Введение в сети хранения данных

Данные

- Рост объемов
- Децентрализация
- Необходимость масштабирования
- Стоимость
- Надежность
- Безопасность
- Сложность управления

Характеристики Систем Хранения Данных

- Объем
- Механизм доступа
- Скорость доступа
- Отказоустойчивость
- Доступность
- Безопасность
- Сложность управления

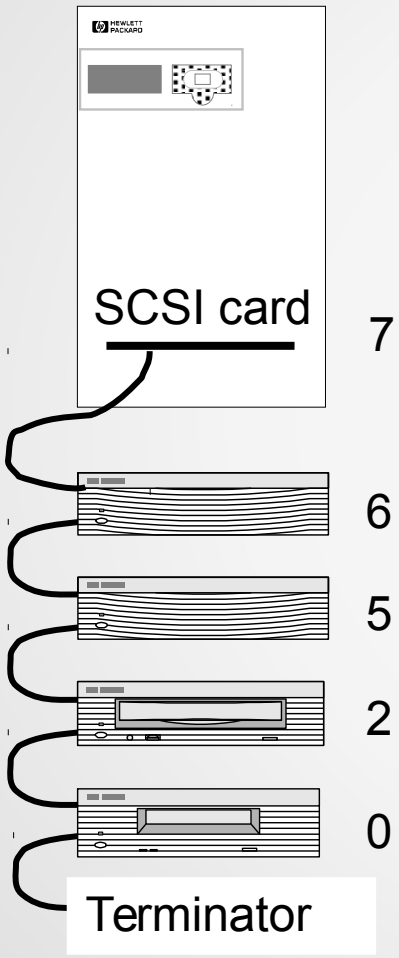
Отказоустойчивость

- Задачи
  - Сохранность данных
  - Обеспечение доступности
- Методы обеспечения отказоустойчивости:
  - Дублирование узлов
  - Избыточность
    - RAID
    - MultiPath

SCSI

SCSI (англ. Small Computer Systems Interface, произносится скази) — интерфейс, разработанный для объединения на одной шине различных по своему назначению устройств, таких как жёсткие диски, накопители на магнитооптических дисках, приводы CD, DVD, стримеры

Концепции и адресация SCSI устройств

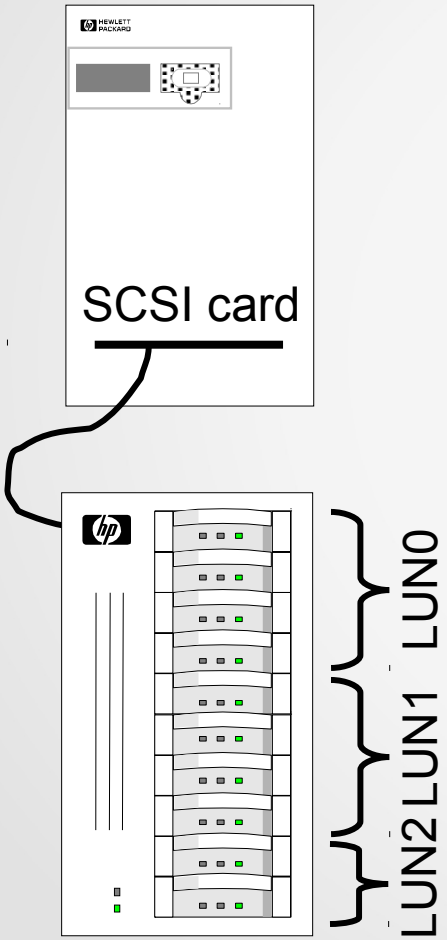


- стандарта SCSI
  - SE single-ended,
  - LVD low-voltage-differential — интерфейс дифференциальной шины низкого напряжения (+большая скорость)
  - HVD high-voltage-differential — интерфейс дифференциальной шины высокого напряжения (+большое растояние)
- Типы шины
  - Узкий" ("Narrow") 8-битные данные
  - "Широкий" ("Wide") 16-битные данные
- **SCSI цепочки**
- **SCSI терминаторы**
- **SCSI таргет адреса**

7 (более высокий приоритет) ----------> 0 ----> 15 ----------> 8 (более низкий приоритет)

Адресация SCSI устройств

HOST . CHANNEL . TARGET . LUN



SCSI Lun (0x00 … 0xff)

SCSI Target (0x00 … 0xff)

SCSI-Bus per HBA (0x00/0x01 for FC)

SCSI/FC HBA (0x00, …) sometimes called SCSI host scsi0...scsiX used in various commands

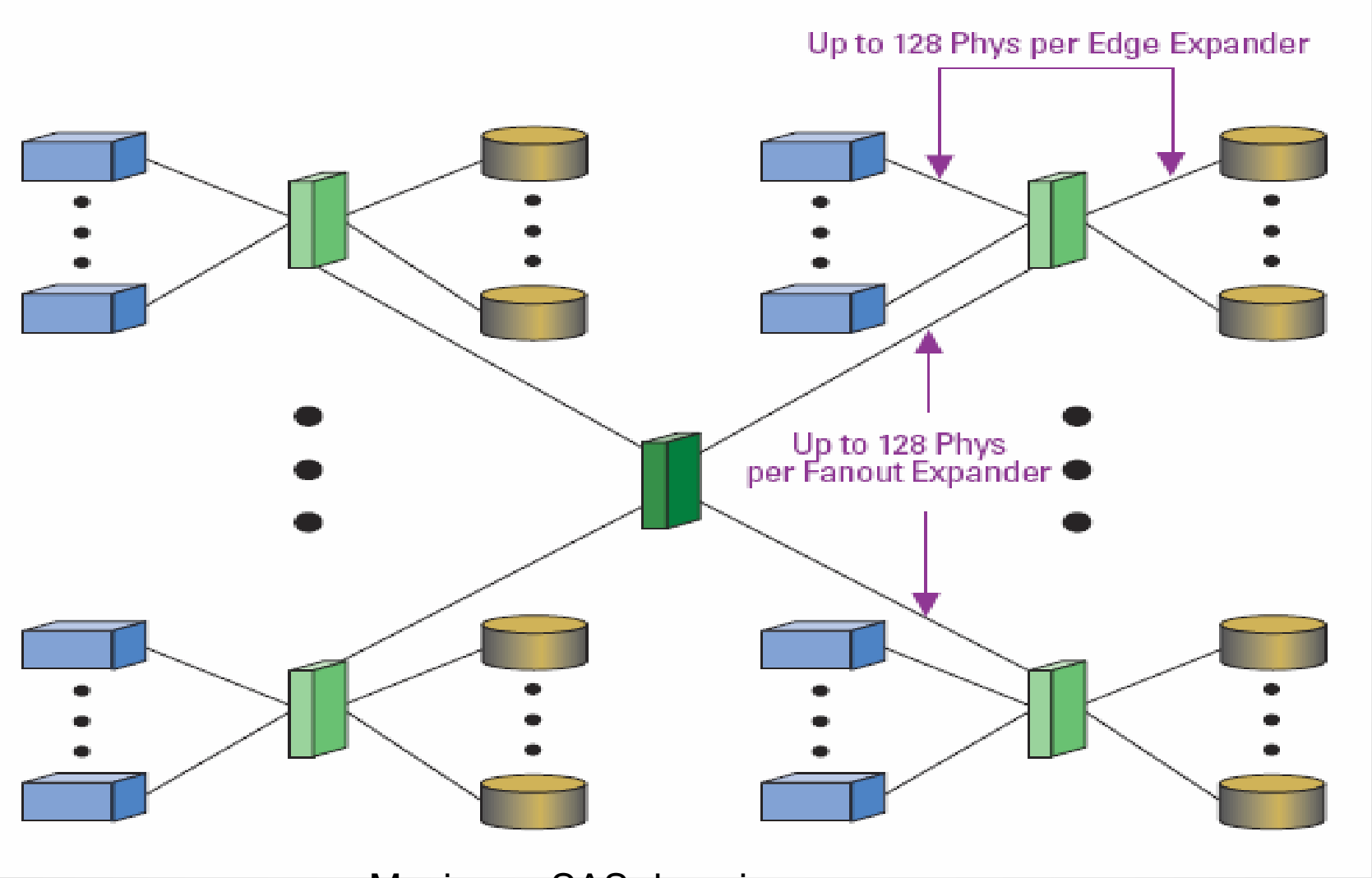
Serial Attached SCSI (SAS)

- компьютерный интерфейс, разработанный для обмена данными с такими устройствами, как жёсткие диски, накопители на оптическом диске и т. д.
- SAS использует последовательный интерфейс для работы с непосредственно подключаемыми накопителями (англ. Direct Attached Storage (DAS) devices).
- SAS разработан для замены параллельного интерфейса SCSI и позволяет достичь более высокой пропускной способности, чем SCSI.
- управления SAS-устройствами используются команды SCSI.

Сравнение SAS и параллельного SCSI

- SAS использует последовательный протокол (меньшее количество сигнальных линий)
- Интерфейс SCSI использует общую шину. SAS использует соединения точка-точка
- SAS не нуждается в терминации шины
- SAS поддерживает большое количество устройств (> 16384)
- SAS поддерживает высокие скорости передачи данных (1,5, 3,0 или 6,0 Гбит/с)

Комутация устройств в SAS



Maximum SAS domain

Методы подключения дискового пространства

- Прямое подключение к хранилищу DAS
- Сетевая система хранения NAS
- Сеть хранения данных SAN

DAS

DAS (Direct Attached Storage) — решение, когда устройство для хранения данных подключено непосредственно к серверу, либо к рабочей станции. Устройства хранения могут быть подключены по одному из интерфейсов: SCSI, FC или SAS.

В случае этой архитектуры отсутствует централизованное управления ресурсами и возможность разделить ресурсы между серверами.

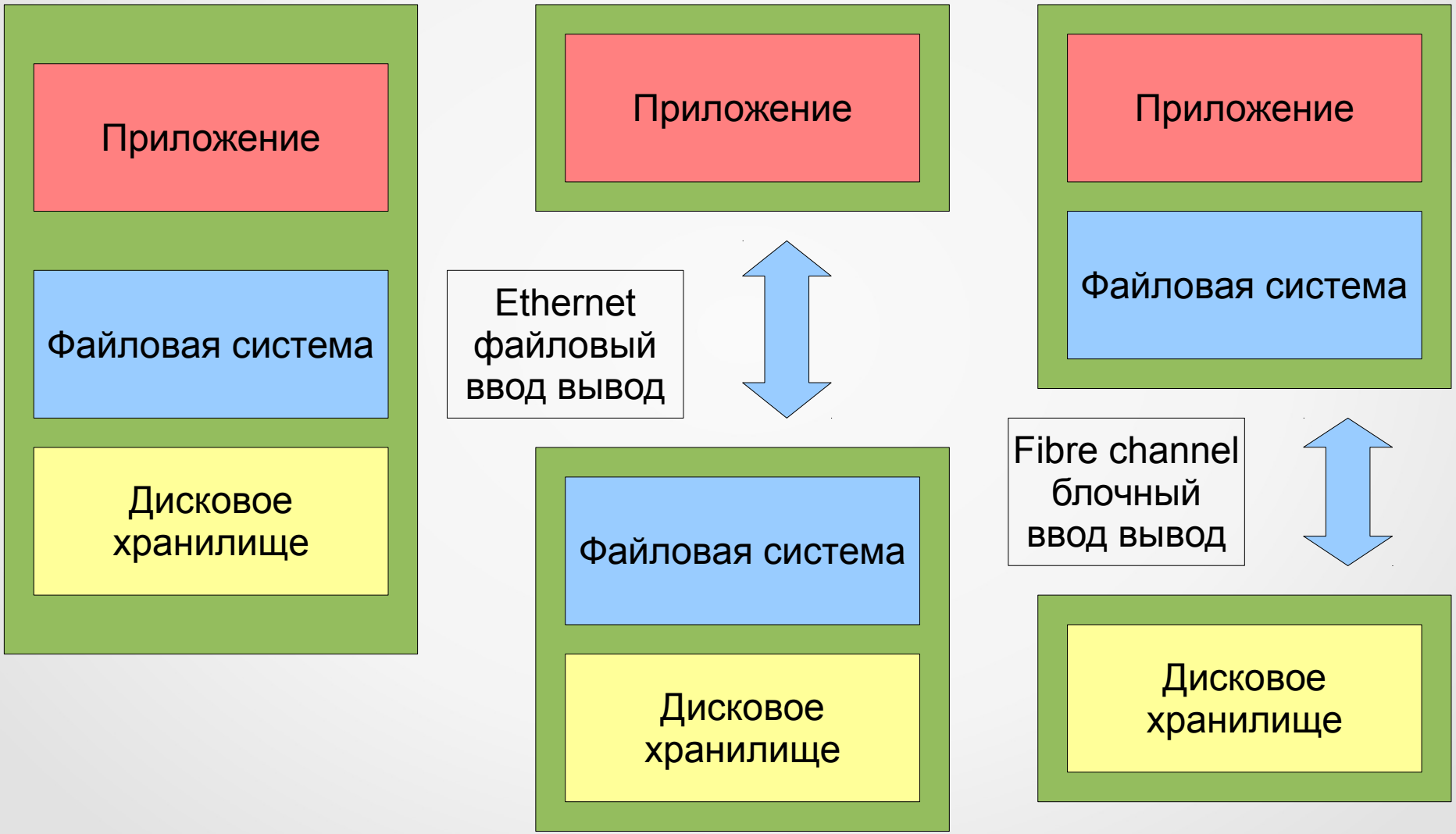
NAS

- NAS (англ. network attached storage) — сетевая система хранения данных.
- используют сетевые протоколы для доступа к файлам (такие как NFS или SMB/CIFS)
- хранилище является удалённым и компьютер запрашивает файл вместо того, чтобы запрашивать блок данных с диска.

SAN

- Storage Area Network (SAN) - это высокоскоростная коммутируемая сеть передачи данных, объединяющая серверы, рабочие станции, дисковые хранилища и ленточные библиотеки.
- Для обмена данными чаще всего используется протокол Fibre Channel.
- Fibre Channel оптимизирован для быстрой гарантированной передачи сообщений и позволяет передавать информацию на расстояние от нескольких метров до сотен километров.

DAS NAS SAN

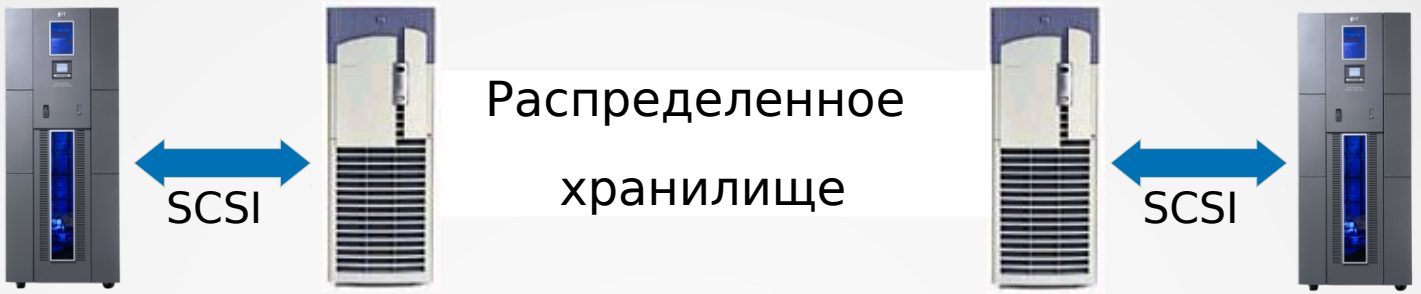

Приложение
Файловая система
Дисковое хранилище
Приложение
Ethernet файловый ввод вывод
Файловая система
Дисковое хранилище
Приложение
Файловая система
Fibre channel блочный ввод вывод
Дисковое хранилище

SAN

Storage Area Network

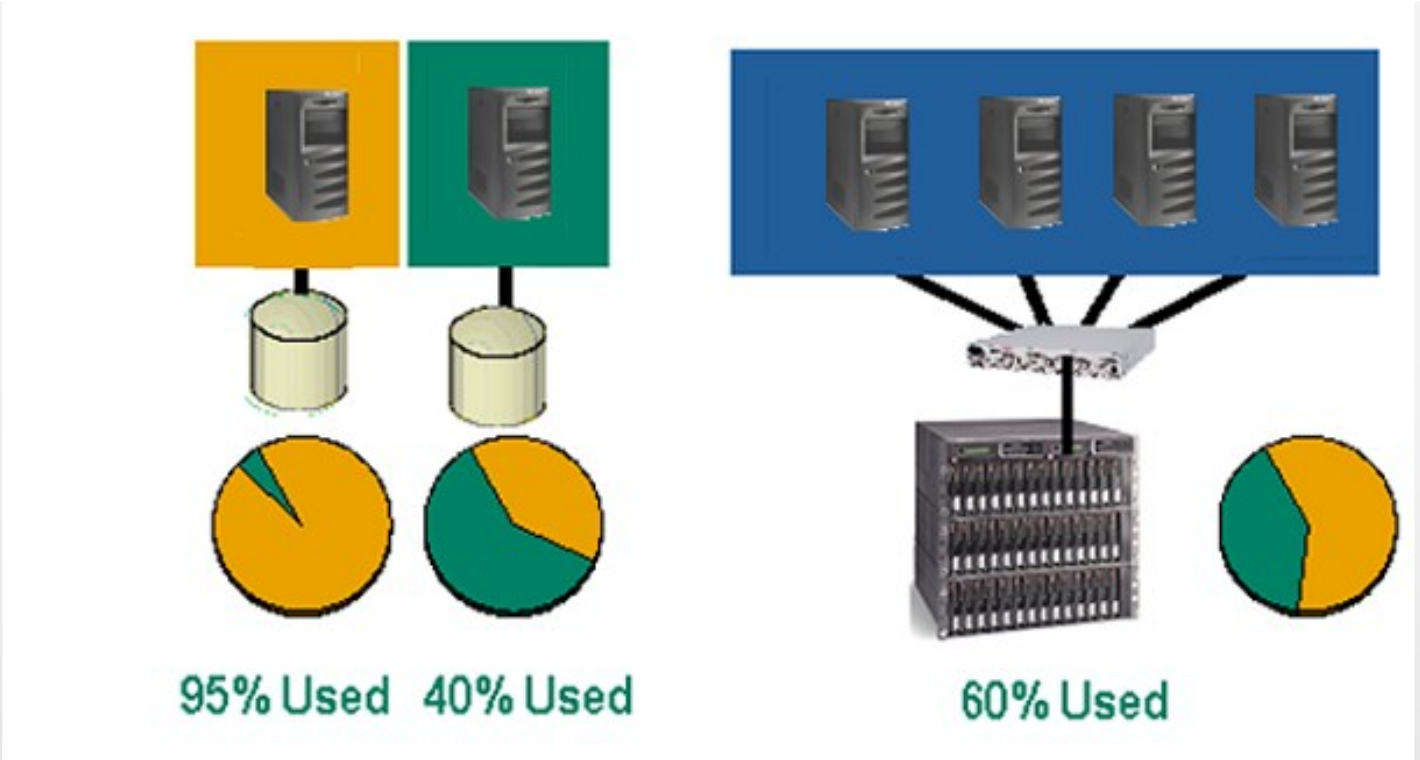
- Доступ к устройствам (RAW)
- Децентрализация
- Поставщики и потребители объединены сетью
- Возможность использования одного устройства несколькими потребителями

Консолидация серверов и систем хранения

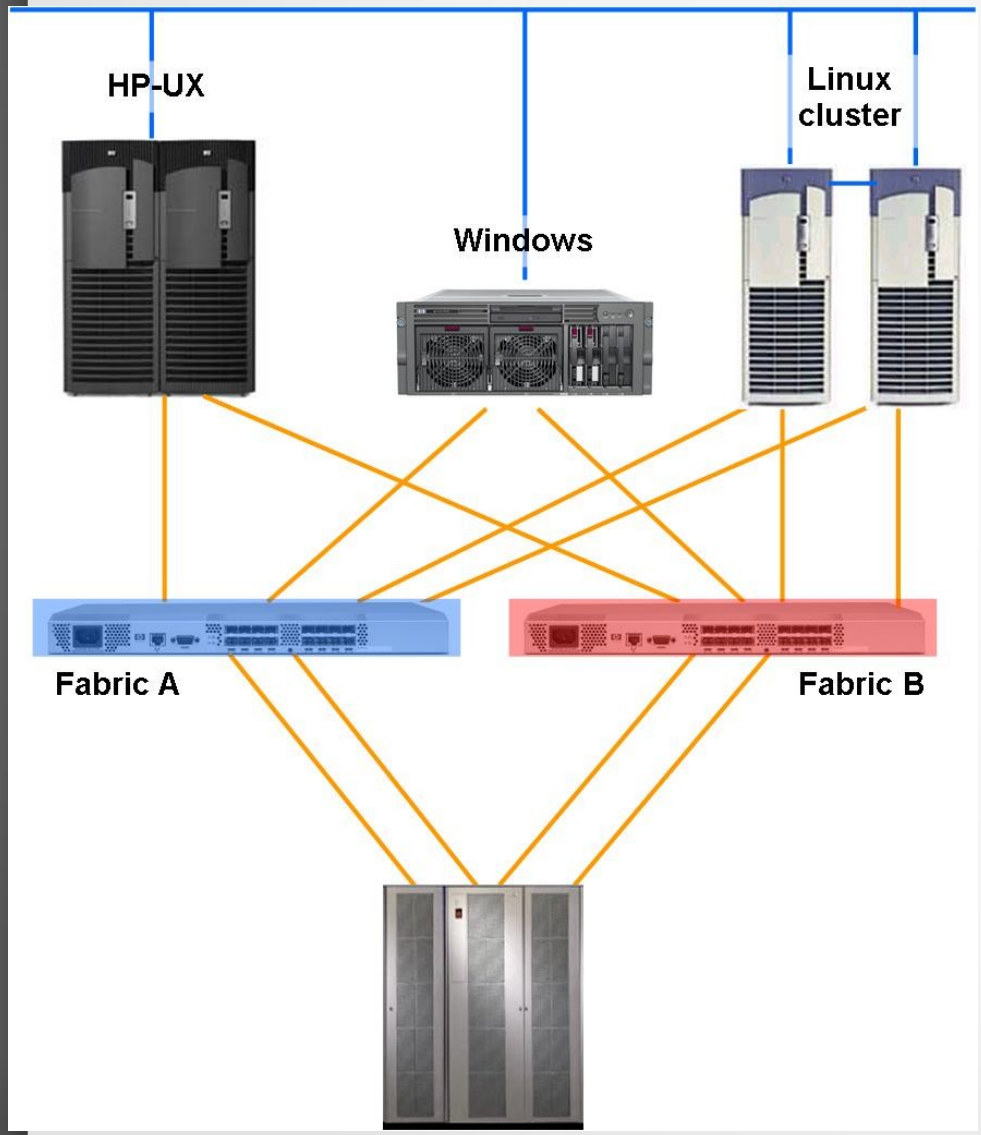


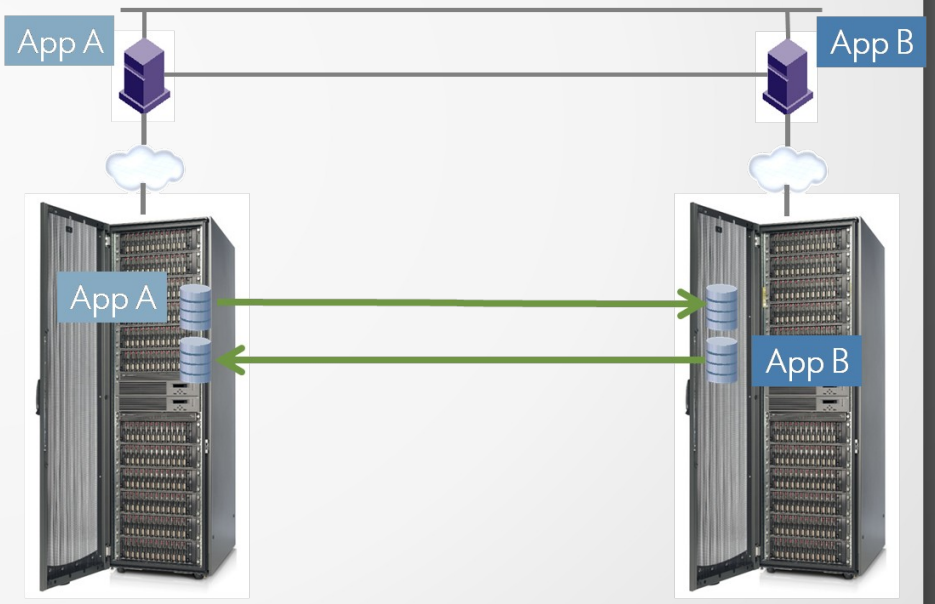
Эффективное управление ёмкостью

- Эффективное использование объема
- Меньше устройств — проще управлять



Высокая Доступность





Компоненты SAN

- Коммутаторы
- Маршрутизаторы, мосты и шлюзы
- Устройства хранения Disk array (target)
- Серверы Host (initiator)
- Среда передачи

Fibre Channel switch

Router

Disk System

Cable

Server

Тип сети SAN

Физические интерфейсы:

- Ethernet
- FibreChannel

Протоколы:

- ATA over Ethernet
- iSCSI (Internet Small Computer Systems Interface)
- FC
- iFCP (Internet Fibre Channel Protocol)
- FCIP (Fibre Channel over TCP/IP)

Fibre Channel

- Fibre Channel или FC — высокоскоростной интерфейс передачи данных, используемый для взаимодействия рабочих станций, мейнфреймов, суперкомпьютеров и систем хранения данных.
- Топология: Порты устройств могут быть подключены
  - напрямую друг к другу (point-to-point) FC-P2P
  - в управляемую петлю (arbitrated loop) FC-AL
    - публичная петля (public loop)
    - частная петля (private loop)
  - в коммутируемую сеть, называемую «тканью» (англ. fabric. Часто на сленге просто «фабрика») FC_SW
- Можно различать топологию по двум критериям
  - есть ли цикл
  - есть ли комутатор

loop	fabric	topology
yes	no	private (arbitrated) loop
yes	yes	public loop
no	no	direct point-to-point
no	yes	switched point-to-point (*)

Структура и заголовок FC фрейма

SEQ_ID
Unique ID allocated to any given sequence within a specific exchange

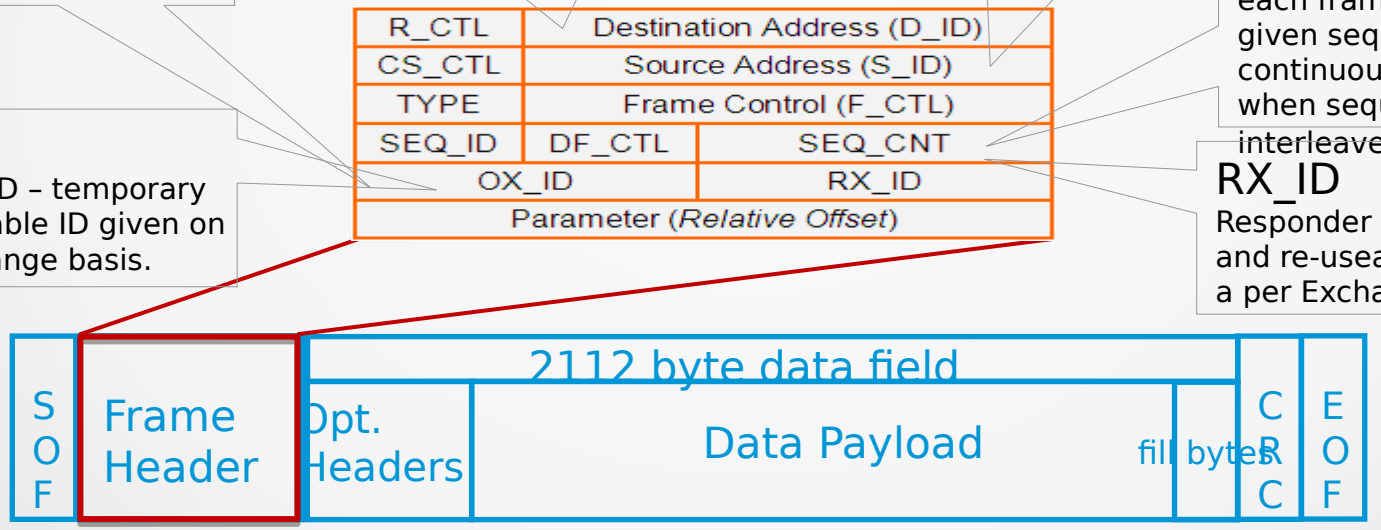
D_ID
24 bit FC_ID of destination port

S_ID
24 bit FC_ID of transmitting port

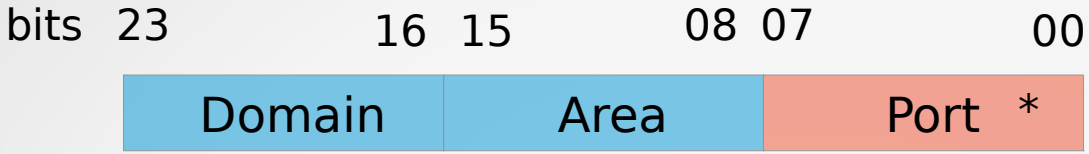
SEQ_CNT
Incremented by x0001 for each frame sent within a given sequence. May continuously increment when sequences are interleaved.

OX_ID
Originator ID – temporary and re-useable ID given on a per Exchange basis.

RX_ID
Responder ID – temporary and re-useable ID given on a per Exchange basis.



Fibre Channel адресация (для FC-SW)



FC-SW	Domain id of the Switch	Port number on the switch	Vendor specific entry*
FC-AL	Domain id of the Switch	Port number on the switch	AL-PA of the NL port

24 bit FC_ID address field

R_CTL	Destination Address (D_ID)	
CS_CTL	Source Address (S_ID)	
TYPE	Frame Control (F_CTL)	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter (*Relative Offset*)		

Frame Header

* ***Vendor specific field FC-SW***

Switch vendor	***Port* field entry**
Brocade	00
McData	13

Обмен, последовательности и кадр на примере SCSI операции запись

Initiator

Target

Exchange

FCP_CMND_W

Frame1, Sequence1

FCP_XFER_RDY

Frame1, Sequence2

FCP_DATA

FCP_DATA

FCP_DATA

FCP_DATA

FCP_DATA

Frame1, Sequence3

Frame2, Sequence3

Frame3, Sequence3

Frame4, Sequence3

Frame5, Sequence3

FCP_RSP
Status Good

Frame1, Sequence4

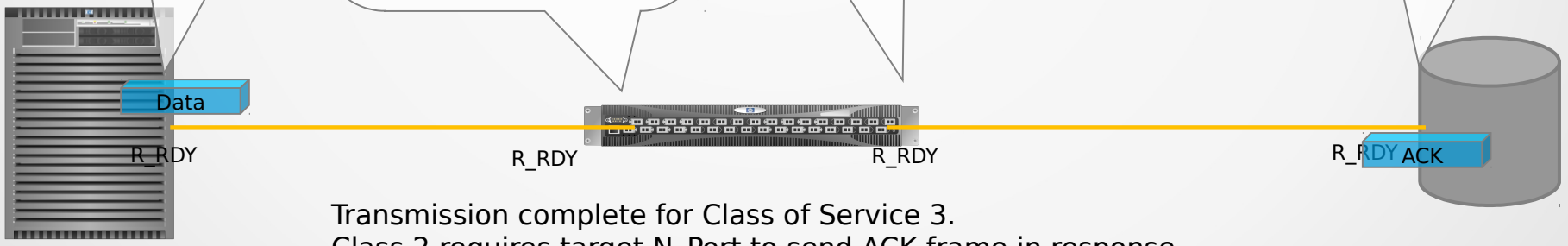
FC контроль передачи

N_Port transmits Data Frame to disk target. Decrements F_Port buffer credit count by one.

F_Port clears occupied buffer. Sends R_RDY to transmitting N_Port to increment buffer count.

F_Port sends Data Frame on to link and decrements target N_Port buffer credit count by one.

N_Port receives Data Frame and sends R_RDY to transmitting F_Port to increment it’s buffer count.

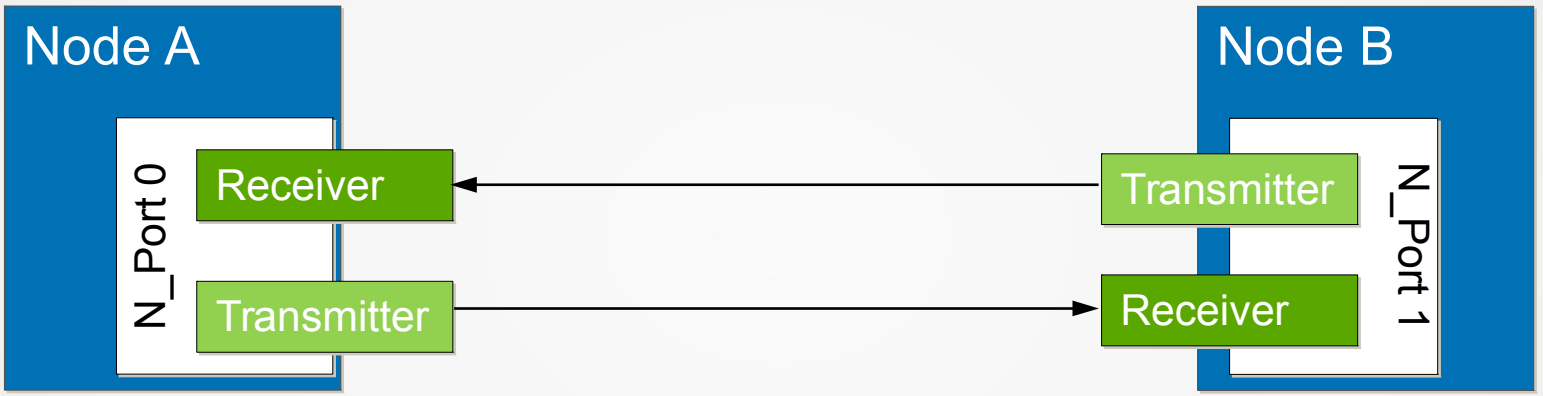


Transmission complete for Class of Service 3.
Class 2 requires target N_Port to send ACK frame in response…
Class 2 frame transmission complete:

- Buffer to Buffer Flow Control

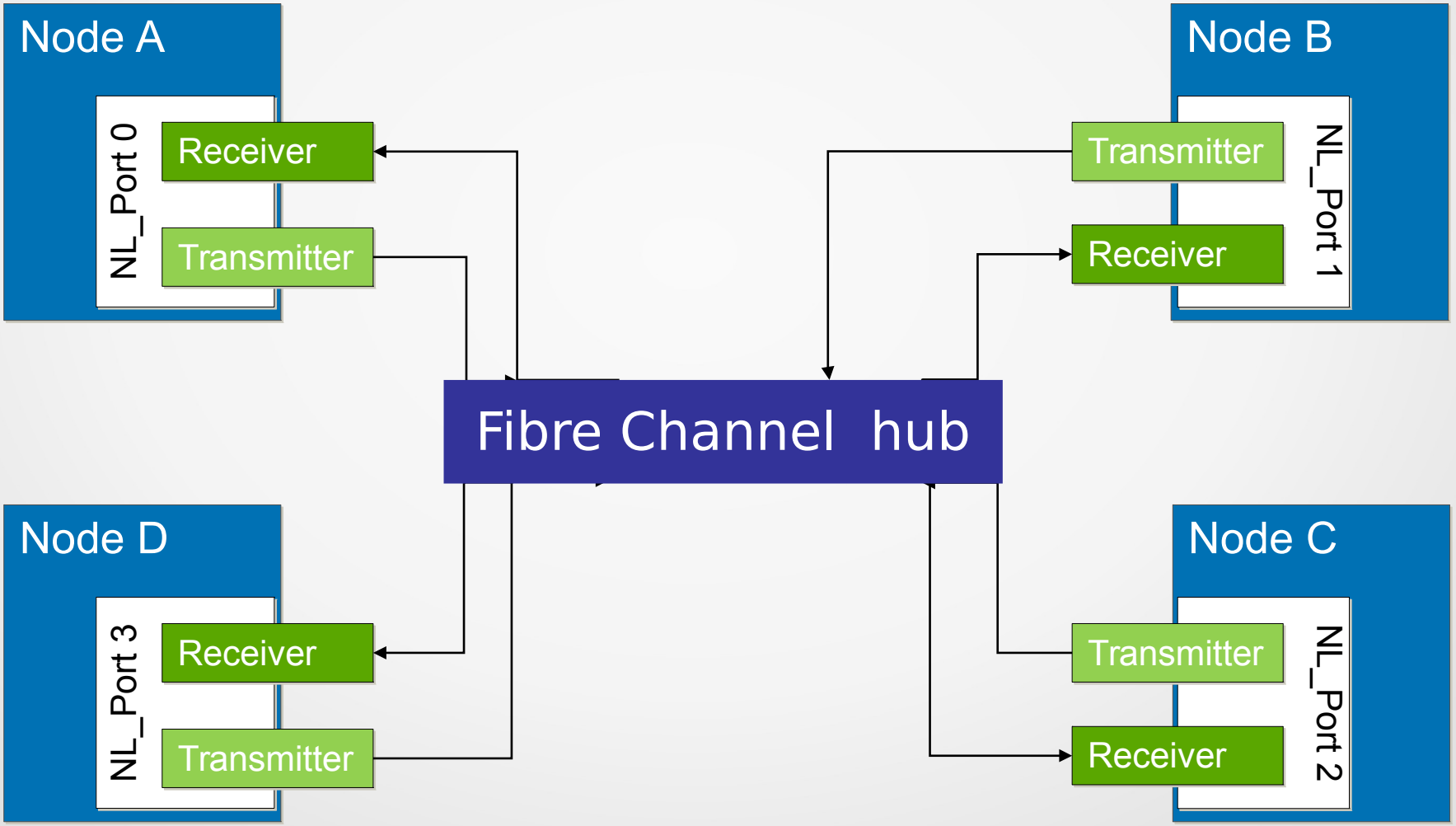
  AND
- End to End Flow Control

прямое подключение (point-to-point) FC-P2P

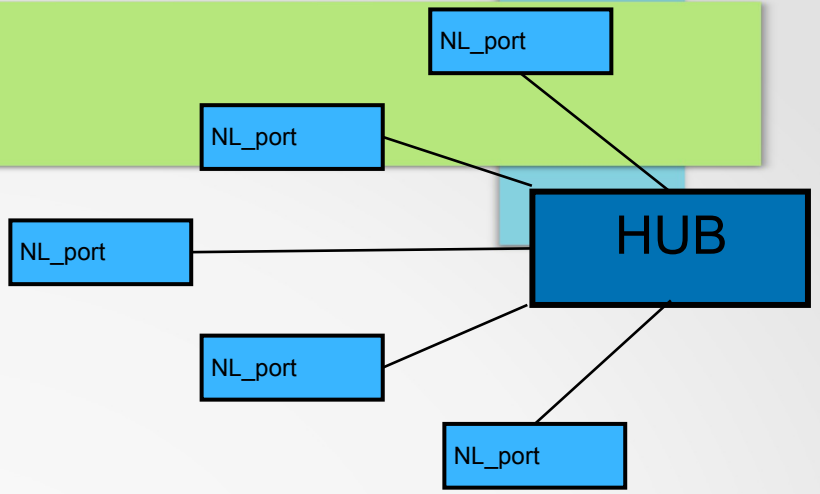


+дешего +монопольное использование канала
- комутация только двух устройств

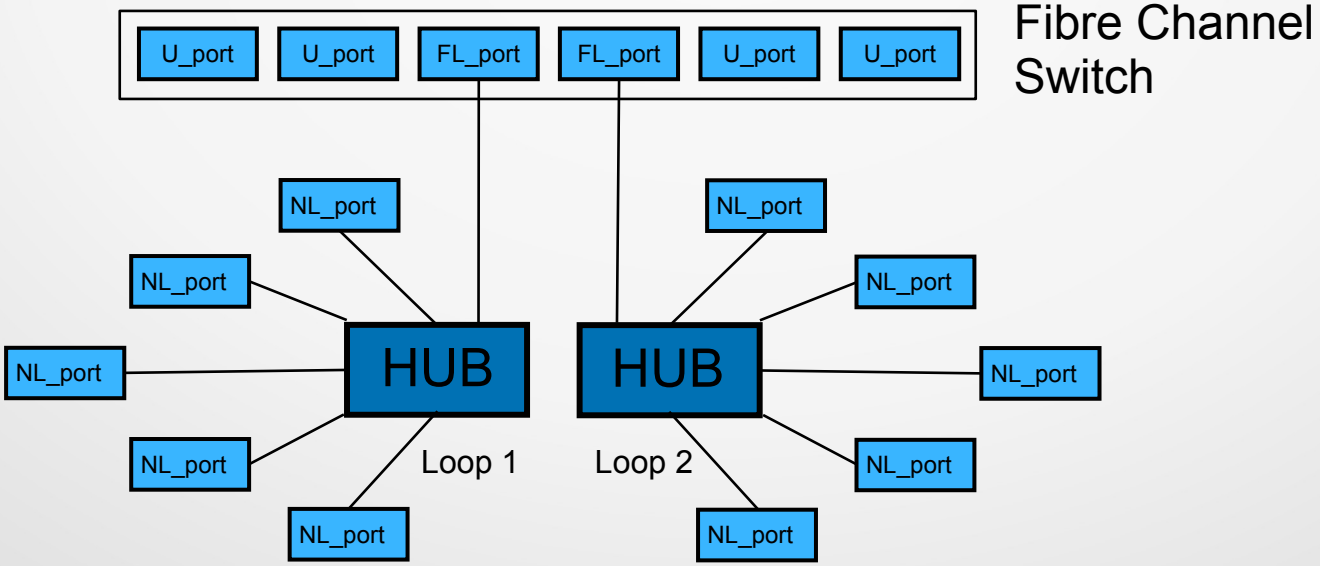
управляемая петля (arbitrated loop) FC-AL



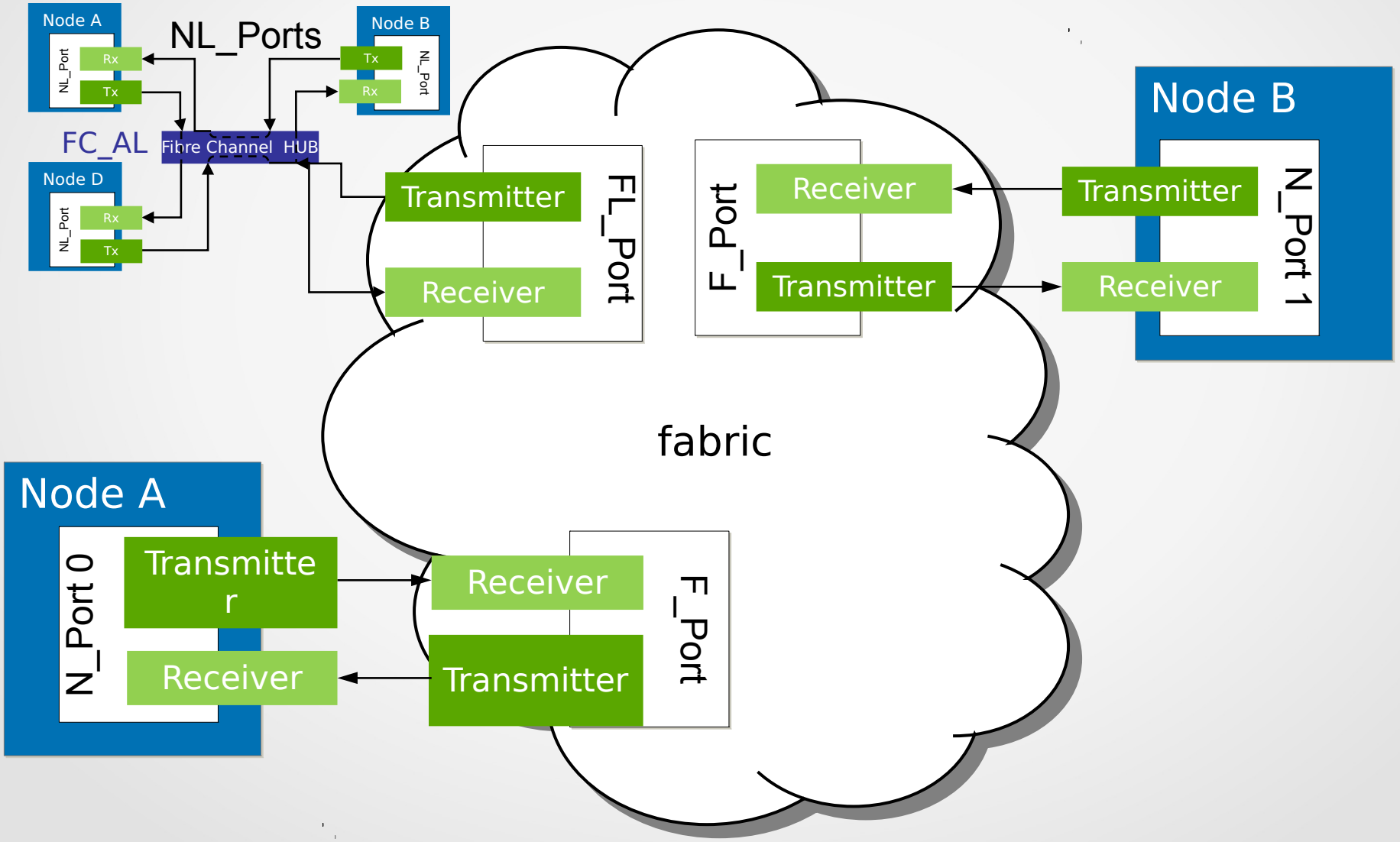
Частная петля (Private loop)



Публичная петля (Public loop)

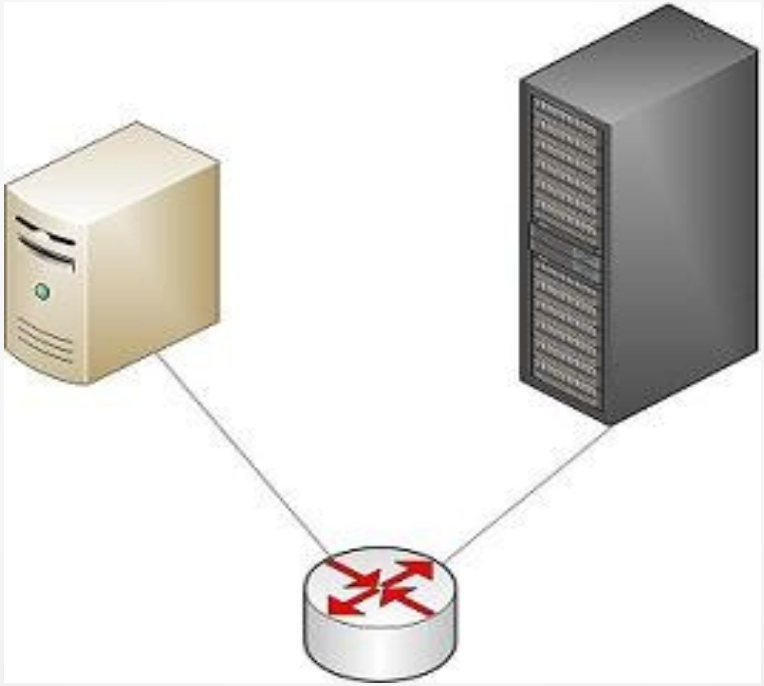


коммутируемая сеть, «ткань» («фабрика») FC_SW

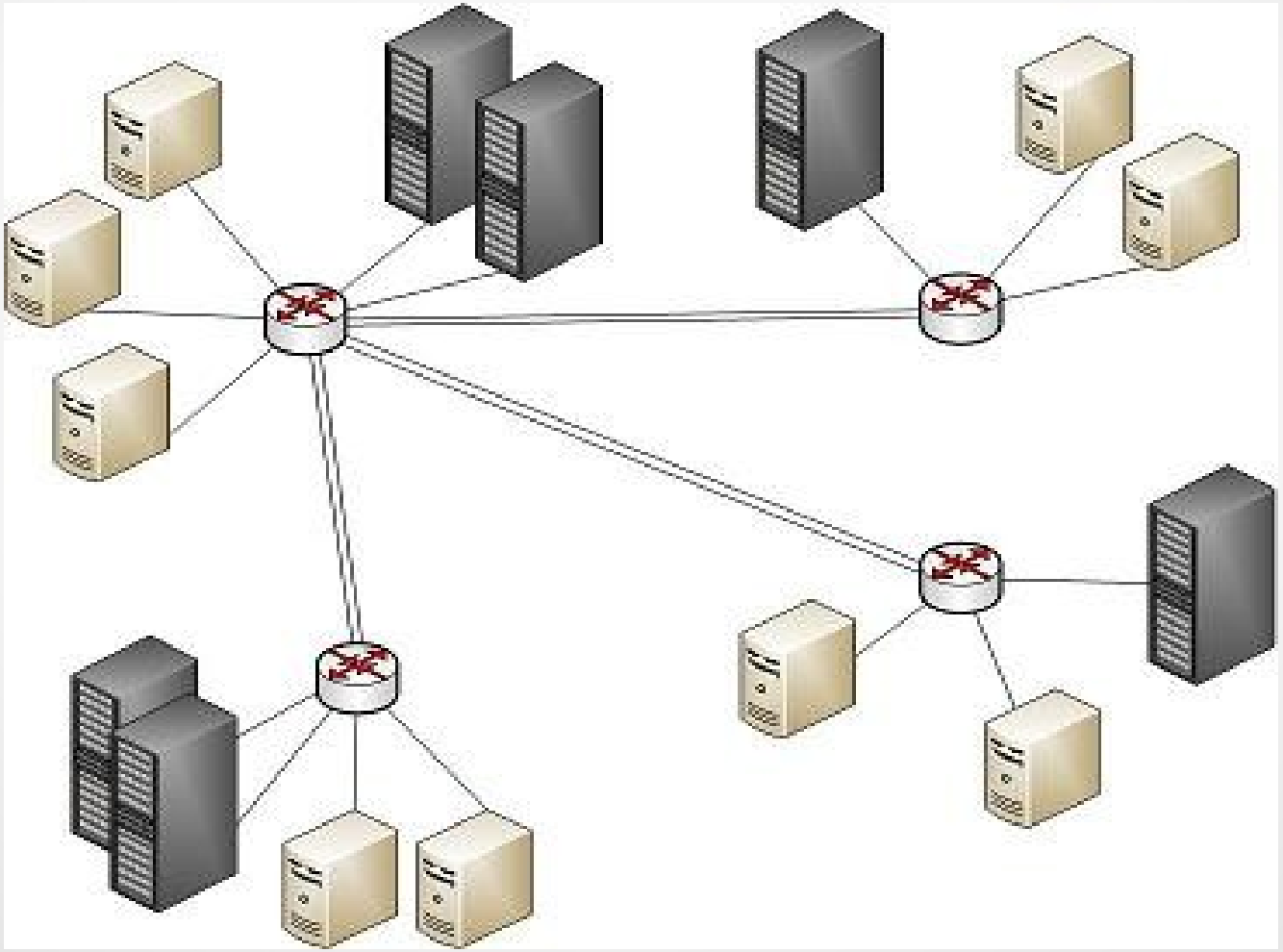


Различные топологии «ткани» («фабрики»)

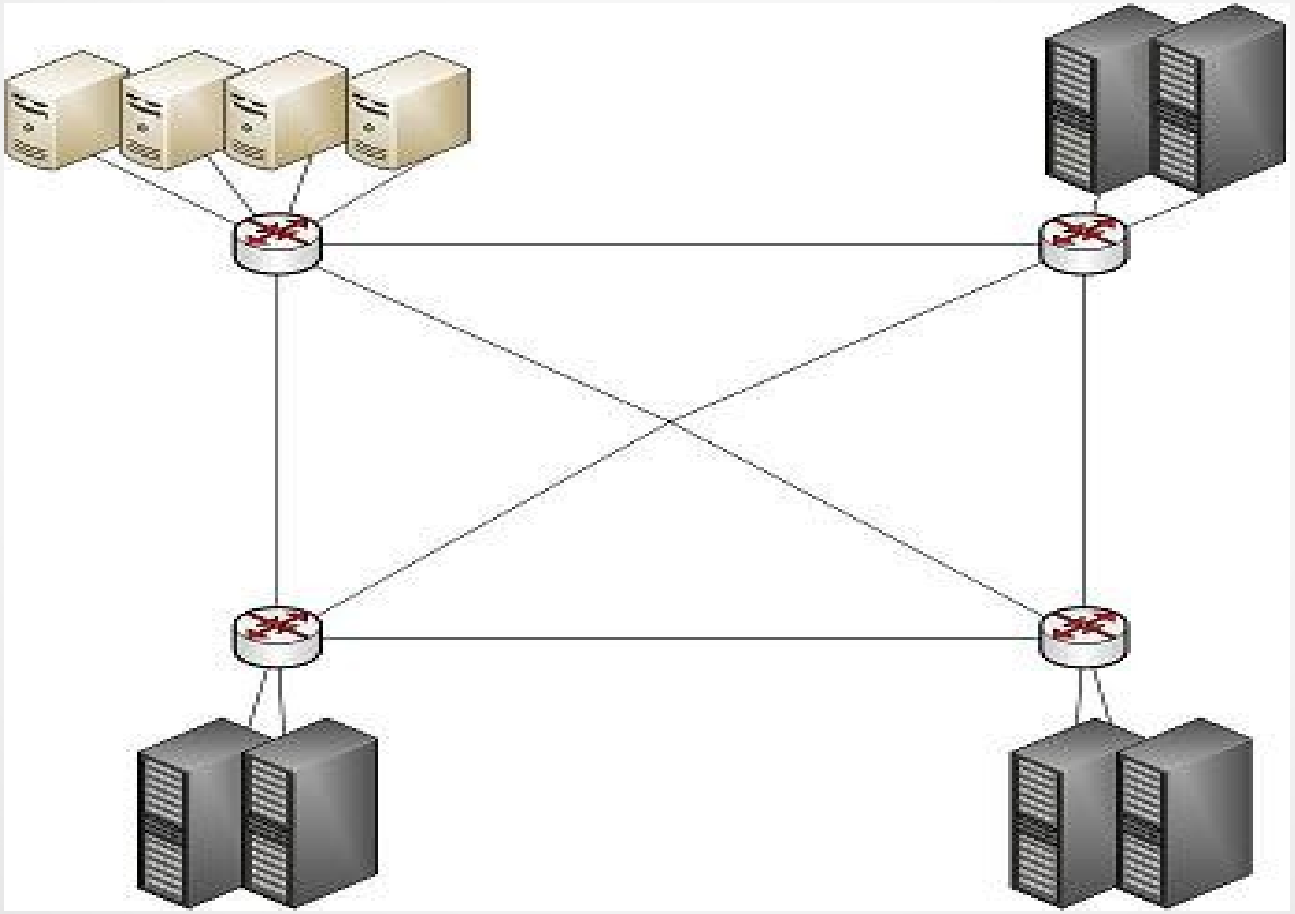
«Одно-коммутаторная» структура
Single-switch fabric



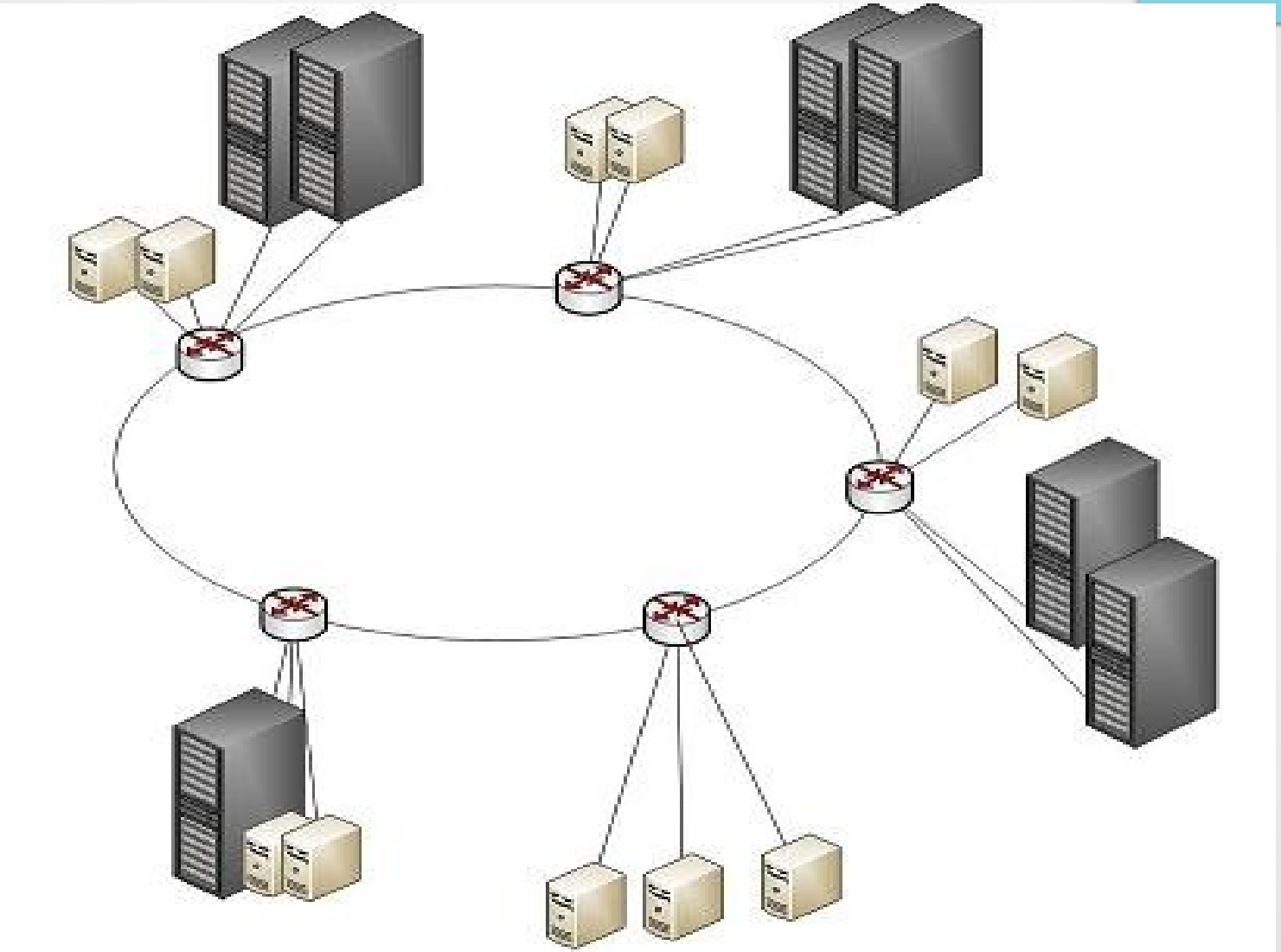
Древовидная или Каскадная структура Cascaded fabric



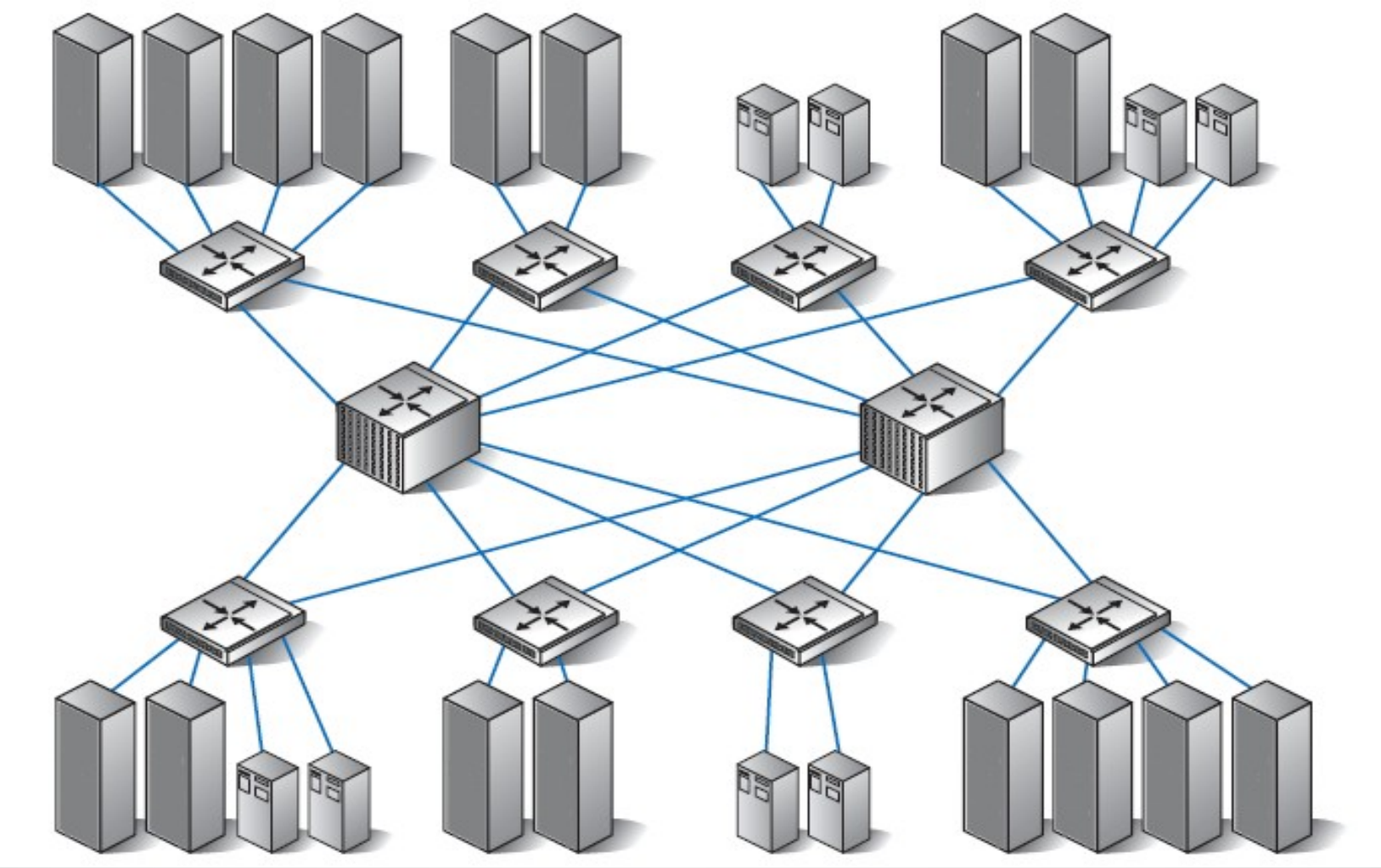
Решётка
Meshed fabrics



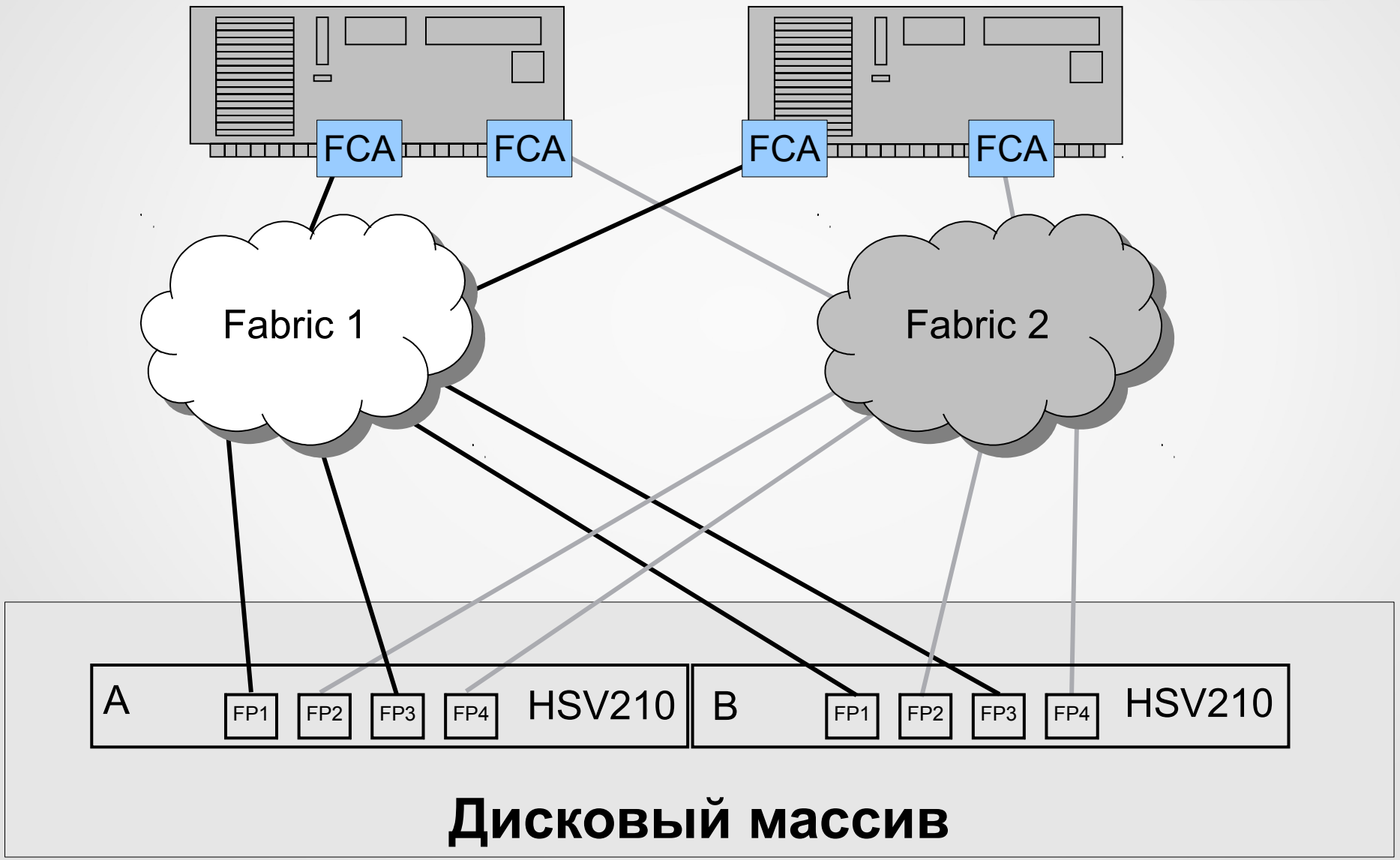
Кольцо
Ring fabric



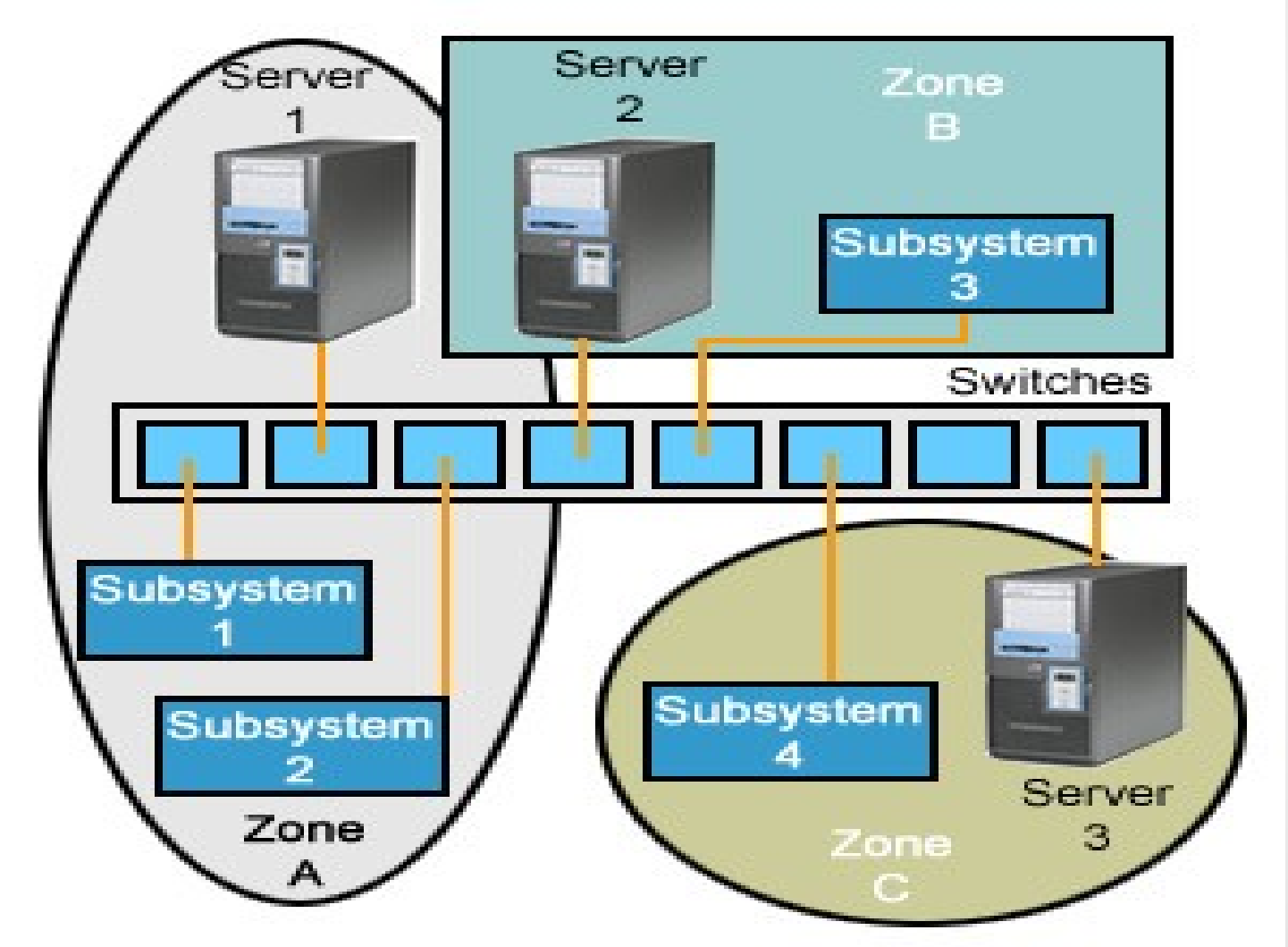
Core-edge fabric



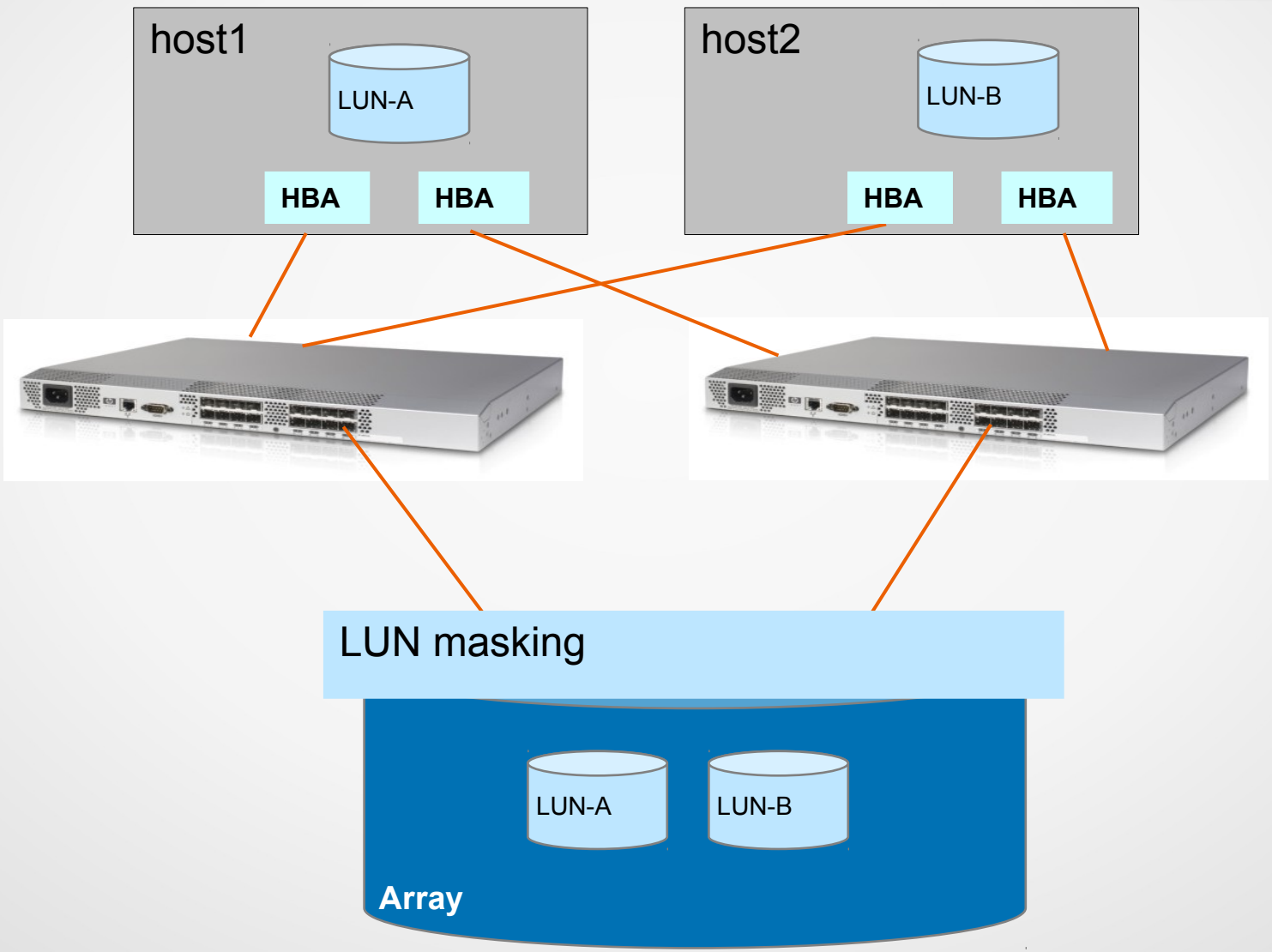
Избыточность
множественные пути к хранилищу



Зонирование «ткани»



Маскирование LUN
или выборочная презентация хранилищ (SSP)



Логические типы портов

- Порты узлов:
  - **N_Port (Node port)**, порт устройства с поддержкой топологии «Точка-Точка».
  - **NL_Port (Node Loop port)**, порт устройства с поддержкой топологии «Ткань» (Fabric).
- Порты коммутатора/маршрутизатора (только для топологии FC-SW):
  - **F_Port (Fabric port)**, порт ткани. Используется для подключения портов типа N_Port к коммутатору.
  - **FL_Port (Fabric Loop port)**, порт ткани с поддержкой петли. Используется для подключения портов типа NL_Port к коммутатору.
  - **E_Port (Expansion port)**, порт расширения. Используется для соединения коммутаторов. Может быть соединён только с портом типа E_Port.
  - **G_port (Generic port)**

Уникальный адрес устройства

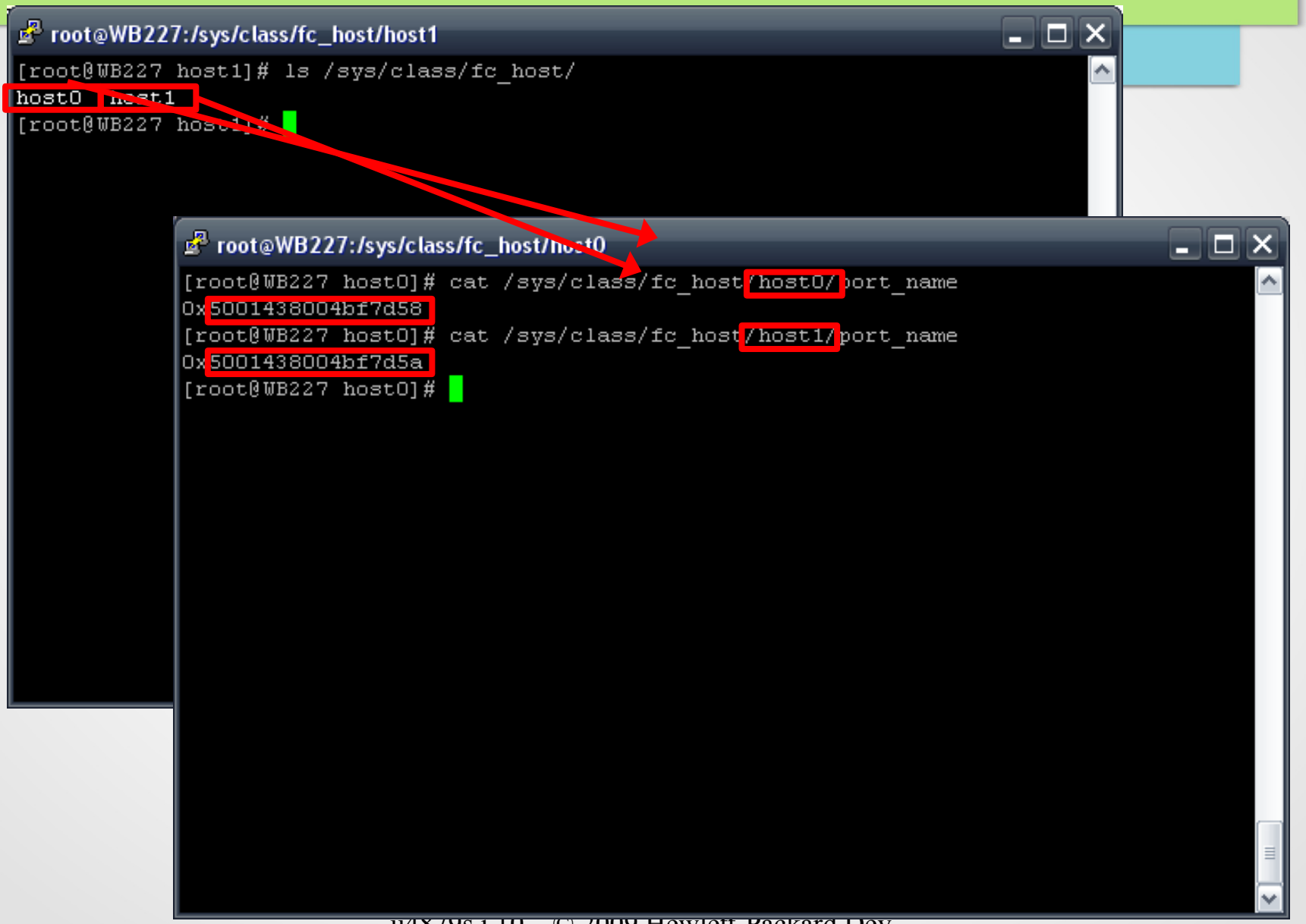
Каждое устройство имеет уникальный 8-байтовый адрес, называемый NWWN (Node World Wide Name), состоящий из нескольких компонент:

```
A0:00:BB:BB:BB:CC:CC:CC
||    |        |
||    |        ±- Назначаются производителем устройства.
||    ±-- Назначаются IEEE для каждого производителя.
|±------- Всегда 0:00 (Зарезервировано стандартом)
±-------- Число произвольно выбирается производителем.
```

Fibre Channel WWN

- WWN может использоваться для
  - Зонирования — для описания членства портов устройств в зонах.
  - Маскирования LUN — для определения доступности хостам LUN на системе хранения
- WWN не используется для адресации и доставки фрейма внутри фабрики

Collecting port WWN (RHEL)



Verifying LUN presentation (RHEL)

`fdisk -l`

```
root@WB227:~
[root@WB227 ~]# fdisk -l

Disk /dev/cciss/c0d0: 73.3 GB, 73372631040 bytes
255 heads, 63 sectors/track, 8920 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes

           Device Boot      Start         End      Blocks   Id  System
/dev/cciss/c0d0p1   *           1          13      104391   83  Linux
/dev/cciss/c0d0p2              14        8920    71545477+  8e  Linux LVM
```

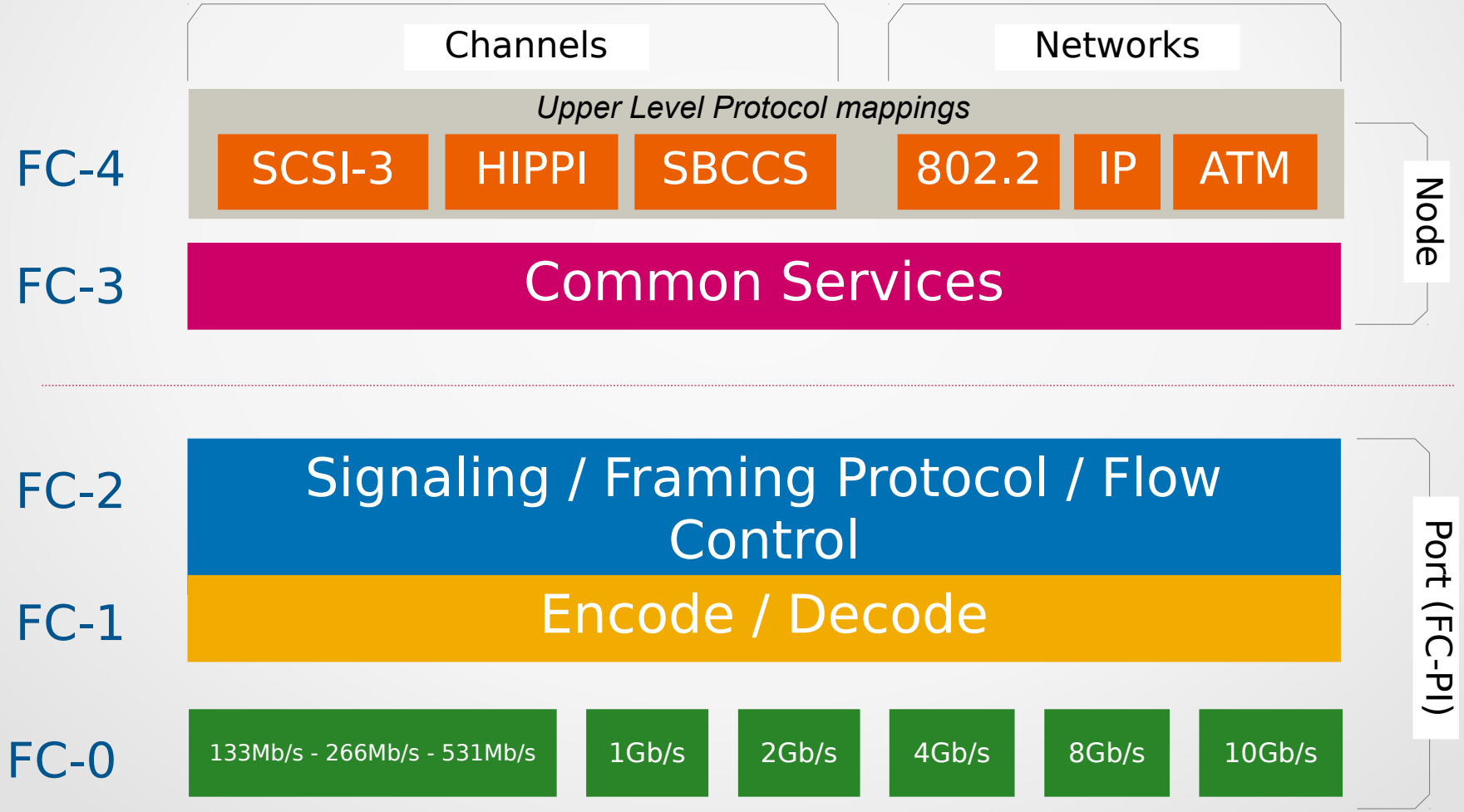
If no new LUNS detected, use:

`echo "- - -" > /sys/class/scsi_host/hostX/scan`

where `X` is HBA number checked earlier and try again with `fdisk -l`

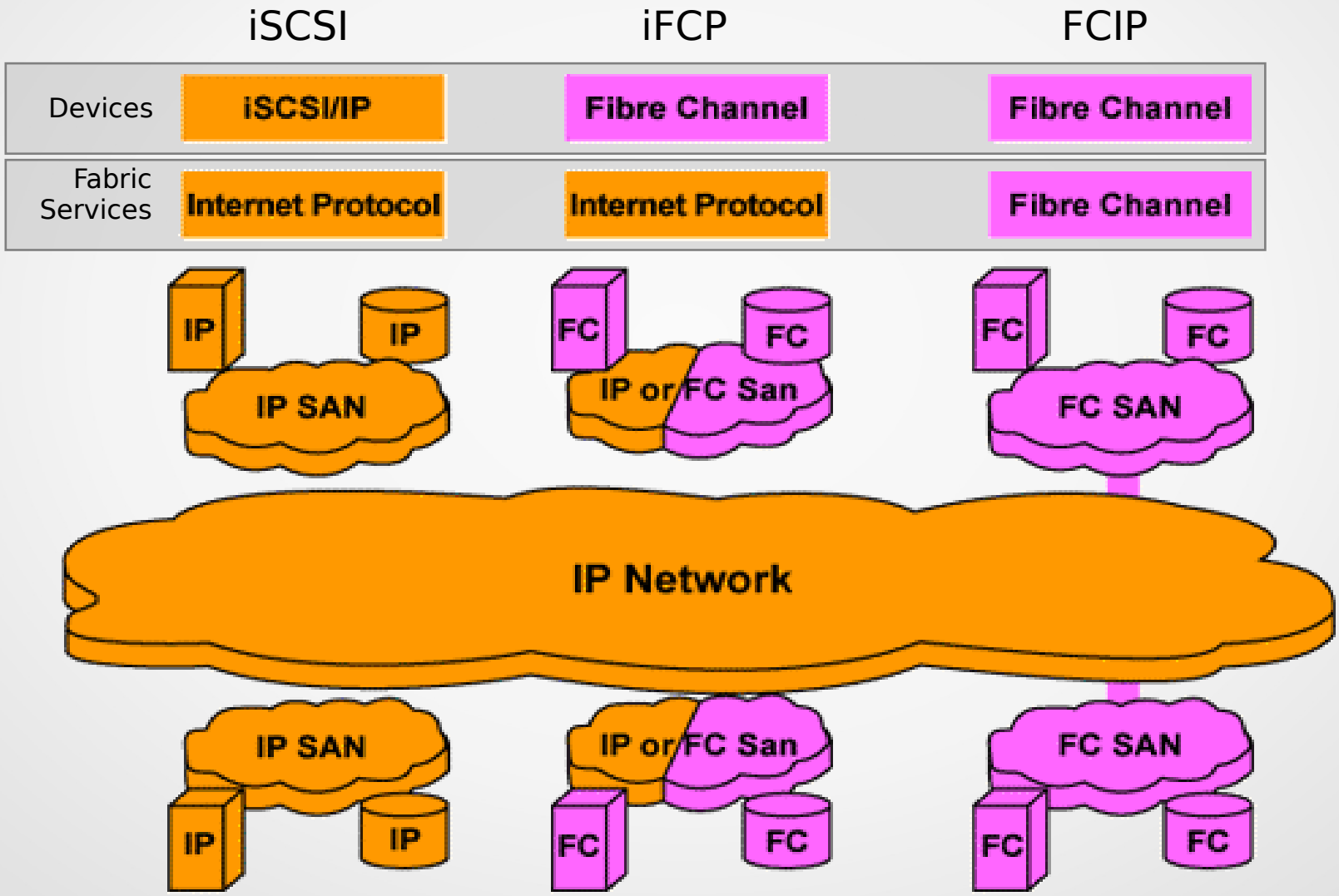
```
[root@WB227 ~]# echo "- - -" > /sys/class/scsi_host/host0/scan
[root@WB227 ~]# echo "- - -" > /sys/class/scsi_host/host1/scan
[root@WB227 ~]# fdisk -l

Disk /dev/cciss/c0d0: 73.3 GB, 73372631040 bytes
255 heads, 63 sectors/track, 8920 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes

           Device Boot      Start         End      Blocks   Id  System
/dev/cciss/c0d0p1   *           1          13      104391   83  Linux
/dev/cciss/c0d0p2              14        8920    71545477+  8e  Linux LVM

Disk /dev/sda: 1073 MB, 1073741824 bytes
34 heads, 61 sectors/track, 1011 cylinders
Units = cylinders of 2074 * 512 = 1061888 bytes
```

Сетевая модель Fibre Channel

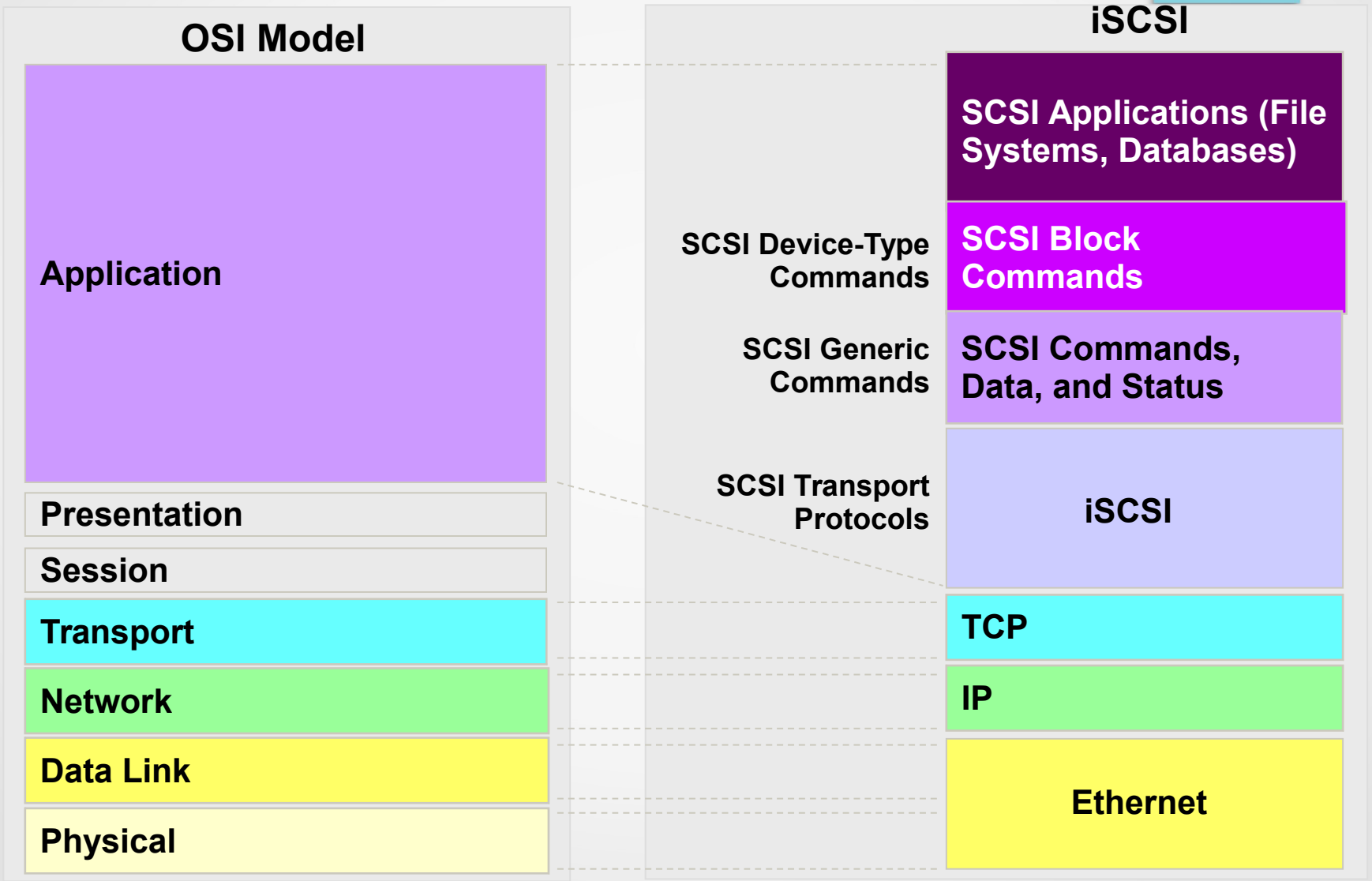


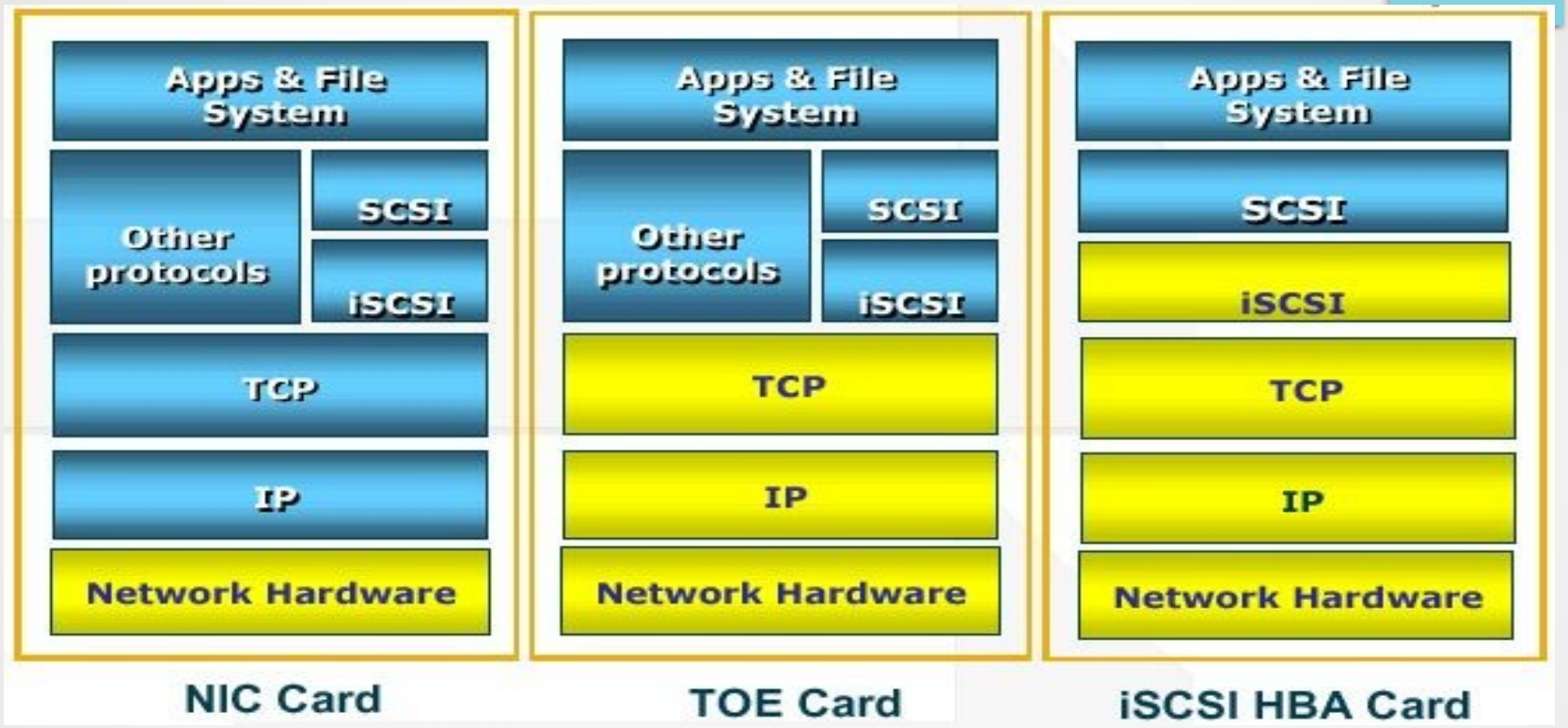
Сетевая модель Fibre Channel

- FC-0 Описывает среду передачи, трансиверы, коннекторы и типы используемых кабелей.
- FC-1 Описывает процесс 8b/10b Кодирования (каждые 8 бит данных кодируются в 10-битовый символ (Transmission Character)), специальные символы и контроль ошибок.
- FC-2 Описывает сигнальные протоколы. На этом уровне происходит разбиение потока данных на кадры и сборка кадров. Определяет правила передачи данных между двумя портами, классы обслуживания
- FC-3 Определяет такие особенности, как: расщепление потока данных (striping), шифрования, компрессия, избыточность
- FC-4 Предоставляет возможность переноса других протоколов (SCSI, ATM, IP, HIPPI FDDI, Token Ring, AV, VI, IBM SBCCS и многих других.)


iSCSI
iFCP
FCIP
Devices
iSCSI/IP
Fibre Channel
Fibre Channel
Fabric Services
Internet Protocol
Internet Protocol
Fibre Channel
IP
IP
IP SAN
FC
FC
IP or FC San
FC
FC
FC SAN
IP Network
IP SAN
IP
IP
IP or FC San
FC
FC
FC SAN
FC
FC

Стек протоколов iSCSI



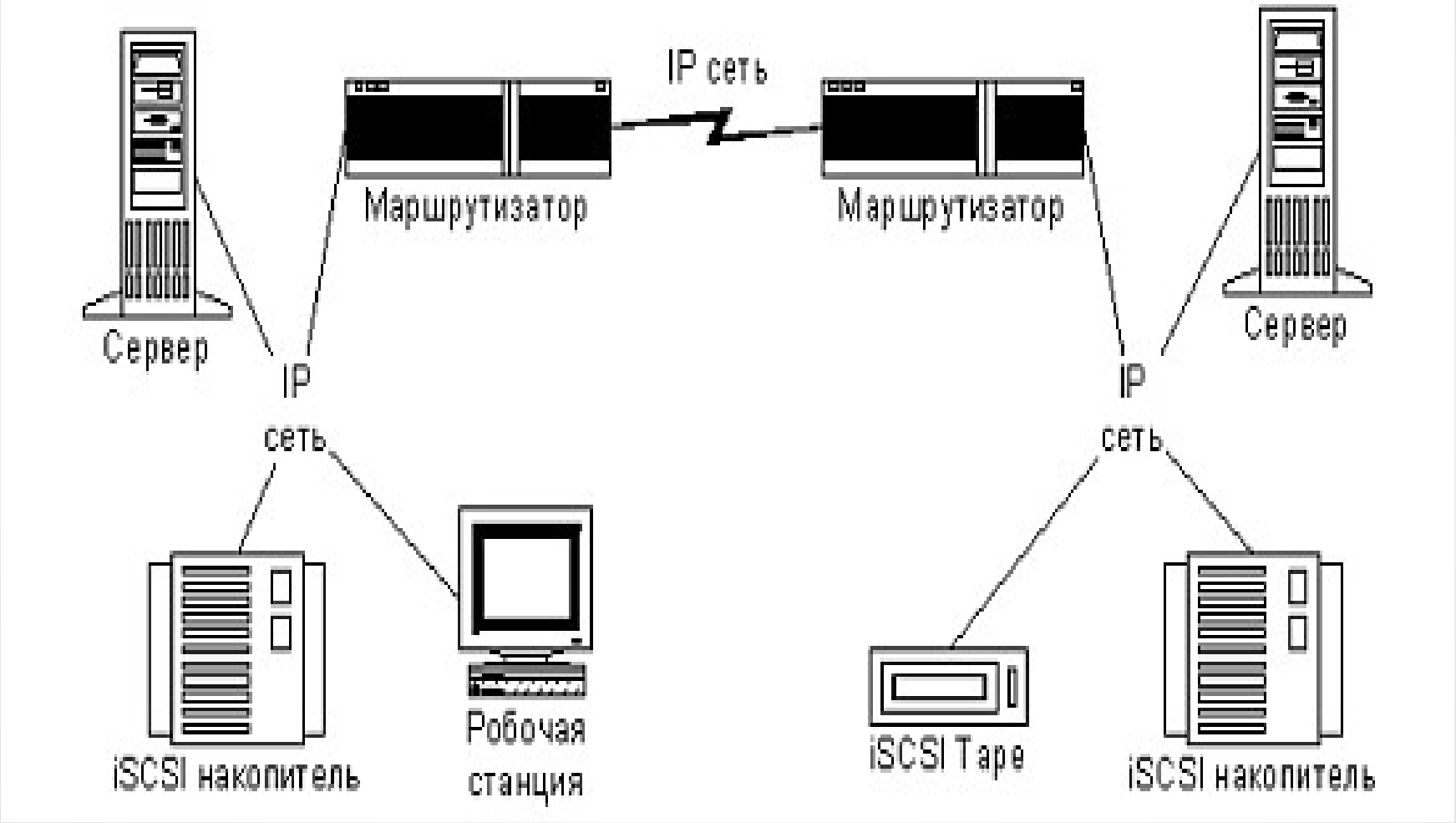


Good → Better → Best

Processed in the Network Card

Processed in the Server

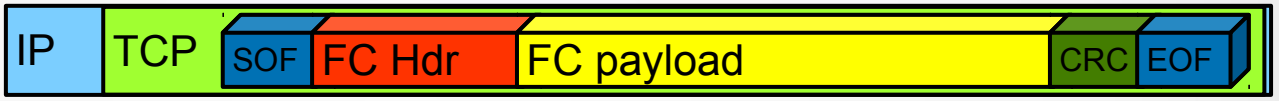
iSCSI сеть

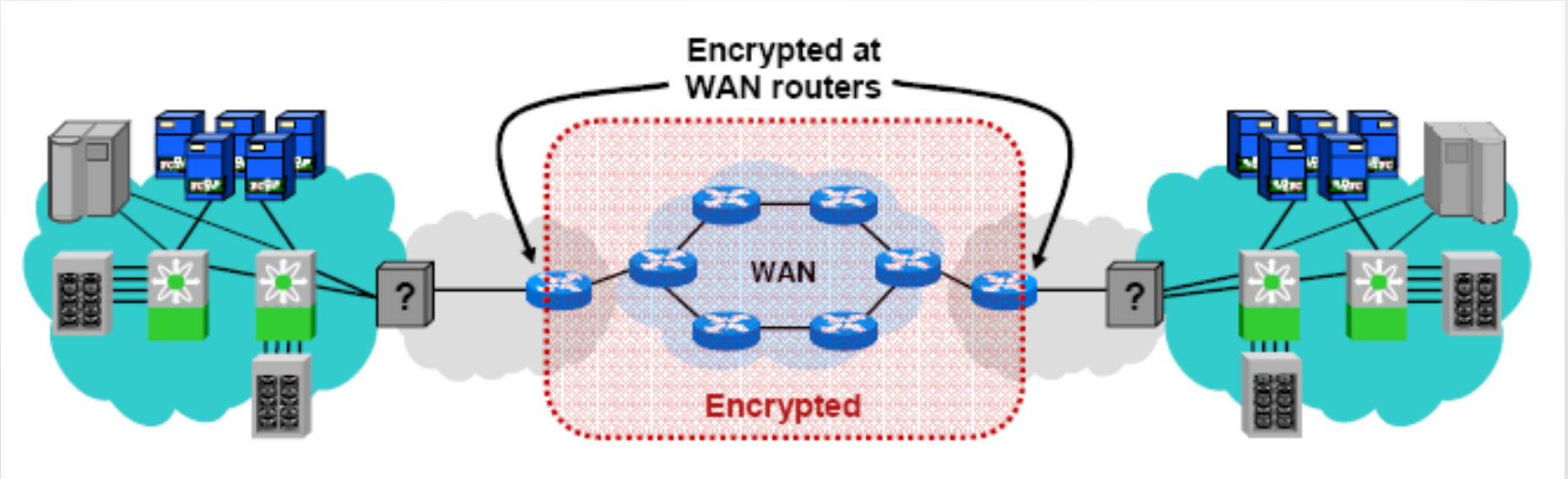

IP сеть
Маршрутизатор
Маршрутизатор
Сервер
Сервер
IP
сеть
IP
сеть
iSCSI накопитель
Робочая
станция
iSCSI Tape
iSCSI накопитель

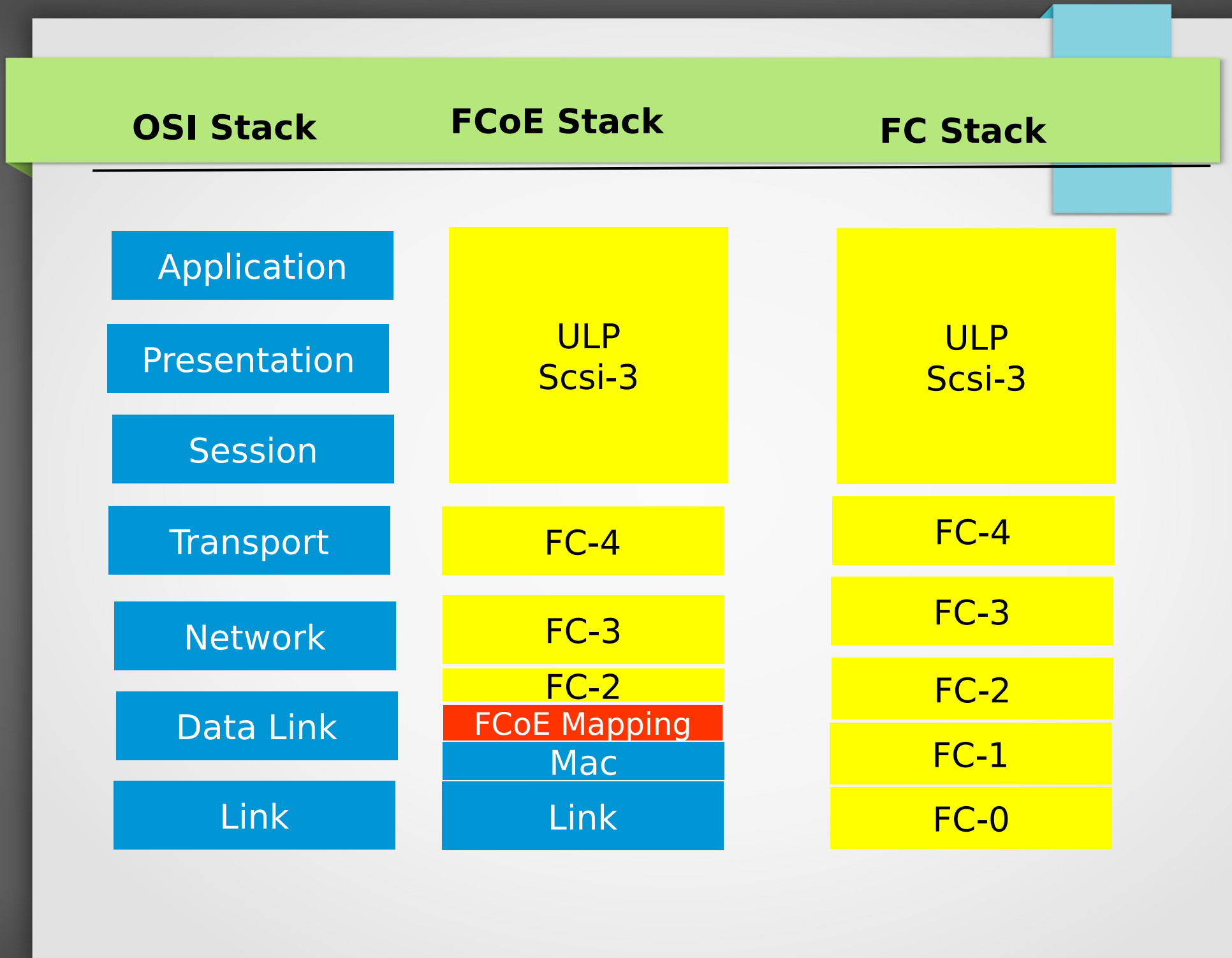
FCIP туннелирование

FC frame

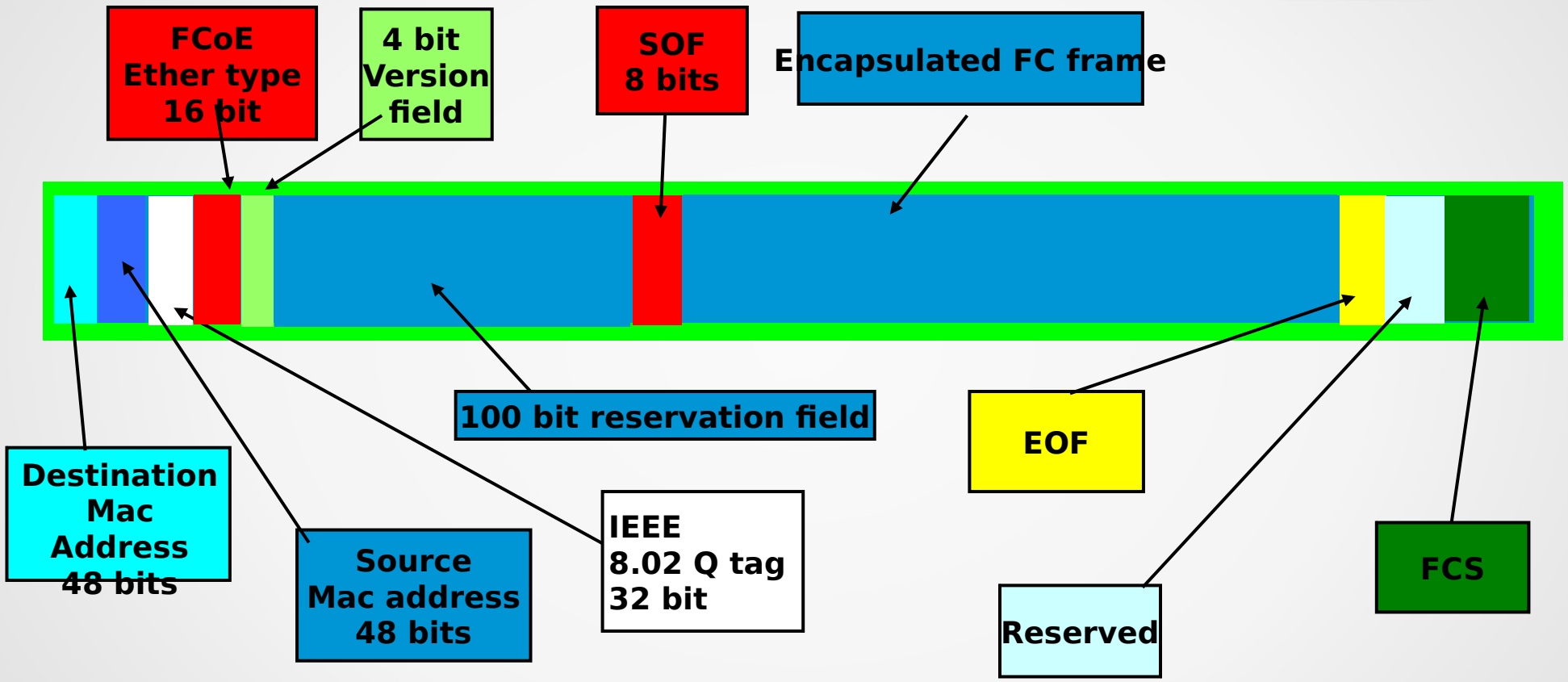
FCIP frame






OSI Stack
FCoE Stack
FC Stack
Application
Presentation
Session
Transport
Network
Data Link
Link
ULP
Scsi-3
FC-4
FC-3
FC-2
FCoE Mapping
Mac
Link
ULP
Scsi-3
FC-4
FC-3
FC-2
FC-1
FC-0

FCoE инкапсуляция



Lossless Ethernet