Введение в сети хранения данных

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.

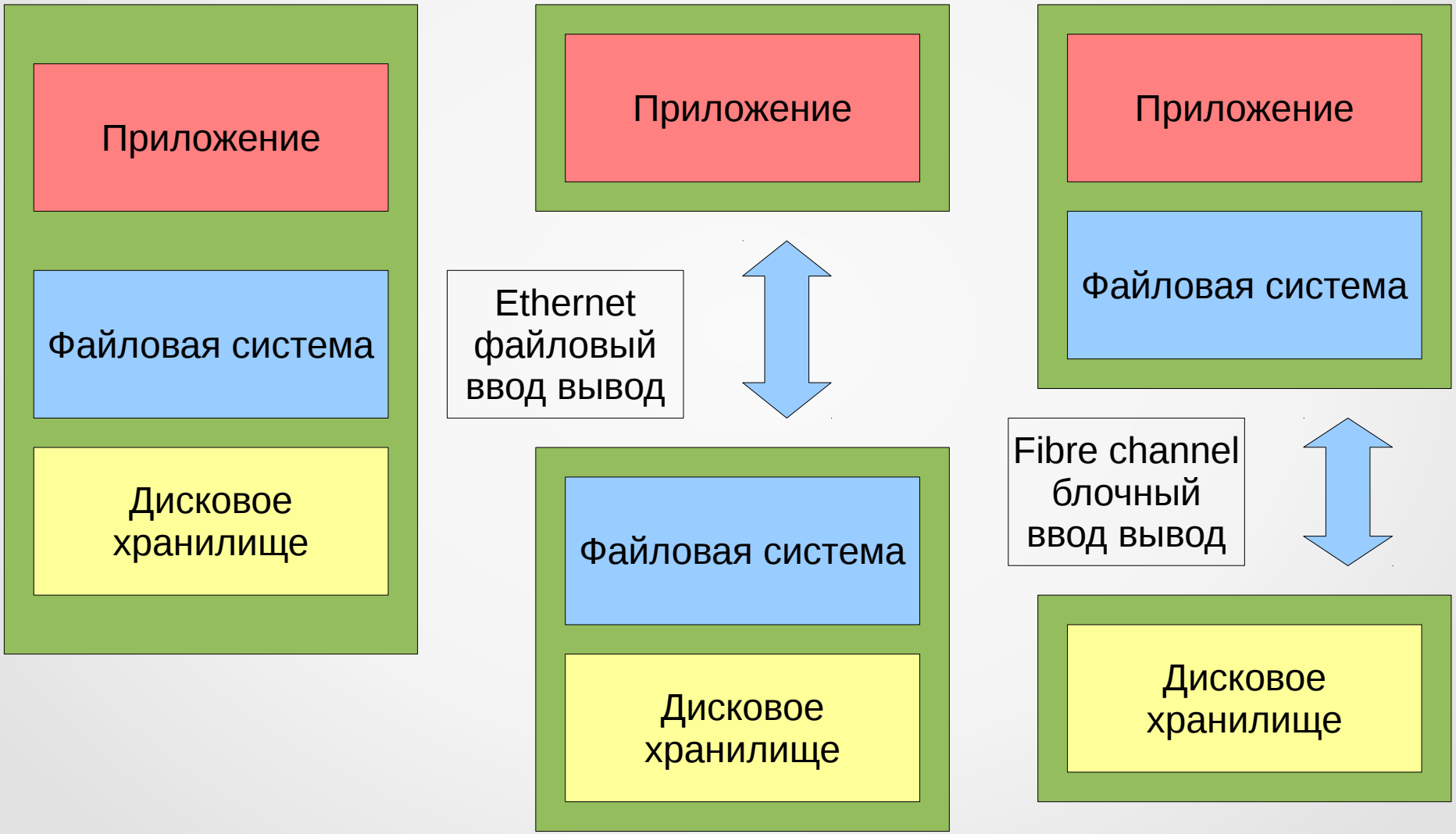
SAN

•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

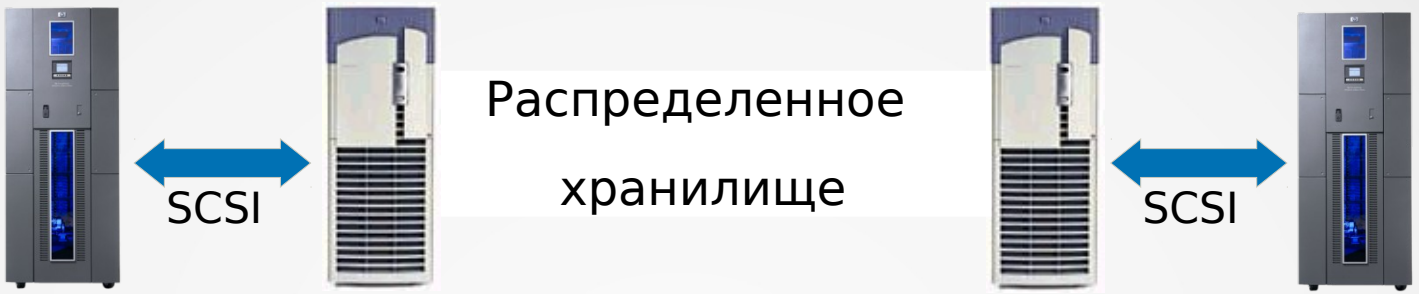

Приложение
Файловая система
Дисковое хранилище
Приложение
Ethernet файловый ввод вывод
Файловая система
Дисковое хранилище
Приложение
Файловая система
Fibre channel блочный ввод вывод
Дисковое хранилище

SAN

Storage Area Network

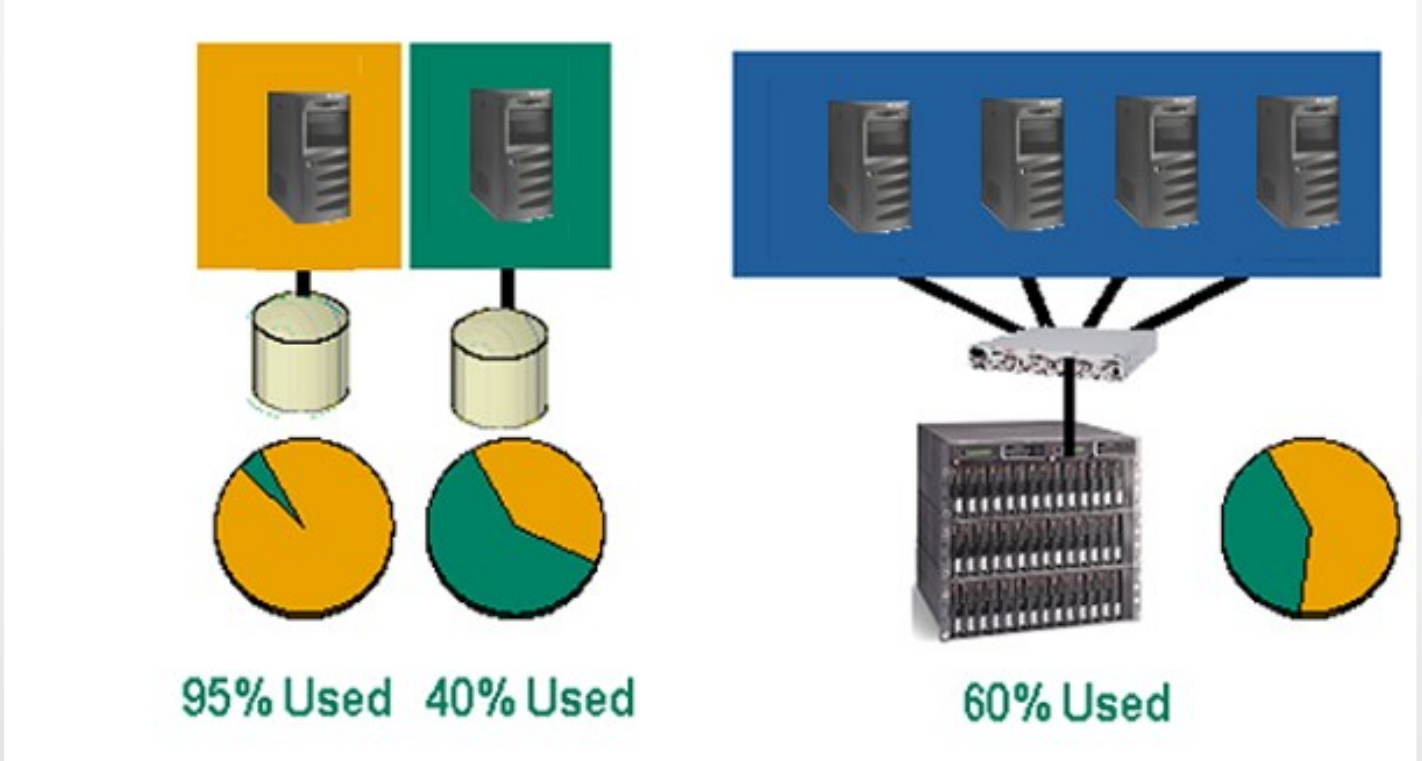
- Доступ к устройствам (RAW)
- Децентрализация
- Поставщики и потребители объединены сетью
- Возможность использования одного устройства несколькими потребителями

Консолидация серверов и систем хранения

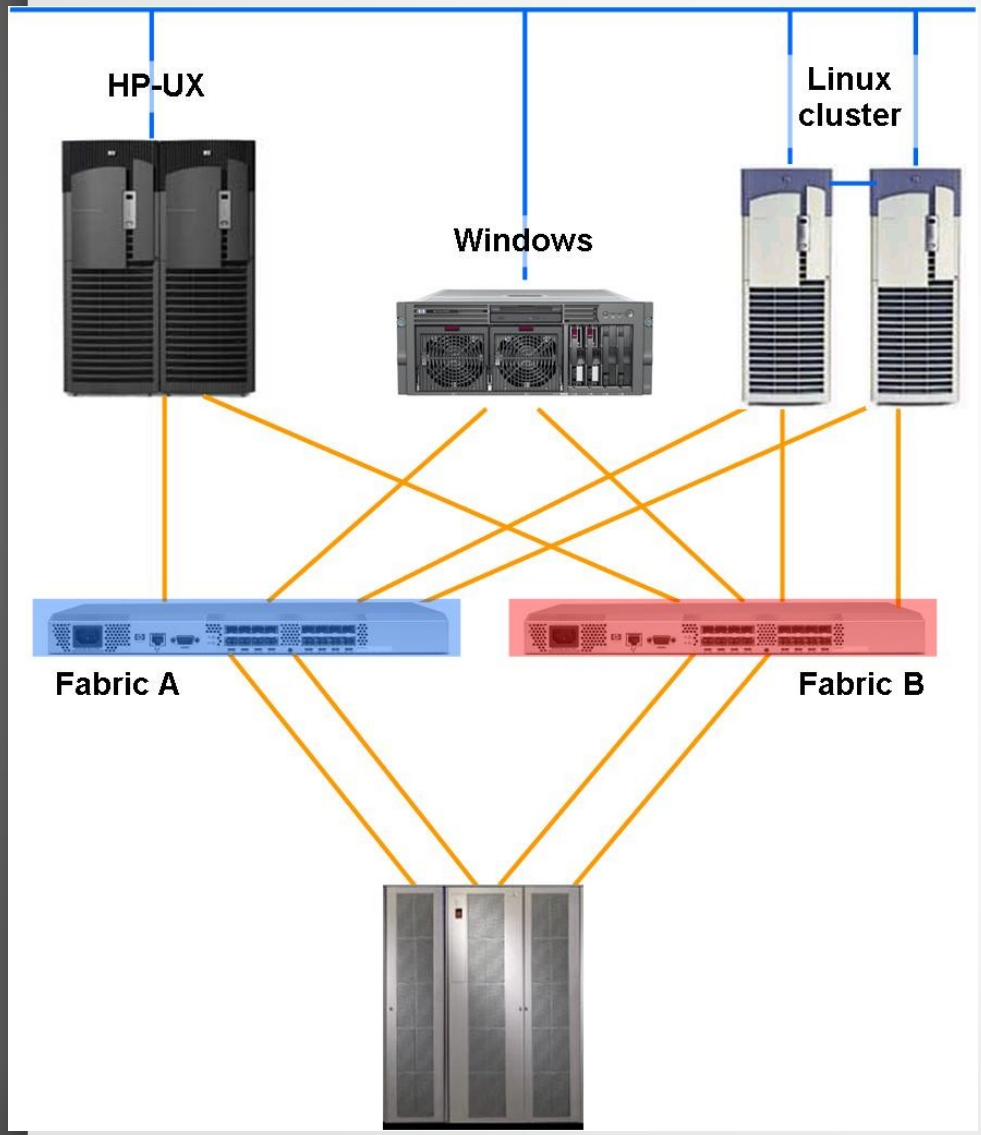


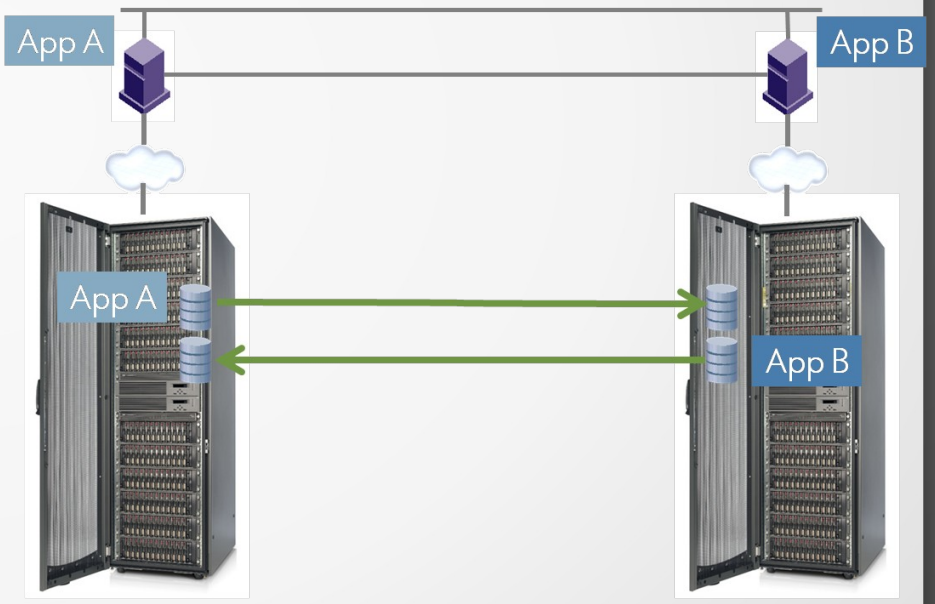
Эффективное управление ёмкостью

- 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:

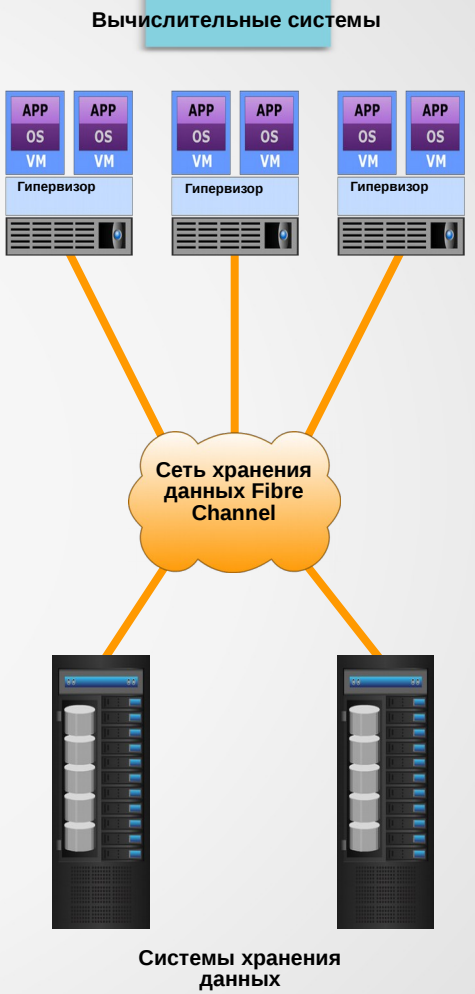
- Ethernet
- FibreChannel

Протоколы:

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

Что такое сеть хранения данных Fibre Channel?

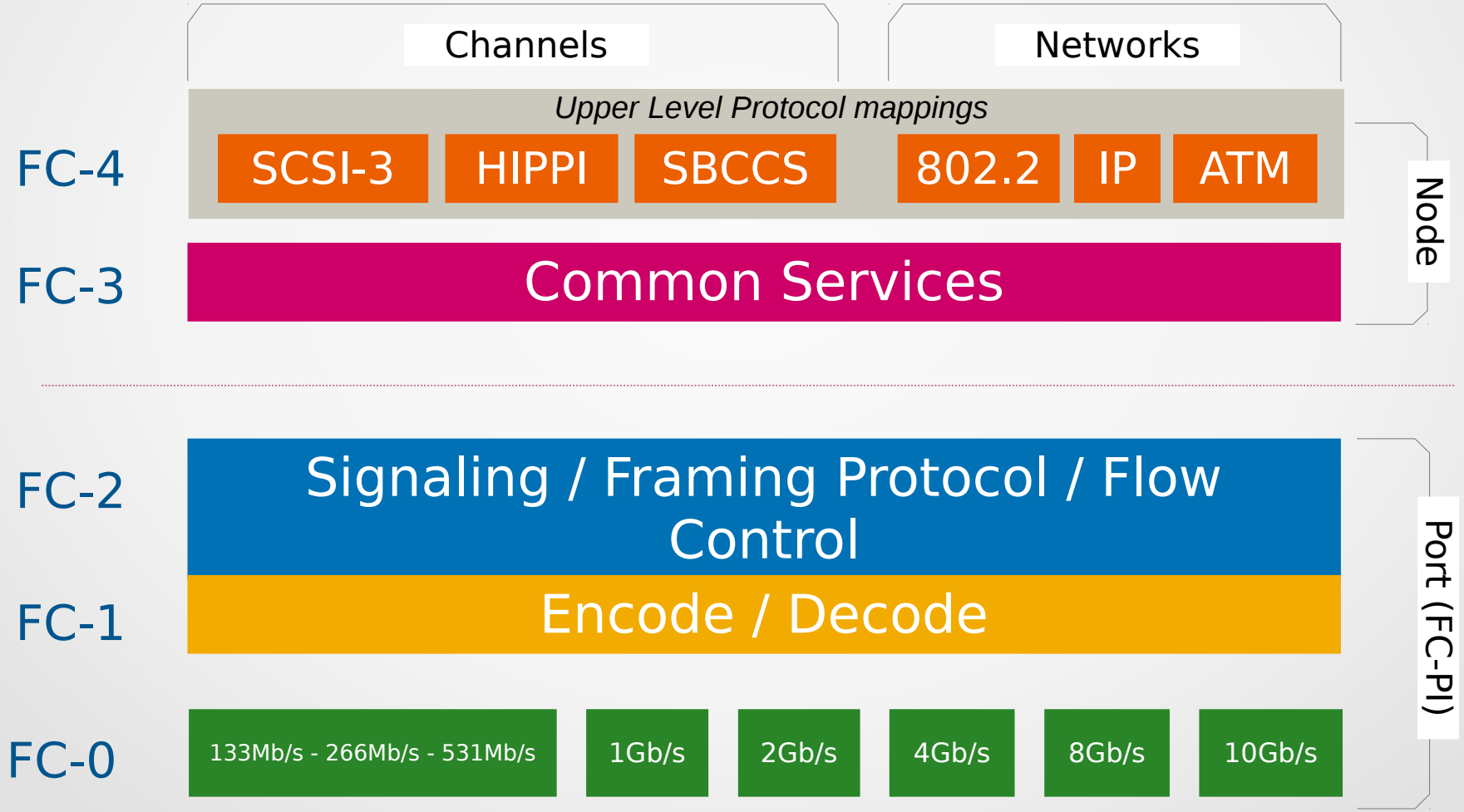
- Сеть хранения данных, использующая для обмена данными протокол Fibre Channel (FC)
- Поддерживает скорость передачи данных до 16 Гбит/с
- Обеспечивает передачу данных без сброса пакетов
- Обеспечивает высокую масштабируемость
  - Теоретическая возможность размещения примерно 15 миллионов устройств



OSI vs FC stack

Layer	Title	Fibre Channel
7	Application	SCSI-3, IPI, HIPPI, IP, VI, AE, AV, IPI, ATM, FICON, …
6	Presentation	
5	Session	
4	Transport	FC-4 Protocol Interface ULP
3	Network	FC-3 Encryption Authentication
2	Data Link	FC2 Framing Flow Control Class of Service
		FC – 1 Encoding Link Control
1	Physical	FC-0 Physical

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.)

World Wide Name (WWN)

– 64 bit IEEE structured address

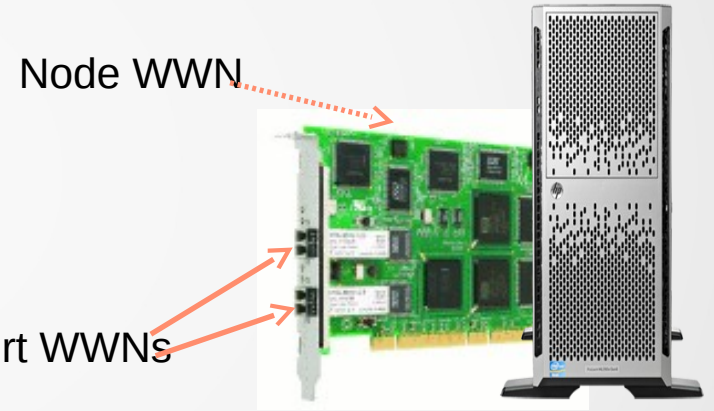
– Vendor specific bit variables

– Used to preserve identity of a node if its FC-2 or FC-3 layer address is changed

Initiator example:
1 X Node WWN
2 X Port WWN

Target example:
1 X Node WWN (assigned to the node)
4 X Port WWN (port derivatives of Node WWN)

General	
Rate	8 Gbps
Adapter type	EMULEX LPe12002
CNA port	No
WWN	20110002AC002E2D
Port type	Host
Mode	Target
Mode change	Allowed




```
N_Port Node World Wide Name = 0x50060b000024a149
N_Port Port World Wide Name = 0x50060b000024a148
```

Уникальный адрес устройства

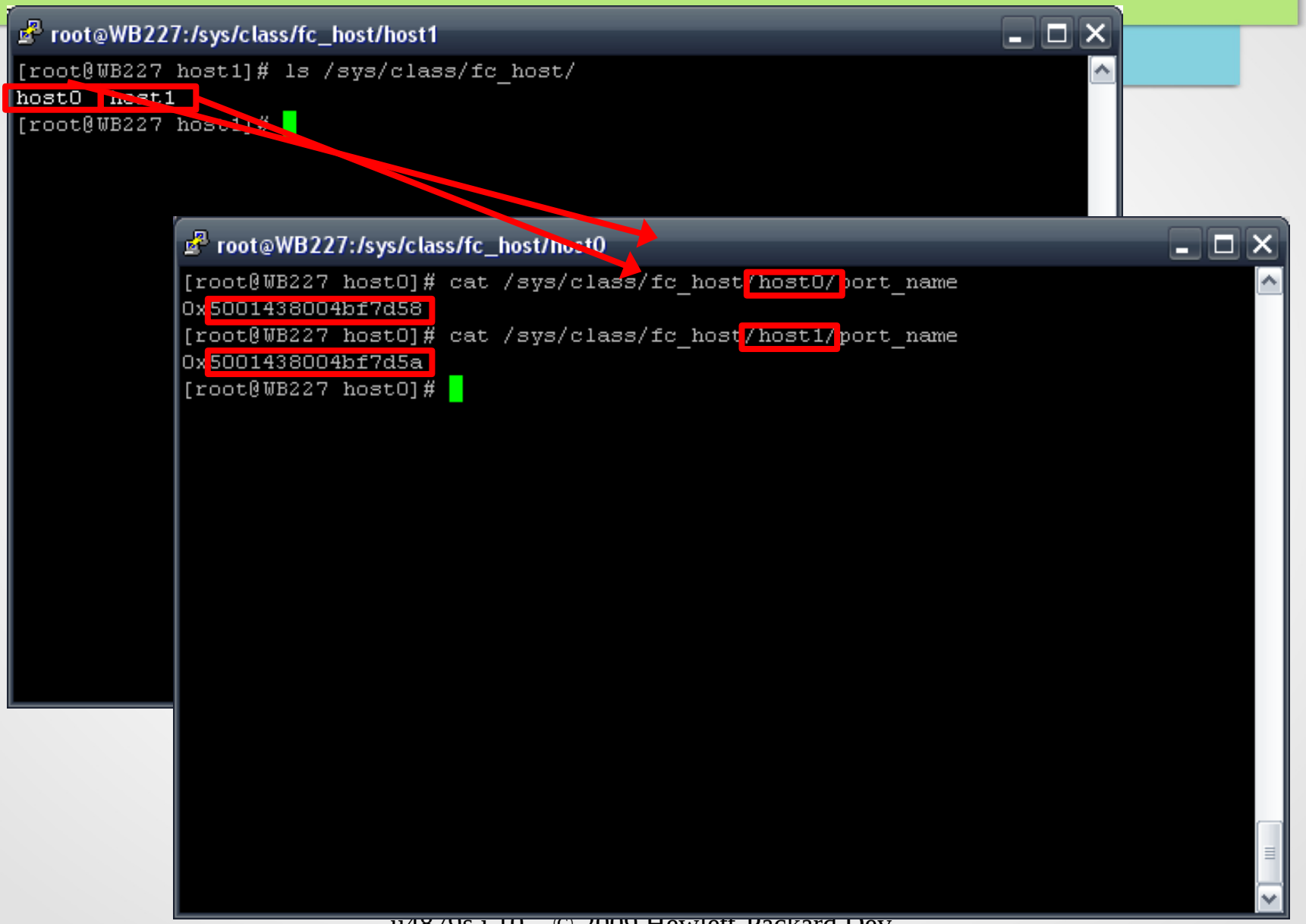
Каждое устройство имеет уникальный 8-байтовый адрес, называемый NWWN (Node World Wide Name), состоящий из нескольких компонент:

```
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

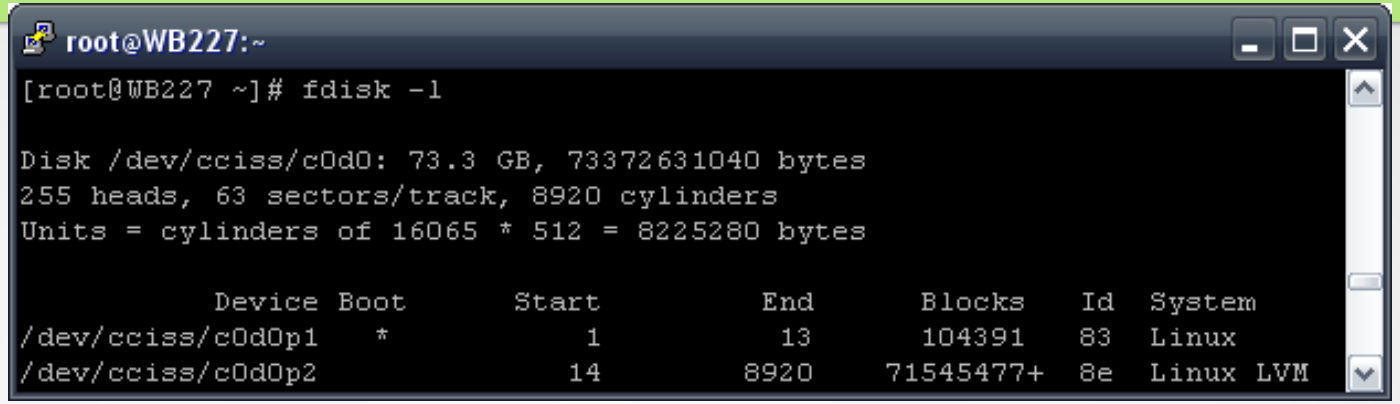
- WWN может использоваться для
  - Зонирования — для описания членства портов устройств в зонах.
  - Маскирования LUN — для определения доступности хостам LUN на системе хранения
- WWN не используется для адресации и доставки фрейма внутри фабрики

Collecting port WWN (RHEL)



Verifying LUN presentation (RHEL)

`fdisk –l`



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`

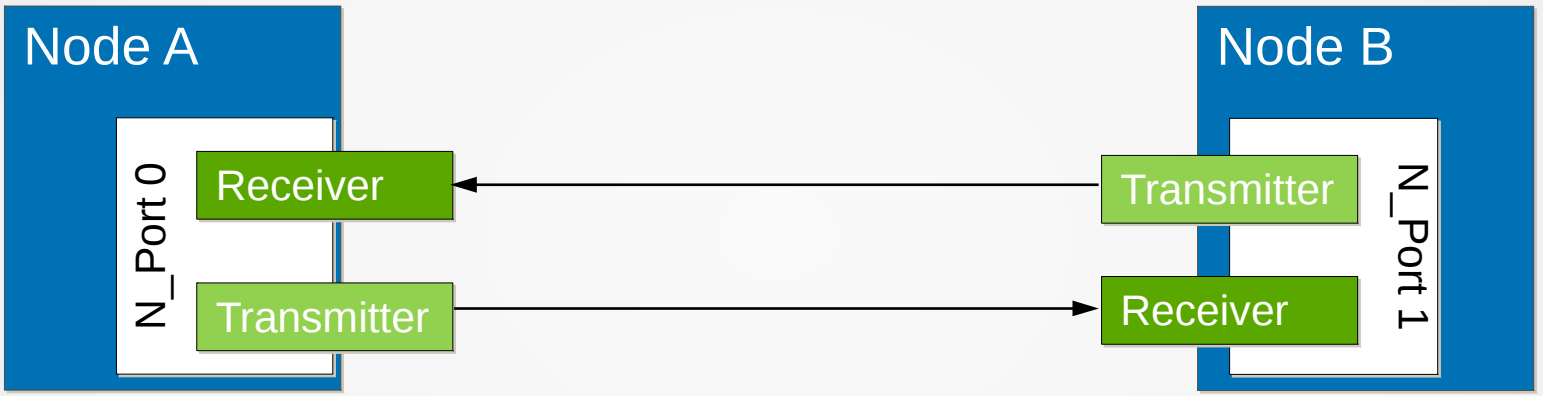
```
[root@WB227 ~]# echo "- - -" > /sys/class/scsi_host/host0/scan
[root@WB227 ~]# echo "- - -" > /sys/class/scsi_host/host1/scan
[root@WB227 ~]# fdisk -l

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
```

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

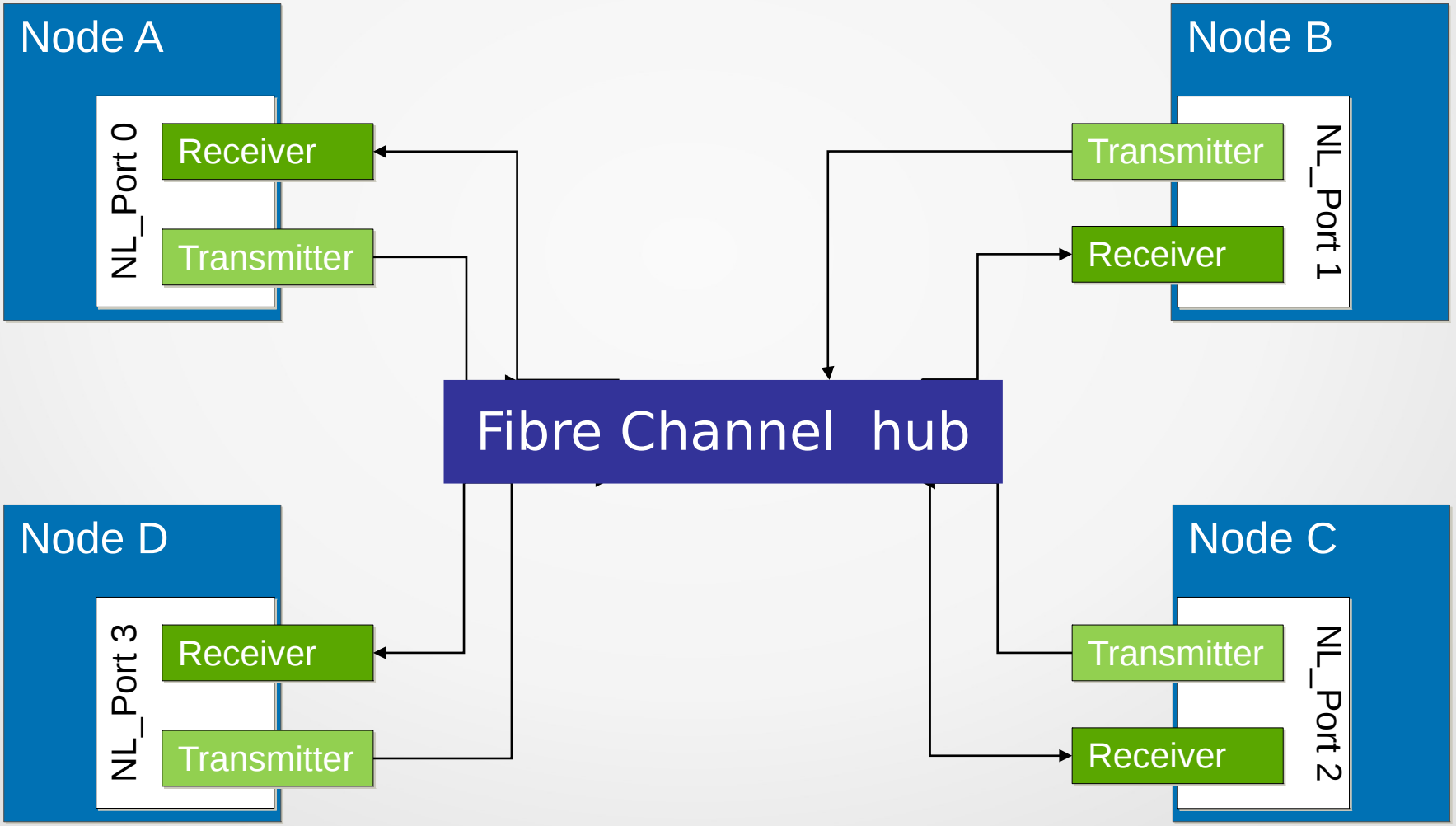
loop	fabric	topology
yes	no	private (arbitrated) loop
yes	yes	public loop
no	no	direct point-to-point
no	yes	switched point-to-point (*)

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

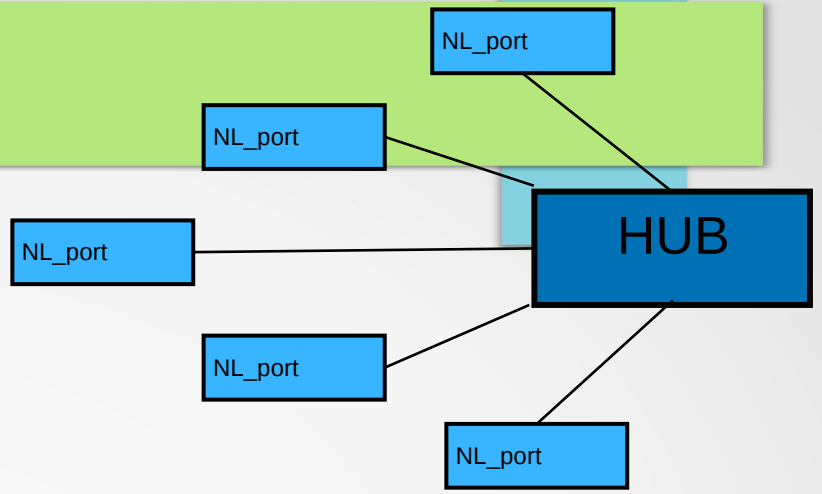


+дешего +монопольное использование канала
- комутация только двух устройств

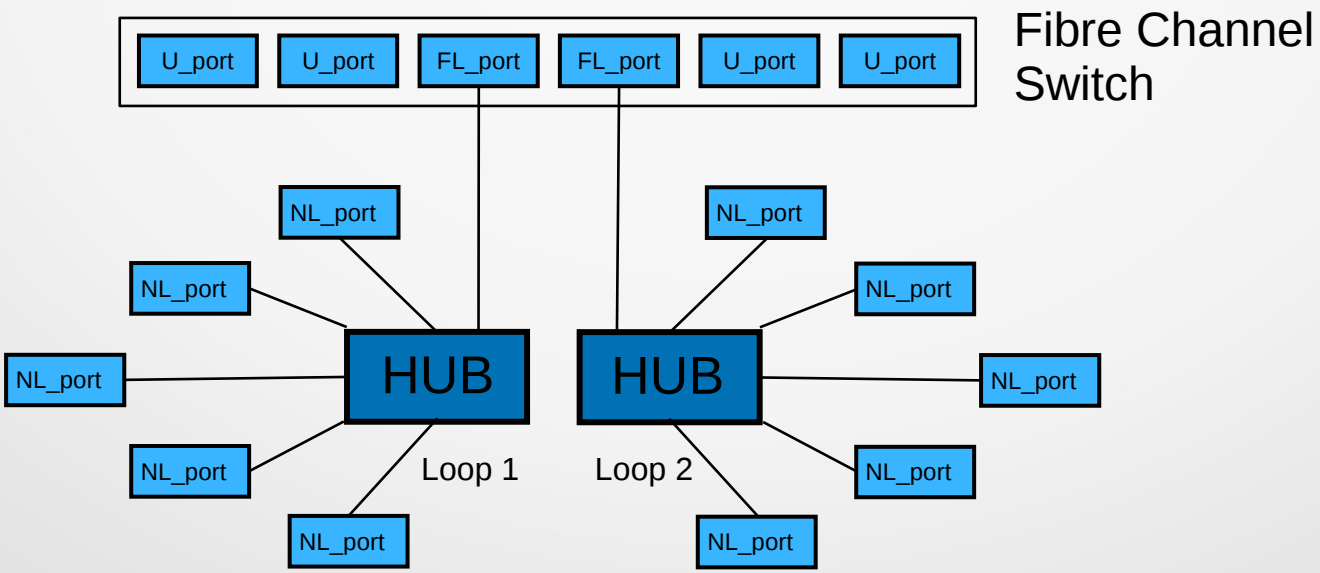
управляемая петля (arbitrated loop) FC-AL



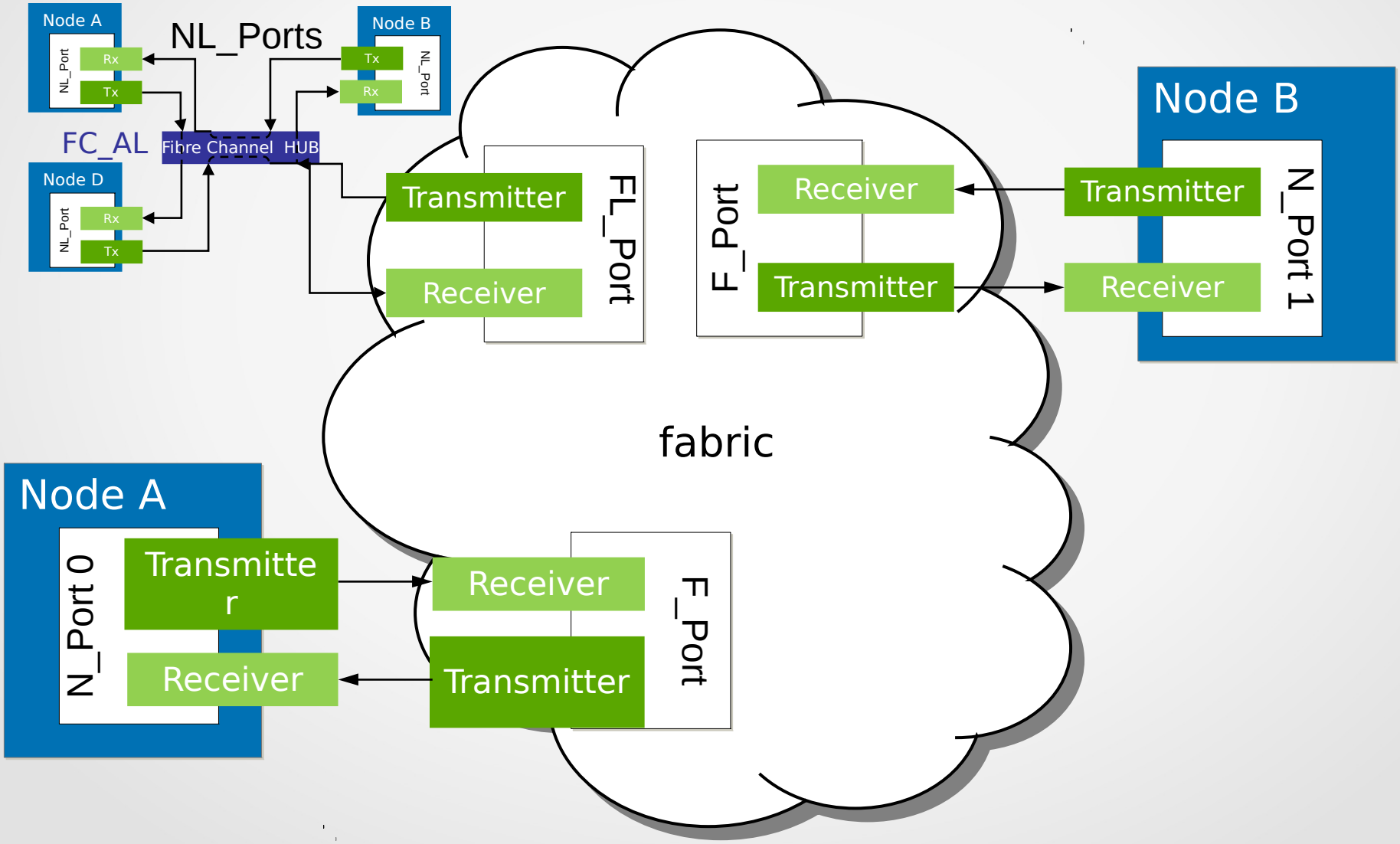
Частная петля (Private loop)



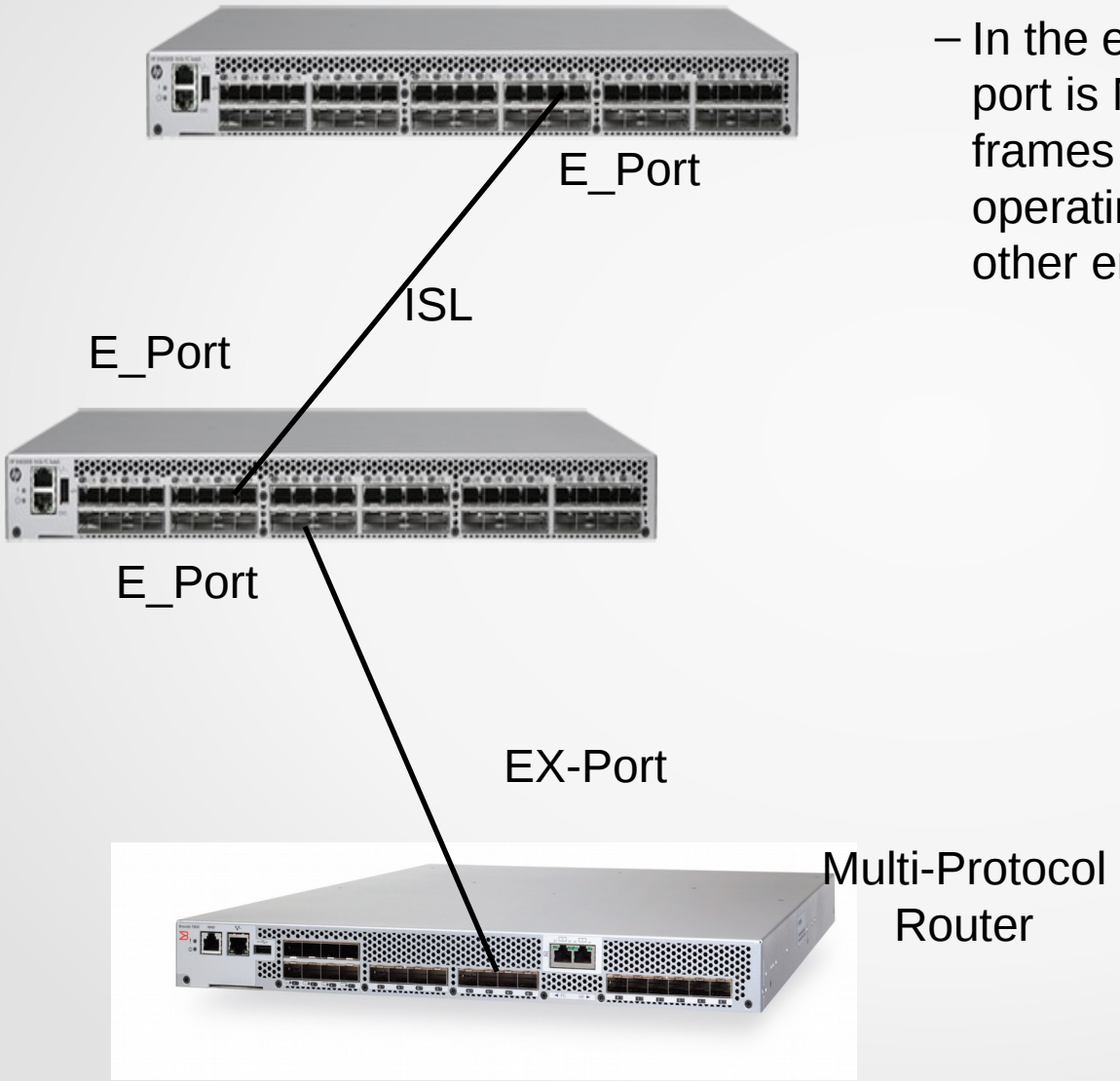
Публичная петля (Public loop)



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



Fibre Channel port types



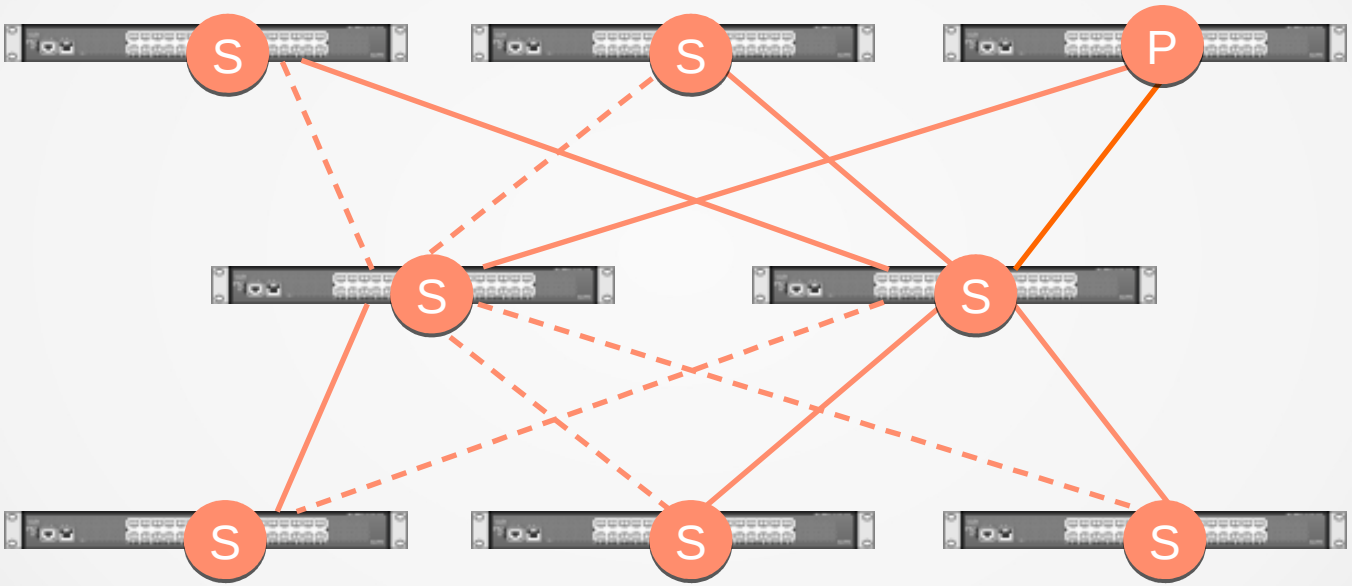
– In the event that the recently connected port is NOT an Nx-port Link Service frames are exchanged to determine the operating parameters of the device at the other end of the link.

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)**

Principal switch

–



Структура и заголовок FC фрейма

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

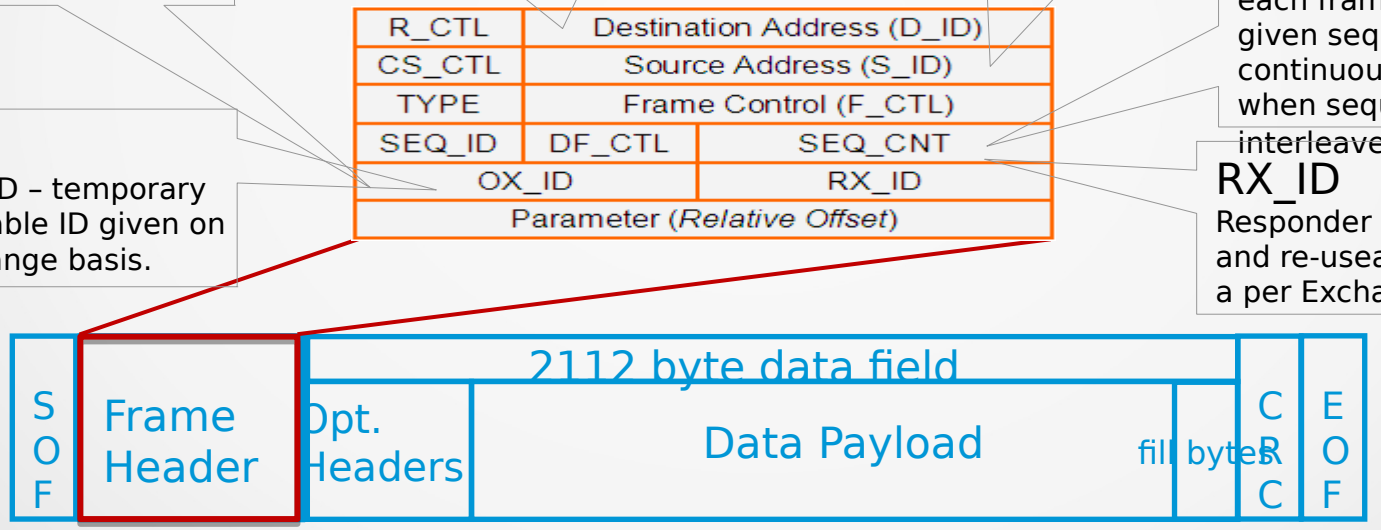
D_ID
24 bit FC_ID of destination port

S_ID
24 bit FC_ID of transmitting port

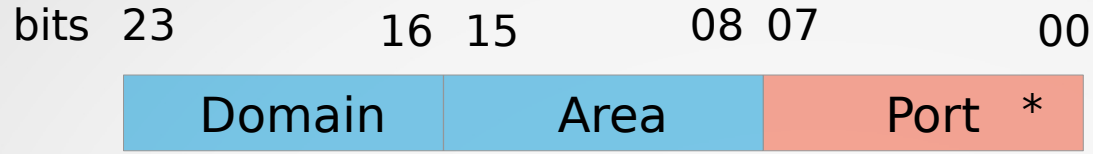
SEQ_CNT
Incremented by x0001 for each frame sent within a given sequence. May continuously increment when sequences are interleaved.

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

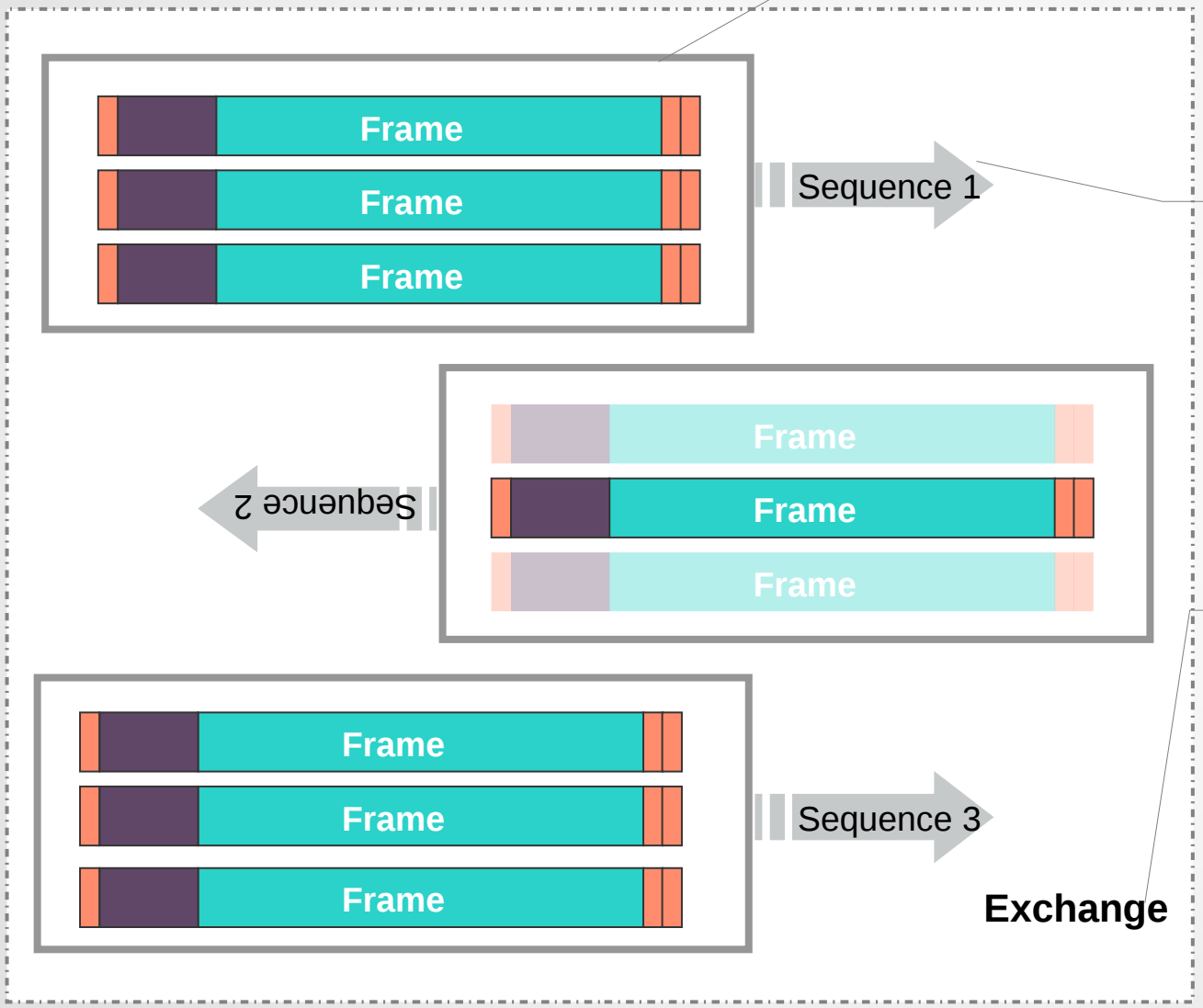
R_CTL	Destination Address (D_ID)	
CS_CTL	Source Address (S_ID)	
TYPE	Frame Control (F_CTL)	
SEQ_ID	DF_CTL	SEQ_CNT
OX_ID		RX_ID
Parameter (*Relative Offset*)		

Frame Header

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

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

Fibre Channel terminology



Frame
Used to carry ULP data in the payload

Sequence
- Single direction only
- May consist of only one frame
- Equates directly to SCSI Information Unit

Exchange
- Bi-directional
- Equates to SCSI Read or SCSI Write
- Contains multiple sequences

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

Class of Service

- Indicates frame delivery importance to transport layer
- Most widespread CoS is Class 3
- Most fabric switches and HBA's support CoS 2 and 3
- Disk arrays tend to support Class 3 only

Class of service	Fibre Channel description
Class 1	▪ Dedicated connection ▪ In-order delivery, acknowledge first frame only ▪ No flow control after first frame of connection
Class 2	▪ Connectionless ▪ Frame switched ▪ Out-of-order delivery possible ▪ Acknowledge each frame ▪ Buffer-to-buffer and end-to-end flow control for all frames
Class 3	▪ Frame switched ▪ Out-of-order delivery possible ▪ No acknowledgments ▪ Buffer-to-buffer frame control for all frames
Class 4	▪ Connection oriented ▪ Virtual circuit ▪ In-order delivery
Class 5	*True Isynchronous* – no longer used
Class 6	▪ Connection oriented ▪ Multicast service
Class F	▪ Connectionless and acknowledged (similar to Class 2) ▪ Used by switches for fabric related traffic ▪ Out-of-order delivery possible ▪ Fabric may reject frames if not delivered within ED_TOV

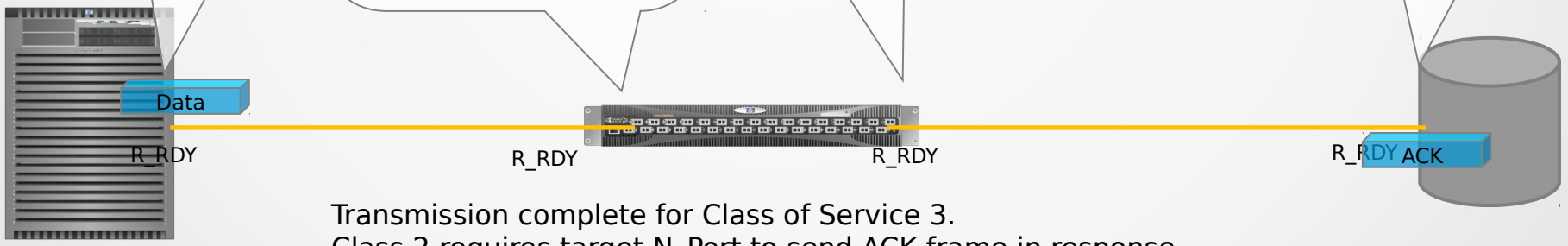
FC контроль передачи

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.

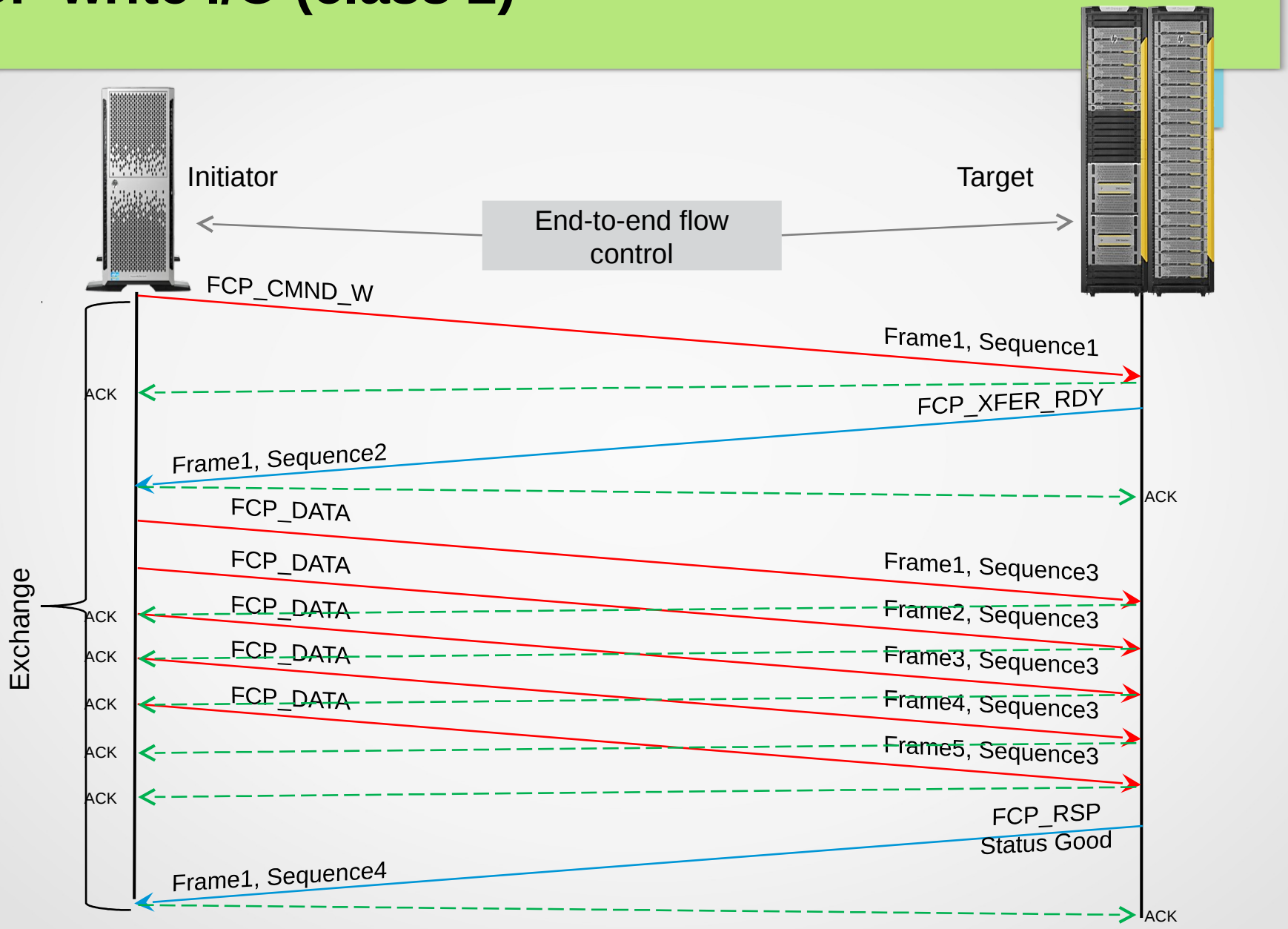


Transmission complete for Class of Service 3.
Class 2 requires target N_Port to send ACK frame in response…
Class 2 frame transmission complete:

- Buffer to Buffer Flow Control

  AND
- End to End Flow Control

FCP write I/O (class 2)



FSPF in practice

Domain_ID: 1

Port:
8

Link Descriptor
Owning Domain_ID: x'01'
Output Port Index: x'08'
Domain_ID of Neighbor: x'03'
Neighbor Port Index: x'0A'
Link Cost: 500

NOTE: the Brocade default link cost metric value for 2 ,4, 8, 10 and 16 Gbps is 500

Domain_ID: 3

Port: 10

Port: 5

Link Descriptor
Owning Domain_ID: x'03'
Output Port Index: x'0A'
Domain_ID of Neighbor: x'01'
Neighbor Port Index: x'08'
Link Cost: 500

Link Descriptor
Owning Domain_ID: x'05'
Output Port Index: x'0A'
Domain_ID of Neighbor: x'03'
Neighbor Port Index: x'05'
Link Cost: 500

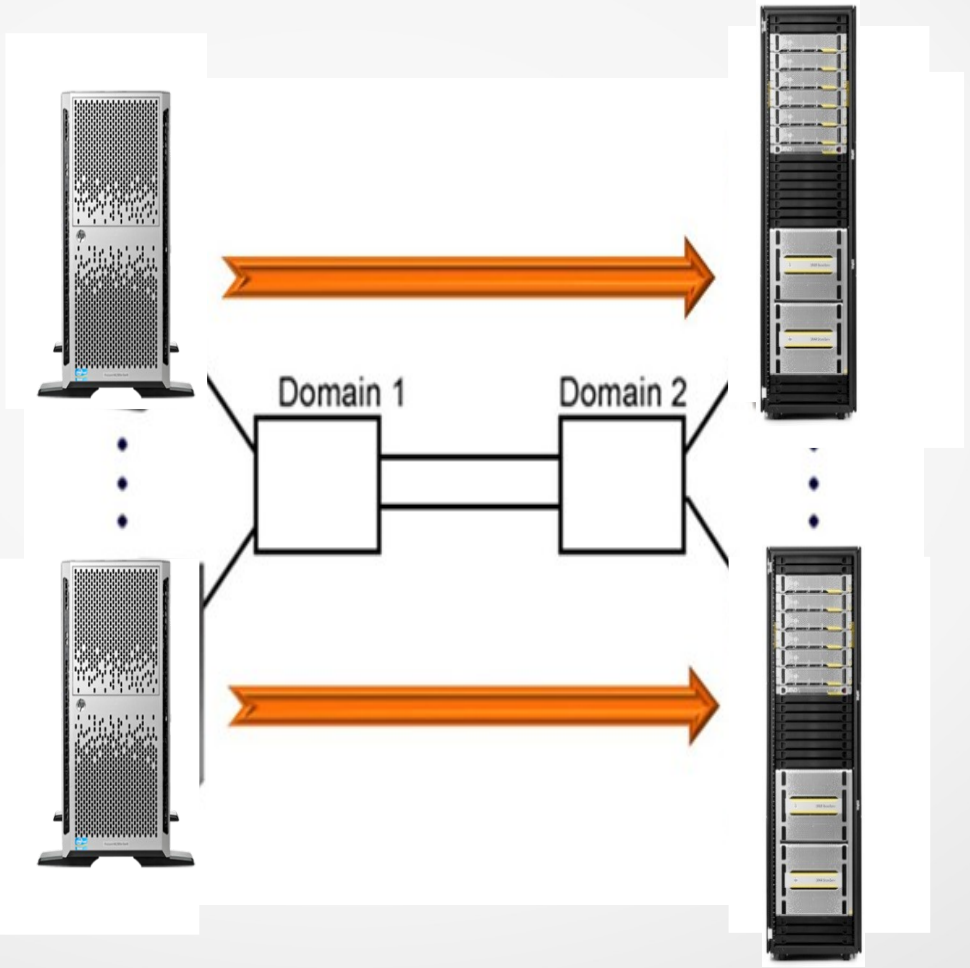
Link Descriptor
Owning Domain_ID: x'03'
Output Port Index: x'05'
Domain_ID of Neighbor: x'05'
Neighbor Port Index: x'0A'
Link Cost: 500

Domain_ID: 5

Port: 10

Port Based Routing

– The choice of routing path is based only on the incoming port and the destination domain.



Equal cost routes – exchange-based routing

- Equal cost routes are considered for routing on a ‘per exchange’ basis.
- Switch routing handler examines frame header and selects ISL based on OX_ID/RX_ID.
- All frames within a specific exchange are routed across the same ISL, preserving In Order Delivery (IOD).
- Frame delivery still takes place using D_ID.
- Brocade’s 4, 8, 10, 16 Gbps ASICS can use the FSPF protocol, port-based routing or Exchange-based routing.
- Exchanged-base routing is Brocade’s factory default setting.

All frames within an exchange (OX_ID/RX_ID pair) are routed via the same ISL. In which case:

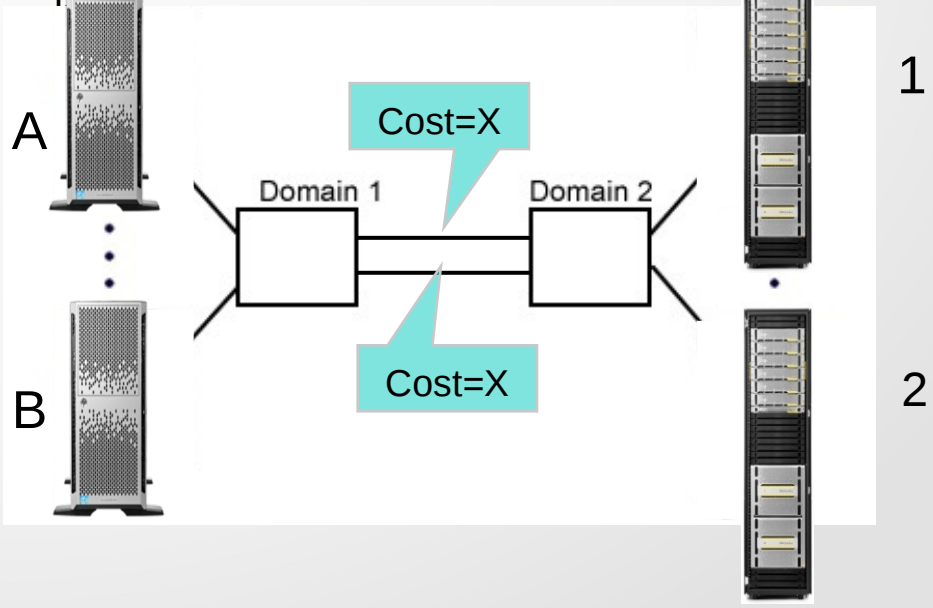
A→1 uses TOP or BOTTOM ISL for each exchange

A→2 uses TOP or BOTTOM ISL for each exchange

B→1 uses TOP or BOTTOM ISL for each exchange

B→2 uses TOP or BOTTOM ISL for each

In Order Delivery is preserved within the exchange and switches can more evenly *share* equal cost



ISL bandwidth aggregation

– Switch routing controller may use flow-based or exchange-based routing to aggregate the bandwidth across separate, equal cost links.

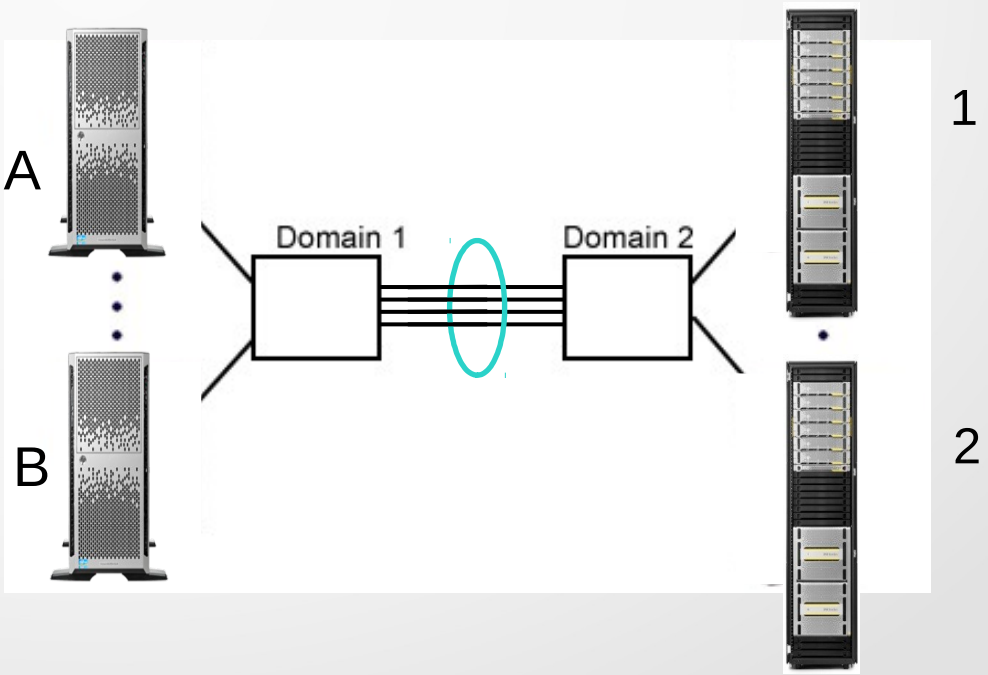
– Link failure causes fabric reconfiguration and topology rediscovery (rebuild LSD) is Masterless Trunking is not used

– Can more equally balance load by combining bandwidth of multiple, equal cost ISLs

– Ensures that frames arrive at the destination In Order

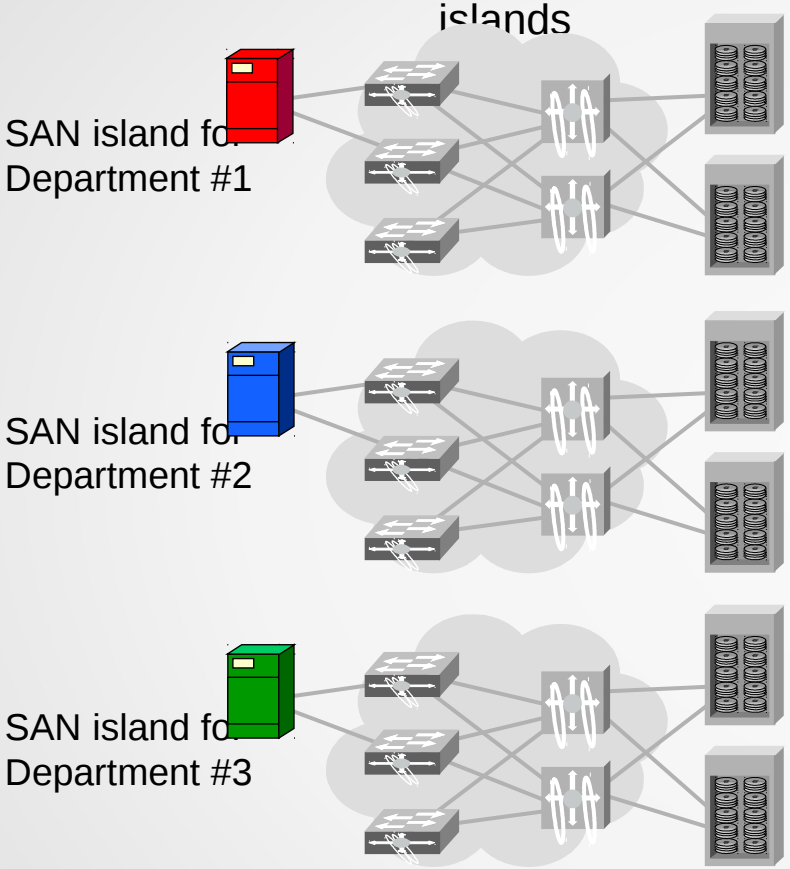
– Brocade – Trunk

– Cisco – Port Channel

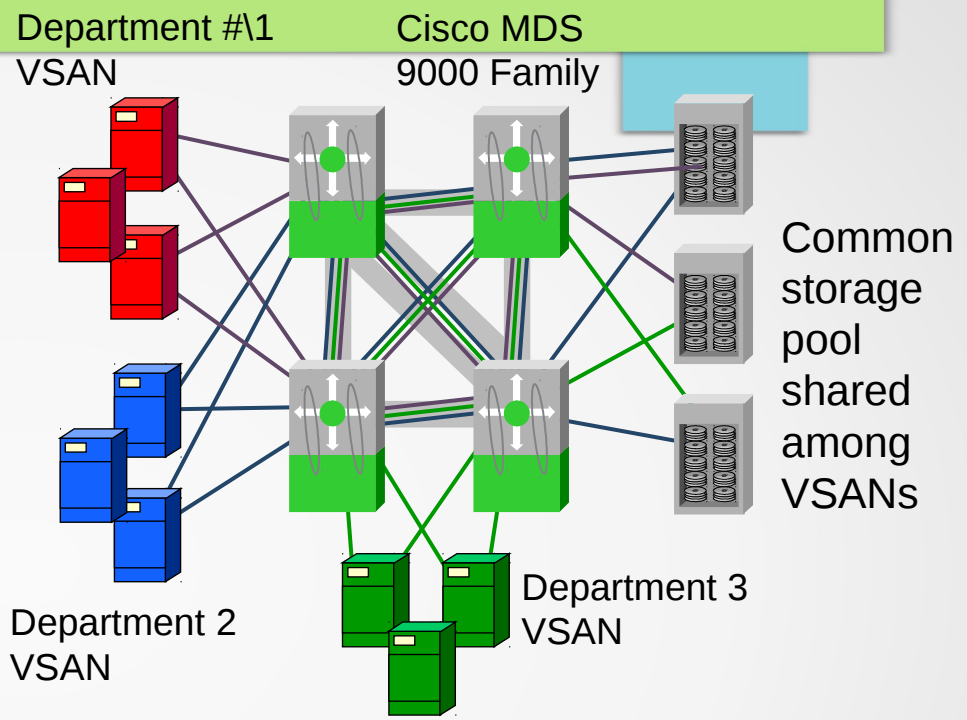


Cisco virtual SANs (VSANs)

Application/Department-based SAN islands



- Separate physical fabrics
- Over-provisioning ports on each island
- High number of switches to manage



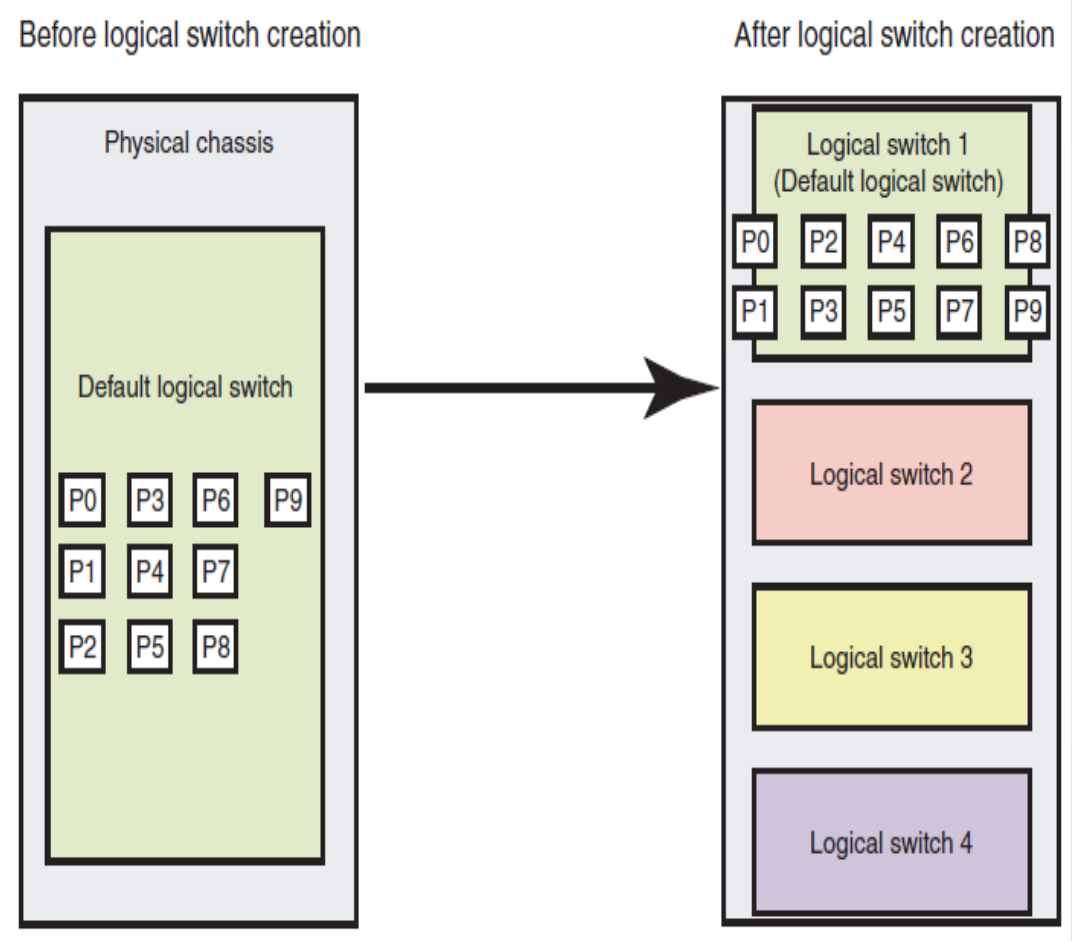
Collapsed fabric with VSANs

Less over-provisioning required—lower $$

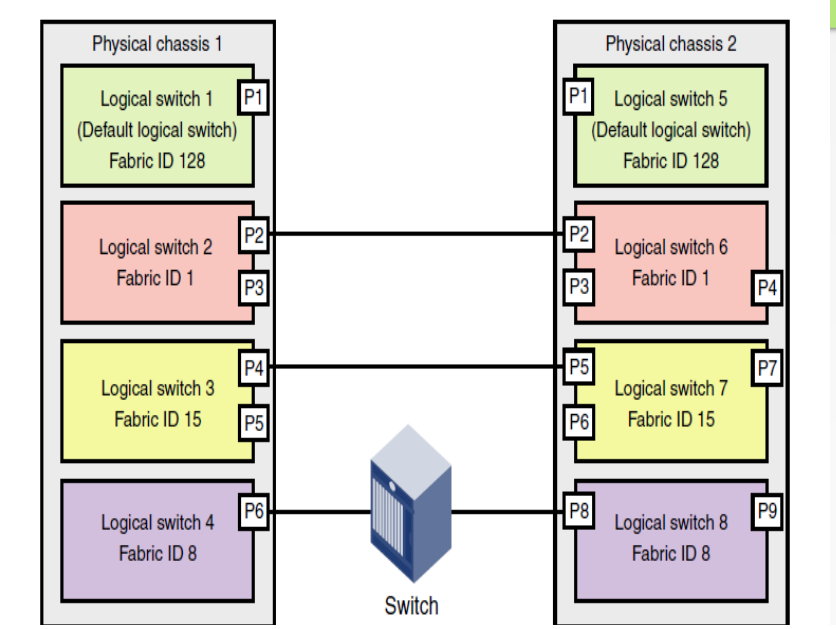
- Common redundant physical infrastructure
- Fewer switches to manage
- Move unused ports non-disruptively
- Addressing per VSAN
- Can overlap across FC backbone areas

Virtual Fabrics (Brocade)

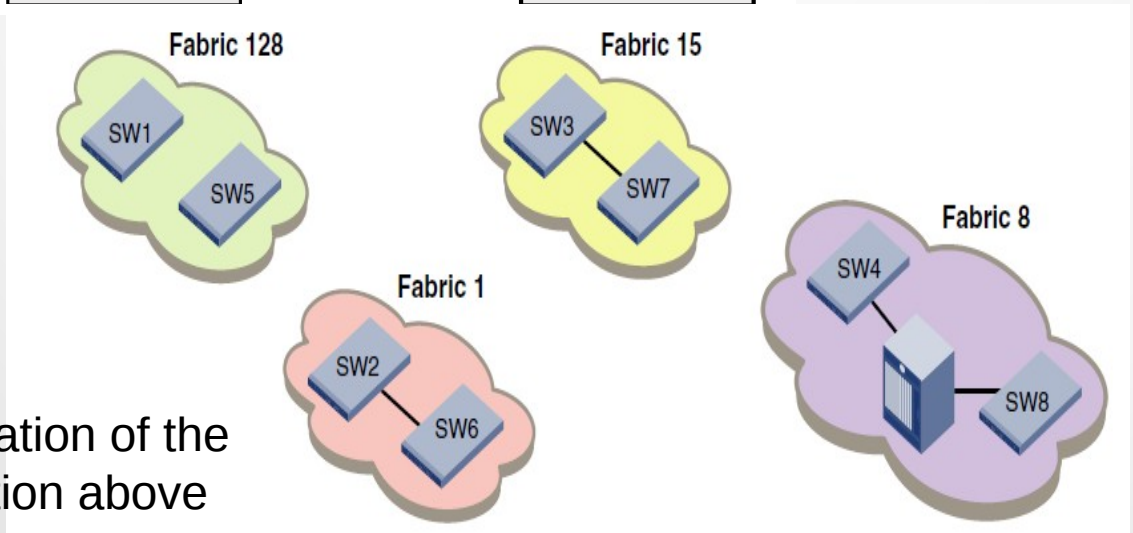
- Can create up to eight logical switches, depending on the switch model.
- Initially, all ports belong to the default logical switch. When you create additional logical switches, they are empty and you must assign ports to those logical switches



Logical fabrics and ISLs



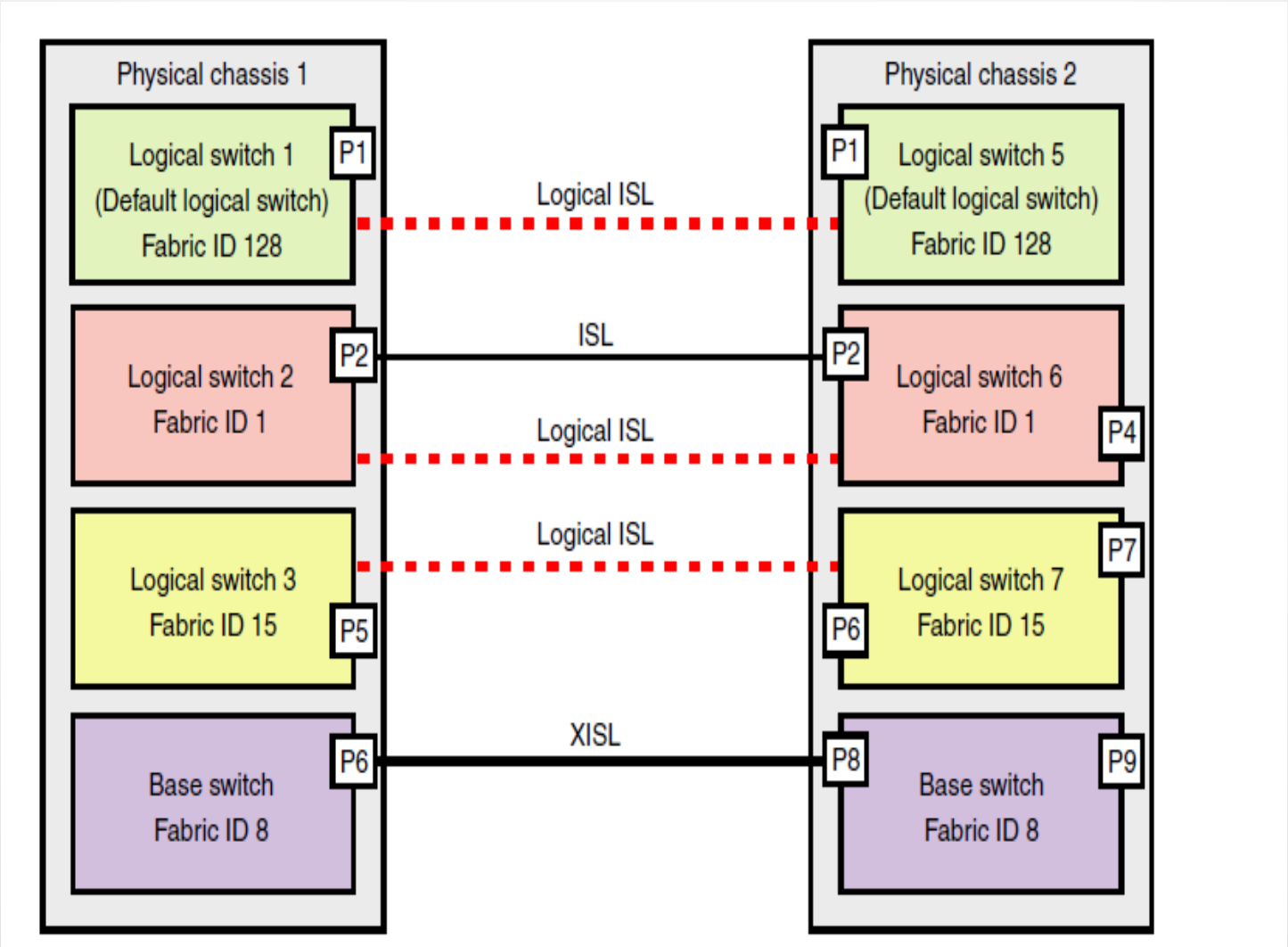
The ISLs between the logical switches are dedicated ISLs because they carry traffic only for a single logical fabric.



Logical representation of the configuration above

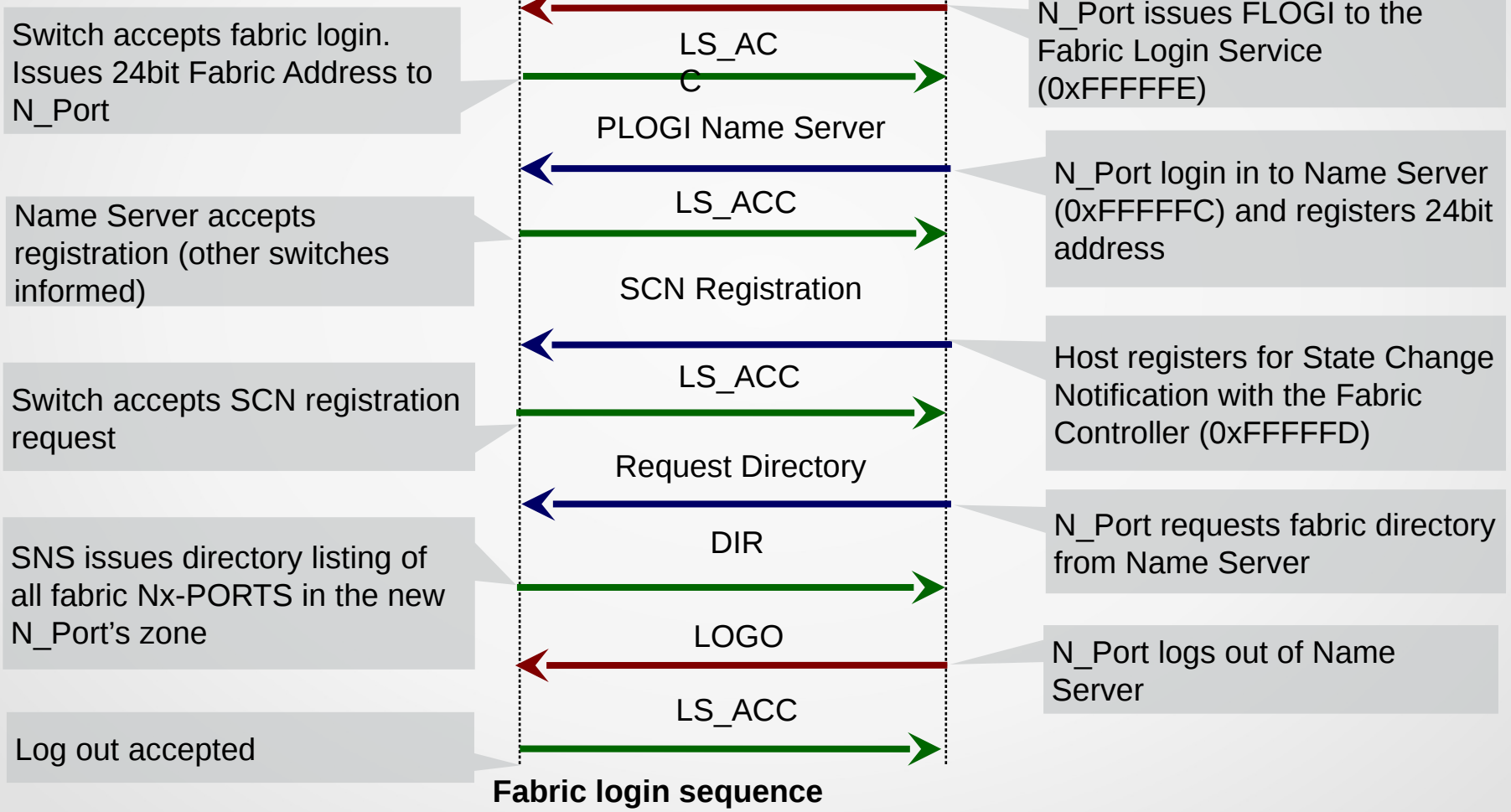
Logical fabric and ISL sharing

– Another way to connect logical switches is using extended ISLs and base switches.



Fabric login sequence

Switch



Fabric login sequence

N_Port login sequence

Storage (Target)

Switch

Host (Initiator)

PLOGI

N_Port issues PLOGI to all 24bit fabric addresses in the directory list one after another. PLOGI frame includes CoS and WWN's and buffer credits.

LS_ACC

N_Port responds with same information in payload.

N_Ports have exchanged service parameters and can now establish upper level protocol capabilities

PRLI

Process Login: Establishes ULP roles and capabilities, e.g. SCSI Initiator in this case.

LS_ACC

N_Port responds acknowledging SCSI support and target status thereby establishing an Image Pair.

N_Port login sequence

Well known addresses

– Well known addresses are all set within the reserved Domain_ID byte x'FF'.

x'FFFFF5' – Multicast Server
x'FFFFF6' – Clock Synchronisation Server
x'FFFFF7' – Key Distribution Server
x'FFFFF8' – Alias Server
x'FFFFF9' – QoS Facilitator (typically Class 4)
x'FFFFFA' – Management Server
x'FFFFFB' – Time Server
x'FFFFFC' – Directory Server
x'FFFFFD' – Fabric Controller
x'FFFFFE' – Fabric Login Server
x'FFFFFF' – Broadcast Address

x'FFFC*(01-EF)'* – Domain controller (each switch) address used in S_ID and D_ID frame header fields for Class F traffic where *(01-EF)* is the Domain_ID of that switch.

x'FFFFFD'
Required service in all fabrics

x'FFFFFE'
Required service in all fabrics

x'FFFFFC'
Optional service

x'FFFFFA'
Optional service with Fabric Zone Server as a sub-function

Fabric services

– Simple Name Service

– Fabric login

– Alias and multicasting (may not be implemented)

– State Change Notification

– Zoning

SNS: Port and Nodes attributes

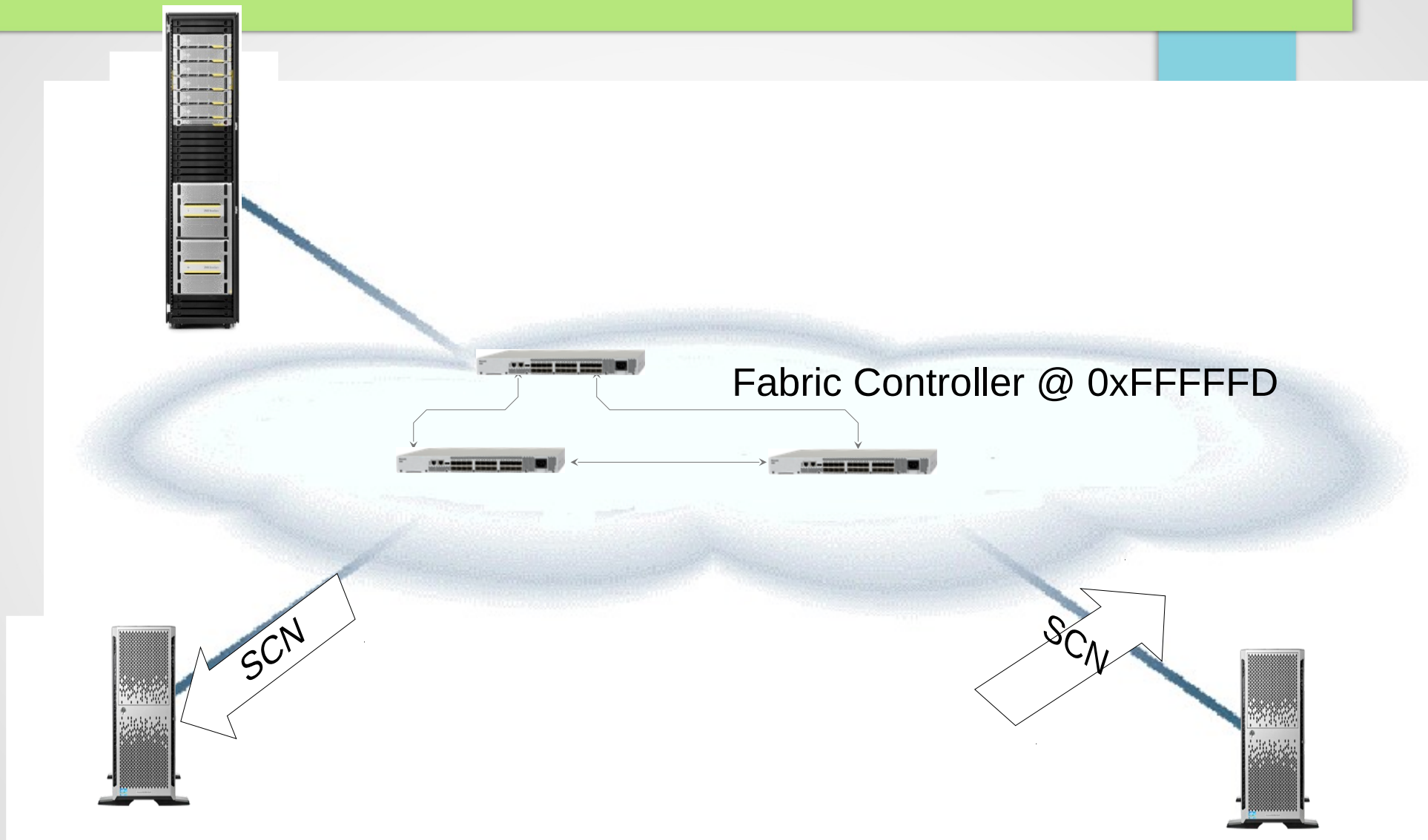
- Port attributes
  - Native port address ID
  - Port name (World Wide Port Name / WWPN)
  - Class of service supported
  - FC-4 types
  - Port type
- Node attributes
  - Node name (World Wide Node Name / WWNN)
  - Device name

Name server detail – Brocade

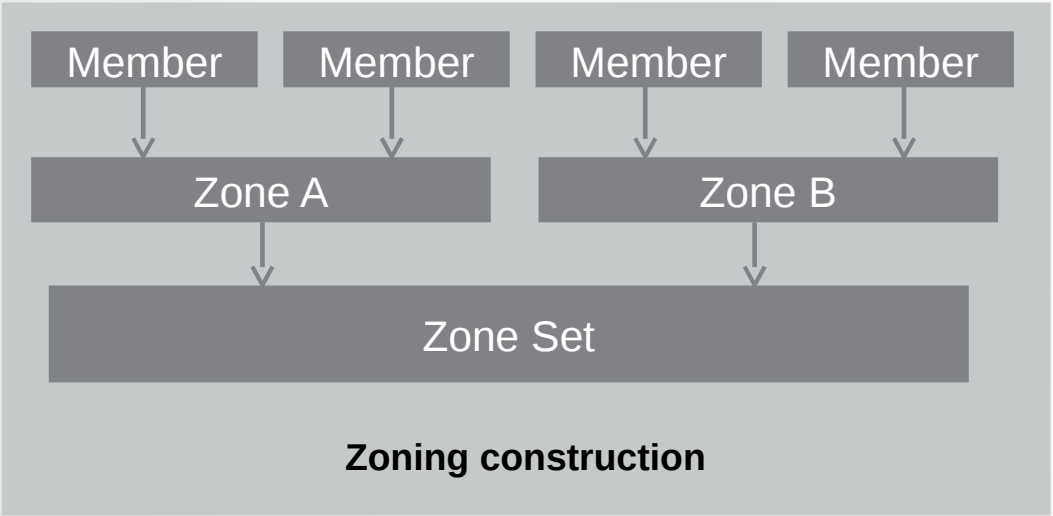
nsshow

```
EdgeA1:admin> nsshow
{
Type Pid    COS     PortName                NodeName                TTL(sec)
N    030400;    2,3;10:00:00:00:c9:4d:33:9d;20:00:00:00:c9:4d:33:9d; na
FC4s: FCP
NodeSymb: [47] "Emulex 394757-B21 FV1.91A2 DV5-5.20A10 BL20-362"
Fabric Port Name: 20:04:00:05:1e:03:62:57
Permanent Port Name: 10:00:00:00:c9:4d:33:9d
Port Index: 4
Share Area: No
Device Shared in Other AD: No
Redirect: No
N    030500;    2,3;10:00:00:00:c9:4d:24:dd;20:00:00:00:c9:4d:24:dd; na
FC4s: FCP
NodeSymb: [47] "Emulex 394757-B21 FV1.91A2 DV5-5.20A10 BL20-363"
Fabric Port Name: 20:05:00:05:1e:03:62:57
Permanent Port Name: 10:00:00:00:c9:4d:24:dd
Port Index: 5
Share Area: No
Device Shared in Other AD: No
Redirect: No
N    030600;      3;50:06:0b:00:00:29:e5:aa;50:06:0b:00:00:29:e5:ab; na
FC4s: FCP
Fabric Port Name: 20:06:00:05:1e:03:62:57
Permanent Port Name: 50:06:0b:00:00:29:e5:aa
Port Index: 6
Share Area: No
Device Shared in Other AD: No
Redirect: No
N    030700;      3;50:05:08:b3:00:90:f2:81;50:05:08:b3:00:90:f2:80; na
FC4s: FCP
NodeSymb: [23] "HP StorageWorks MSA1000"
Fabric Port Name: 20:07:00:05:1e:03:62:57
Permanent Port Name: 50:05:08:b3:00:90:f2:81
Port Index: 7
Share Area: No
Device Shared in Other AD: No
Redirect: No
The Local Name Server has 4 entries }
```

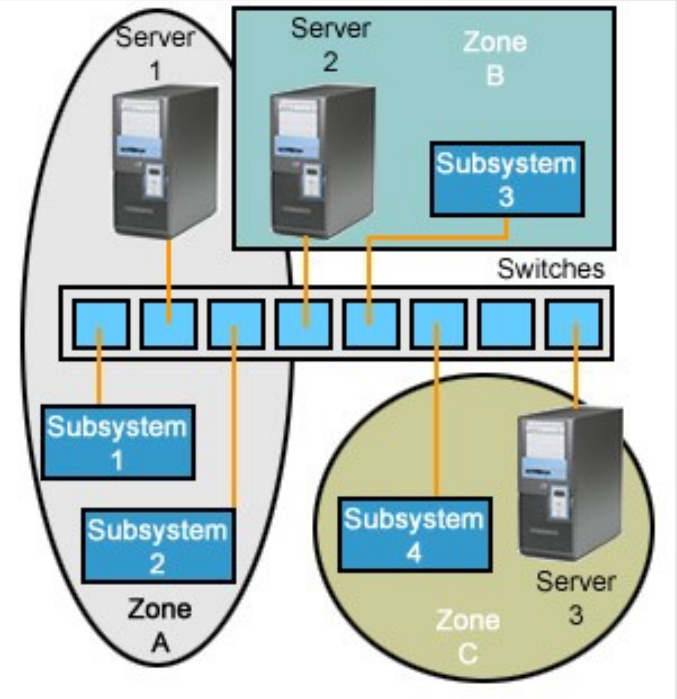
Registered State Change Notification



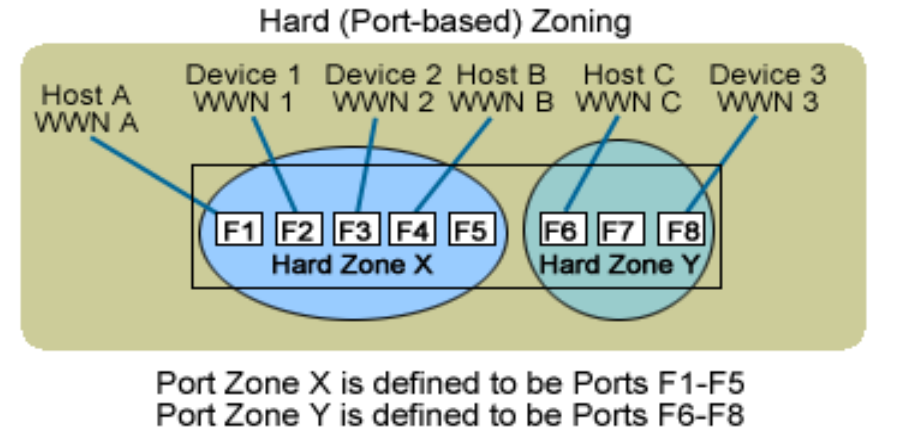
Зонирование «ткани»

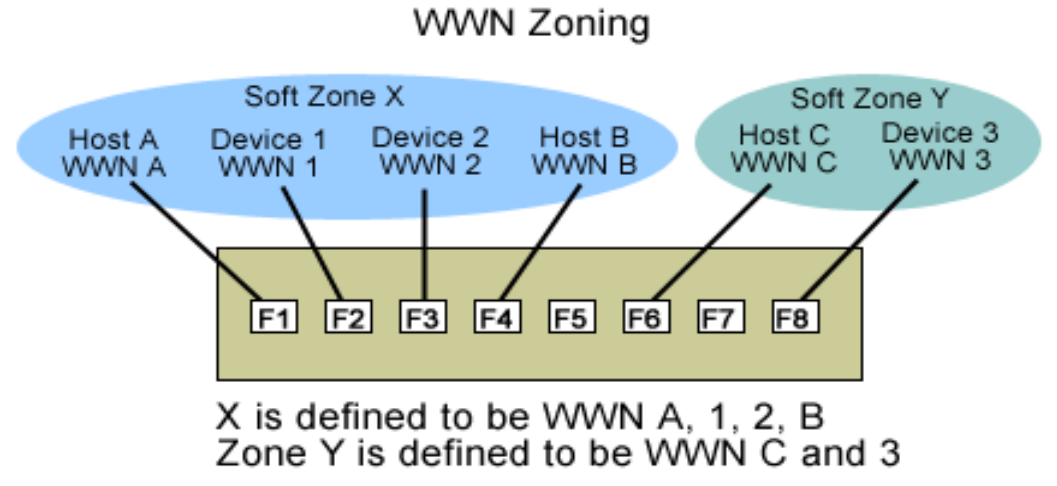


Zoning construction

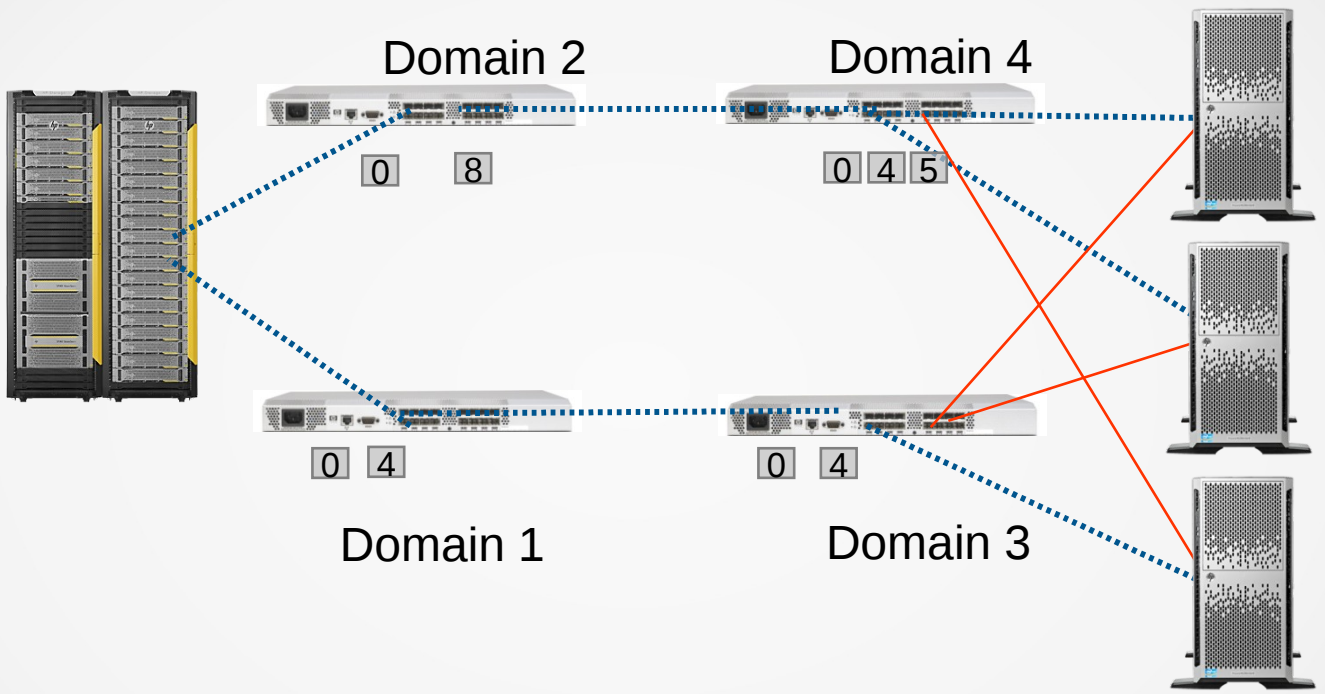


Типы Зон


Hard (Port-based) Zoning
Host A
WWN A
Device 1
WWN 1
Device 2
WWN 2
Host B
WWN B
Host C
WWN C
Device 3
WWN 3
F1
F2
F3
F4
F5
F6
F7
F8
Hard Zone X
Hard Zone Y
Port Zone X is defined to be Ports F1-F5
Port Zone Y is defined to be Ports F6-F8

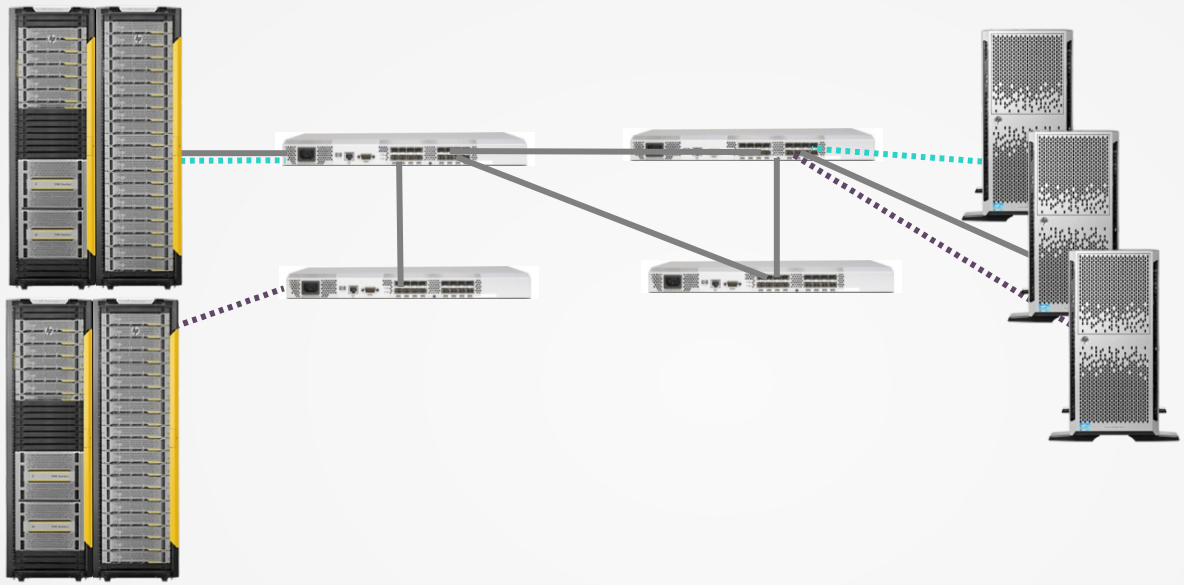

WWN Zoning
Soft Zone X
Host A
WWN A
Device 1
WWN 1
Device 2
WWN 2
Host B
WWN B
Soft Zone Y
Host C
WWN C
Device 3
WWN 3
F1
F2
F3
F4
F5
F6
F7
F8
X is defined to be WWN A, 1, 2, B
Zone Y is defined to be WWN C and 3

Traffic Isolation (TI) Zones




```
zone --create -t objtype [-o optlist] name -p "portlist
zone --create -t ti Dom1_3_zone -p "1,0; 1,4; 3,0; 3,4"
zone --create -t ti Dom2_4_zone -p “2,0; 2,8; 4,0; 4,4;4,5"
```

Quality of Service (QoS) Zones

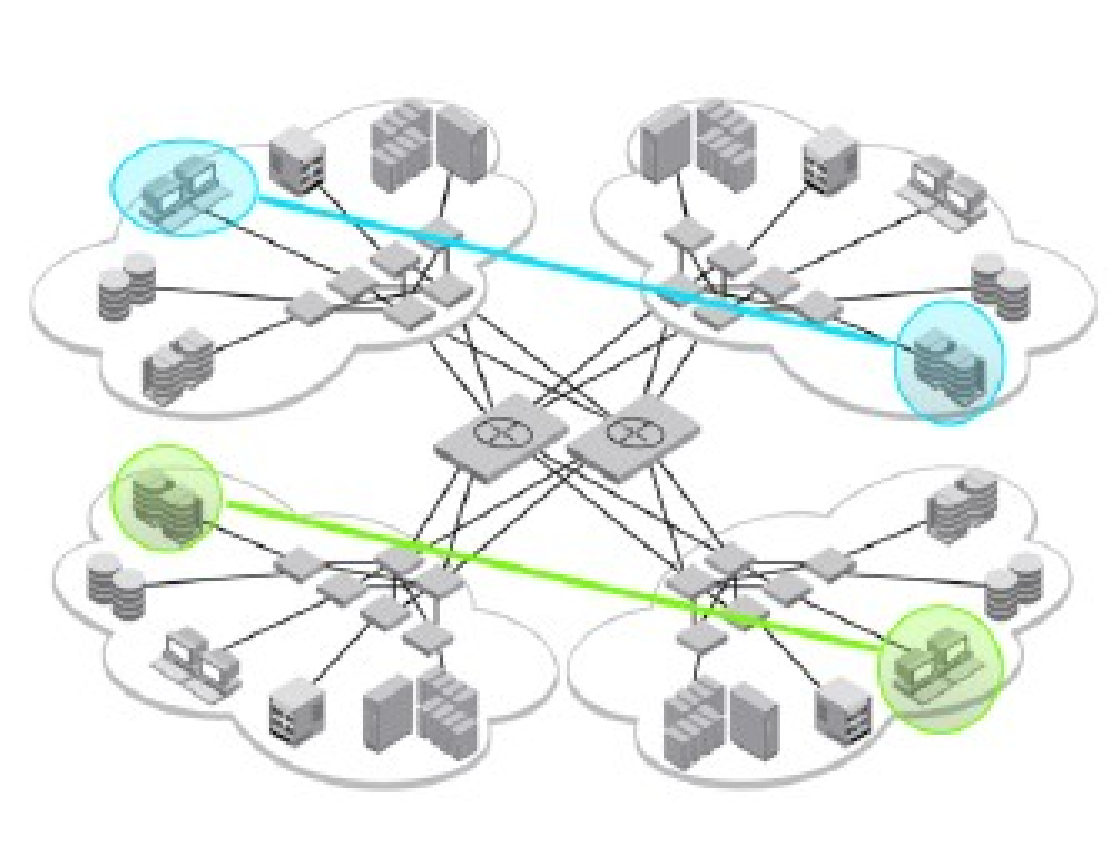


= High Priority

= Medium Priority

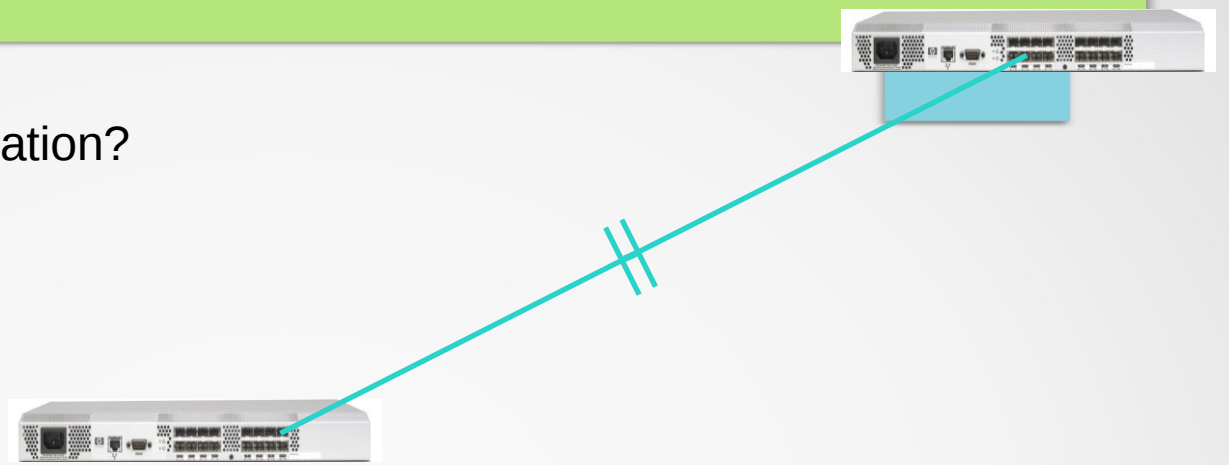
= Low Priority

LSAN Zones



Fabric Segmentation

- What causes Fabric Segmentation?
  - Zone type mismatch
  - Zone content mismatch
  - Zone configuration mismatch
  - Duplicate Domain IDs
  - Different ‘Time Out Value’ or other fabric parameters

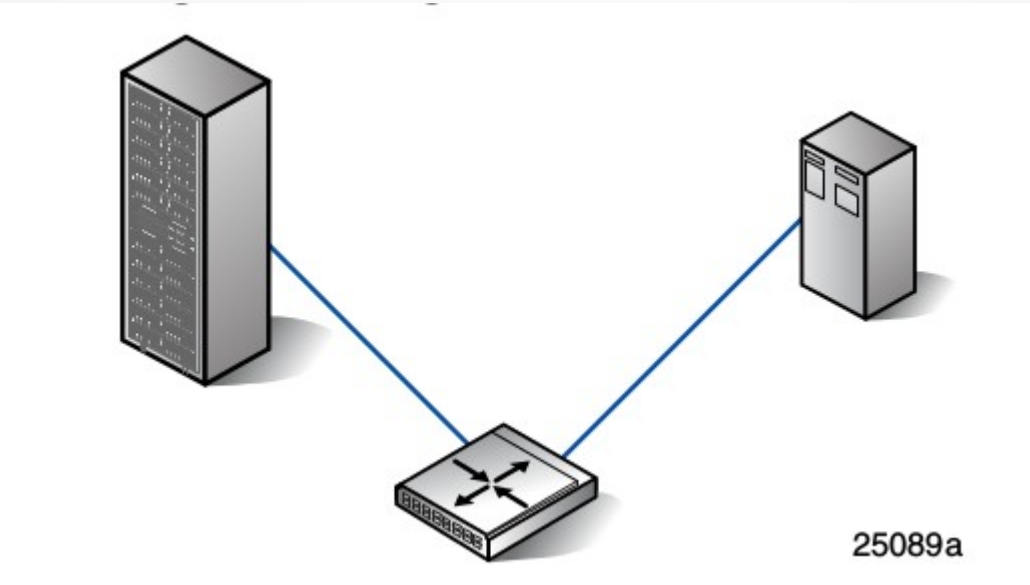


Switches use SW-ILS (Switch Internal Link Services) and Class F frames to exchange fabric parameters, zoning and routing information and reconfiguration data for distributed fabric services

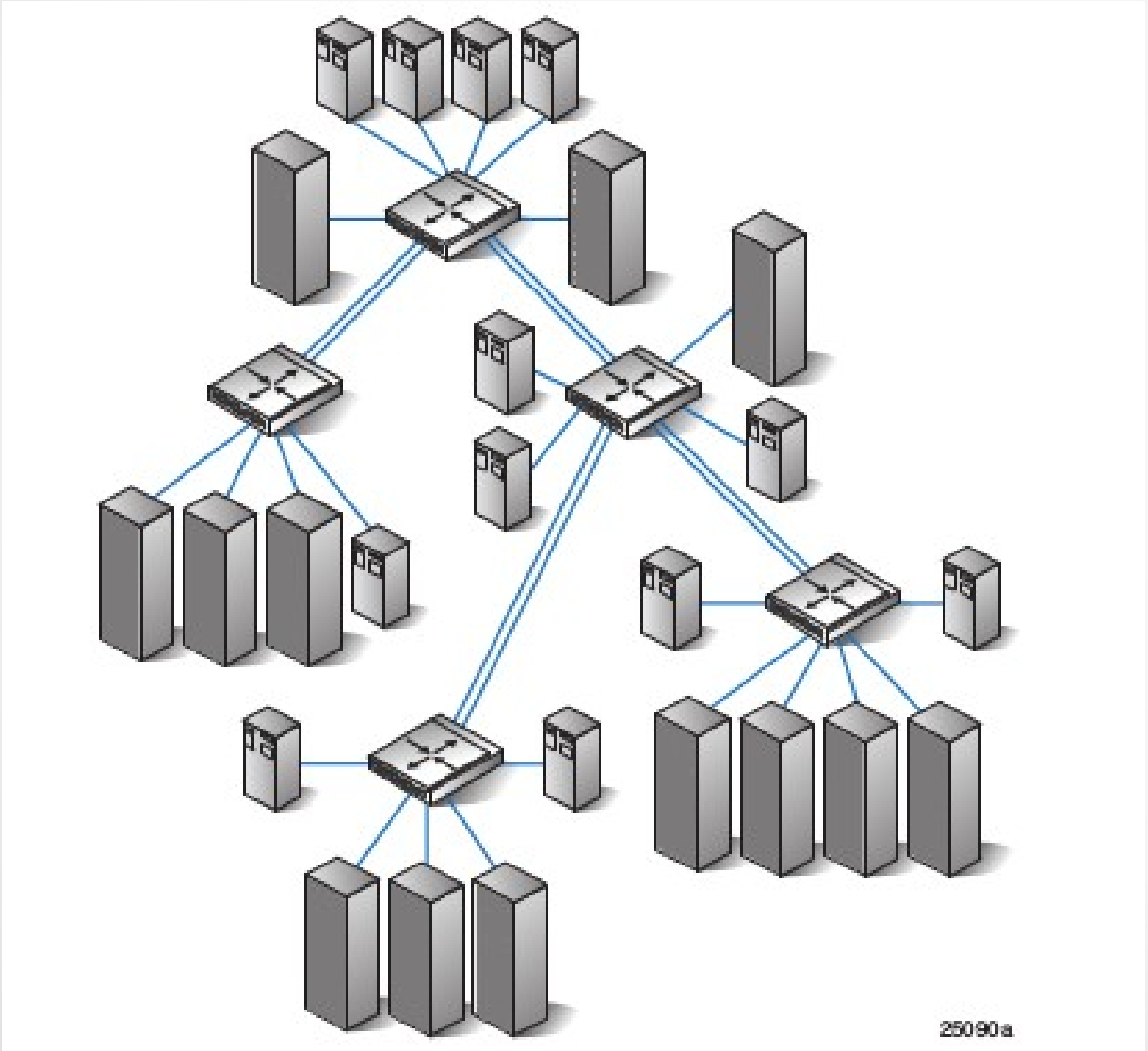
NOTE: Fabrics may segment if you make administrative changes to a single switch in a multi-switch fabric.

Различные топологии «ткани» («фабрики»)

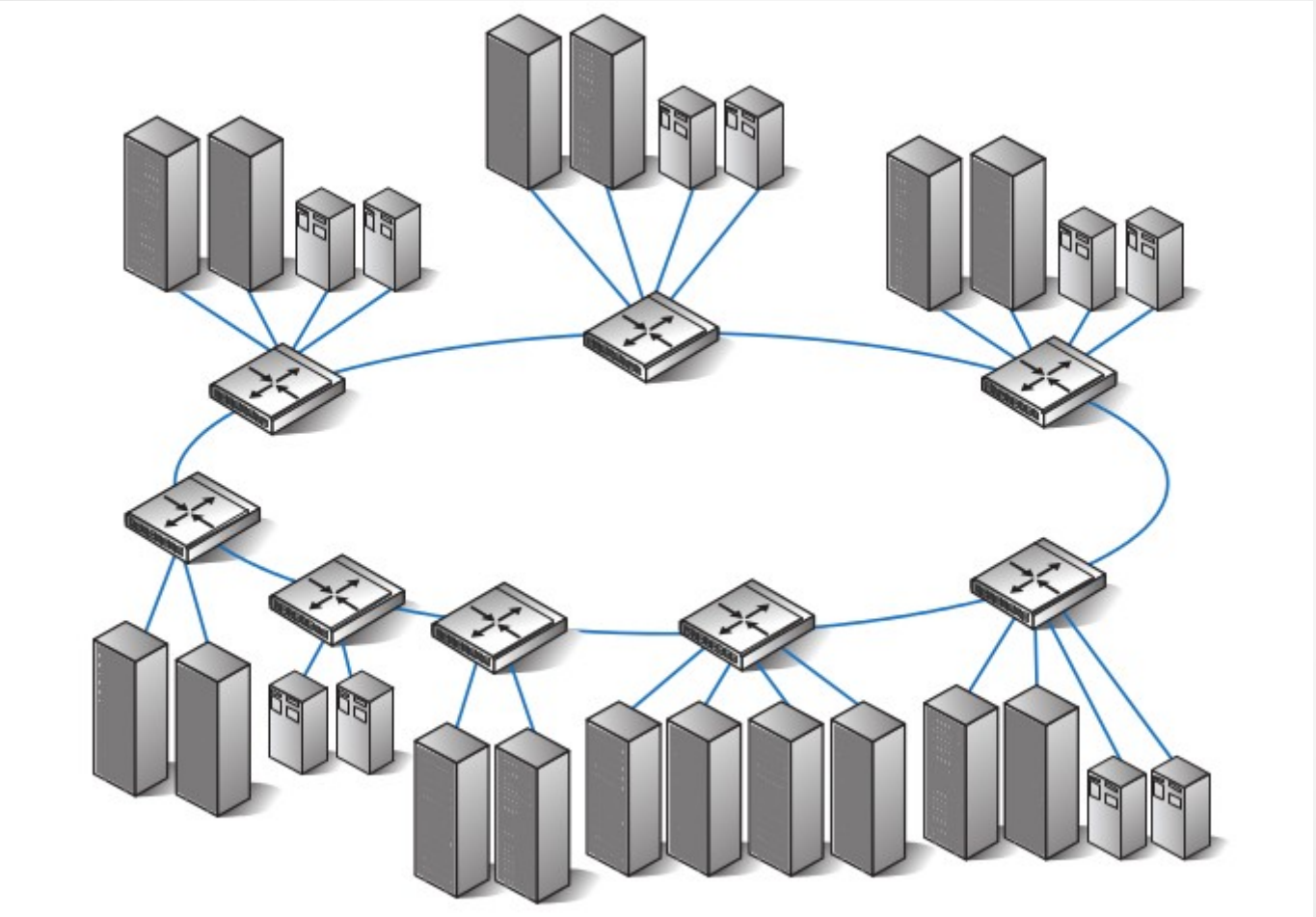
«Одно-коммутаторная» структура
Single-switch fabric

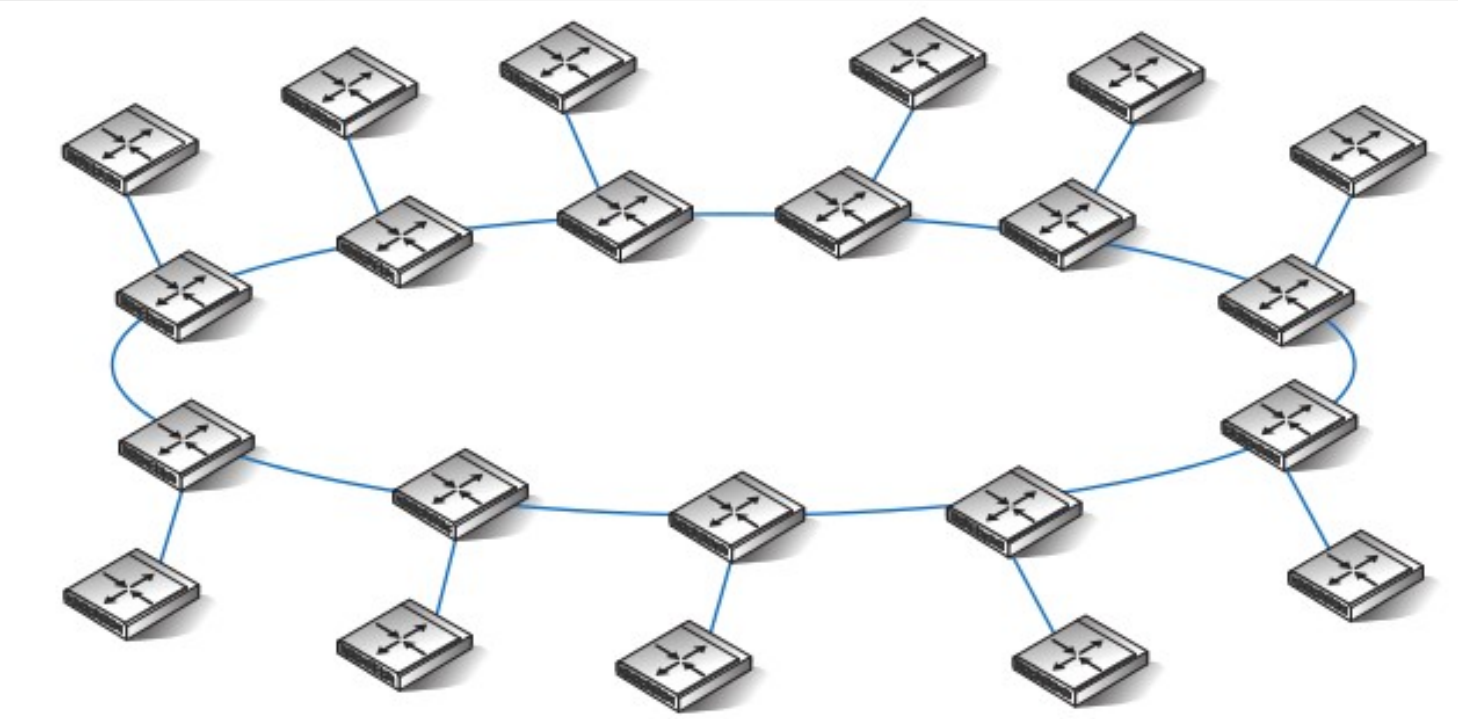


Древовидная или Каскадная структура Cascaded fabric

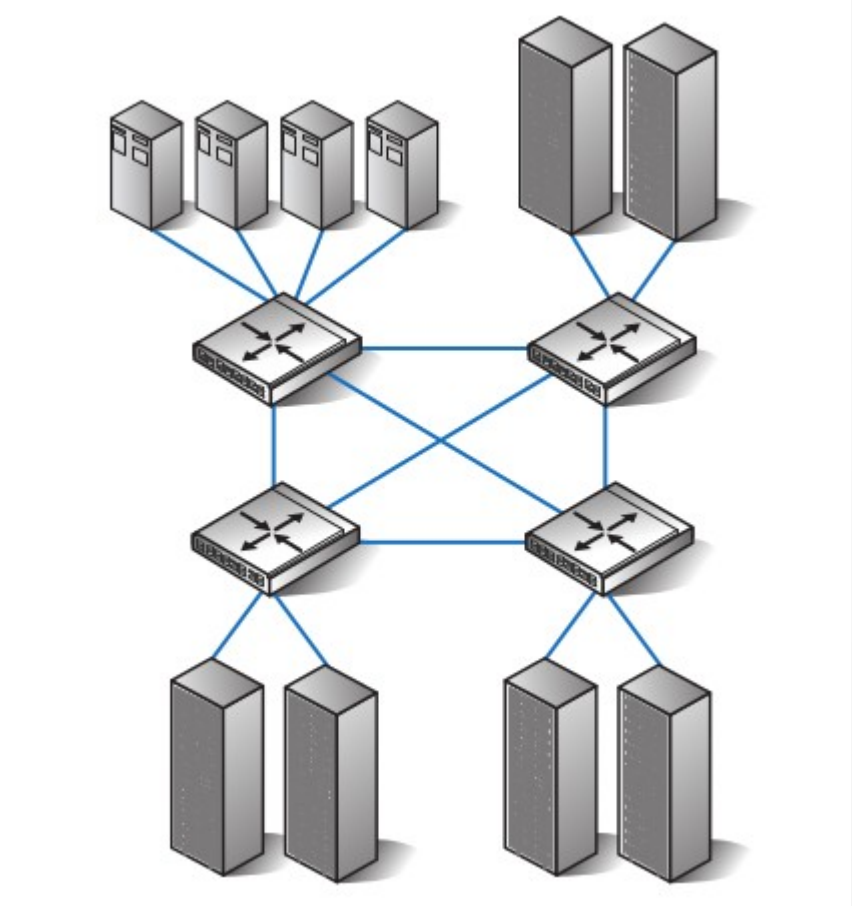


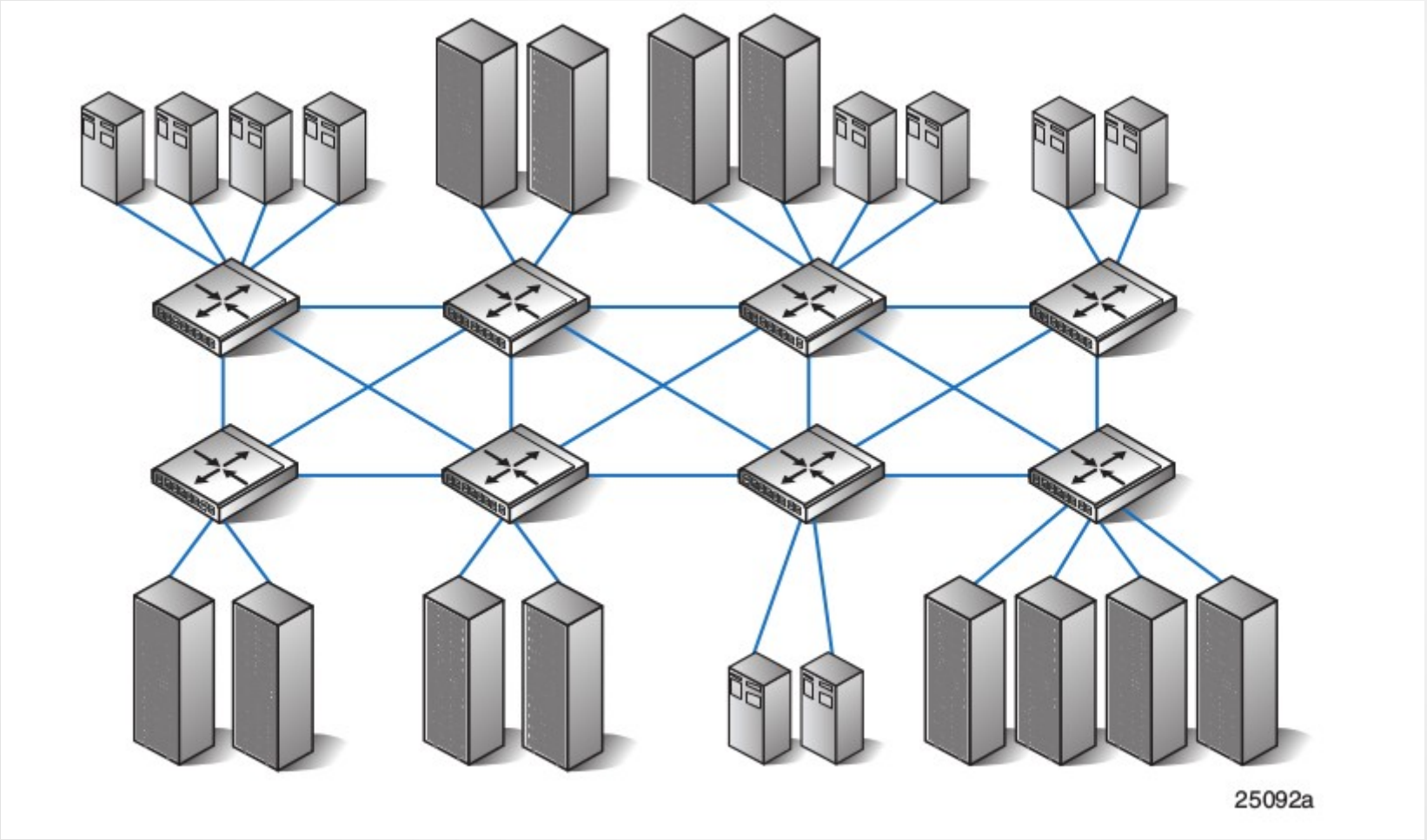
Кольцо
Ring fabric



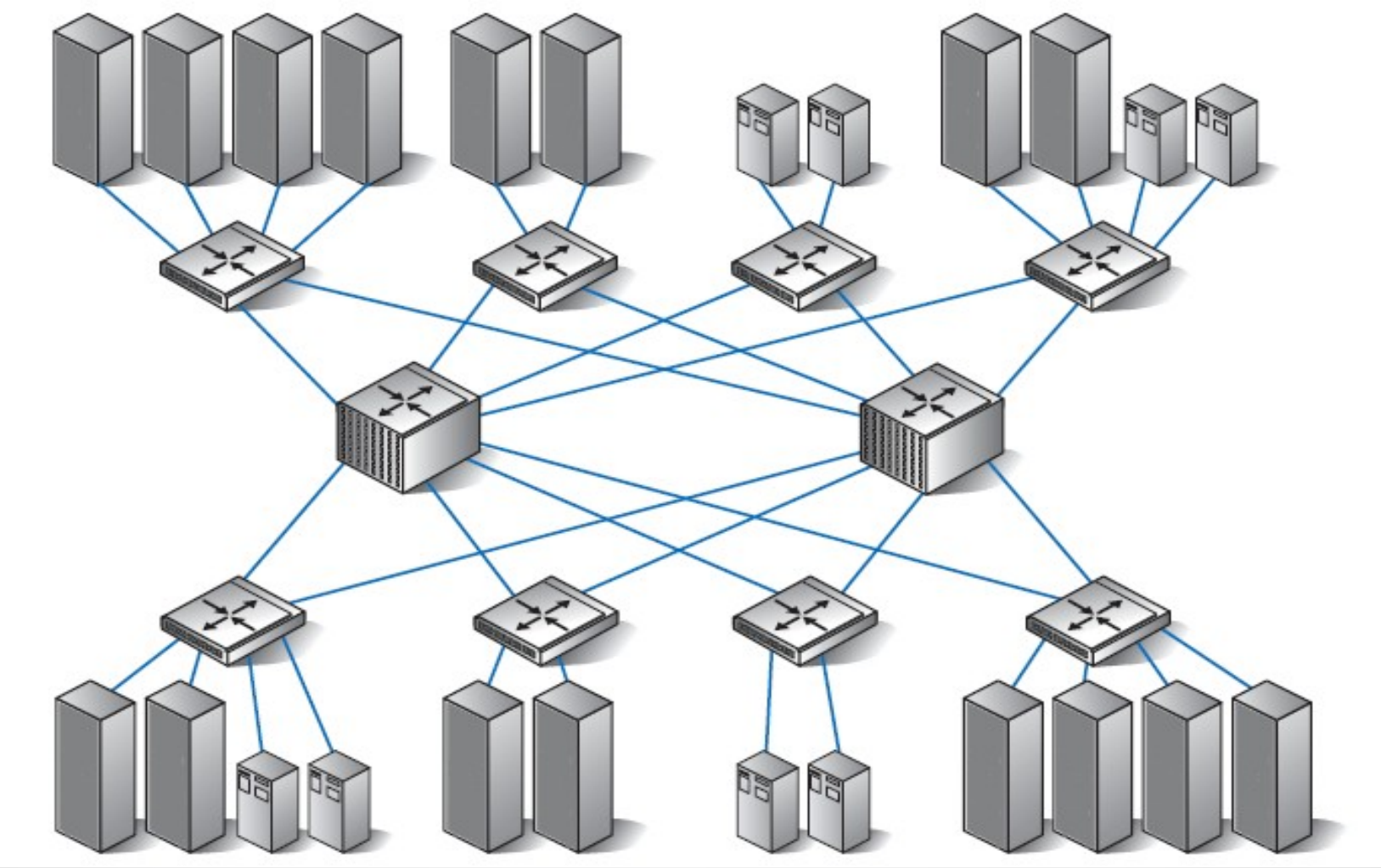


Решётка
Meshed fabrics

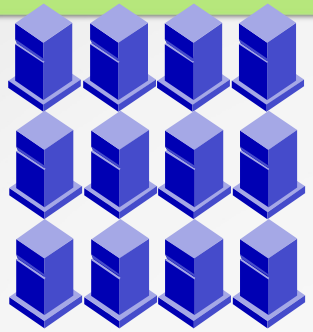



25092a

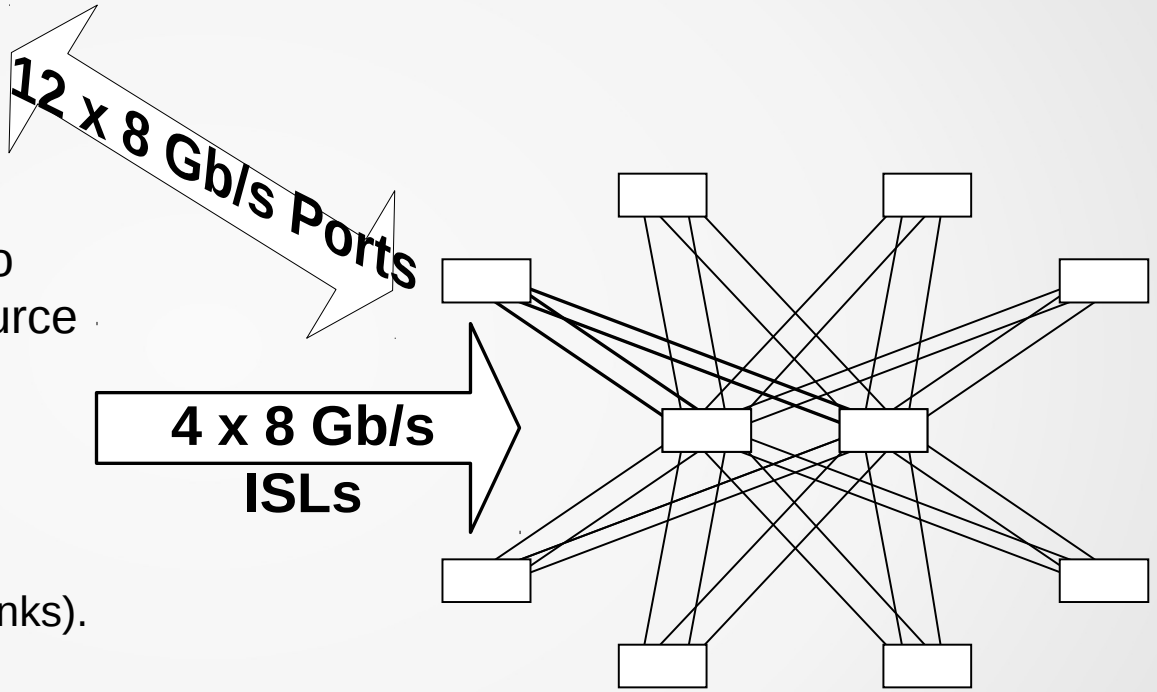
Core-edge fabric



ISL oversubscription

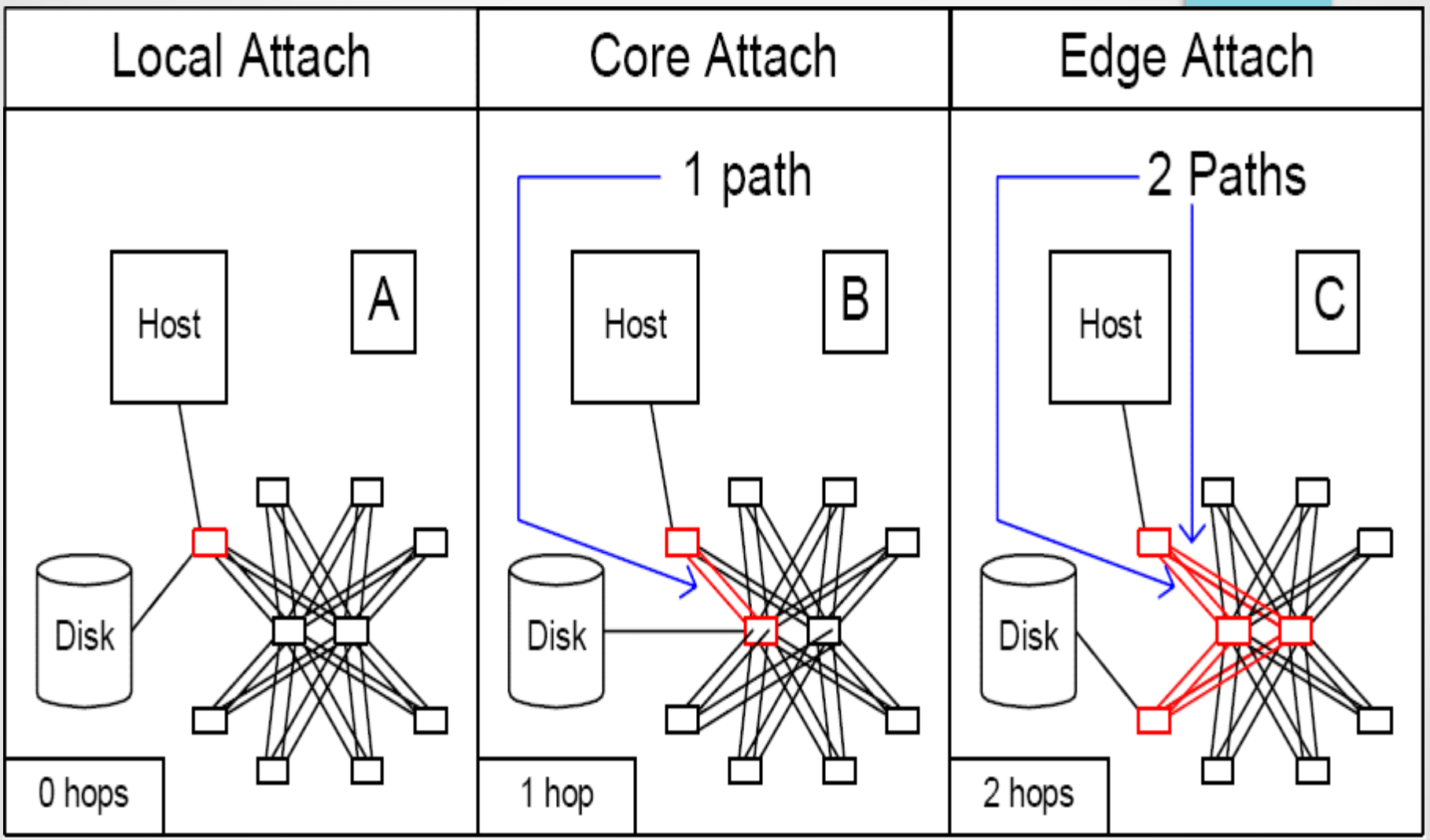


3 to 1 ISL oversubscription

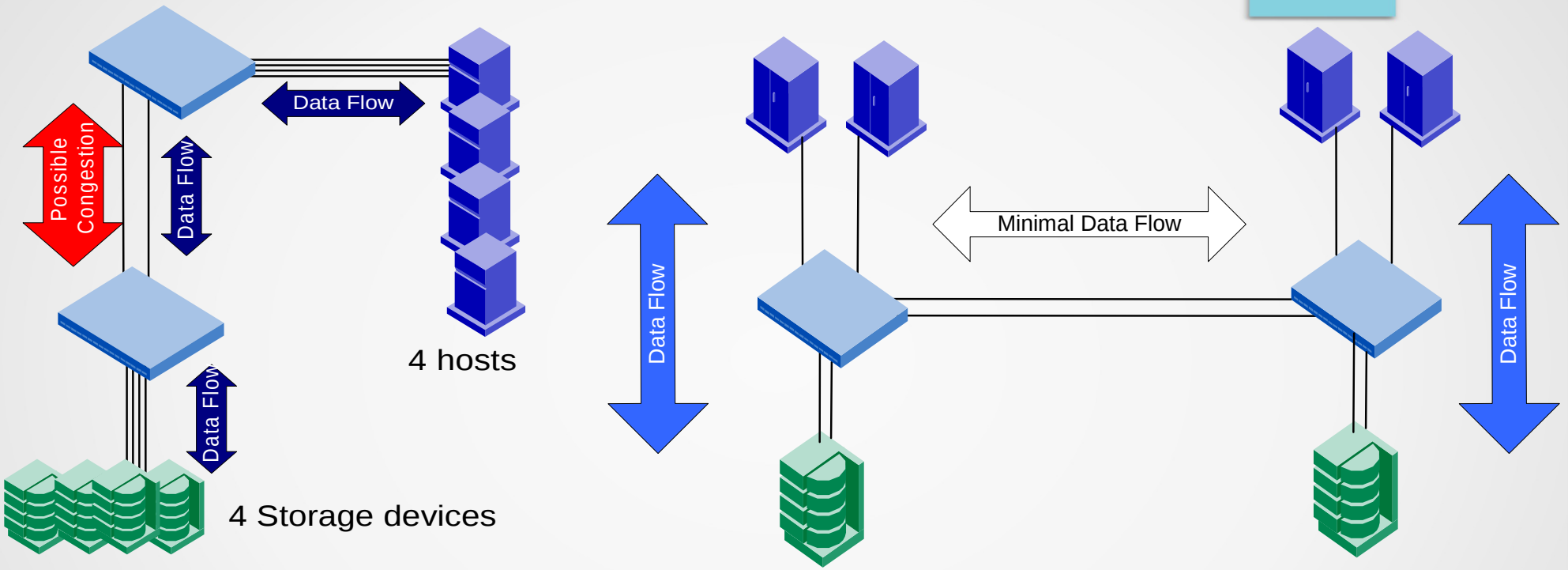


- Oversubscription - A condition in which more devices might need to access a resource than that resource can fully support.
- Oversubscription ratios:
  - Host to ISL (1:1, 3:1, 7:1)
  - Edge switch to core switch (use trunks).
  - Storage to ISL (7:1, 15:1)

Device attachment points



Data Locality

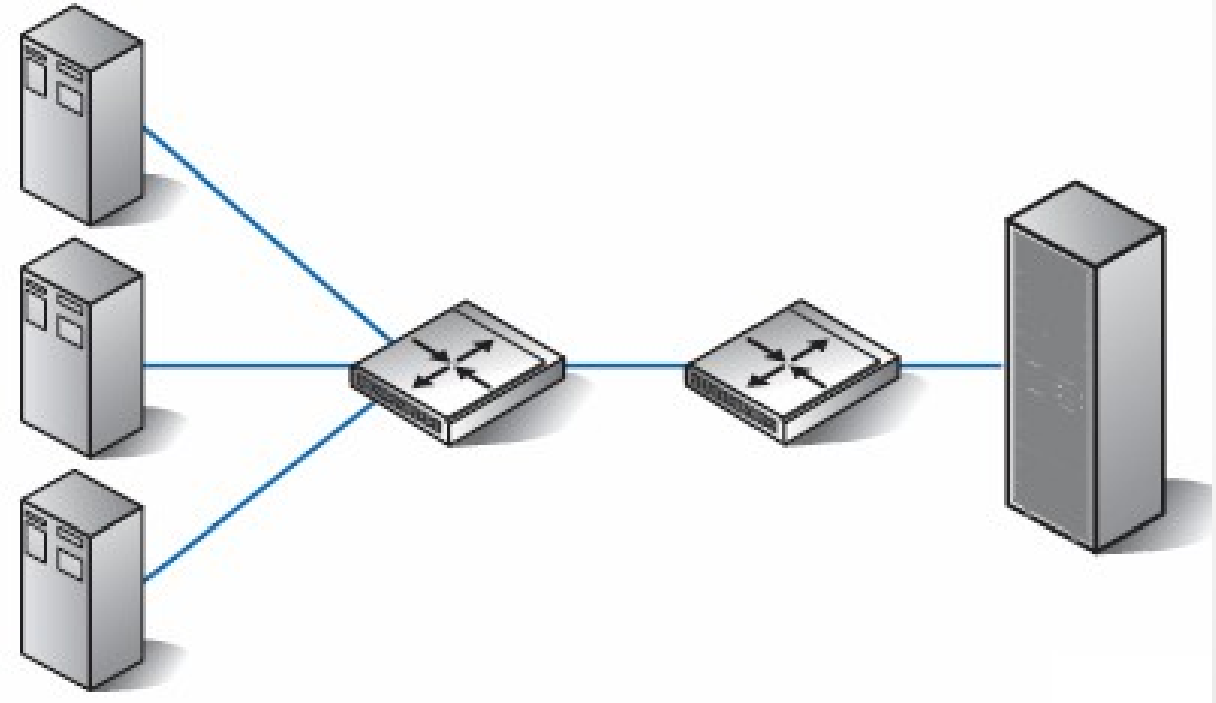


– Data path includes one or more hops
– Potential for ISL oversubscription

– Design to keep data ‘local’
– Minimises utilisation of ISLs
– Improves performance

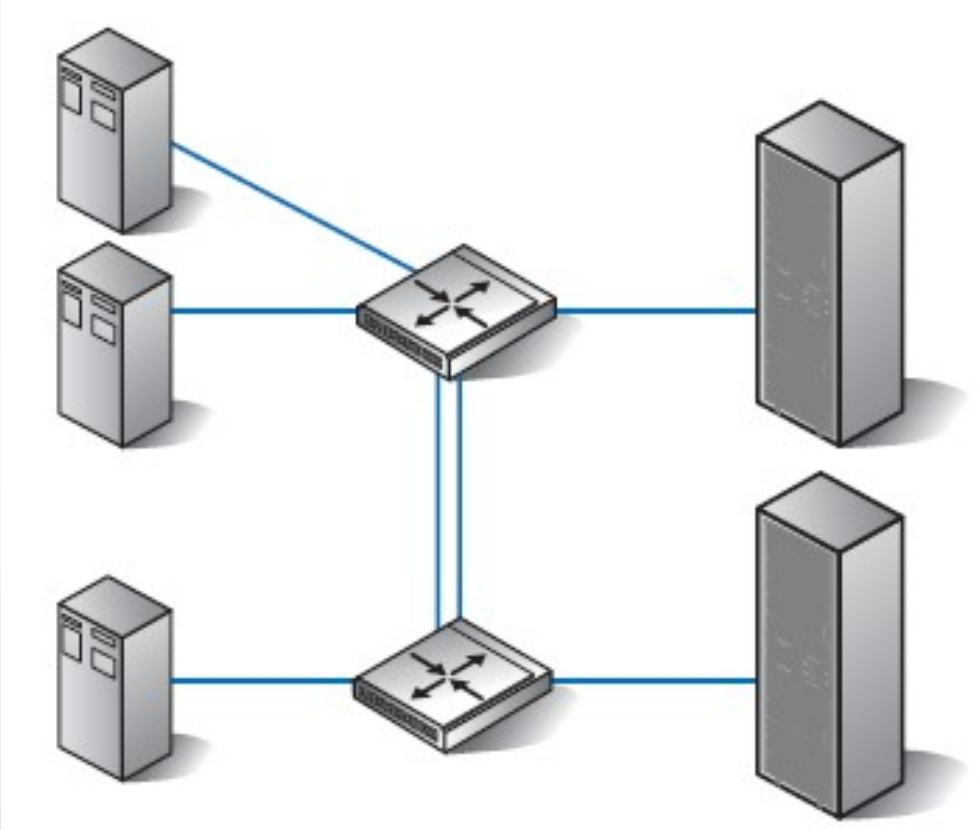
Level 1 – Single connectivity fabric

- Level 1 provides maximum connectivity but does not provide fabric resiliency or redundancy
- Each switch has one path to other switches in the fabric. Each server and storage system has one path to the fabric



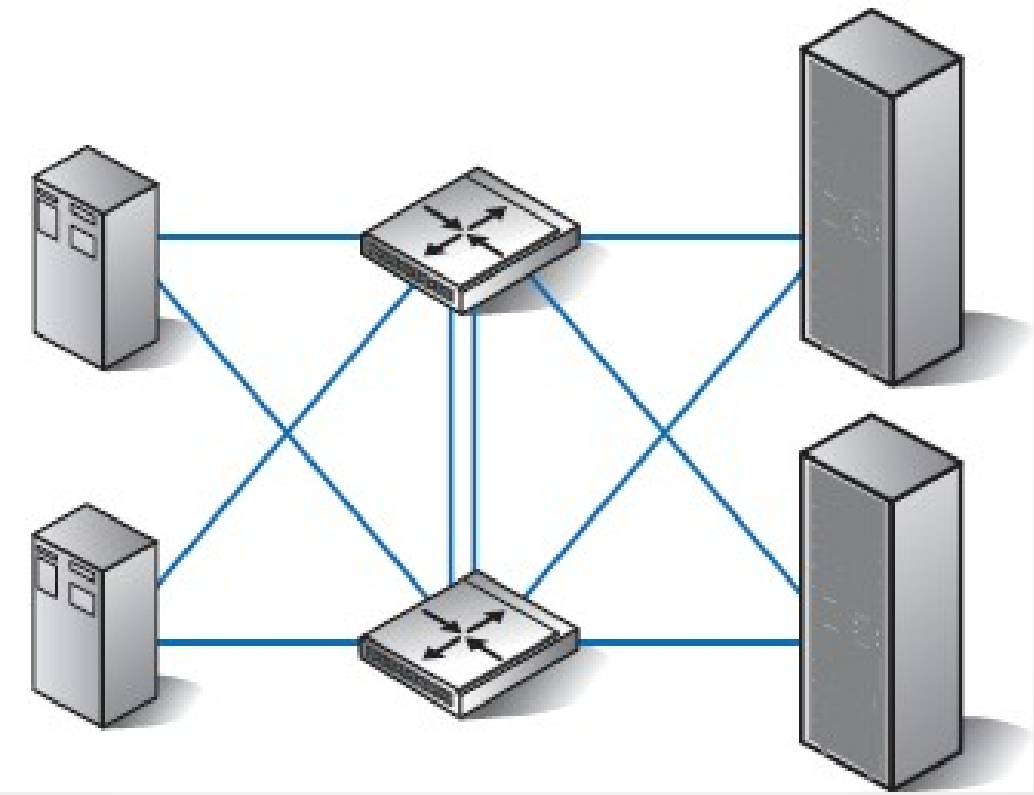
Level 2 – Single resilient fabric

- Level 2 provides fabric path redundancy by using:
  - Multiple ISLs between switches
  - Multiple paths to all switches in the fabric, or both
- Each server and storage system has one path to the fabric.



Level 3 – Single resilient fabric with multiple device paths

– Level 3 is the same as level 2 but also provides multiple server and storage system paths to the fabric to increase availability.



Level 4 – Multiple fabrics and device paths (NSPOF)

– Level 4 provides multiple data paths between servers and storage systems, but unlike level 3, the paths connect to physically separate fabrics.

– This level ensures the highest availability with no-single-point-of-failure (NSPOF) protection.