Die wunderbare Welt auf Schicht 2

GUUG Hamburg

Dipl.-Ing. A. la Quiante
TSA, Public Sector Operation,
Germany
alaquian@cisco.com
7817

Agenda

TOI := Transfer of Information

Erklärung der Technik
(Abkürzungen und Begriffe)

Grundlage, Voraussetzung und Motivation

Unified I/O und Unified Fabric
DCE, CEE und DCB und die Standards

Quo Vadis Layer 2
VSS, vPC, FabricPath und TRILL

ca. 60-90 Minuten

Motivation I

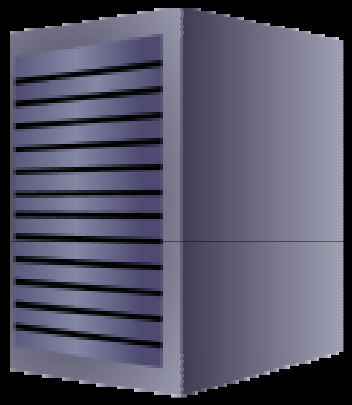
Server
Anwendung
Speicher

Verbindung

Fat-Client

Ist-Stand (vielfach)
Dezentrale Systeme
100M/1G NIC + 2xHBA

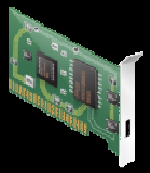
Motivation II



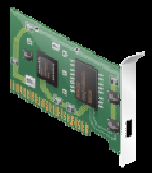
NIC 1 für LAN (VLAN 42 := Produktion)

NIC 2 für LAN (VLAN 42 := Produktion)

HBA 1 für SAN (VSAN 33 := Produktion)



HBA 2 für SAN (VSAN 33 := Produktion)

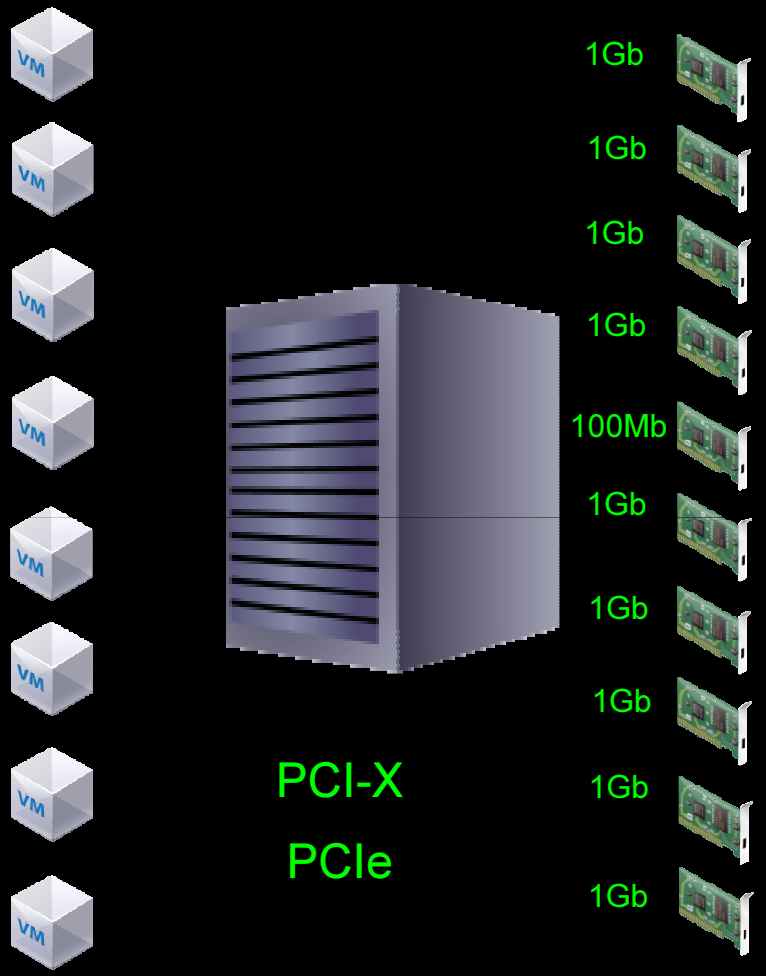


Managementzugang

VLAN := virtual LAN
VSAN := virtual SAN

Realer Server,
der nächste Schritt: Virtualisierung

Motivation III



NIC 1 für LAN (VLAN 42 := Produktion)

NIC 2 für LAN (VLAN 42 := Produktion)

HBA 1 für SAN (VSAN 33 := Produktion)

HBA 2 für SAN (VSAN 33 := Produktion)

Managementzugang

NIC 3 für VMkernel

NIC 4 für VMkernel

NIC 5 für Service Console

NIC 6 für Service Console

NIC 7, ... für LAN (VLAN 43 ... n)

Bei einem Verhältnis von hier 8:1 und ehemals 550 Mbit/s pro Server benötigen wir 4.4 Gbit/s für die Uplinks der VM-Uplinks (Produktion)

Eine vmnic pro VM?

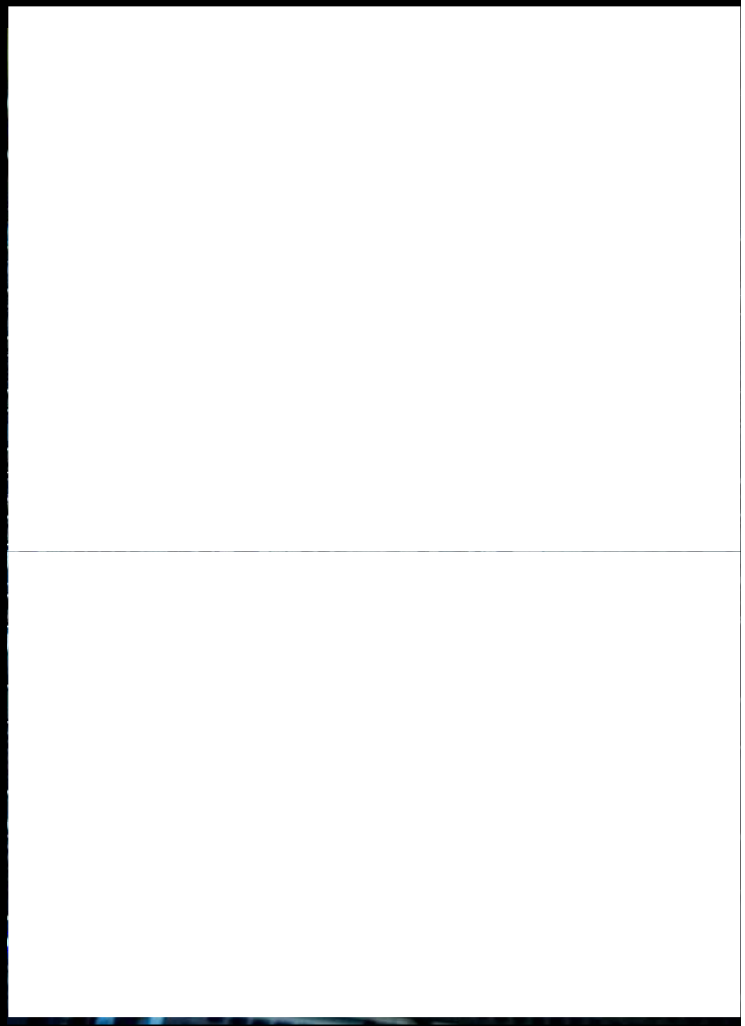
The Case for 10GE to the Server

Multi-Core CPU architectures allowing bigger and multiple workloads on the same machine

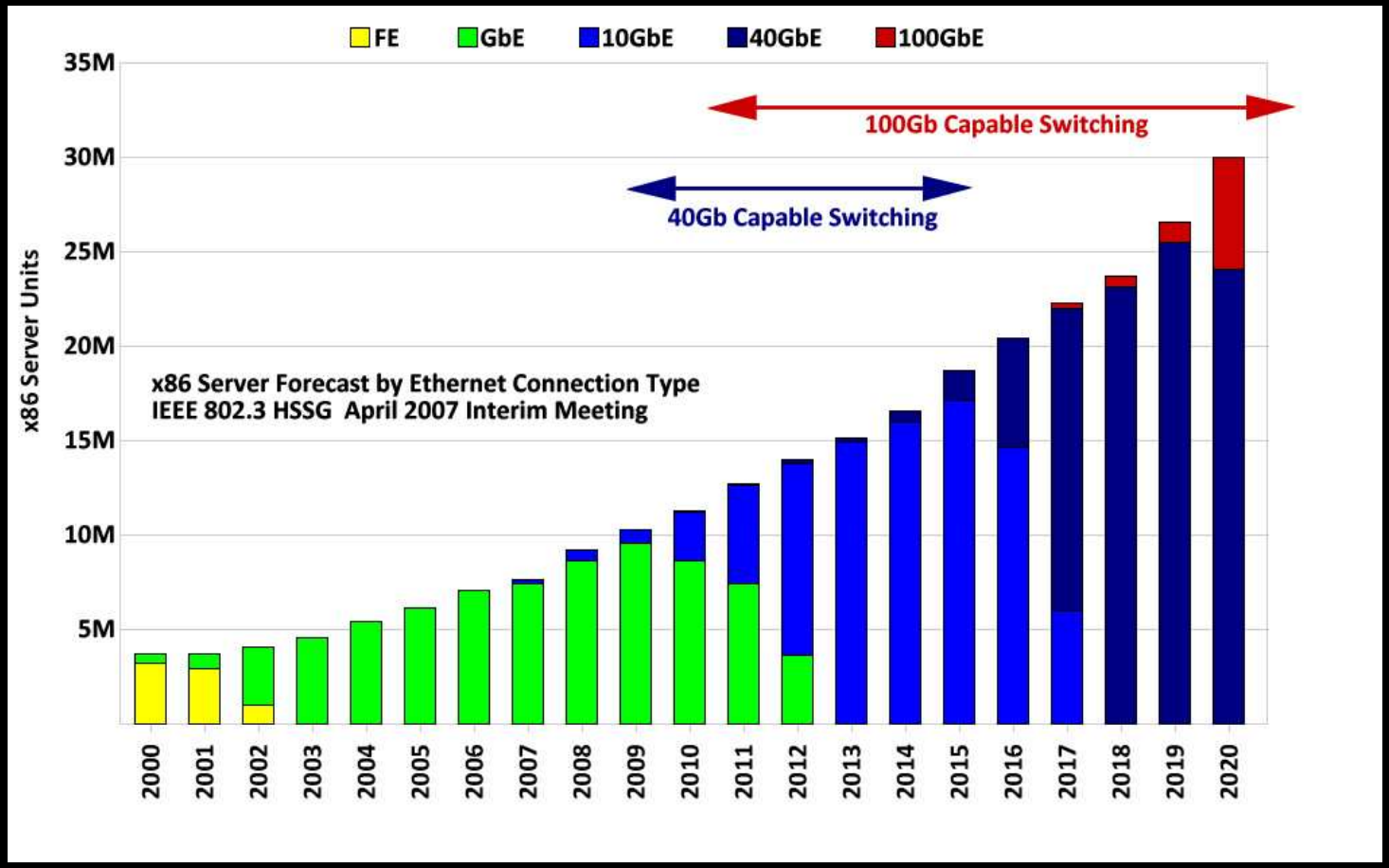
Server virtualization driving the need for more I/O bandwidth per server

Growing need for network storage driving the demand for higher network bandwidth to the server

10GE LAN on server Motherboards (LoM) beginning mid-2008 *(source: Broadcom)*



10/40/100 Gigabit Ethernet



Prognose vom Frühjahr 2007

Marketing und Standards

DCE := Data Center Ethernet

CEE := Converged Enhanced Ethernet

DCB := Data Center Bridging

DCB Data Center Bridging

Feature / Standard	Benefit
Priority Flow Control (PFC) **IEEE 802.1Qbb**	Enable multiple traffic types to share a common Ethernet link without interfering with each other
Bandwidth Management **IEEE 802.1Qaz**	Enable consistent management of QoS at the network level by providing consistent scheduling
Congestion Management **IEEE 802.1Qau**	End-to-end congestion management for L2 network
Data Center Bridging Exchange Protocol (**DCBX**)	Management protocol for enhanced Ethernet capabilities
L2 Multipath for Unicast and Multicast	Increase bandwidth, multiple active paths. No spanning tree (TRILL IETF, L2MP Cisco)

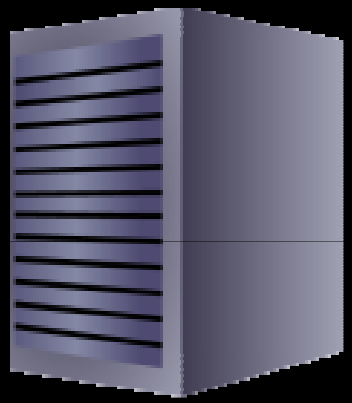
FCoE
IEEE 802.3Qbh

T11 FC-BB-05 Standard since 03-JUN-09
(VN-TAG)

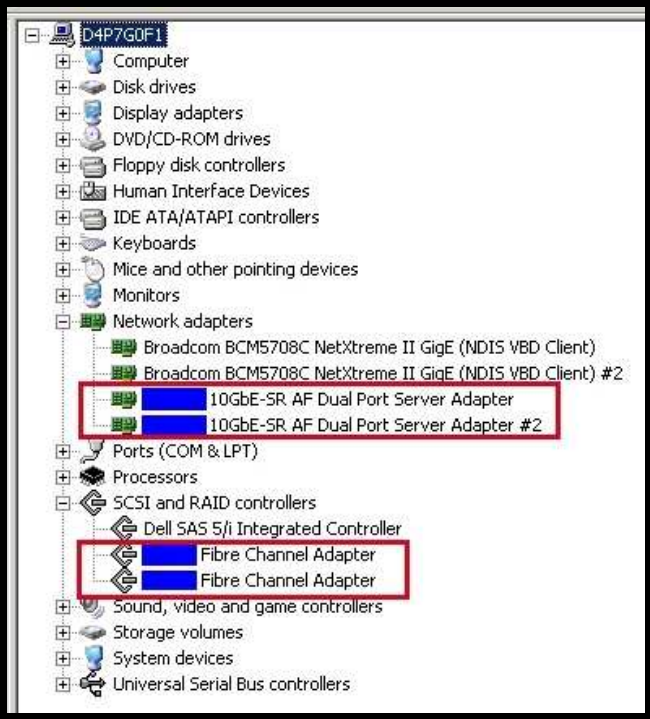
Adapter

NIC
HBA
CNA
VIC

Unified I/O



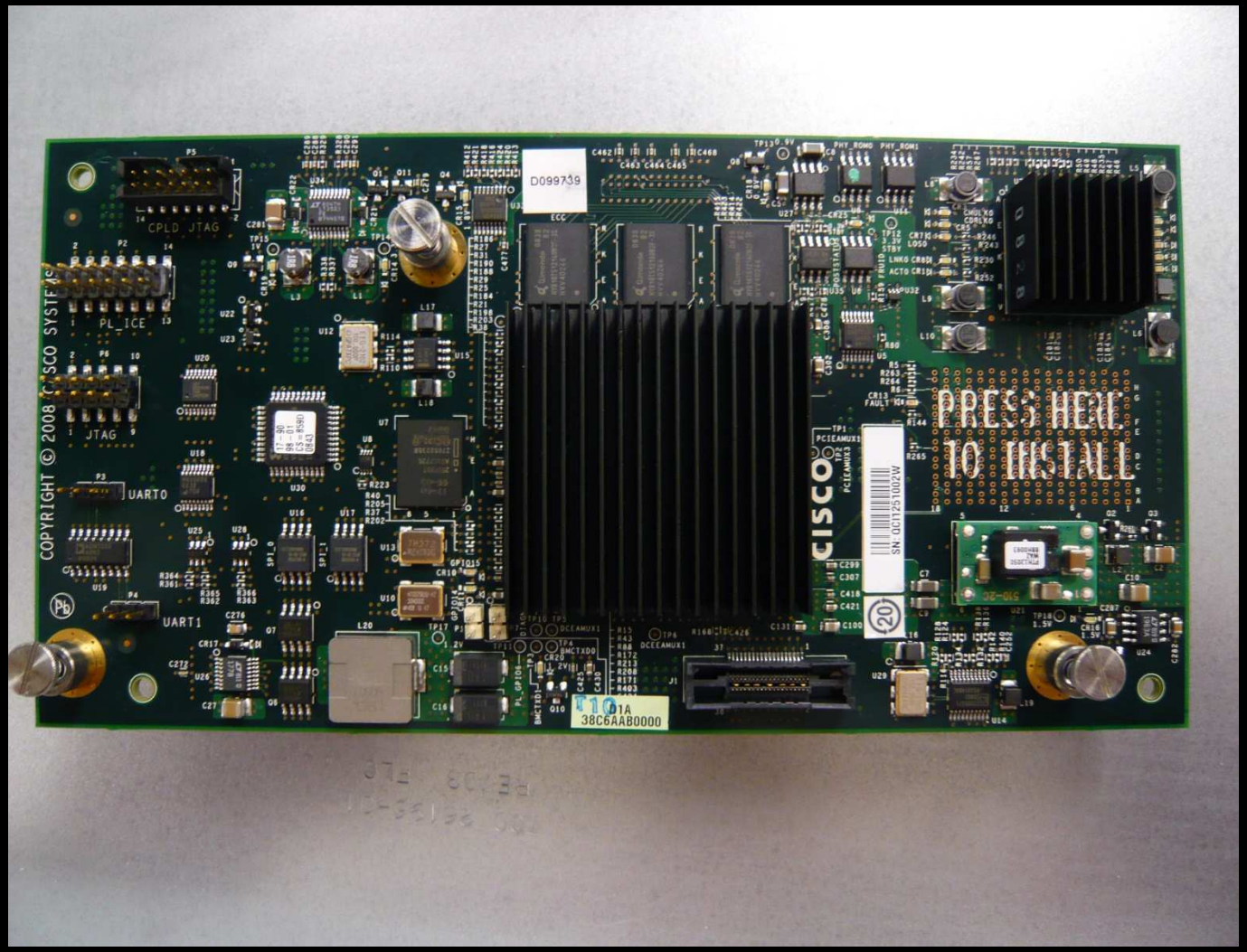
CNA



CNA := Converged Network Adapter

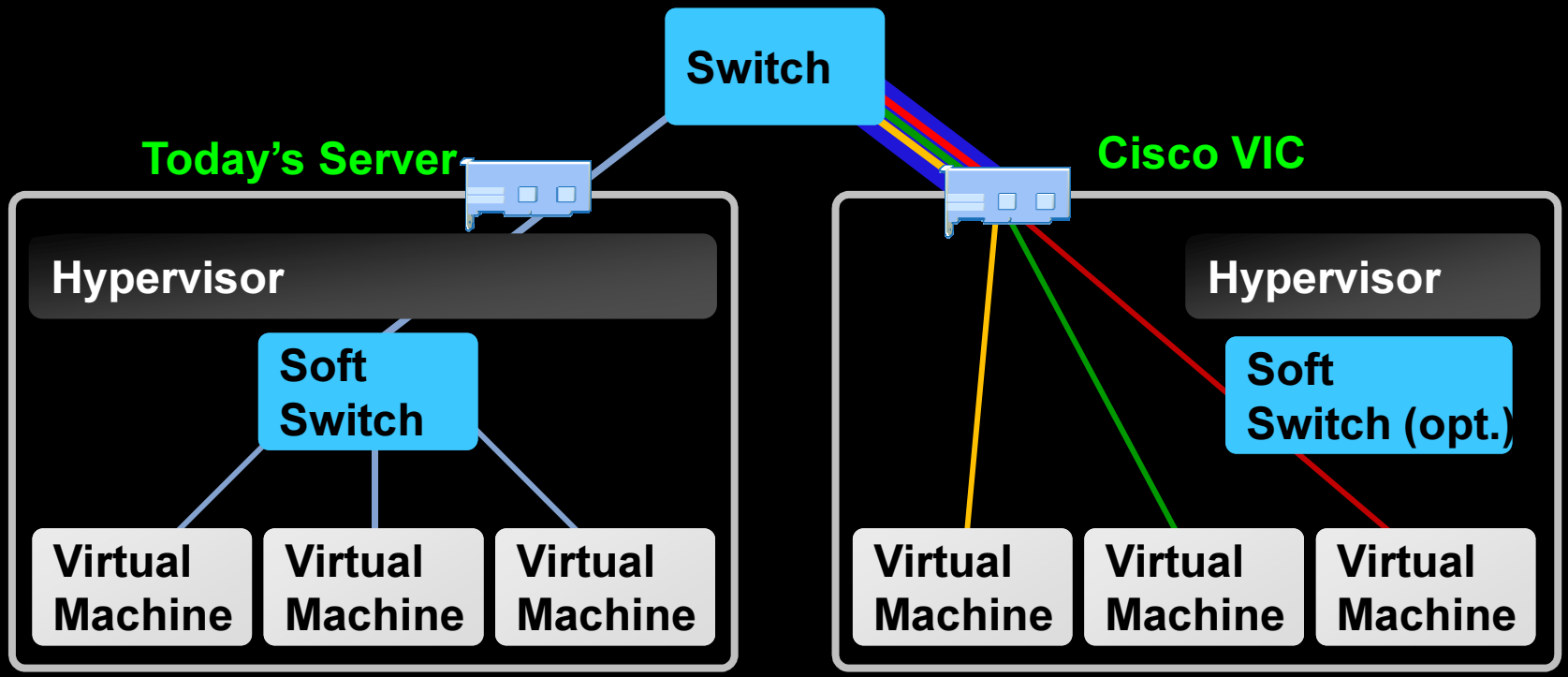
CNA := NIC + HBA

Unified I/O und Unified Fabric



VIC M81KR

Cisco Virtualized Adapter



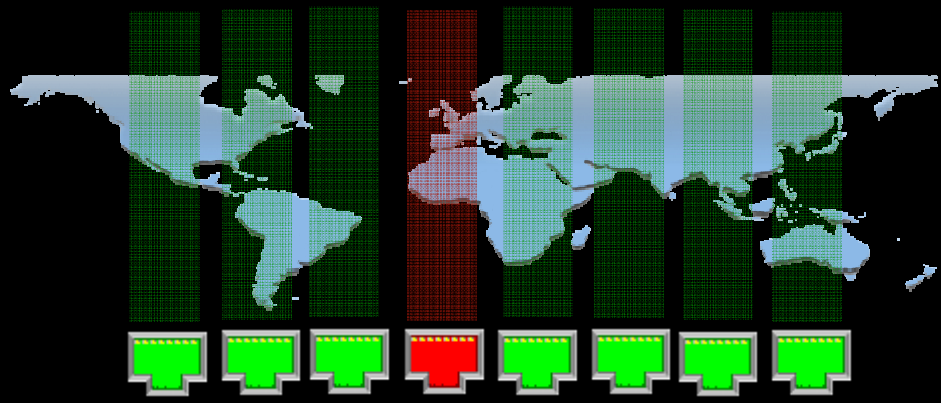
True wire once architecture – highly dynamic

Network policy and visibility brought to VMs

Hypervisor bypass support – increases performance

Unified I/O

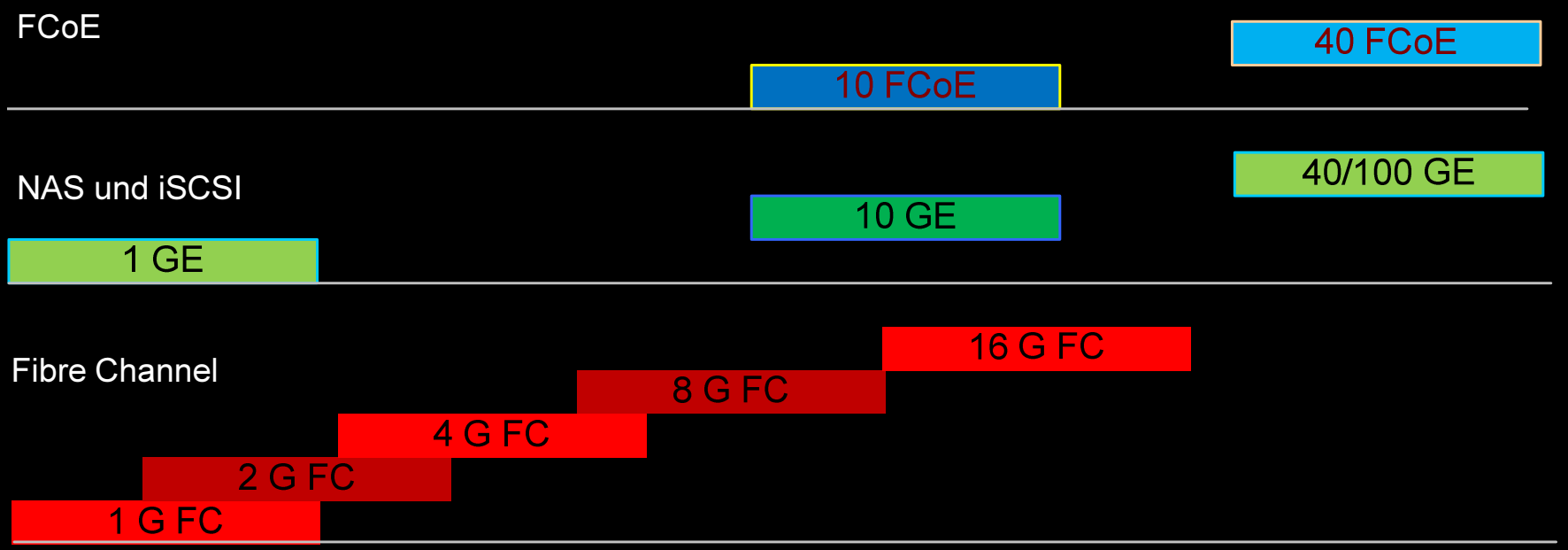
Strategie



Gedankenspiel: Weltweiter Absatz

FC 100.000 Ports

Ethernet 100.000.000 Ports

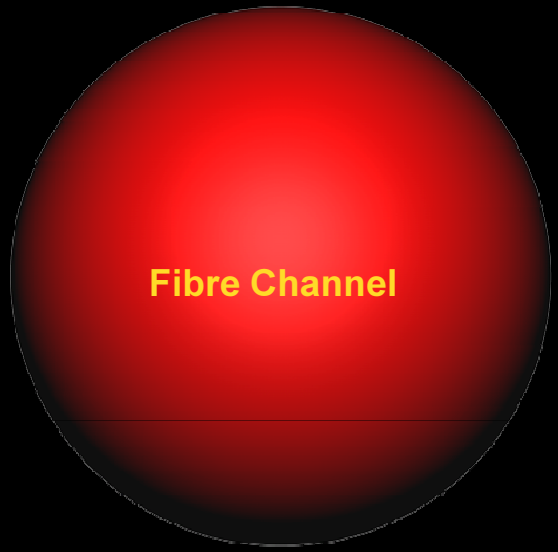


Welche Technologie wird zukünftig das bessere Preis/Leistungsverhältnis aufweisen?

Who is Who

T11 Group: Fibre Channel Standards

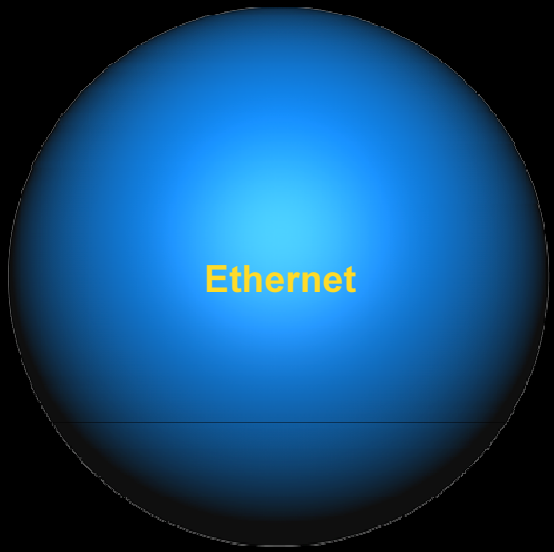
- Part of INCITS
- Has defined FC technologies for over 10 years; FCIA markets them
- Focuses on all things FC:
  - Physical Layer
  - Switching
  - Framing
  - Security
- Why it's important to FCoE:
  - Standardized method of transporting Fibre Channel frames over Ethernet
  - Standardized method for multi-hop FCoE



Who is Who

IEEE 802.1 Working Group:
LAN Bridging Standards

- Part of IEEE 802 (LAN and MAN committee)
- Defines LAN Bridging technologies
  - E.g., all about Ethernet switching
- The Data Center Bridging (DCB) Task Group is inside IEEE 802.1
  - DCB developed bridging extensions relevant for the Data Center environment
- Why it's important to FCoE:
  - Those bridging extensions enable I/O consolidation with FCoE



When Are Standards “Done?”

Standard is technically stable, a.k.a "Done," when it moves from Development to Approval phase

1. Investigation
2. Development
3. Approval
4. Publication

Status of the Standards

FC-BB-5

Technically stable in October, 2008

Completed in June 2009

Published in May, 2010

PFC

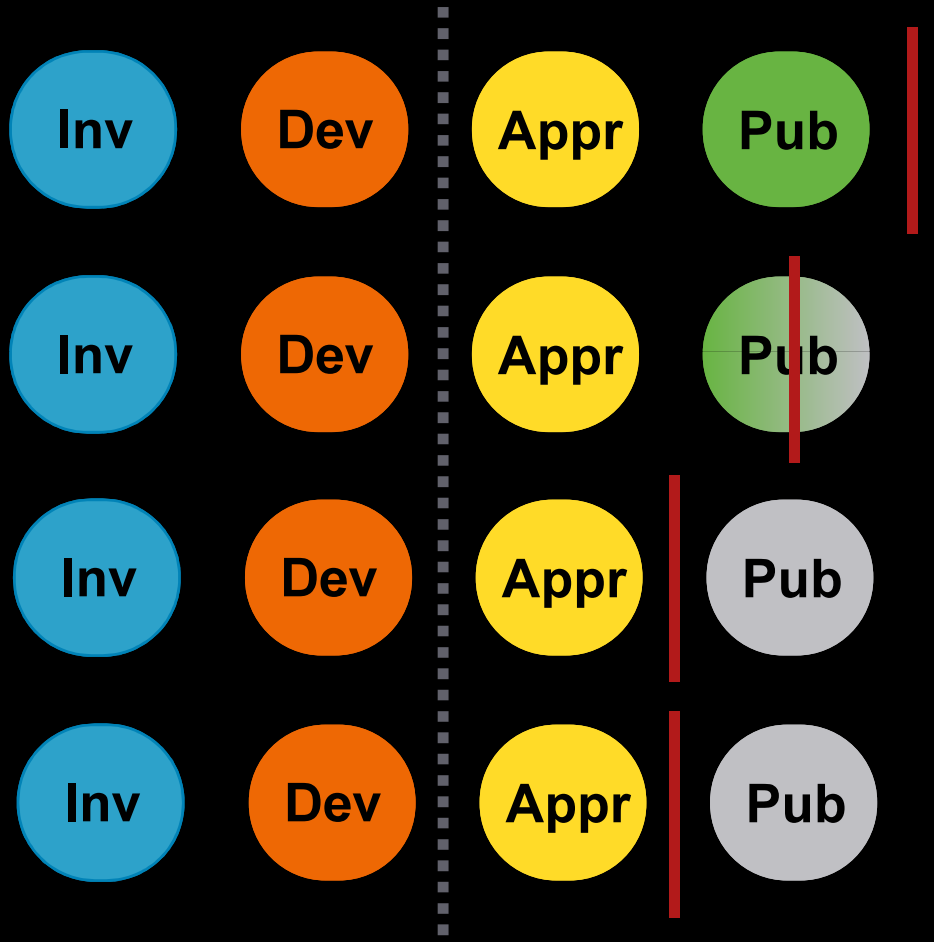
Completed in July 2010, awaiting publication

ETS

Completed in July 2010 (completing Approval Phase 3)

DCBX

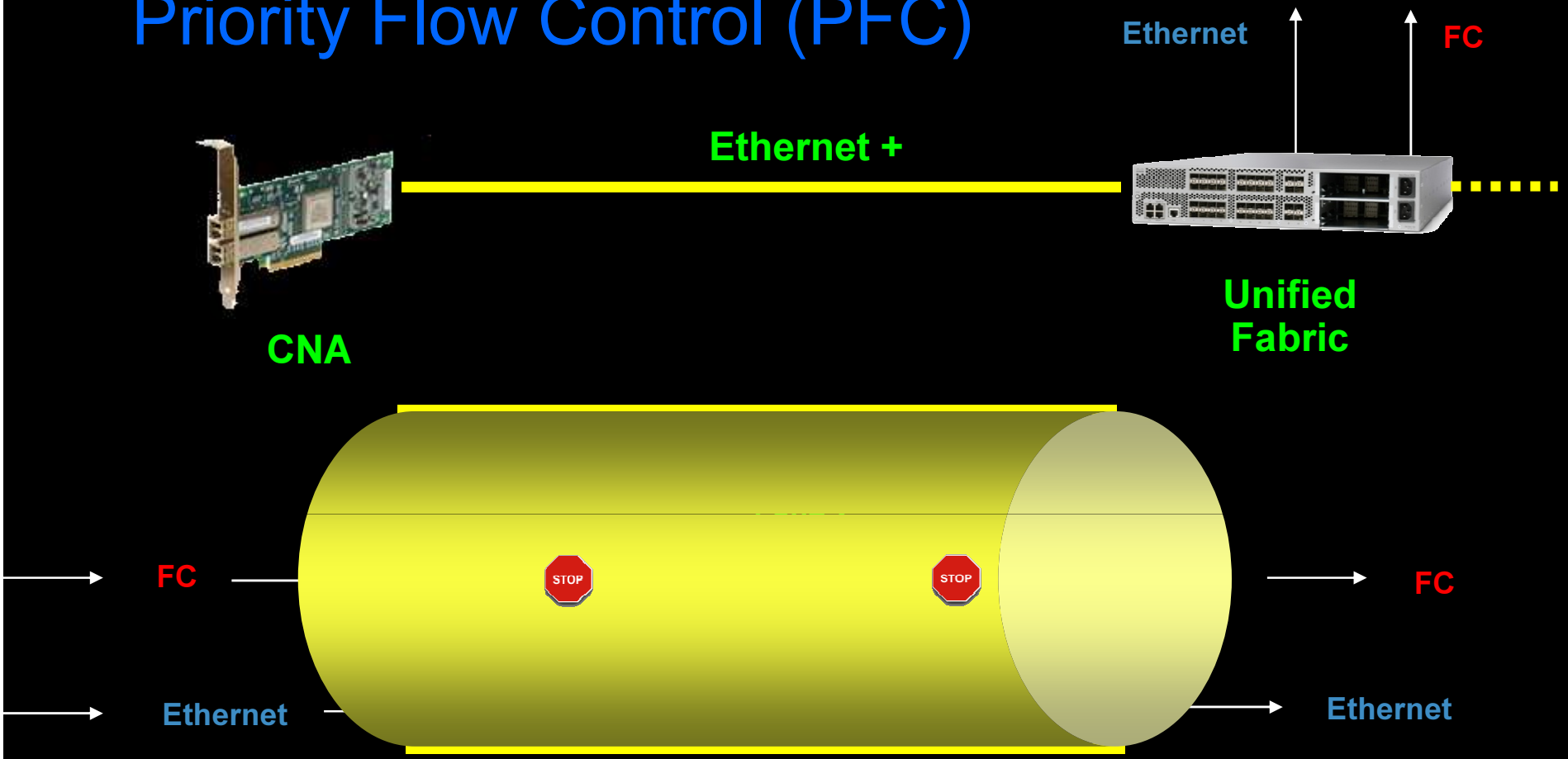
Completed in July 2010 (completing Approval Phase 3)



Priority Flow Control (PFC)

Priority Flow Control (PFC)
IEEE 802.1Qbb

Priority Flow Control (PFC)



DCB := Data Center Bridging

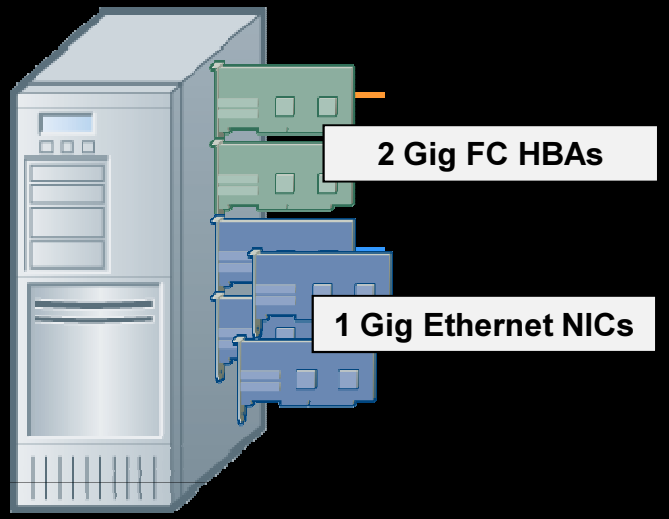
FCoE := Fibre Channel over Ethernet

Bandwidth Management

Bandwidth Management
IEEE 802.1Qaz

Enhanced Transmission Selection
Bandwidth Management

- Once feature fcoe is configured, 2 classes are made by default
- By default, each class is given 50% of the available bandwidth

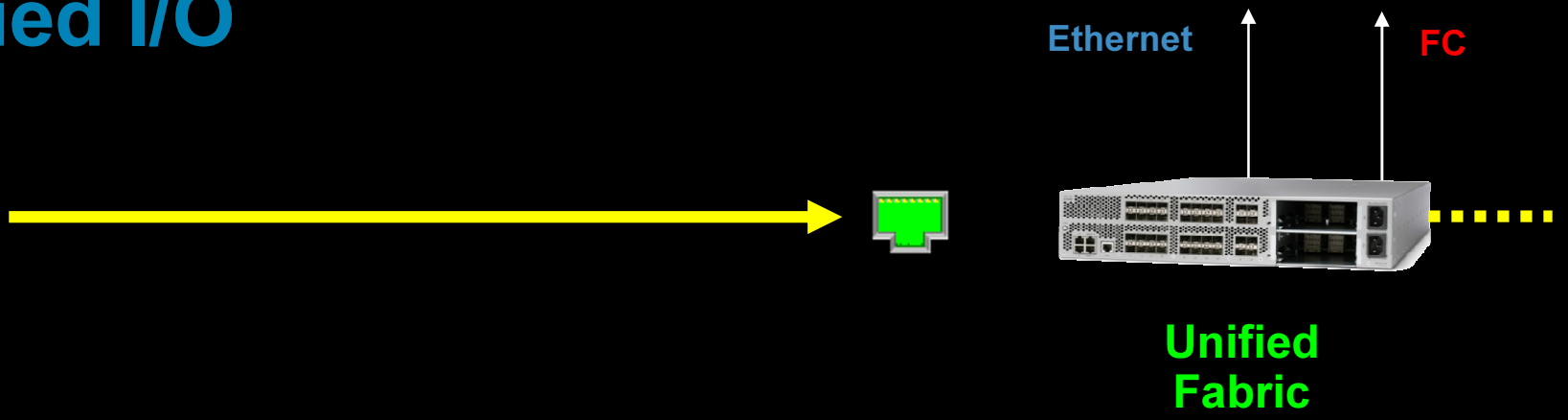



```
tme-n5k-2# show queuing interface eth 1/3

Interface Ethernet1/3 TX Queuing
qos-group  sched-type  oper-bandwidth
    0        WRR              50
    1        WRR              50
```

- Best Practice : FCoE and Ethernet each receive 50%
- Can be changed through QoS settings when higher demands for certain traffic exist (i.e. HPC traffic, more Ethernet NICs)

Unified I/O



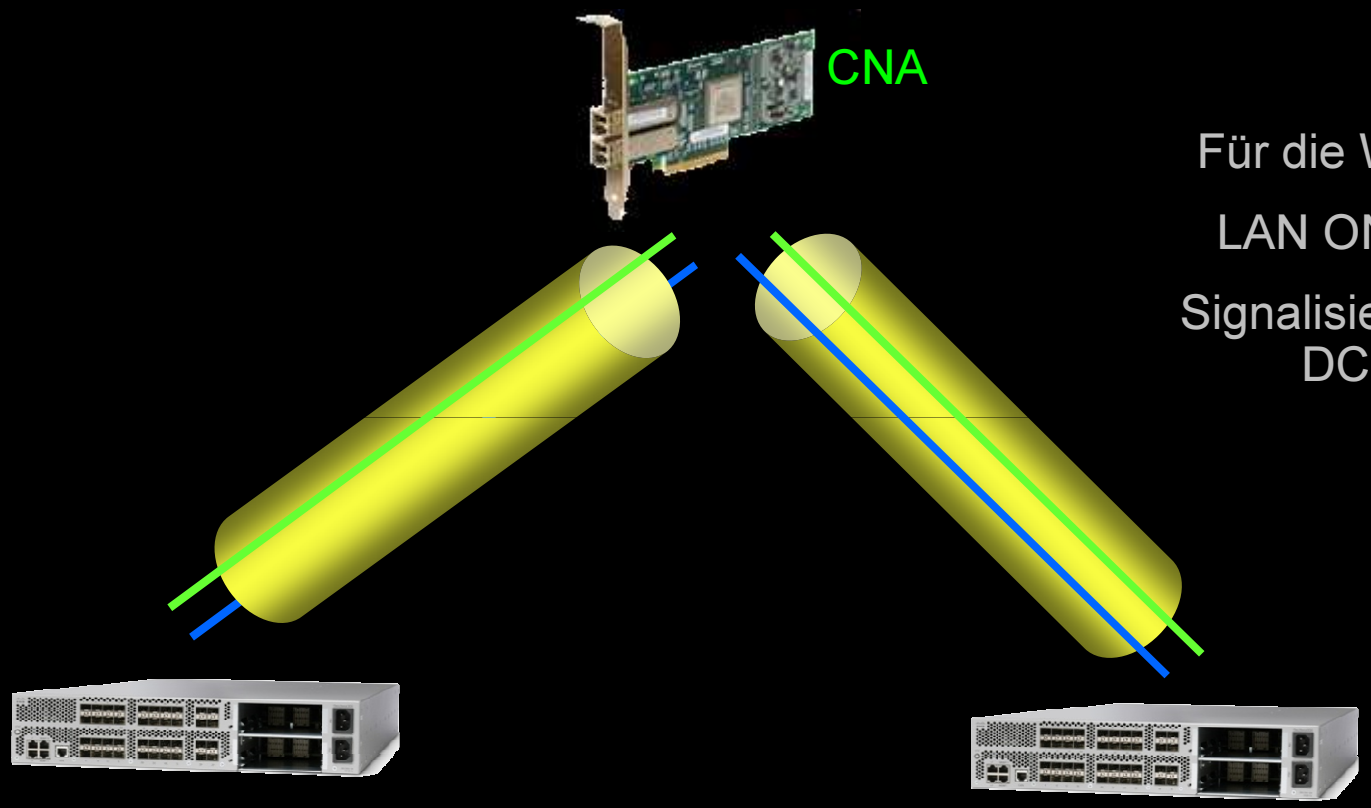

```
N5k(config)# policy-map type queuing policy-fcoe
N5k(config-pmap-que)# class type queuing class-default
N5k(config-pmap-c-que)# bandwidth percent 40
N5k(config-pmap-c-que)# class type queuing class-fcoe
N5k(config-pmap-c-que)# bandwidth percent 60
N5k(config-pmap-c-que)# system qos
N5k(config-sys-qos)# service-policy type queuing input policy-fcoe
```

input policy either on the interface the CNA is attached to or globally to the system qos

Data Center Bridging Exchange Protocol

Data Center Bridging Exchange Protocol
(**DCBX**)

DCBX

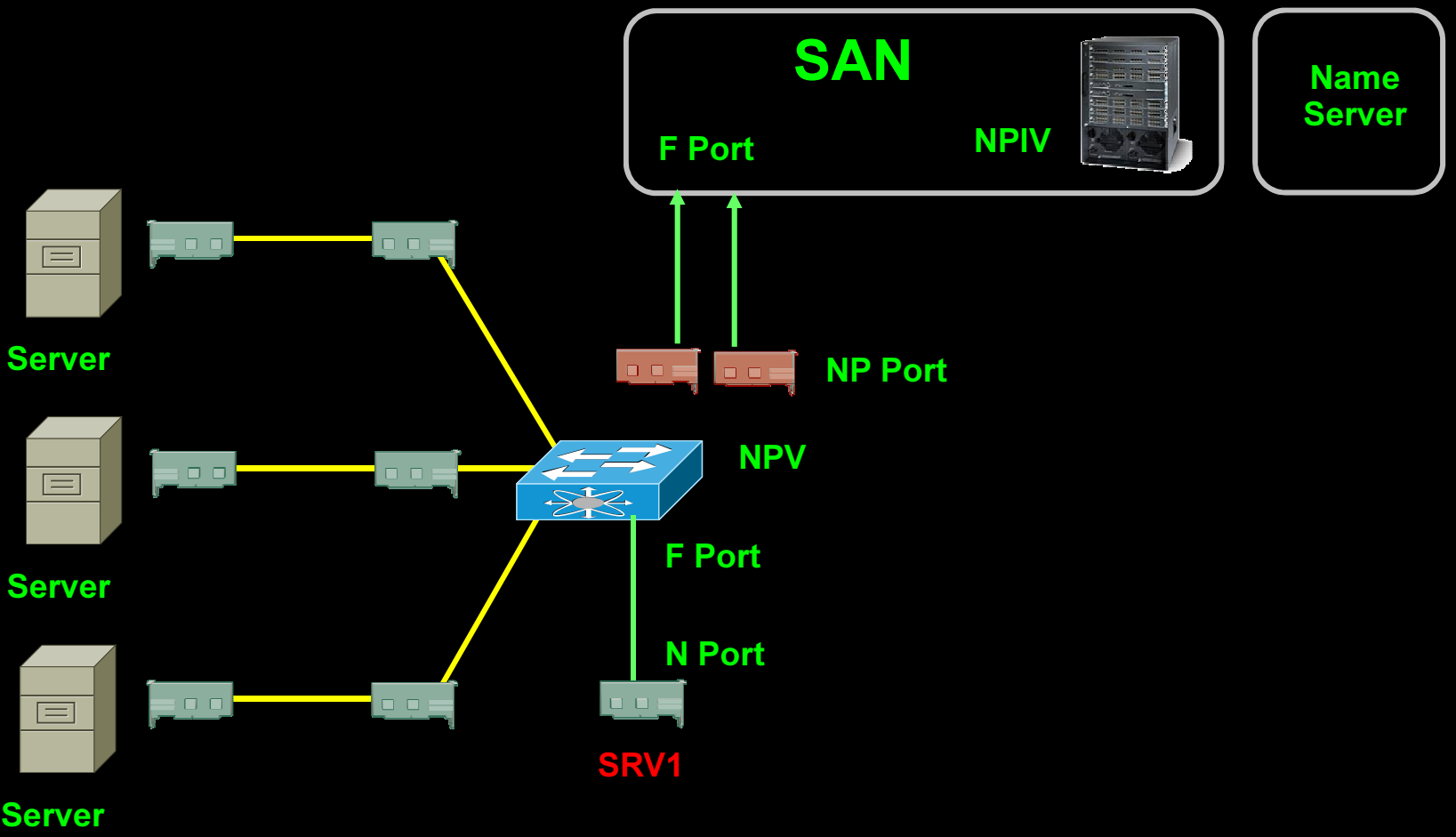


Für die Wartung

LAN ON/OFF:

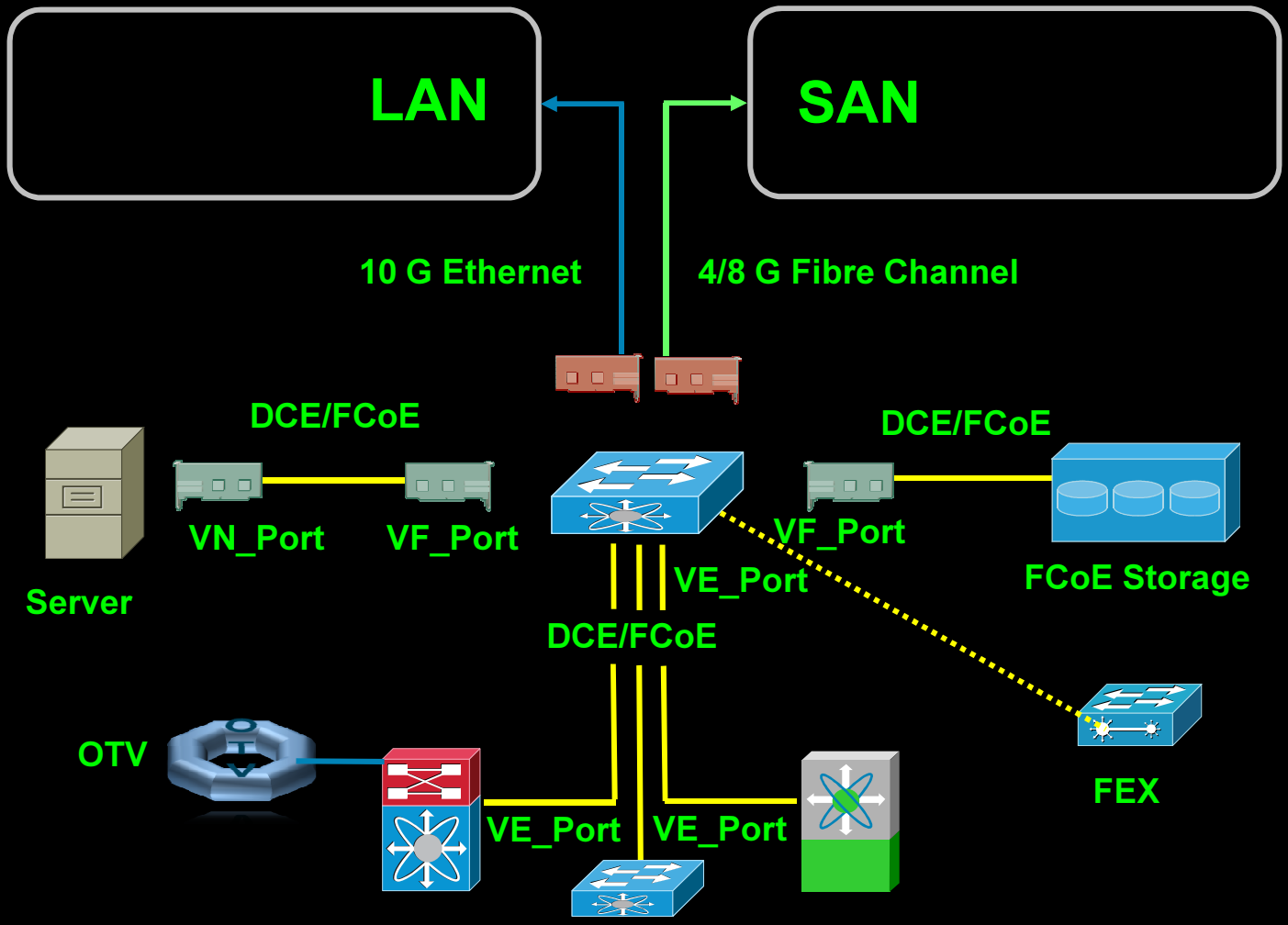
Signalisierung via DCBX

I/O Consolidation (NPV or FC Mode)



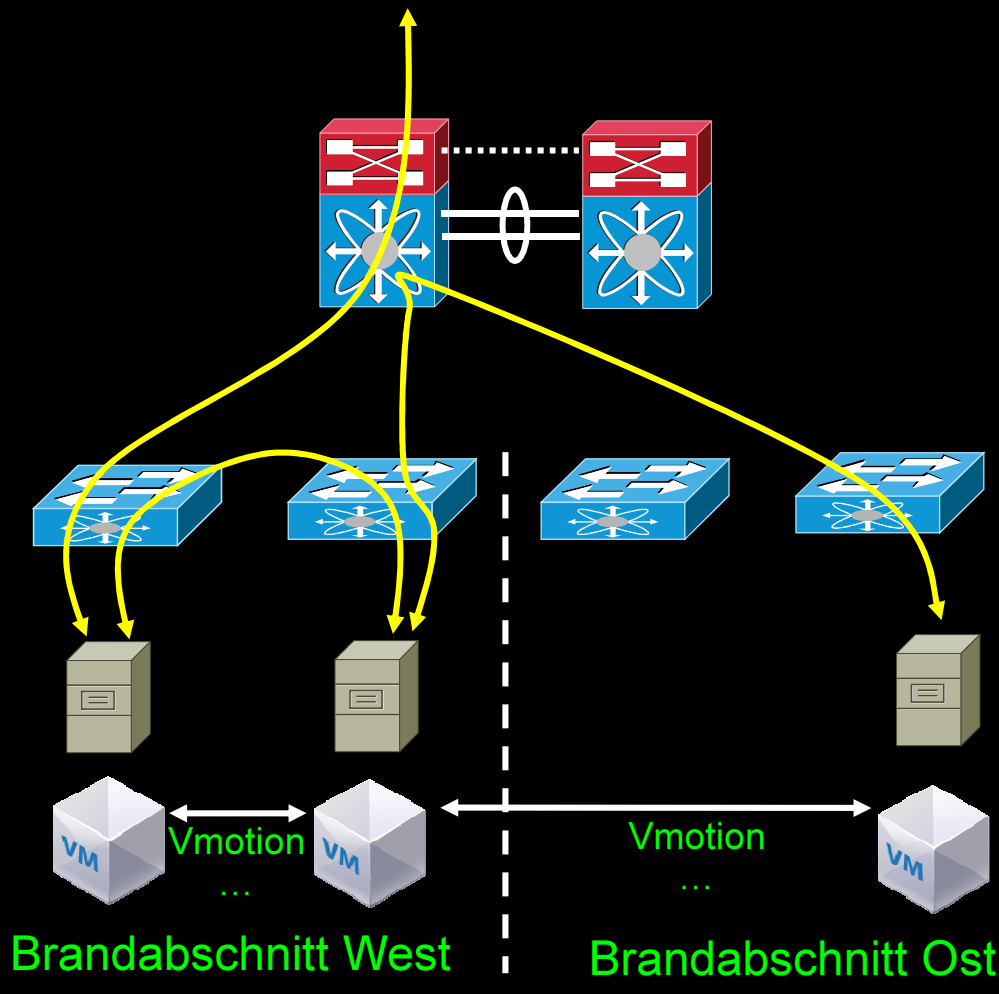
NPIV := N Port ID Virtualization (Fabric Device Feature)
NPV := N Port Virtualizer (Edge Device Feature)

Unified Fabric (VE Ports)



Schicht 2

Kommunikationsbeziehungen



Cluster, Vmotion & Co erfordern „wieder“ Layer 2 End-to-End

Order evtl. LISP

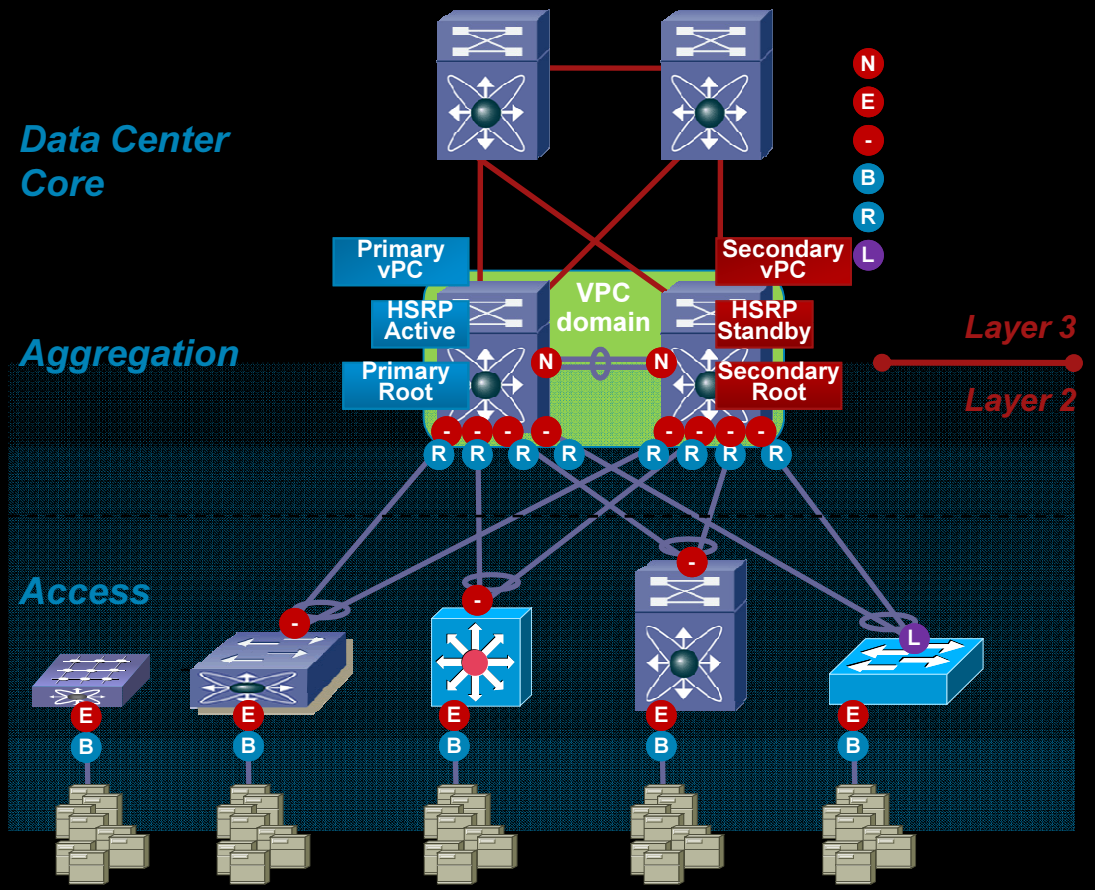
VSS
Catalyst 6500
vPC
Nexus 5000 & 7000

TRILL
(evtl. 1H2011?)
IETF WG
L2MP / FabricPath
(jetzt)
Cisco

Ehemals Silo-Architektur und Nord-Süd Kommunikation.
Zunehmend zusätzlich Ost-West Kommunikation

Constraints for Scaling Layer 2 Network

- Port Density on Switches
- Over-subscription Ratio
- Complex STP Configuration
- X-Chassis Port-Channel
- MAC Table Size



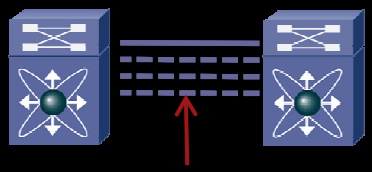
Existing Options for Layer 2 Expansion

More I/O slots

More ports per I/O podule

Higher Port-Density

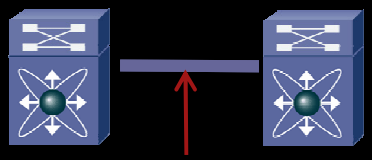
Wasted Bandwidth



STP only allows single active link between 2 devices

Multiple Inter-Switch Links

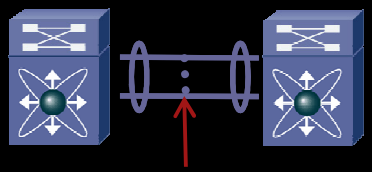
Higher Port Cost



Use interface with speed equals to the combination of multiple lower-speed links

Higher Interface Speed

Limited Scale

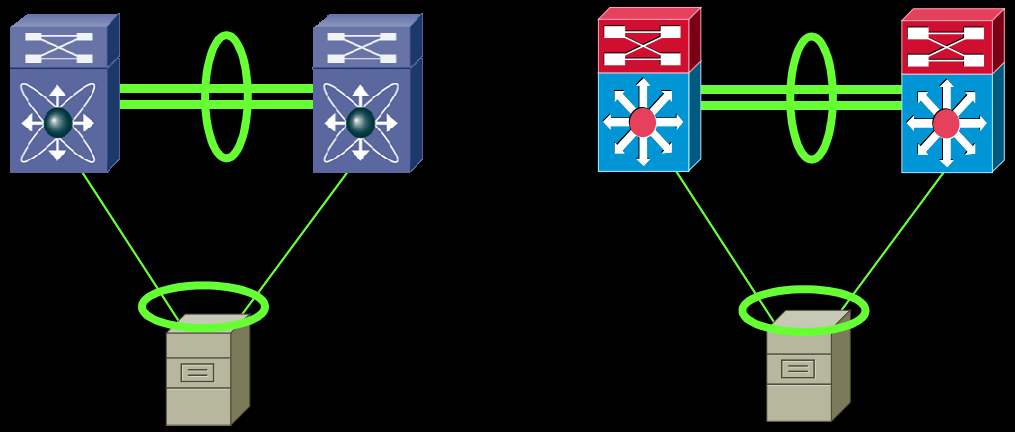


More ports in a bundle (up to 16-port today)

Port-Channel/Link-Aggregation

VSS und vPC

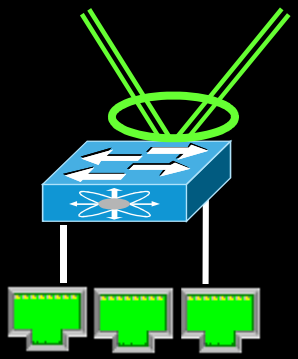
- VSS := Virtual Switching System
- vPC := Virtual Port Channel (EC)



- NIC
  Teaming: Active-Passive
  Teaming: TLB
  Teaming: EC

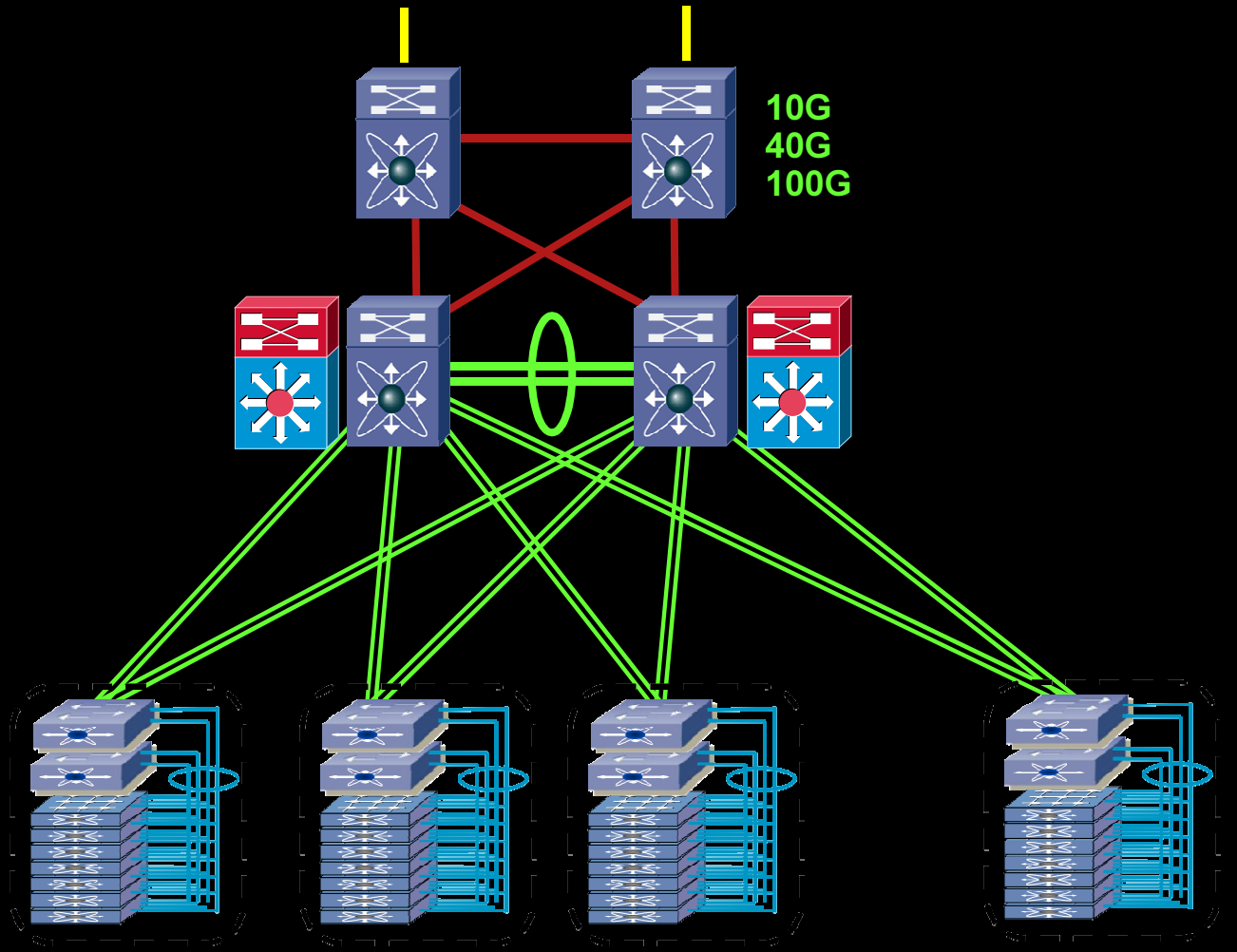
VSS & vPC

- Limitierung:
  zwei und genau
  zwei zentrale
  Systeme



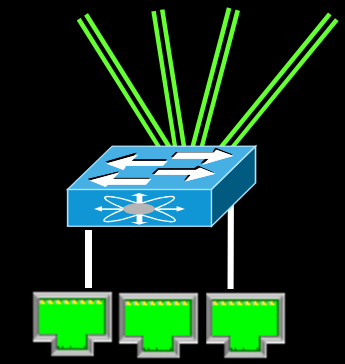
Port 1 … Port 24
Port 1 … Port 48

- Limitierung:
  8 oder 16 Member
  per EtherChannel

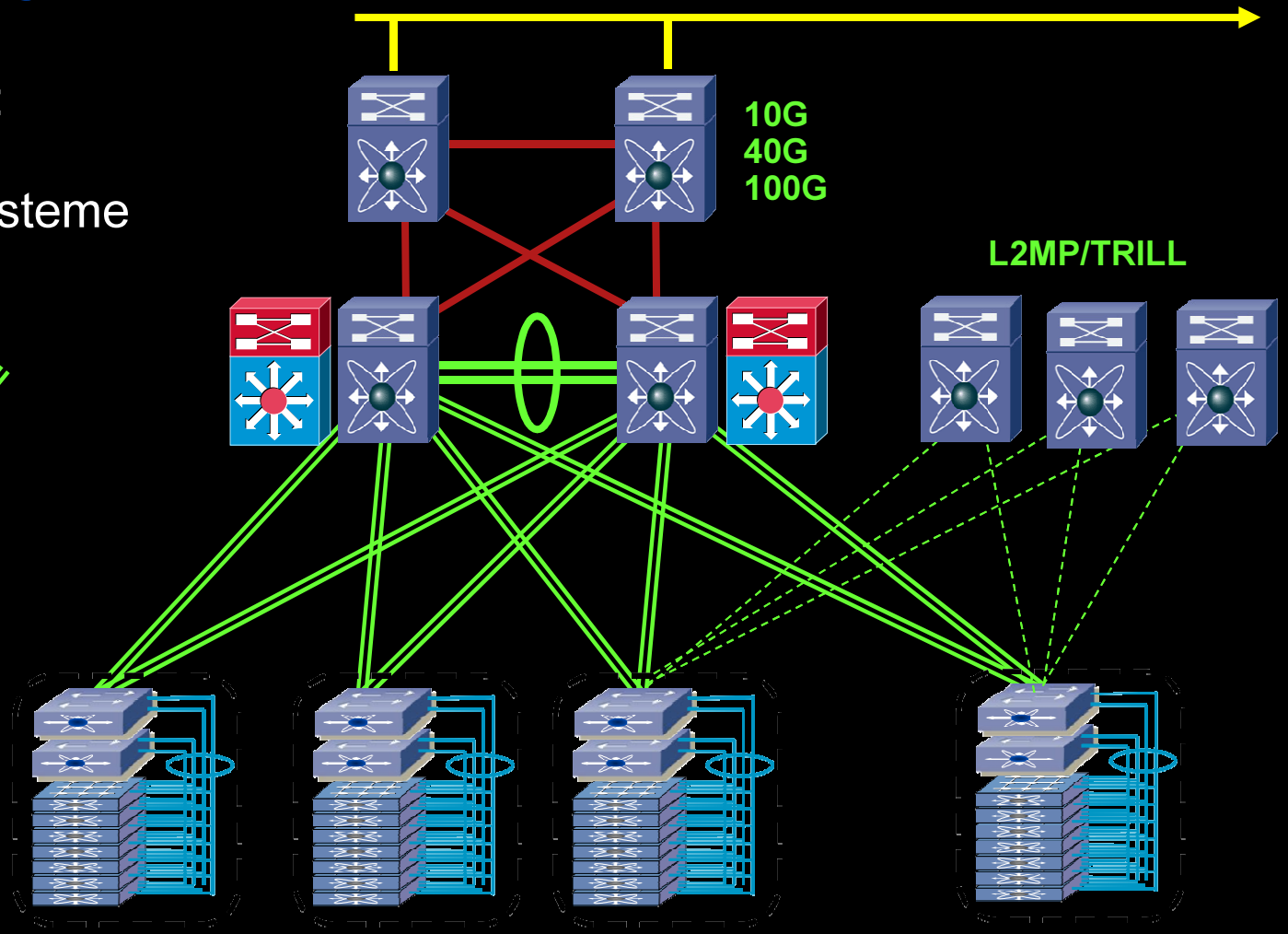


Skalierungsoption: FabricPath

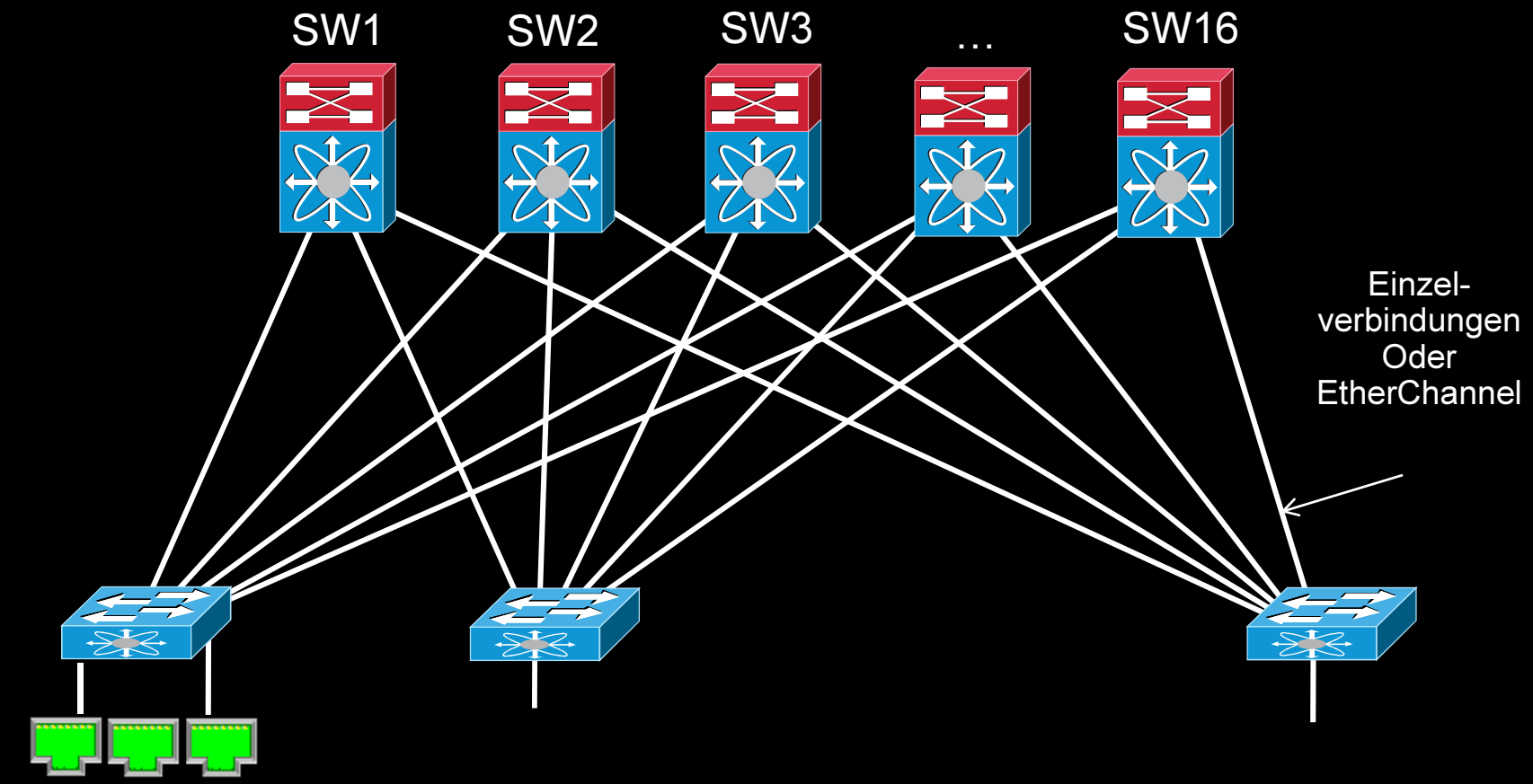
- Limitierung:
  zwei bis 16
  zentrale Systeme



Port 1 … Port 24
Port 1 … Port 48



Maximal skalierbare Schicht 2 Umgebungen



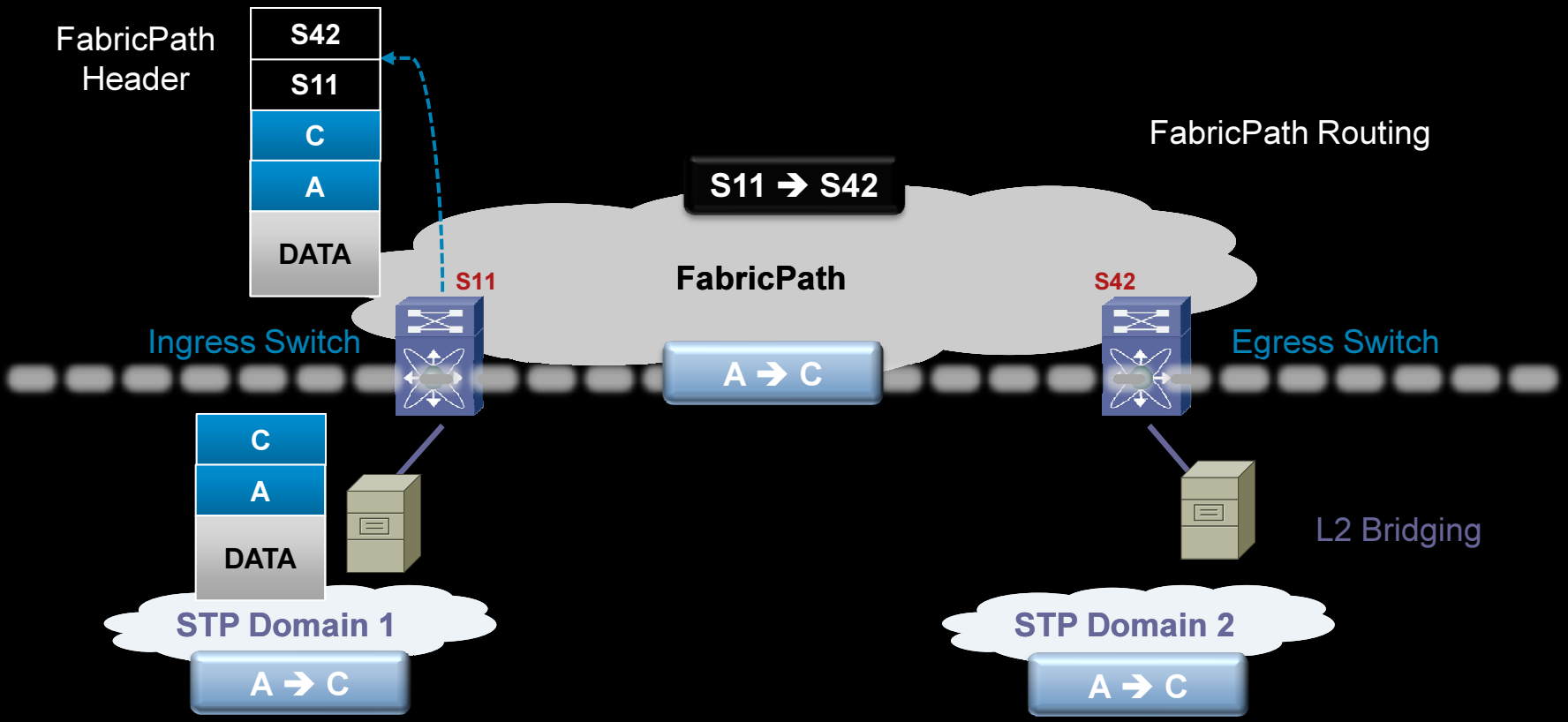
Aufgabe:

Mehr als 2 zentrale Systeme (Spines) aber schliefenfreiheit auf Schicht 2 ohne Spanning-Tree (STP)

Data Plane Operation

Encapsulation to creates hierarchical address scheme

- FabricPath header is imposed by ingress switch
- Ingress and egress switch addresses are used to make "Routing" decision
- No MAC learning required inside the L2 Fabric



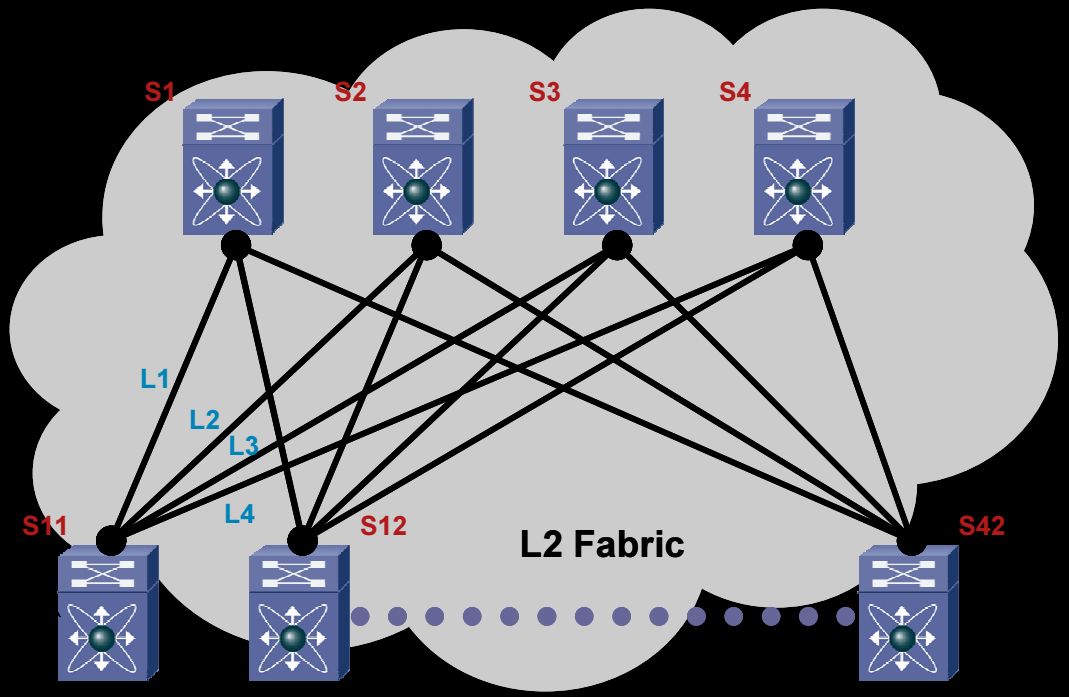
Control Plane Operation

Plug-N-Play L2 IS-IS is used to manage forwarding topology

- Assigned switch addresses to all FabricPath enabled switches automatically (no user configuration required)
- Compute shortest, pair-wise paths
- Support equal-cost paths between any FabricPath switch pairs

FabricPath Routing Table

Switch	IF
S1	L1
S2	L2
S3	L3
S4	L4
S12	L1, L2, L3, L4
…	…
S42	L1, L2, L3, L4



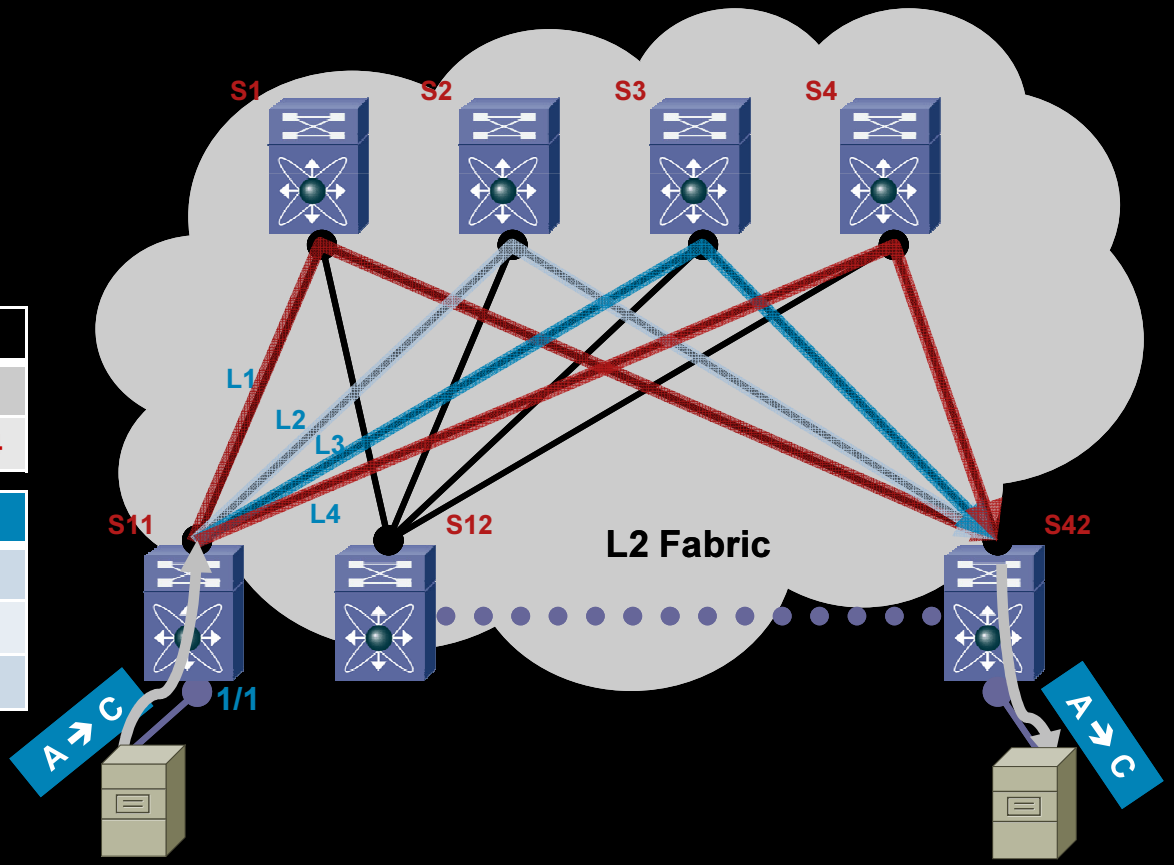
Unicast with FabricPath

Forwarding decision based on ‘FabricPath Routing Table’

- Support more than 2 active paths (up to 16) across the Fabric
- Increase bi-sectional bandwidth beyond port-channel
- High availability with N+1 path redundancy

Switch	IF
…	…
S42	L1, L2, L3, L4

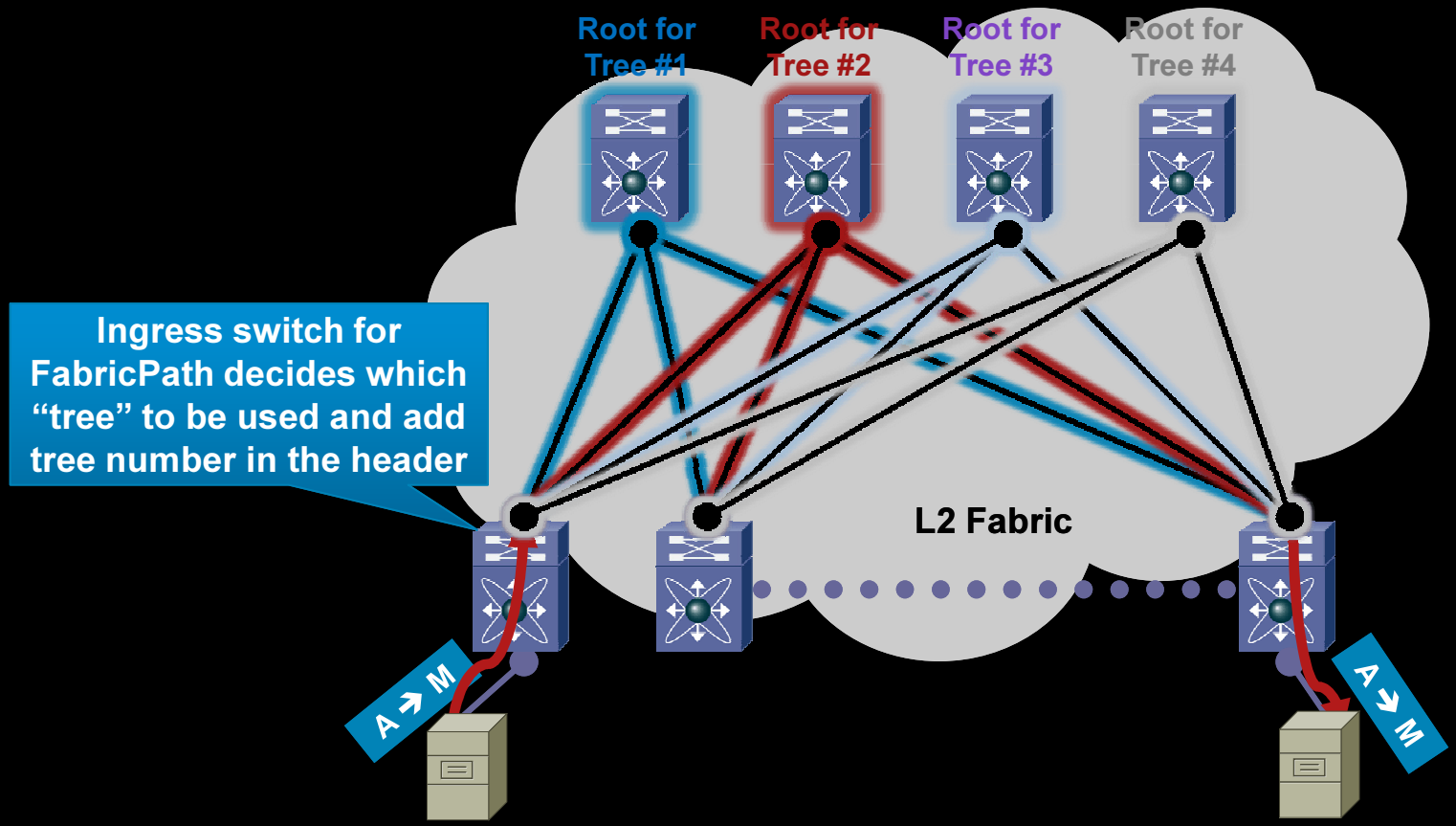
MAC	IF
A	1/1
…	…
C	S42



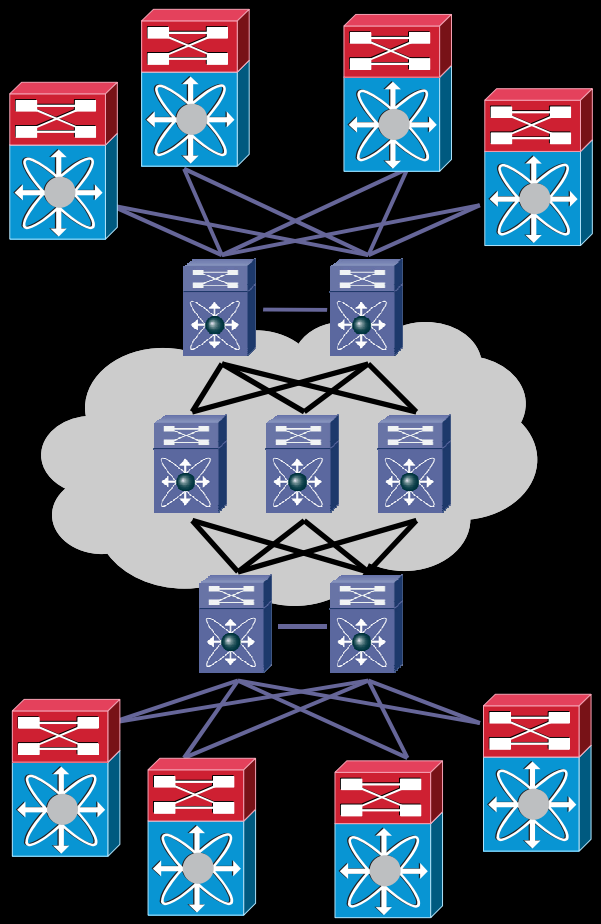
Multicast with FabricPath

Forwarding through distinct ‘Trees’

- Several ‘Trees’ are rooted in key location inside the fabric
- All Switches in L2 Fabric share the same view for each ‘Tree’
- Multicast traffic load-balanced across these ‘Trees’



Use Case: L2 Internet Exchange Point



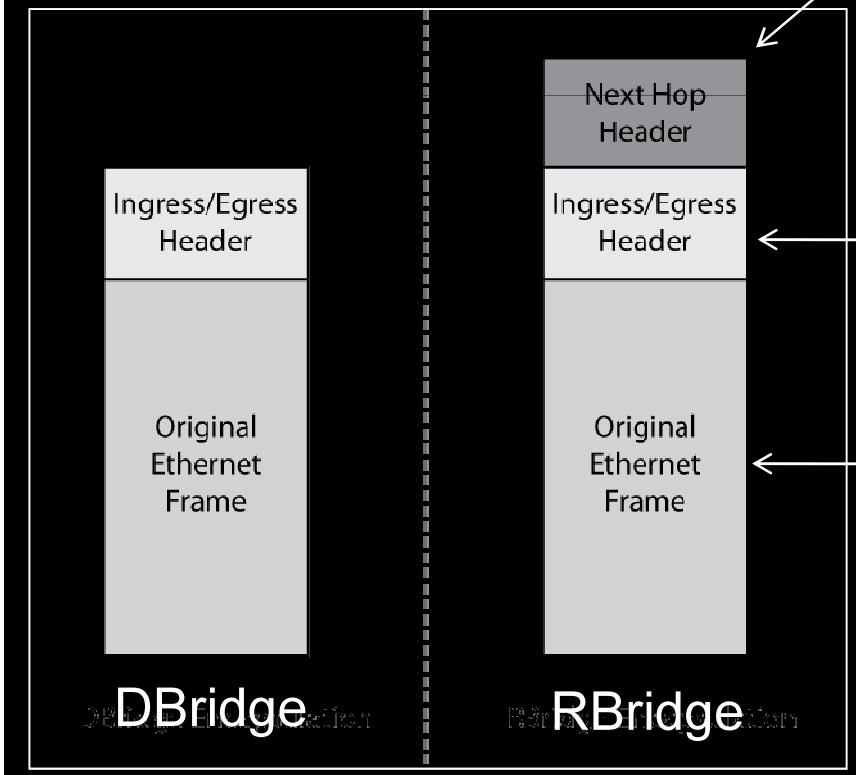
IXP Requirements

- Layer 2 **P**eering enables multiple providers to peer their internet routers with one another
- **10GE non-blocking fabric**
- Scale to thousands of ports

FabricPath Benefits for IXP

- Transparent Layer 2 fabric
- **Scalable to thousands of ports**
- **Bandwidth not** limited by chassis / port-channel limitations
- **Simple to manage, economical to build**

L2MP Bridges require an additional Ingress/Egress header



RBridges also require a next hop header to traverse a legacy Ethernet cloud

It contains Ingress and Egress D/Rbridge switchIDs, i.e., the entry and exit points of the L2MP cloud

This is unchanged, except FCS

Schicht 2

Schicht 2 zwischen Rechenzentren

Die Aufgabe

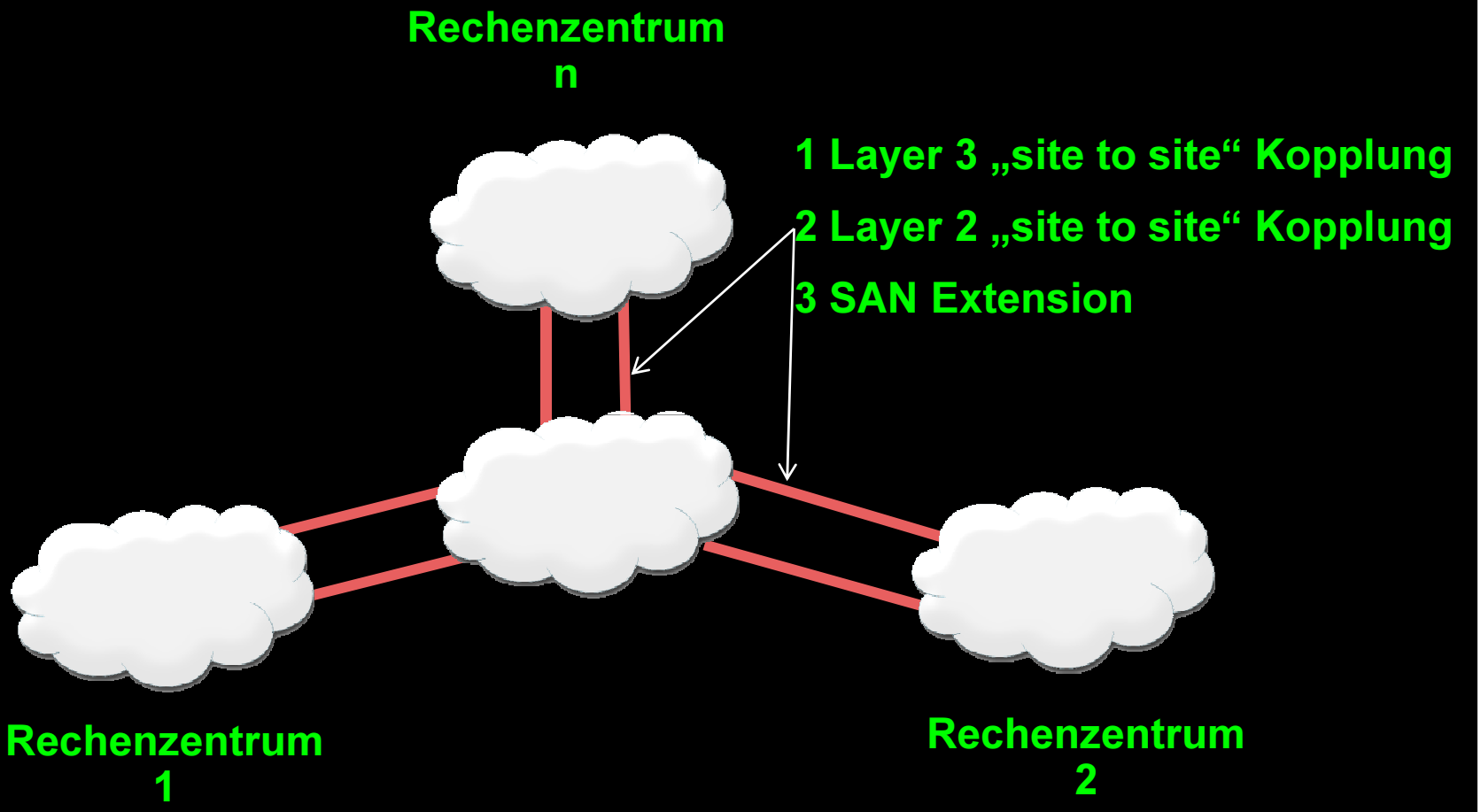
Rechenzentrum 1

Verbindung

Rechenzentrum 2

Die simple Aufgabe: 2 Rechenzentren und eine Verbindung
„Tür an Tür“

Die Aufgabe



Die Aufgabe für Fortgeschrittene:
n Rechenzentren und redundante Verbindung

Die Möglichkeiten für DCI

L2TPv3

EoMPLS

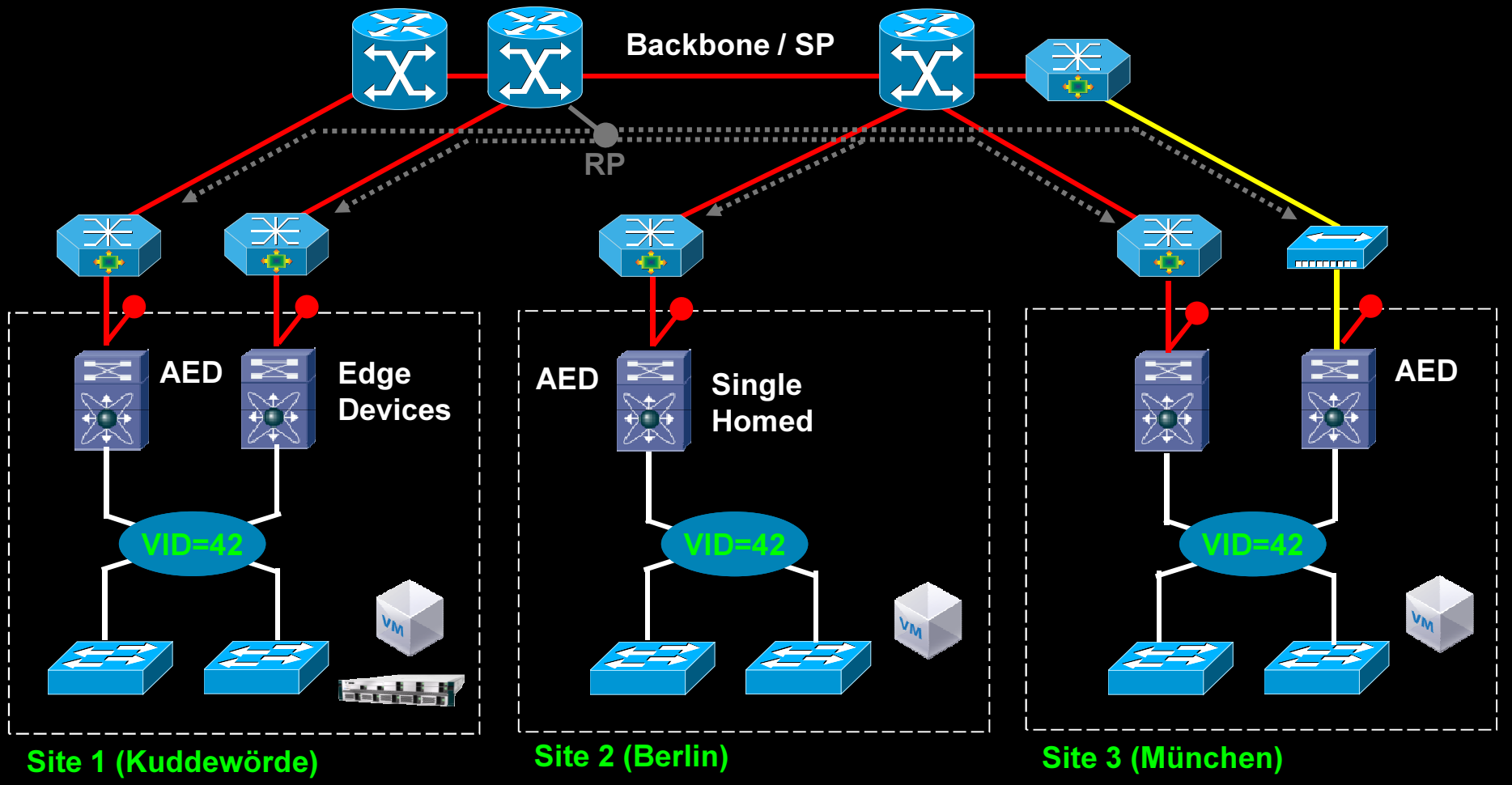
L2MP / TRILL

VPLS

VPLSoGRE

OTV

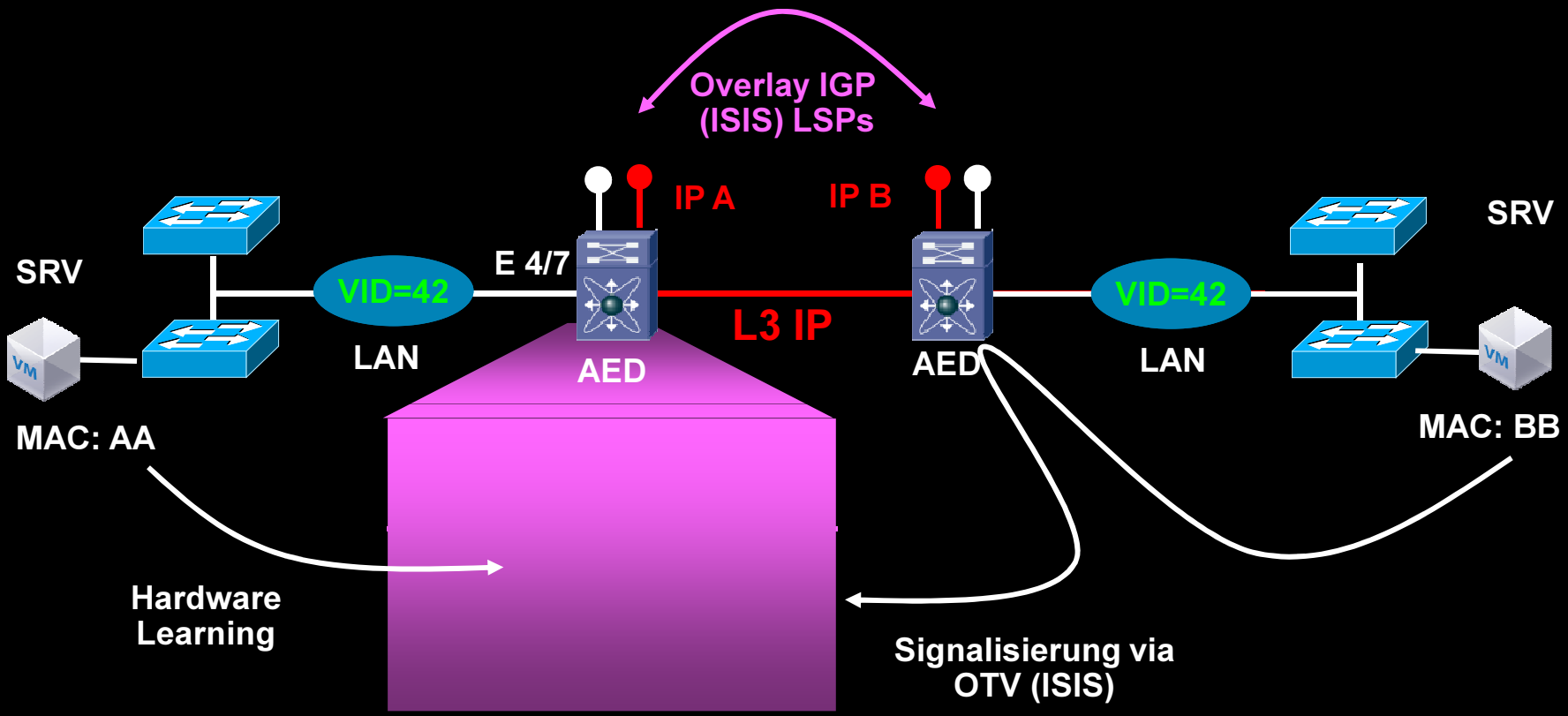
Bestandteile der Lösung



AED := Authorized Edge Device (pro VLAN)

L3
Metro Ethernet
L2 (LAN)
MC Tree

OTV = MAC Bridging + MAC routing

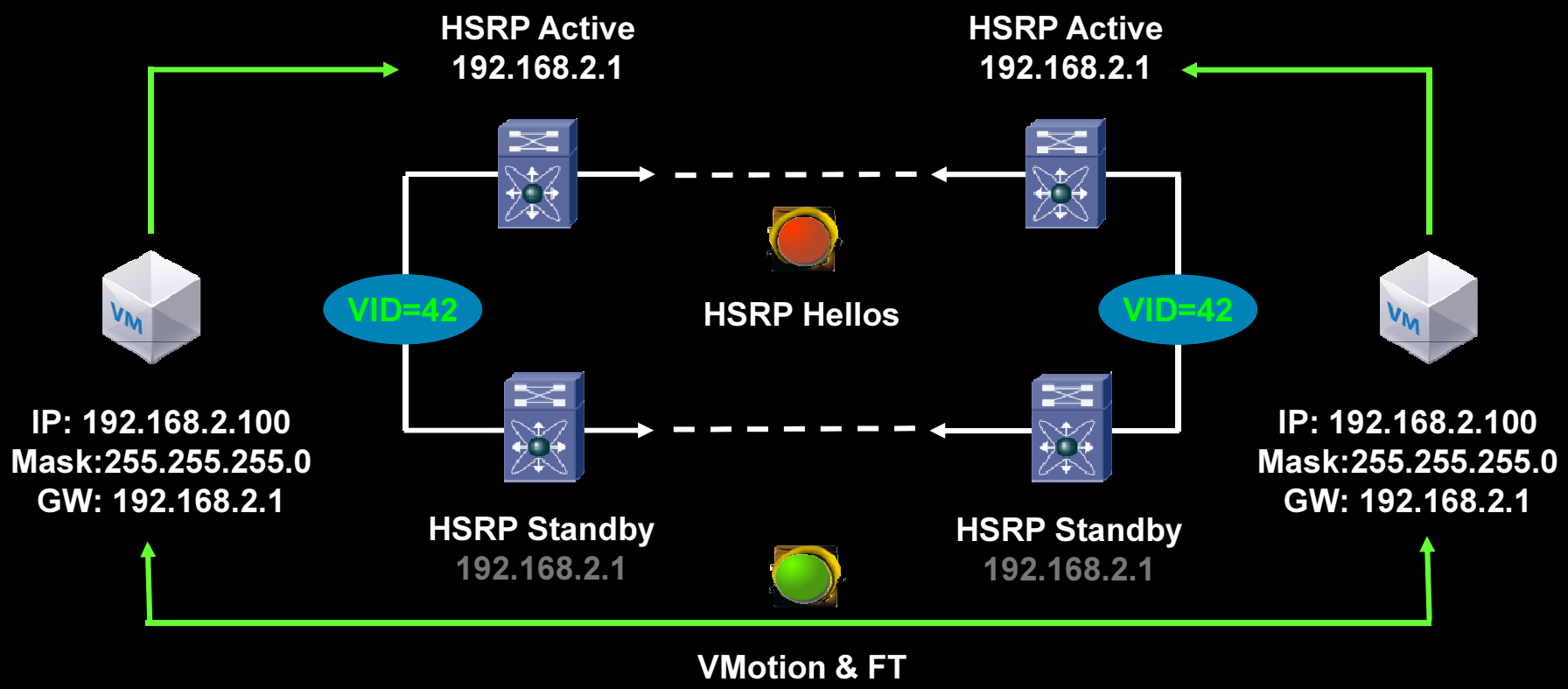


Sites: MAC Bridging
Core: “MAC routing” via “overlaid IGP”
MAC table: “destination ports” und “IP routes”

Overlay Interface

Internal Interface

FHRP



Standardverhalten:
Filter der HSRP Nachrichten auf den OTV Stecken


CISCO