FCoE介绍

TSE Song Zhongru

议题

FCoE - Fiber Channel Over Ethernet
DCB - Data Center Bridging
FCoE Products Introduction and Deployment

云计算时代，数据中心的网络融合趋势

FC不能直接与以太网互通：

服务器配置多网卡，高散热，高能耗，运维成本高。

FC行业封闭：寡头经济

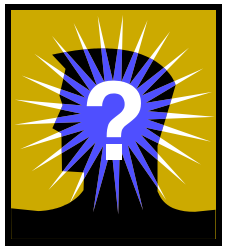
技术发展缓慢，目前还停留在8G。

高速以太网时代已经到来：

10GE以太网已经开始规模应用，目前正朝着40GE/100GE演进。

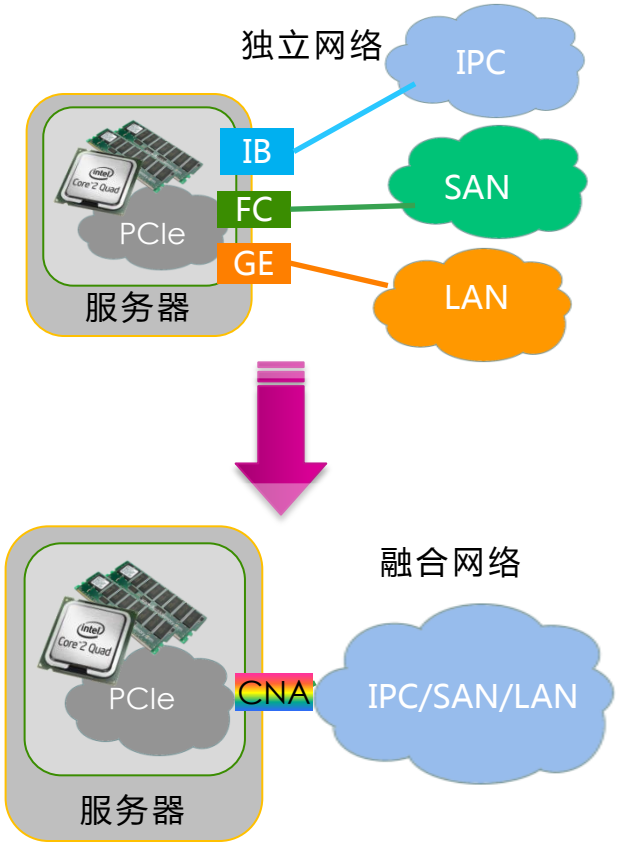
增强以太网技术出现：

FC在以太网传输可以确保无丢包。



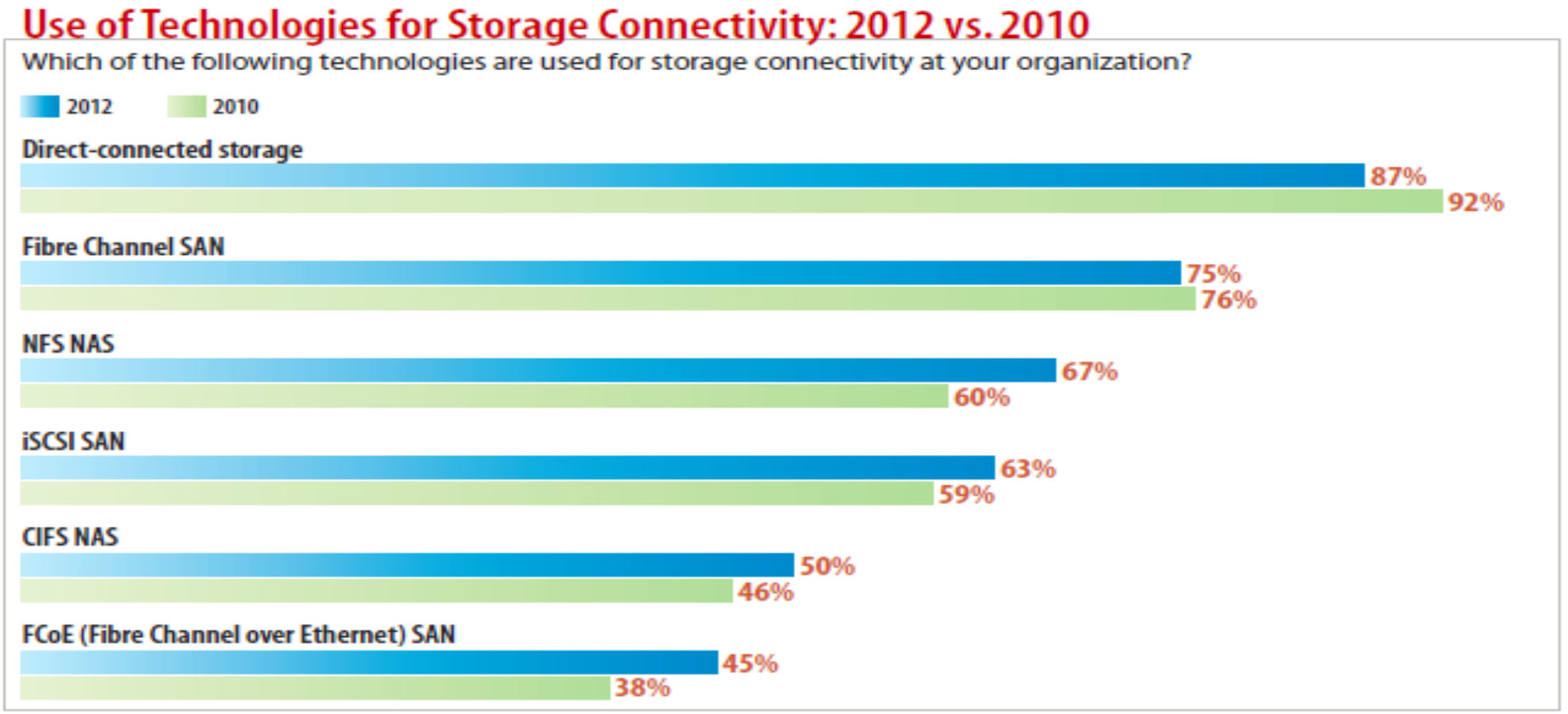
企业如何保留现有的FC 基础架构？

如何以更低的价格提供相同的性能？



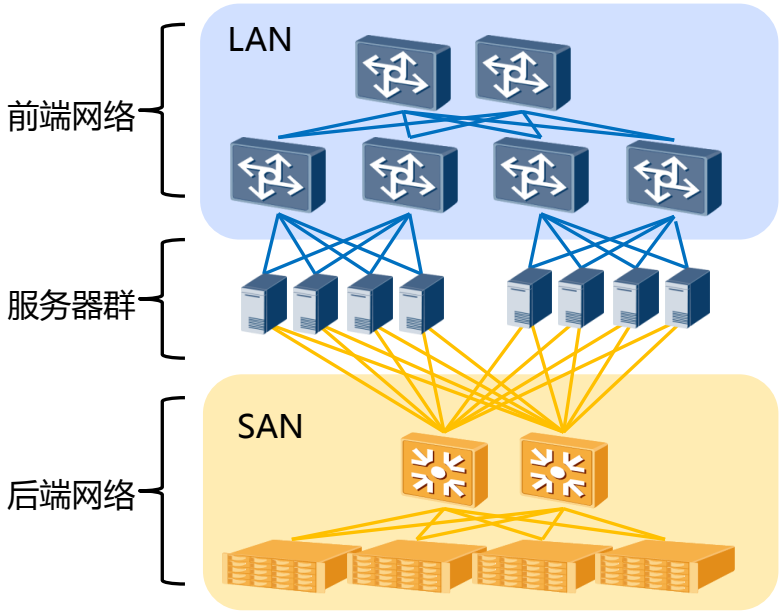
存储连接技术

Direct-connect storage, FC SAN, and NAS are the top 3



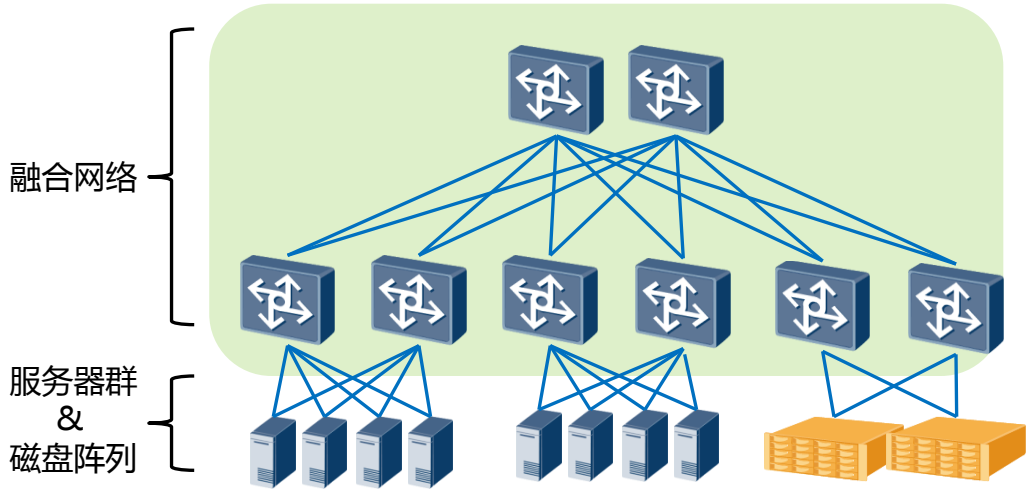
FCOE—数据中心的融合网络

传统的数据中心架构



融合后

数据中心融合网络架构



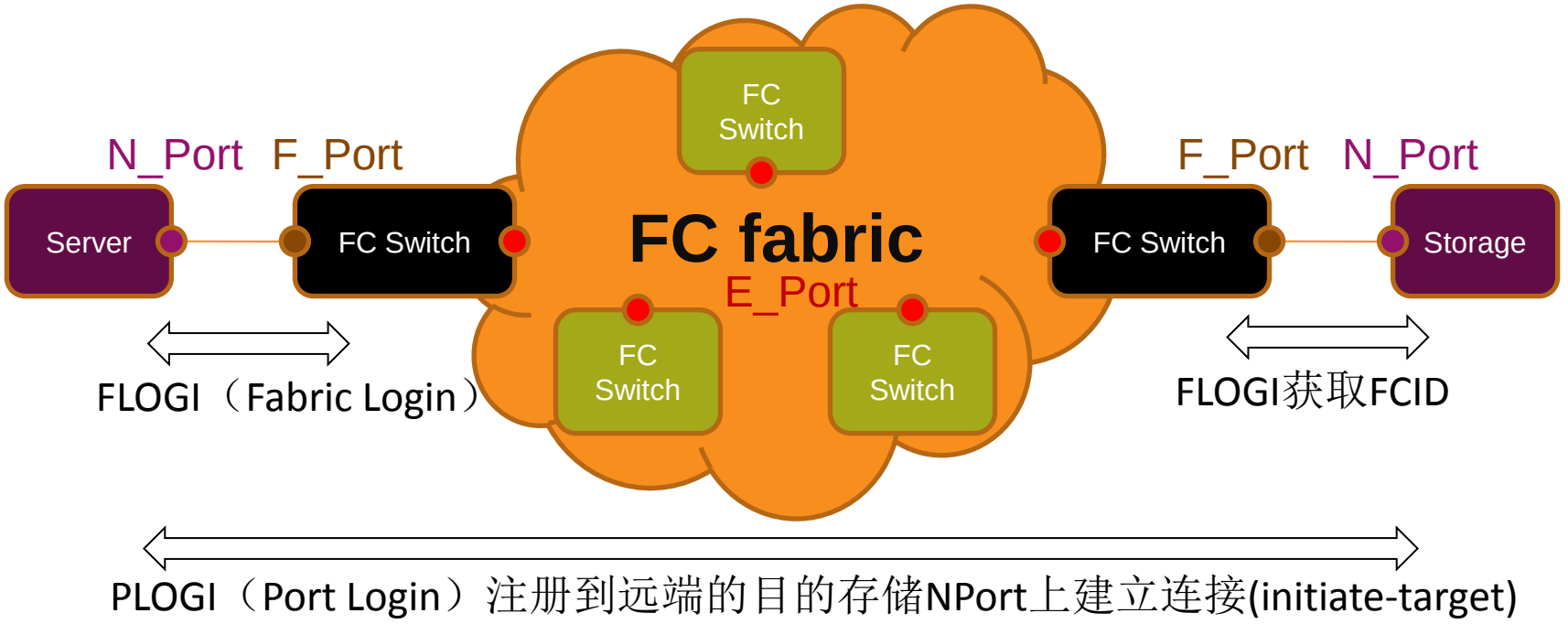
当前数据中心架构的问题

- 网络复杂，LAN/SAN独立部署，扩展困难
- 能效比低，服务器上至少配置4~6块网卡，增加功耗

融合网络

- 网络简化，LAN/SAN融合，统一交换
- 低TCO，服务器配置CAN融合网卡

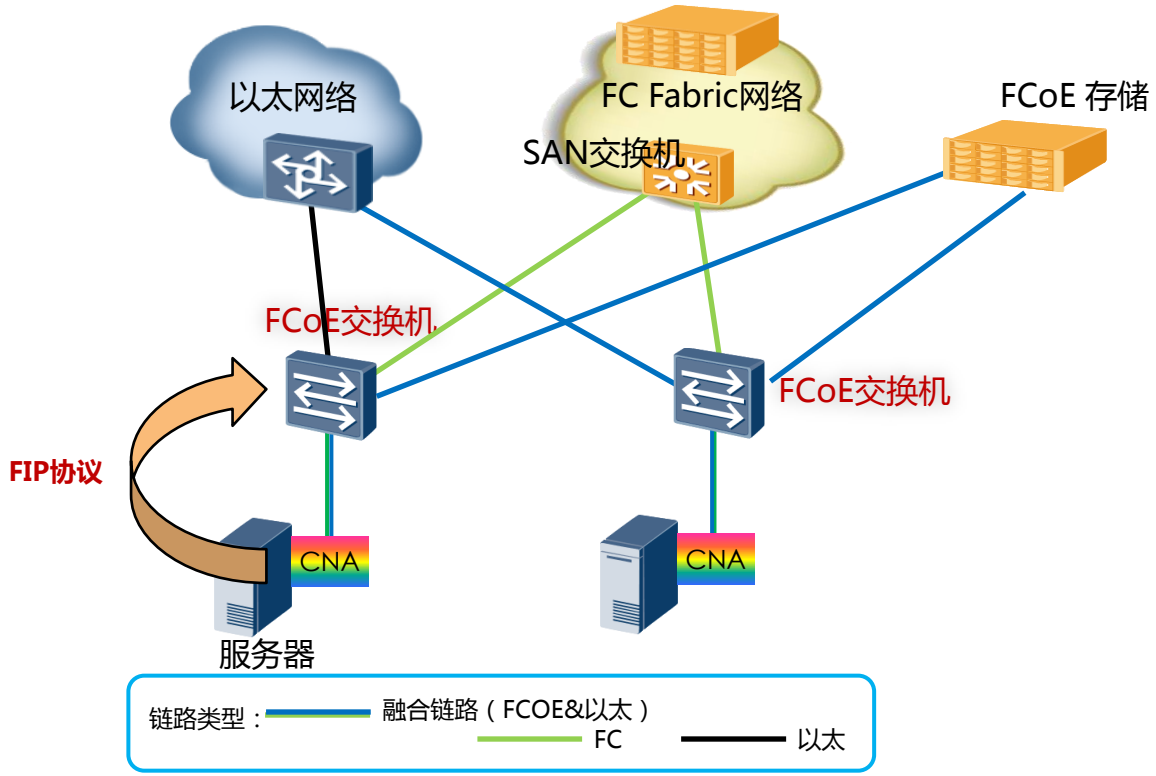
SAN网络组件和工作流程



Nport是服务器或存储等终端节点连接FC网络的的接口
Fport是FC交换机设备连接服务器或存储等终端节点的接口
EPort是FC交换机互联接口

FC ID：包含Domain ID，是基于接口的，每个NPort的FC ID是由**直连的FC Switch动态分配。FC ID的主要作用就是供数据报文在FC网络中寻址转发，是端到端不变的，类似IP地址**
FSPF: FC网络用Fabric Shortest Path First进行FC ID的寻址学习
WWN：节点和每个接口唯一的，WWN的作用是为了身份识别和安全控制，在zone中区分

FCOE基本组件



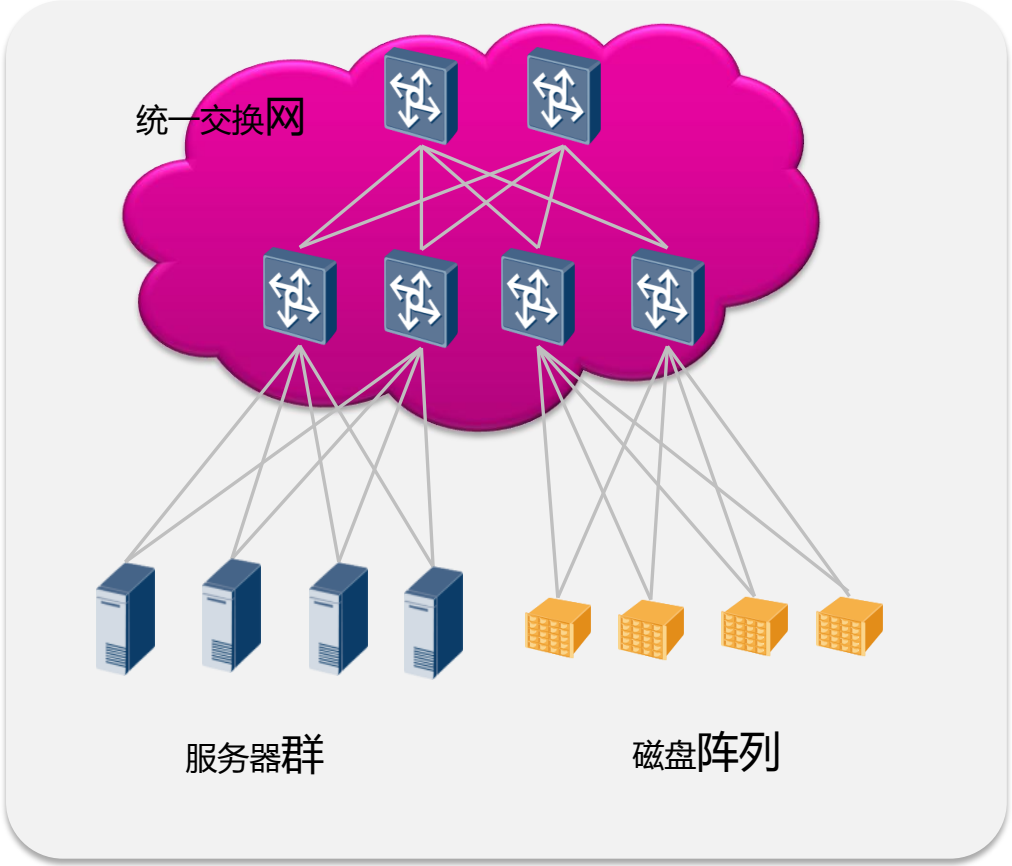
FCoE组件

- 服务器通过10G融合网卡接入到FCoE交换机。
- FCoE交换机对以太业务和FC业务进行分流。
- 通过FIP协议，由交换机完成用户的FCoE的初始化，分配FCID和MAC地址。

FCoE交换机的分类

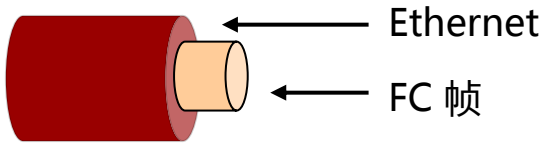
- 物理形态：FCoE交换机可以支持FCoE，FC两种端口。
- 功　　能：按照不同功能分为FCF,NPV,FDF,FSB。

FCOE



什么是FCoE？

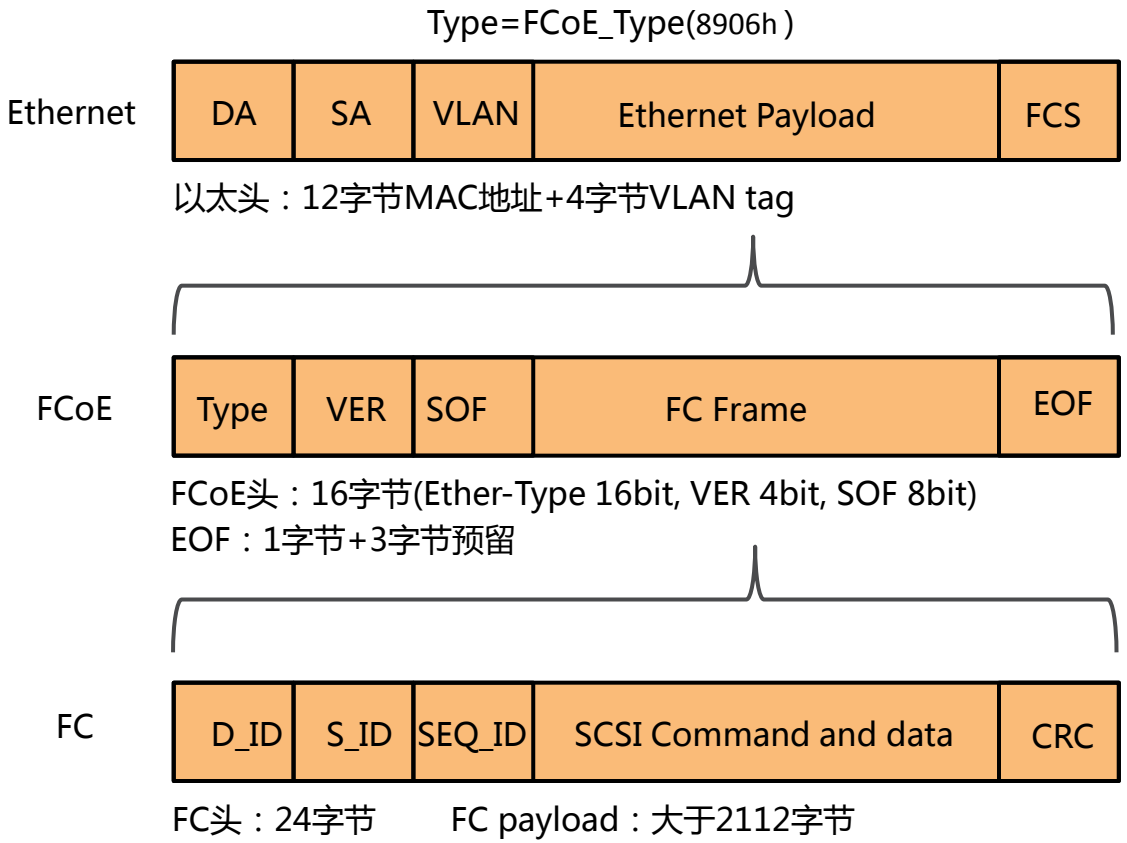
- FC over Ethernet,将光纤通道架构运行于增强以太网（CEE）上，从而形成融合网络。



融合网络的特点

- 服务器配置CNA融合网卡,减少网卡数量，降低功耗；
- FCoE不会改变FC基础架构，与现有FC基础设施无缝互通，实现对FC SAN的投资保护；
- 拥塞情况下仍提供无丢包的可靠传输，从而引入了数据中心桥接（DCB），也称融合增强以太网（CEE）。

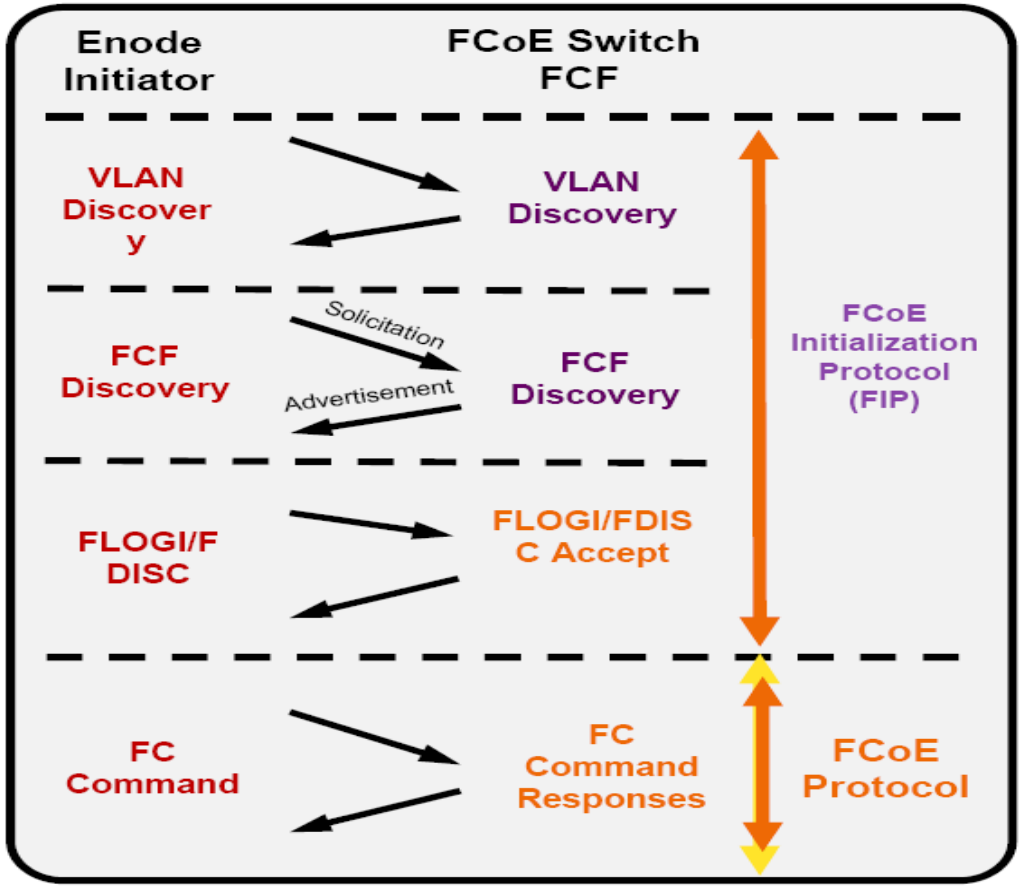
FCOE—FCoE报文格式



FCoE报文格式

- DA：目的MAC，单播为下一跳FCF的MAC，组播为保留MAC。
- SA：源MAC，为每一跳FCF的MAC或终端MAC。
- VLAN：FIP协议中指定 VLAN或FCoE数据VLAN。
- Type：以太类型，取值为8906h时，对应的是FCoE报文。8914对应FCIP报文
- FCS：以太帧校验。
- VER:版本号
- SOF:开始标志
- EOF:结束标志
- D_ID：目的FCID地址。
- S_ID：源FCID地址。
- SEQ_ID：Sequence号。

FCoE—FCIP协议

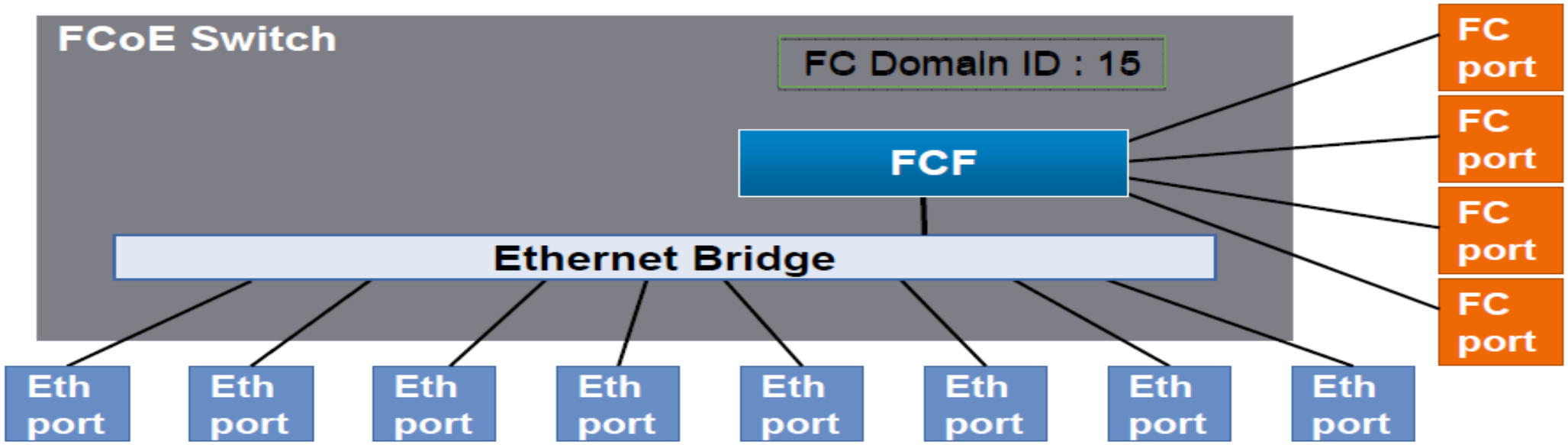


•FIP（FCoE Initialization Protocol）：
进行初始化连接
FIP运行于VFPort和VNPort之间或VEPort之间

•FIP在接口使能后一共做了三件事：
1、使用本地VLAN（如VLAN1）确认FCoE数据报文将要使用的VLAN ID。
2、和FCF建立连接
3、FLOGI/FDISC（Discover Fabric Service Parameters，FC节点设备第一次向FC交换机注册请求FC ID时使用FLOGI，后面再续约或请求其他FC ID时都使用FDISC）

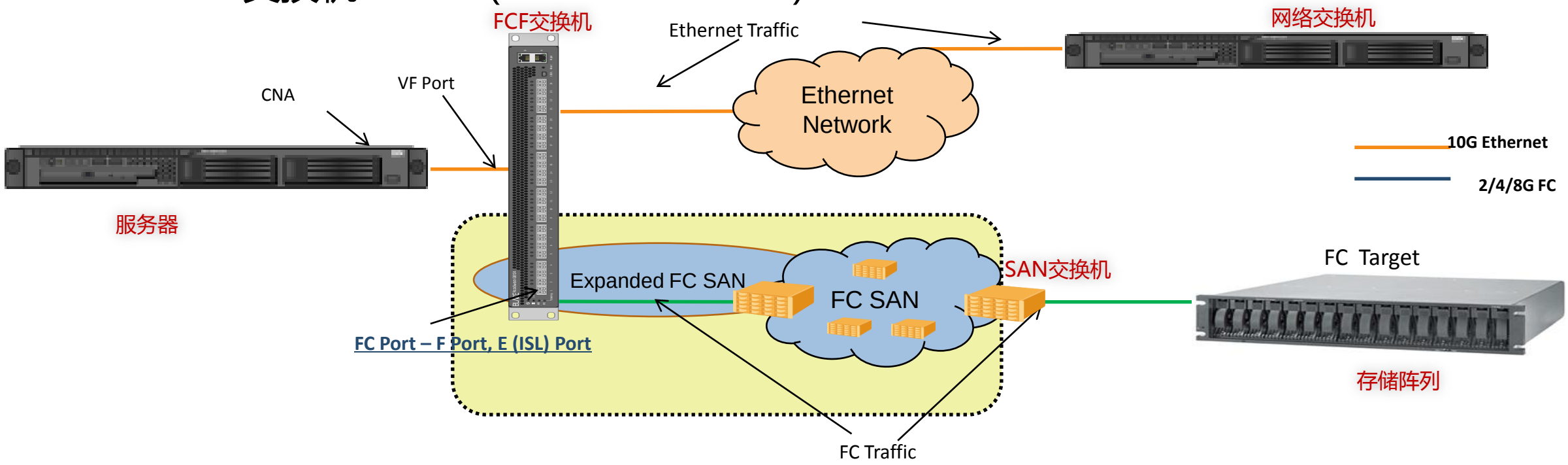
FCF(Fibre Channel Forwarder)

FCoE交换机的逻辑结构：



FCF（Fibre Channel Forwarder）是FCoE里面重要的角色，可以是软件或者芯片硬件实现，需要占用Domain ID，处理FCoE交换机中所有与FC相关的工作，功能类似SAN交换机OS，如**封装解封装和FLOGI等**，每个FCF都会有自己的MAC,Enode MAC也是由FCF分配的并具有唯一性，叫做**FPMA（Fabric Provided MAC Address**）FPMA由两部分组成，FC-MAP与FC ID
FCF分配FC ID给Enode
根据FCID支持本地转发，参与计算FSPF，占Domain ID

FCoE—FCoE交换机： FCF（ FC Forwarder ）



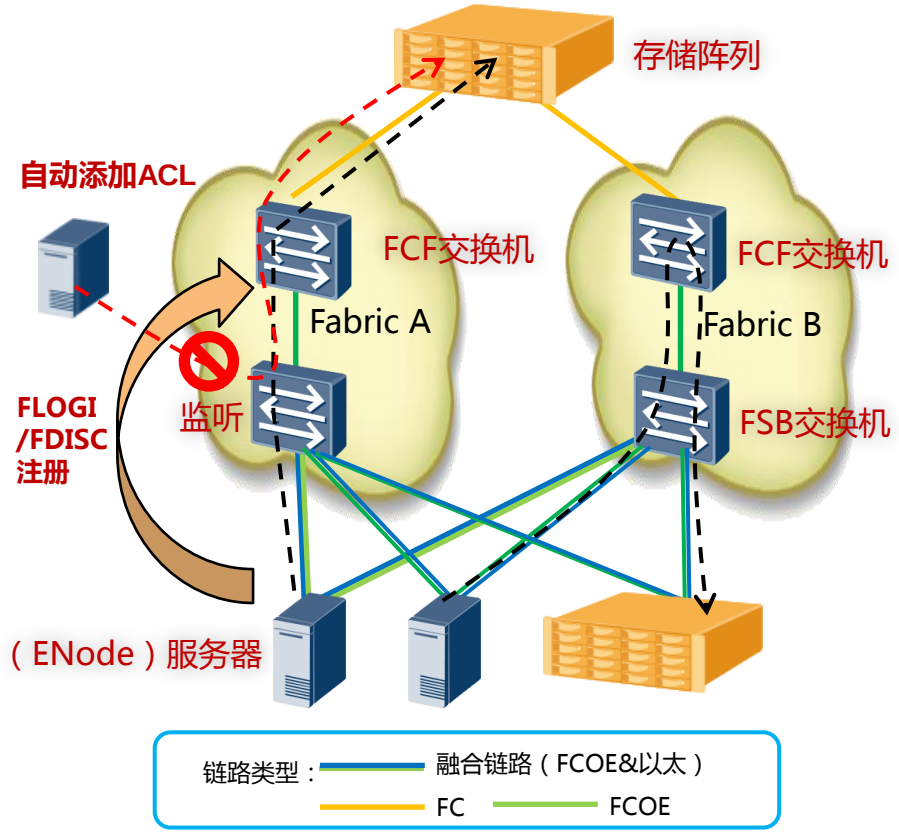
FCF功能介绍

- FCF（ FC Forwarder ）FC转发器
- 控制面：提供FIP注册服务，并向所在Fabric网络提供NS（名字服务），RSCN（用户状态变更通知），FSPF（FC最短路径转发），zoning（分区）。
- 转发面：通过FCID进行转发，支持本地转发。
- 作用：提供FCoE初始化功能（FIP），并提供Fabric服务。

FCF的优缺点

- 优点：支持与传统FC网络对接，能够独立组网。
- 同时提供FC和10G两用接口
- 缺点：受限于domain ID的规格，限制了网络规模。

FCoE—FCoE交换机：FSB（FIP Snooping Bridge）



FSB功能介绍

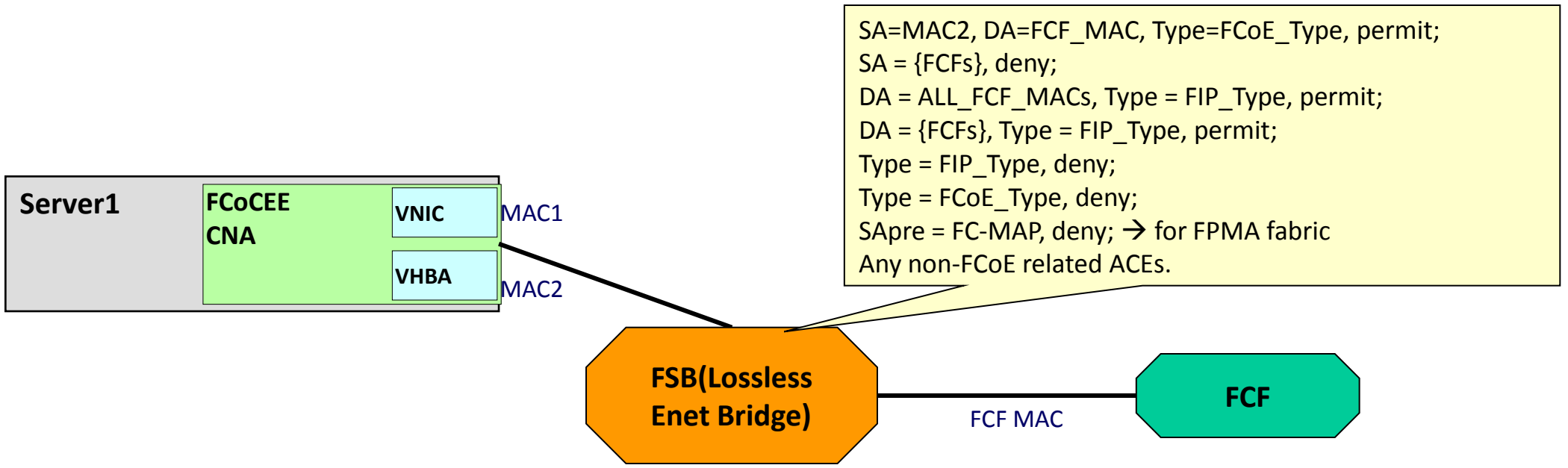
- FSB（FIP Snooping Bridge）FIP监听桥
- 控制面：不参与FIP控制协议，但监听FIP消息，并依据监听的消息控制链路的访问权，增强安全性。
- 转发面：通过MAC进行转发。
- 作用：提供10GE FCoE融合接入并进行以太透传。

FSB的优缺点

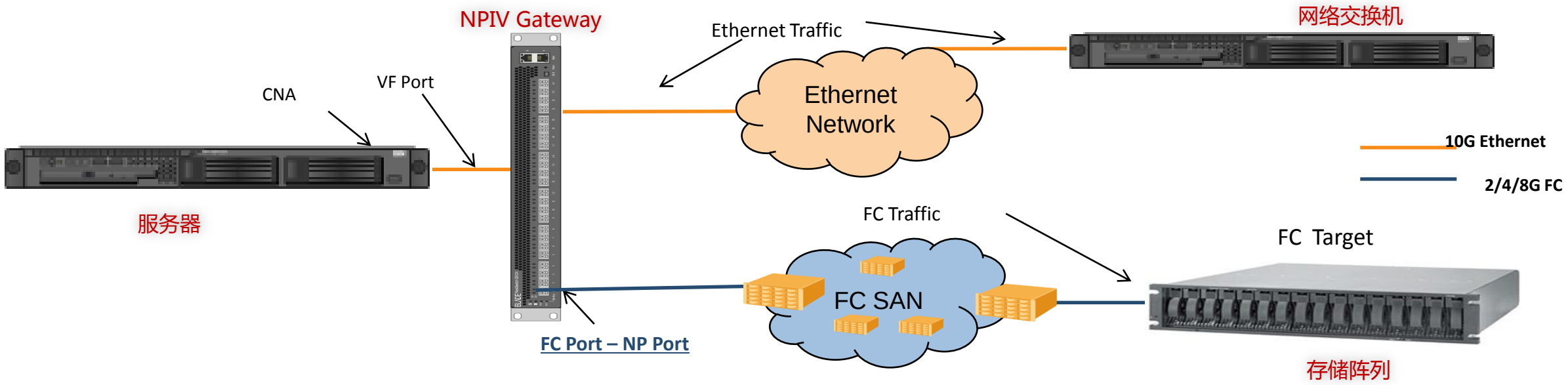
- 优点: 部署简单，基于MAC转发，效率高。
- 缺点：不能独立组网，仅支持FCoE端口，不能兼容FC网络接入，从Source到Target的存储流量必须经过FCF，流量回绕。

FSB (FIP Snooping Bridge)-自动添加访问控制

- FIP Snooping Bridge
  - Snoops FIP messages (VLAN Disc, FCF Disc and FLOGI)
  - Secure host-FCF path with dynamic ACL

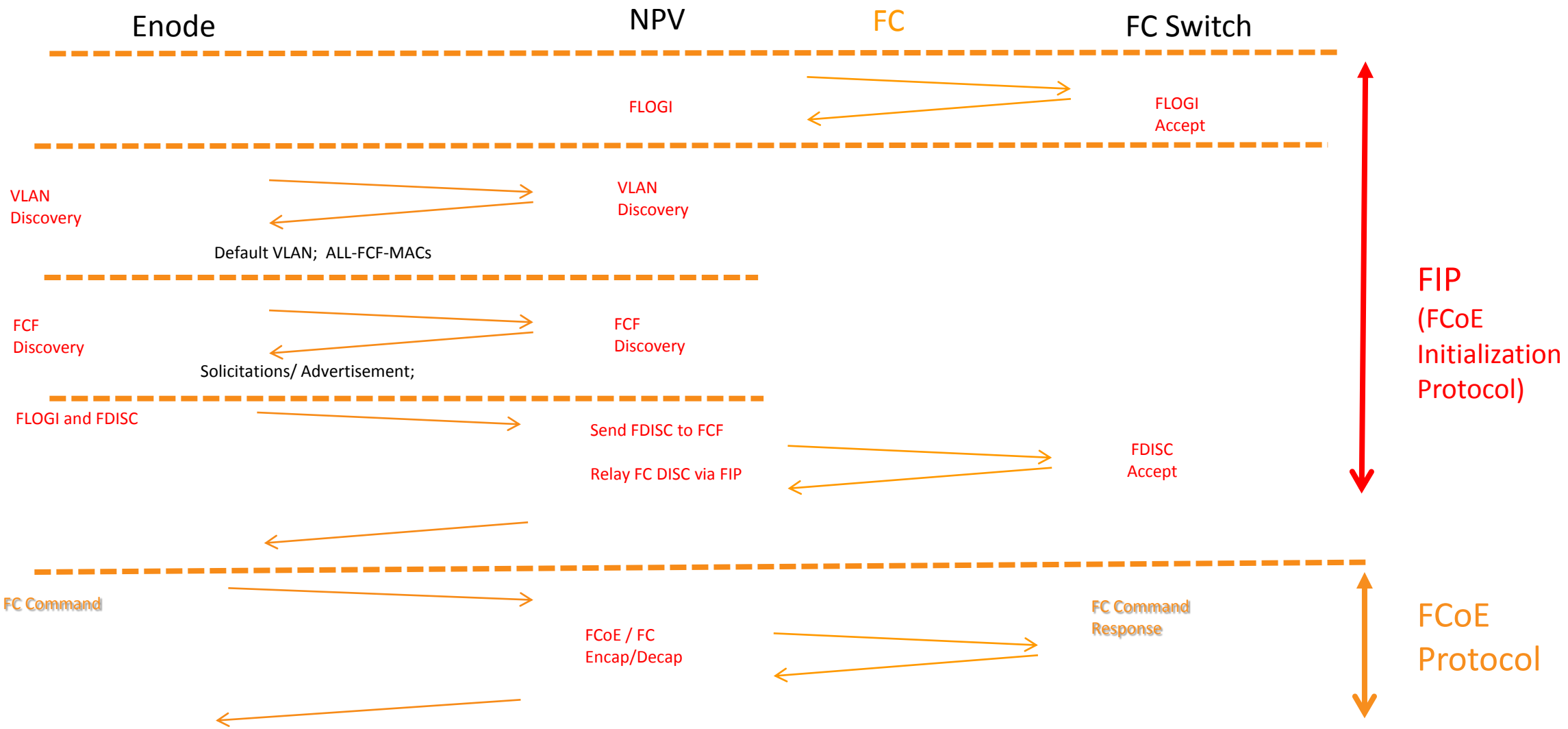


FCoE—FCoE交换机：NPIV Gateway

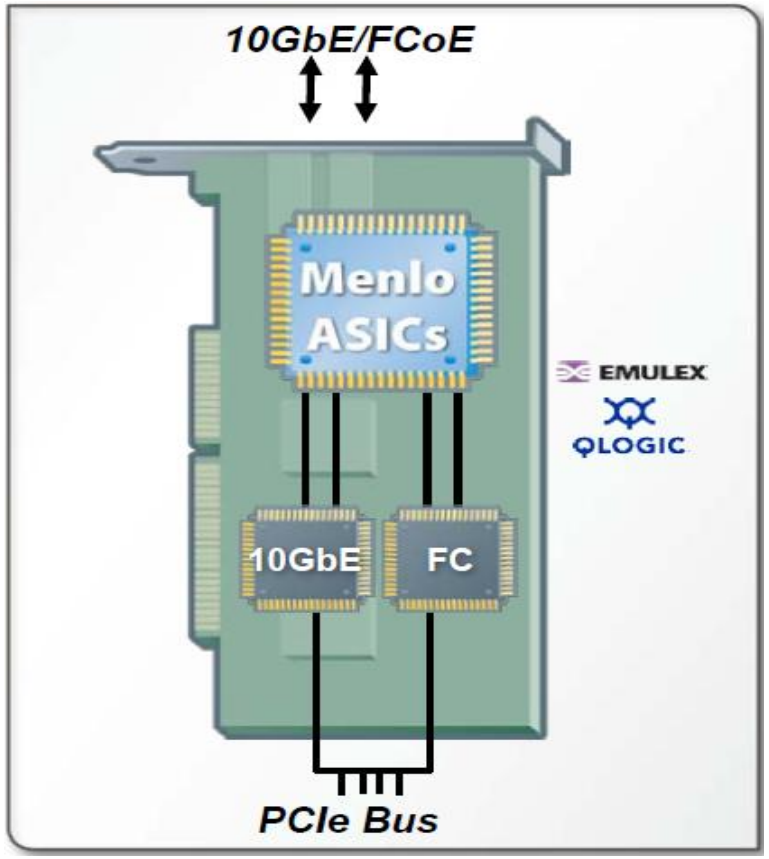


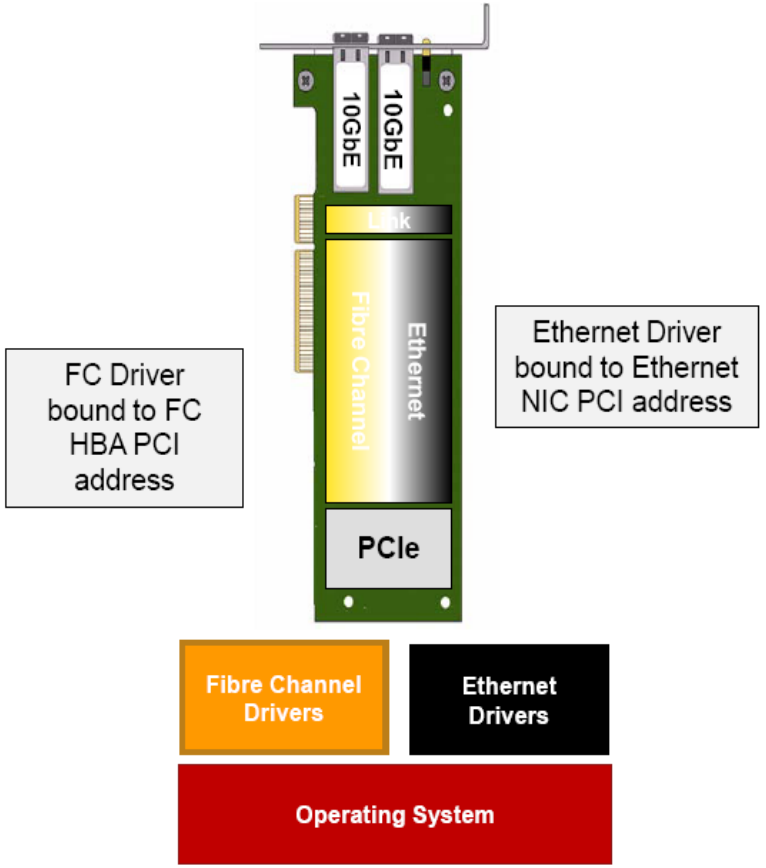
- FCoE NPIV Gateway
  - 代理ENODE节点FC login请求，从SAN交换机获取FCID给ENODE
  - Provides NP Port functionality to connect to FC SAN
  - 不占用**domain-id**
  - 不需要实现FCF，不提供FABRIC服务，不用管FC转发，不计算FSPF
  - 在CNA网卡和FC Switch之间对转发的数据报文进行FCoE头的封包解包

NPV Gateway



FCoE-CNA网卡

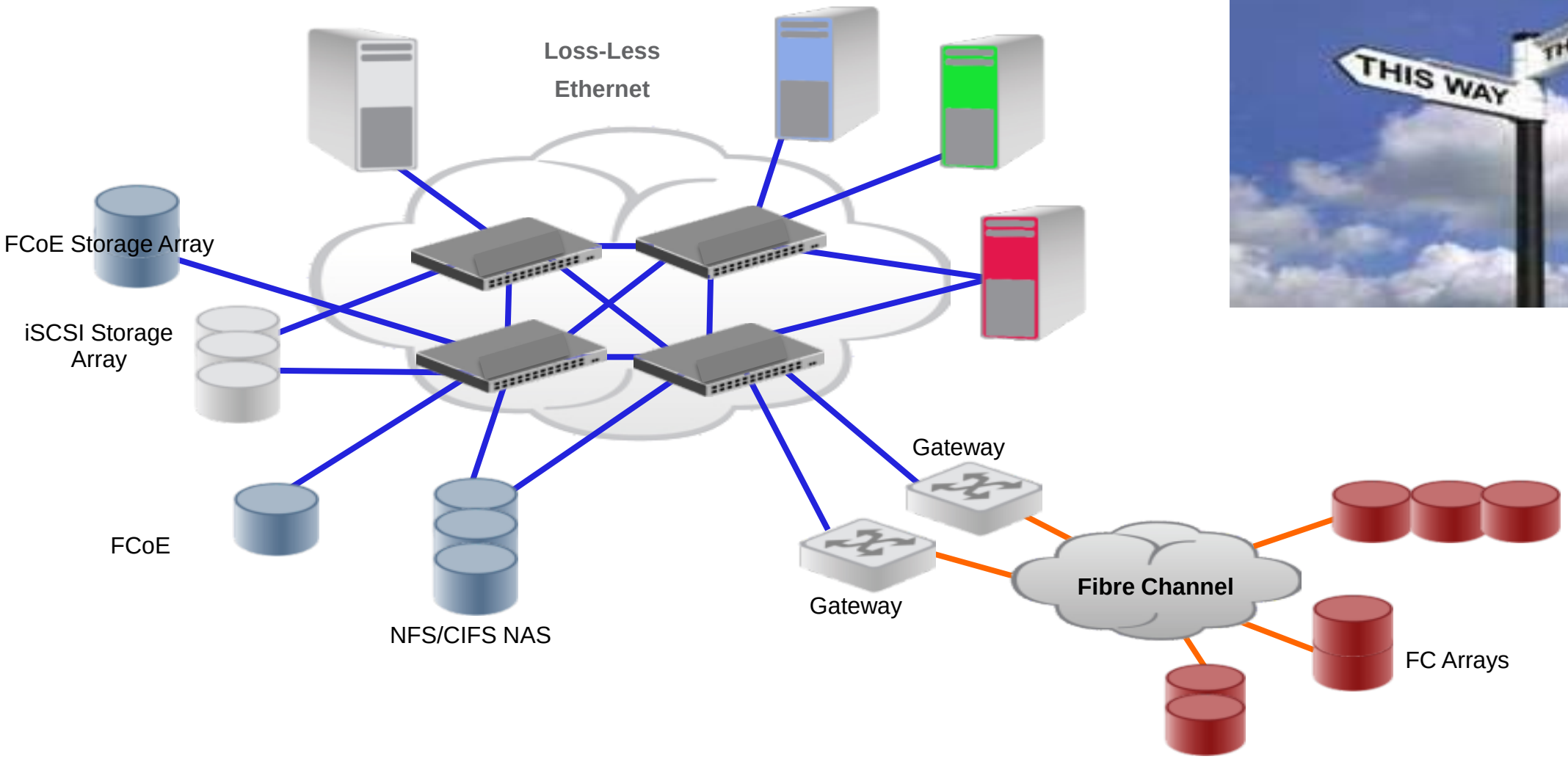




CNA（Converged Network Adapter）
万兆Ethernet和FC HBA（Host Bus Adapter）网卡的合体，里面包含两个独立芯片处理Ethernet和FC各自的流量
在操作系统上看到的就是两个独立的Ethernet和FC网络接口，其上再增加第三个芯片进行流量混合封包处理

DCB - Data Center Bridging

存储融合网络呼唤无损以太网



DCB—以太技术的完善

传统以太

- 拥塞控制：基于TCP的丢包重传以及滑动窗口机制和于目的端的流控。
- Pause机制：以太现有技术，通过Pause帧通告下游停止发送。

传统FC

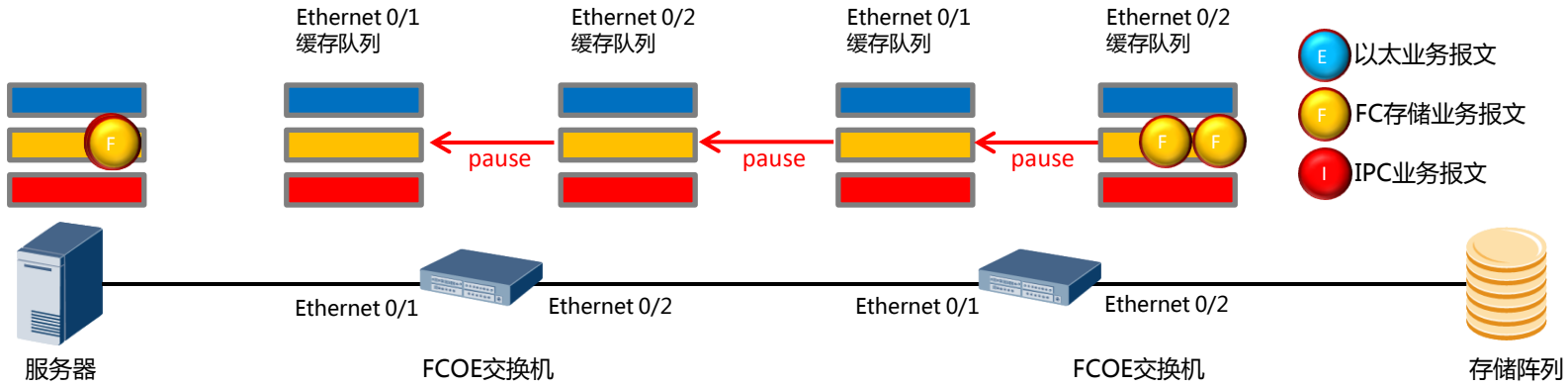
- 拥塞控制：基于传输节点间BB_Credit和端到端EE_Credit的流量拥塞控制(类似令牌机制，针对带宽传输能力进行传输节点间的流控)。

增强以太网

- 拥塞控制：利用以太Pause机制，并参考FC的流控实现方法，定义了基于PFC,ETS的逐级流量拥塞控制，和基于CN的端到端流量拥塞控制机制。

存储业务要求在网络传输中不丢包，传统的FC通过Credit机制来保证，而传统的以太技术只能在传输中做到“尽力而为”，通过引入DCB技术，改进以太网在上述方面的不足，为FCoE融合业务提供了传输质量保证。

DCB—PFC(Priority Flow Control)802.1qbb



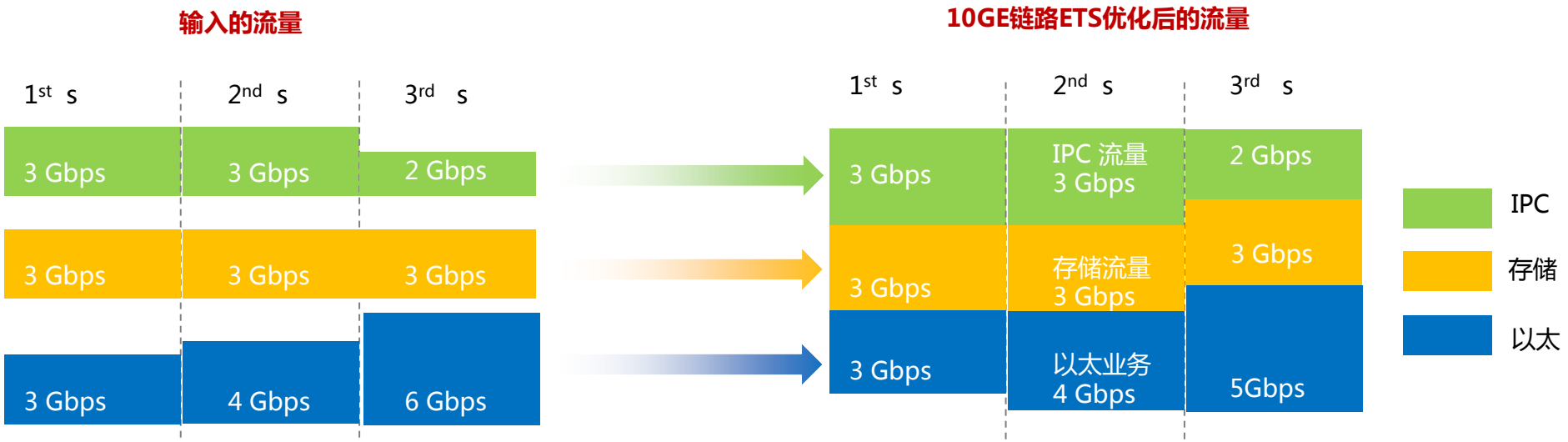
PFC业务识别

- 网络规划设计时，预先定义不同业务的优先级，并为不同的业务设定预定队列阈值。
- 基于业务优先级（0~7）的拥塞控制，超过既定阈值发生拥塞时不影响其他业务的正常处理和转发。

逐级反压

- 通过以太Pause机制，在发生拥塞时，针对该类业务向下游发送反压信号。

DCB—ETS(Enhanced Transmission Selection)802.1qaz



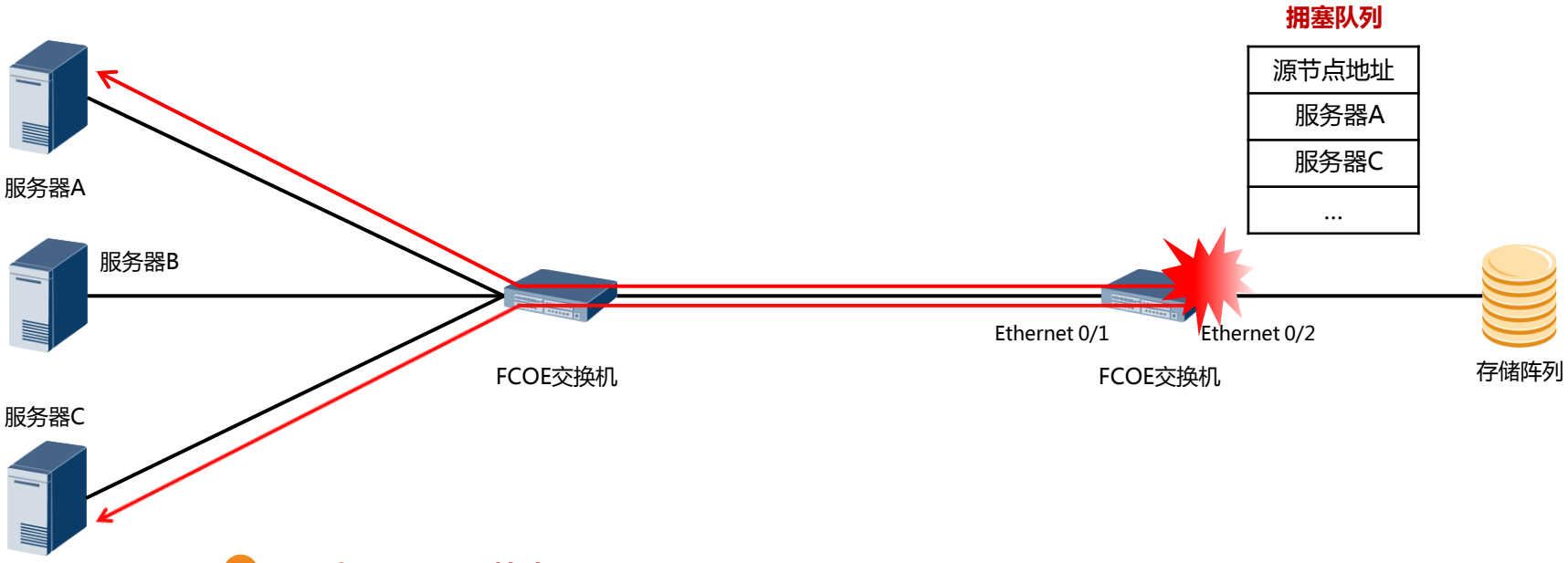
ETS业务识别

- 网络规划设计时，预先定义不同业务的优先级，并为不同的业务设定调度方式和带宽。
- 对于时延高的IPC业务设定优先级调度，对SAN,LAN设定轮询调度。

配合CN,PFC反压机制

- 低时延、高可靠类业务比如IPC,SAN，在超出预设带宽时造成的队列拥塞，通过PFC或CN向下游反压减缓流量压力。
- 对于时延，丢包不敏感的LAN业务，不进行反压，由TCP/IP保证业务重传。

DCB—CN(Congestion Notification)

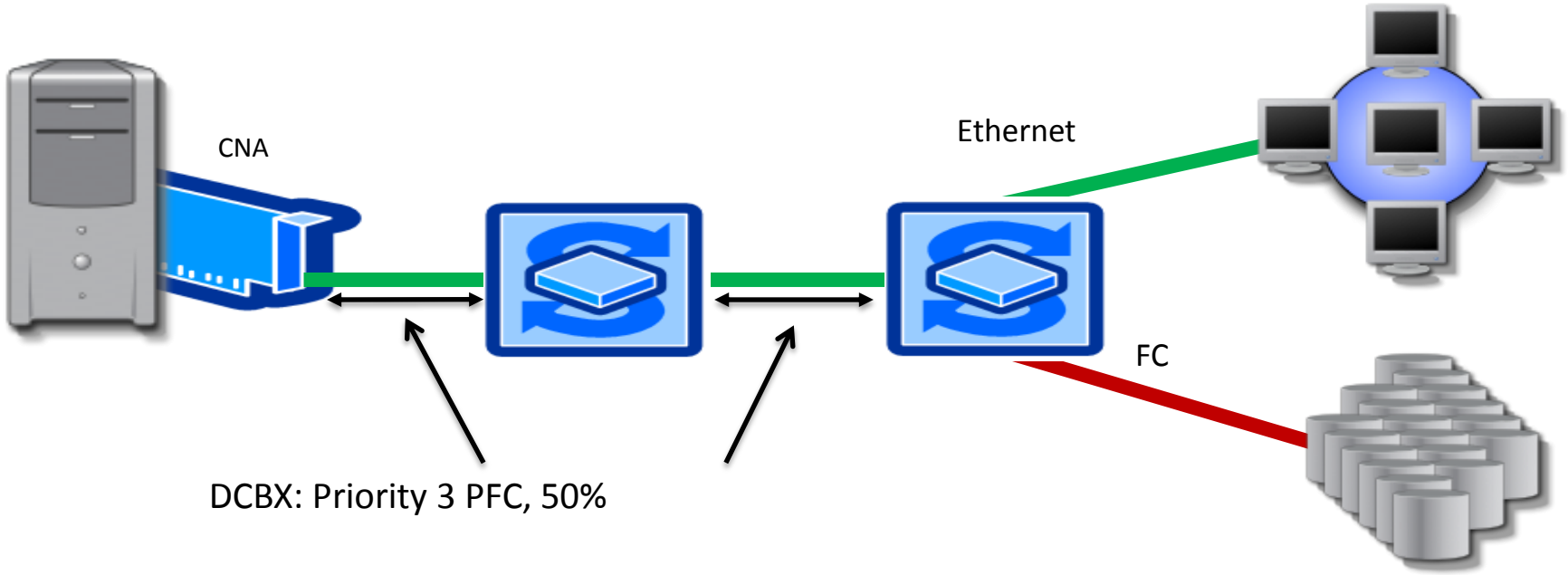


CN实现原理及特点

- 实现原理：发生拥塞时，从拥塞队列中查找到引发拥塞的报文源节点，直接向源节点发送拥塞通知，直到拥塞解除。
- CN特点：直接命中引发拥塞的源节点，阻止其继续向网络发送报文。

DCBX (IEEE 802.1Qaz)

- **Data Centre Bridging eXchange – DCBX**
  - Negotiates PFC and ETS at every connection in the data path, an extension of the established LLDP protocol
  - DCBX determines which 802.1p priorities are to be made lossless by applying PFC, and negotiates bandwidth guarantees for each priority group using ETS



FCoE Products Introduction and Deployment

IBM FCoE产品

- FSB
  EN4093R
  G8124E
  G8264

Figure 1. IBM System Networking RackSwitch G8124E

– FCoE Multihop
– CEE support

- FCF
  G8264CS

System x MTM: 7309-DRX, 7309-DFX
Power Systems MTM: 1455-64F

Benefits	Key Features
• **Business flexibility**	• 12 Omni Ports – 10GbE or 4/8Gb Fibre Channel • 36 10GbE ports; 4 40GbE ports

– FCoE Multihop
– FC SAN
– Direct attach FCoE storage
– CEE support

- NPV
  G8264CS
  CN4093

- CNA网卡

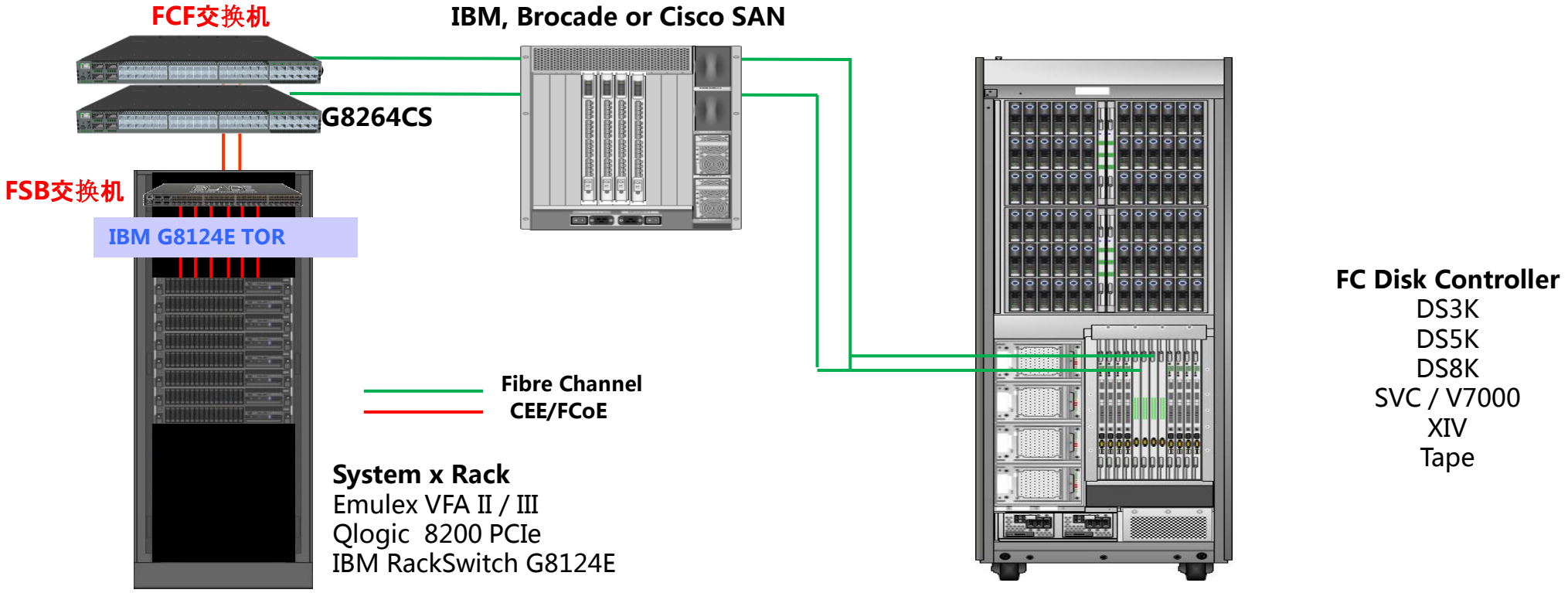
Emulex CNA

QLogic 8100 CNA

Brocade CNA

– FCoE and ISCSI(FOD)
– CEE(PFC ETS CN)
– Virtual fabric
–Bandwidth allocation

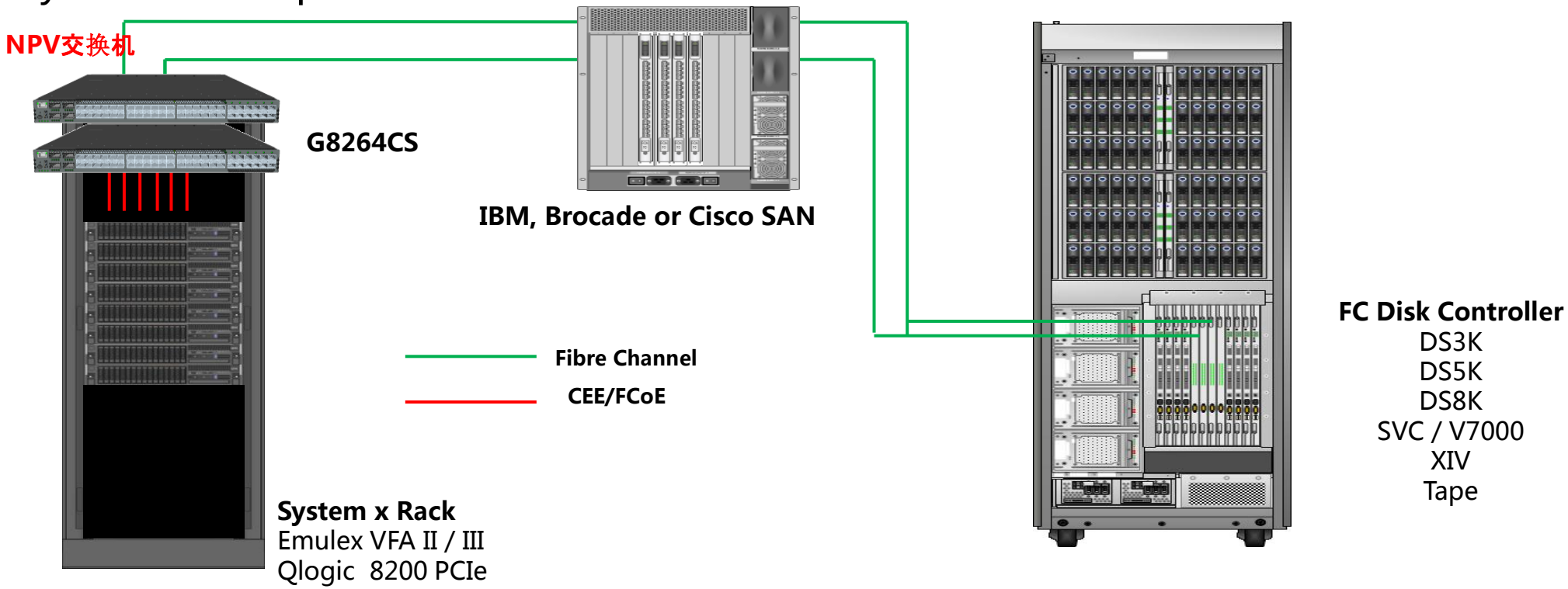
Rack level FCoE Aggregation to upstream FCoE switch splitting to FC SAN



Adapter	NIC Config	Transit	FCoE Switch	SAN Switch	Storage Target	OS levels
Emulex VFA II/III (Adapter + FOD Key) 49Y7950 + 49Y4274 49Y7940 + 49Y4274 95Y3751 (Included) 90Y6456 + 90Y5178 95Y3762 + 95Y3760 88Y7429 + 95Y3760	pNIC vNIC2	G8124E	G8264CS	Cisco SAN Brocade SAN	FC: Storwize V3700/7V000, SVC, DS3K/5K, DS8K, Tape, XIV	Win2008, WS2012, ESX 4/5, RHEL 5/6, SLES 10/11
Qlogic 8200 PCIe (Adapter + FOD Key) 90Y4600 + 00Y5624	pNIC vNIC2	G8124E	G8264CS	Cisco SAN Brocade SAN	FC: Storwize V3700/7V000, SVC, DS3K/5K, DS8K, Tape, XIV	Win2008, WS2012, ESX 4/5, RHEL 5/6, SLES 10/11

Target Availability
2/15/2013
2/18/2013

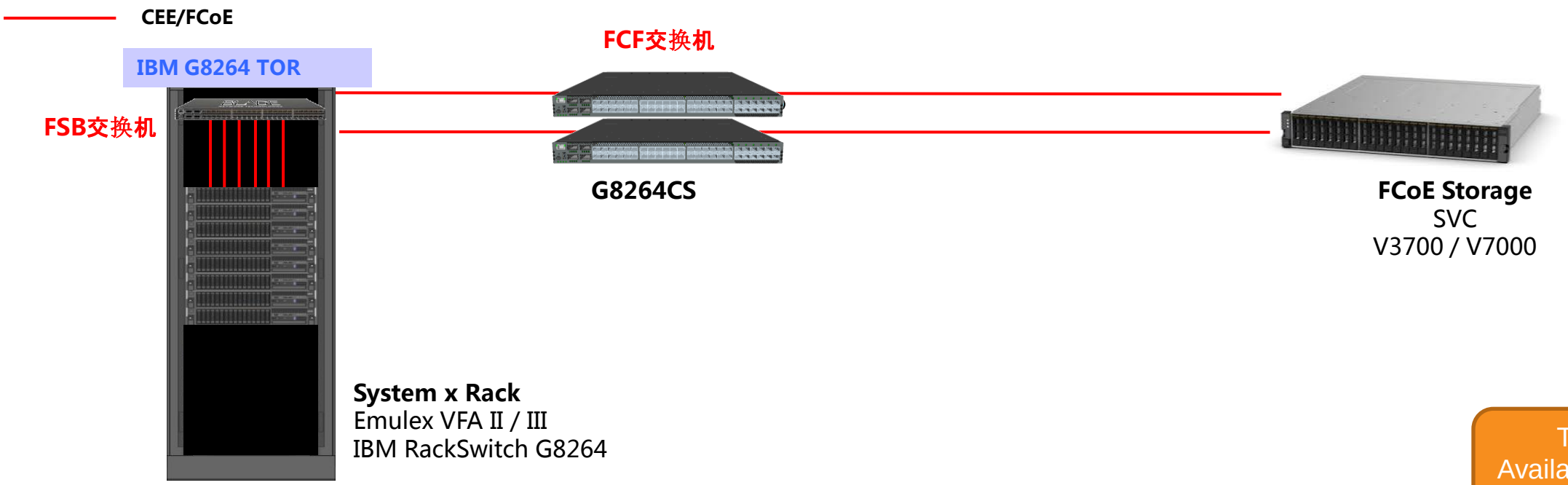
NPV Gateway in Rack to Upstream FC SAN



Adapter	NIC Config	FCoE Switch	SAN Switch	Storage Target	OS levels
Emulex VFA II/III (Adapter + FOD Key) 49Y7950 + 49Y4274 49Y7940 + 49Y4274 95Y3751 (Included) 90Y6456 + 90Y5178 95Y3762 + 95Y3760 88Y6454 + 95Y3760	pNIC vNIC2	G8264CS	Cisco SAN Brocade SAN	FC: Storwize V3700/V7000, SVC, DS3K/5K, DS8K, Tape, XIV	Win2008, WS2012, ESX 4/5, RHEL 5/6, SLES 10/11
Qlogic 8200 PCIe (Adapter + FOD Key) 90Y4600 + 00Y5624	pNIC vNIC2	G8264CS	Cisco SAN Brocade SAN	FC: Storwize V3700/7V000, SVC, DS3K/5K, DS8K, Tape, XIV	Win2008, WS2012, ESX 4/5, RHEL 5/6, SLES 10/11

Target Availability
2/15/2013
2/18/2013

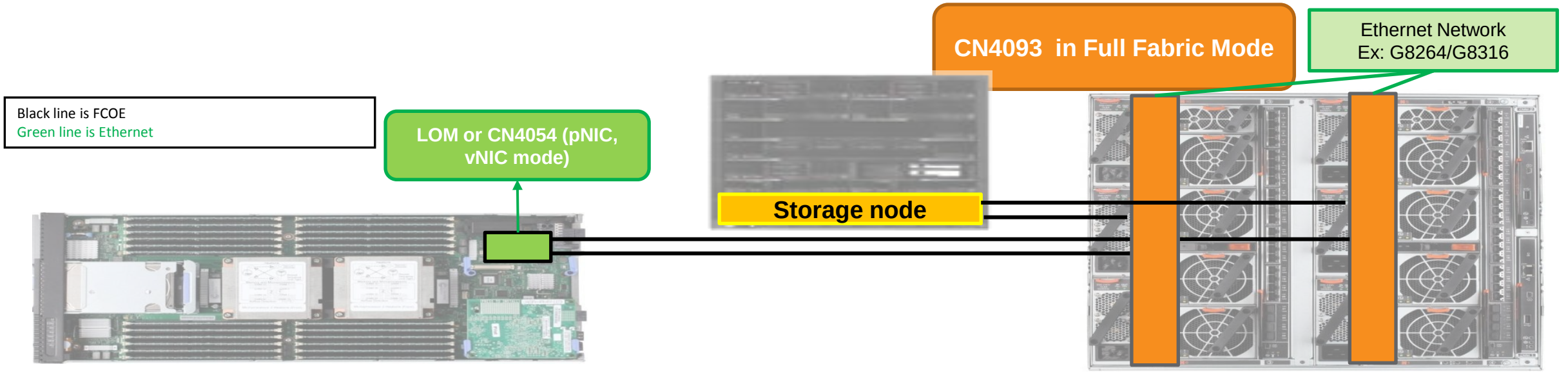
IBM End-to-End FCoE Configuration



Target Availability June 2013

Adapter	NIC Config	Transit	FCoE Switch	SAN Switch	Storage Target	OS levels
Emulex VFA II/III (Adapter + FOD Key) 49Y7950 + 49Y4274 49Y7940 + 49Y4274 95Y3751 (Included) 90Y6456 + 90Y5178 95Y3762 + 95Y3760 88Y6454 + 95Y3760	pNIC vNIC1 vNIC2	G8264	G8264CS	None	FCoE: Storwize V7K, V3700, SVC	WS2012, ESX 5, RHEL 6, SLES 11

Flexsystem -- CN4093 in Full Fabric FC/FCoE Switch (End-to-End FCoE)



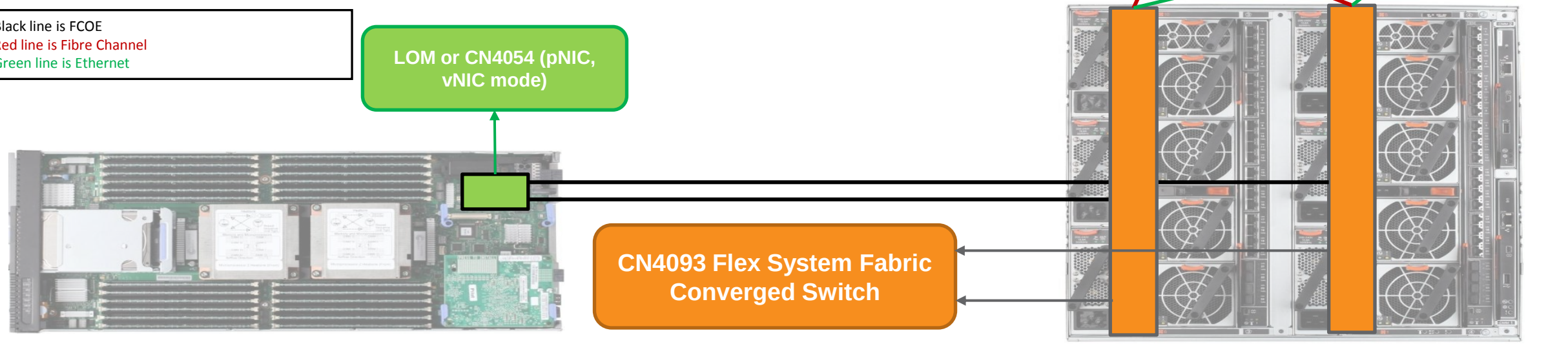
Full Fabric FC/FCoE Switch

- Services Login Request FLOGIs
  - Consumes a Domain ID
- Provides FC Function and Services
  - Zoning, Name Server, RSCN (Registered State Change Notification)
- Forwarding based on FC Domain IDs
- FC/FCoE Encap/Decap (optional)

Flexsystem -- CN4093 in NPV Gateway

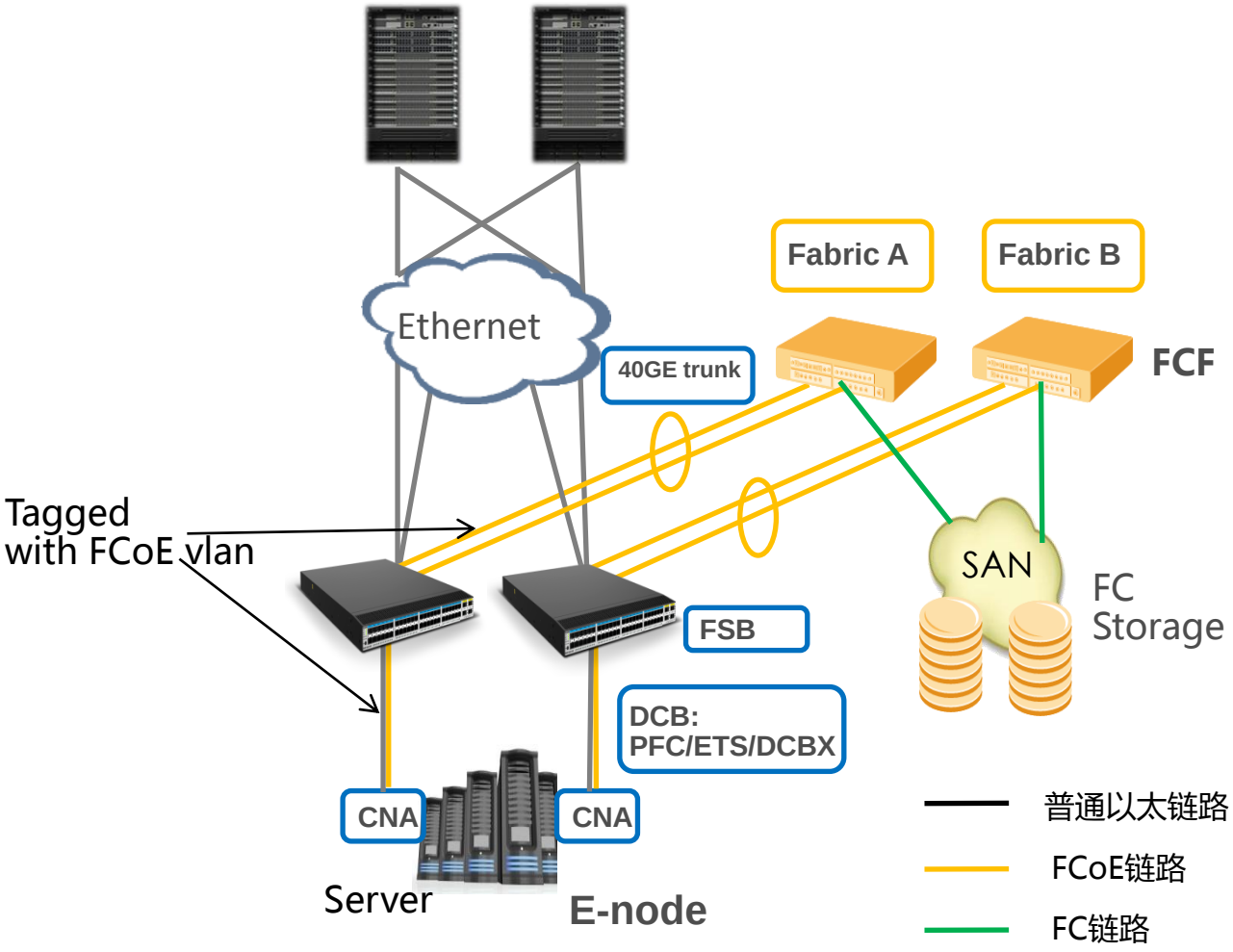
Key value:
- Acquisition costs
  - Reduce number of director class SAN switch ports
- Leverage existing external SAN infrastructure

Black line is FCOE
Red line is Fibre Channel
Green line is Ethernet



Adapter	Integrated Switch	FCoE Top Of Rack Switch	SAN Switch	Storage Target	OS levels
LOM & CN4054 4-port adapter (BE3) pNIC and vNIC modes	CN4093 10Gb Switch NPIV mode	N/A	Cisco SAN Brocade SAN	Storwise V7K, FC: SVC, DS3K/5K, DS8K, Tape, XIV	Win2008, ESX 4/5, RHEL 5/6, SLES 10/11

接入存储融合的FCoE组网



无丢包

- 部署DCB（利用PFC/ETS/DCBX等技术）
- 提高FCoE存储业务的传输优先级，保证在融合链路中存储业务的零丢包

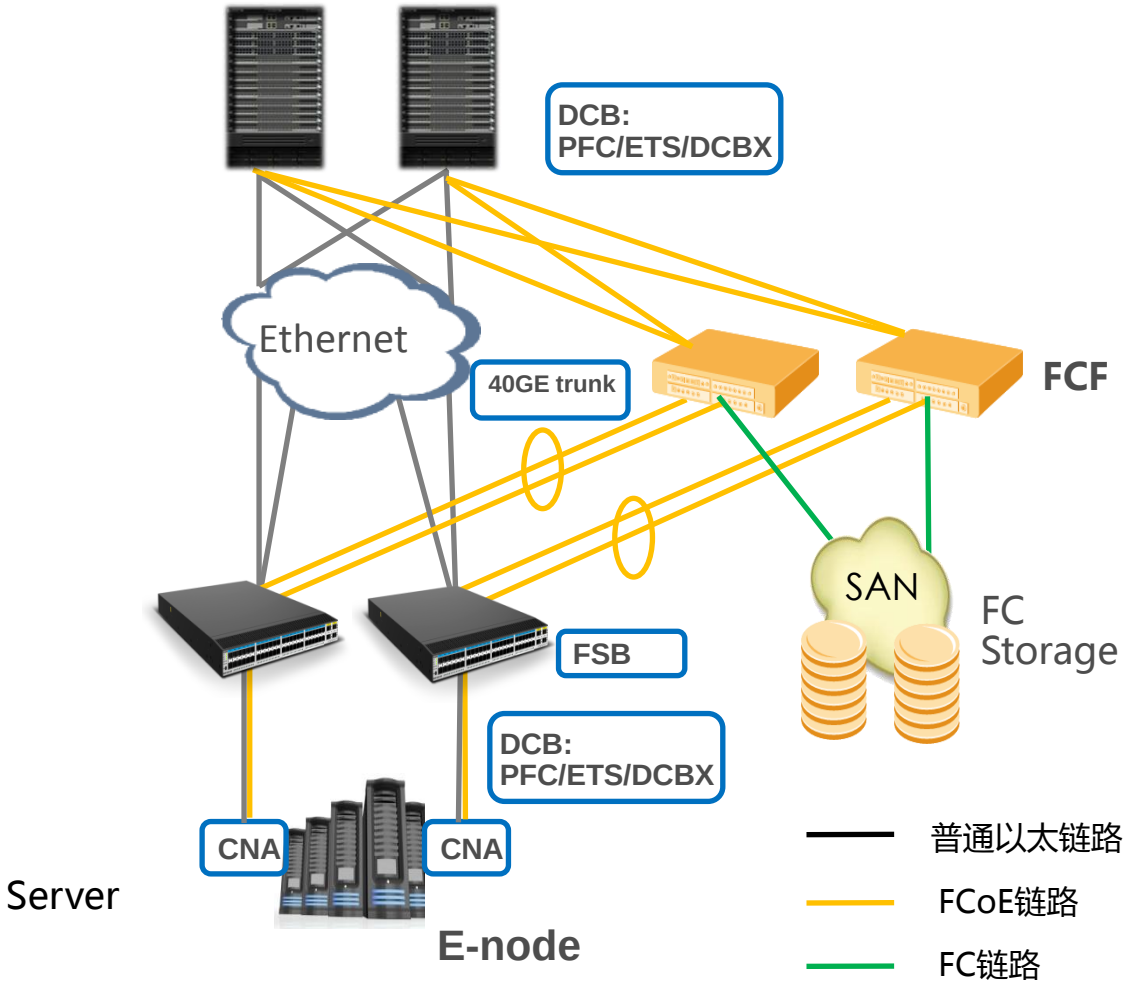
高安全

- 监听FIP注册过程,根据FIP Snooping的结果,拒绝非法用户访问存储区

高可靠

- FSB与FCF配置trunk链路，排除以太单链路故障对FCoE业务的影响
- FSB配合传统存储的双平面拓扑设计，服务器通过双网卡接入到不同的TOR交换机（FSB），并最终接入到不同的Fabric平面

网络融合方案的演进方向



阶段1：接入存储融合

- TOR支持FCoE/FC接口和DCB无丢包以太网
- 存储流量在TOR直接分流到SAN存储网络

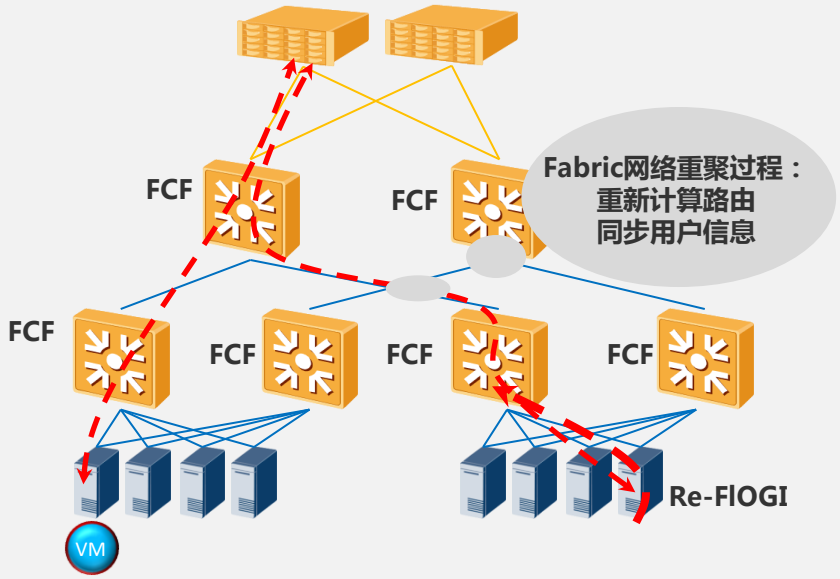
演进

阶段2：核心存储融合

- TOR/核心都支持FCoE/FC接口和DCB无丢包以太网
- 存储流量在核心分流到SAN存储网络

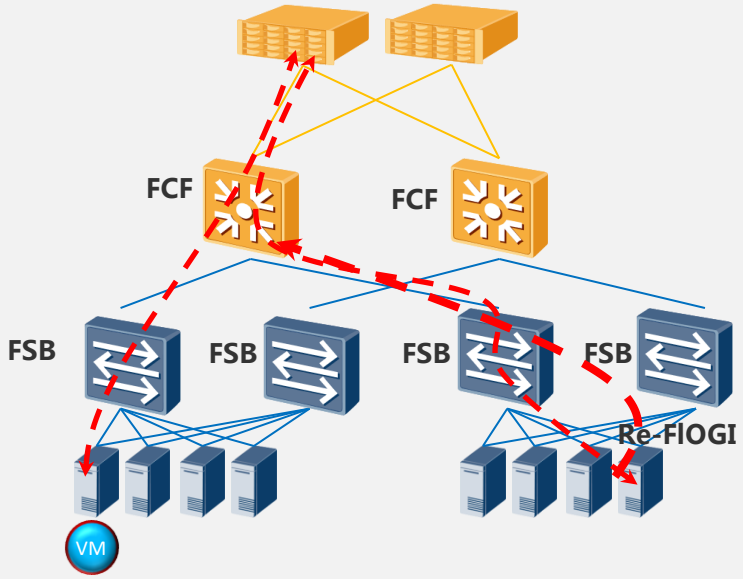
分布式和集中式FCoE网关

分布式FCoE网关部署方案



- VM跨TOR热迁移：VM其接入的FCoE网关发生了变化，重新注册（Re-FILOGI），FCID地址变更，存储业务挂起,，业务中断，待Fabric网络重聚后存储业务恢复。

集中式FCoE网关部署方案



- VM跨TOR热迁移：迁移前后，FCoE网关没有改变，重新注册（Re-FILOGI），保留FCID地址，Fabric网络无需重聚，存储业务快速上线。

Song Zhongru
Technical Support Engineer

+86(0)18926544460
supers@cn.ibm.com

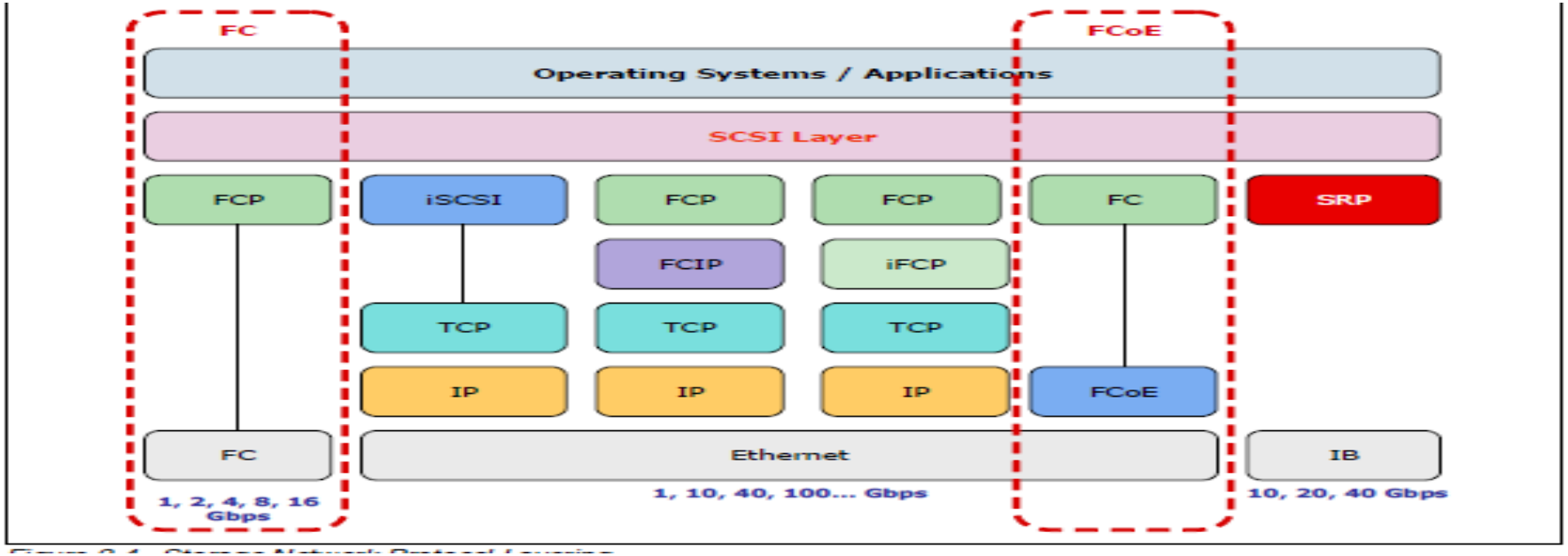
Thank You!

FSB, NPV and Full Fabric FC/FCoE Switch

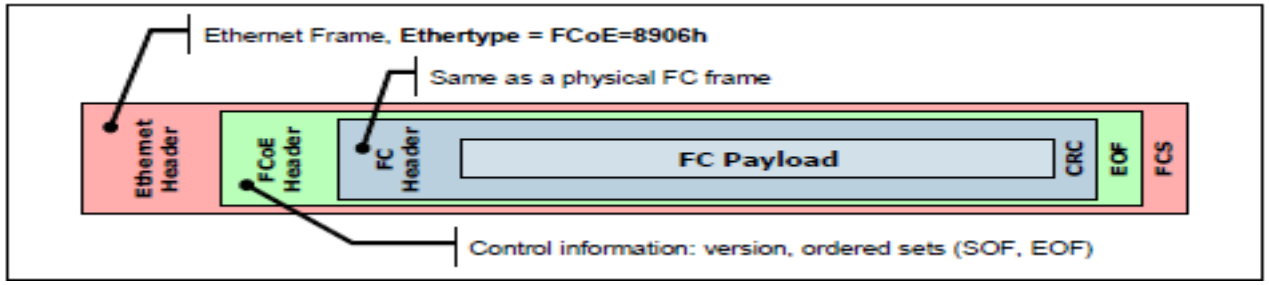
Features	FSB	NPV Gateway	Full Fabric FC/FCoE Switch
FCoE End Device Connectivity	√	√ (F)	√ (F)
FC Connectivity		√ (NP) and (NP-F)*	√ (F)*
FCoE/FC Gateway (Encap/Decap)		√	√*
Security	√ (FIP ACL)	√ (FIP ACL)	√ (FIP ACL, Zoning)
Zoning			√
S_FCID/D_FCID Routing			√

* == Not available in Nov. 2012

FCoE Network Layering and FCoE Frame



Based on FC Model



No FC Frame Fragmentation; Jumbo frames

Source: IBM Redbooks
http://www.redbooks.ibm.com/redbooks/pdfs/sg247986.pdf