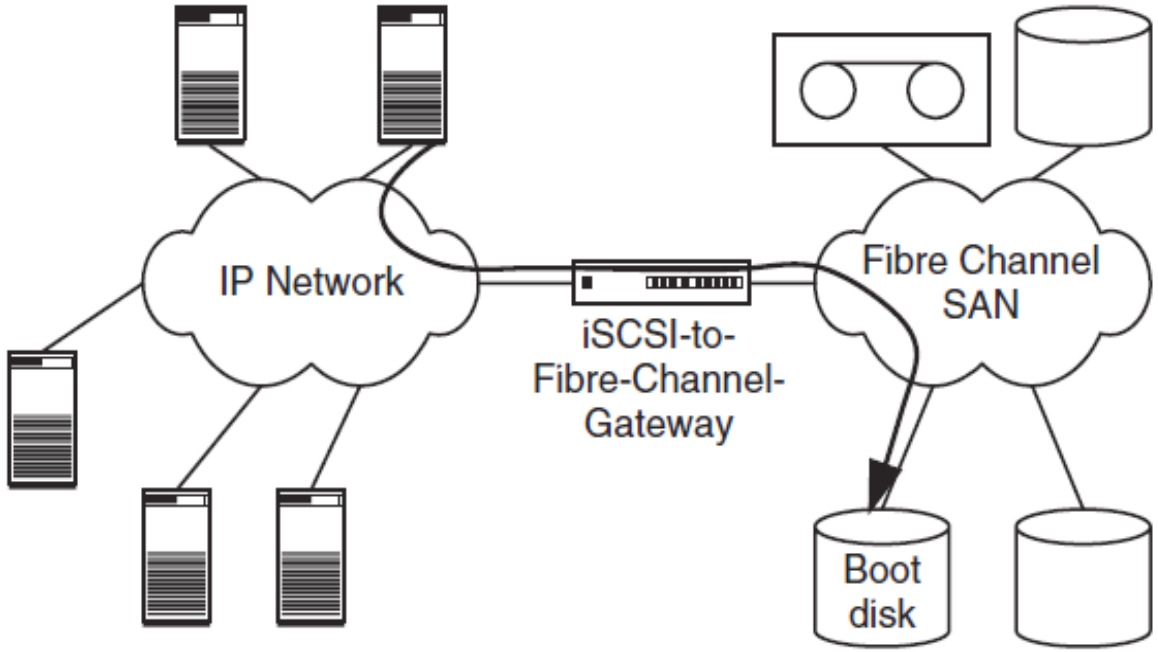
- Имена и адреса в FC
- У всех устройств FC сети есть 64 бит. имена
- WWN vs FCN
- WWN: WWPN. WWNN
- FLOG – 24 bit port address
- S_ID vs D_ID
- KcA 8 bit AL_PA (Arbitrated Loop Physical Address)



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

Metro SAN и IP_FC SAN





ВОРОСЫ?

http://arccn.ru/

smel@arccn.ru

+7 (495) 240-50-63

@ArccnNews