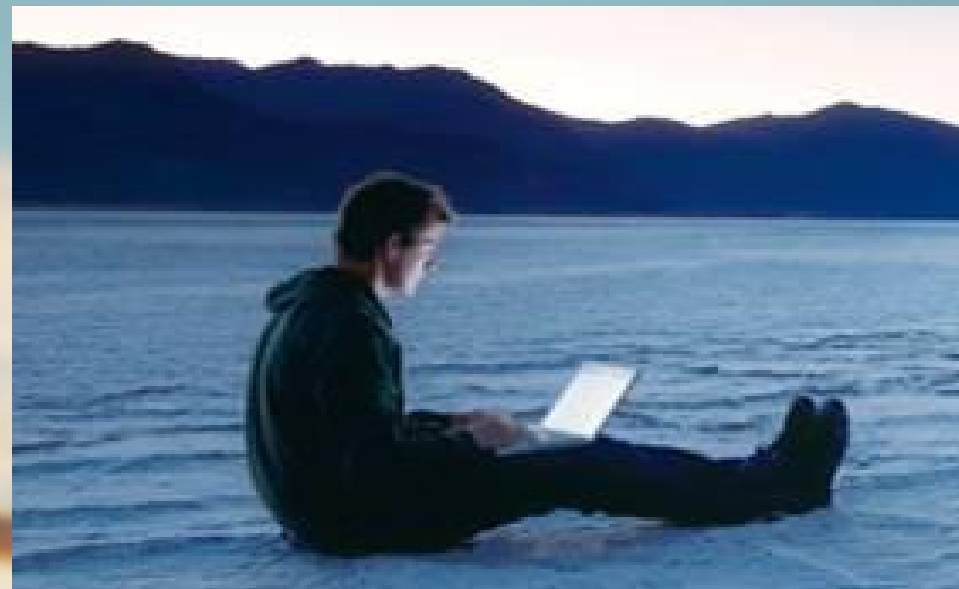




# SAN网络和融合网络介绍



- 周涛
- QQ: 53408031 Mobile: (86)18611846551
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站: [www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024 腾讯课堂: <https://ielab.ke.qq.com/>



## SAN网络和融合网络介绍

周涛

QQ: 53408031

Mobile:

(86)18611846551

Site: [www.ie-lab.cn](http://www.ie-lab.cn)

YY直播: 58761024

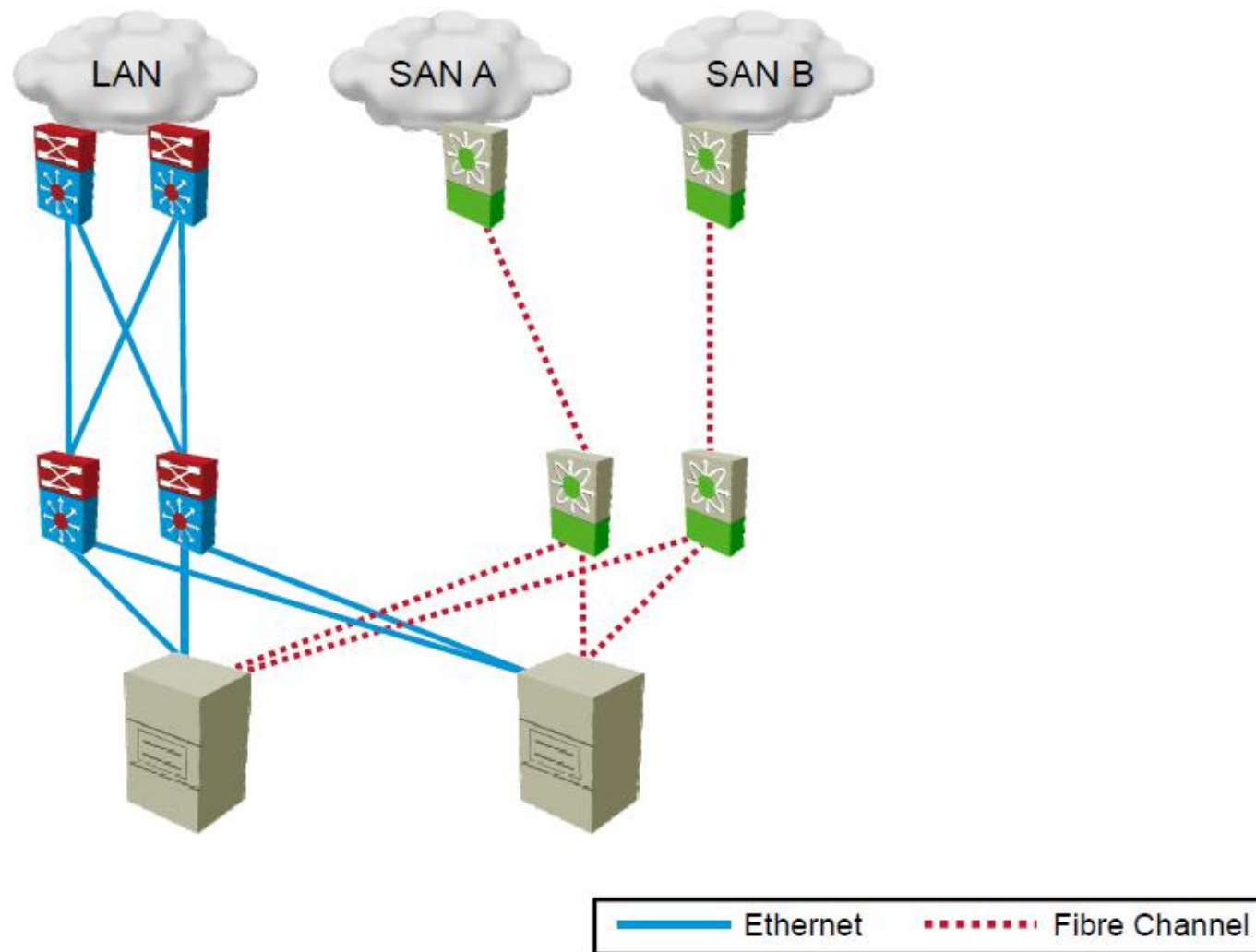
## SAN网络和融合网络介绍

- 1. 存储简介
- 2. Fibre Channel 术语
- 3. Fibre Channel 工作原理
- 4. 融合网络
- 5. FCoE的术语
- 6. FCoE的工作原理
- 7. FCoE的标准集
- 8. 传统数据中心网络向融合网络迁移



# 传统的数据中心网络

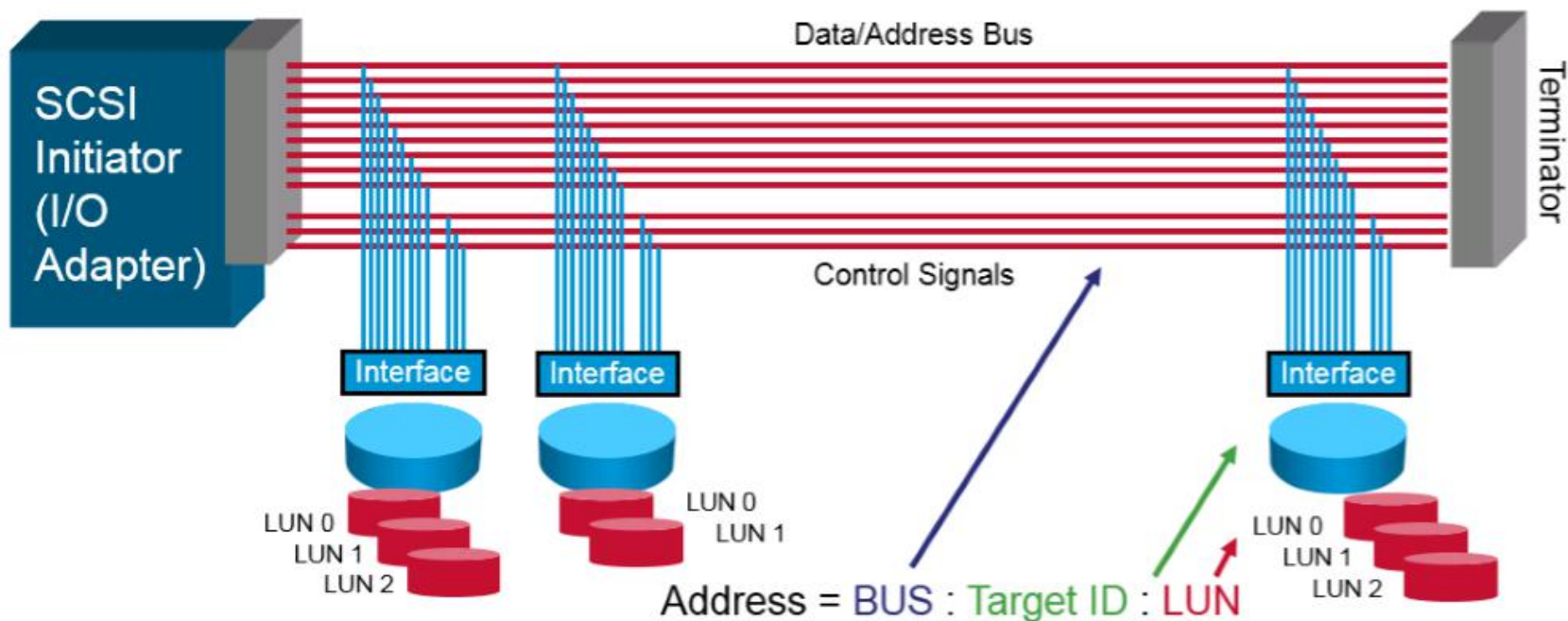
Traditional Deployment





## 什么是存储

- 通过介质存储信息(数据), 我们现在常用的是硬盘
- **SCSI Protocol**主要负责在**Initiators**和**Target**之间传输命令、状态和数据块





## SCSI Protocol:

- **SCSI 使用并行总线架构:**
  - 数据同时从多条线路上传输
  - Half-duplex – 在总线上要么传输数据，要么接收数据

并行传输的缺陷:

- 通信双方之间的距离足够短
- 导线电阻不均衡，导致速度出现差距
- 价格
- 干扰

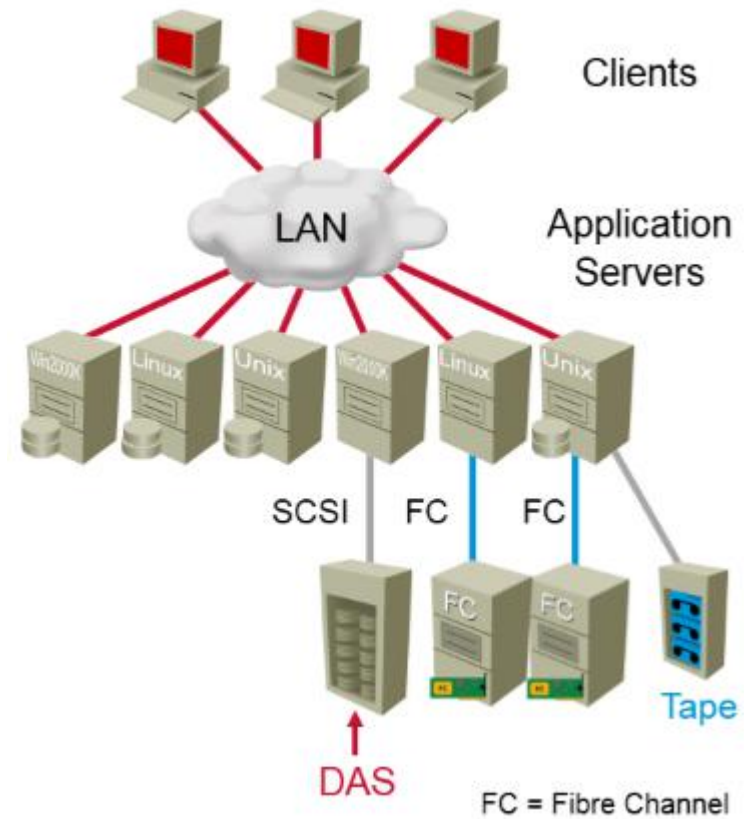


## 存储网络

把硬盘从主机内拿出来:

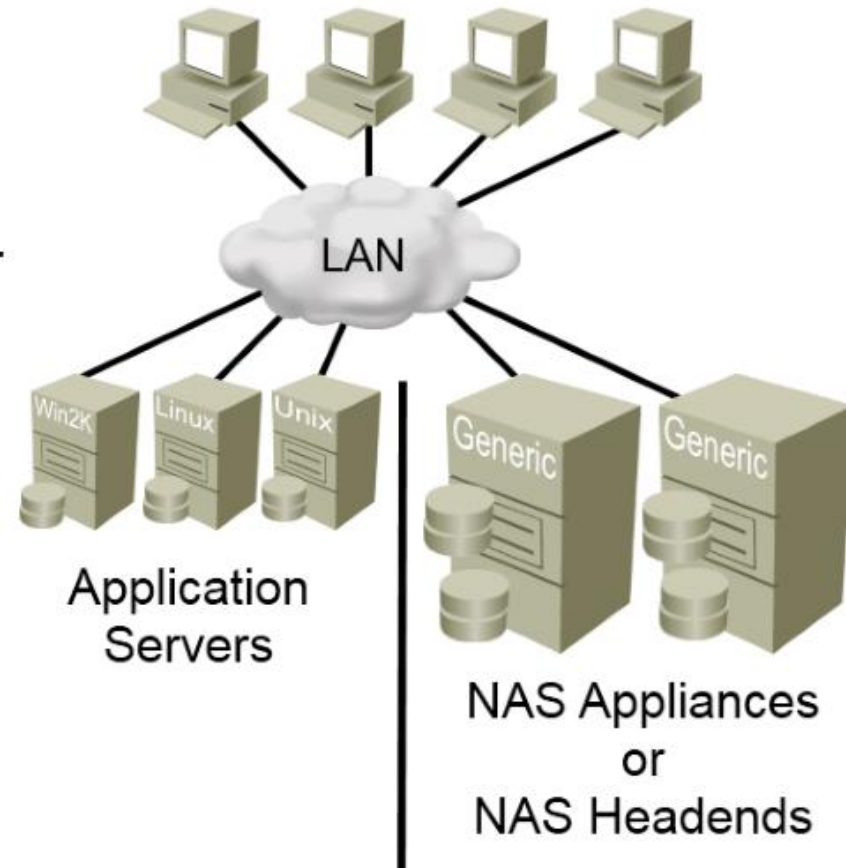
- DAS
- NAS
- Fabric

- DAS – Direct-Attached Storage 直连存储
- 存储直接连接在服务器后面，很少移动
- 扩展性很差
- 不可能进行有效的共享



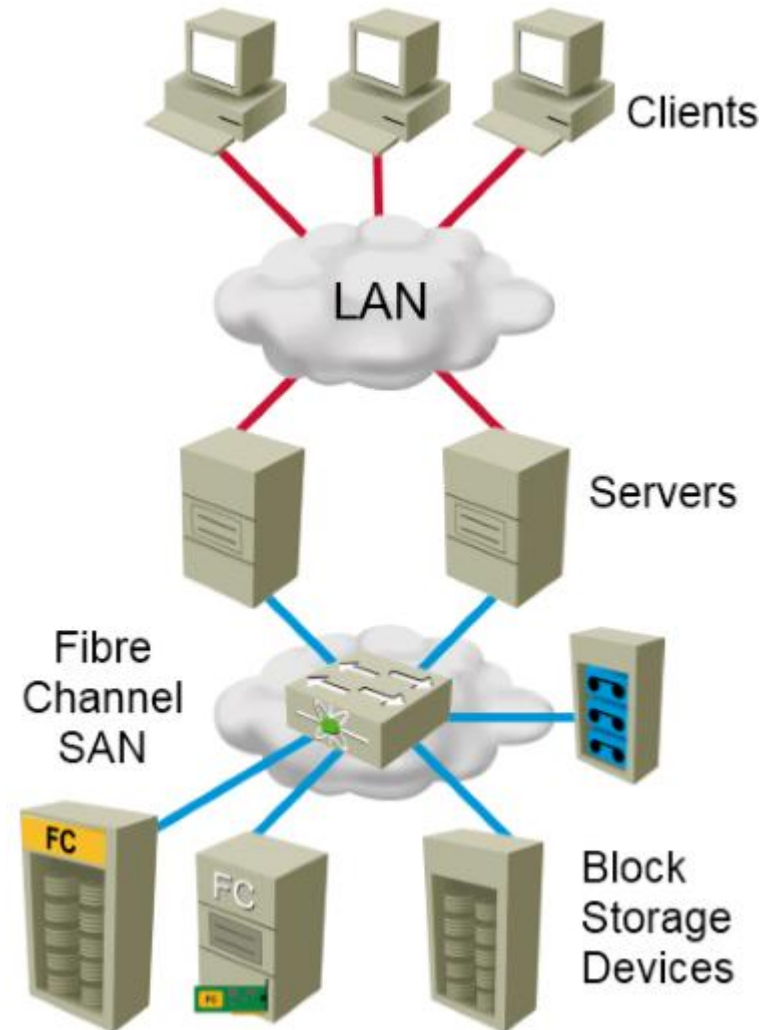
## 存储网络 – NAS

- NAS – Network-Attached Storage  
网络附着存储
- 通过IP网络访问存储
- 通过NFS或CIFS，以文件级别访问存储设备
- NAS速度相对DAS来说，比较慢

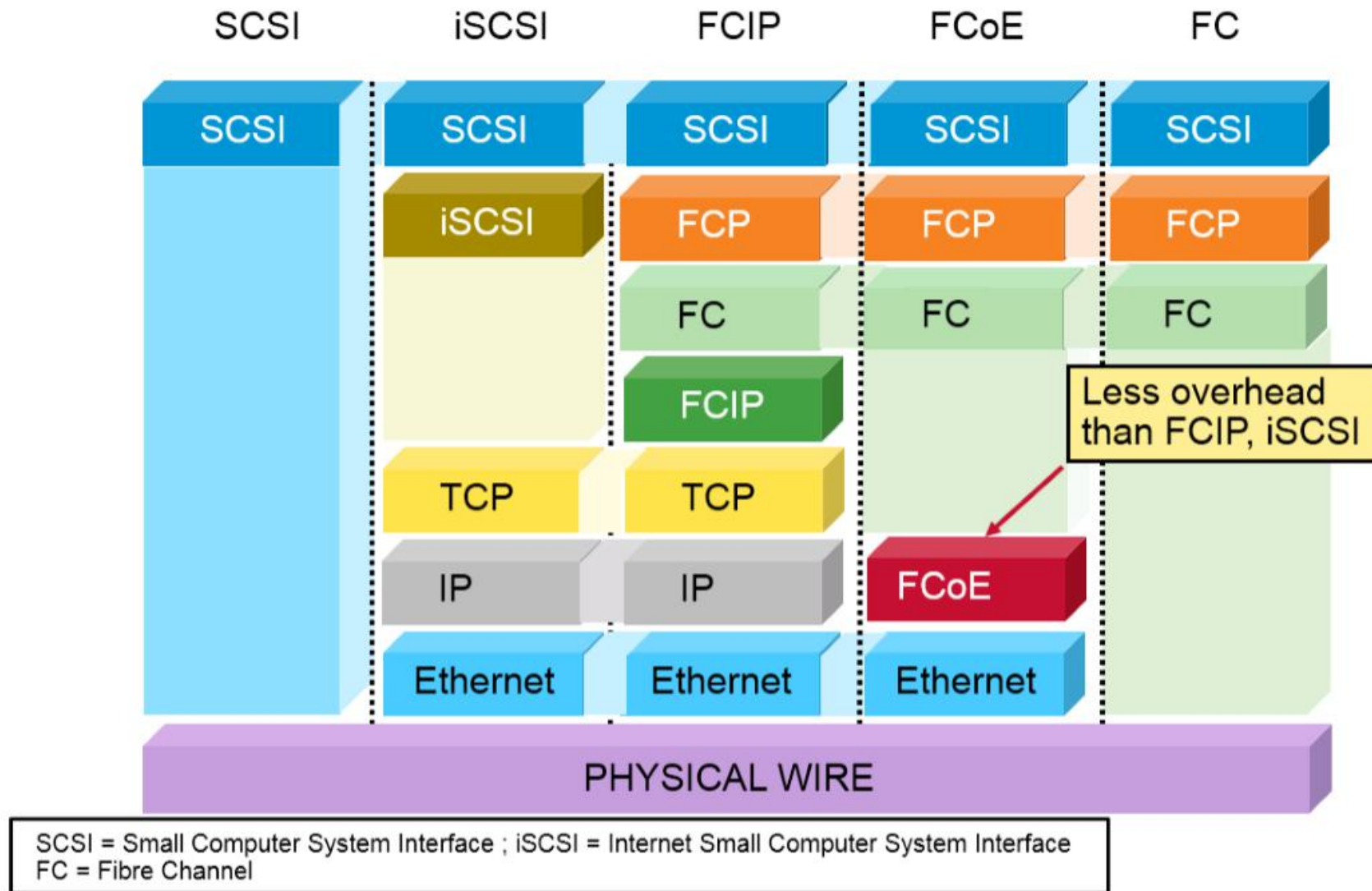




- SAN – Storage Area Network 存储网络
- 高性能互联，可以提供高速I/O传输
- 通过SCSI，以块(block)级别访问存储设备
- 存储可以给服务器共享
- 可能多厂商之间互联的时候会有一些兼容性问题
- 使用Fibre Channel来传输SCSI指令或数据

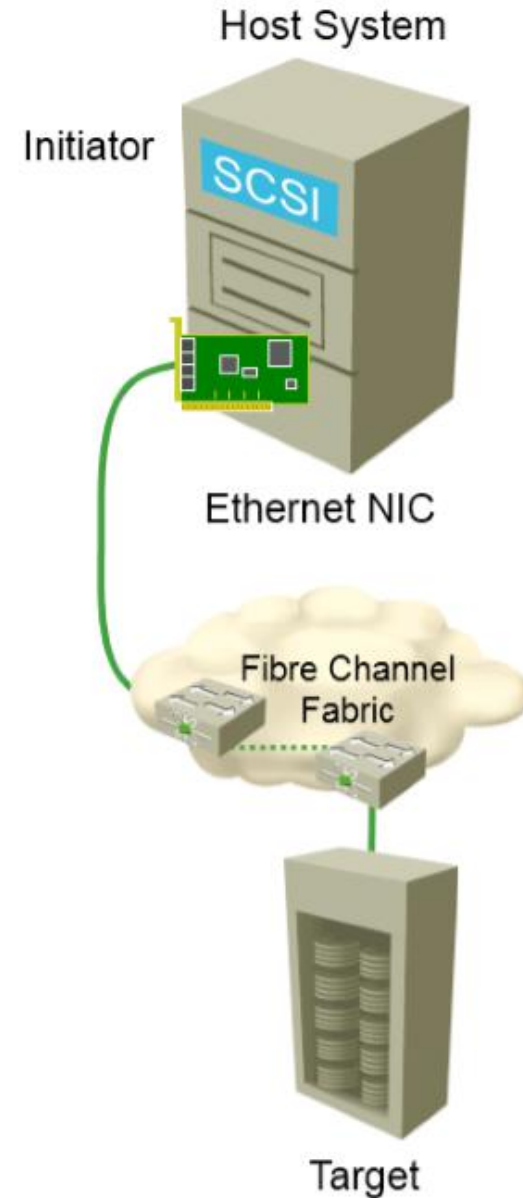


# 存储网络架构比较



## Fibre Channel Overview:

- 扩展和网络化了SCSI
  - 为SCSI Payload提供高速传输
  - 使用串行标准
- Fibre Channel优点:
  - 地址可以支持16M 设备
  - Loop – 仲裁环(共享)、fabric(交换式)
  - 速率可以到达1、2、4、8、16、32、128G
  - 可以延伸到10公里
  - 多协议支持
- 将Channel和network的优点结合起来





## SAN网络和融合网络介绍

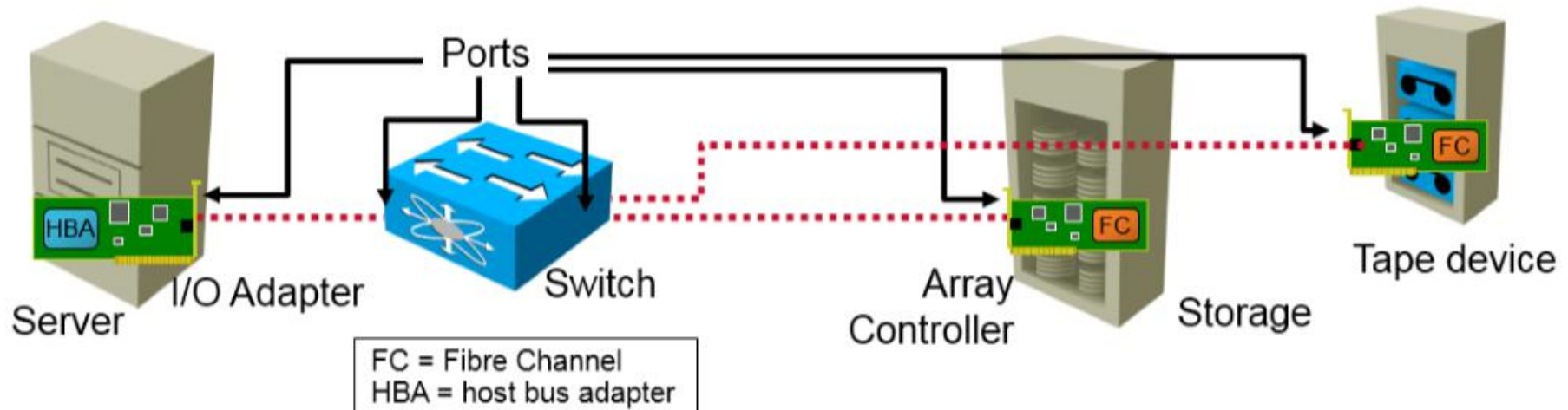
### SAN网络和融合网络介绍

1. 存储简介
- ➔ 2. Fibre Channel 术语
3. Fibre Channel 工作原理
4. 融合网络
5. FCoE的术语
6. FCoE的工作原理
7. FCoE的标准集
8. 传统数据中心网络向融合网络迁移



## Fibre Channel : Fibre Channel

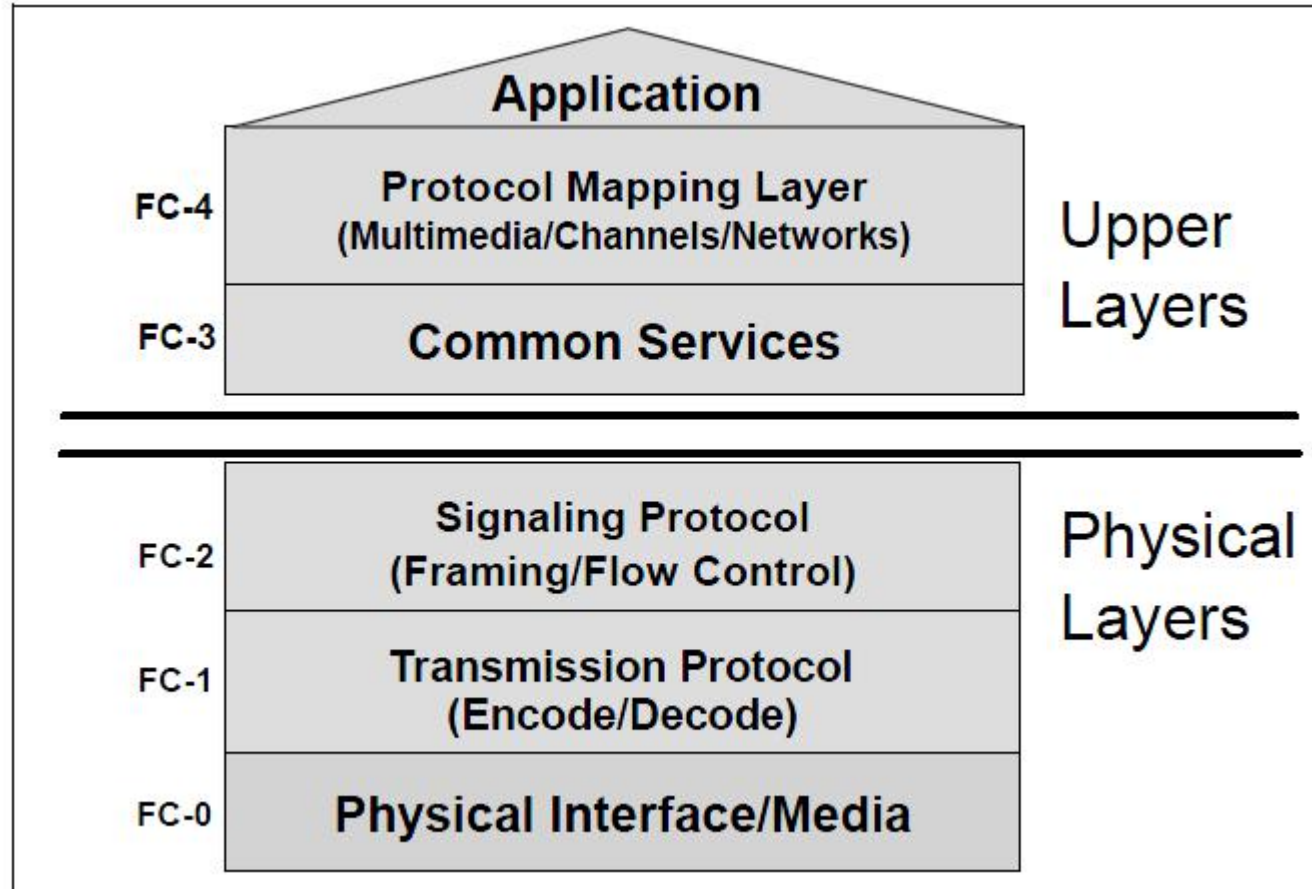
- Fibre Channel: 是用于存储网络的工业标准
- 可以mapped多种协议:
  - Fibre Channel Protocol (FCP)
  - SCSI
  - FICON





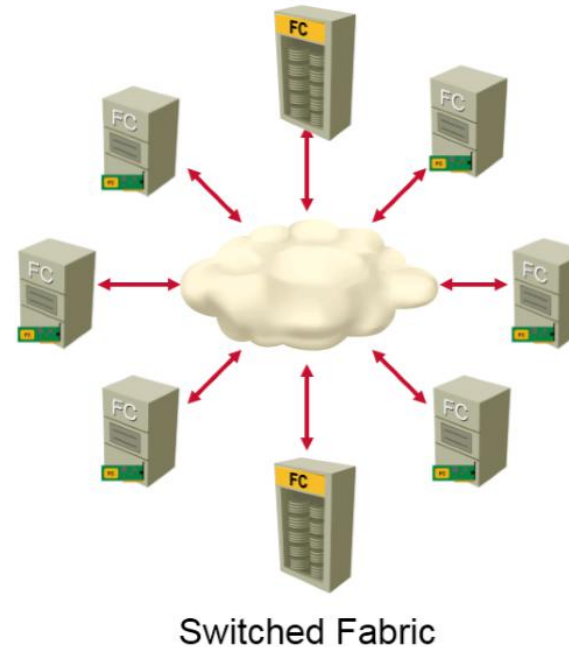
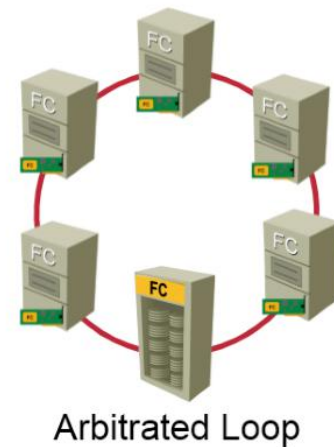
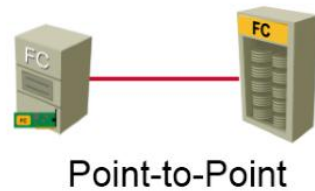


# Fibre Channel: Fibre Channel 分层架构



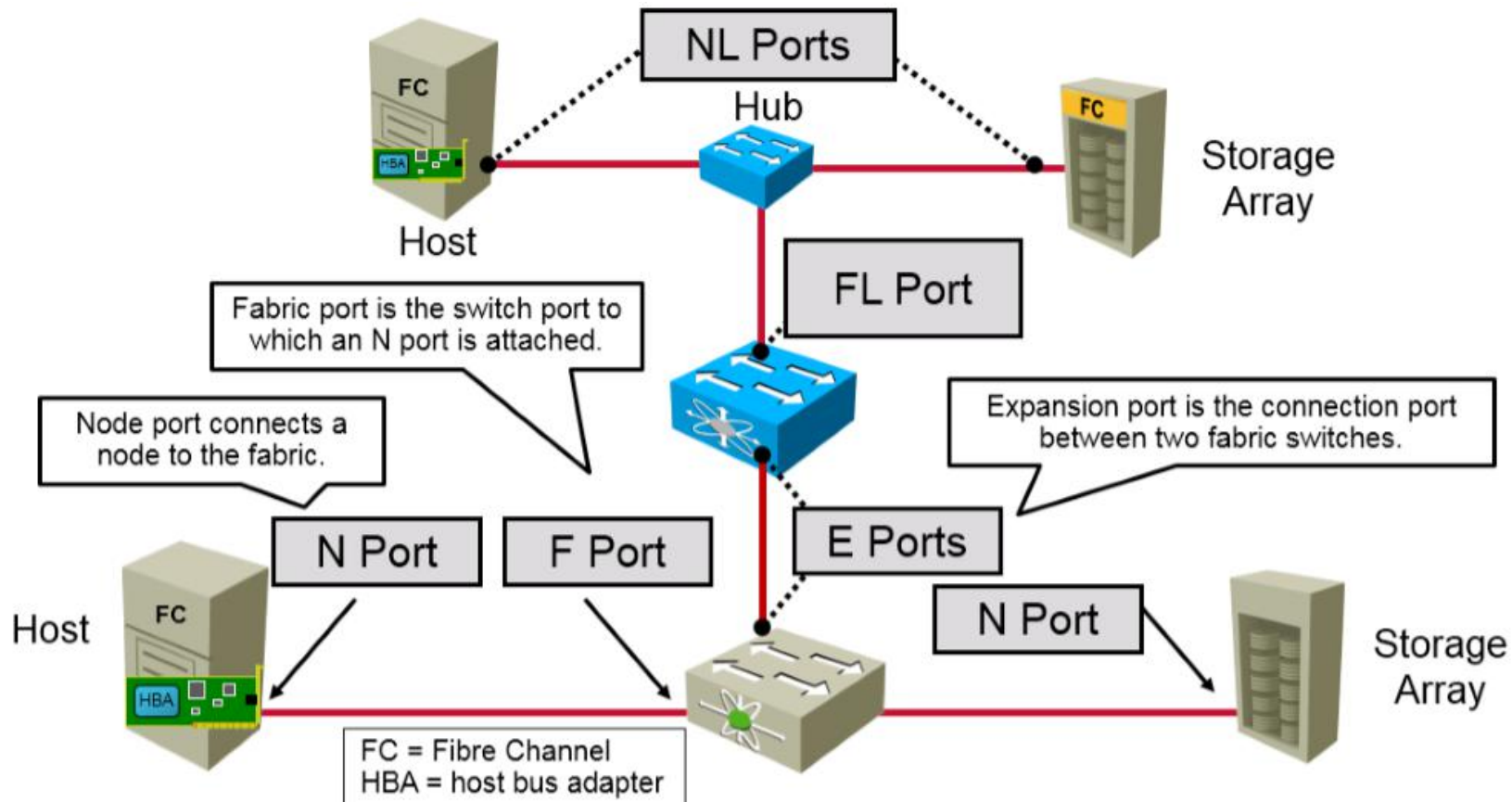
## Fibre Channel : Fibre Channel FC-2

- FC-2定义了frame结构和signaling protocol，可以支持FC帧的可靠传输
- FC-2特指数据传输，与ULP没有关系
- FC-2支持point-to-point, arbitrated loop, switched三种拓扑环境



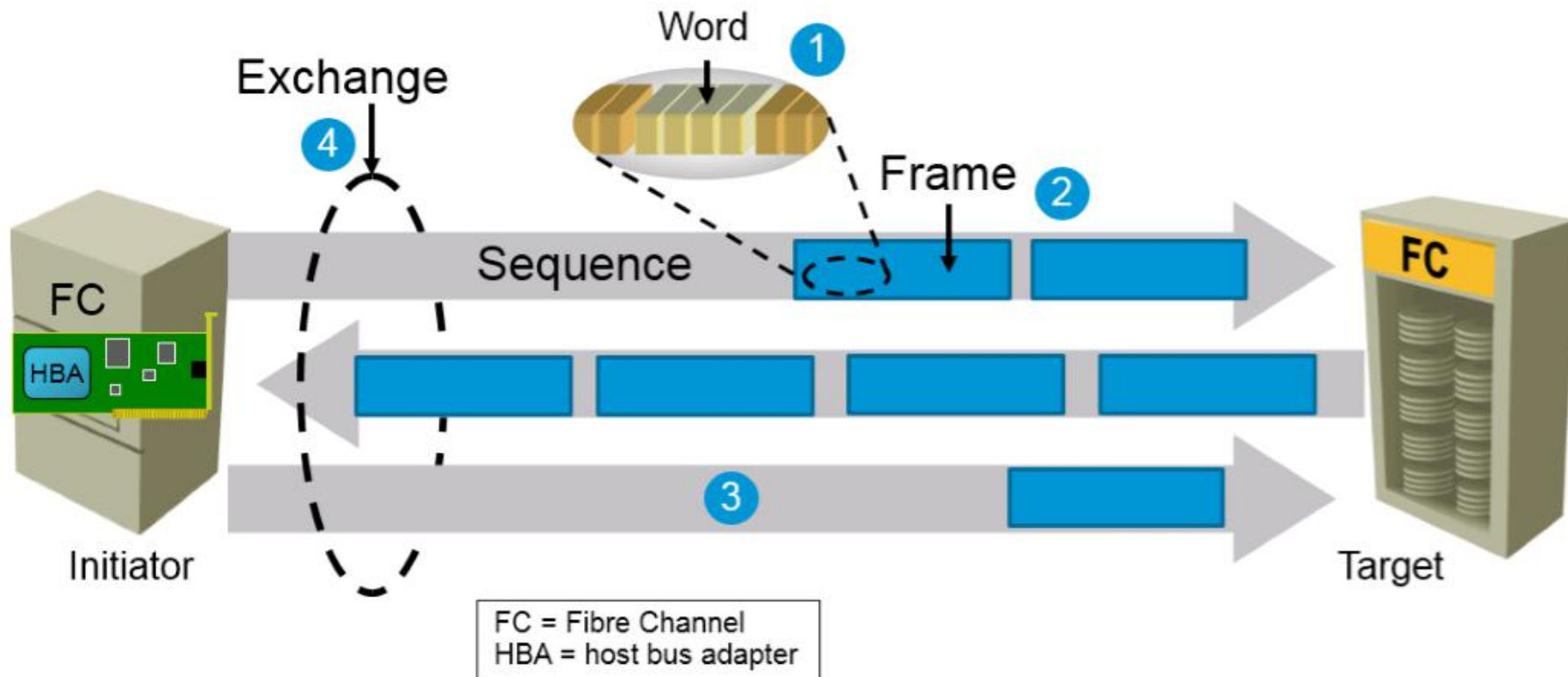
# Fibre Channel : Fibre Channel 的接口类型

- 普通接口: *node (N), fabric (F), expansion (E)*
- 其他接口: *node loop (NL)* 和 *fabric loop (FL)*



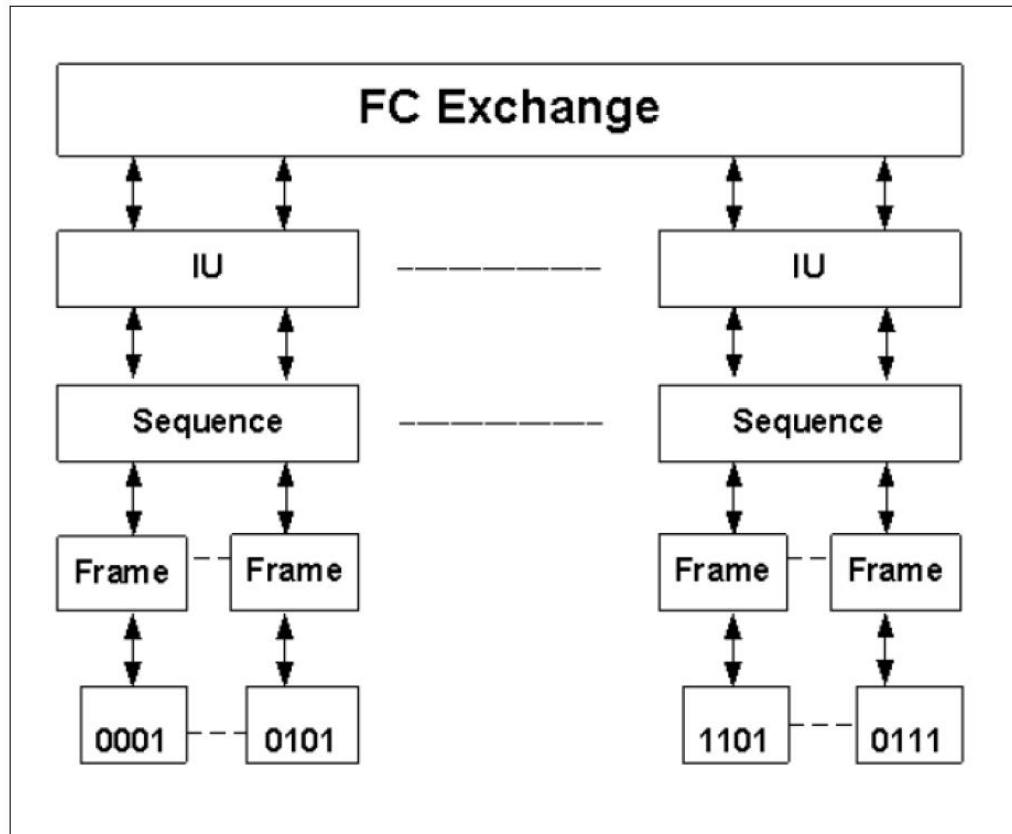
## Fibre Channel : Fibre Channel Frame Structure

- Word: 数据中的最小数据单元，这个单元4个传输字符，将32bits编码进40bits
- Frame: 是一串words的集合，通过SOF和EOF分割，跟IP包差不多
- Sequence: 单向的一系列frame的集合
- Exchange: 在两个Node之间发送的一系列Sequence



# Fibre Channel : Fibre Channel Frame Structure

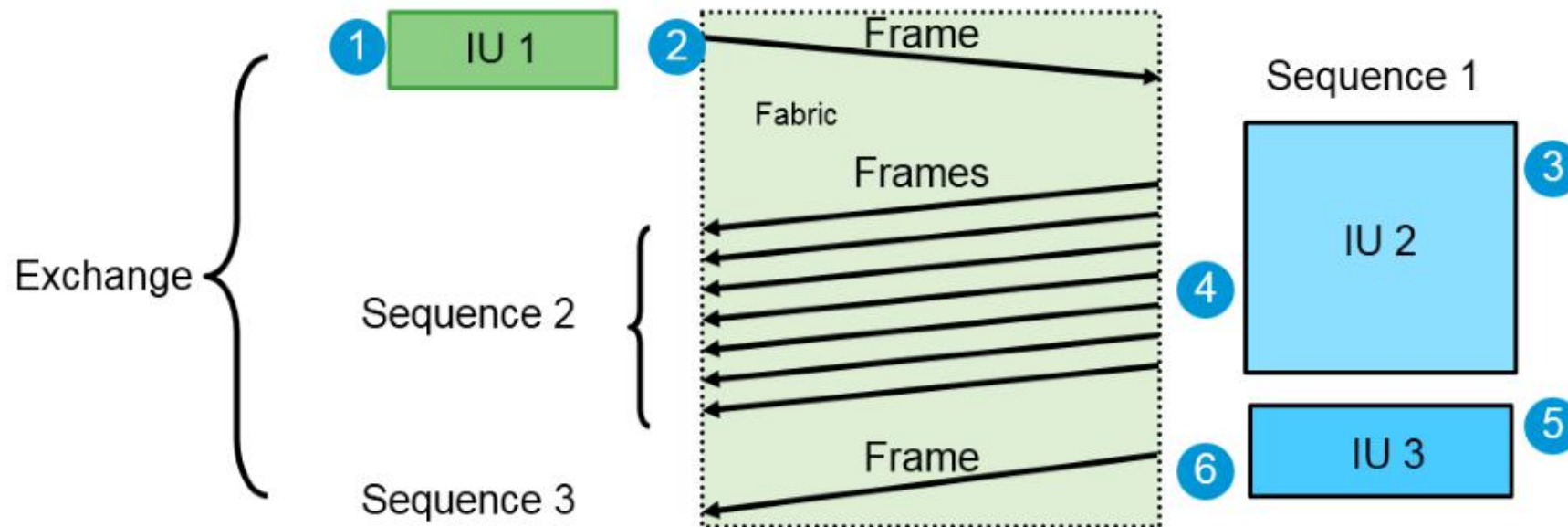
- 一个Exchange负责管理一个操作，可能会分成多个sequences。在initiator和target之间，同一时间可以有多个exchange。
- 以SCSI举例，一个SCSI的任务就是一个exchange，一个SCSI的任务包括多个IU (Information unit)。下面的IU相关SCSI任务。
  - Command IU
  - Transfer ready IU
  - Data IU
  - Response IU





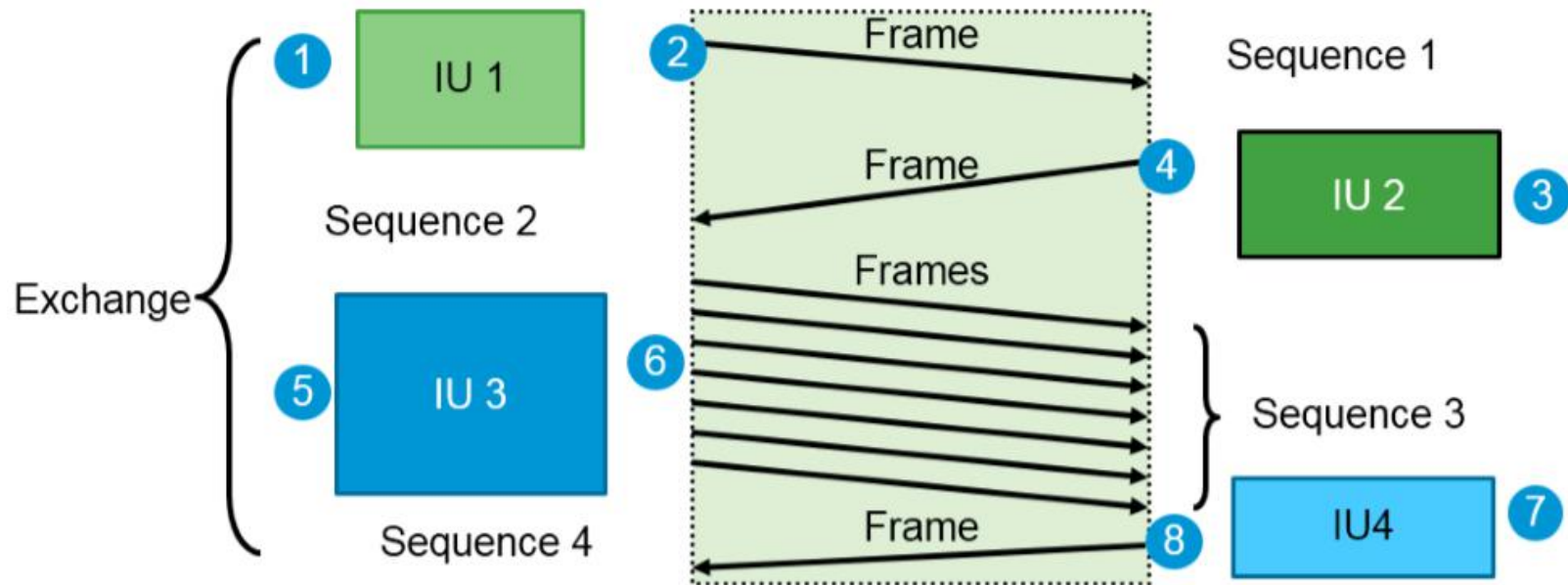
## Fibre Channel : Fibre Channel Read Operation

- 1.IU 1 承载SCSI read request操作 (FCP\_CMD)
- 2.Initiator FC-2层发送IU 1为一个单独的帧
- 3.Target检索被请求的数据(FCP\_DATA)，并且打包成IU2
- 4.Target FC-2层发送IU2为frame sequence
- 5.Target产生状态命令(FCP\_RSP)，并且打包成IU3
- 6.Target发送IU3



# Fibre Channel : Fibre Channel Write Operation

- 1. IU1包括SCSI write request操作 (FCP\_CMD)
- 2. Initiator FC-2层发送IU1为一个单独的帧
- 3. Target回应包括SCSI write request response (FCP\_XFR\_RDY)
- 4. Target FC-2层发送IU2为一个单独的帧
- 5. Initiator从ULP buffer中检索数据(FCP\_DATA)
- 6. Initiator FC-2层将IU3转换成个或多个数据块传输
- 7. Target产生状态命令(FCP\_RSP)确认这次exchange结束
- 8. Target FC-2层转换IU4为一个单独的帧，并发送出去作为sequence 4

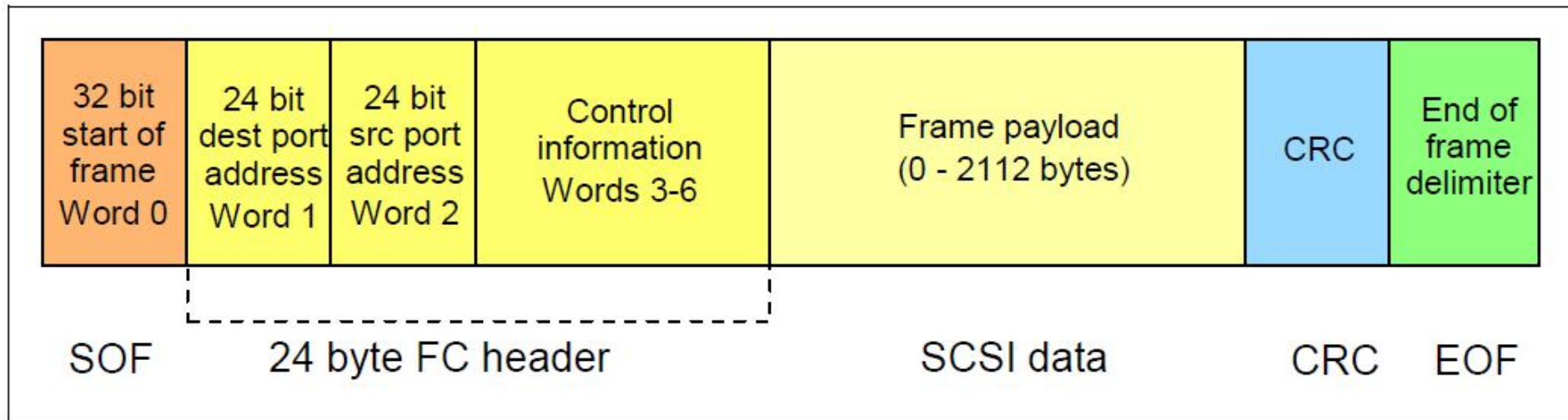




## Fibre Channel : Fibre Channel Frame format

一个Frame包括以下元素:

- *SOF*
- *Frame Header*
- *Optional headers and payload (data field)*
- *CRC*
- *EOF*



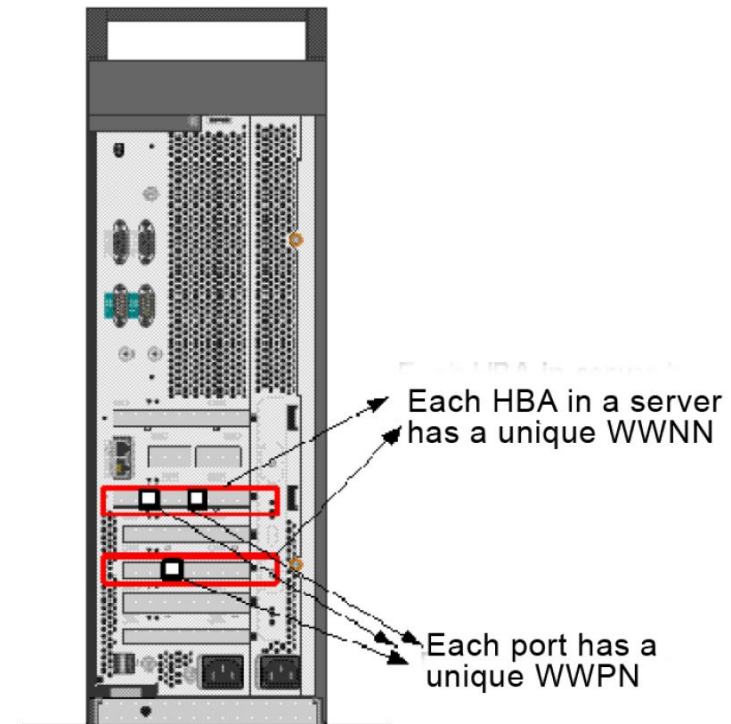
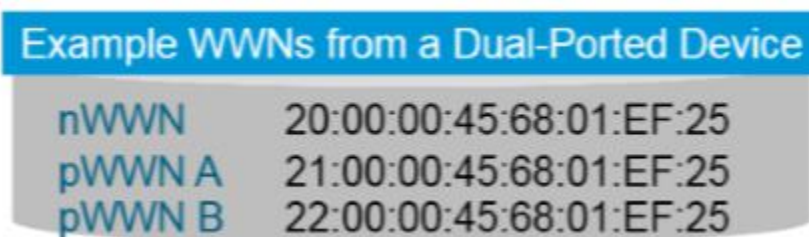
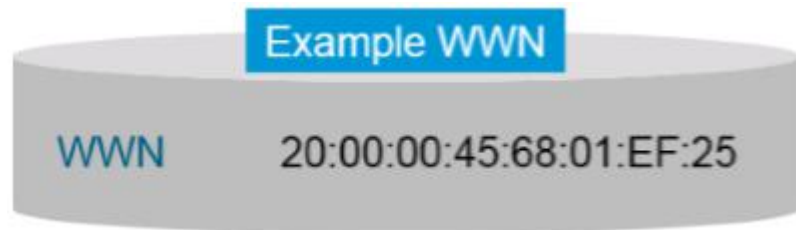


## Fibre Channel : Fibre Channel Frame Header

	Byte 0	Byte 1	Byte 2	Byte 3
Word 0	R_CTL	Destination_ID (D_ID)		
Word 1	Reserved	Source_ID (S_ID)		
Word 2	Type	Frame Control (F_CTL)		
Word 3	SEQ_ID	DF_CTL	Sequence Count (SEQ_CNT)	
Word 4	Originator X_ID (OX_ID)		Responder X_ID (RX_ID)	
Word 5	Parameter			

# Fibre Channel : Fibre Channel Address – WWNs

- 每个Fibre Channel的端口和节点都有一个硬件地址称之为WWN(world wide name)
  - 由IEEE分配给厂家，厂商分配给每个设备
  - 64 or 128bits(现在大多数是128bits)
- Switch的Name Server会将WWNs和FC Address做一个映射
  - nWWNs用于标识设备
  - pWWNs用于标识设备的每个端口

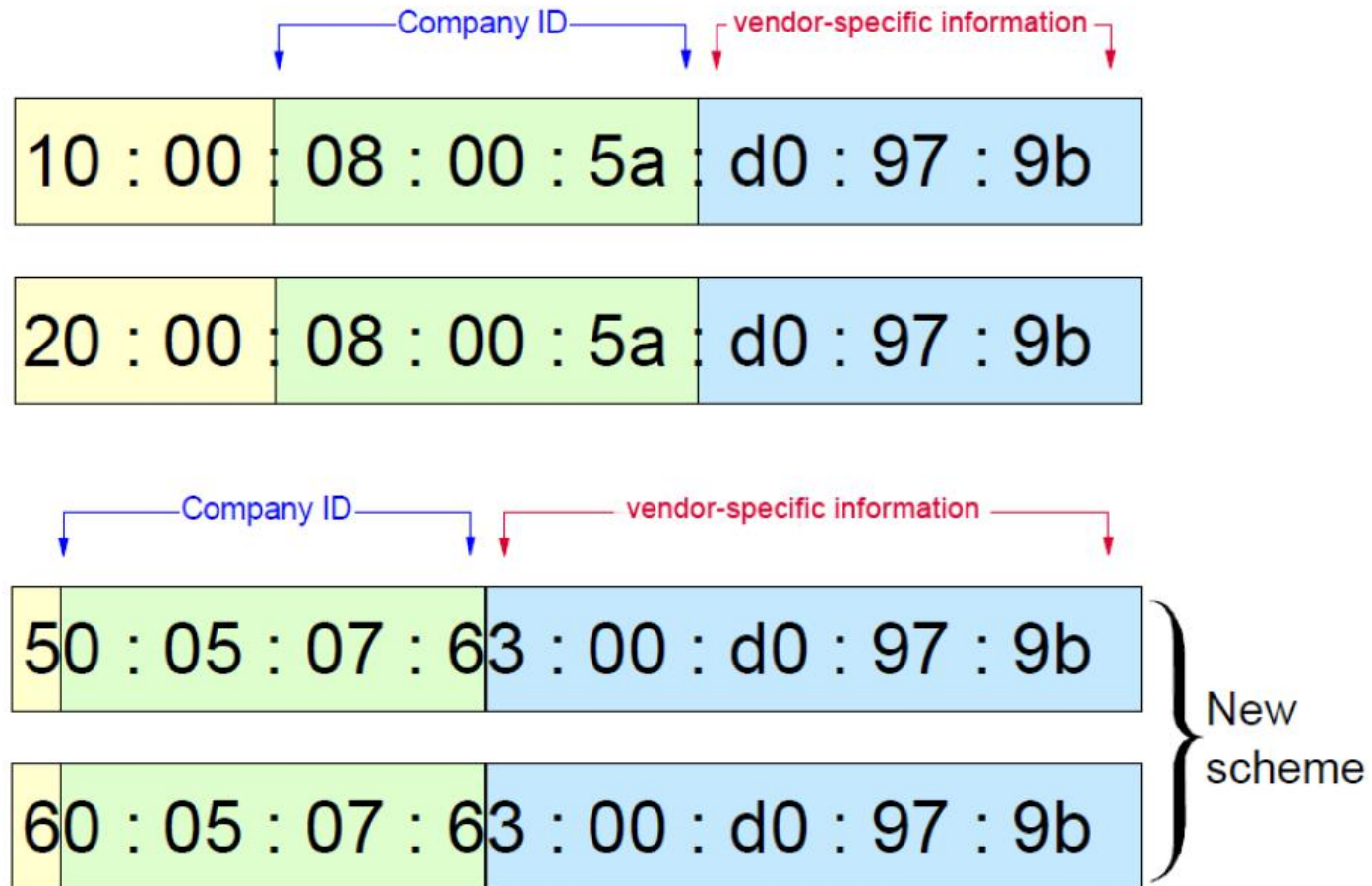






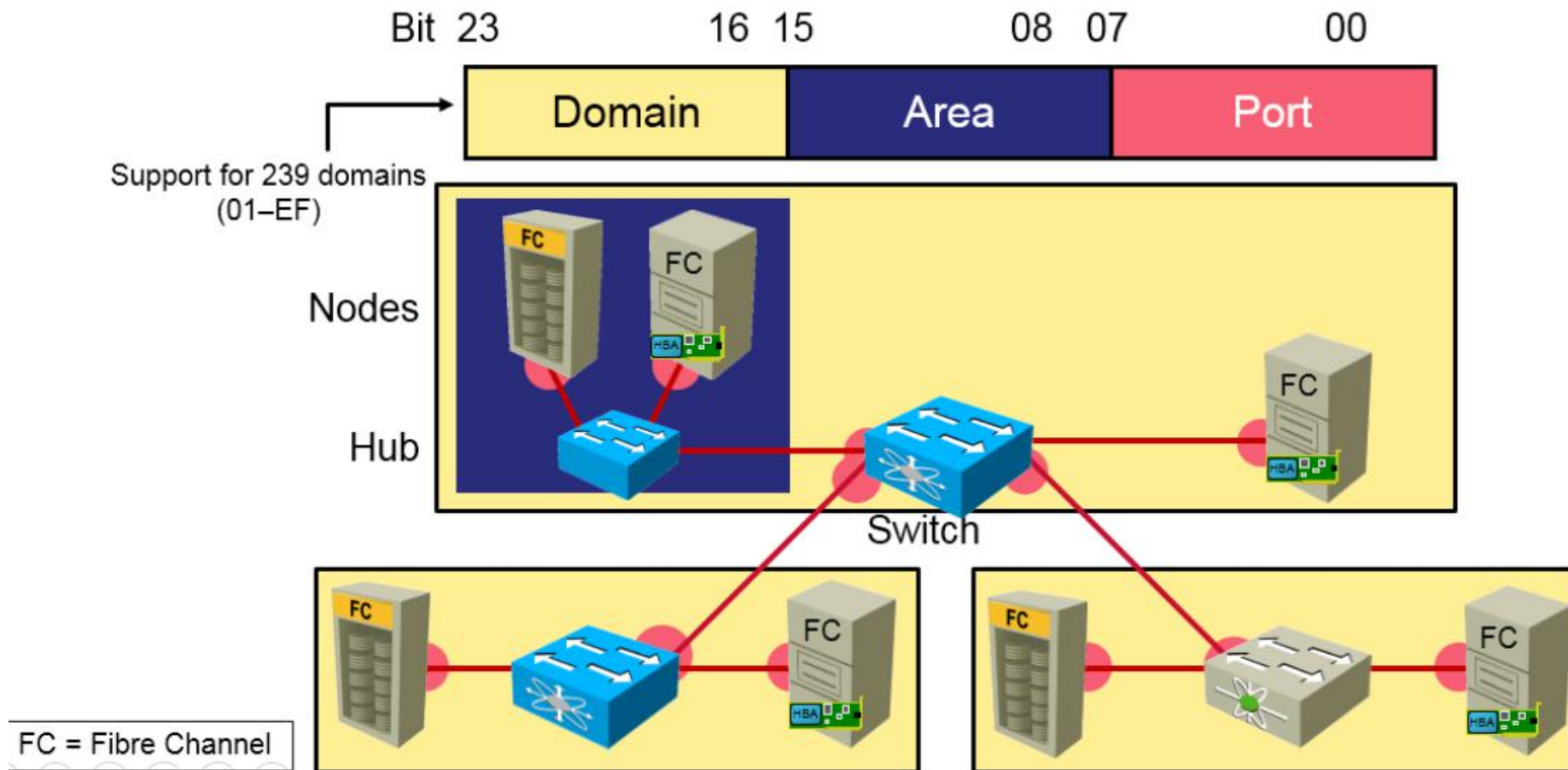
# Fibre Channel : Fibre Channel Address – WWNs

## WWN Addressing scheme



# Fibre Channel : Fibre Channel Address Format

- Domain ID用于定义一个交换机，每个Fabric中的交换机有一个唯一的Domain ID。
- Area ID用于标识一个端口组。
- Port ID用于标识设备上连接的单独的设备。





## SAN网络和融合网络介绍

### SAN网络和融合网络介绍

1. 存储简介
2. Fibre Channel 术语
- ➔ 3. Fibre Channel 工作原理
4. 融合网络
5. FCoE的术语
6. FCoE的工作原理
7. FCoE的标准集
8. 传统数据中心网络向融合网络迁移



# Fabric

## 网络初始化



- 周涛
- QQ: 53408031 IE-LAB公开课群:79791756
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站:[www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024



## Fibre Channel : Fabric Service

Fabric Service是一系列的服务集。有很多种下面列出来的是一些主要的Fabric Service，它们使用保留的Fibre Channel Address:

- Management Services (0xFFFFFA)
- Time Services (0xFFFFFB)
- Simple name server (0xFFFFFC)
- Login Services (0xFFFFFE)
- Registered State Change Notification (RSCN) (0xFFFFFD)

以上只是一部分Fabric Services，还有很多。



# Fibre Channel : Fabric 配置过程

在SAN Fabric可以使用FC设备前，网络必须初始化并且处于正常操作状态。以下是基本Fabric初始化时的步骤：

## ■1.连接初始化

- 交换端口建立不同端口之间交换连接参数（达到帧同步状态）

## ■2.侦测端口操作模式

- 交换机开始试图发现端口是FL\_Port, E\_Port还是F\_Port。在程序的最后才知道交换机端口类型，如果被确定为E\_Port, 连接参数和缓冲信用将初始成为商定设置。

## ■3.选择主交换机（Principal Switch）

- 如果交换机发现自己处于一个多交换机环境，将会引发一个选举进程，为Fabric选出一个主交换机控制Domain ID的分配。

## ■4.分配Domain ID

- 新的主交换机成为Domain ID地址管理者，它将发出一个AAI（地址标识通告；Announce Address Identifier），用于通知Fabric内的所有交换机新的主交换机已经选出，现在可以申请Domain ID了，所有的交换机将向Domain ID地址管理者申请一个。最终，所有的交换机都将拥有一个Domain ID。

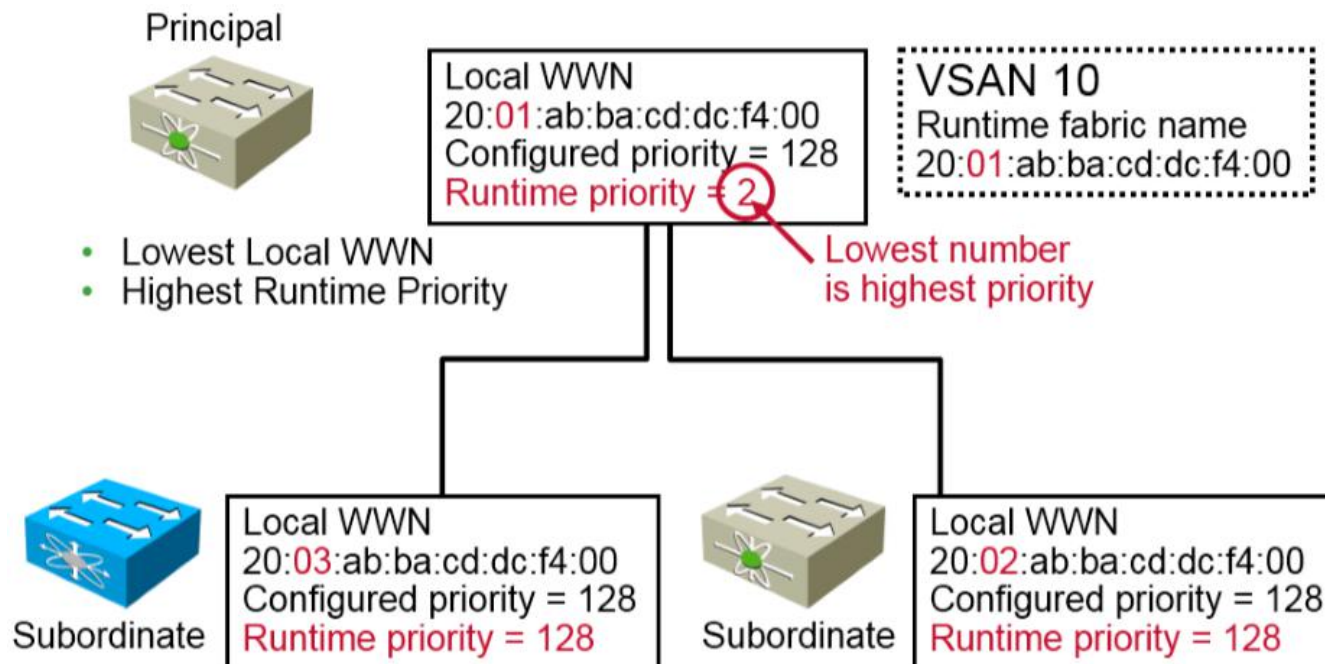
## ■5.路径选择

- Fabric内的交换机将利用FSPF算法找出所有路径信息。交换机将通过这一进程知道到达目标的最佳路径。



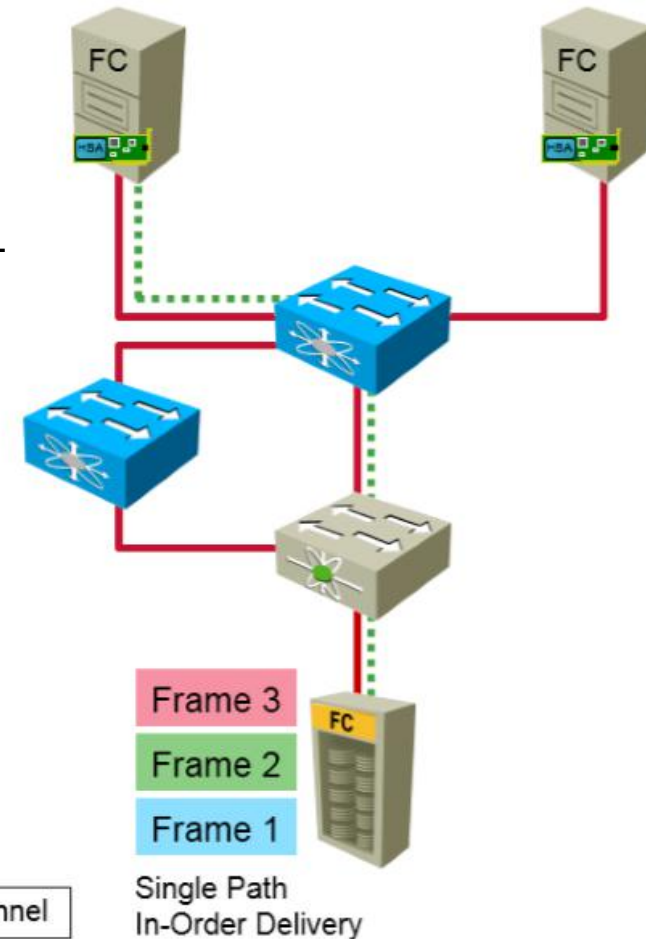
# Fibre Channel : Fabric 配置过程 – 选择Principal Fabric Switch

- Fabric Channel domain (fcdomain) feature执行principal switch的选择, domain ID的分发, FCID的分配, 和fabric重新配置
- 在Fabric中只有一个Principal Switch, 在Fabric初始化过程中, 主交换机为每个交换机分配8bits的Domain ID。
- 如果主交换机因为任何原因发生故障, Fabric将产生一个新的选举进程推选一个新的主交换机。在Fabric中WWN最小的交换机将成为主交换机。
- 对于已经拥有主交换机的Fabric, 新加入的并拥有最小WWN的交换机将不会生成选举进程。



# Fibre Channel : Fabric 配置过程 - 路径选择 - FSPF

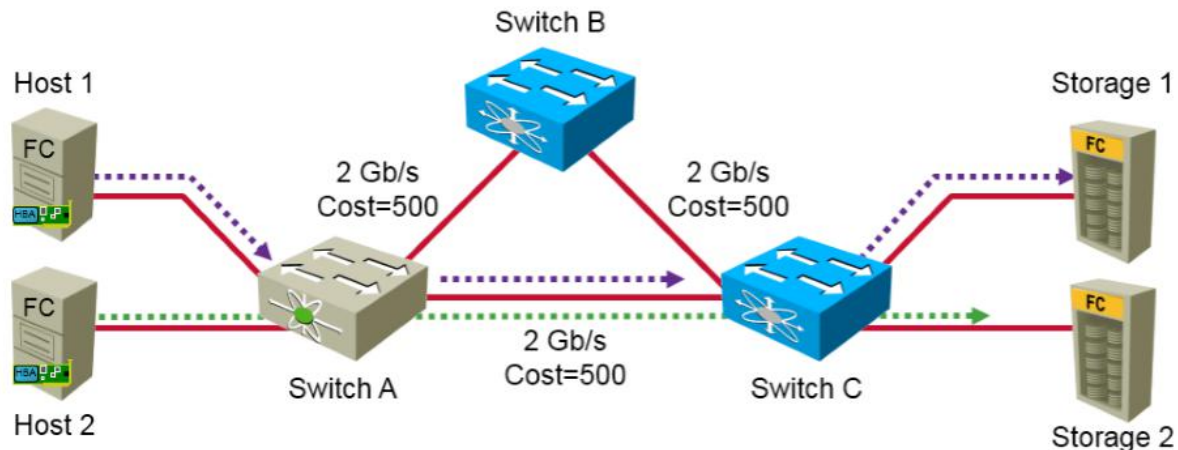
- FSPF将交换机结合在一起并建立一个受到保护的网络的的基础
- 自动Fabric拓扑分析和路径选择
- 在SAN配置变化时动态配置路由
- 可以设定静态路由
- 发送故障自动选择下一条最优路径
- 与OSPF一样，FSPF算法基于接口cost值。而接口cost与ISL的speed有关。
- 这样使Frame沿着一个固定的路径传输，换句话说，相同initiator和target之间的数据包走相同路径，是有序的。



# Fibre Channel : Fabric 配置过程 – 路径选择 – FSPF

FSPF由四个主要部分组成:

- Hello协议用于建立相邻交换机之间的信任关系 (双向, 全双工通讯)
  - (第一部分结束后, 交换机将知道哪条ISL正在工作并且可以用于双向帧传输)
- 复制拓扑数据库初始化, 维护和同步
  - 当交换机启动或出现一条新的ISL将会导致初始数据库同步
  - 连接状态变更(Link State Updates; LSU)和连接状态确认(Link State Acknowledgements; LSA)用于维持复制数据库的完整性
  - 变更机制可以由连接状态变更引发, 也会基于时间或周期性进行.
- 路径选择使用基于权值的路径算法
  - 连接权值根据数据传输速率设定, 使用专用公式:  $\text{Link Cost} = A * (1.0625E12 / \text{Baud Rate})$  进行计算, A的缺省值为1, 可以根据对某一连接的好恶进行更改. 例如, 1Gbps连接, 权值为  $1 \times (1.0625E12 / 1.0625E9) = 1,000$ .
- 路由表是为一条通道上从跳跃到跳跃的路由选择而建立和更改
  - 连接状态记录包括所有相关Domain的ISL连接状态信息.



## FC-2 端口注册过程



- 周涛
- QQ: 53408031 IE-LAB公开课群:79791756
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站:[www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024



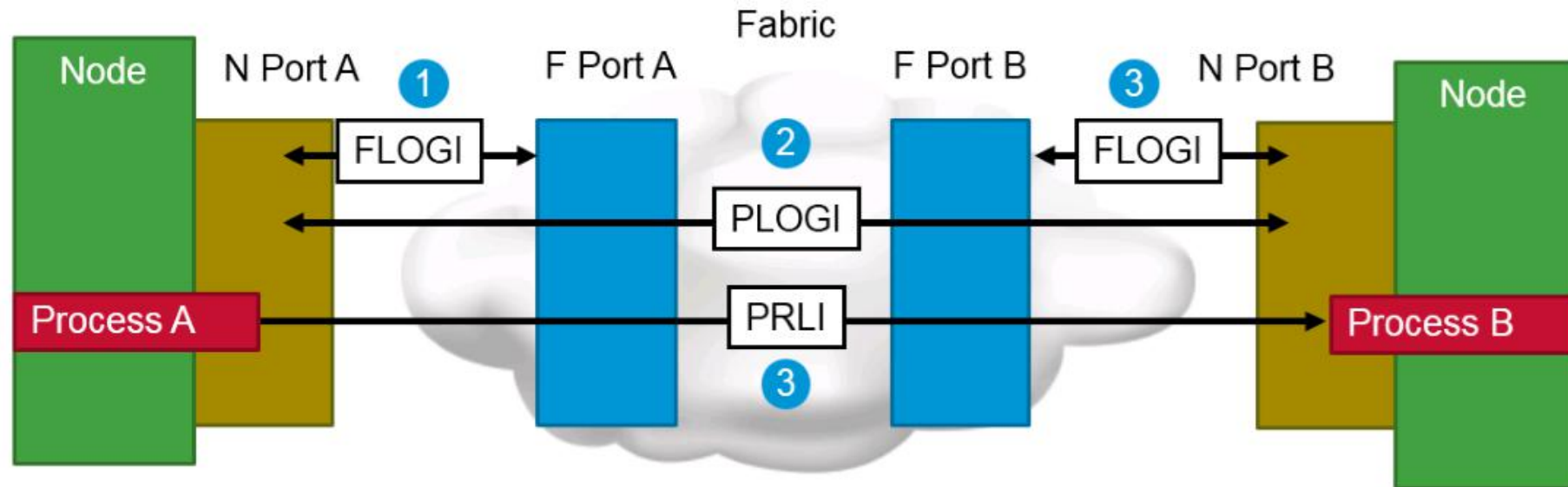
## Fibre Channel : FC-2端口注册过程 – Login Server & Name Server

- Login Server (0xFFFFFE)，用于执行Fabric login
- Name Server (0xFFFFFC)，是FC-GS规范中的组成部分
- 负责关于Fabric连接设备的目录信息
- 设备连接到网络必须在Name Server注册
- Name Server使用FC-CT（普通传输）协议在整个Fabric内部分发信息（动态扩展能力）
- Fibre Channel设备可以向Name Server查询网络资源信息（简单请求和响应模式）
- Name Server特性：
  - 没有单点故障
  - 每个分布式名称服务器只维护“自己的”本地信息，同时从其他分布式名称服务器上通过服务器到服务器协议（基于FC-CT）获取异地信息
  - 服务器到服务器通讯传输到外部名称服务器客户端（initiator或者target）
  - 每个分布式名称服务器可以缓冲存储异地信息并保持一定时间
  - 快速、有效的设备发现能力
  - 最新设备信息，通过注册，名称服务器可以立即提供设备的主要信息。当任意交换机或设备停止工作后（注销）所有相关名称服务器信息都将立即删除。

## Fibre Channel : FC-2端口注册过程

Initiator和Target之间互相通信之前，首先要经历三个过程：

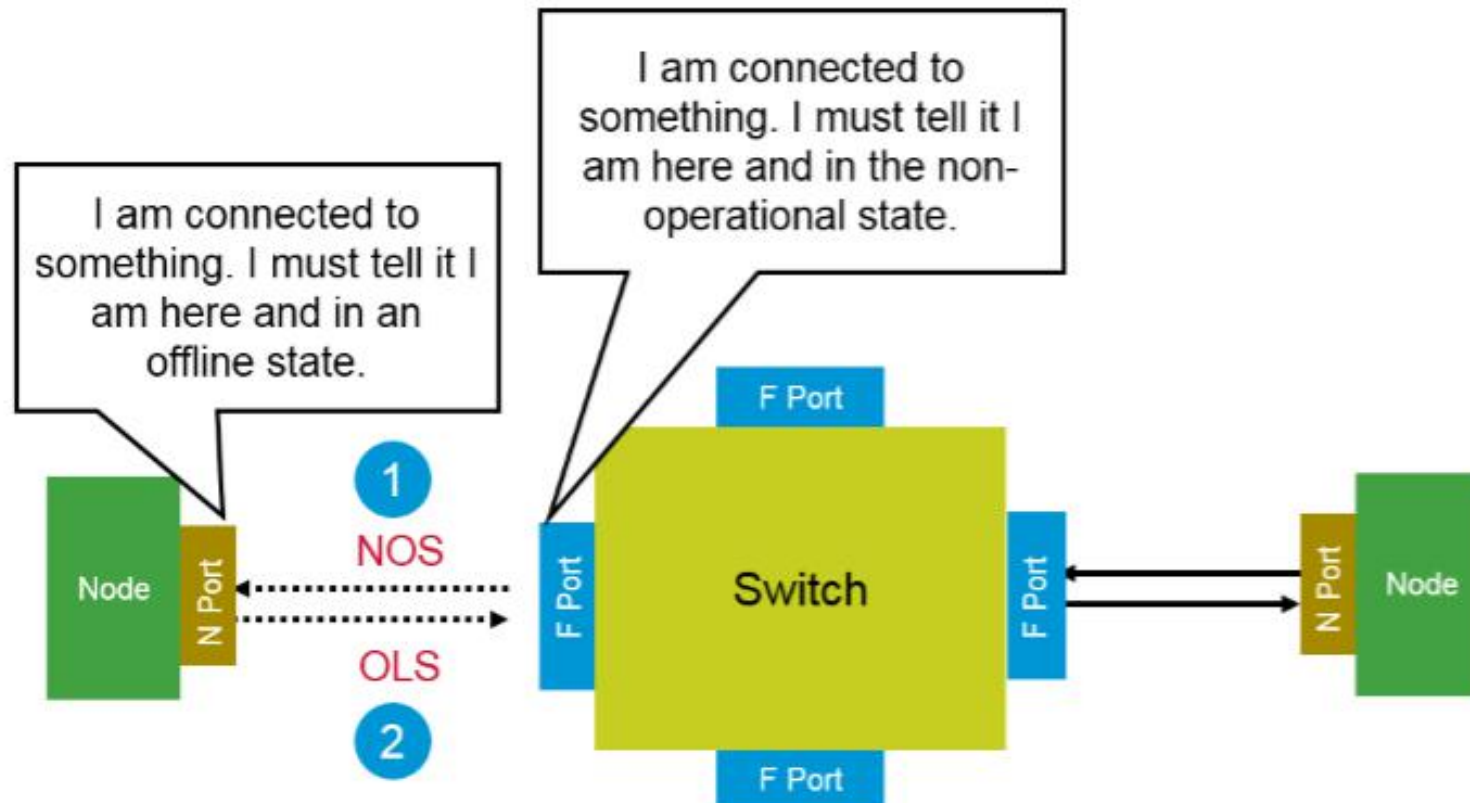
- 1. N\_Port必须登陆连接到相应的F\_Port上 – fabric login (FLOGI)
- 2. N\_Port必须登陆连接到目的N\_Port上 – port login (PLOGI)
- 3. N\_Port必须与target N\_Port之间交换ULP的支持信息
  - 确保initiator和target的process可以通信
  - 这个过程叫process login (PRLI)





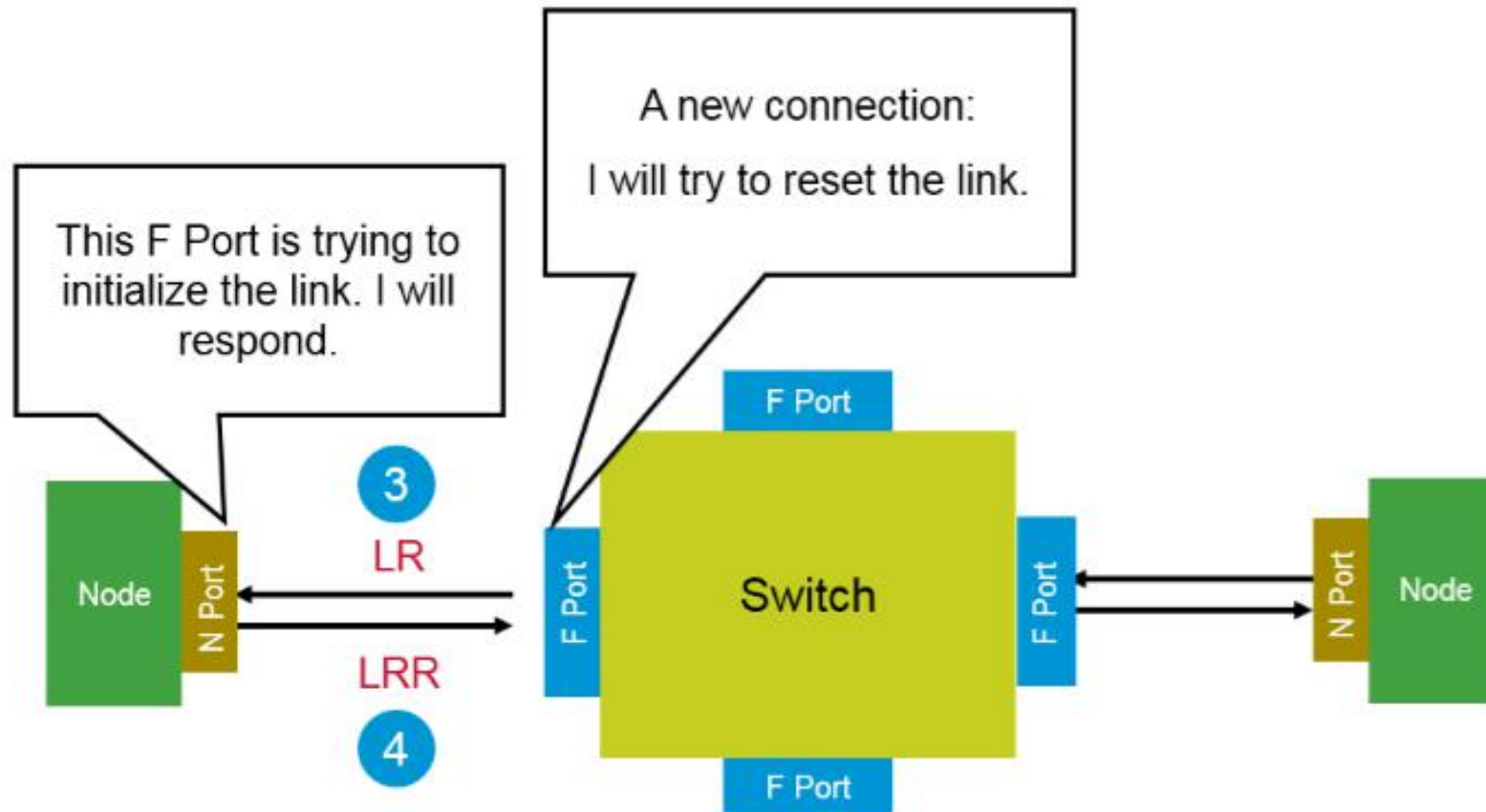
## Fibre Channel : FC-2端口注册过程 - FLOGI

- 1. F\_Port 向N\_Port发送一个NOS (not-operational) sequence
- 2. 当N\_Port收到NOS sequence, 会相应一个OLS (offline state) sequence, 进行链路初始化



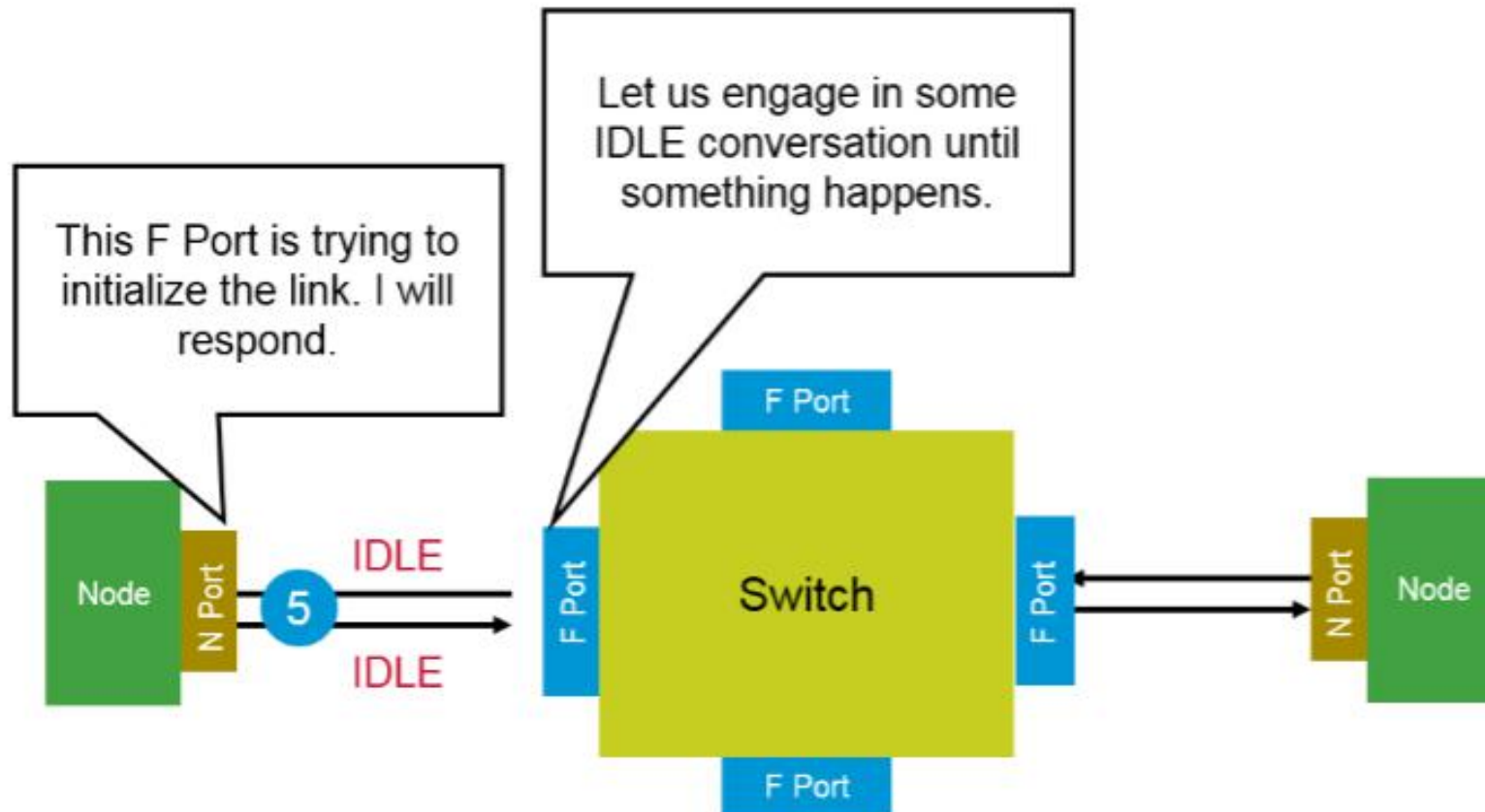
## Fibre Channel : FC-2端口注册过程 - FLOGI (con.)

- 3. F\_Port 通过发送一个LR (link reset) 来reset接口
- 4. N\_Port响应一个LRR (link reset response)



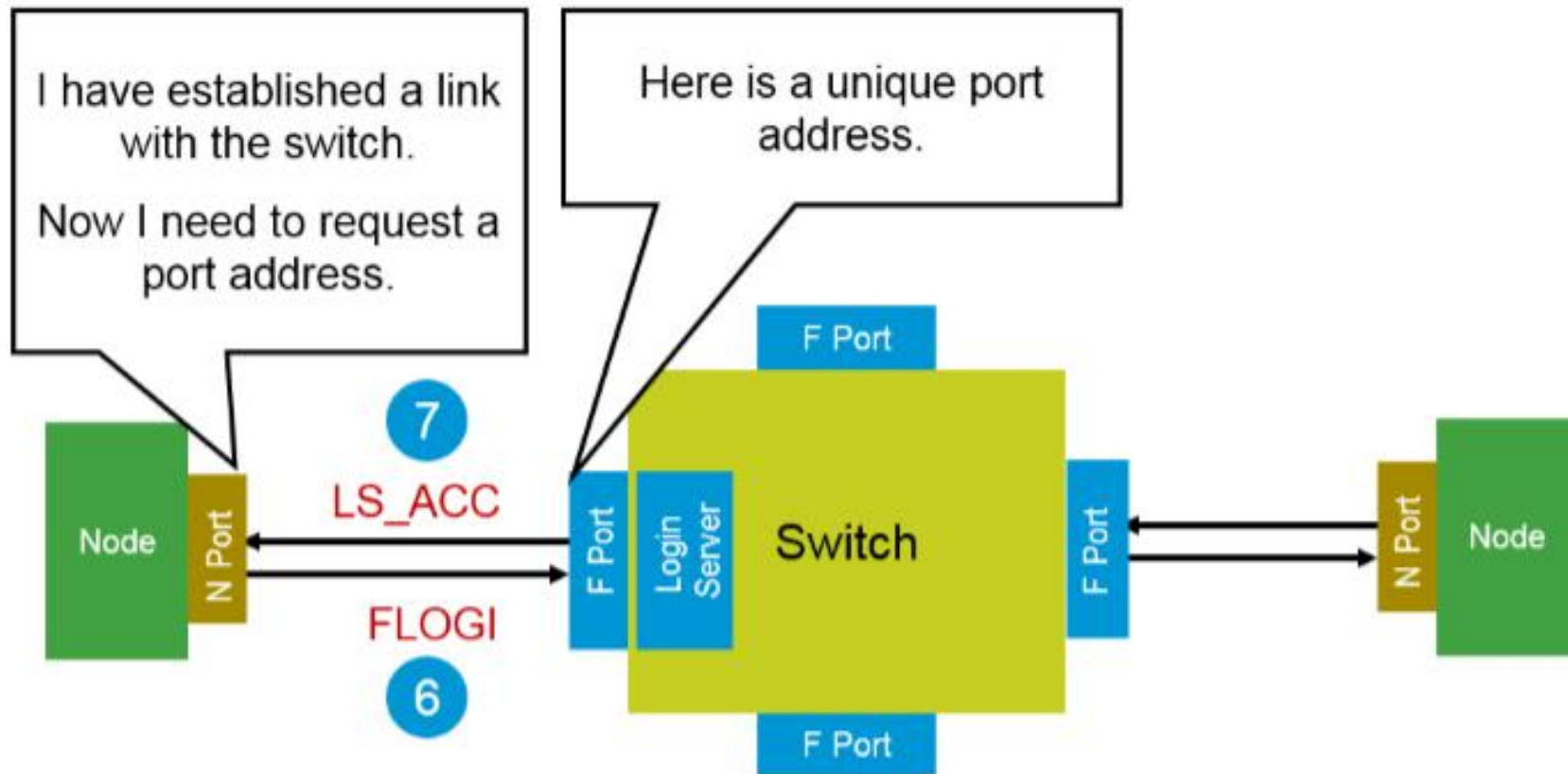
## Fibre Channel : FC-2端口注册过程 - FLOGI (con.)

- 5. link已经active, 然后IDLE会填充在words里
- 这时N\_Port的Source ID为0x000000



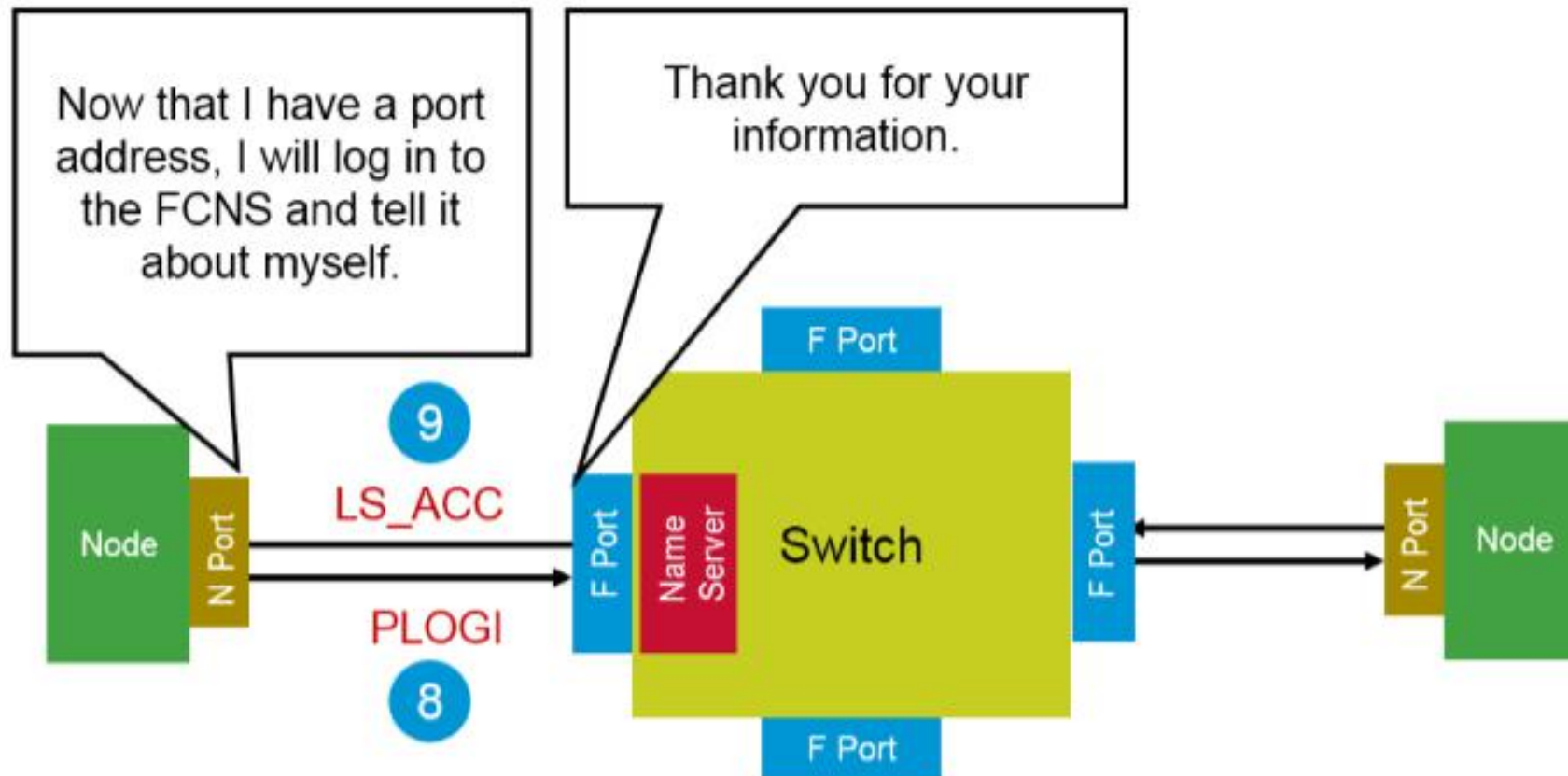
## Fibre Channel : FC-2端口注册过程 – FLOGI (con.)

- 6. 在N\_Port与F\_Port建立连接后，N\_Port希望获得一个port address。N\_Port发送FLOGI link service发往Login Server (0xFFFFF $\bar{E}$ )
- 7. Login Server发送一个LS\_ACC(link state accept)，这里面包括N\_Port的FCID



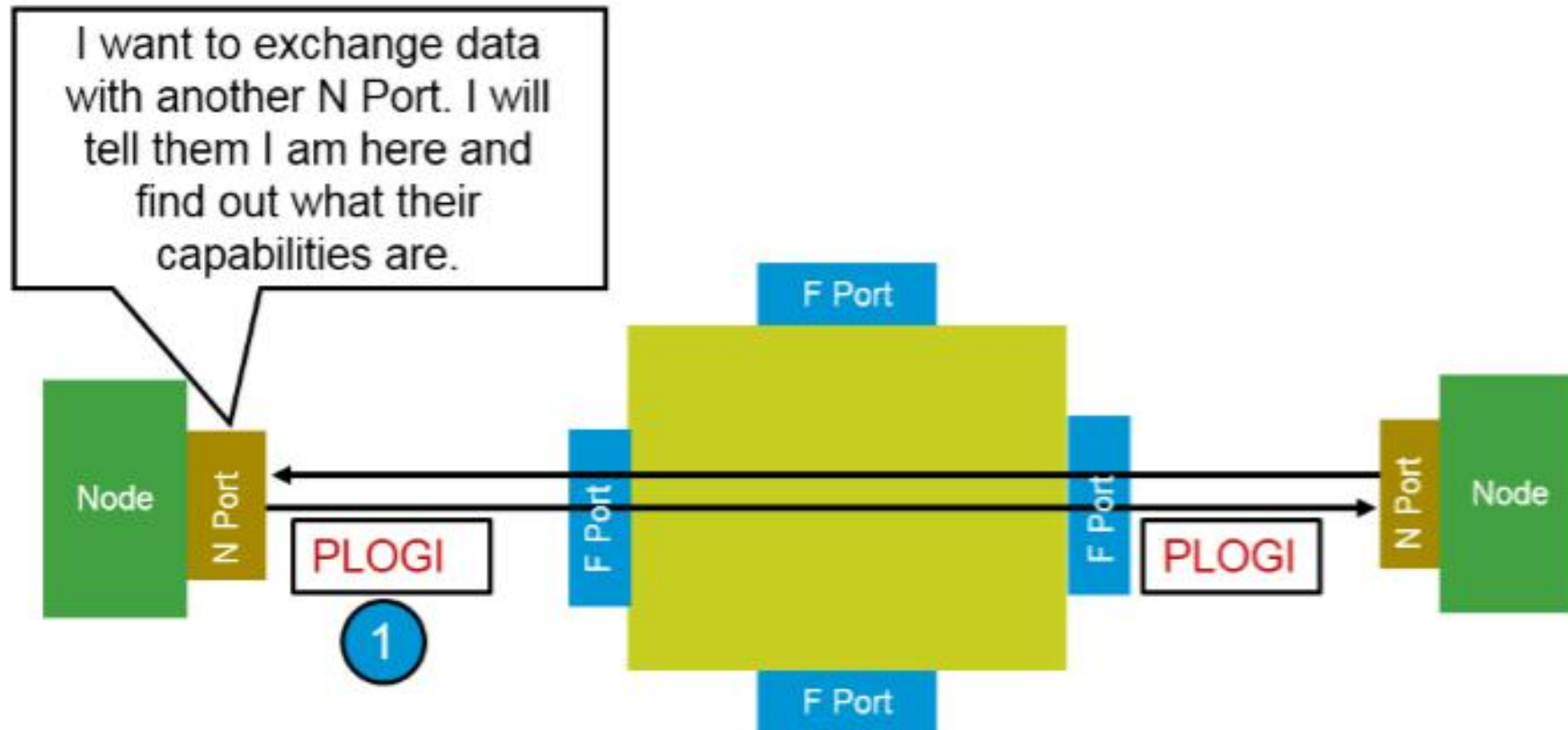
## Fibre Channel : FC-2端口注册过程 - PLOGI

- 8. 在收到FCID后，N\_Port访问Name Server（0xFFFFFC），N\_Port发送它的服务参数，比如buffer credit的number，最大的payload大小和支持的CoS等
- 9. Name Server发送一个LS\_ACC(link state accept)



## Fibre Channel : FC-2端口注册过程 - PLOGI (con.)

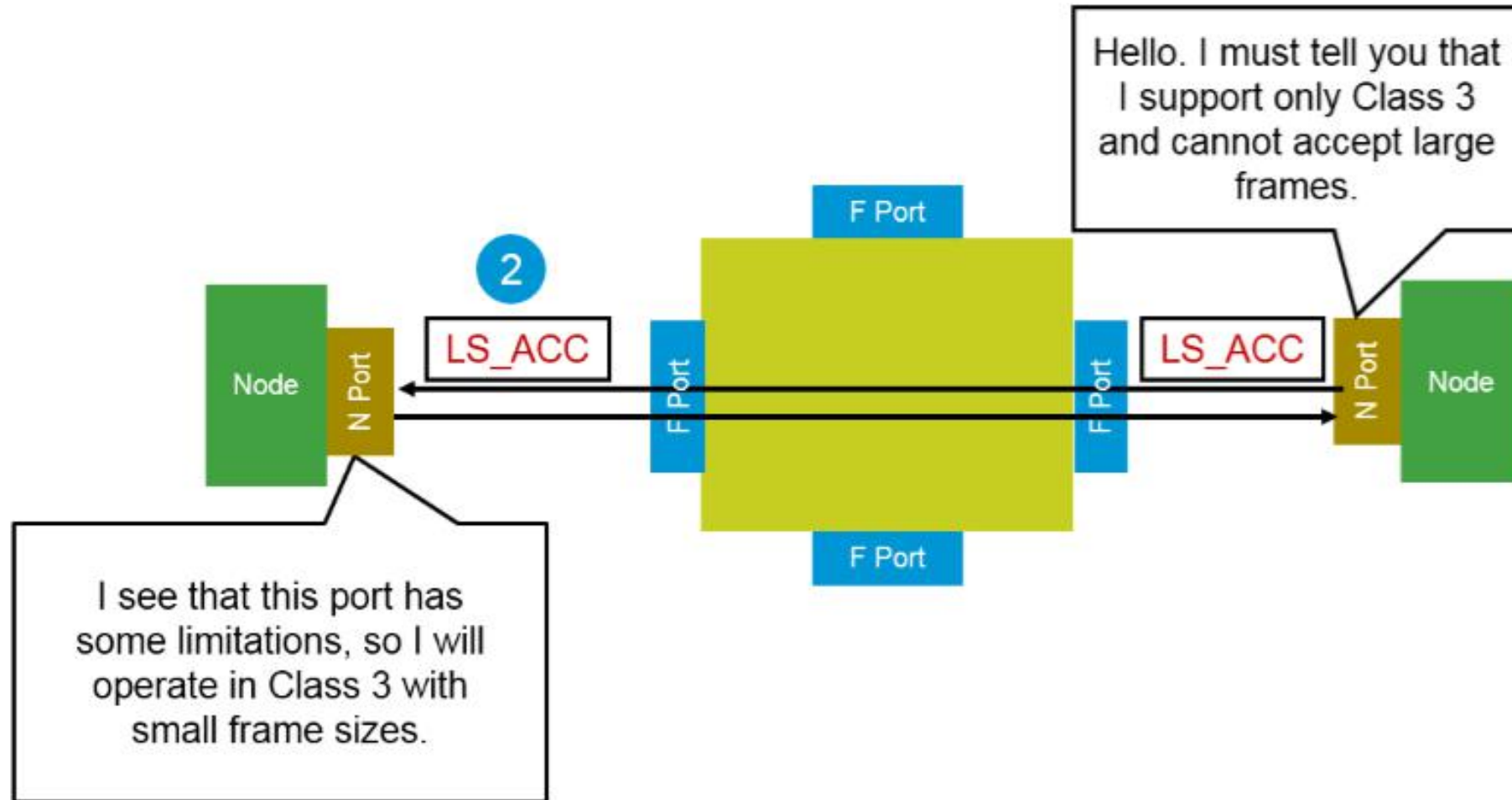
- PLOGI主要用于在initiator和target之间创建一个通讯的通道，并且设置和交换operational参数
- 在FLOGI之后，N\_Port通过PLOGI访问其他的N\_Port。在通信之前必须完成PLOGI。
- 1. Initiator N\_Port发送一个PLOGI帧并且封装N\_Port的操作信息





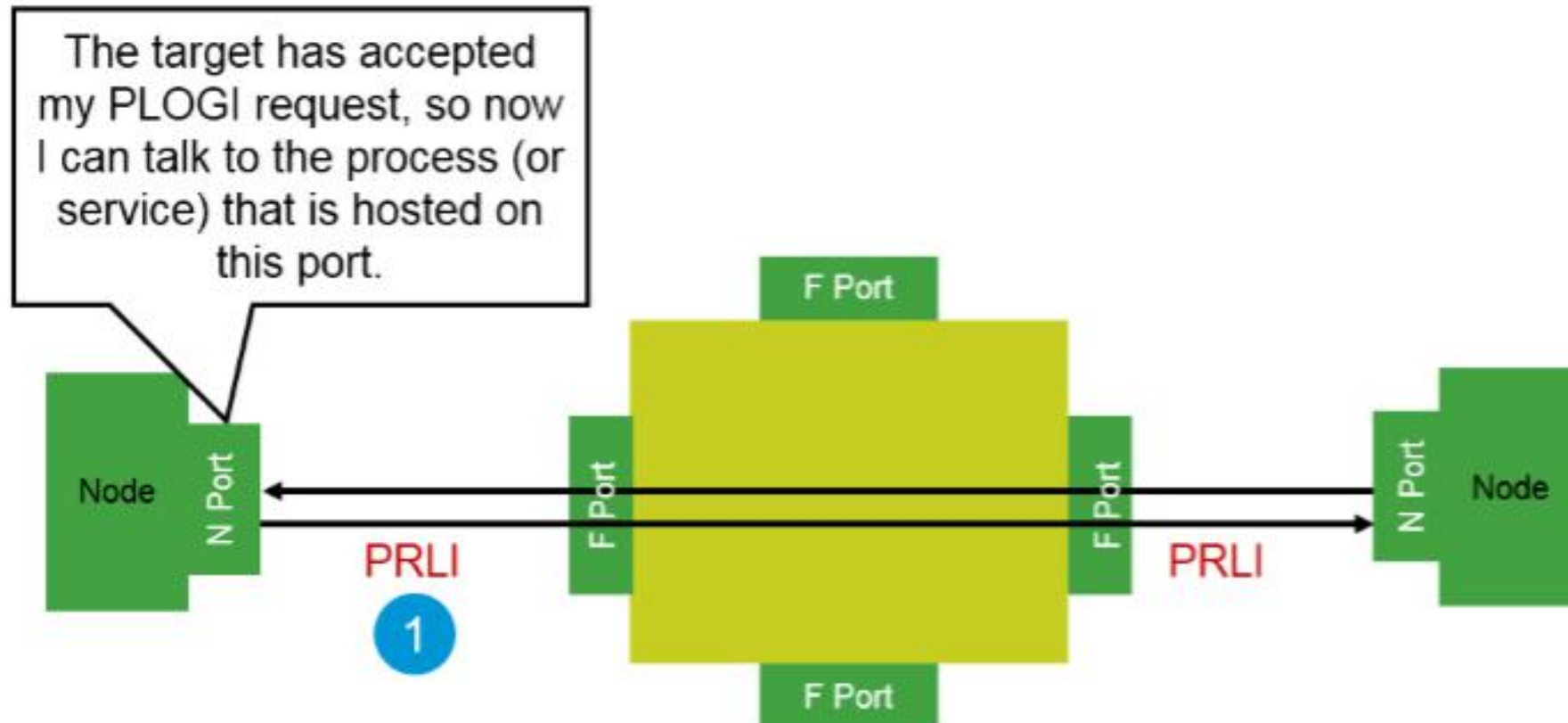
## Fibre Channel : FC-2端口注册过程 - PLOGI (con.)

- 2. Target N\_Port发送一个ACC（里面包括target的操作参数），响应Initiator N\_Port。操作系统驱动会管理initiator N\_Port保存信息，并且做一些参数上的修改。



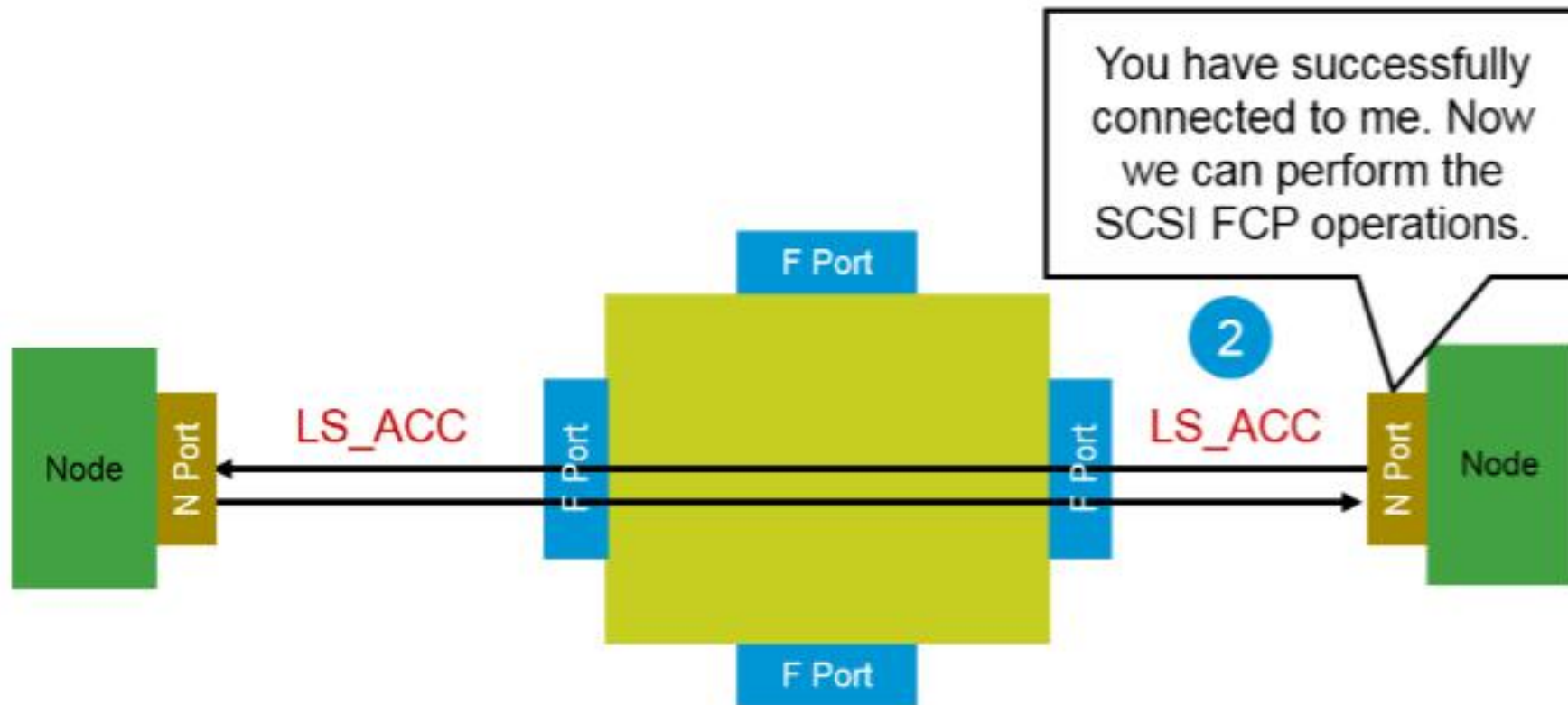
## Fibre Channel : FC-2端口注册过程 - PRLI

- 在PLOGI完成后，N\_Port之间互相知道对方的操作参数。这时initiator可以使用PRLI协议开启一个通道。PRLI协议用于在FC-4建立一个session。
- 1. initiator 发送一个PRLI帧包括ULP的支持信息



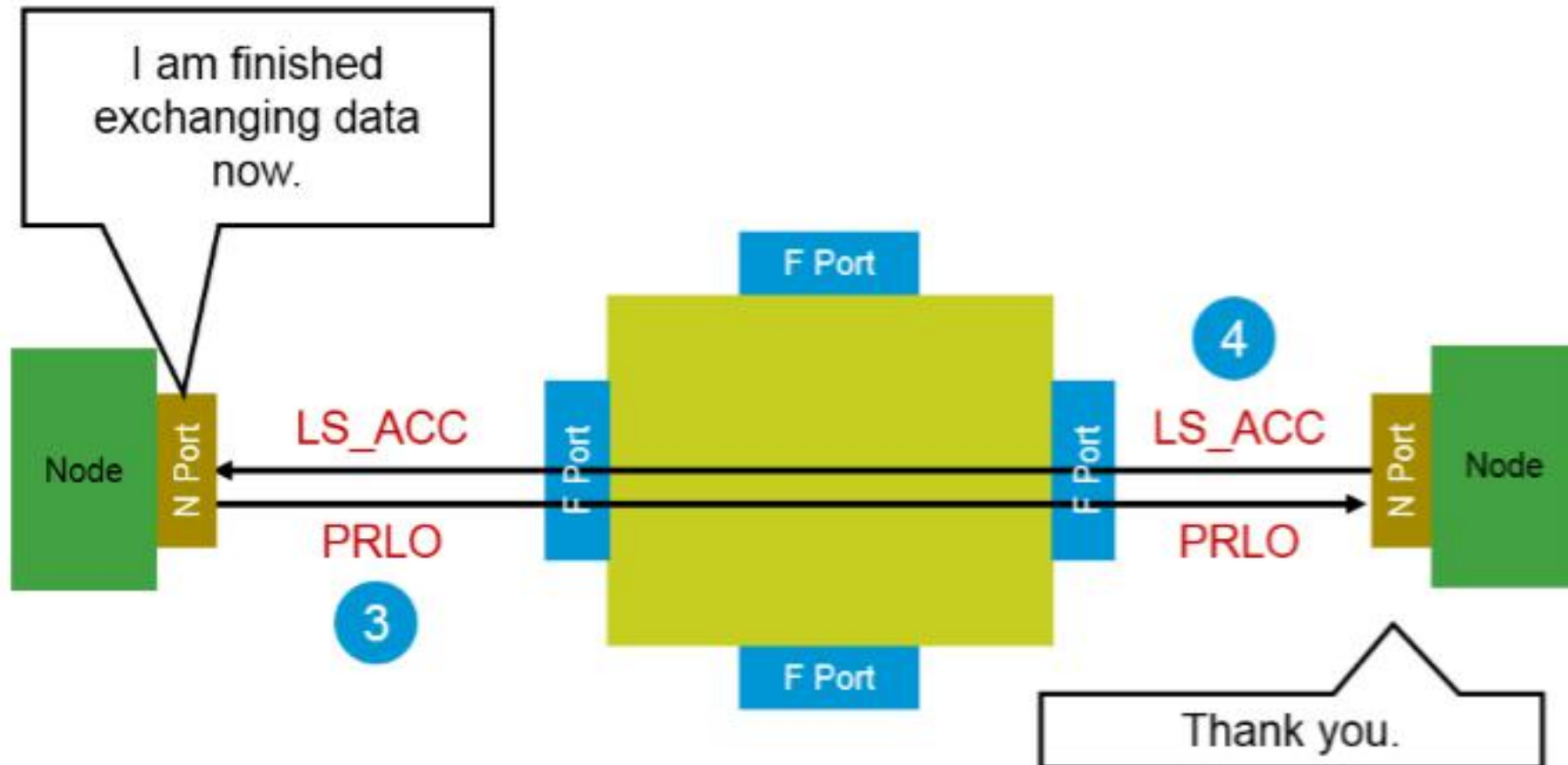
## Fibre Channel : FC-2端口注册过程 - PRLI (con.)

- 2.Target响应一个ACC包括ULP的支持的详细信息。这样一个channel建立成功，这个又称为image pair



## Fibre Channel : FC-2端口注册过程 - PRLI (con.)

- 3. 当Initiator与target之间完成exchange后，initiator发送一个PRLO帧
- 4. target响应一个ACC帧， image pair就终结了。



## VSAN

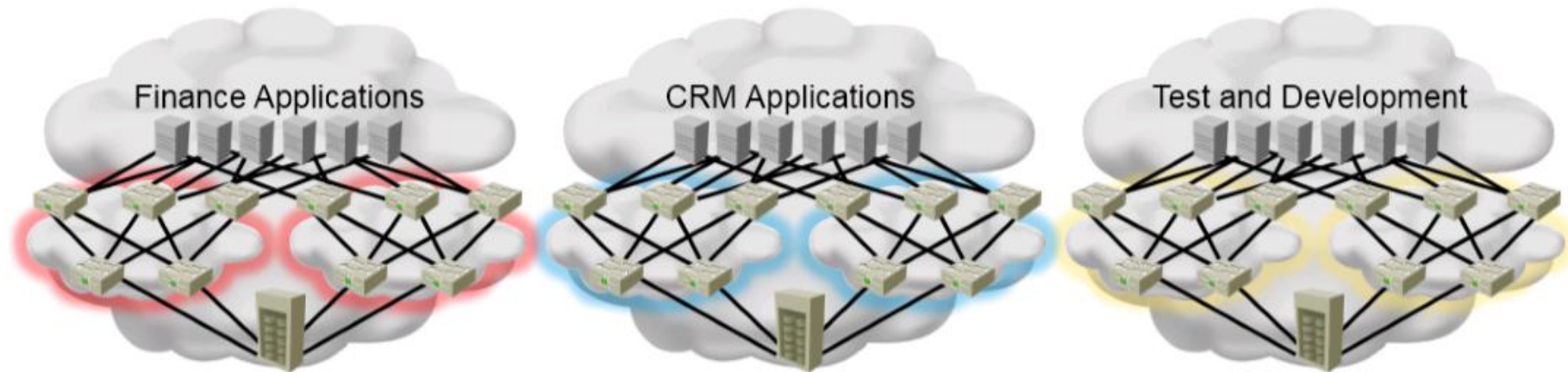


- 周涛
- QQ: 53408031 IE-LAB公开课群:79791756
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站:[www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024



## Fibre Channel : VSAN – Virtual Fabric

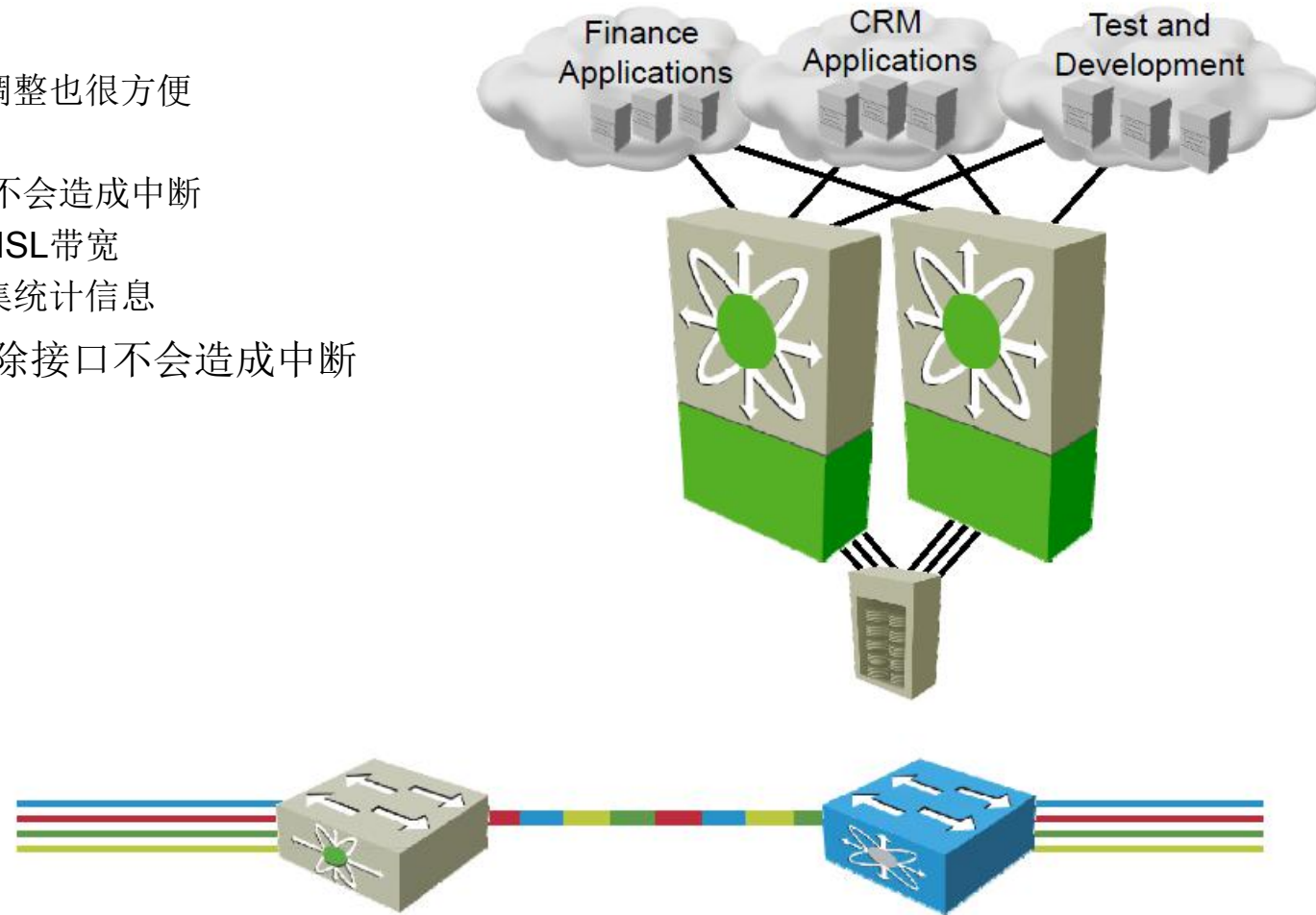
- 企业应用通常需要进行区分隔离：
  - 物理隔离、冗余的SAN
  - 在应用之间提供安全和隔离
  - 易于扩展和管理
- 限制：
  - 共享数据和资源比较困难
  - 资源利用率很低
  - 设备很多，需要用到很多的空间、电力和制冷





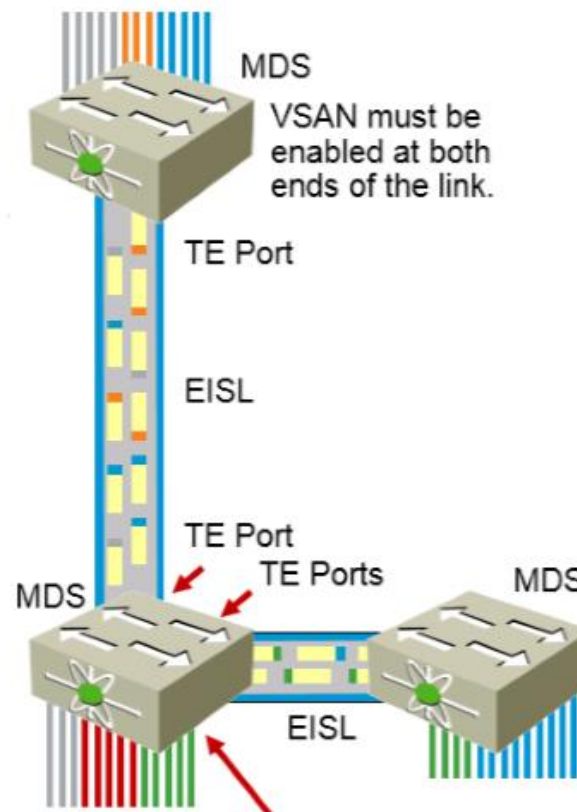
# Fibre Channel : VSAN – Virtual Fabric

- VSAN (官方叫法是Virtual Fabric), 与VLAN很类似。
  - 创建独立的virtual fabric, 并将接口划入其中
- Features:
  - 动态提供VSAN、调整也很方便
  - 提高端口利用率
  - 重新分配的时候, 不会造成中断
  - 通过trunking共享EISL带宽
  - 基于每个VSAN收集统计信息
- 从VSAN中添加、删除接口不会造成中断



# Fibre Channel : VSAN – TE端口和EISL

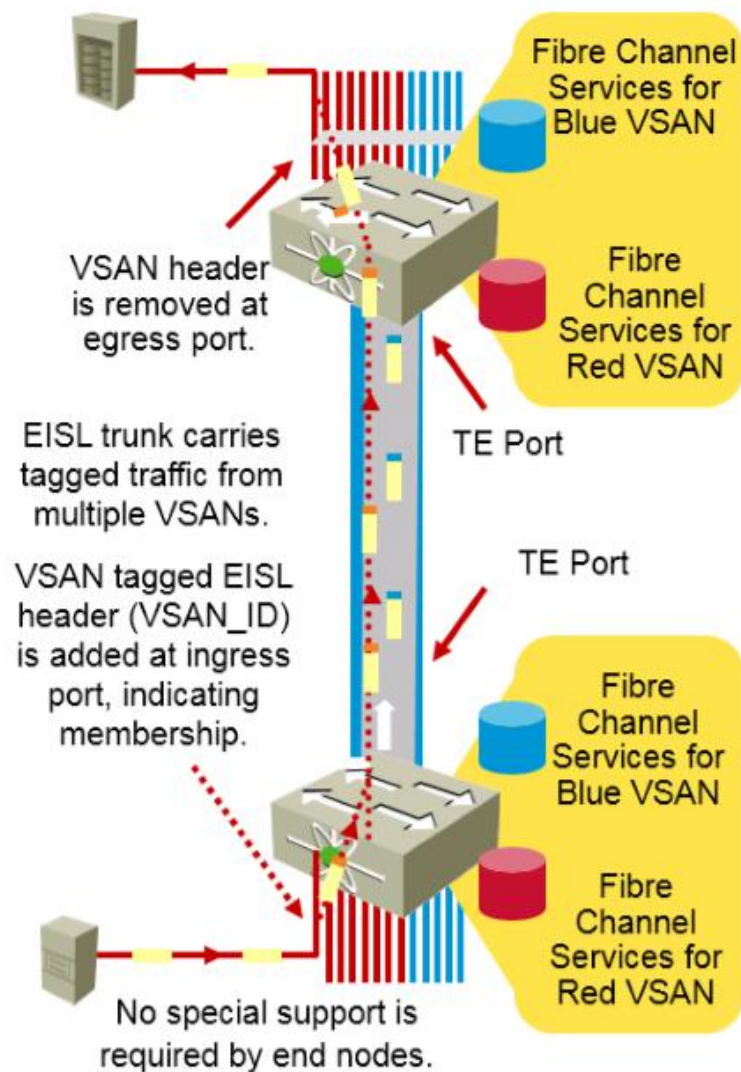
- TE端口和EISL跟Ethernet的Trunk的功能一致
- TE Port
  - 承载多个VSAN的数据帧
  - 默认允许所有VSAN(1-4093)通过
  - 可以通过VSAN-allowed list来控制哪些VSAN流量可以通过
  - E-port默认有native VSAN可以操作
- EISL – Enhanced ISL
  - 连接两个TE端口的链路
  - 承载每个VSAN单独的控制协议：FSPF、name server等



No devices are connected to blue VSAN, but blue VSAN must be enabled to allow routing between end switches.

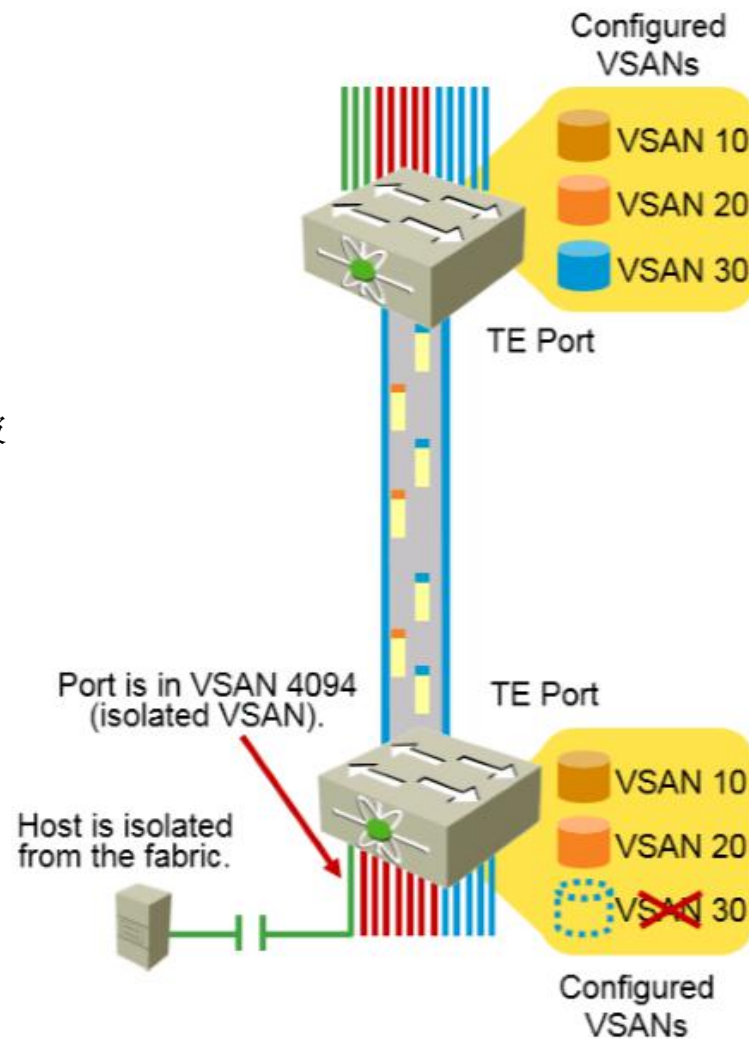
## Fibre Channel : VSAN 的主要功能

- 流量基于硬件隔离
- 对于终端设备是透明的，不需要额外的驱动和配置
- 帧在Trunk上(EISL)传递的时候，需要打上tag
- 每个VSAN有独立的Fabric Services
  - Name Server
  - Zone Server
  - Domain controller
  - Alias Server
  - Login Server
  - FSPF routing
  - Management



# Fibre Channel : VSAN 的号码

- VSAN 1:
  - 默认存在，交换机自动配置，不能被删除
  - 默认所有的端口都属于VSAN 1
- VSAN 2 – VSAN 4093:
  - 用户可配置的VSAN
  - 每个物理VSAN可以有4000个VSAN
- VSAN 4094 (Isolated VSAN):
  - 端口所属的VSAN被删除后，会被放入VSAN 4094，被隔离
  - 这个端口的流量不能通过交换机进行传递
  - 这个VSAN一直存在不能被删除



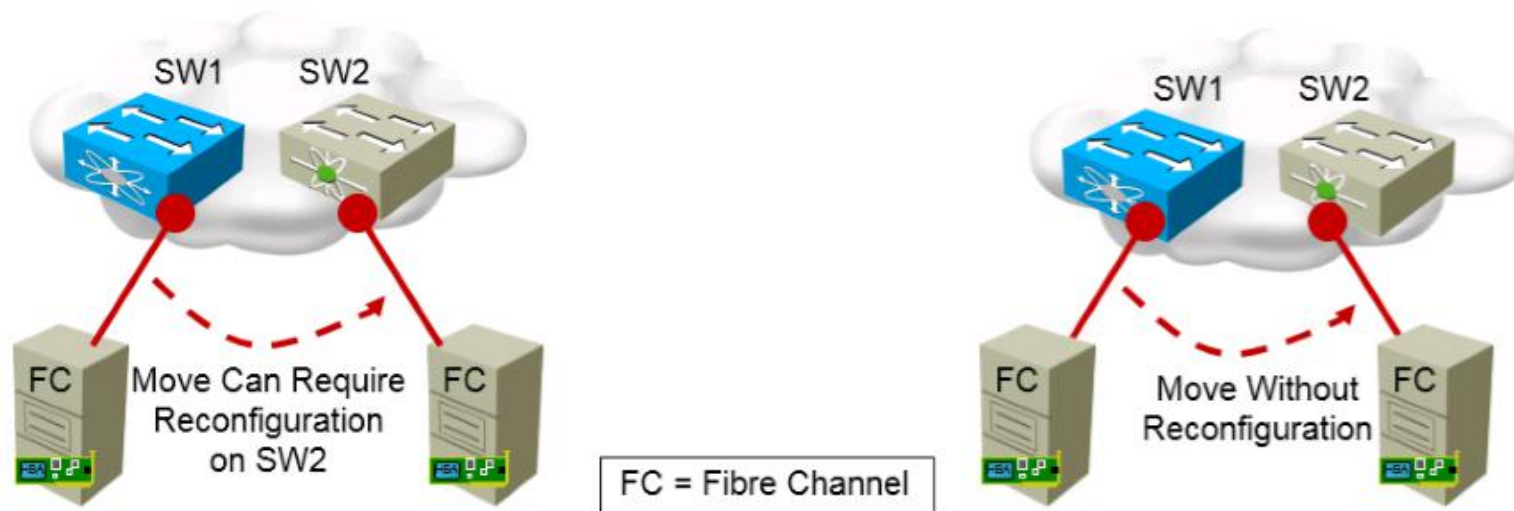
# Fibre Channel : Port- 和 WWN-based VSANs

## Port-based VSANs:

- 设备属于哪个VSAN基于交换机端口的配置
- 如果将Server或Storage移动到其他的switch上需要重新配置
- 交换机端口属于特定的VSAN

## WWN-based VSANs:

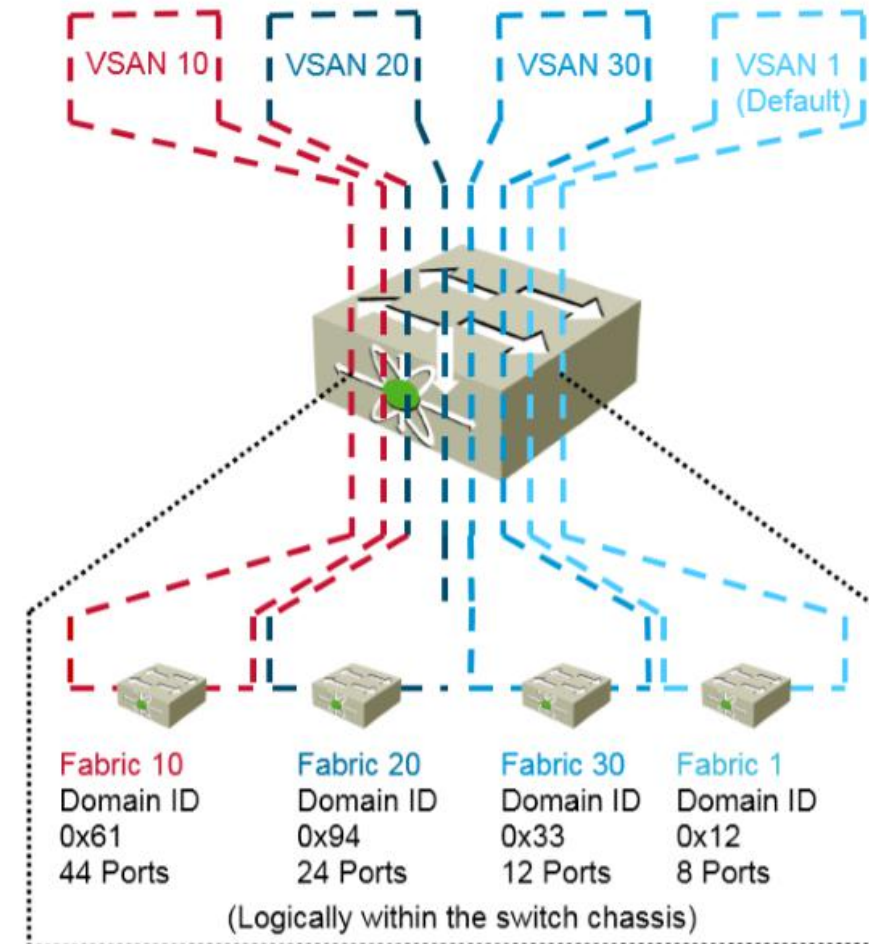
- 设备属于哪个VSAN基于Server或Storage的pWWN
- 通过CFS(Cisco Fabric Services)发布配置
- 当Server或Storage移动的时候不需要重新配置
- 终端设备的端口属于特定的VSAN





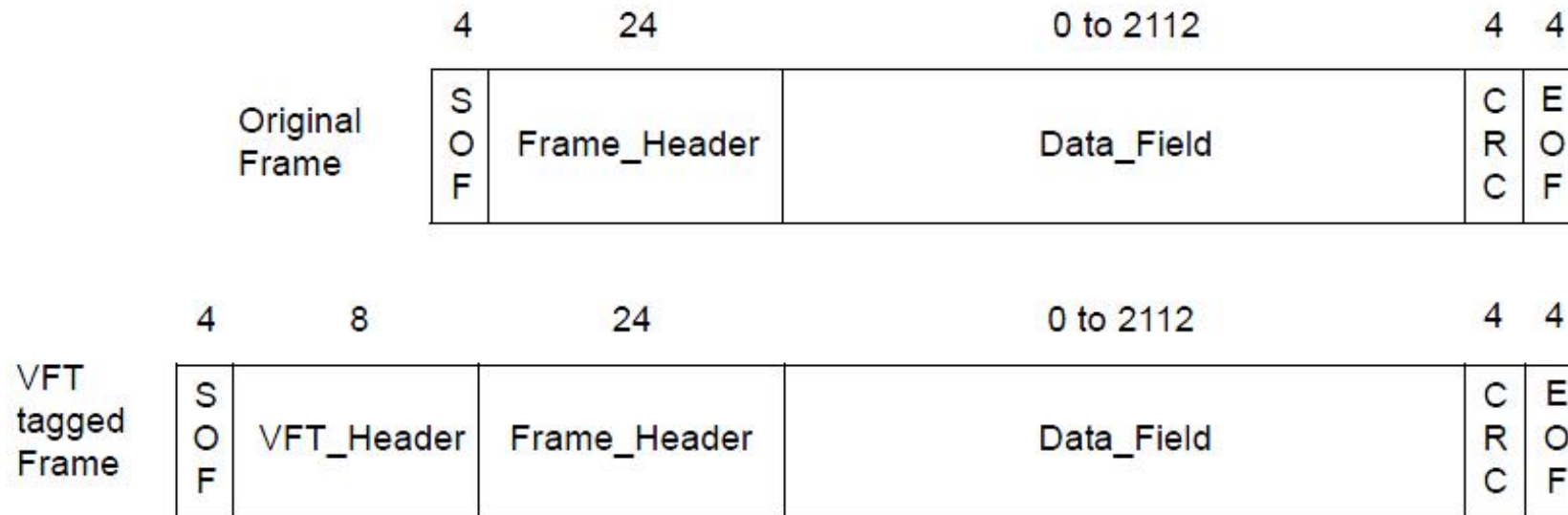
# Fibre Channel : VSAN Tagging

- 流量隔离：
  - 控制流量的入向和出向端口
- 在Fabric中的每个帧都需要唯一标识：
  - 在ingress-port添加一个labeled包括VSAN\_ID
  - 通过E-port, VSAN\_ID将被剥离
  - 通过TE-port, VSAN\_ID会被保留
- 在FC帧的头部包含priority用于做QoS
- 多个VSAN中可以重用FC-ID





# Fibre Channel : VSAN Tagging







## Fibre Channel : VSAN Tagging

Bits Word	31 .. 24	2 3	2 2	21 .. 18	1 7	1 6	15..13	12 .. 01	0
0	R_CTL	Ver		Type	R	E	Priority	VF_ID	R
1	HopCt	Reserved							

- R\_CTL: 8个bits, 应该设置为0x50, 表示是一个VFT\_Header扩展包头
- Version: 2个bits, 两个比特为0
- Type: 4个bits, 0x0
- R: 保留, 1个bit, 0
- E: 1个bit, 表示ESP帧头是否存在。设置为0, 表示没有ESP\_Header; 设置为1, 表示VFT\_Header后面有一个ESP\_Header
- Priority: 3个bits, 用于QoS与以太帧头的CoS的格式和意义一样
- VF\_ID: 12bits, 表示VSAN
- HopCt: 8bits, 表示帧在丢弃前, 还有多少跳可以被传输

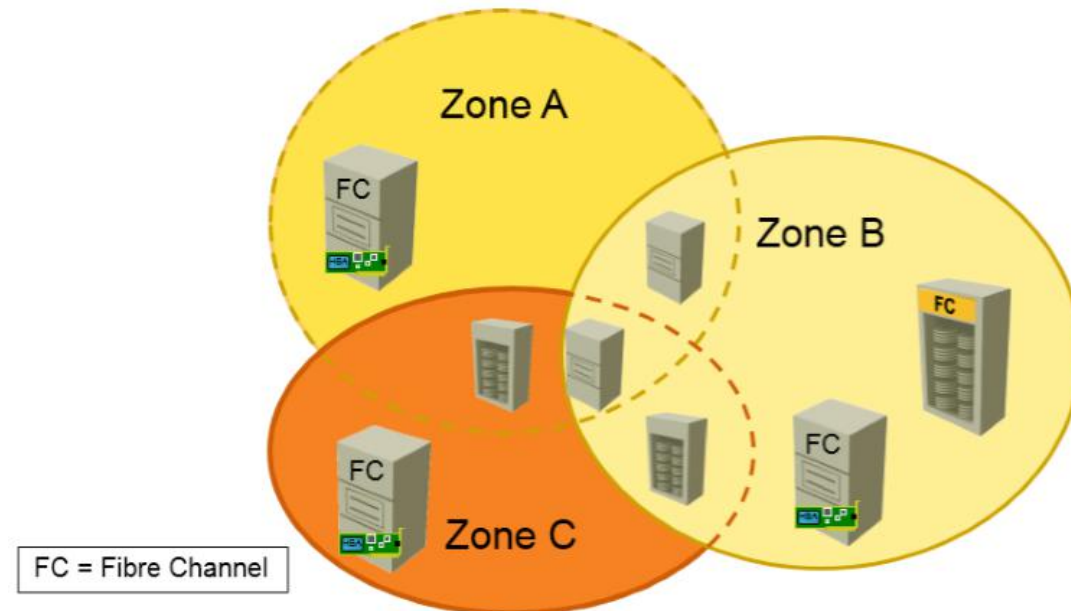
## Zoning



- 周涛
- QQ: 53408031 IE-LAB公开课群:79791756
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站:[www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024

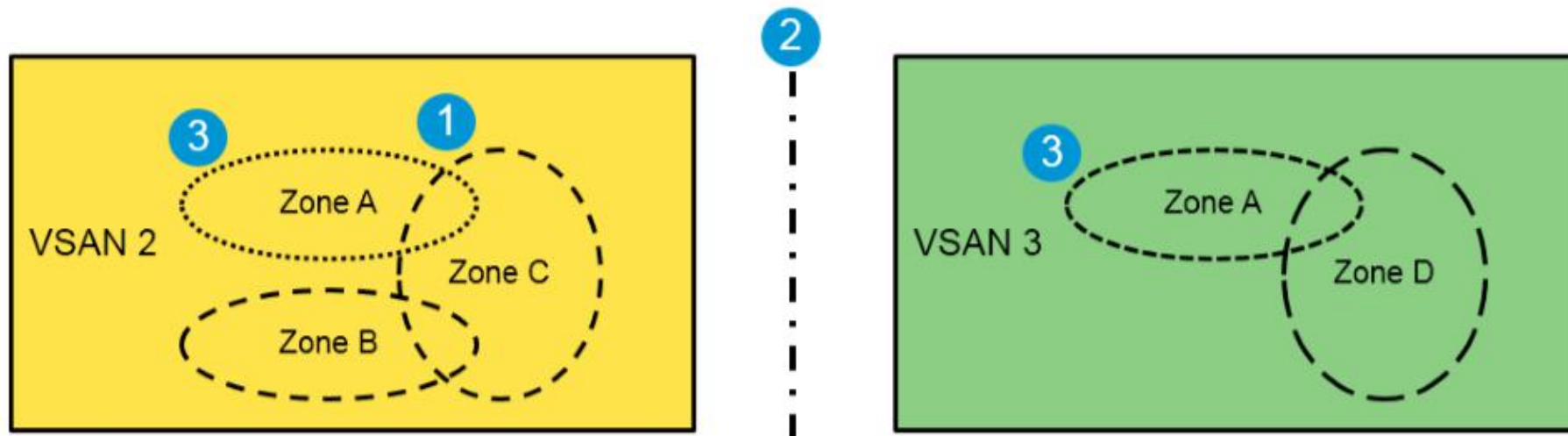
# Fibre Channel : Zoning简介

- Zone用于限制在相同Fabric或VSAN中哪些设备可以互访:
- Zone由一个或多个zone member组成
- Zoneset由一个或多个zone组成
- 在一个zone内，大多数进行本地化管理



# Fibre Channel : Zoning规则

- 1.Zone可以overlap
- 2.Zone通常不会在VSAN之间配置
  - Zone的操作通常在一个VSAN内部完成
  - 除非做了IVR，这样zone才能在VSAN间操作
- 3.Zone每VSAN有效
  - 两个VSAN都有Zone A，但是它们之间是不同的，同时也是分离的





## Fibre Channel : Zoning成员和执行方式

- **Soft Zoning:**
  - 在交换机软件中部署，并且由name server执行
  - Name Server会响应同一个Zone内的discovery query
- **Hard Zoning:**
  - 在端口ASIC由ACL执行
  - 对所有数据流执行操作
- **Zone成员类型:**
  - pWWN、fWWN，FCID，interface和sWWN
  - Domain ID和port number
  - IP address
  - Symbolic node name (比如iSCSI-qualified name)
  - Fibre Channel 或 device alias



## Fibre Channel : VSAN和Zoning的对比

	VSAN	Zone
路由和名称空间	每个VSAN使用它们单独的路由和名称空间	在同一个VSAN中的Zone使用相同的routing database
流量泛洪	有限制的单播、组播和广播流量	有限制的单播
成员	基于物理端口或pWWN	基于pWWN
终端	一个HBA或存储设备可以属于一个单独的VSAN	一个HBA或storage可以存在于多个zone内
哪些接口支持VSAN	源端口、目的端口和E-port	只有源和目的端口
范围	在大的环境中部署	一组initiators和targets, zone外是不可见的
配置改变	根据需求改变	可以经常改变 (比如备份)
推荐使用	按照应用或按照部分来划分	Single-initiator zoning

# Fibre Channel Flow Control

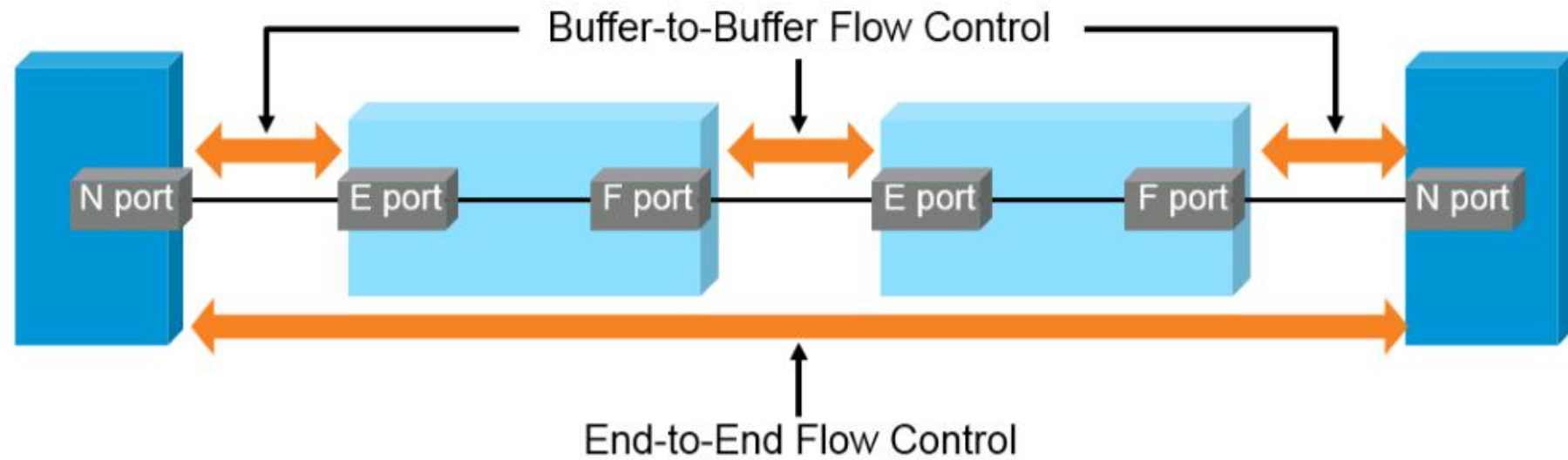


- 周涛
- QQ: 53408031 IE-LAB公开课群:79791756
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站:[www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024



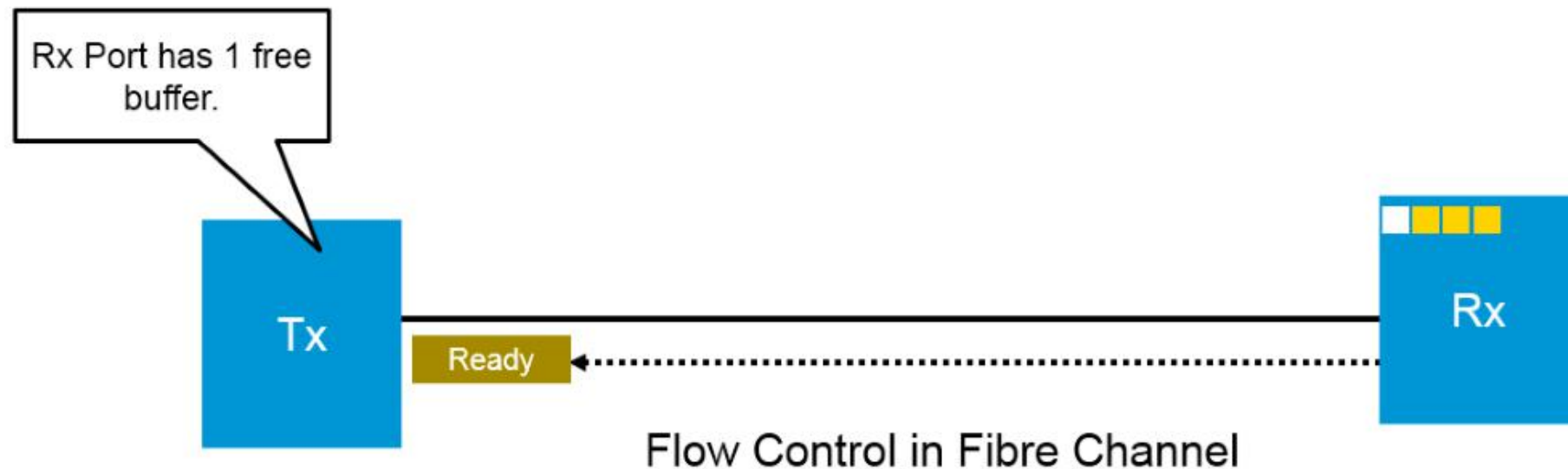
## Fibre Channel : Flow Control

- Fibre Channel定义了两种flow control的方法:
- Buffer-to-Buffer (port-to-port)
- End-to-End (source-to-destination)



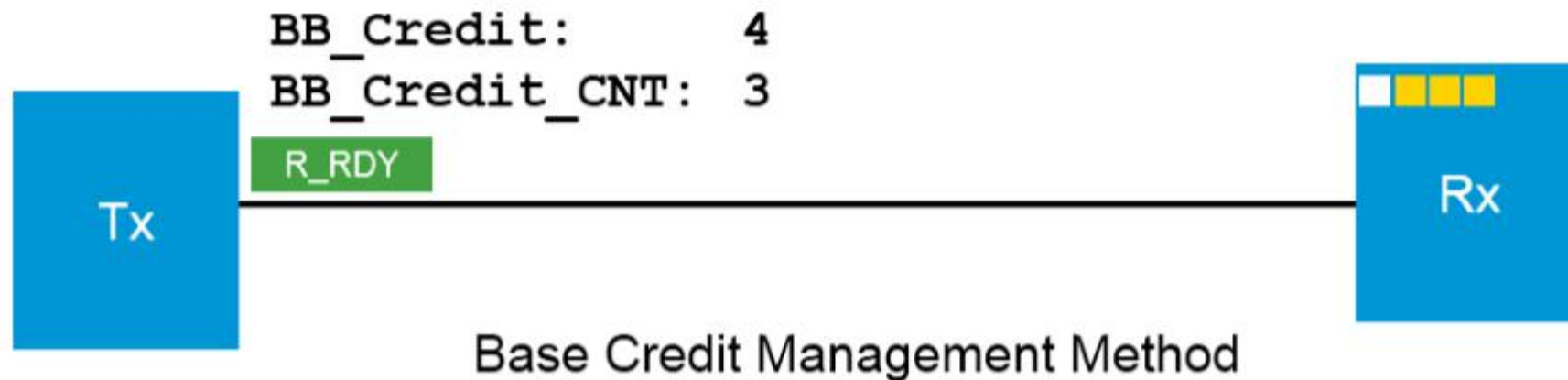
## Fibre Channel : Flow Control – Buffer-to-Buffer Flow Control

- Fibre Channel使用credit-base的方法：
  - 发送者必须等接收者告诉传输者它可以接收另外一个数据帧，发送者才可以发送。永远由接收端控制流量的发送。
- 优点：
  - 防止因为buffer overruns (缓冲区溢出) 而造成的丢包
  - 在高速流量负载下可以达到最大性能



## Fibre Channel : Flow Control – Buffer-to-Buffer Flow Control

- 1. 当Tx端口发送一个PLOGI request, Rx端口会响应一个ACC帧。在ACC帧里面包含接收端的frame buffer的大小和数目。这个叫做(BB\_Credit)。Tx端口保存BB\_Credit的值在表项内部
- 2. Tx端口同时保存另外一个值, 叫做BB\_Credit\_CNT。这个值表示buffer credit已经使用的数字。在端口完成PLOGI之后, BB\_Credit\_CNT设置为0。
- 3. 每次Tx端口发送一个帧, 同时增加BB\_Credit\_CNT的数值。
- 4. 接收端收到帧以后, Rx端口处理帧并且移到ULP缓冲区中。如果buffer是可用的, Rx端口会发送一个“接收者准备好”(R\_RDY) 确认信令返回给Tx端口。
- 5. 当接收端收到R\_RDY信令, 它会减少BB\_Credit\_CNT的数值。



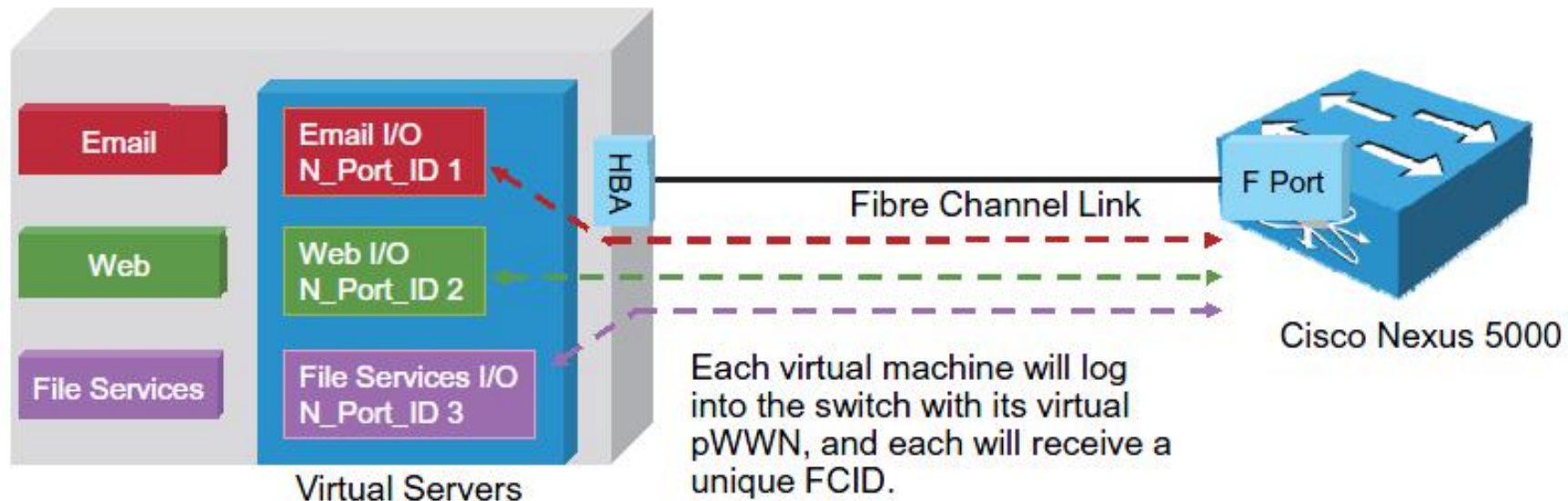
## NPV & NPIV



- 周涛
- QQ: 53408031 IE-LAB公开课群:79791756
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站:[www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024

## N-Port ID Virtualization

- N-Port ID Virtualization(NPIV)提供一种给一个N Port分配多个FCID的方法
- NPIV允许多个应用共享同一个HBA
- 每个应用(VM)拥有自己的pWWN，允许基于这些pWWN做访问控制、Zoning和端口安全等
- 对于服务器虚拟化特别有用，诸如：Vmware，Microsoft Hyper-V和Citrix XenServer





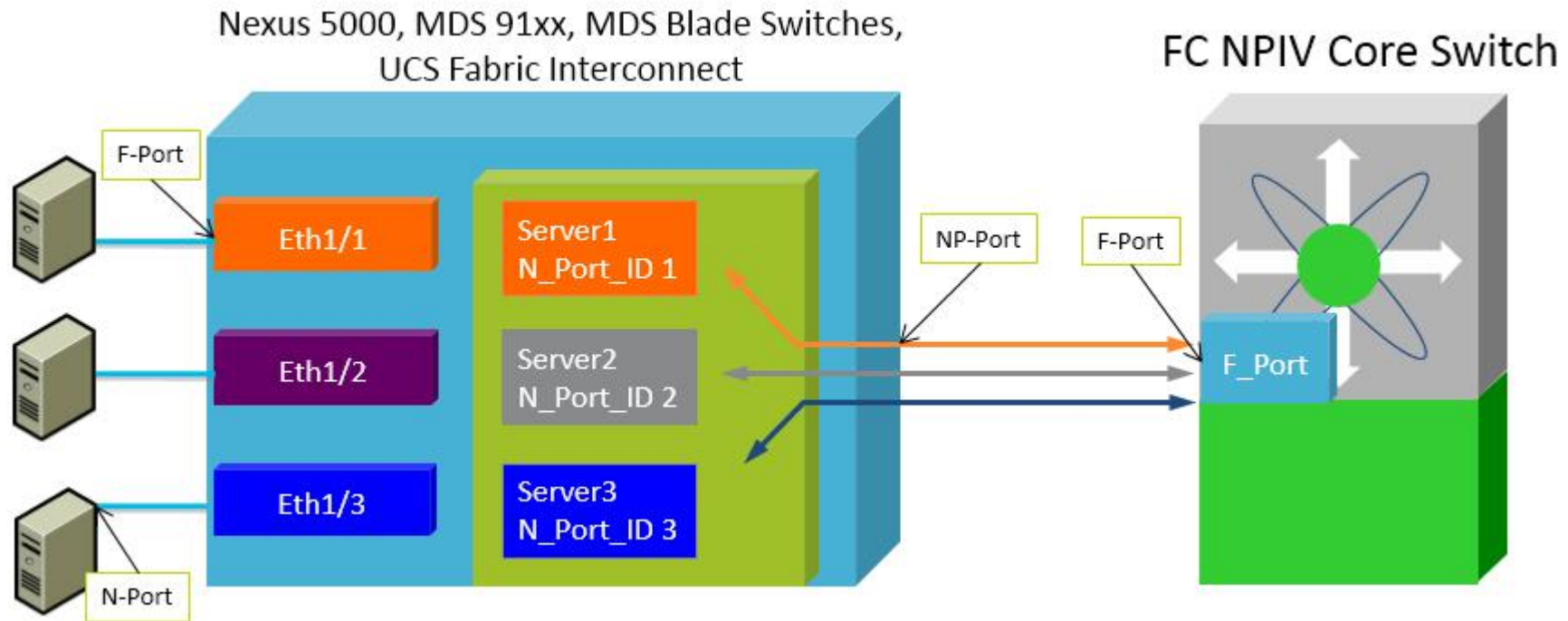
## NPV和NPIV支持的设备

- Edge Switches (NPIV/NPV)
  - Cisco 9124, MDS 9134, MDS 9148
  - IBM和HP的Fibre Channel blade switches
  - Cisco Nexus 5000 和 5500系列交换机
  - Cisco UCS 6100 和 6200 FI
- Core Switches(NPIV)
  - Cisco MDS 9500系列Multilayer Directors
  - Cisco MDS 9216 Multilayer Fabric Switch and 9222i
  - Cisco MDS 9124 , 9134和9148交换机
  - 第三方交换机



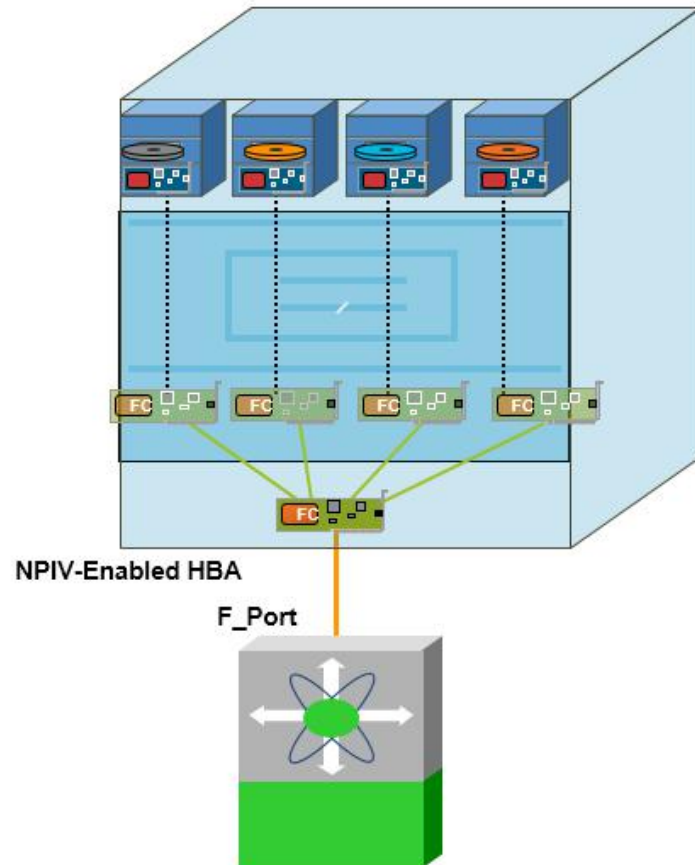
## N-Port Virtualization

- 对NPIV进行扩展
- 允许刀片交换机或ToR交换机虚拟成一块HBA卡连接到FC交换机上
- NPV设备聚合本地连接的N Port的信息到一条或多条上行链路(pseudo-interswitch links) 到核心交换机
- FLOGI/FDISC

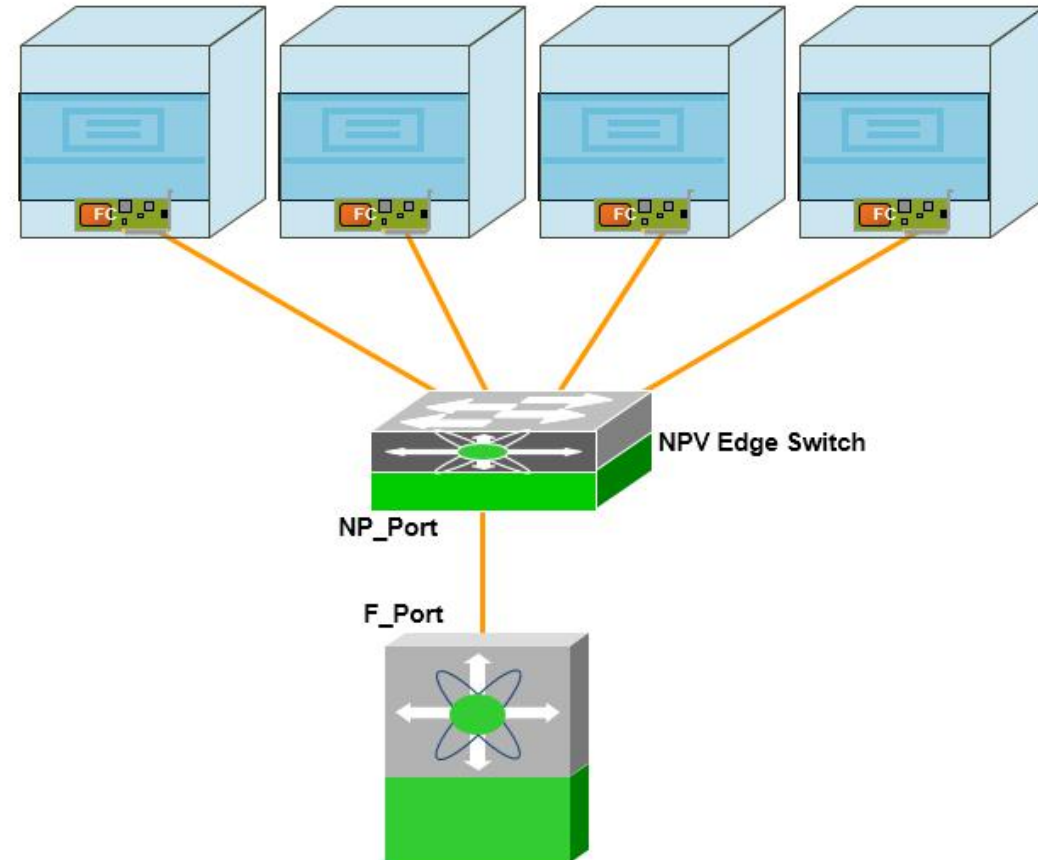




## Virtual Machine Aggregation



## 'Intelligent Pass-Thru' (NPV)





## N-Port Virtualization

Feature	普通的交换机	NPV
Fibre Channel Service	<ul style="list-style-type: none"><li>支持所有的FC Service(FLOGI、name server、zoning、domain server、FSPF等)</li><li>FSPF, zoning和name Server database分发到所有直连的交换机上</li></ul>	大多数服务是关闭的
Switching operation	本地交换机	<ul style="list-style-type: none"><li>设备相当于一个代理</li><li>相当于是上游交换机的下属</li></ul>
Domain ID	每个交换机都拥有Domain ID	不用Domain ID
扩展性和可管理性	<ul style="list-style-type: none"><li>ISL是FSPF路由表中的路径</li><li>一个port-channel最多可以有16个接口</li><li>理论上在一个Fabric中可以有239台交换机</li></ul>	<ul style="list-style-type: none"><li>不需要管理员做一些SAN的管理(比如FSPF、FC policy)</li><li>VSAN的扩展</li></ul>
QoS	支持	不支持



## SAN网络和融合网络介绍

周涛

QQ: 53408031

Mobile:

(86)18611846551

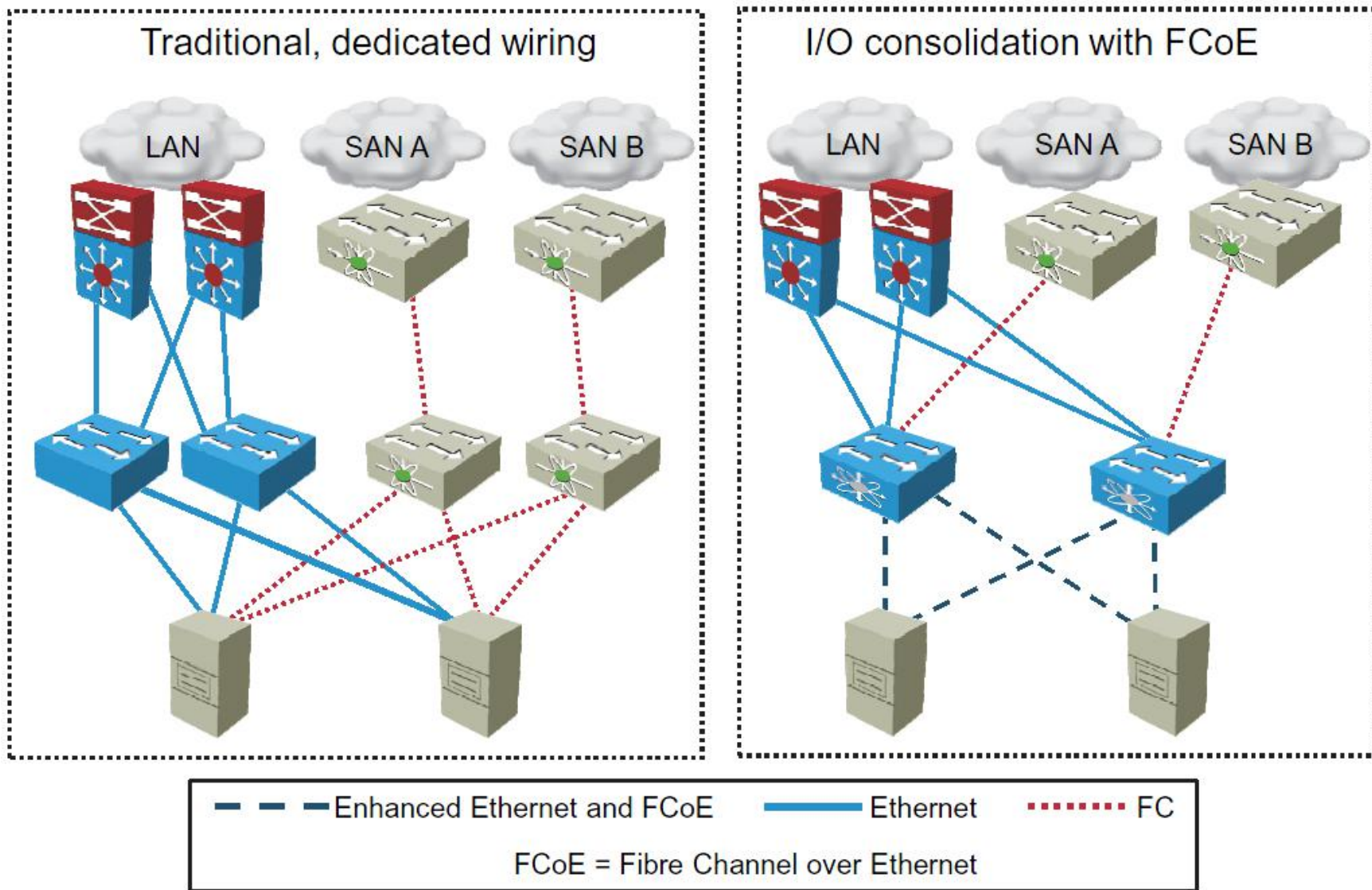
Site: [www.ie-lab.cn](http://www.ie-lab.cn)

YY直播: 58761024

## SAN网络和融合网络介绍

1. 存储简介
2. Fibre Channel 术语
3. Fibre Channel 工作原理
- 4. 融合网络
5. FCoE的术语
6. FCoE的工作原理
7. FCoE的标准集
8. 传统数据中心网络向融合网络迁移







# 融合网络 – Ethernet vs. Fibre Channel

## Ethernet Economic Model

- 直接连接在主板上
- 驱动集成到O/S中
- 多厂商支持
- 主流技术
- 懂的人多
- 互操作性很好

## FC Economic Model

- 永远是单独的接口卡
- 特殊驱动
- 支持的厂商很少
- 特殊技术
- 需要特殊的专业知识
- 互操作性需要测试



# 了解FCoE

Fibre Channel is to FCoE

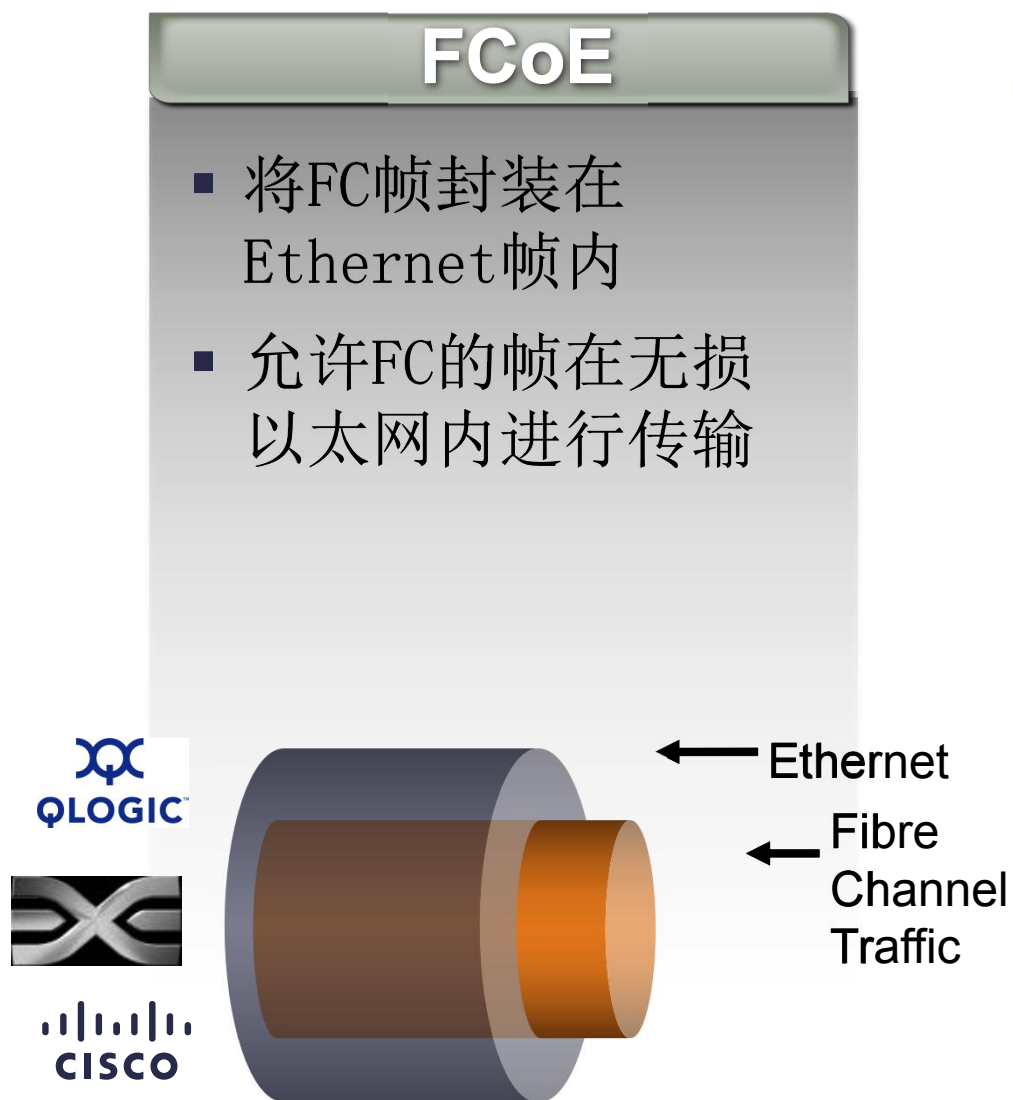
As



is to



# FC over Ethernet (FCoE)



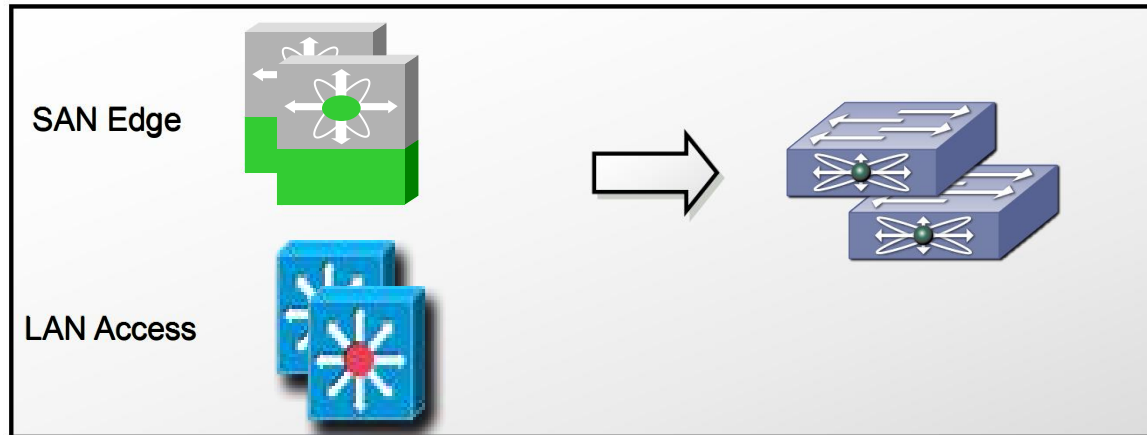
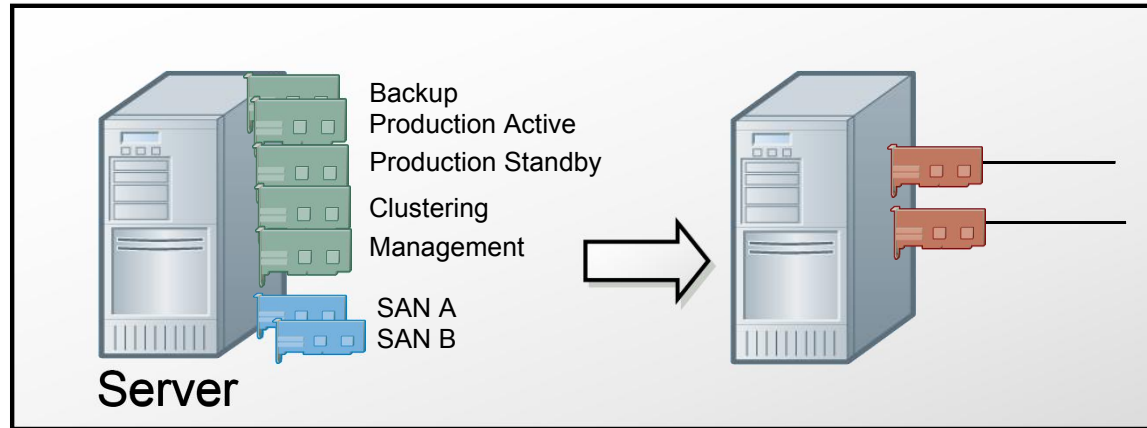
## 优点

- 用更少的线路
  - block I/O & Ethernet traffic 共存在同一线路中
- 用更少的适配器
- 耗电小
- 可以与现存的SAN网络互操作
  - 保持管理的一致性
- 没有网关

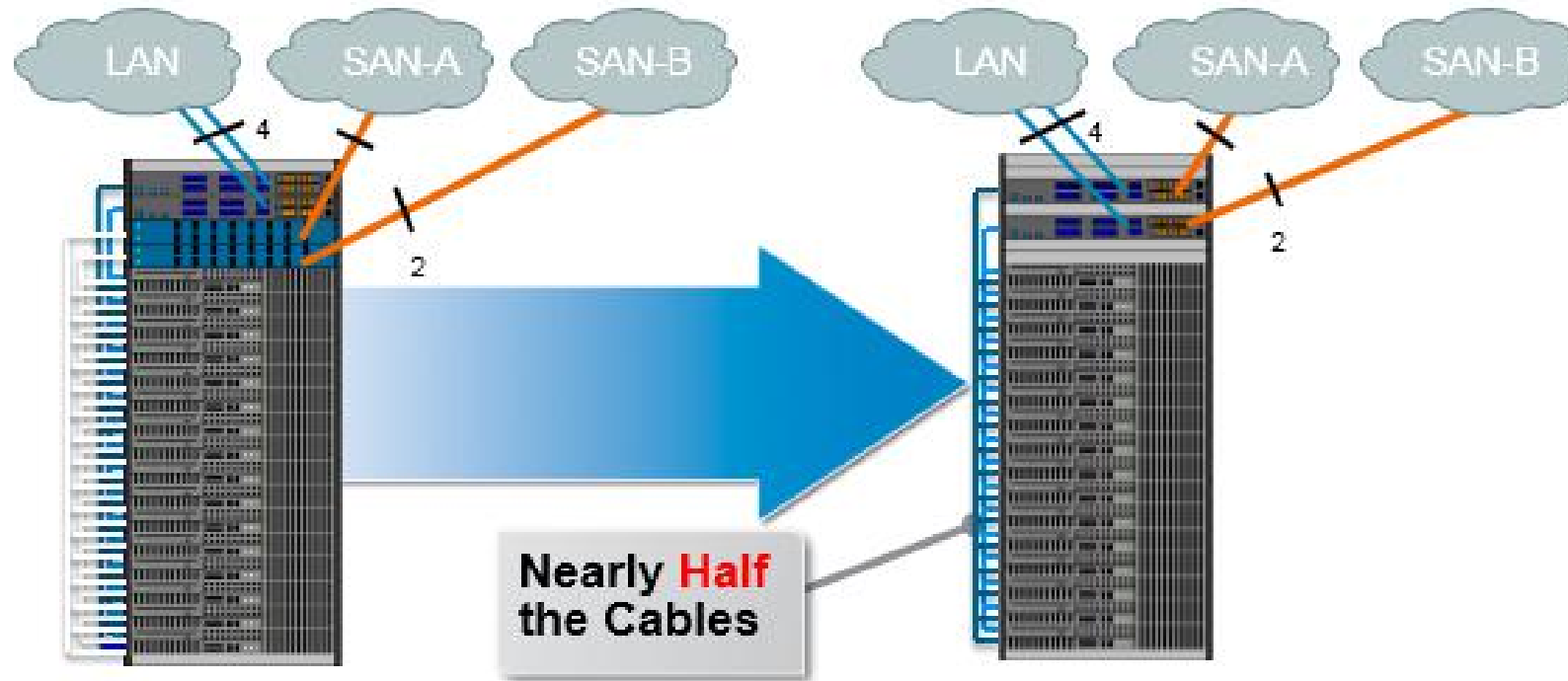




# FCoE使能Unified Fabric



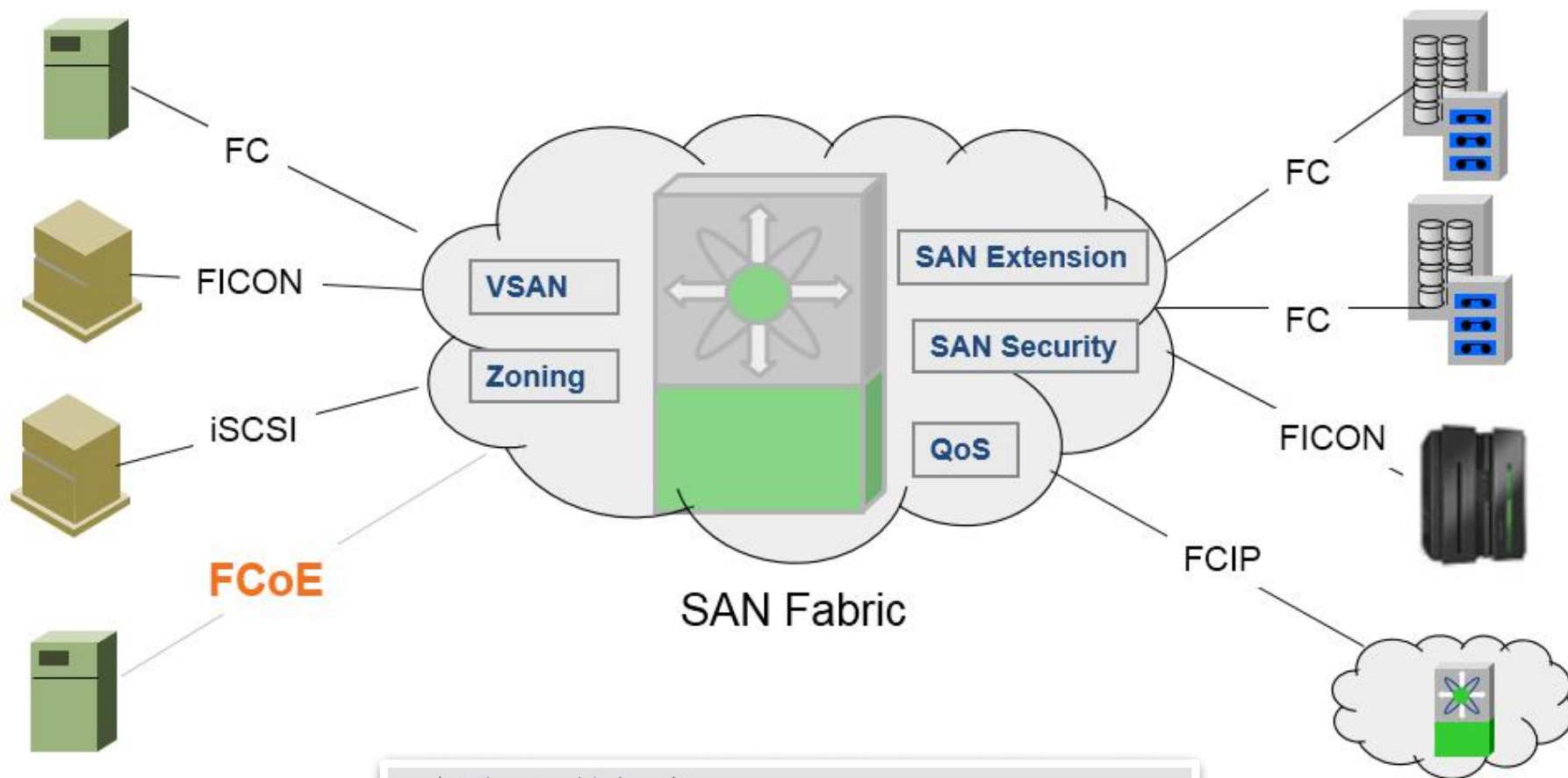
# FCoE Cabling Reduction



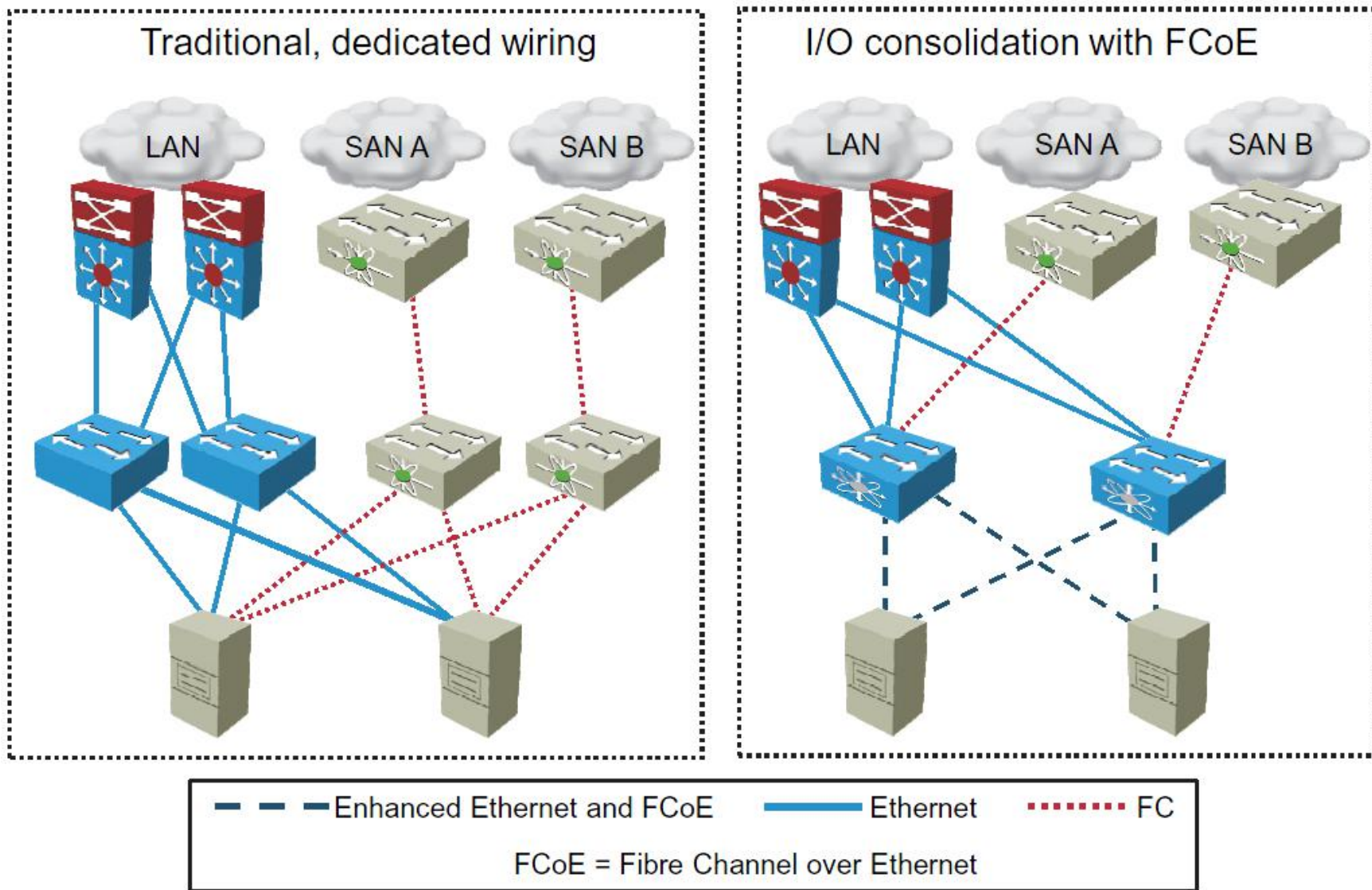
16 Servers	Enet	FC	Total
Adapters	16	16	32
Switches	2	2	4
Cables	36	36	72
Mgmt. Pts.	2	2	4

16 Servers	Enet	FC	Total
Adapters	16	0	16
Switches	2	0	2
Cables	36	4	40
Mgmt. Pts.	2	0	2

# FCoE 是 FC SANs 一种扩展

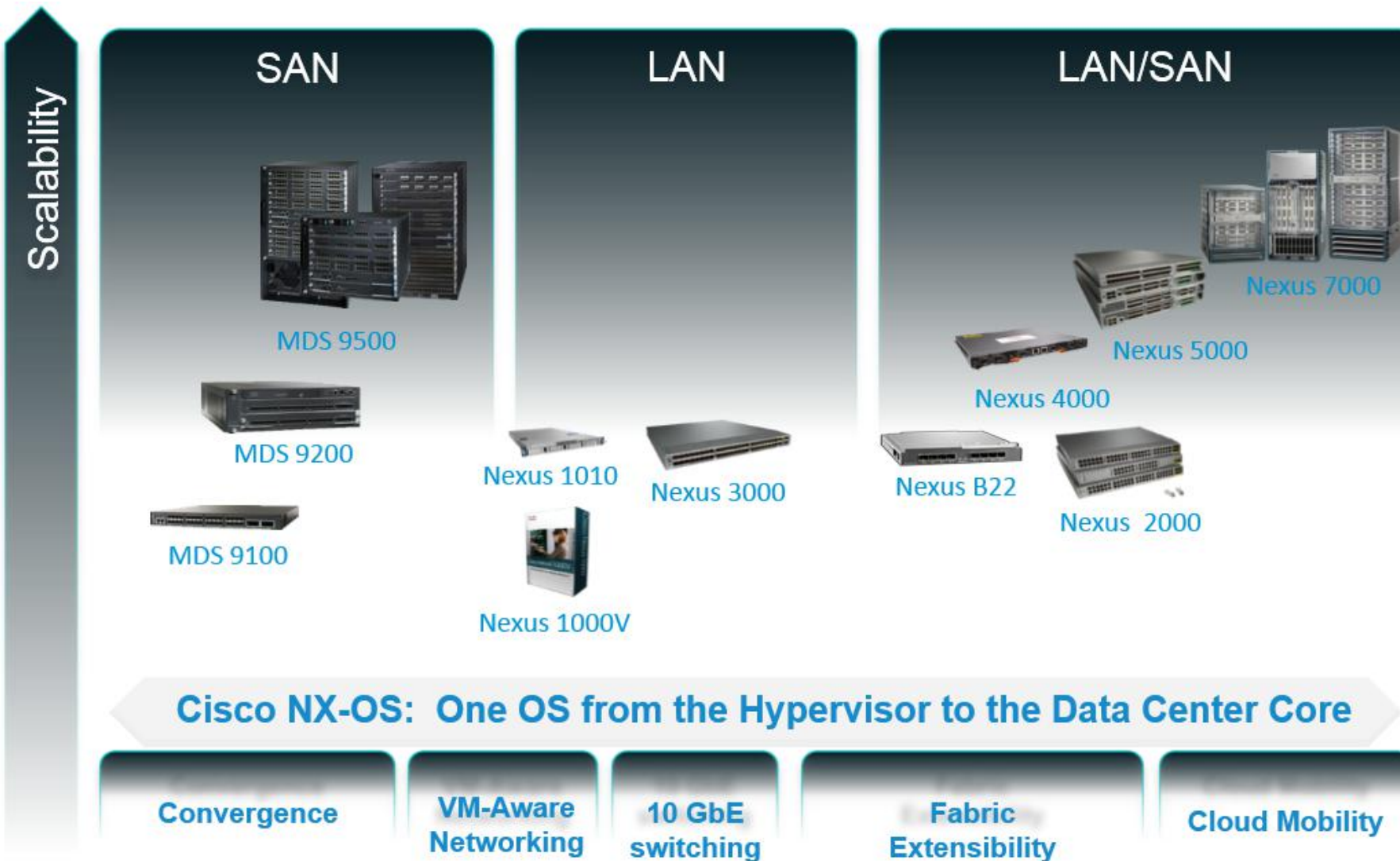


- 保留FC的投资
- 简化了服务器连接SAN的连接方式





# Cisco支持FCoE的产品



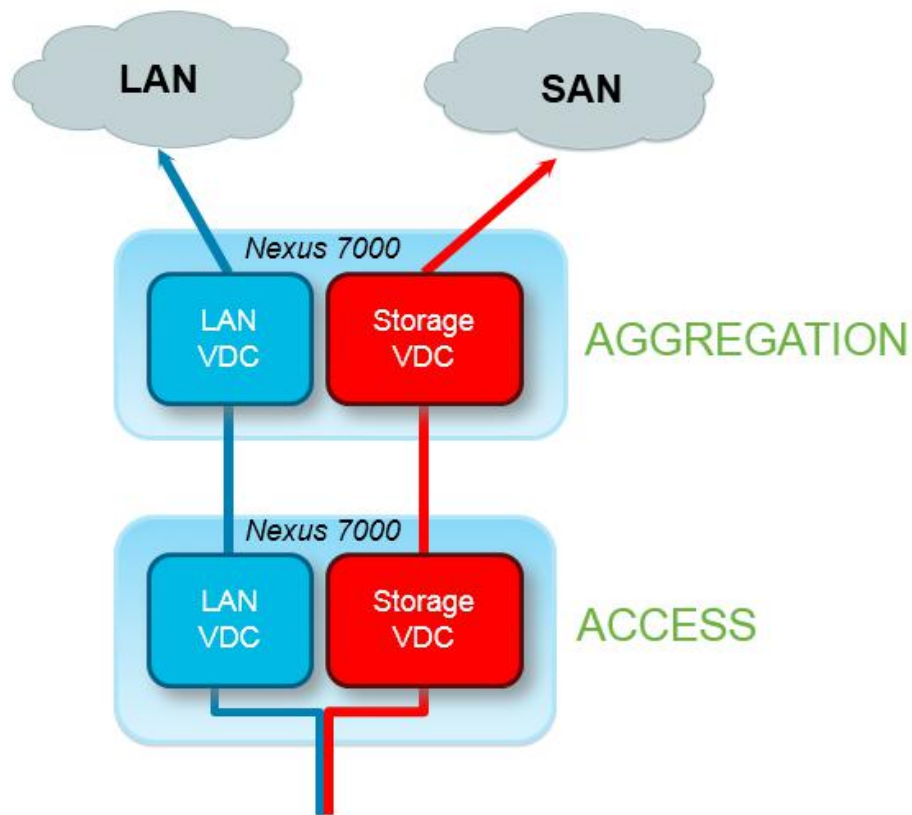
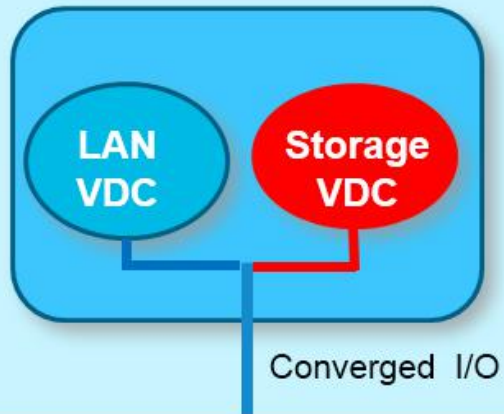


# Nexus 7000使用Storage VDC

## A Virtual MDS

### Dedicated Storage VDC – Converged Interfaces

- 连接host/target的接口，不是 ISLs
- 单独支持Storage相关协议的VDC
- 入向以太流量基于0x0800
- 入向FCoE流量被Storage VDC所处理



VDCs 提供高可用性和错误隔离



## SAN网络和融合网络介绍

周涛

QQ: 53408031

Mobile:  
(86)18611846551

Site: [www.ie-lab.cn](http://www.ie-lab.cn)

YY直播: 58761024

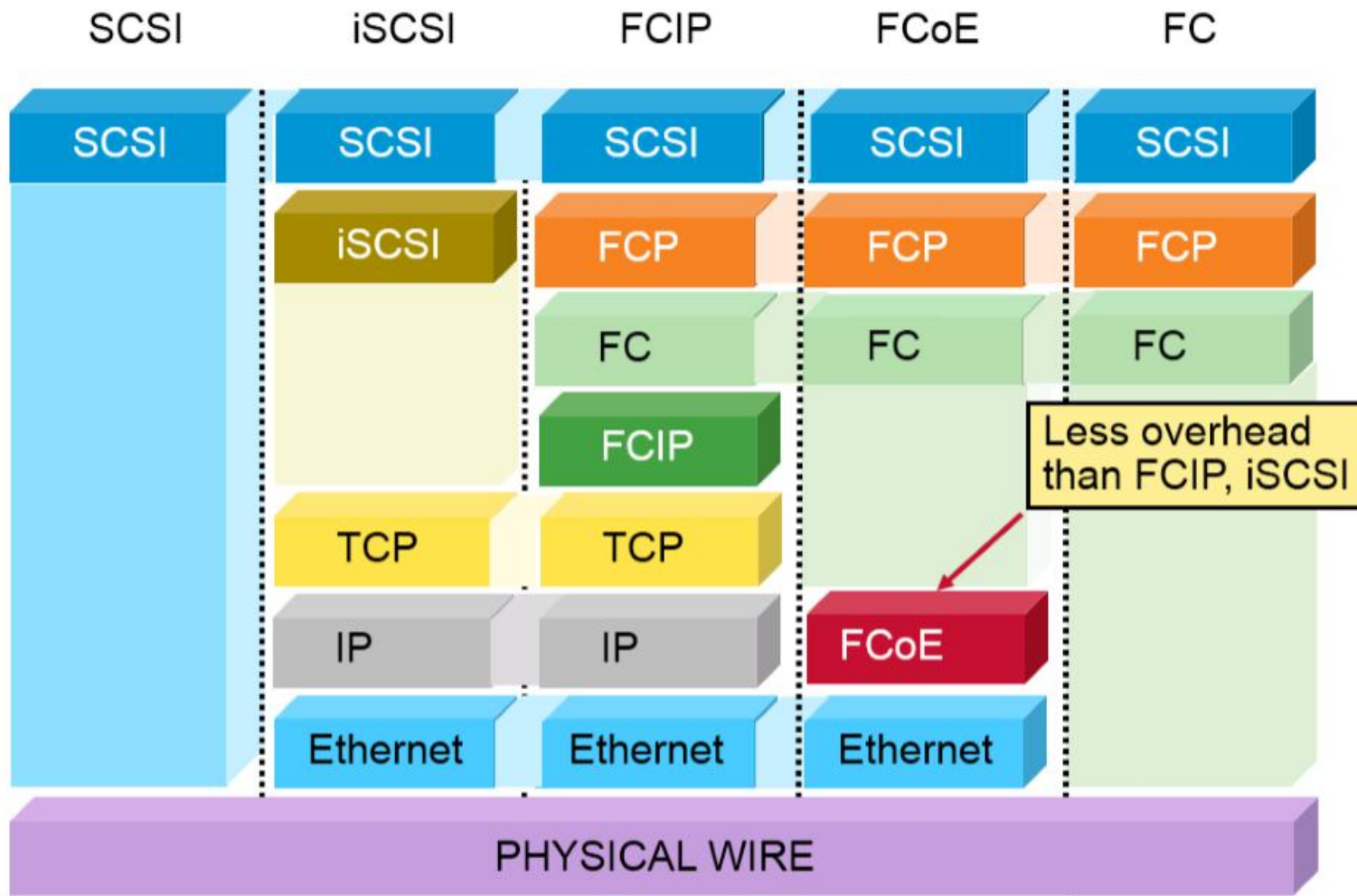
## SAN网络和融合网络介绍

1. 存储简介
2. Fibre Channel 术语
3. Fibre Channel 工作原理
4. 融合网络
- ➔ 5. FCoE的术语
6. FCoE的工作原理
7. FCoE的标准集
8. 传统数据中心网络向融合网络迁移





# 架构比较



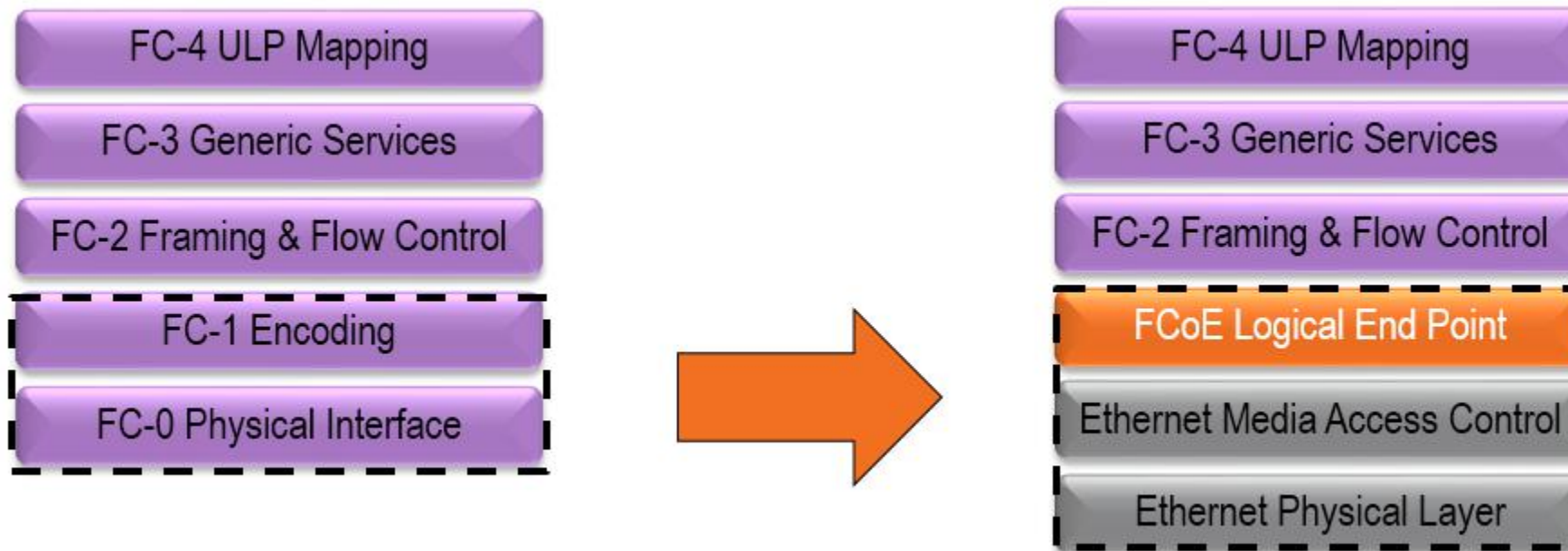
SCSI = Small Computer System Interface ; iSCSI = Internet Small Computer System Interface  
 FC = Fibre Channel



## FCoE工作在第几层

- Ethernet用于替代FC的FC-0和FC-1
- FC-2分成FC-2P, FC-2M, FC-2V
  - 去掉FC-2P和FC-2M保留FC-2V
  - 用FCoE LEP替代

Fibre Channel Protocol Stack	FCoE Protocol Stack
FC-4	FC-4
FC-3	FC-3
FC-2V	FC-2V
FC-2M	FCoE Entity
FC-2P	
FC-1	Ethernet MAC
FC-0	Ethernet PHY



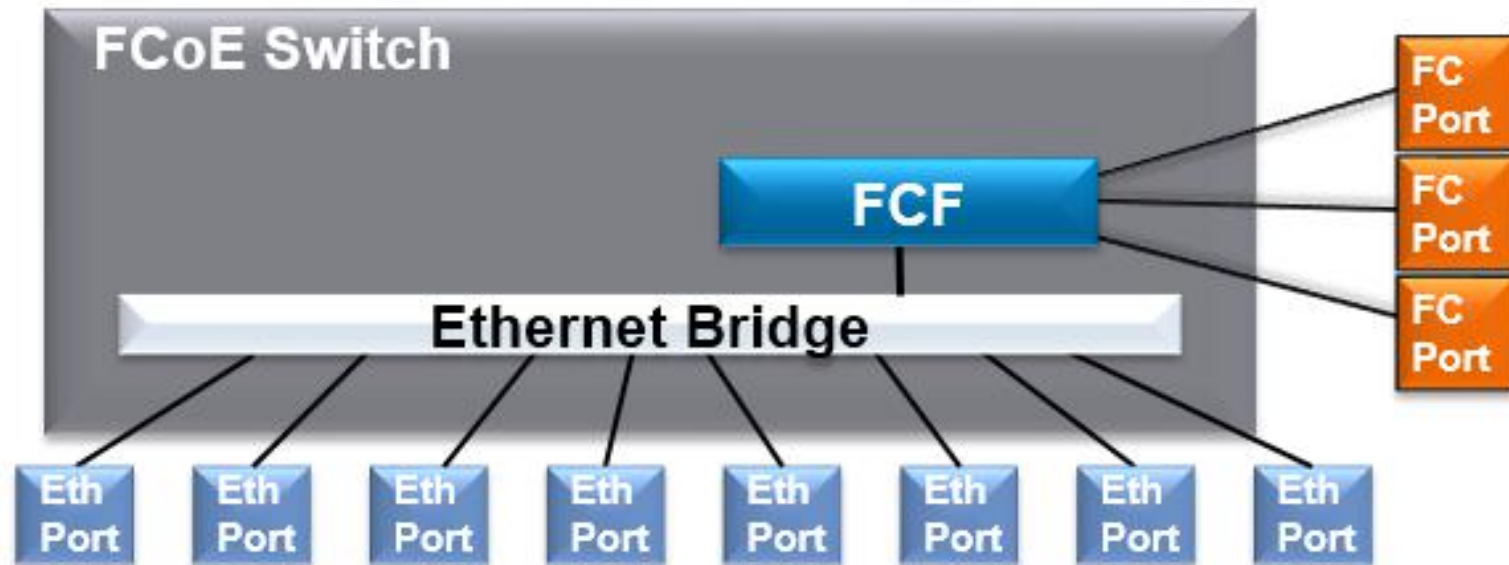


## FCoE and FIP协议

- FCoE有两个不同的协议：FCoE和FIP
- 他们的帧格式不一样
- 都在FC-BB-5内定义

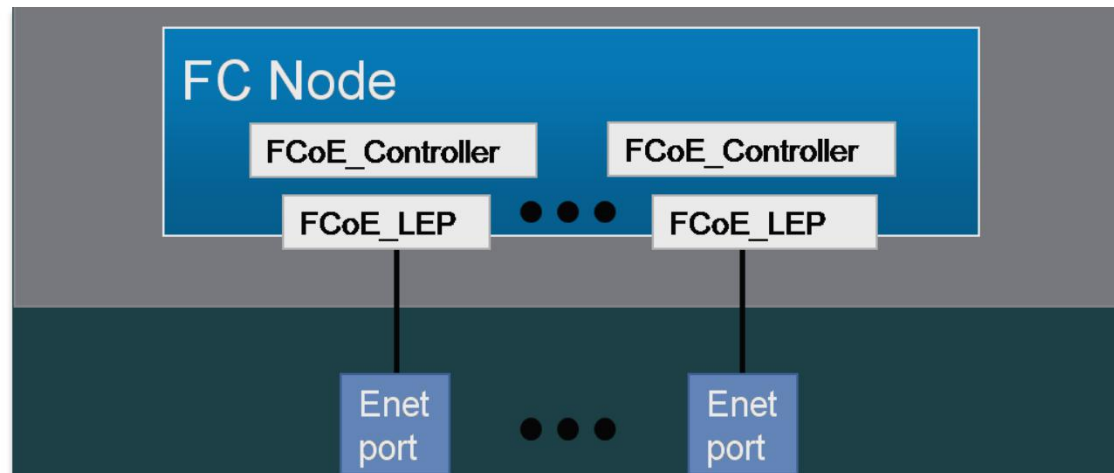
	FCoE	FIP
Plane	Data plane	Control Plane
Purpose	<ul style="list-style-type: none"><li>• 传输大多数FC的帧</li><li>• 传输所有的SCSI流量</li></ul>	<ul style="list-style-type: none"><li>• 发现连在FCoE VLAN中的FCFs</li><li>• 在FCF之间或者FCF和ENode之间建立virtual link</li><li>• 连入Fibre Channel Fabric</li></ul>
EtherType	0x8906	0x8914

- FCF(Fibre Channel Forwarder)是一个在FCoE交换机内的逻辑的FC交换机
  - ENode和FCF执行FLOGI
  - 包括FCF-MAC
  - 获得一个Domain-ID
- 执行FCoE帧的封装和解封装



## FCoE的端口类型

- “Virtual” Port类型
  - VN\_Port
  - VF\_Port
  - VE\_Port
  - VNP
- 在FCF上有一个逻辑端口LEP(link endpoint), 用于处理FC帧的封装和解封装等





## SAN网络和融合网络介绍

周涛

QQ: 53408031

Mobile:  
(86)18611846551

Site: [www.ie-lab.cn](http://www.ie-lab.cn)

YY直播: 58761024

### SAN网络和融合网络介绍

1. 存储简介
2. Fibre Channel 术语
3. Fibre Channel 工作原理
4. 融合网络
5. FCoE的术语
- ➔ 6. FCoE的工作原理
7. FCoE的标准集
8. 传统数据中心网络向融合网络迁移



# FCoE 原理

## Control-Plane

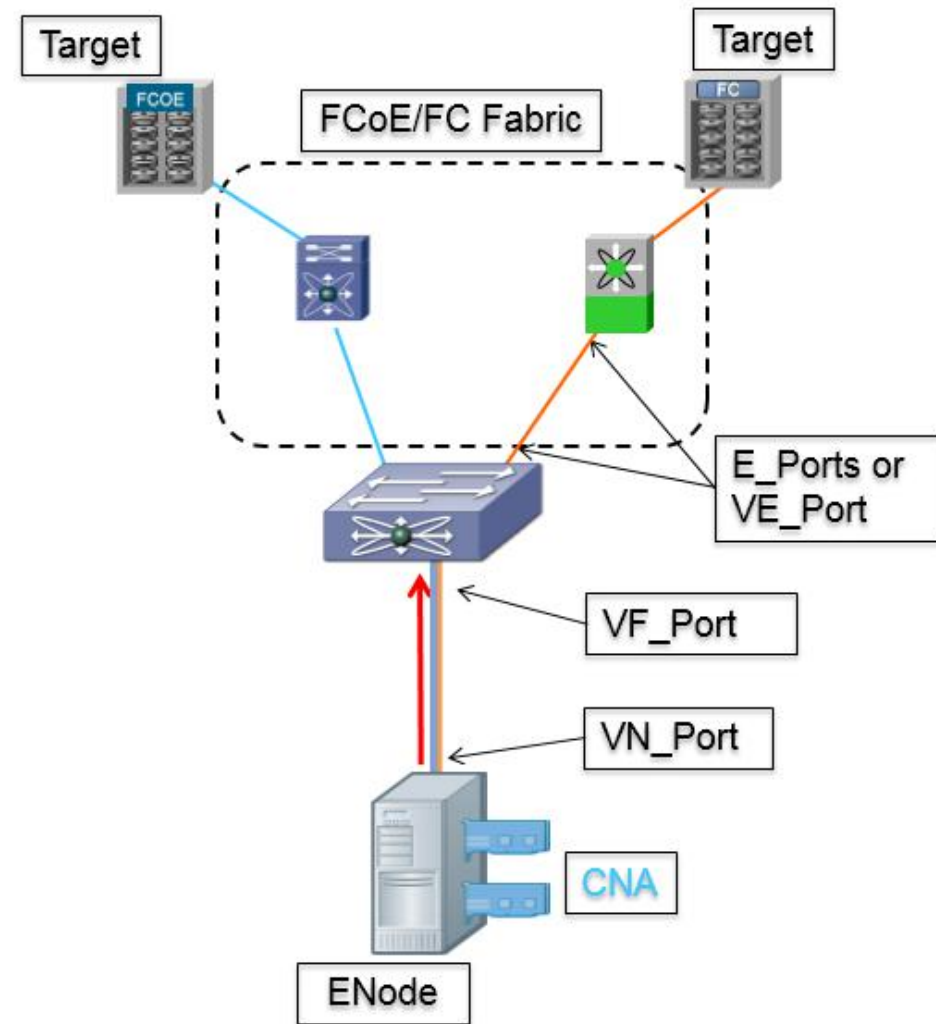


- 周涛
- QQ: 53408031 Mobile: (86)18611846551
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站: [www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024 腾讯课堂: <https://ielab.ke.qq.com/>



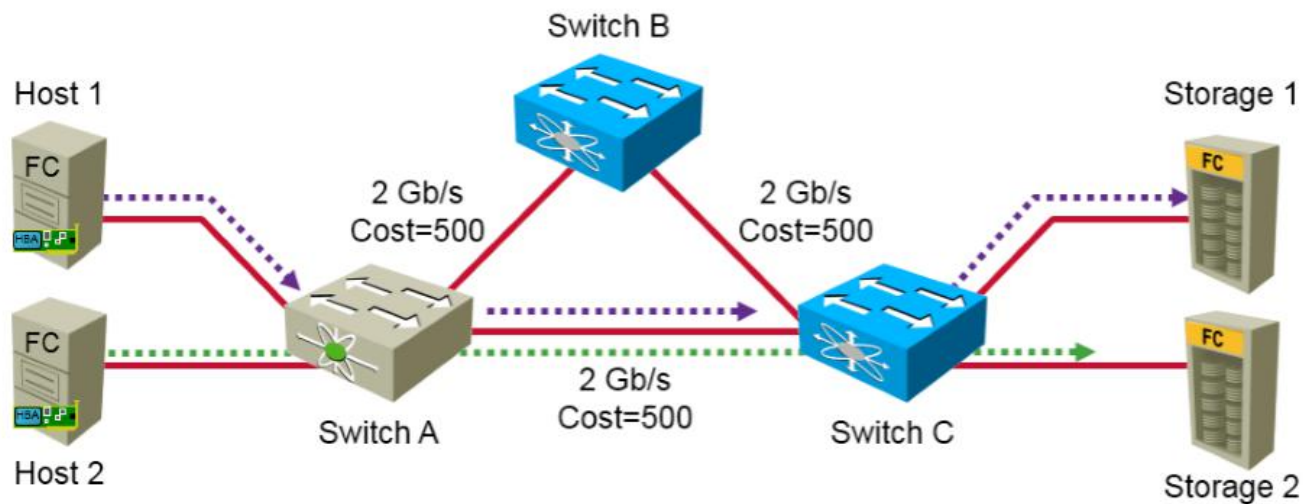
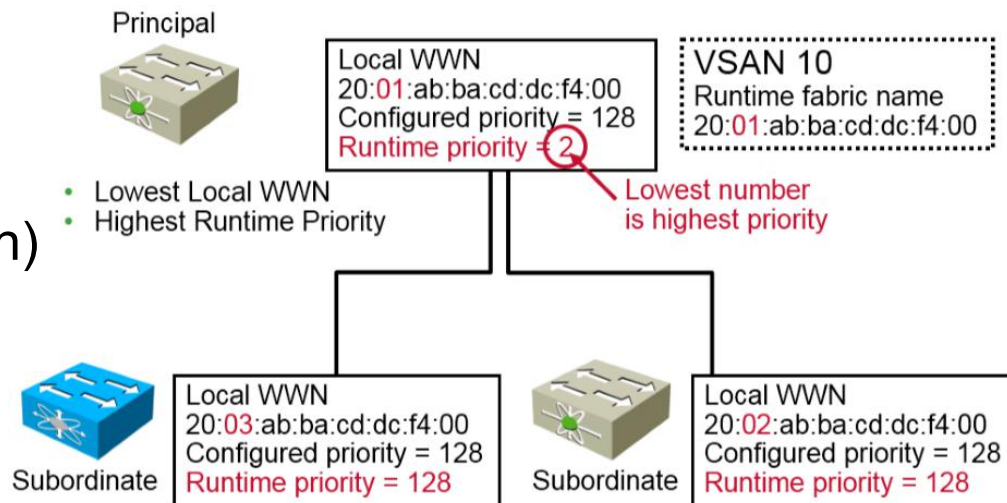
## FCoE网络的拓扑

- Initiator可以连接FCoE-Switch, 然后通过FCoE或FC的Fabric连接到Target



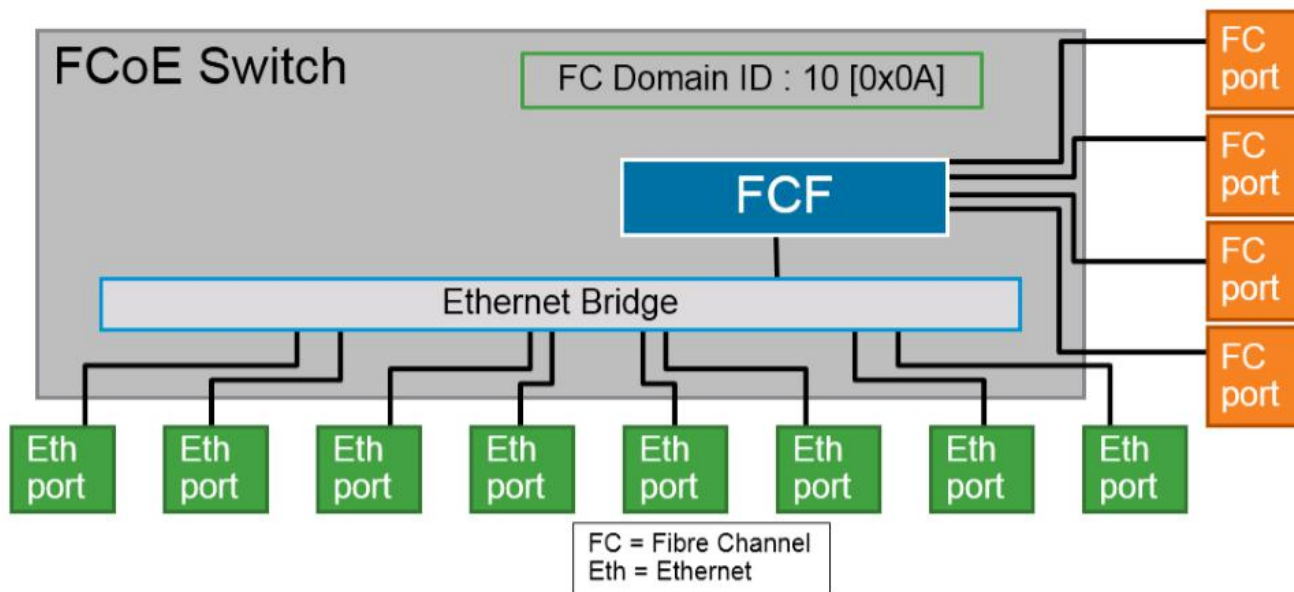
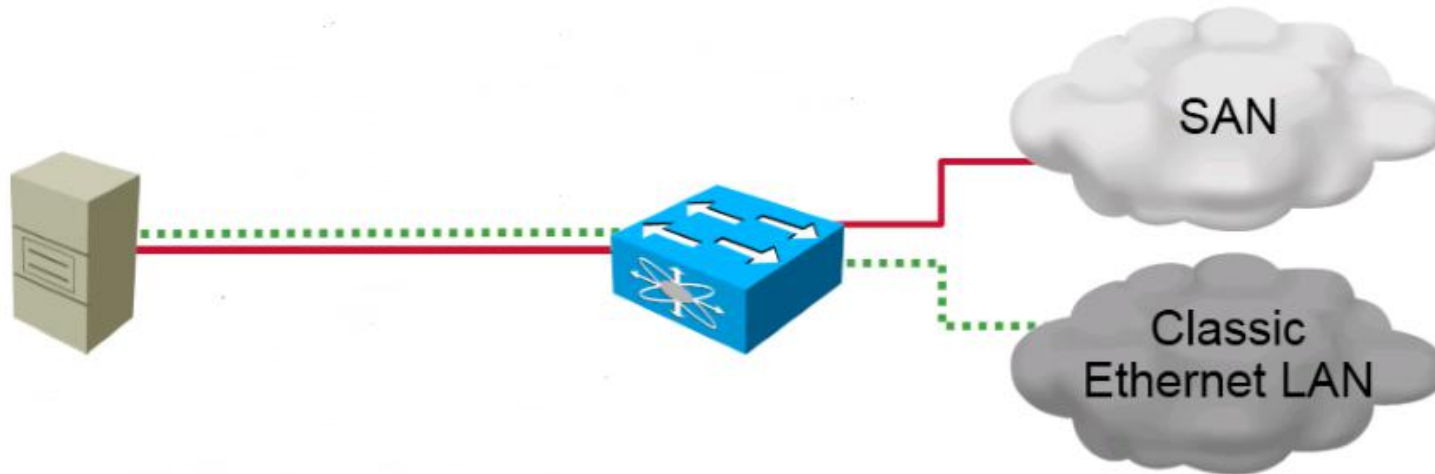
# Fabric网络初始化

- 1.连接初始化
- 2.侦测端口操作模式
- 3.选择主交换机(Principal Switch)
- 4.分配Domain ID
- 5.路径选择(FSPF)



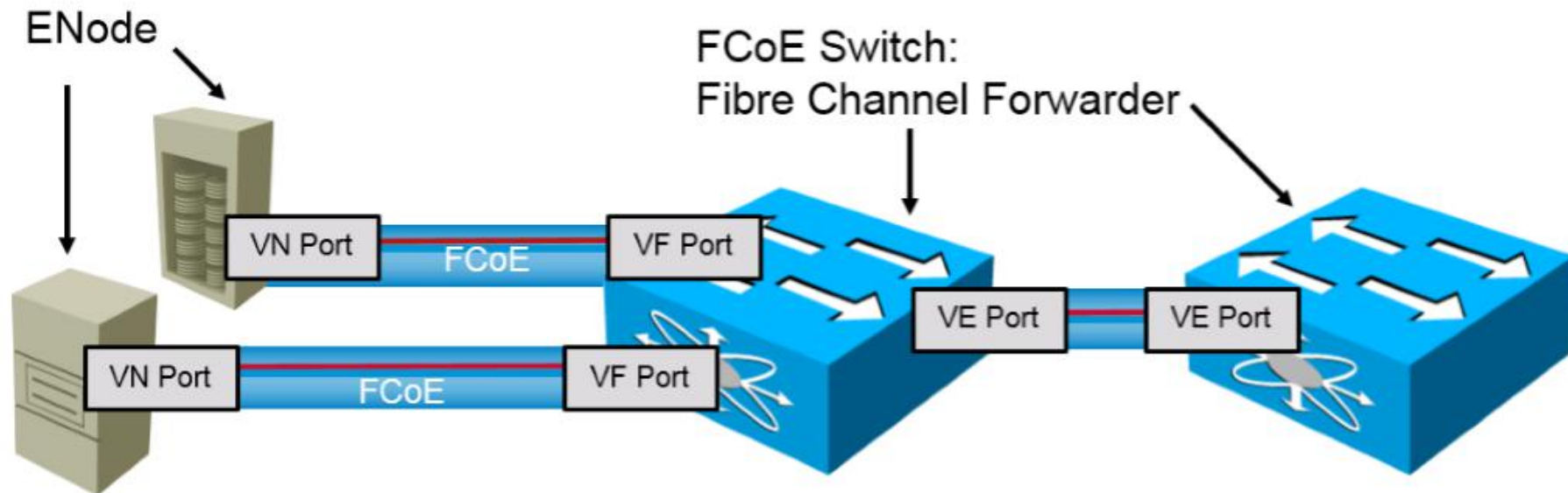


# FCoE设备连入Fabric



## FCoE设备连入Fabric(续)

- FCoE-Switch之间、FCoE-Switch和ENode之间使用FIP协议，将FCoE-Switch和ENode连入Fabric





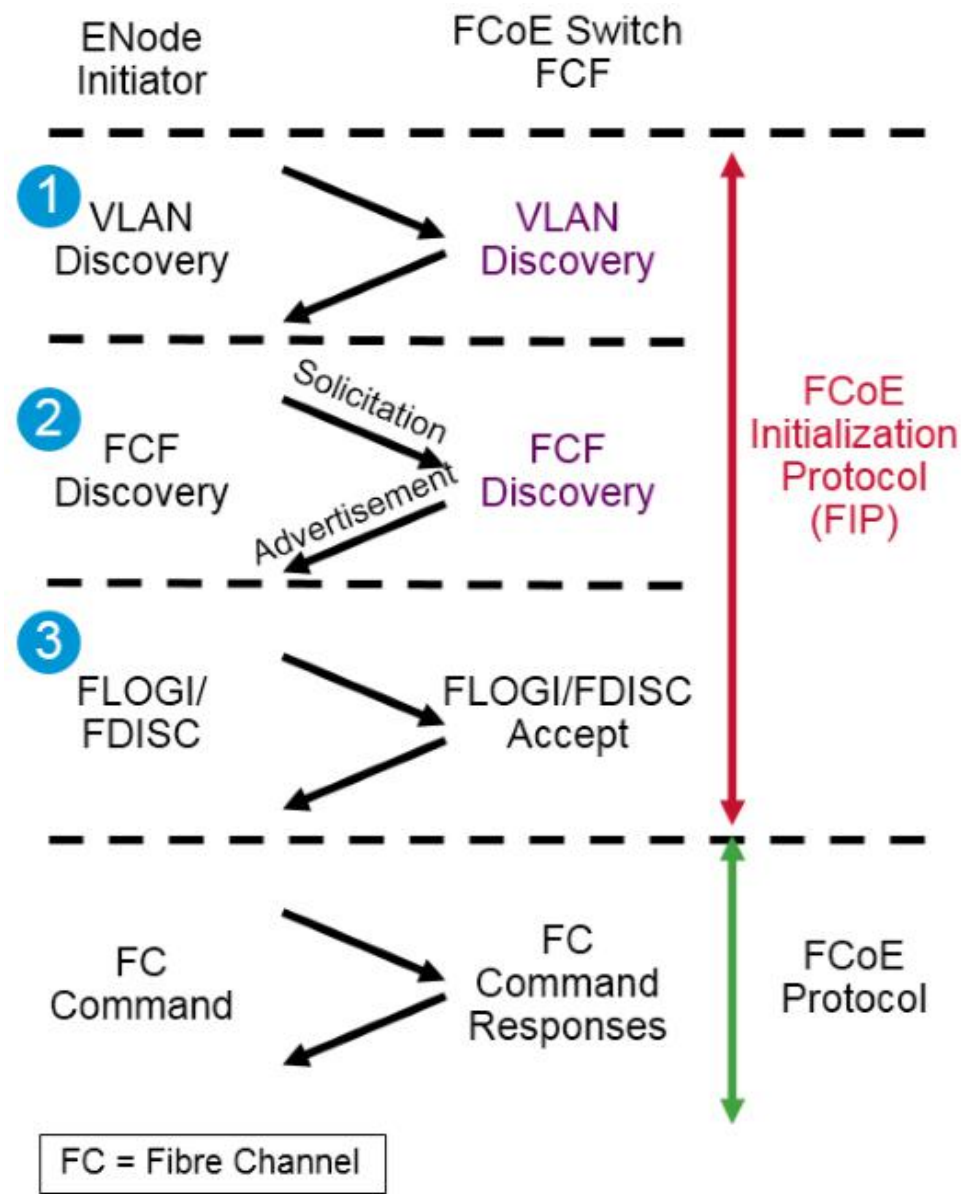
## FCoE and FIP协议

- FCoE有两个不同的协议：FCoE和FIP
- 他们的帧格式不一样
- 都在FC-BB-5内定义

	FCoE	FIP
Plane	Data plane	Control Plane
Purpose	<ul style="list-style-type: none"><li>• 传输大多数FC的帧</li><li>• 传输所有的SCSI流量</li></ul>	<ul style="list-style-type: none"><li>• 发现连在FCoE VLAN中的FCFs</li><li>• 在FCF之间或者FCF和ENode之间建立virtual link</li><li>• 连入Fibre Channel Fabric</li></ul>
EtherType	0x8906	0x8914

## FIP建立Virtual Link

- FCoE的VLAN Discovery
  - FIP发送一个组播帧发往所有FCF的组播MAC地址，发现FCoE VLAN
  - FIP帧使用native VLAN发送
- FCF Discovery
  - FIP在FCoE VLAN中发送组播帧，目的地址为所有FCF的地址，用于发现FCFs
  - FCF通过自己的MAC地址来响应
- Fabric Login
  - FIP发送FLOGI request 去往 FCF\_MAC
  - 在ENode和FCF之间建立virtual link

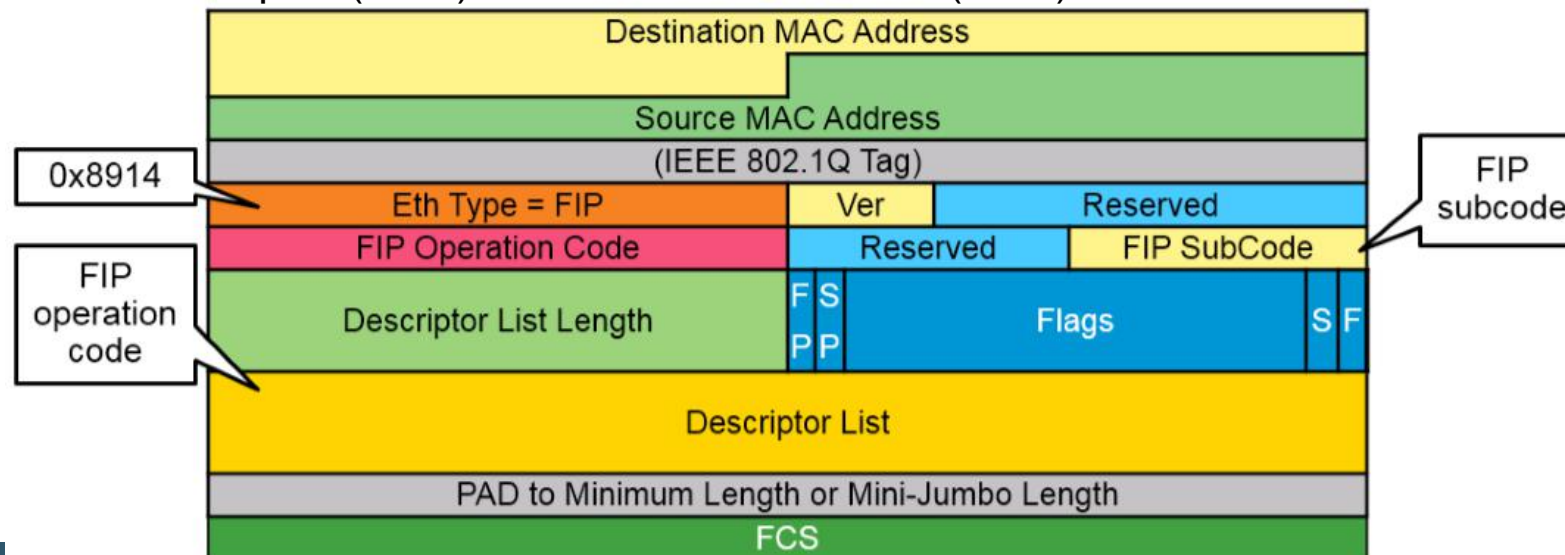




## FIP的帧格式

FIP的operation和subcodes表示消息类型:

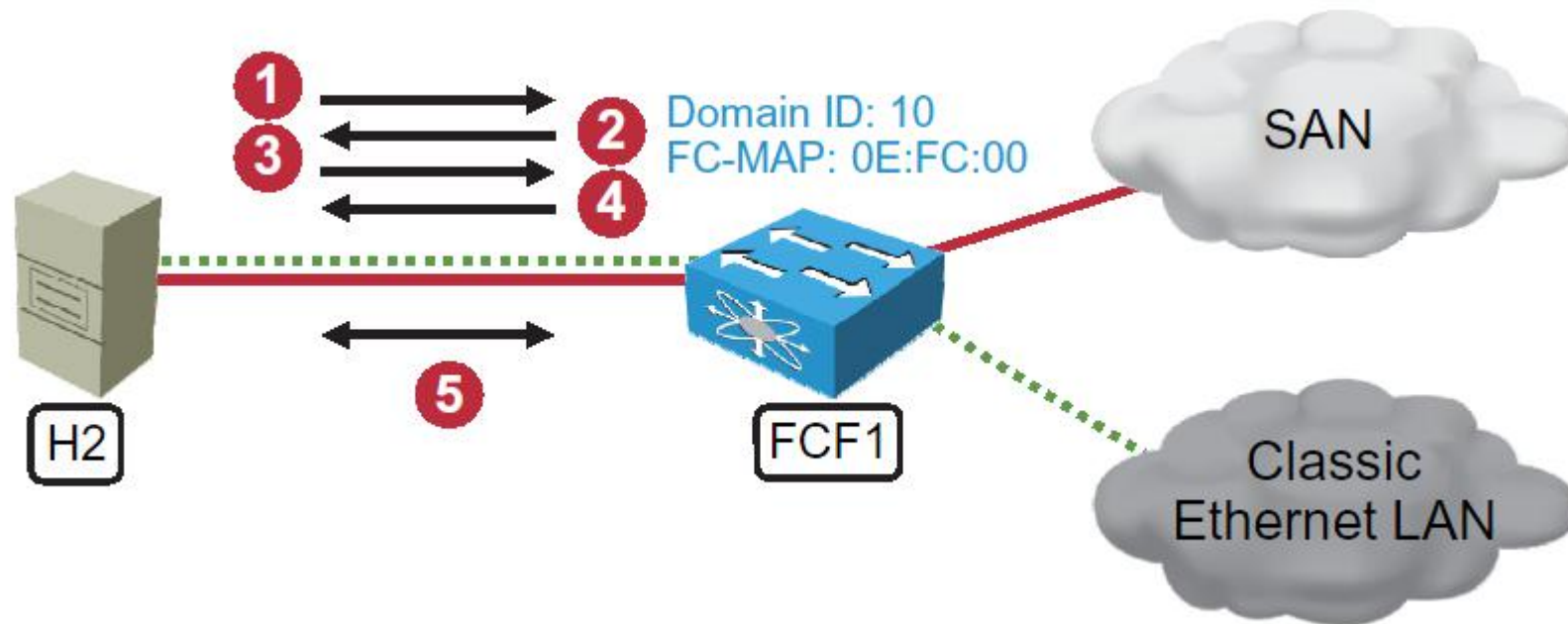
- (0x0001)
  - Discovery Solicitation(0x01) 和 Discovery Advertisement(0x02)
- (0x0002)
  - Virtual Link Instantiation Request(0x01) 和 Virtual Link Instantiation Reply(0x02)
- (0x0003)
  - FIP Keep Alive(0x01) 和 FIP Clear Virtual Links(0x02)
- (0x0004)
  - FIP VLAN Request(0x01) 和 FIP VLAN Notification(0x02)





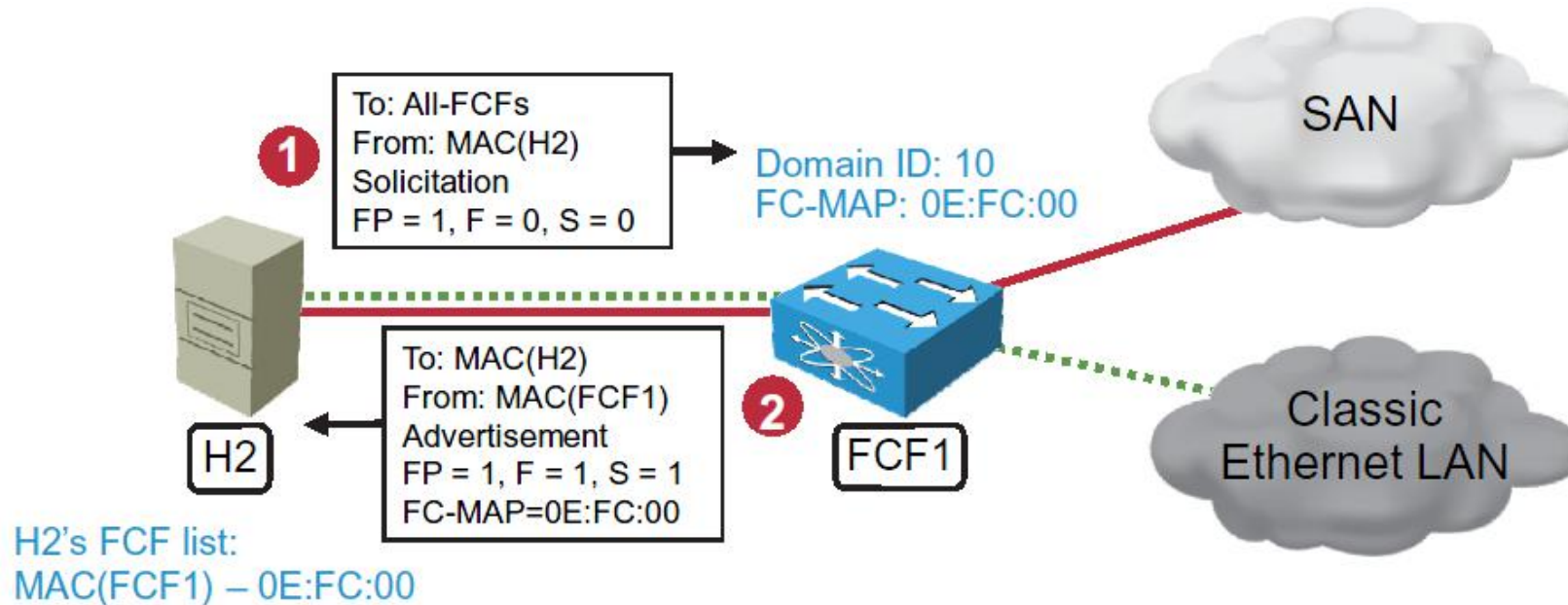
## FCoE设备连入Fabric(续)

- 1.Host Solicitation – 主机请求
- 2.FCoE-Switch提供fabric唯一的FC-MAP
- 3.ENode执行Fabric Login(FLOGI)
- 4.FCF提供FC-ID
- 5.主机使用FPMA做后续的帧传输



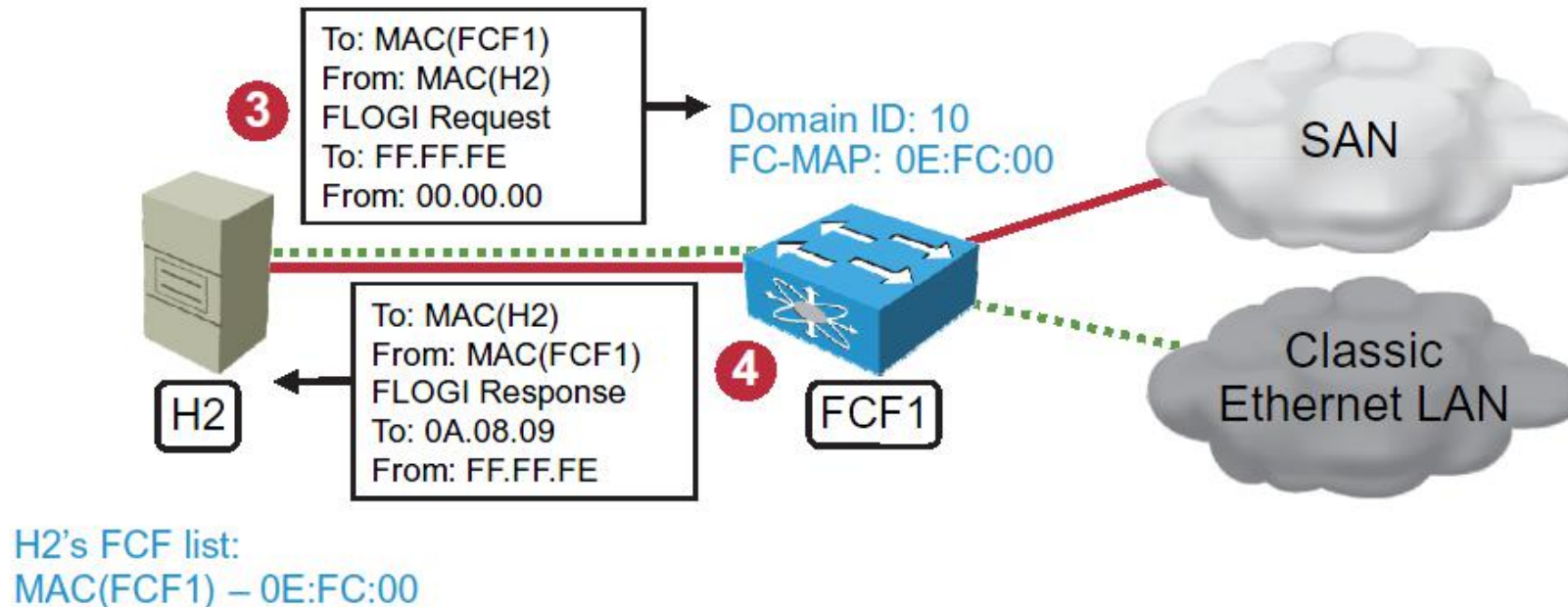
## FCoE设备连入Fabric(续) – Solicitation

- 1. Host Solicitation – 主机请求
  - 表示CNA卡支持MAC类型(FPMA)
  - EtherType=0x8914
- 2. FCoE-Switch提供fabric唯一的FC-MAP
  - 是FPMA的前3个字节
  - FP=1: FPMA的能力; F=1: 这个帧是FCF产生的; S=1响应Solicitation消息
  - EtherType=0x8914



## FCoE设备连入Fabric(续) – FLOGI

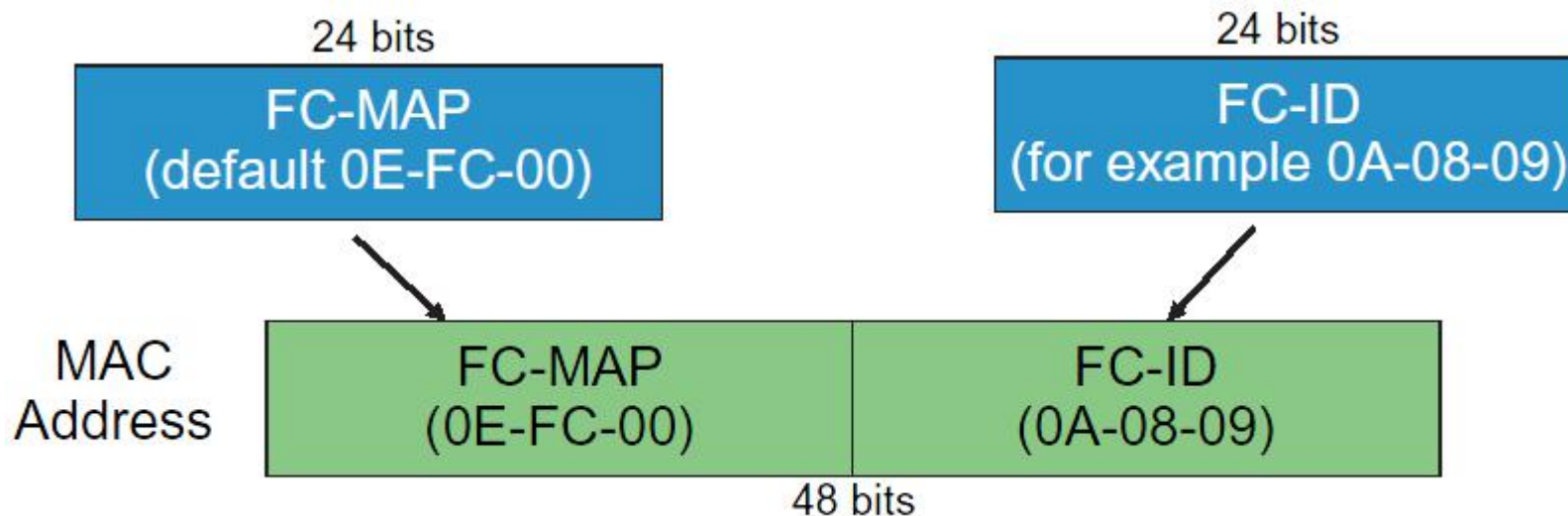
- 3. ENode执行Fabric Login(FLOGI)
  - 现在ENode使用的还是bia-MAC
  - EtherType=0x8914
- 4. FCF提供FC-ID
  - FCF使用bia-MAC来响应
  - EtherType=0x8914





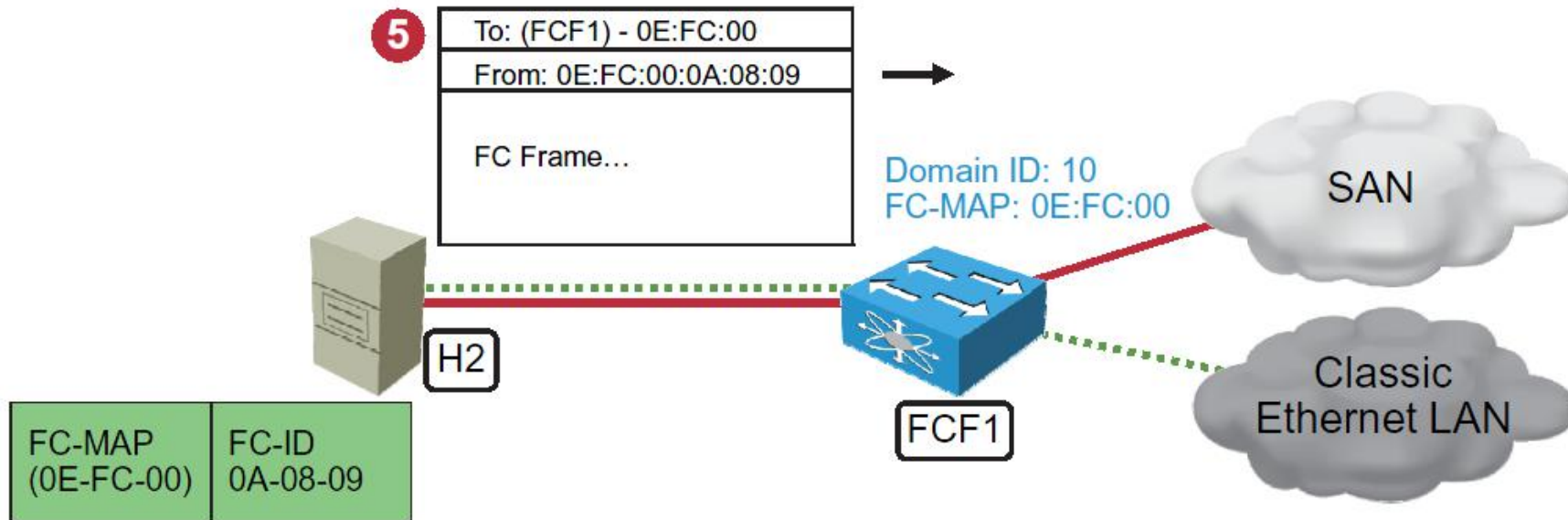
## Fabric Provided MAC Address(FPMA)

- 与Fibre Channel模型保持统一
  - FCoE MAC address prefix (FC-MAP)
  - Fibre Channel ID (FCID)
- 不需要MAC address mapping表
- 多个FC-MAPs支持
  - 每个物理SAN一个
  - 可以与NPIV合用
- 要在FCoE设备上支持



## 后续的FCoE帧传递

- 5. 主机使用FPMA做后续的帧传输
  - FPMA的FCID部分通过FLOGI获得
  - FCID和之前FCF提供的FC-MAP组成FPMA
  - EtherType=0x8906



# FCoE 原理

## Data-Plane



- 周涛
- QQ: 53408031 Mobile: (86)18611846551
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站: [www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024 腾讯课堂: <https://ielab.ke.qq.com/>



## FCoE and FIP协议

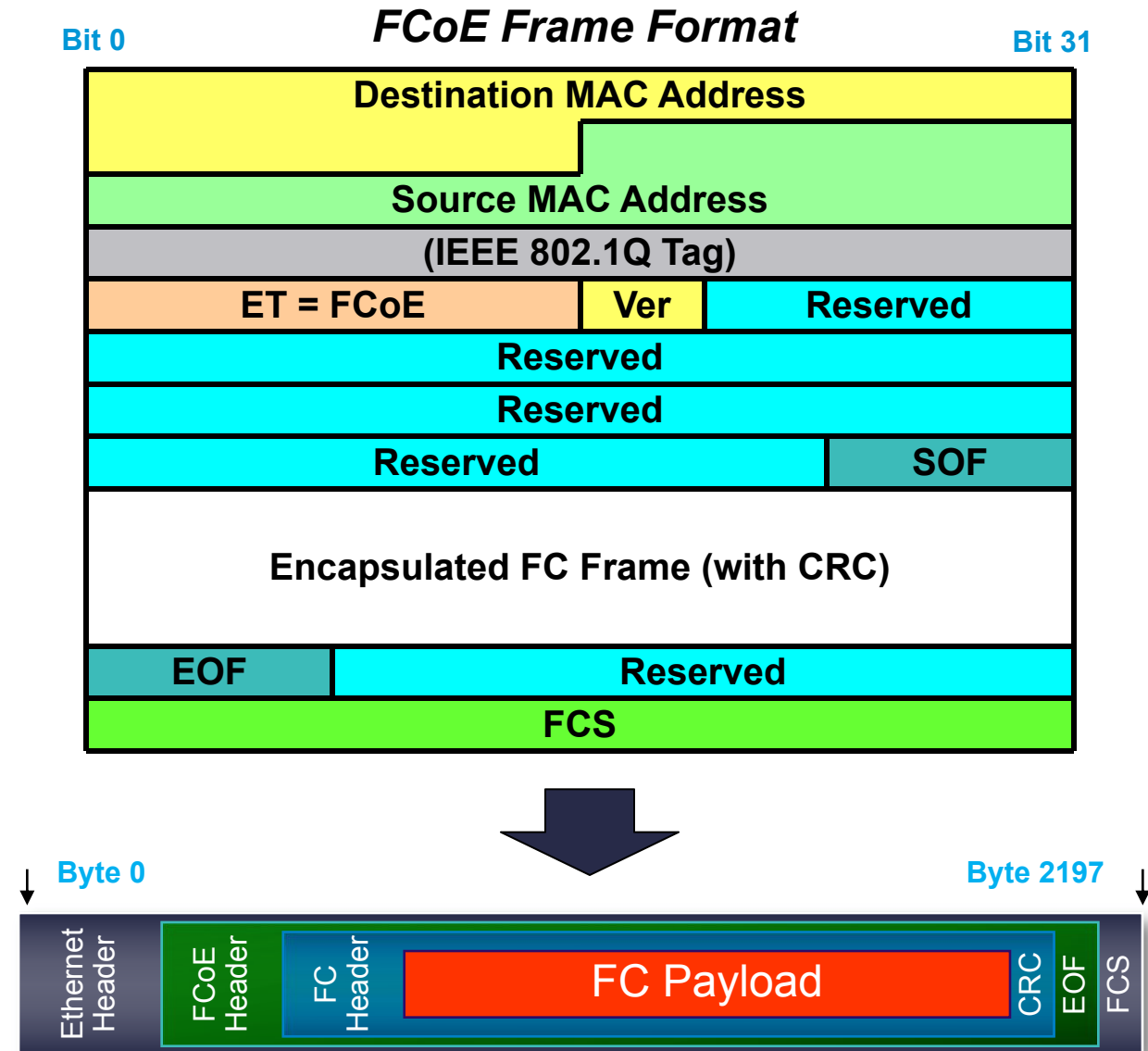
- FCoE有两个不同的协议：FCoE和FIP
- 他们的帧格式不一样
- 都在FC-BB-5内定义

	FCoE	FIP
Plane	Data plane	Control Plane
Purpose	<ul style="list-style-type: none"><li>• 传输大多数FC的帧</li><li>• 传输所有的SCSI流量</li></ul>	<ul style="list-style-type: none"><li>• 发现连在FCoE VLAN中的FCFs</li><li>• 在FCF之间或者FCF和ENode之间建立virtual link</li><li>• 连入Fibre Channel Fabric</li></ul>
EtherType	0x8906	0x8914



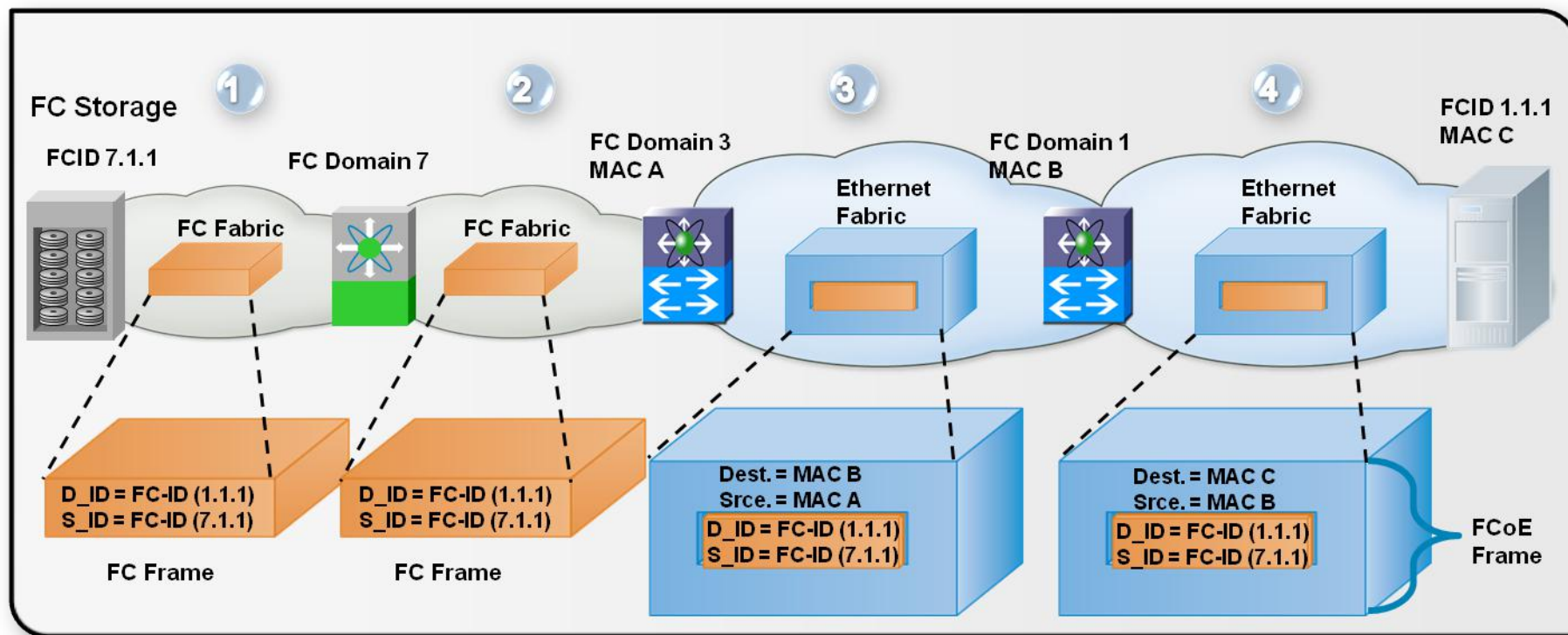


# FCoE 帧结构



## 后续的FCoE帧传递

- FCoE帧:
- MAC addresses (hop-by-hop)
- FC addresses (end-to-end)





## SAN网络和融合网络介绍

周涛

QQ: 53408031

Mobile:  
(86)18611846551

Site: [www.ie-lab.cn](http://www.ie-lab.cn)

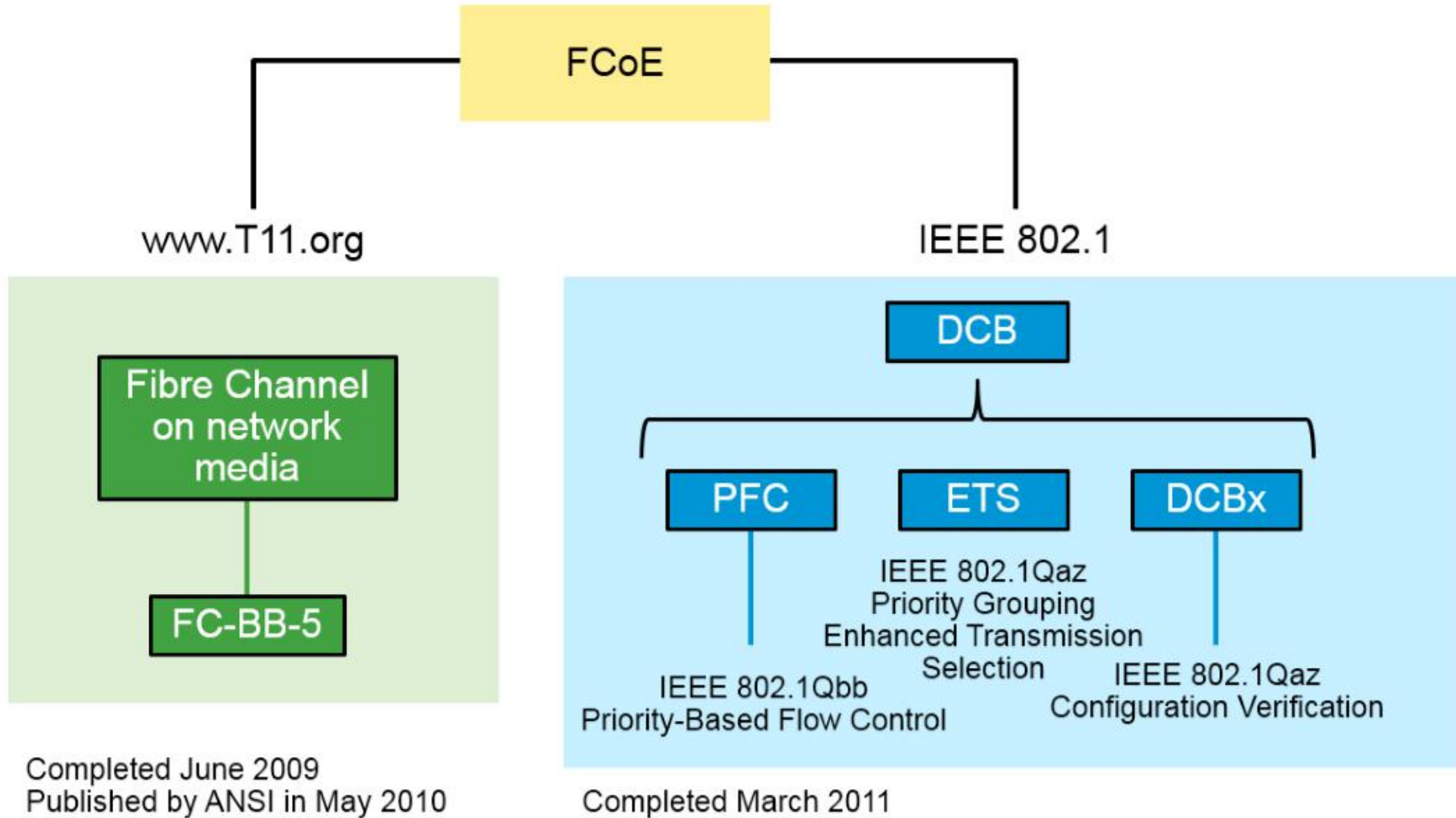
YY直播: 58761024

### SAN网络和融合网络介绍

1. 存储简介
2. Fibre Channel 术语
3. Fibre Channel 工作原理
4. 融合网络
5. FCoE的术语
6. FCoE的工作原理
- ➔ 7. FCoE的标准集
8. 传统数据中心网络向融合网络迁移



# FCoE的标准





## DCB概述

- IEEE对classical Ethernet的增强
- 以太网增强:
  - Priority Group: 对于流量的类别分配虚拟链路和资源
  - 先对流量进行分类, 然后再进行Priority flow control
  - 端到端的拥塞管理和通知
  - Layer 2 multipathing
- 优点:
  - 消除瞬时和永久拥塞
  - Lossless Fabric: no-drop storage links
  - 对于HPC clusters可以确定它们的延迟
  - 为了减少花费和复杂性可以使用融合的Ethernet Fabric



## DCB标准

Technology	IEEE Standard	Description
Priority Flow Control (PFC)	802.1Qbb	对于相应COS值的流量提供无丢弃的传输
Enhanced Transmission Selection (ETS)	802.1Qaz	带宽管理和优先级的选择
Quantized Congestion Notification (QCN)	802.1Qau	拥塞感知和避免(optional)
Data Center Bridging Exchange (DCBX)	802.1Qaz	用于在DCB设备间交换参数通过LLDP提供

## Priority Flow Control

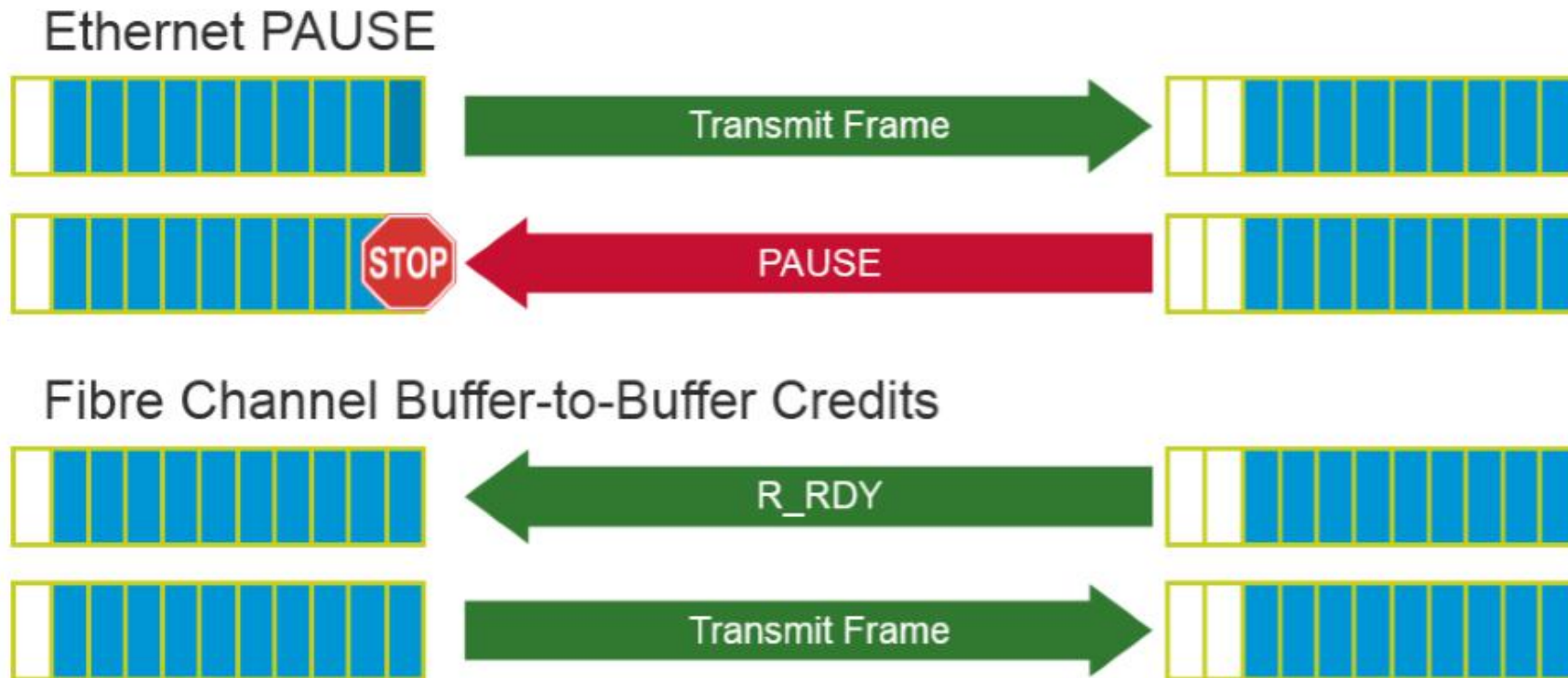


- 周涛
- QQ: 53408031 Mobile: (86)18611846551
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站: [www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024 腾讯课堂: <https://ielab.ke.qq.com/>



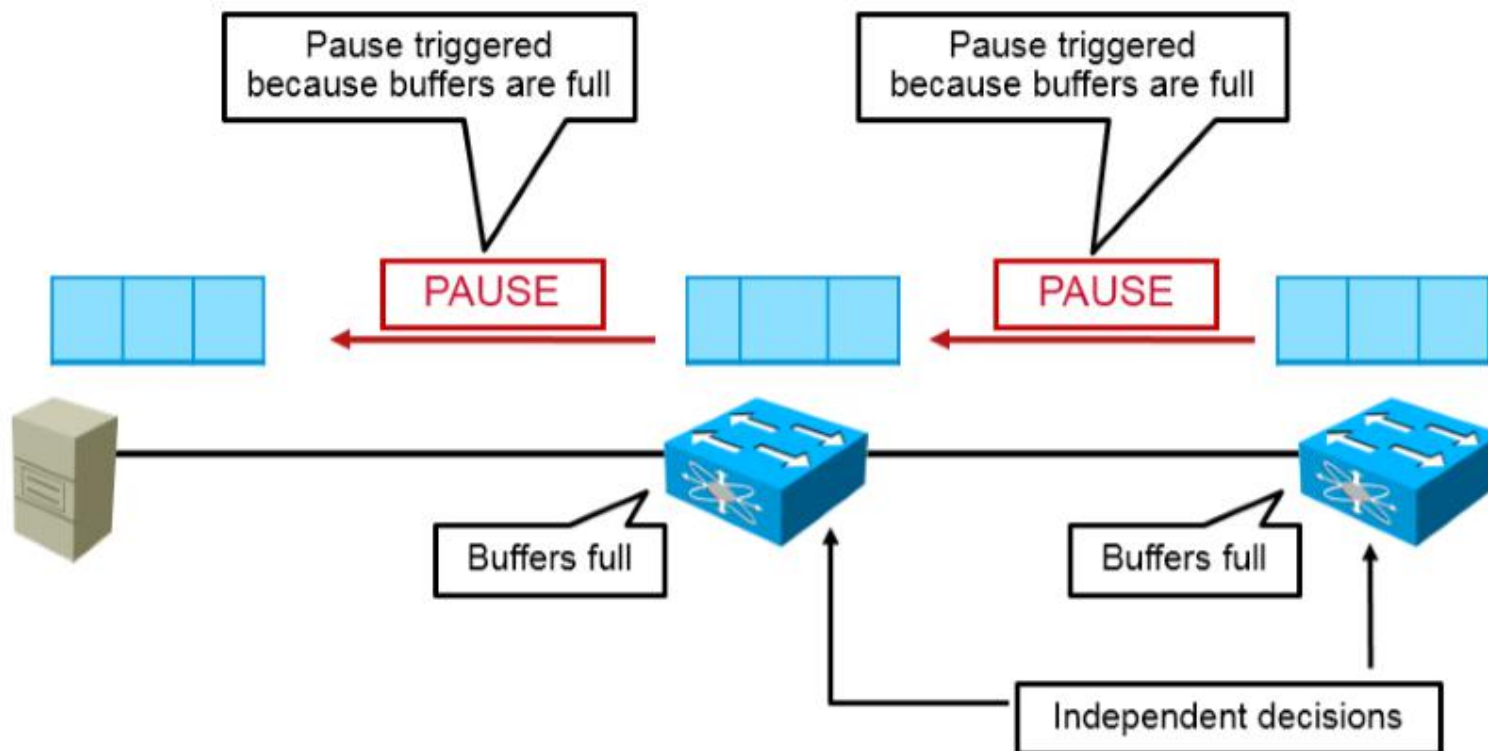
## Link-Level Flow Control

- IEEE 802.3x链路级别流量控制允许发生拥塞的接收端给发送端发送Pause帧，暂停数据传输
- 需要在接口上显式配置，应用于接口上的所有流量



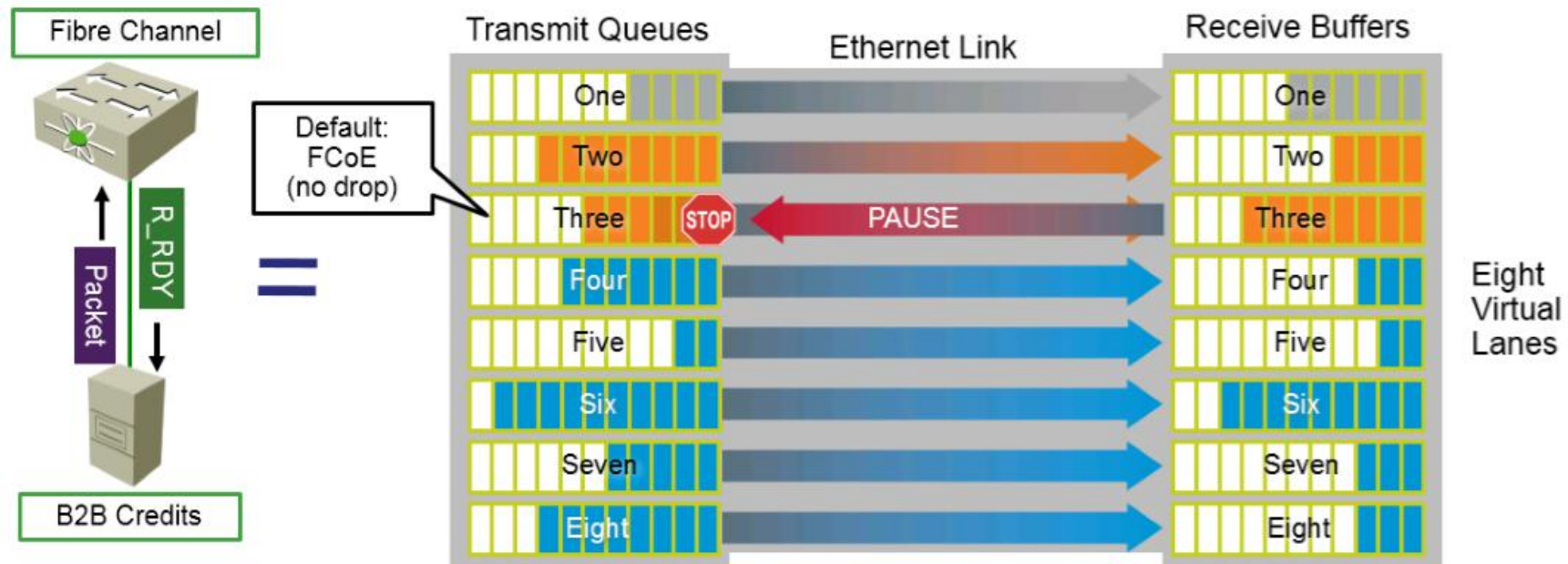
## Pause传输

- PAUSE是hop-by-hop的机制
- 每一跳接收pause帧和发送pause帧都是独立的
- 由可用的buffer space来决定



# Priority Flow Control

- 802.1Qbb, 专门为FCoE设计
- 基于802.1p COS使用PAUSE使能lossless Ethernet
- 当链路拥塞, COS值分配了"no-drop"将会停止
- 其他流量继续发送依靠上层的协议来进行重传
- 这个技术不一定专门用于FCoE



## Enhanced Transmission Selection

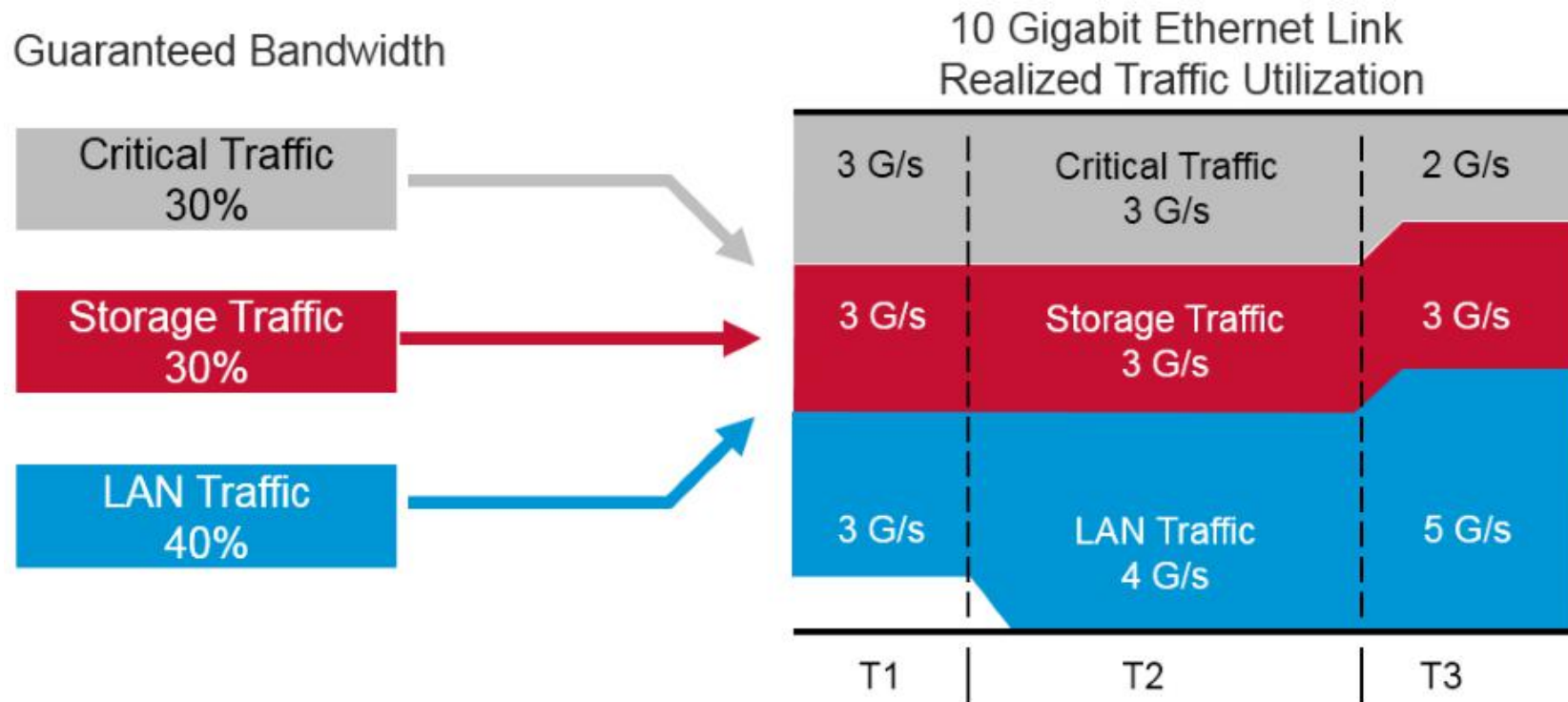


- 周涛
- QQ: 53408031 Mobile: (86)18611846551
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站: [www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024 腾讯课堂: <https://ielab.ke.qq.com/>



## Enhanced Transmission Selection

- 802.1Qaz里定义
- 允许在流量之间智能共享带宽
- 对流量保证一个最小的带宽，如果没有使用，其他流量可以使用
- 可以存在strict priority流量类



## DCBX Protocol

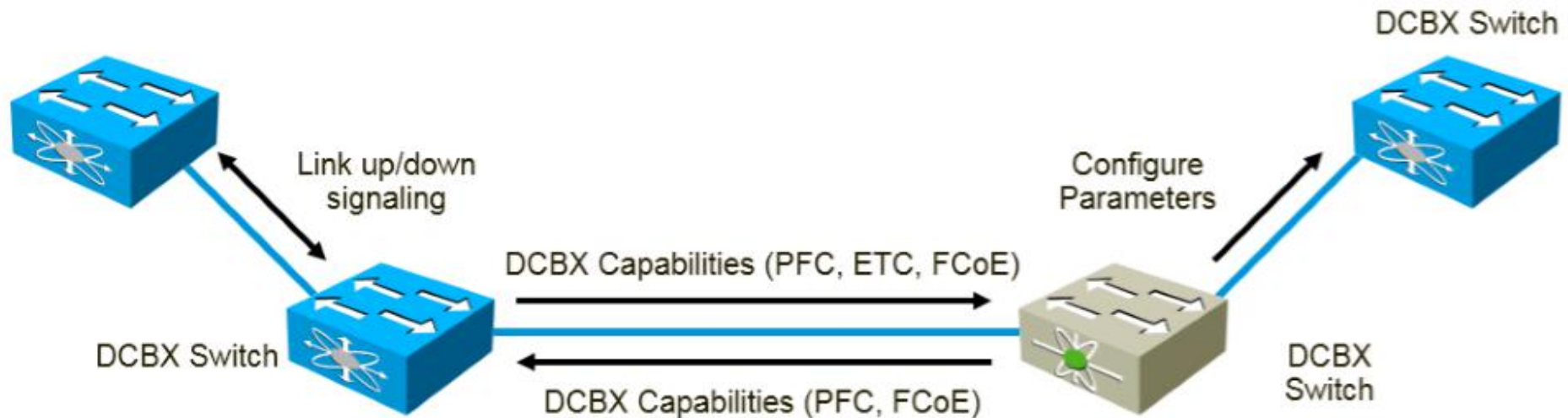


- 周涛
- QQ: 53408031 Mobile: (86)18611846551
- E-MAIL: [53408031@qq.com](mailto:53408031@qq.com) 网站: [www.ie-lab.cn](http://www.ie-lab.cn)
- YY房间: 58761024 腾讯课堂: <https://ielab.ke.qq.com/>



## DCBX Protocol

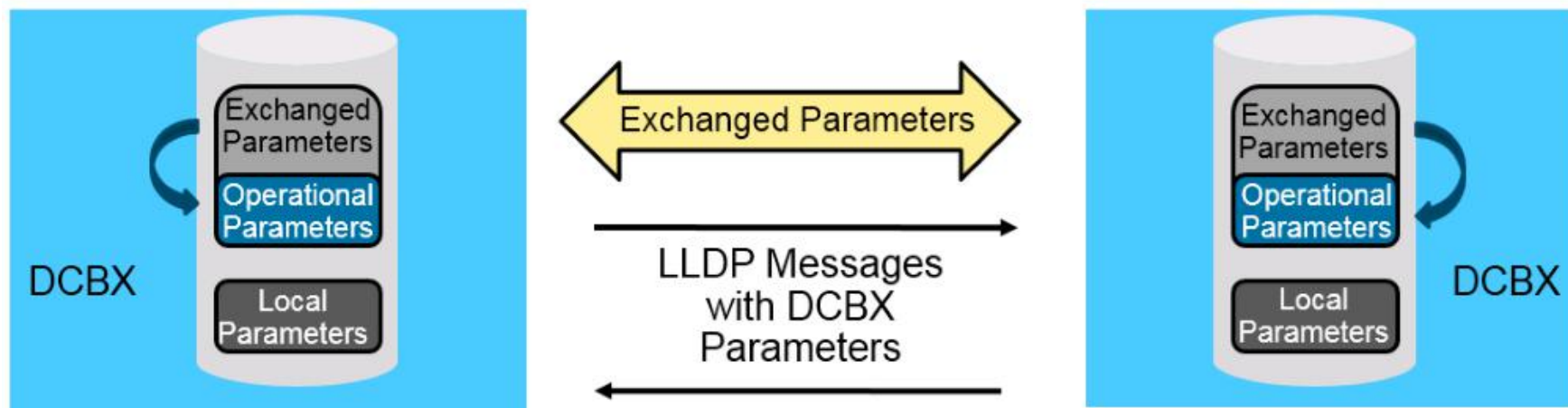
- 802.1Qaz里定义
- 用于P2P链路发现
- 协商能力 – PFC、ETS, FCoE的能力
- 允许从一个节点发送参数到其他设备
- 逻辑链路up/down信令
  - Ethernet接口
  - Fibre Channel接口





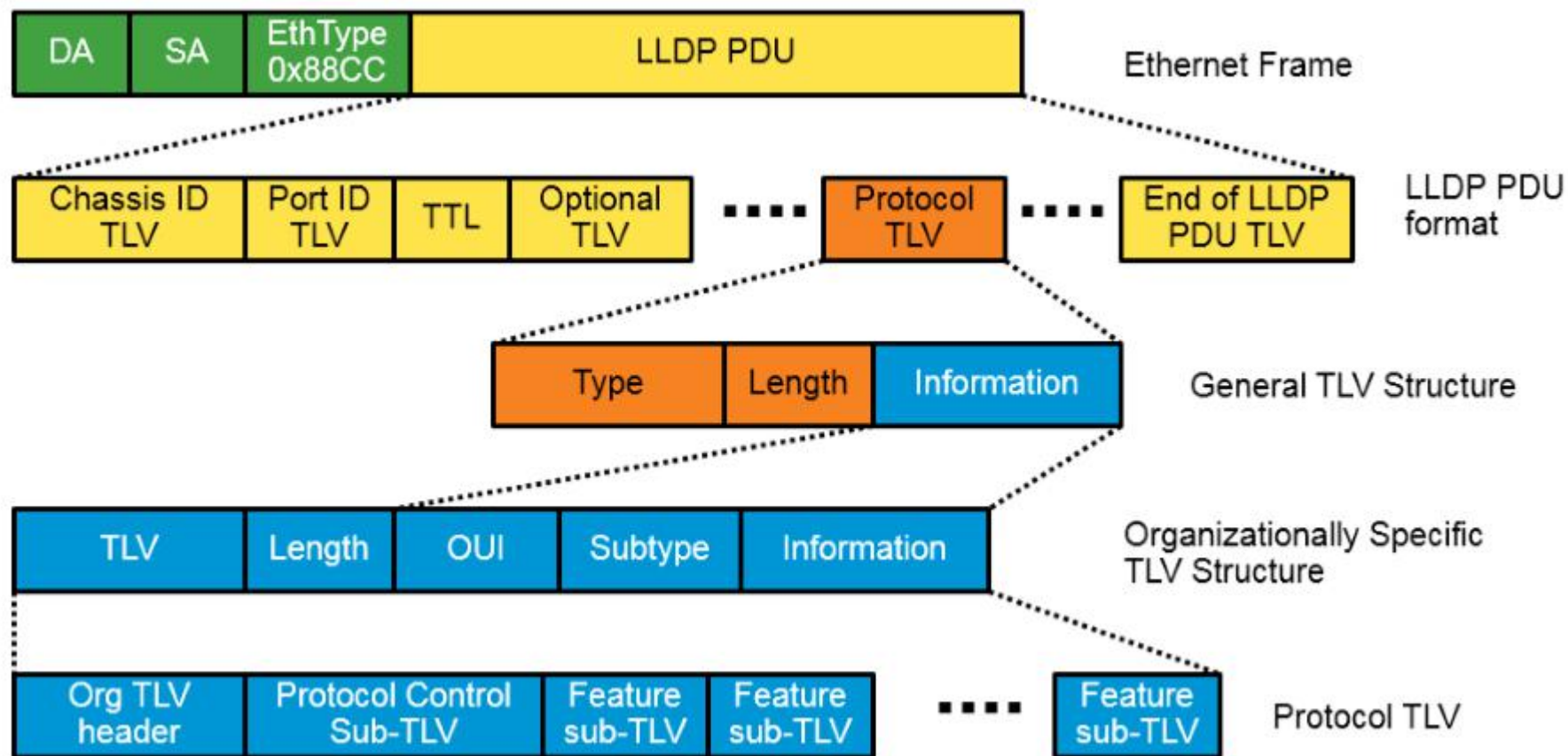
## DCBX 协商

- 发现邻居DCB的能力
- 检查是否有配置不匹配
- Peer Configuration:
  - Administered Parameters – 提供给对端设备
  - Operational Parameters – information purposes only
  - Local Parameters – 不交换
- DCBX协商失败会导致：
  - Per-priority-pause不能使能
  - vFC起不来



## DCBX 报文

- 封装在LLDP报文中传输
- 通过TLV来扩展





## SAN网络和融合网络介绍

周涛

QQ: 53408031

Mobile:

(86)18611846551

Site: [www.ie-lab.cn](http://www.ie-lab.cn)

YY直播: 58761024

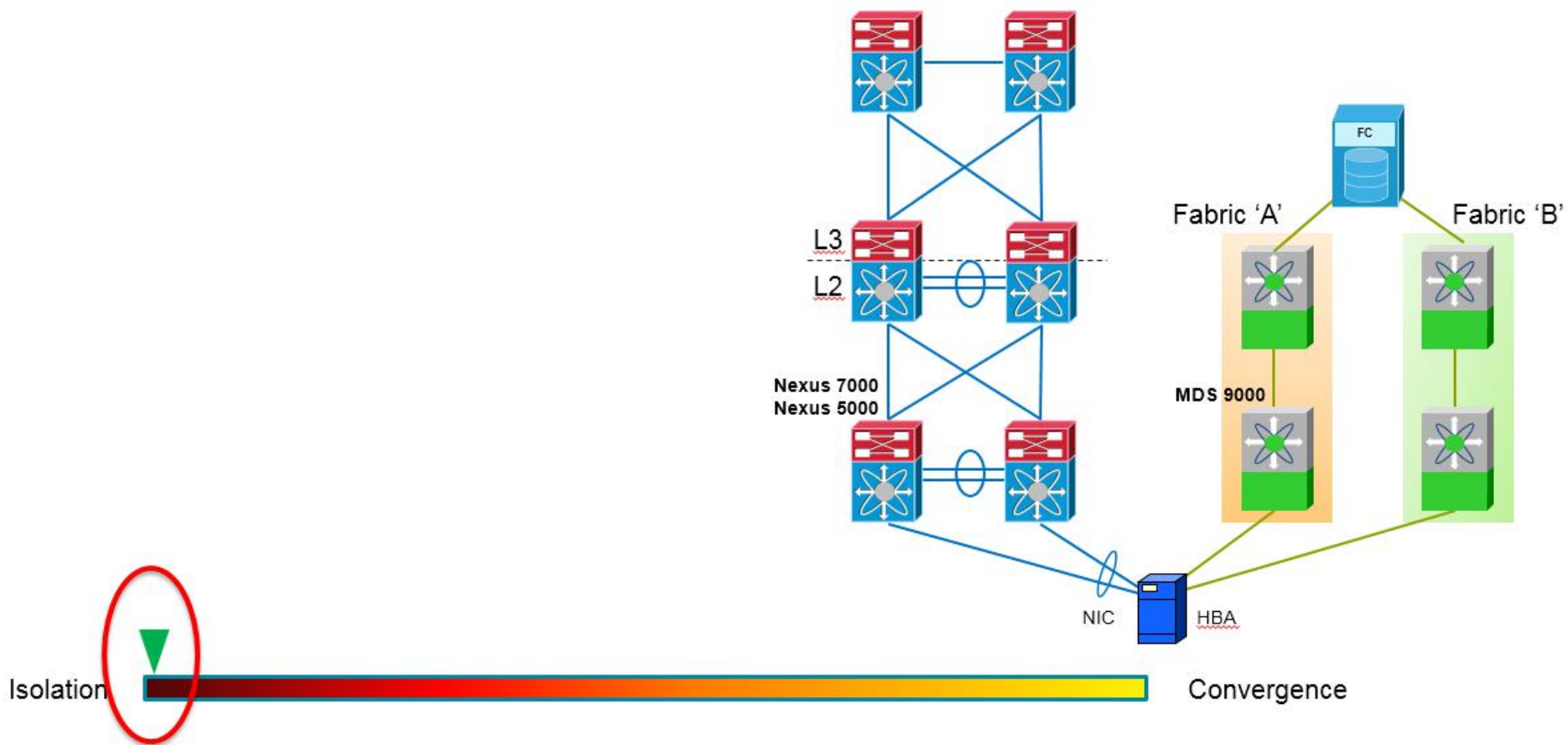
## SAN网络和融合网络介绍

1. 存储简介
2. Fibre Channel 术语
3. Fibre Channel 工作原理
4. 融合网络
5. FCoE的术语
6. FCoE的工作原理
7. FCoE的标准集
- ➔ 8. 传统数据中心网络向融合网络迁移



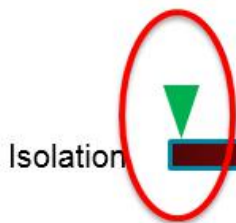
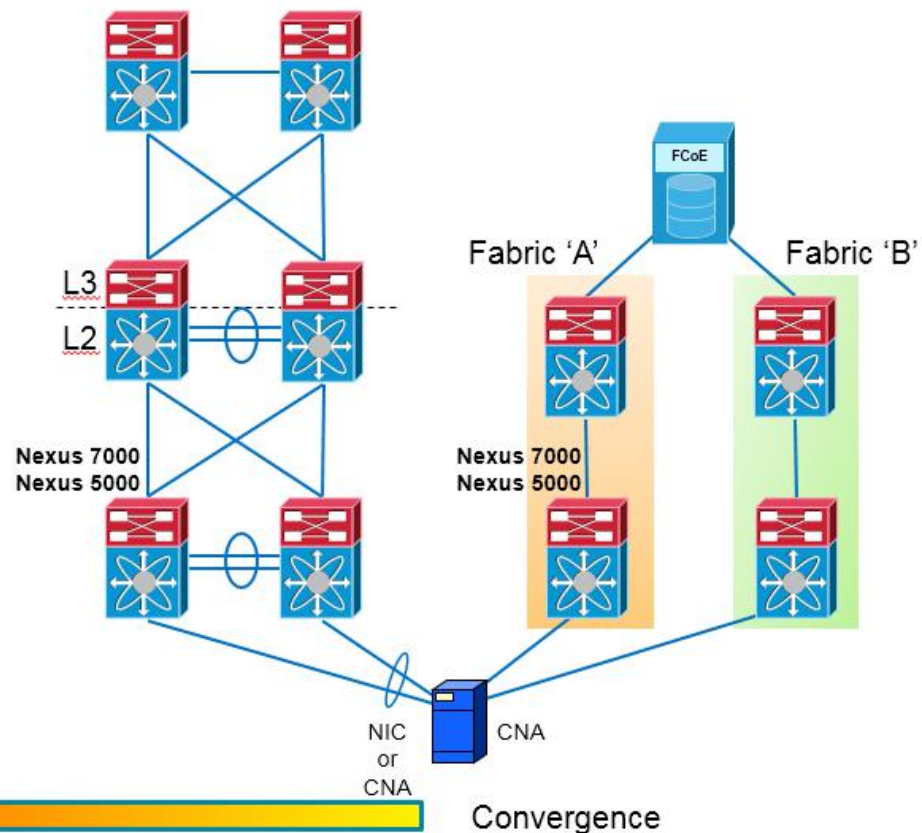
## 传统的数据中心设计

- 物理和逻辑上将LAN/SAN流量分离
- 额外的物理和逻辑上的SAN Fabric



## 数据中心设计 - E-SAN

- 和现存网络的拓扑一样，只不过换成了Nexus Unified Fabric Ethernet交换机
- 物理和逻辑上分离LAN/SAN的流量
- 添加额外的逻辑上和物理上的冗余的SAN的Fabric
- Ethernet SAN Fabric承载FC/FCoE和基于IP的storage流量(iSCSI, NAS,...)
- 共同的组成：Ethernet的能力和花费

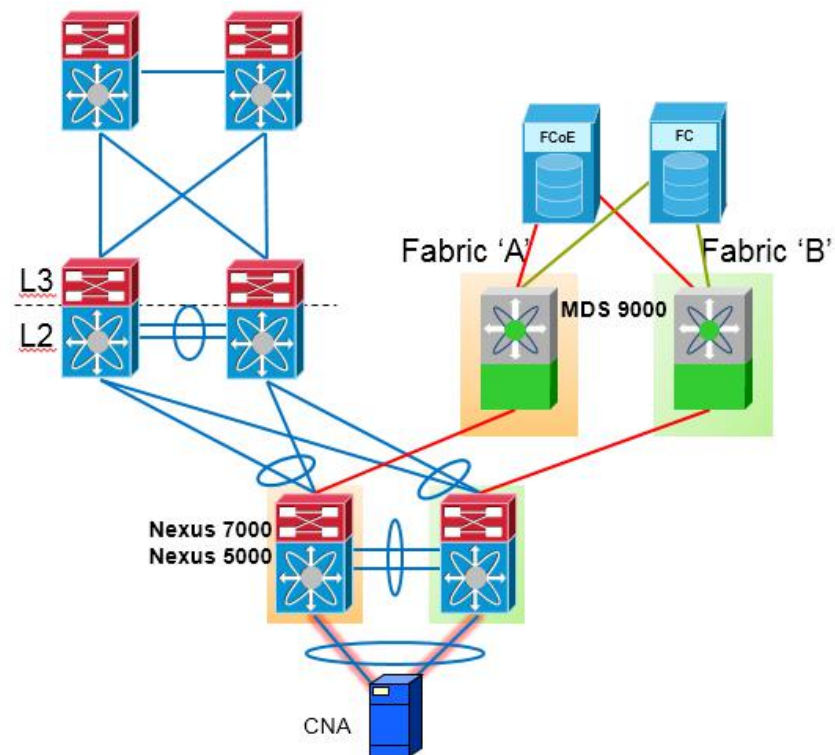


Isolation

Convergence

## 数据中心设计 - 通过vPC实现融合的Access

- 在接入层实现物理设备共享，逻辑上LAN和SAN分离
- 在汇聚层物理和逻辑上分离LAN/SAN的流量
- 添加额外的逻辑上和物理上的冗余的SAN的Fabric
- 可以使用Storage VDC实现管理平面和操作平面的分离



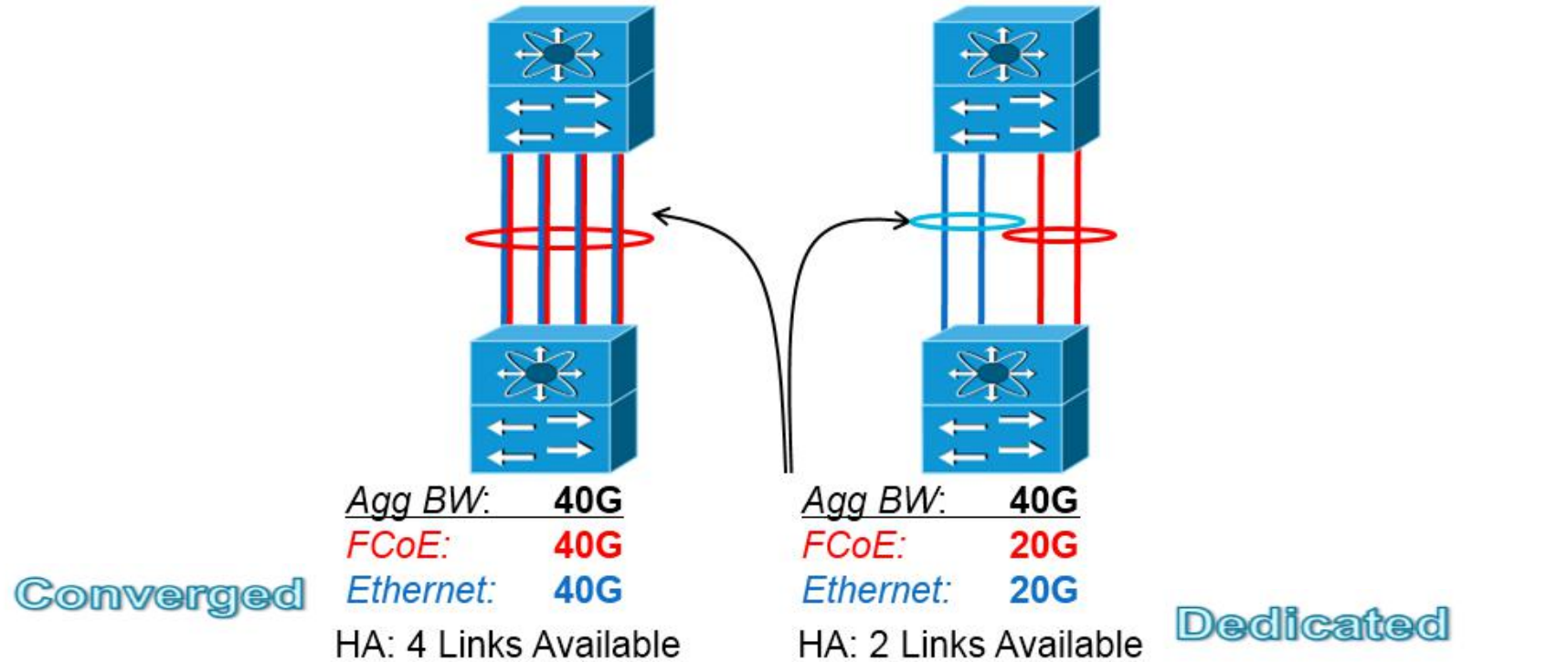
Isolation



Convergence



# Dedicated vs. Converged ISLs



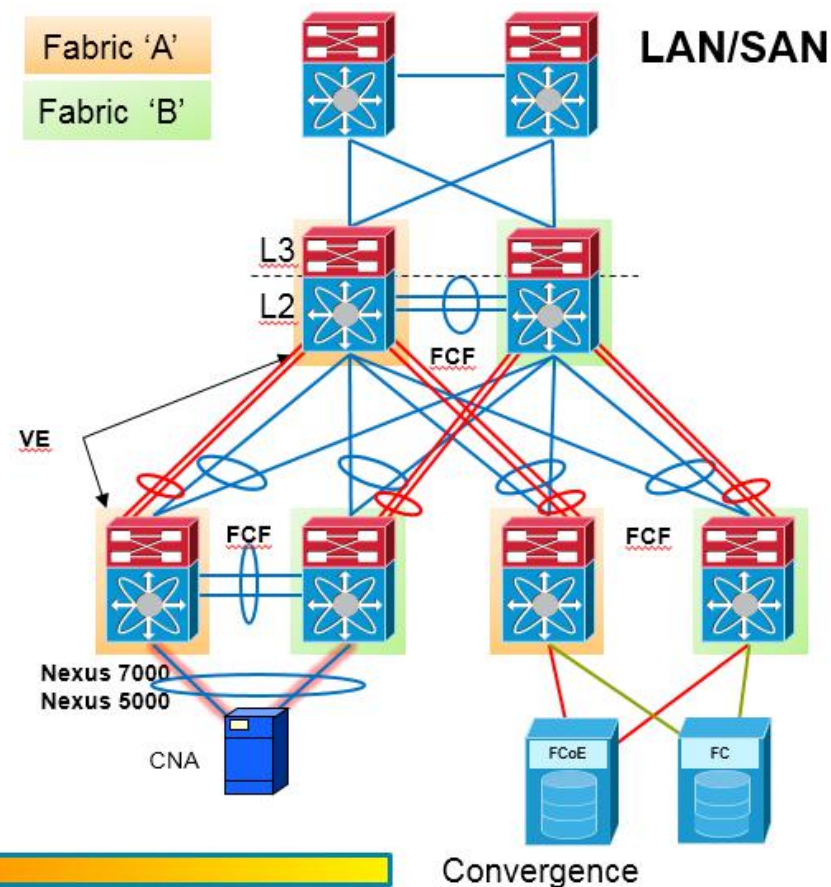
- ✓ One wire for all traffic types
- ✓ QoS guarantees minimum bandwidth allocation
- ✓ No Clear Port ownership
- ✓ Desirable for DCI Connections

- ✓ Dedicated wire for a traffic type
- ✓ No Extra output feature processing
- ✓ Distinct Port ownership
- ✓ Complete Storage Traffic Separation



## 融合的网络 – Dual Fabric 使用专用链路

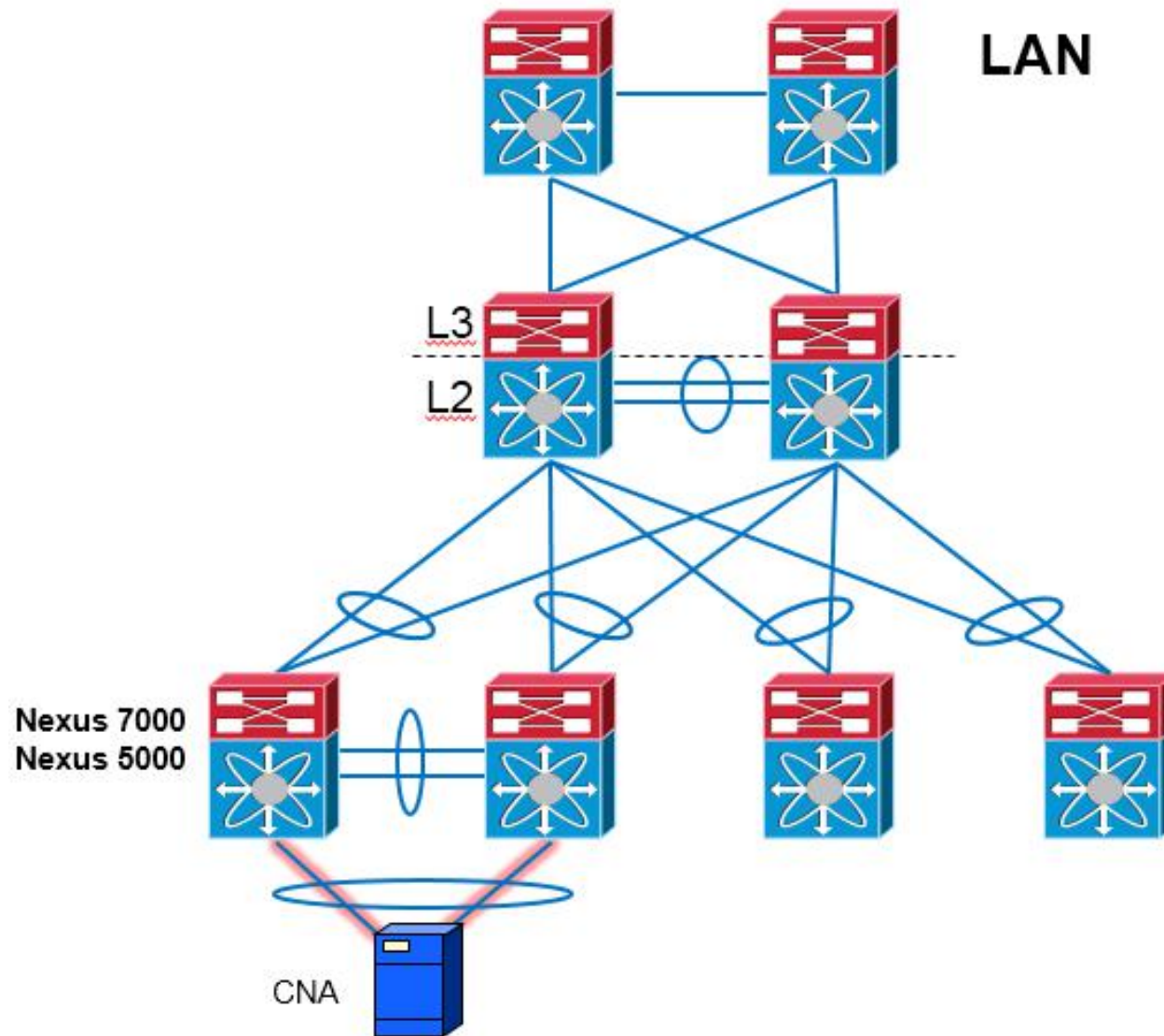
- LAN/SAN流量共享相同的物理设备
- 在交换机之间，LAN/SAN流量使用专用链路
- 所有接入和汇聚层交换机都是FCoE FCF
- 在交换机之间的专用线路是VE\_Port
- N7K可以使用Storage VDC



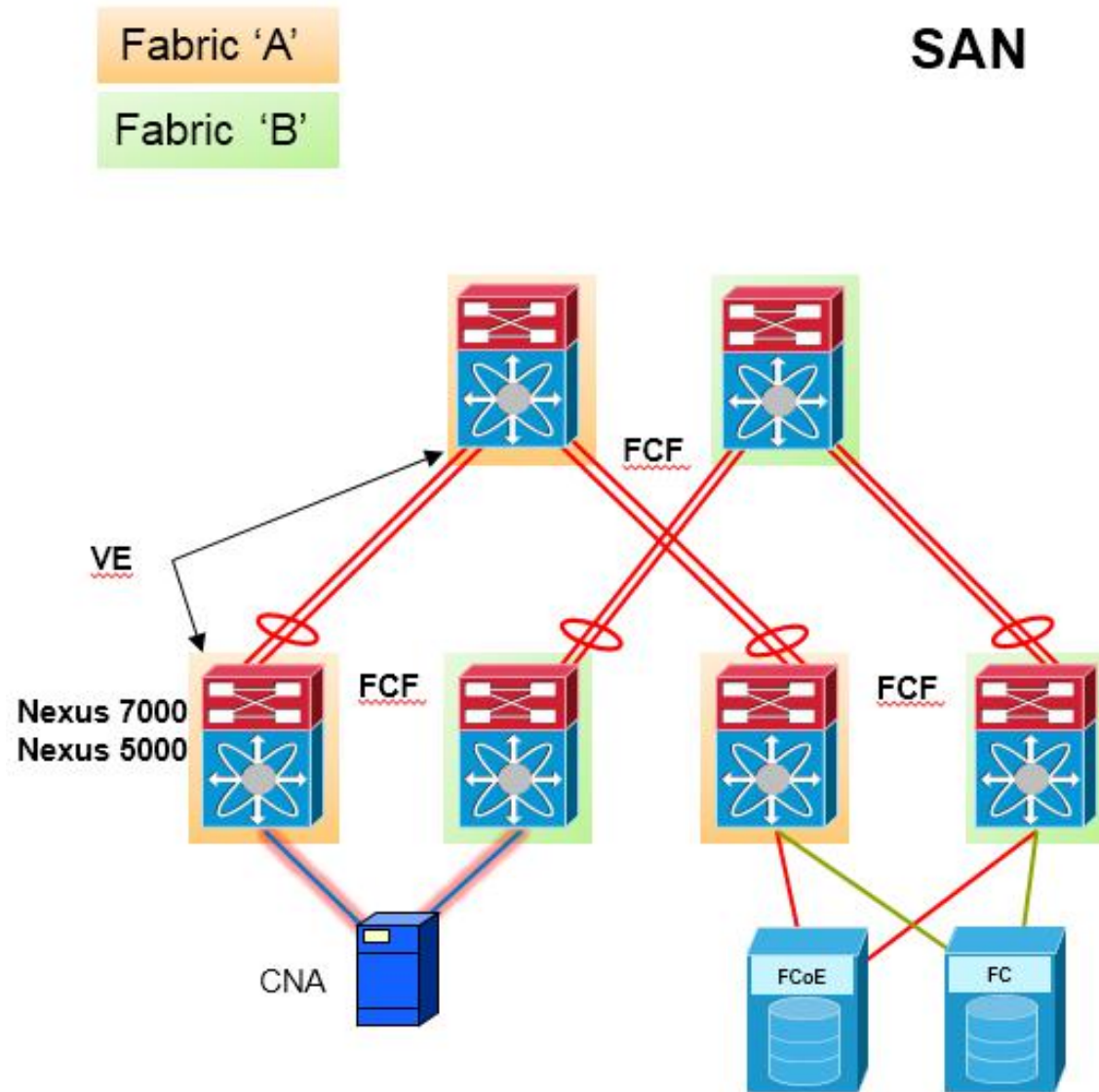
Isolation

Convergence

# 融合的网络 - Dual Fabric 使用专用链路(续)

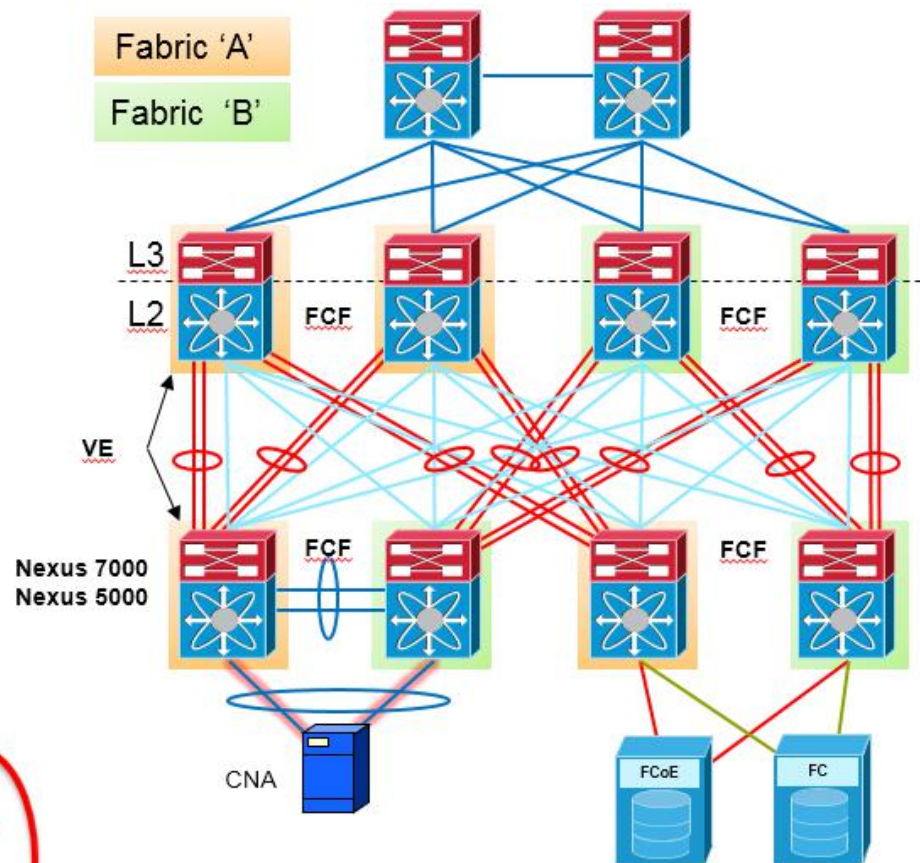


# 融合的网络 - Dual Fabric 使用专用链路(续)



## 融合的网络 – Dual Fabric 使用专用链路(续)

- LAN/SAN流量共享相同的物理设备
- 在交换机之间，LAN/SAN流量使用专用链路
- 所有接入和汇聚层交换机都是FCoE FCF
- 在交换机之间的专用线路是VE\_Port
- N7K可以使用Storage VDC



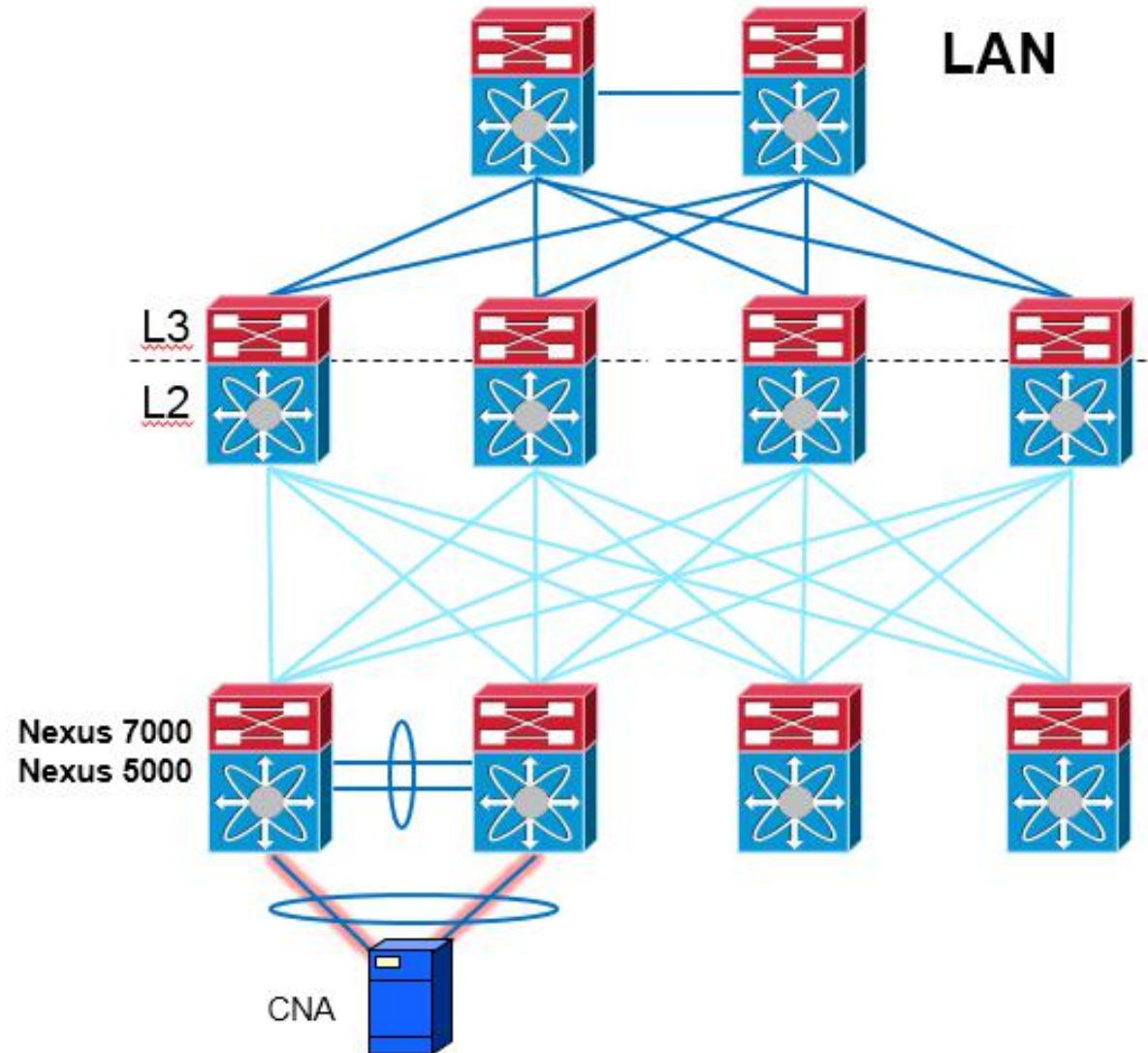
Isolation



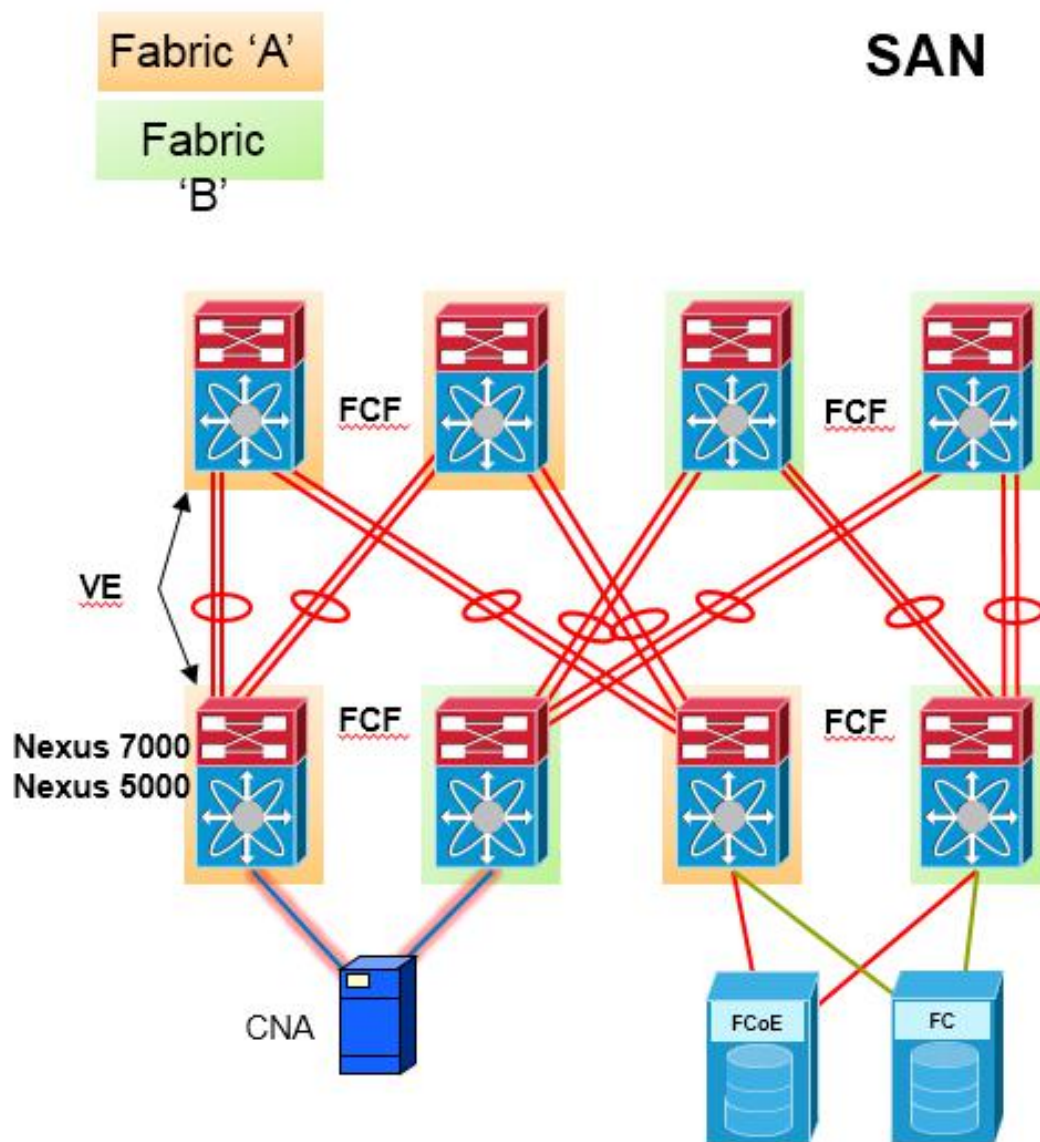
Convergence



# 融合的网络 - Dual Fabric 使用专用链路(续)

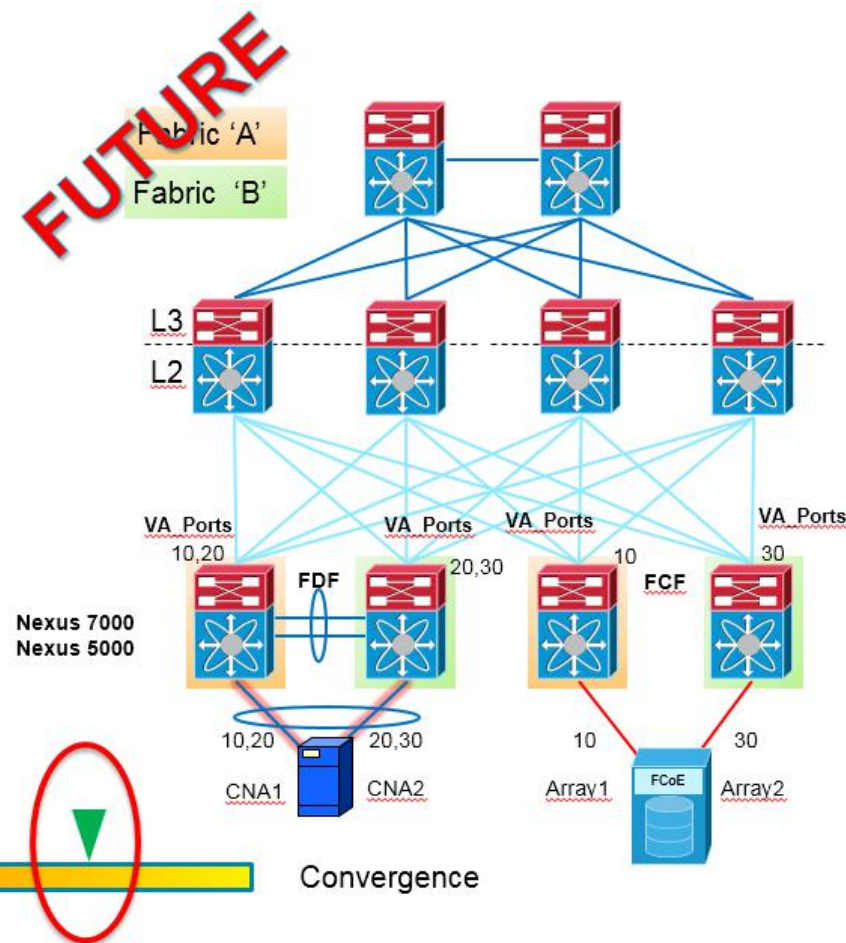


# 融合的网络 - Dual Fabric 使用专用链路(续)



## 融合的网络 – Dual Fabric 使用专用链路(续)

- LAN/SAN流量共享相同的物理设备
- FabricPath enabled
- 在每个邻居FCF交换机之间使用VA\_Ports
- 单独的域
- FDF和FCF做透明的failover
- Single process & database (FabricPath)用于转发
- 对于LAN/SAN提升到N+1冗余
- 共享的链路增加fabric的灵活性和扩展性
- 使用zoning isolation和multipathing冗余来区分SAN A和SAN B



Isolation



Convergence



