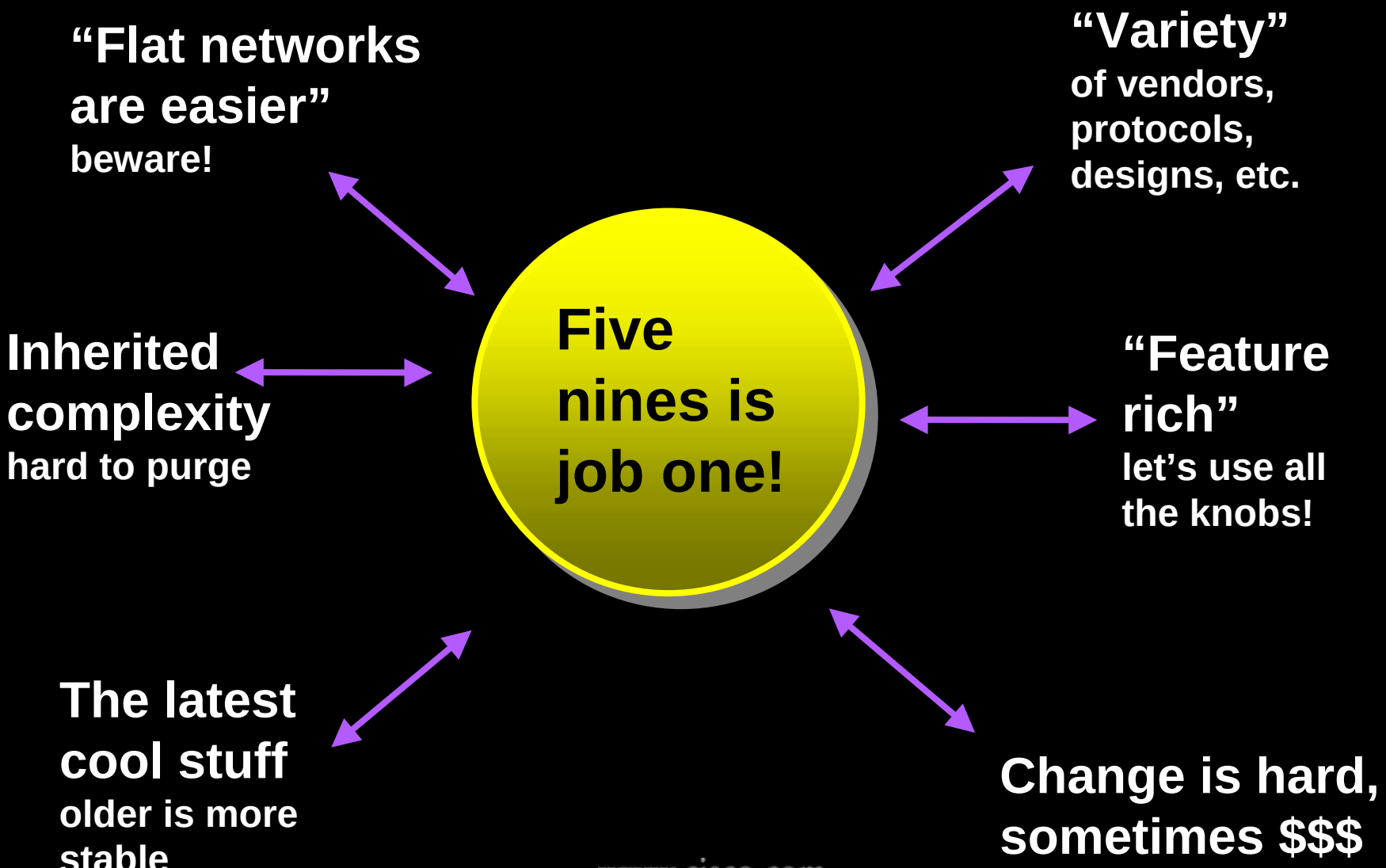
High-Availability Enterprise Network Design

haviland@cisco.com

Cisco Systems

Staying On Target
HA Focus vs Distractions!



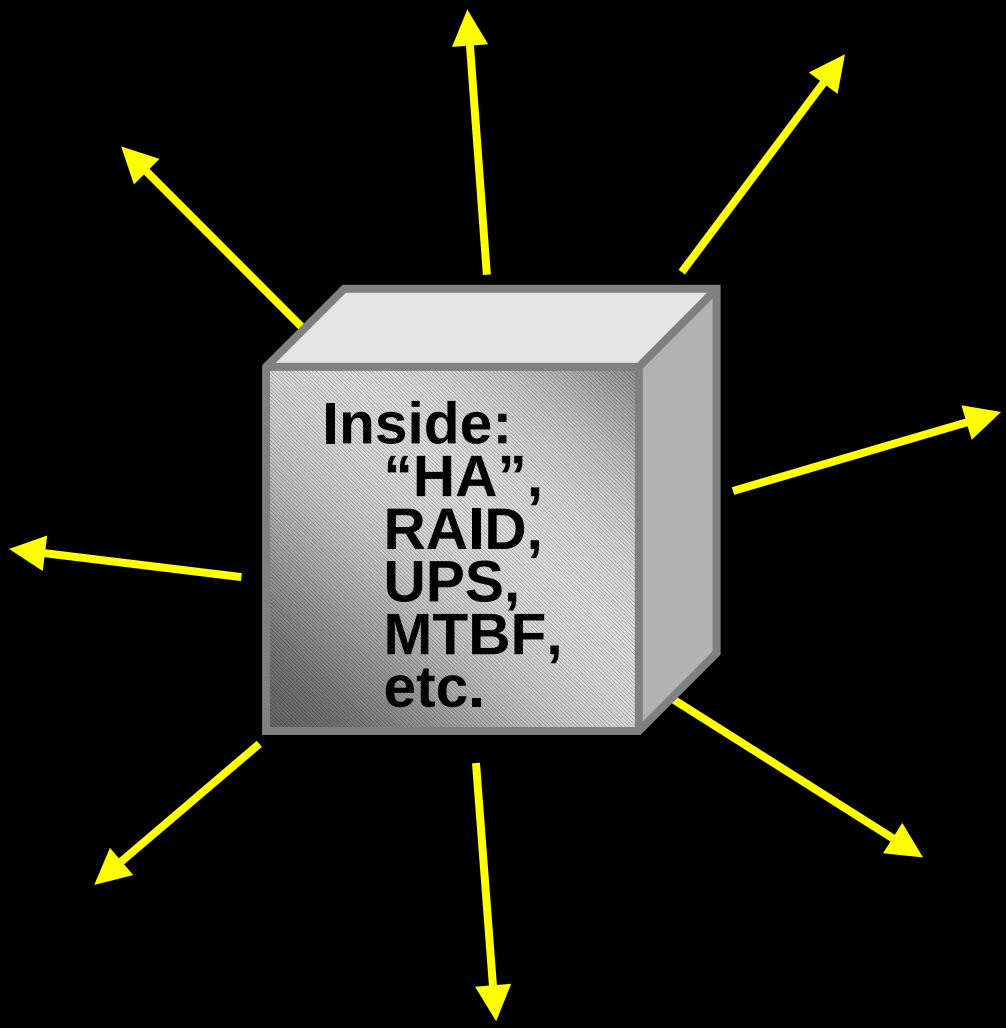
HA Features of the Catalyst 6500
Consider for Backbones & Server Farms

- ✔ **Fabric Redundancy**
  - switch fabric module
  - in CatOS 6.1
- ✔ **Supervisor Redundancy**
  - HA feature in CatOS 5.4.1
  - stateful recovery
  - image versioning on the fly
- ✔ **MSFC Redundancy**
  - config-sync feature
  - IOS 12.1.3 CatOS 6.1
  - HSRP pair

Thinking Outside the Box

For HA/HP design "outside the box"

- the logical design is critical
- network features & protocols
- geophysical diversity is powerful



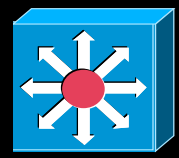
Dramatis Personae
Our Cast of Symbols

- ✔ **Links**
  - GE, DPT, SONET, etc.
- ✔ **L2 switching**
  - L2 forwarding in hardware
- ✔ **L3 switching**
  - L3/L2 forwarding in hardware
- ✔ **Routing**
  - L3 forwarding (SW or HW)
- ✔ **Control plane = IOS**
  - routing protocols & features
- ✔ **QoS where required**
- ✔ **Application intelligence**


GigE


Channel

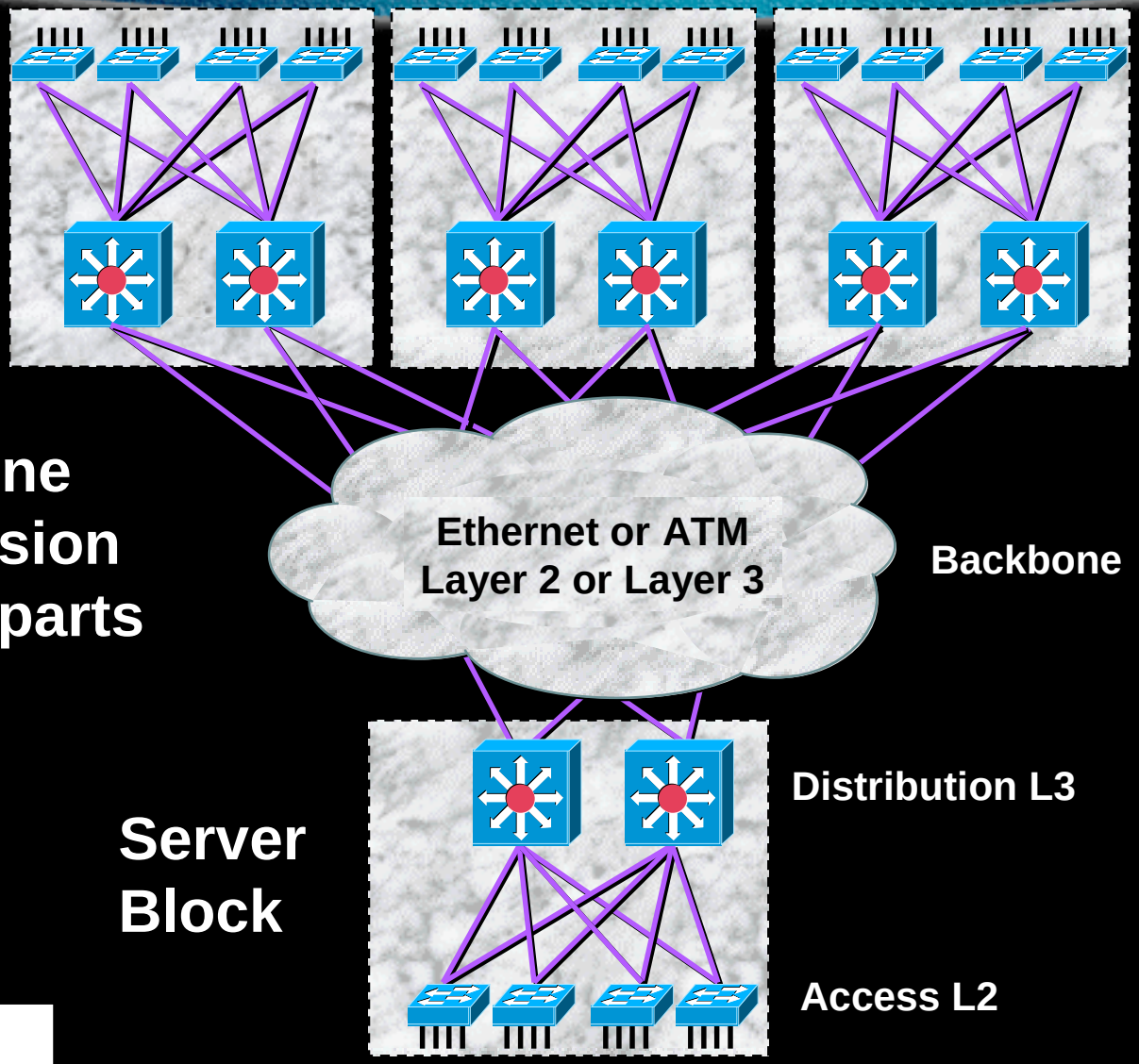

Catalyst 4000


Catalyst 6500


Cisco 7500


Cisco 12000

HA Gigabit Campus Architecture
survivable modules + survivable backbone



- Define the mission critical parts first!

High Availability Design
Why a Modular ABC Approach

- ✔ Many new products, features, technologies
- ✔ HA and HP application operation is the goal
- ✔ Start with modular, structured approach (the "logical" design)
- ✔ Add multicast, VoIP, DPT, DWDM...

Design the Solution
Then Pick the Products

$350
$300
$250
$200
$100

Price per 10/100

Catalyst 4XXX

Catalyst 5XXX

Catalyst 6XXX

10/100 Ports	32-96	24-500+	24-350+
Gigabit Ports	6-12	3-38+	8-64+
Backplane	24 Gbps	1.2-3.6 + 10Gbps	250+ Gbps
Switching Capacity	20 Mpps	Up to 72 Mpps	Up to 150 Mpps

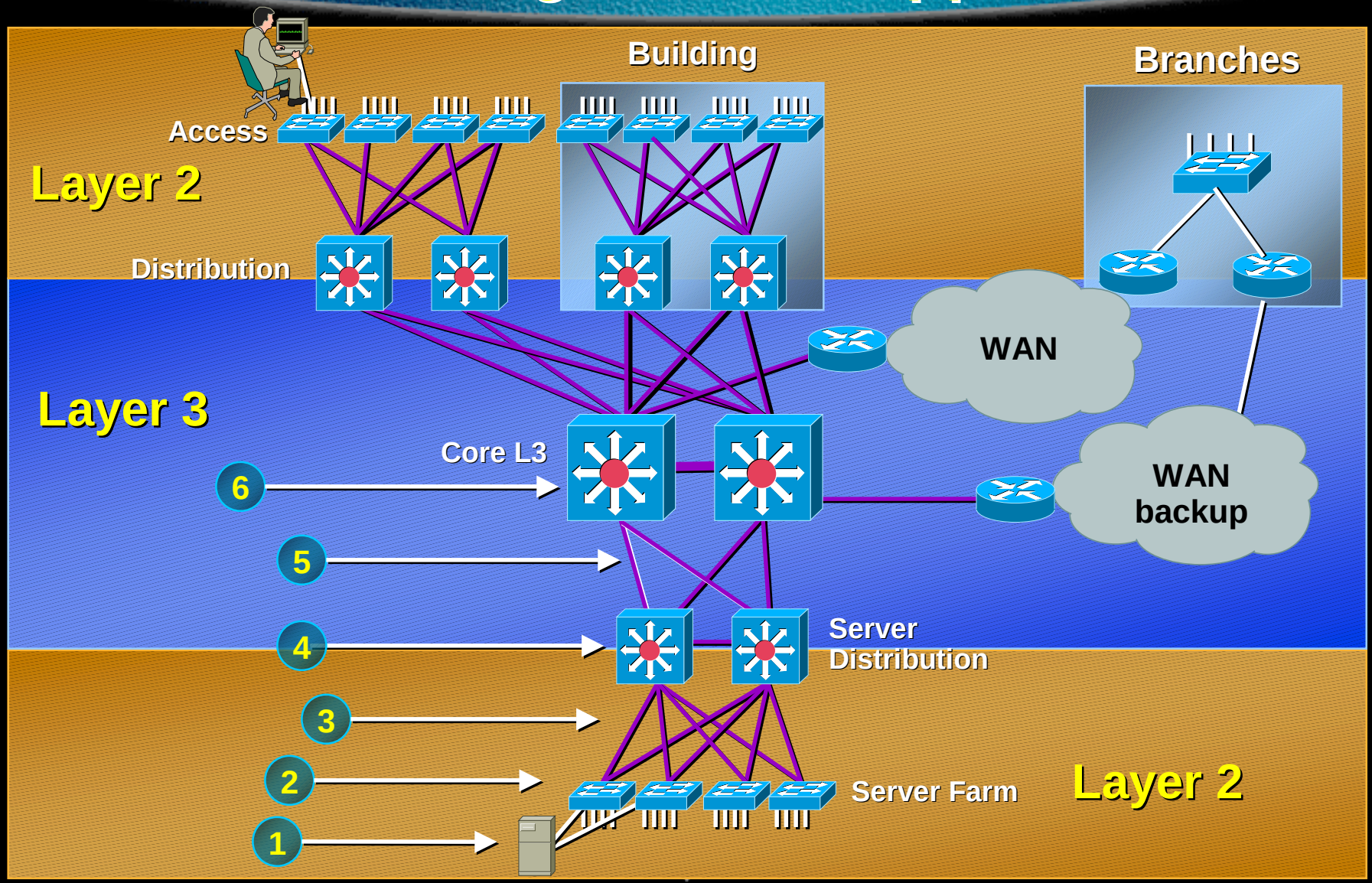
HA Design Reality Check!
Assume Things Fail - Then What?

- ✔ Networks are complex
- ✔ Things break, people make mistakes
- ✔ What happens if a failure occurs?
- ✔ Simple, structured, deterministic design required for fast recovery
- ✔ The “tradeoffs”

  your choices are important

Network Recovery

How Long? What Happens?



Network Recovery Times

If You Follow the Rules

Failure Scenario	Recovery Mode	Recovery Time
1,2 server	Server NIC	< 2 seconds
3,4 uplink	HSRP (& UplinkFast)	tune to 3 seconds
5,6 core	HSRP track	tune to 3 seconds
dual-path L3	alternate path used	< 2 seconds
EtherChannel	channel recovery	< 1 second
L3 routing	EIGRP or OSPF	depends on tuning
L2 general	L2 spanning tree	tune (up to 50 seconds)
DPT	IPS	50 milliseconds

Design for High Availability

How to Build Boring Networks!

- ✔ The Concepts
- ✔ The Rules
- ✔ Design Building Block
- ✔ Design Backbone
- ✔ Notes on Tuning

HA Network Design Concepts
thinking outside the box

1) Simplicity & Determinism
2) Collapse the Sandwich
3) Spanning Tree Failure Domain
4) Map L3 to L2 to L1
5) Scaling and Hierarchy
6) ABCs of Module + Backbone Design
7) The Four Corners

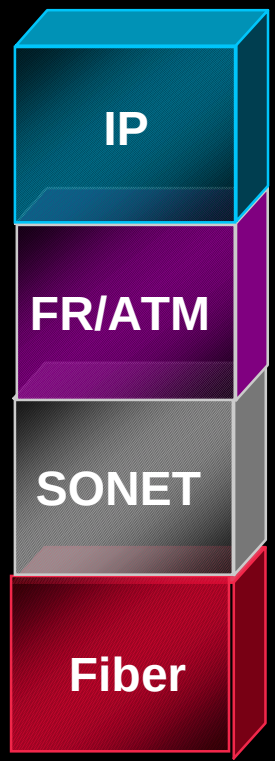
1) Simplicity and Determinism

reducing the degrees of freedom

- ✔Every Choice Affects Availability!
- ✔Determinism or Flexibility?
- ✔Would you support 27 desktop environments?
- ✔Would you support 13 network vendors?
- ✔Would you use 57 varieties of Cisco IOS?

2) Collapse the Sandwich
route IP over glass

Traditional Model



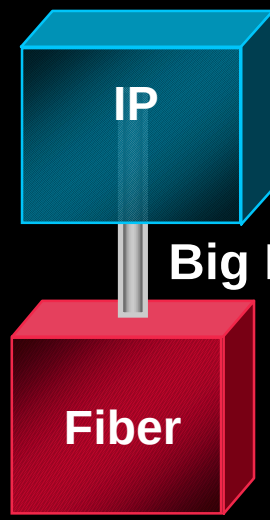
Service

Traffic Eng

Fiber Mgmt

- Lower equipment cost
- Lower operational cost
- Simplified architecture
- Scalable capacity

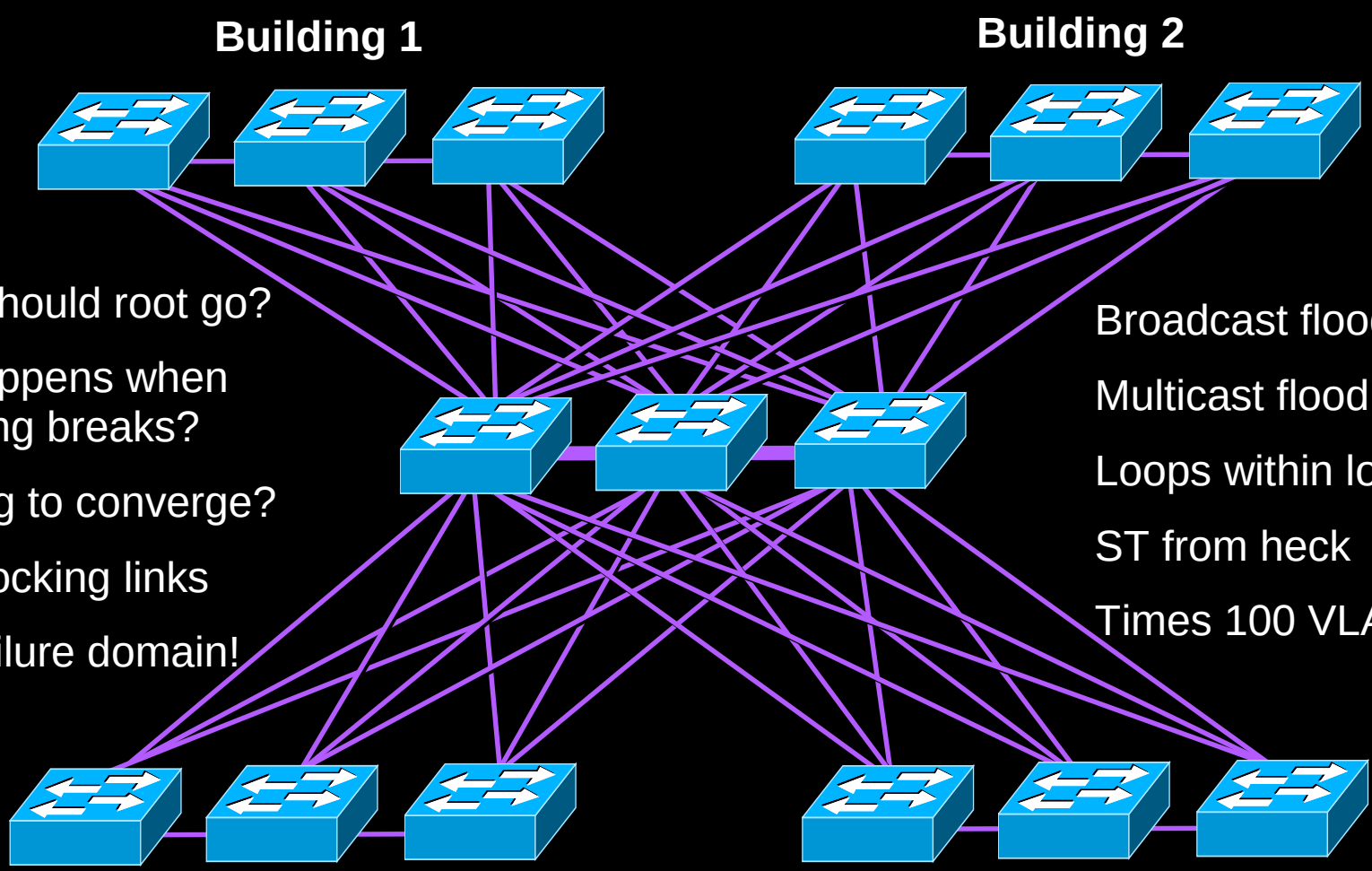
Optical Internetworking



3) Minimize the Failure Domain

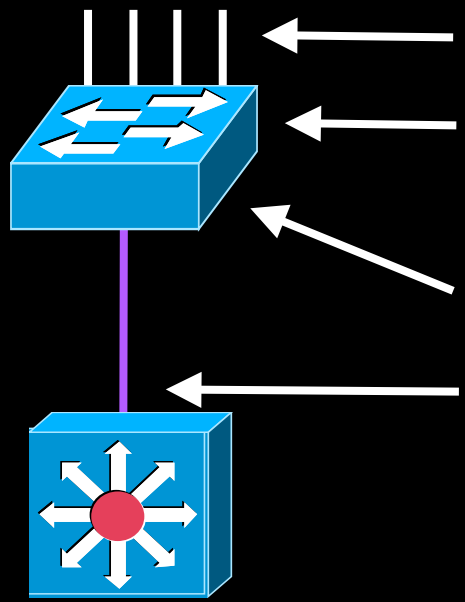
public enemy number one

avoid highly meshed, non-deterministic large scale L2 = VLAN topology



4) Map L3 to L2 to L1

✔Easier administration & troubleshooting



Clients in subnet 10.0.55.0

VLAN 55

wiring closet “55” on floor 55

access switch “55”

interface VLAN 55

all match and life is good

go fishing with your kids

10/100 BaseT

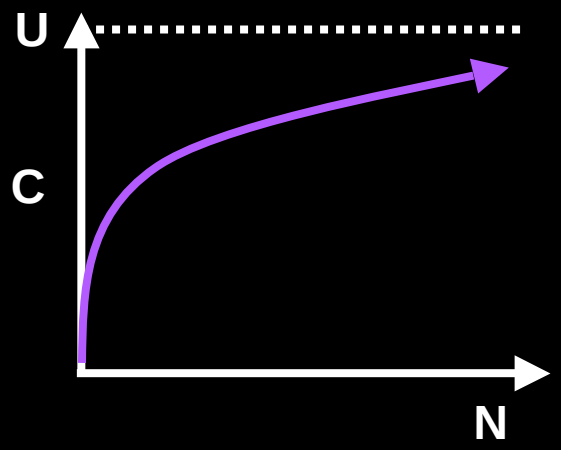
GE or GEC

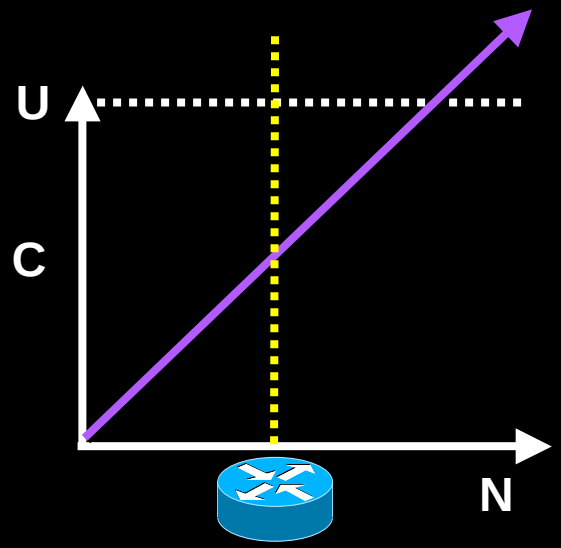
5) Scaling and Hierarchy

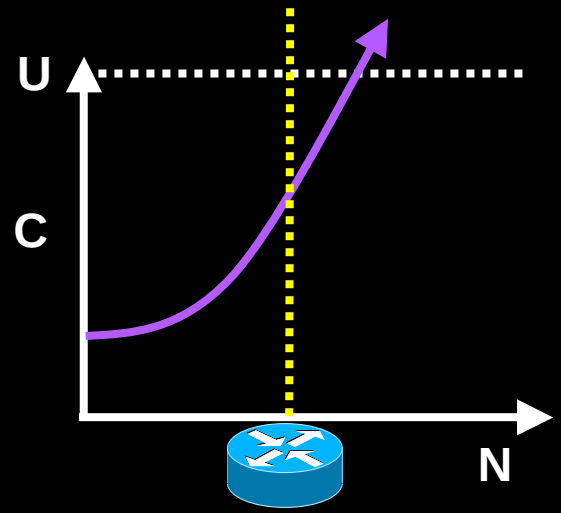
Strong hierarchies like telephone system and Internet segment addressing and therefore scale

Flat L2 Ethernet is easy but does not scale

ATM LANE is logically flat, scales as N squared

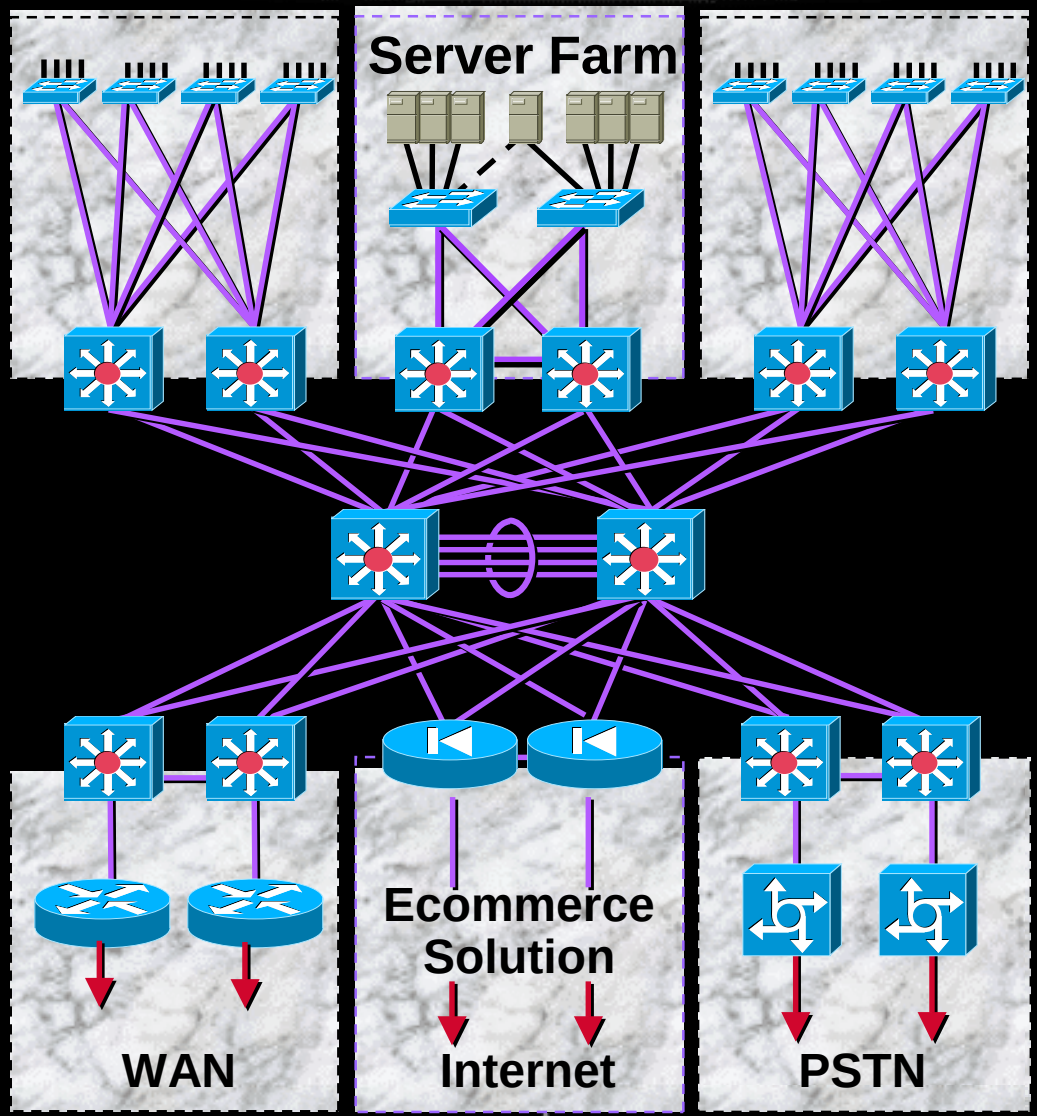






C complexity
U unmanageable
N number of devices

6) Building Block & Backbone Design ABCs



A design bb

B design BB

C connect bb to BB

Divide and conquer

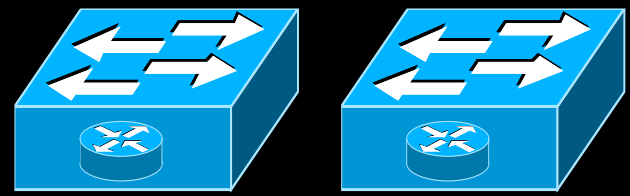
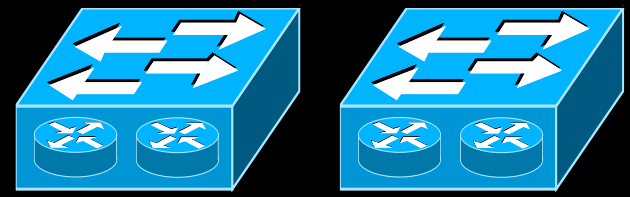
Cookie cutter configuration

Deterministic

L3 demarcation

7) Four Square Network Redundancy

or the Four Corners Problem

L3	One Chassis	Two Chassis
One Supervisor	Simplest No Redundancy 	GeoPhysical Effective 
Two Supervisors	When space is limited  “HA”	Most Complex Belt and Suspenders 

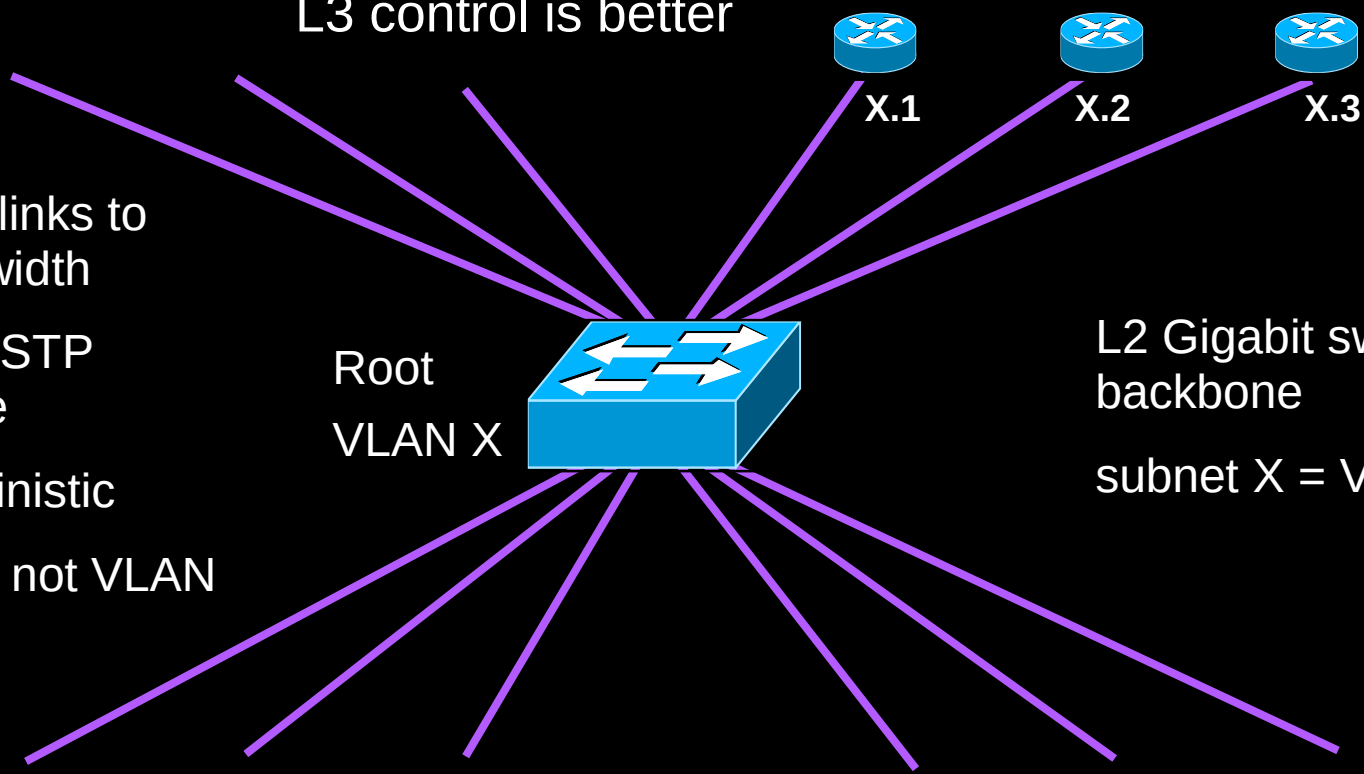
Dos and Don’ts for HA Design

1) Eliminate STP Loops
2) L3 Dual-Path Design
3) EtherChannel Across Cards
4) Workgroup Servers
5) Use HSRP Track
6) Passive Interfaces
7) Issues with Single-Path Design
8) Oversubscription Guidelines
9) HA for single attached servers
10) Protocol Tradeoffs
11) UDLD Protection

Rule 1) Eliminate STP Loops
in the backbone and mission critical points

Too many cooks spoil the broth

L3 control is better



No blocking links to waste bandwidth

Avoids slow STP convergence

Very deterministic

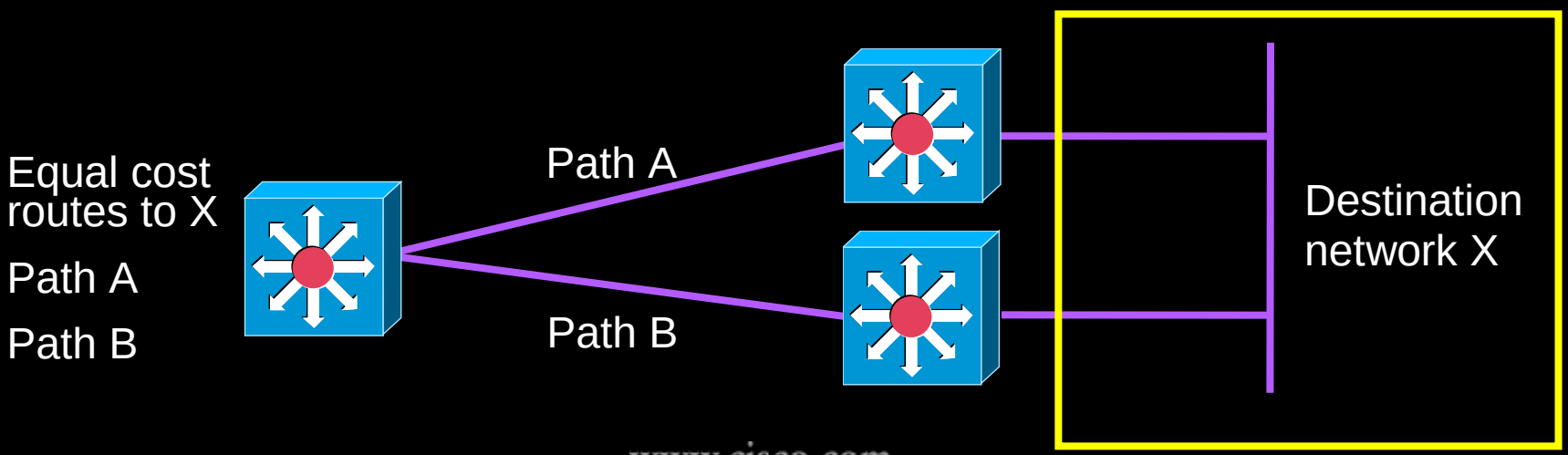
Routed links not VLAN trunks

L2 Gigabit switch in backbone

subnet X = VLAN X

Rule 2) Dual Equal-Cost Path L3

- ✔Load balance - don’t waste bandwidth
  - unlike L1 and L2 redundancy
- ✔Fast recovery to remaining path
  - detect L1 down & purge - about 1s
- ✔Works with any routed fat pipes



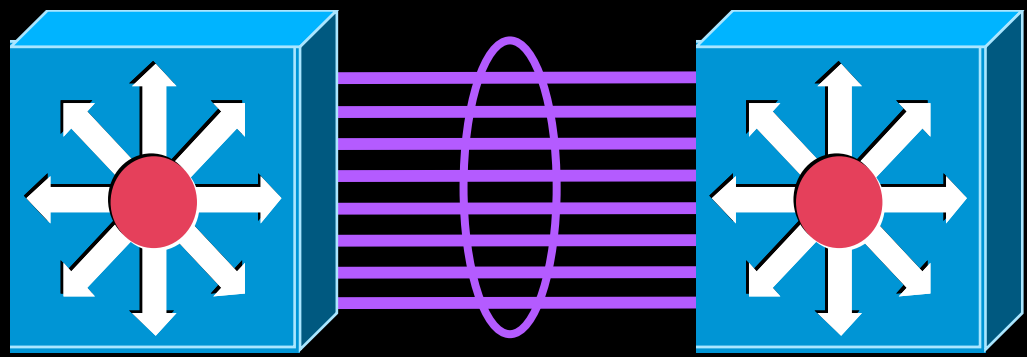
Rule 3) EtherChannel Across Cards

Increased availability

- ✔ Sub second recovery
- ✔ Spans cards on 6500
- ✔ Up to 8 ports in channel

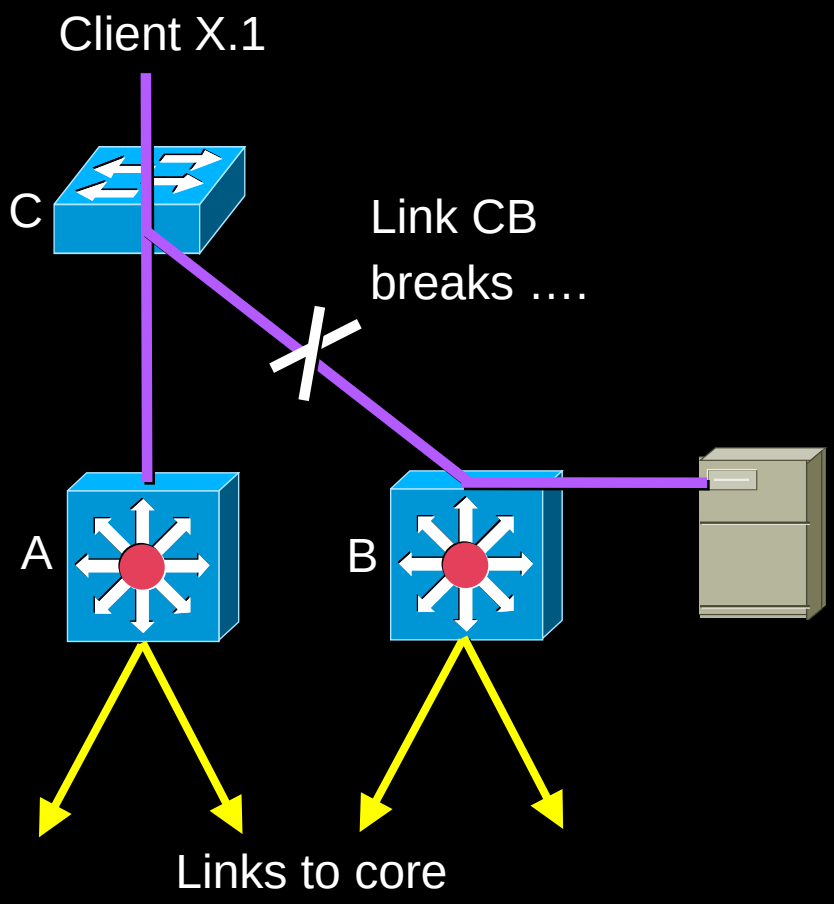
Small complexity increase

- ✔ Single L2 STP link
- ✔ Single L3 subnet
- ✔ less if channel set “on”



Rule 4a) Connect Workgroup Server

- **With no L2 recovery path, what happens if link breaks ….**



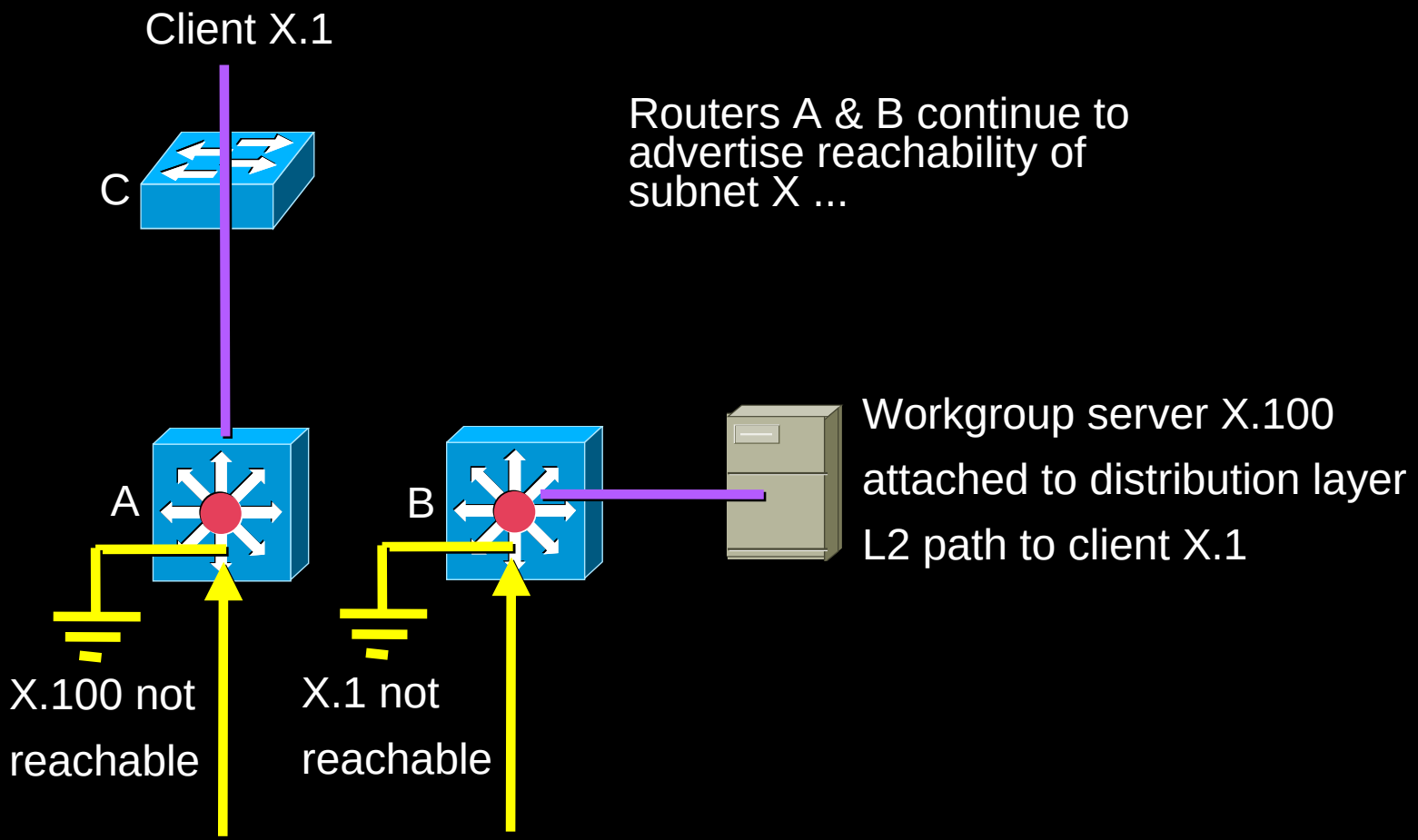
VLAN X in purple includes clients and workgroup servers attached at different places.

Workgroup server X.100 attached to distribution layer

L2 path to client X.1

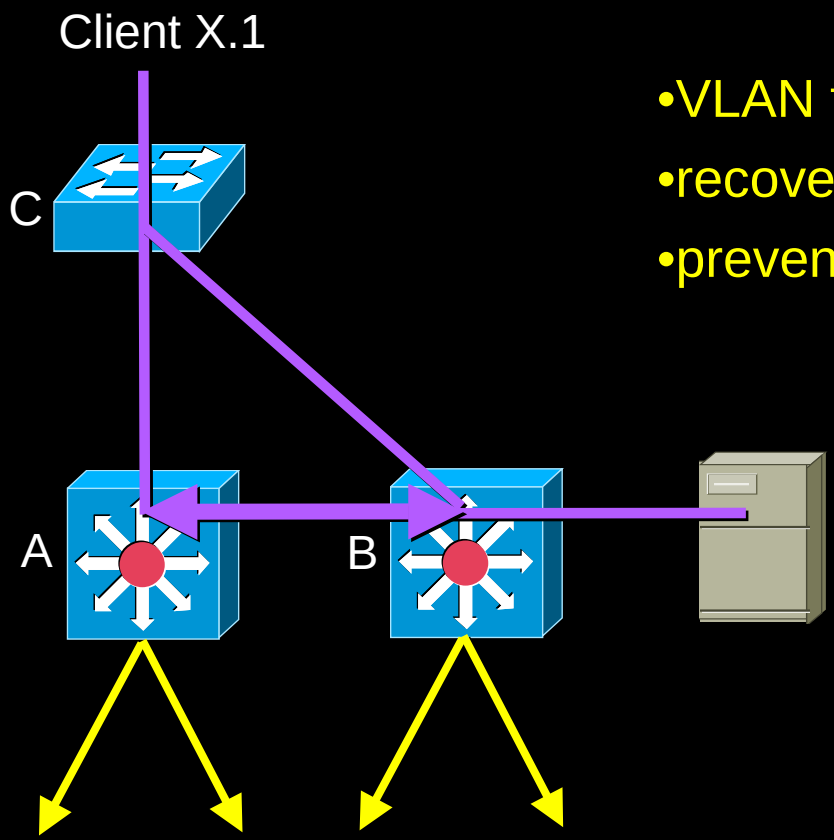
Rule 4b) Connect Workgroup Server

- **Subnet X now discontiguous**
- **Incoming traffic gets dropped**



Rule 4c) Connect Workgroup Server

- **Introduce L2/STP redundancy**
- **Adds a loop (band-aid fix)**



- VLAN trunk AB forms L2 loop
- recovery path for STP
- prevents black hole

Workgroup server X.100
attached to distribution layer
L2 path to client X.1

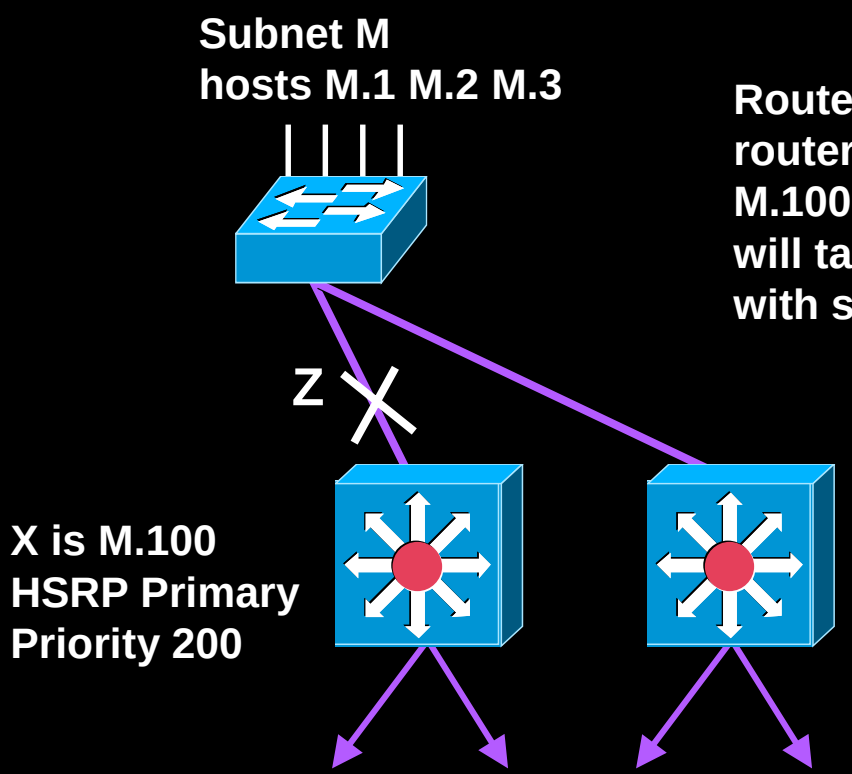
Rule 4d) Connect Workgroup Server

Real Lessons:

- Enterprise Server Farms are better
- L3 demarcation is better
- Example of why extended L2 is difficult

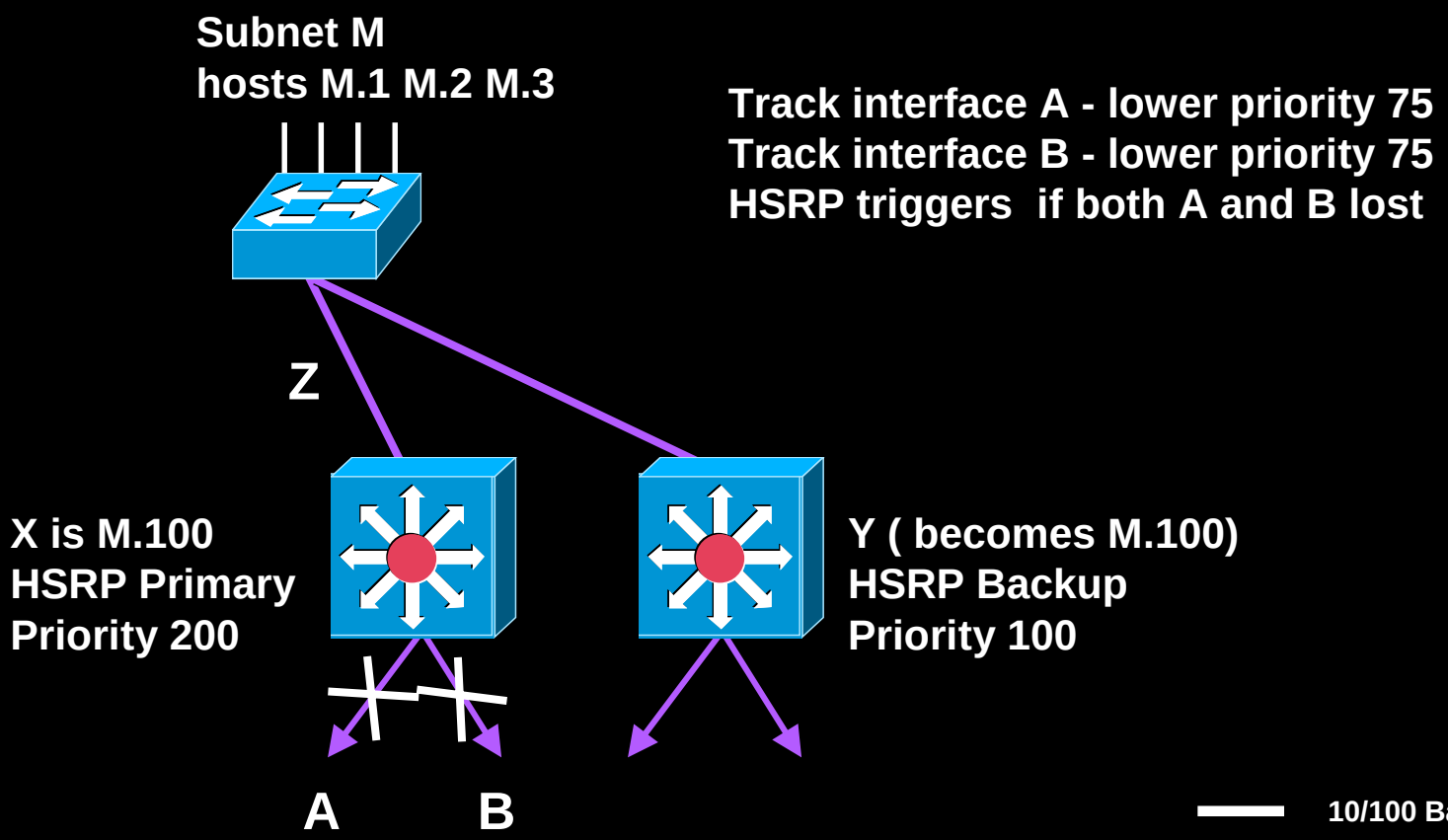
Rule 5a) Use HSRP Track

- Review - Hot Standby Router Protocol
- Fast recovery can be tuned to 3s or less



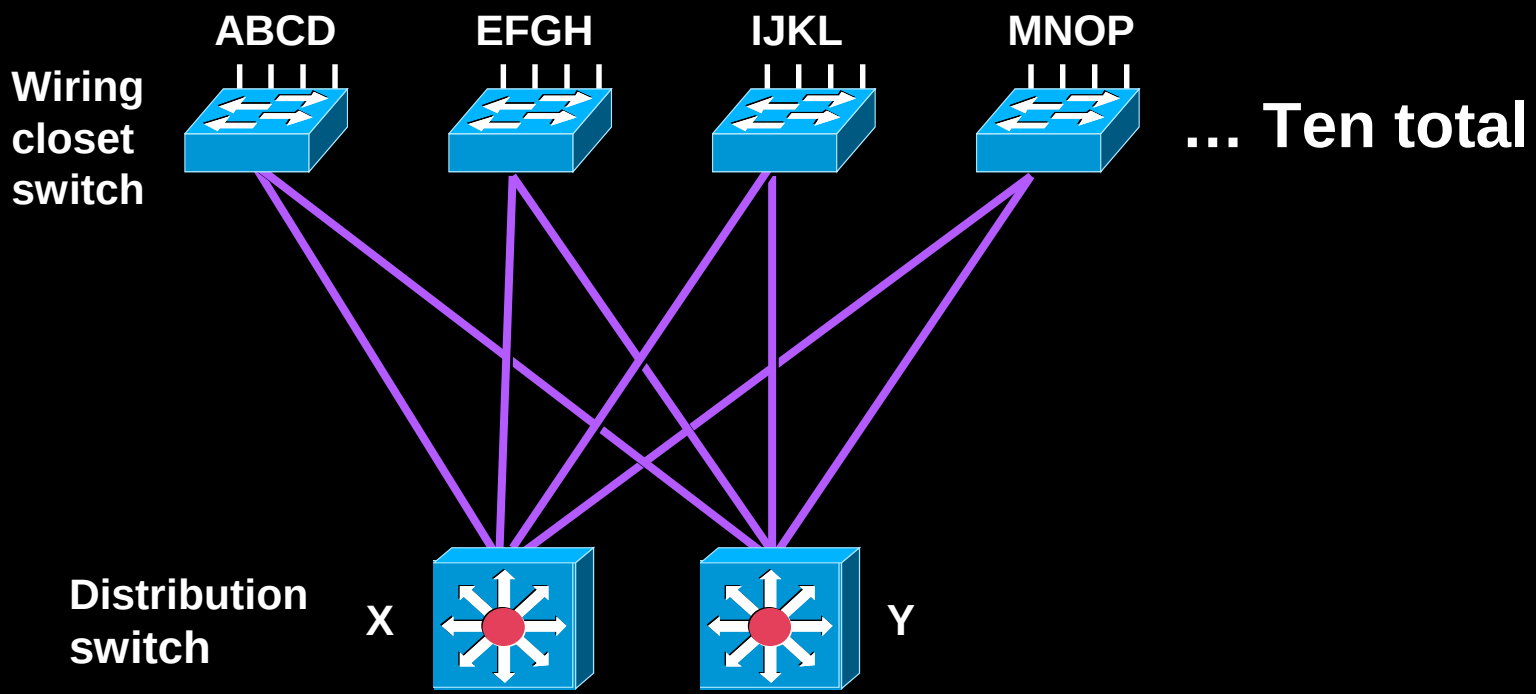
Rule 5b) Use HSRP Track

- Track extends HSRP to monitor links to backbone
- Ensures shortest path - best outbound gateway



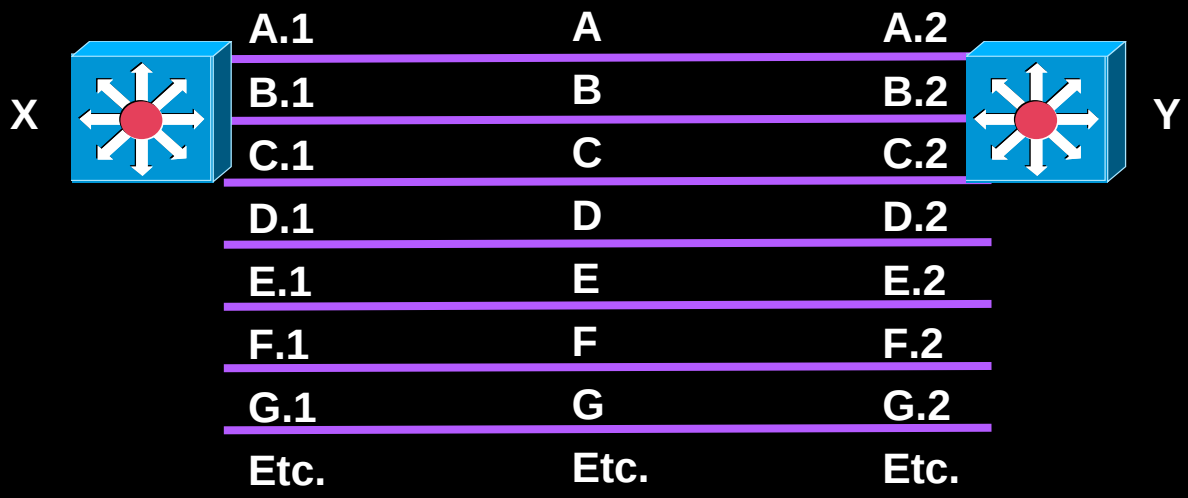
Rule 6a) Use Passive Interfaces

- L3 switches X & Y in distribution layer
- 4 VLANs per wiring closet
- 10 wiring closets



Rule 6b) Use Passive Interfaces

- What X and Y see is 4*10=40 routed links
- Increased protocol overhead & CPU



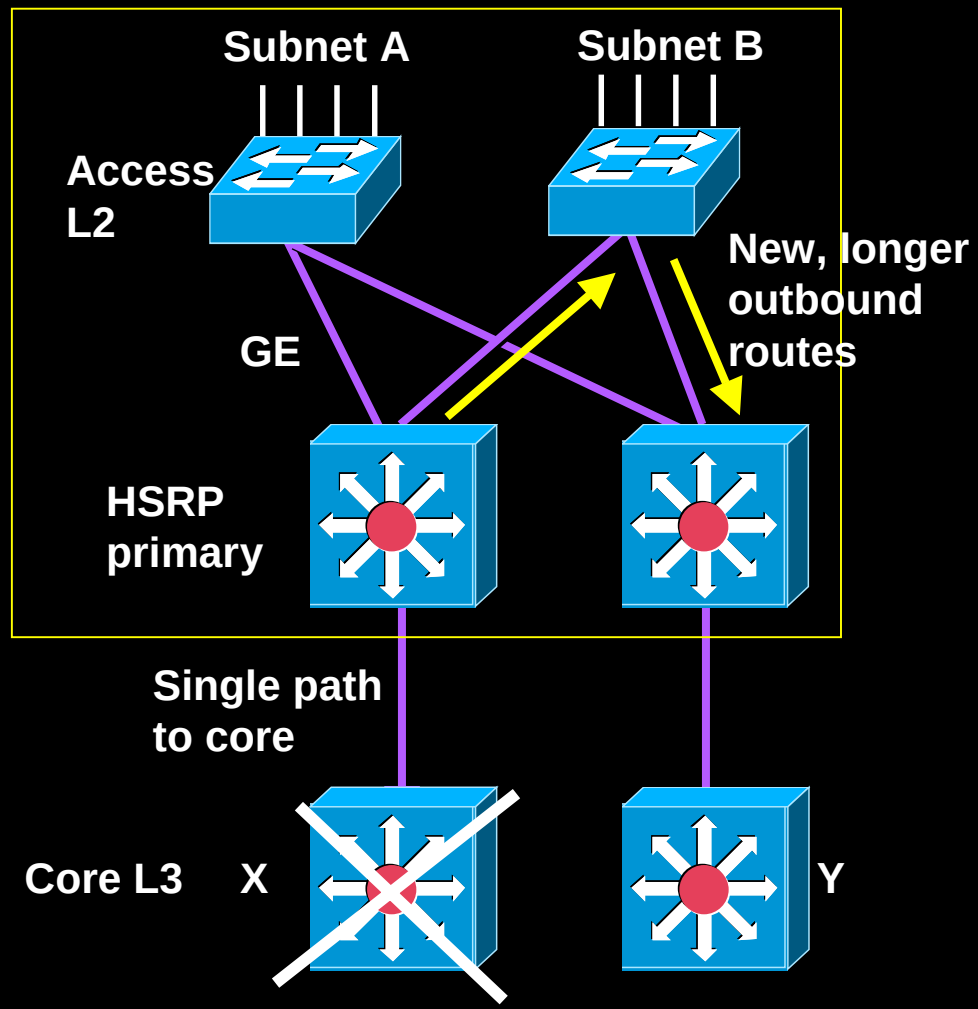
Rule 6c) Use Passive Interfaces

- Turns off routing updates & overhead
- Leave two routed links for redundant paths
- CDP, VTP, HSRP etc. still function on all links

Rule 7a) Issues With Single Path Designs

Outbound case ...

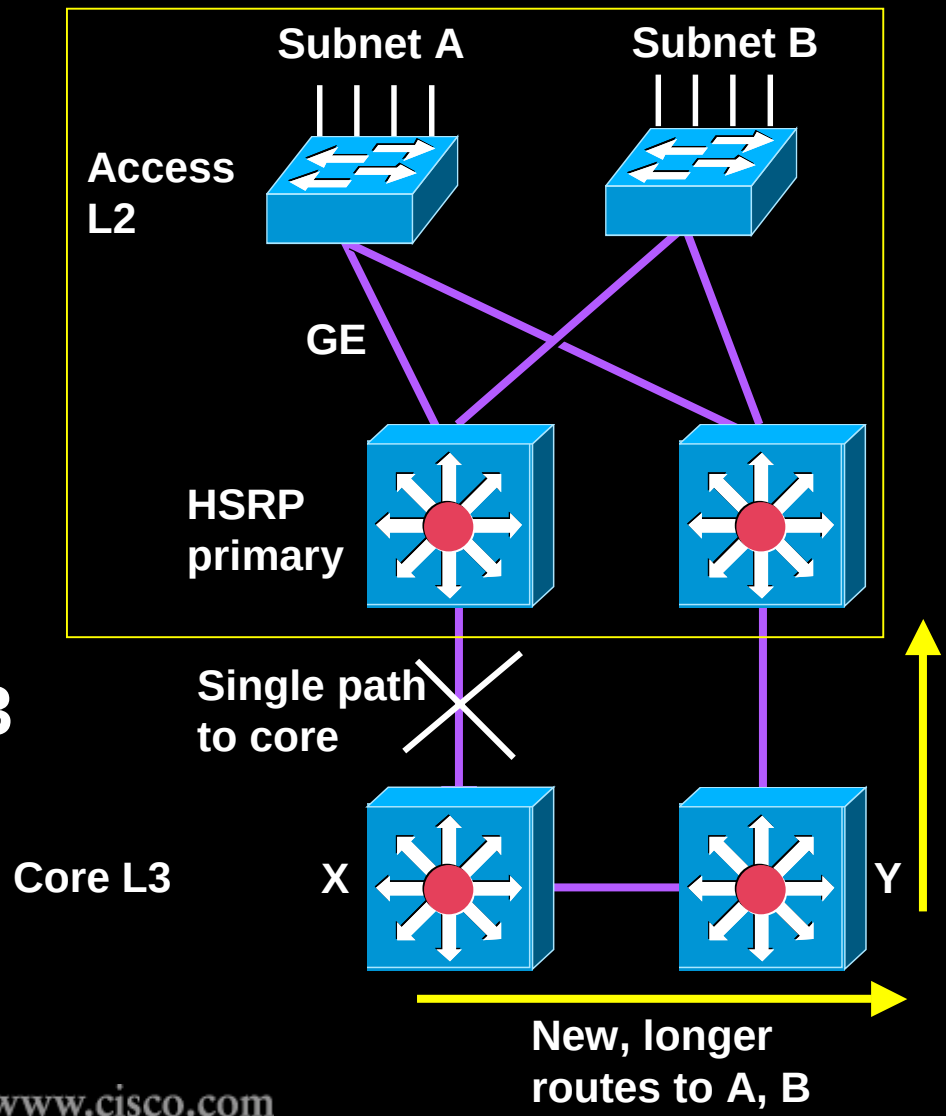
- ✔L3 engine MSFC on core-X reloads
- ✔Lights are on but nobody home - HSRP does not recover
- ✔Remove passive interface to wiring closet subnets A, B
- ✔Provide longer routed recovery path



Rule 7b) Issues with Single-Path Design

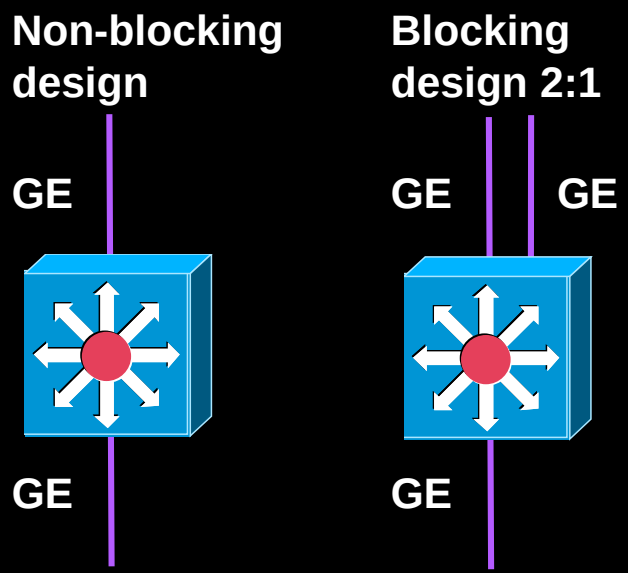
Inbound case ...

- ✔Recovery must take place in both directions
- ✔Routing protocol recovers longer route from X to subnets A, B
- ✔Therefore dual-path L3 is better & faster than single-path



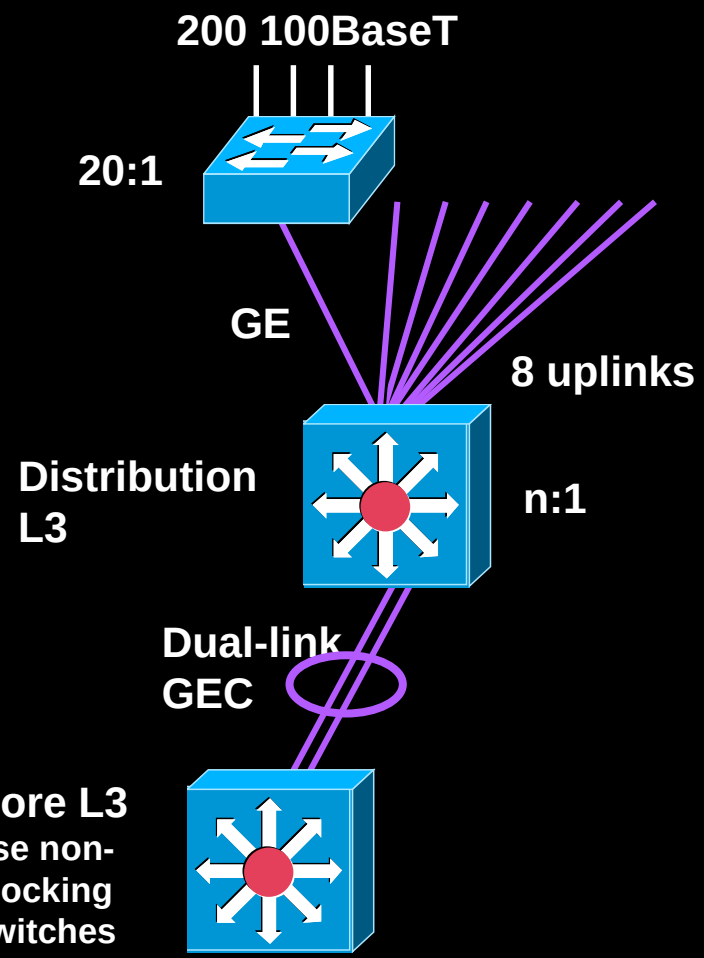
Rule 8a) Oversubscription Guidelines

- ✔Oversubscription part of all networks - not bad
- ✔Non-blocking switches do not mean a non-blocking network
- ✔You determine the amount of “blocking”



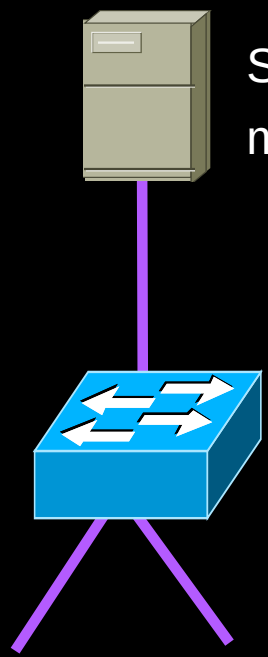
Rule 8b) Oversubscription Guidelines

- ✔ Oversubscription rules of thumb work well
- ✔ 20:1 at wiring closet
- ✔ Less in distribution and server farm
- ✔ QoS required IFF congestion occurs
- ✔ Protect real time flows at congested points



Rule 9) Dual Supervisors
HA for Single Attached Servers

- ✔Single point of failure
- ✔Dual supervisors - fast stateful recovery
- ✔No increase in complexity



Single attached server
mission critical application

HA dual supervisors
Catalyst 6XXX

Redundant uplinks

10/100 BaseT

GE or GEC

Rule 10) Protocol Tradeoffs

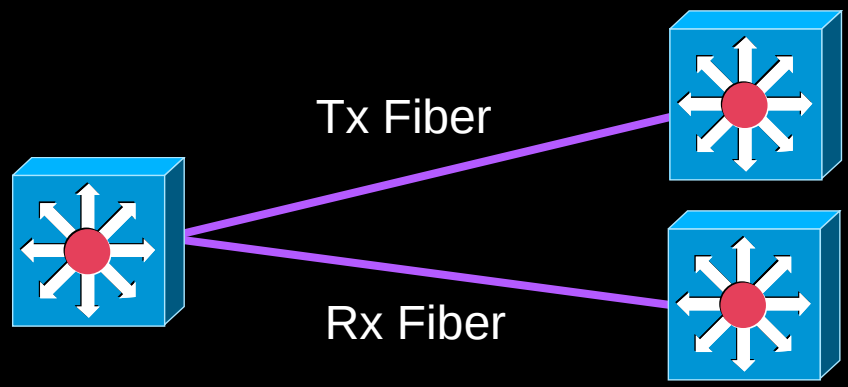
Automatic or Manual Configuration

- ✔ Configuration up front rather than CPU overhead later, for example:
  - set VTP mode transparent
  - set/clear VLANs for each trunk
  - set trunks on or off
  - set channel on or off
- ✔ Choose flexibility or determinism

Rule 11) UniDirectional Link Detection

- ✔UDLD detects mismatch when physical layer checks out OK
- ✔Prevents various failure conditions including crossed wiring

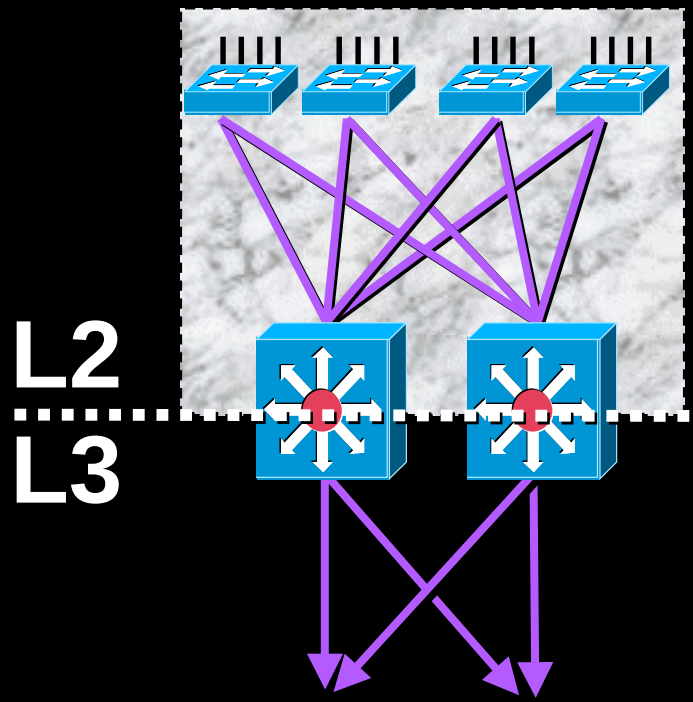
The lights are on, BUT …..



Building Block Means Survivable
Self-contained Backbone

- ✔ Autonomous Survivability Unit - HSRP
- ✔ L3 Broadcast Multicast demarcation
- ✔ Cookie cutter configuration
- ✔ L3 Demarcation of failure domain
- ✔ Simple, repeatable, deterministic
- ✔ Redundancy adds 15% cost at mission critical points like server farm

ASU
delimits failure domain



Building Block Templates

Use "As Is" or Combine

1) **Standard Model**

   simple, structured

2) **VLAN Model**

   more flexible

3) **Large Scale Server Farm Model**

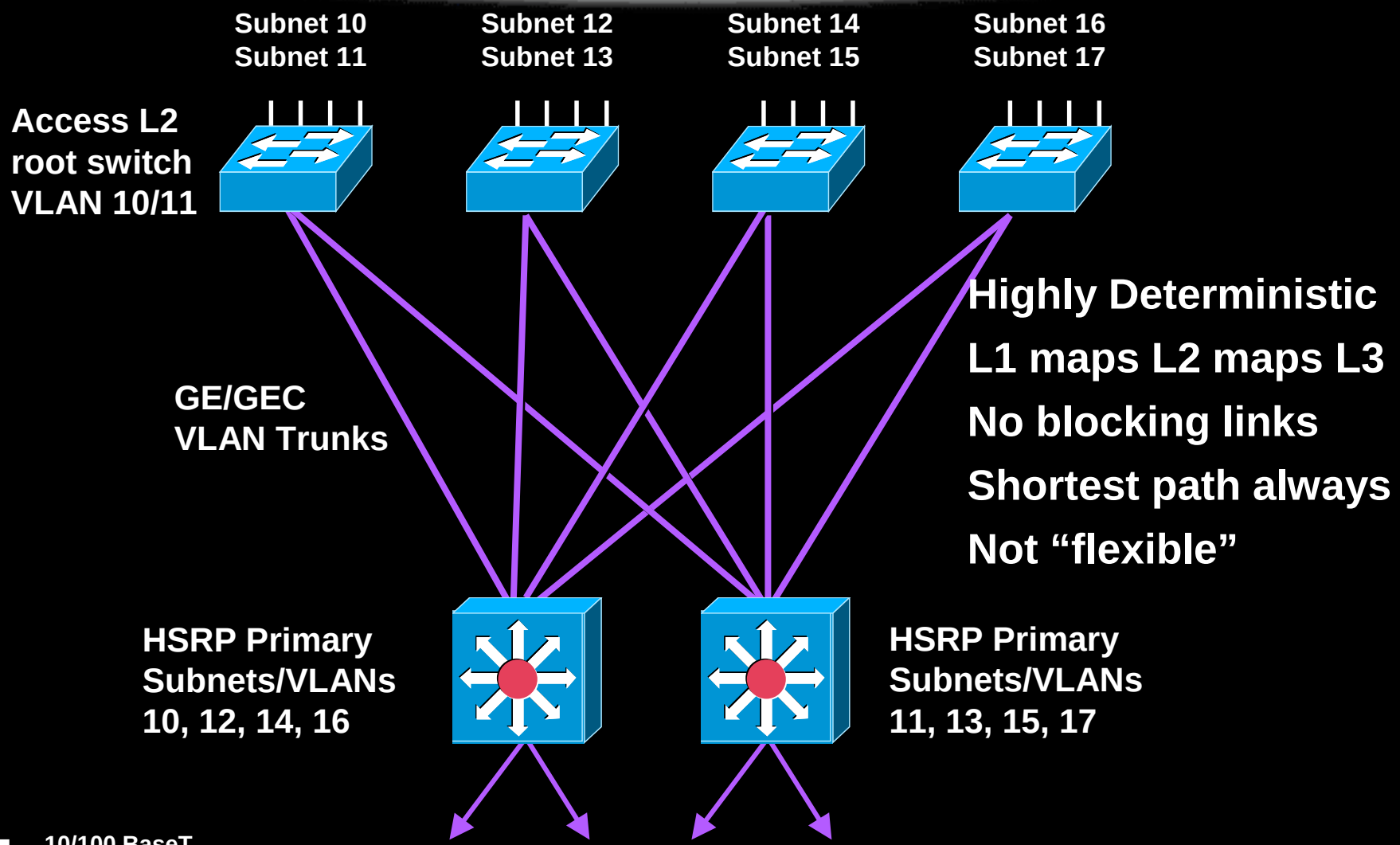
   accommodate dual NIC

4) **Small Scale Server Farm Model**

   accommodate dual NIC

1) Standard Building Block

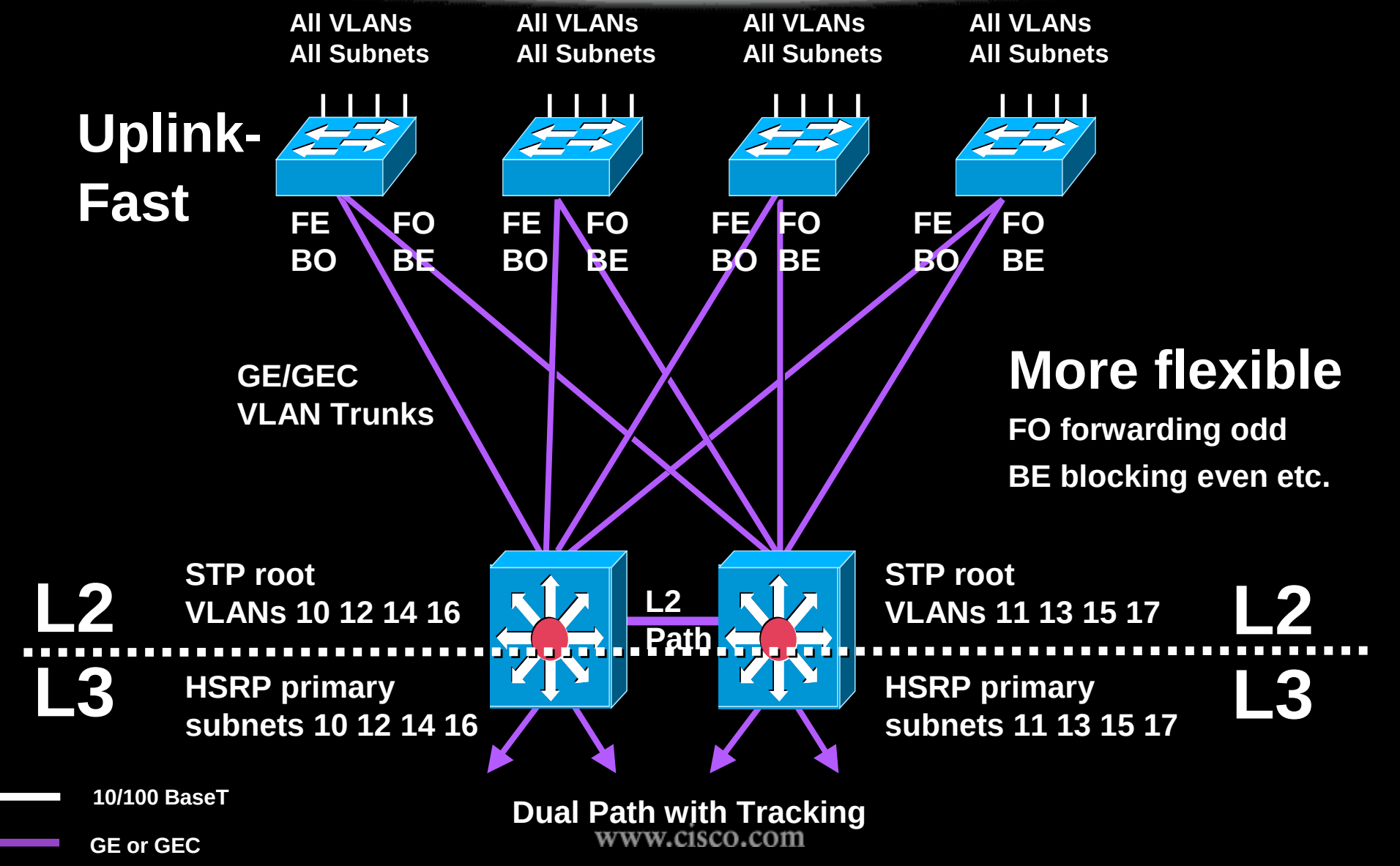
no loops - no STP complexity



2) VLAN Building Block

make L2 design match L3 design

All VLANs terminate at L3 boundary



3) Large-Scale Server Farm
Building Block

based on VLAN building block
aggregates traffic - high BW

Dual-NIC Server
Example Fault Tolerant Mode (FTM)
Same IP Address - seamless recovery

Access L2
UplinkFast

GE/GEC
VLAN Trunks

10/100 BaseT

GE or GEC

L2

STP root
VLANs EVEN

L2
Path

STP root
VLANs ODD

L2

L3

HSRP primary
subnets EVEN

HSRP primary
subnets ODD

L3

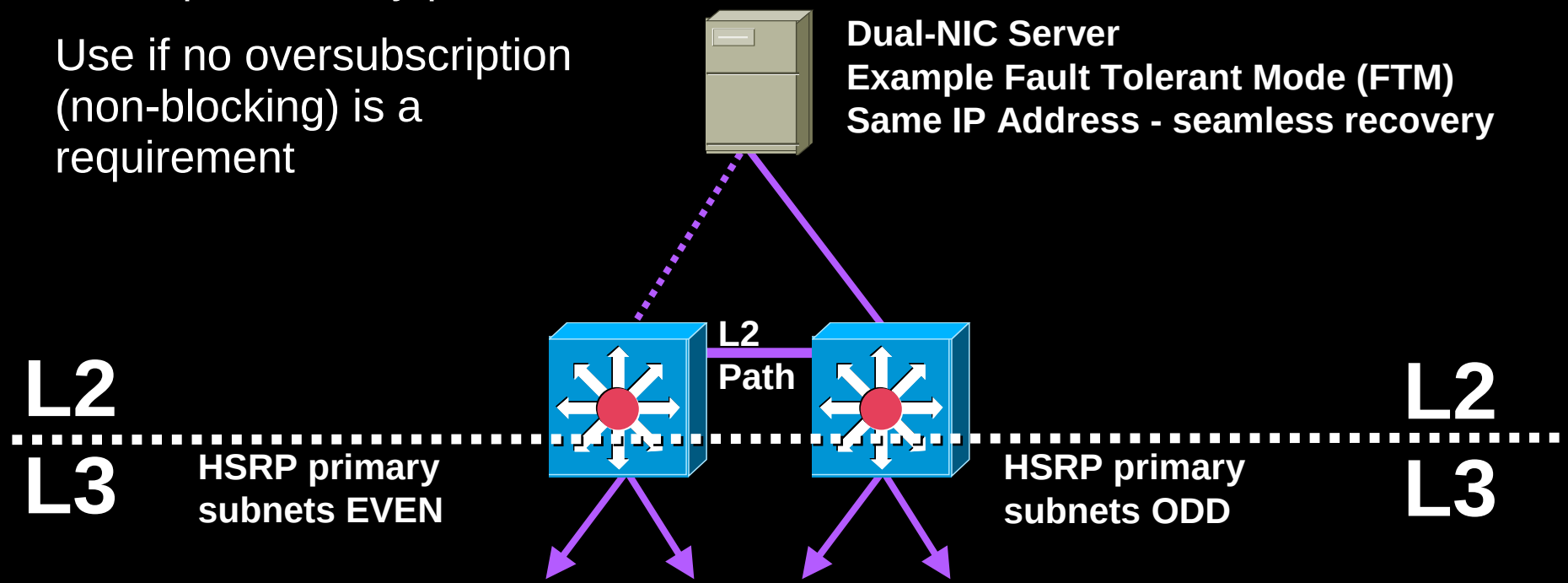
Dual Path with Tracking

4) Small-Scale Server Farm Building Block

Simplified building block with no STP loops

Use if port density permits

Use if no oversubscription (non-blocking) is a requirement

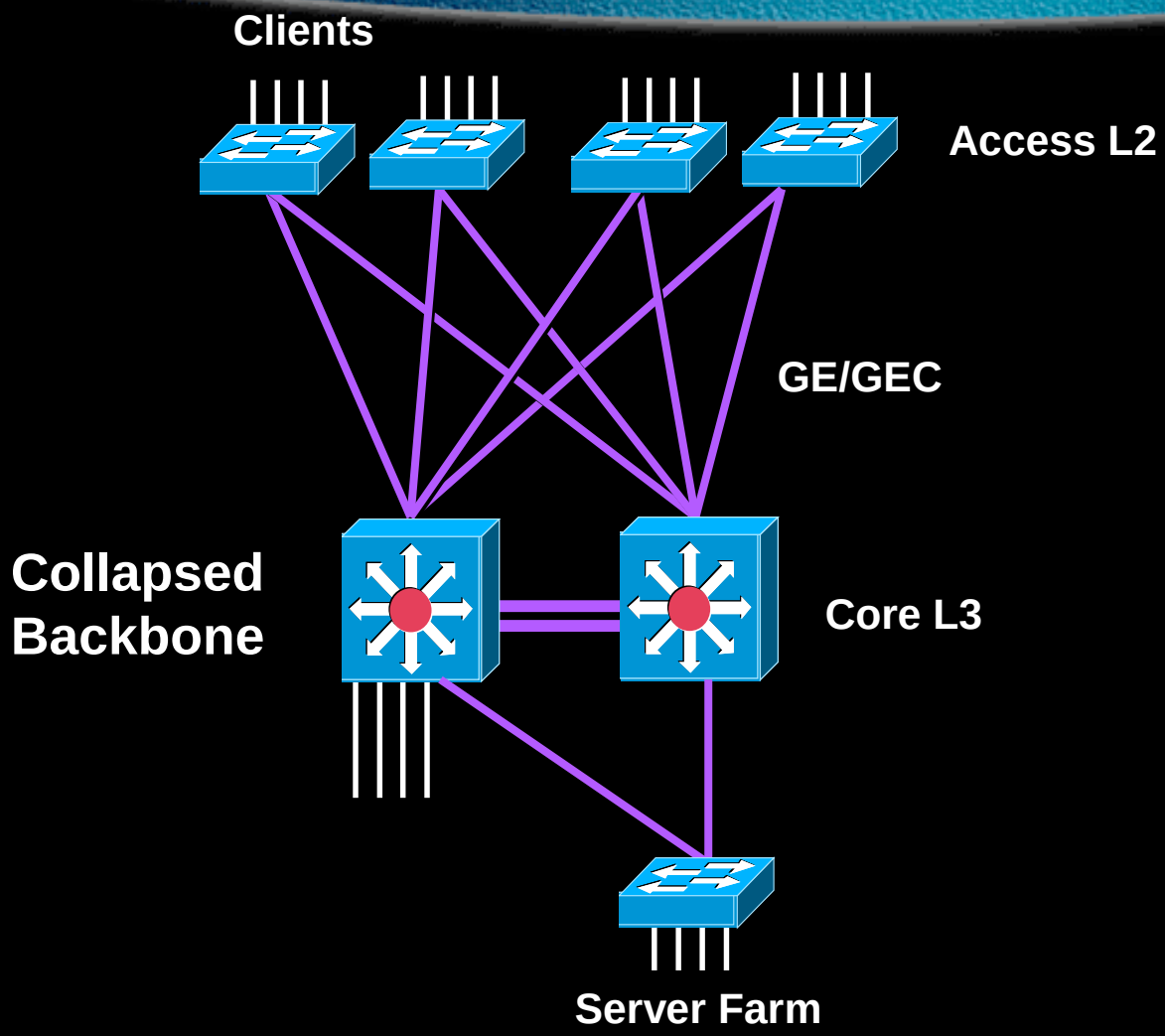


Redundant Backbone Models

all good - increasing scale

1) Collapsed L3 Backbone
2) Full Mesh
3) Partial Mesh
4) Dual-Path L2 Switched
5) Dual-Path L3 Switched

1) Collapsed L3 Backbone
large building or small campus

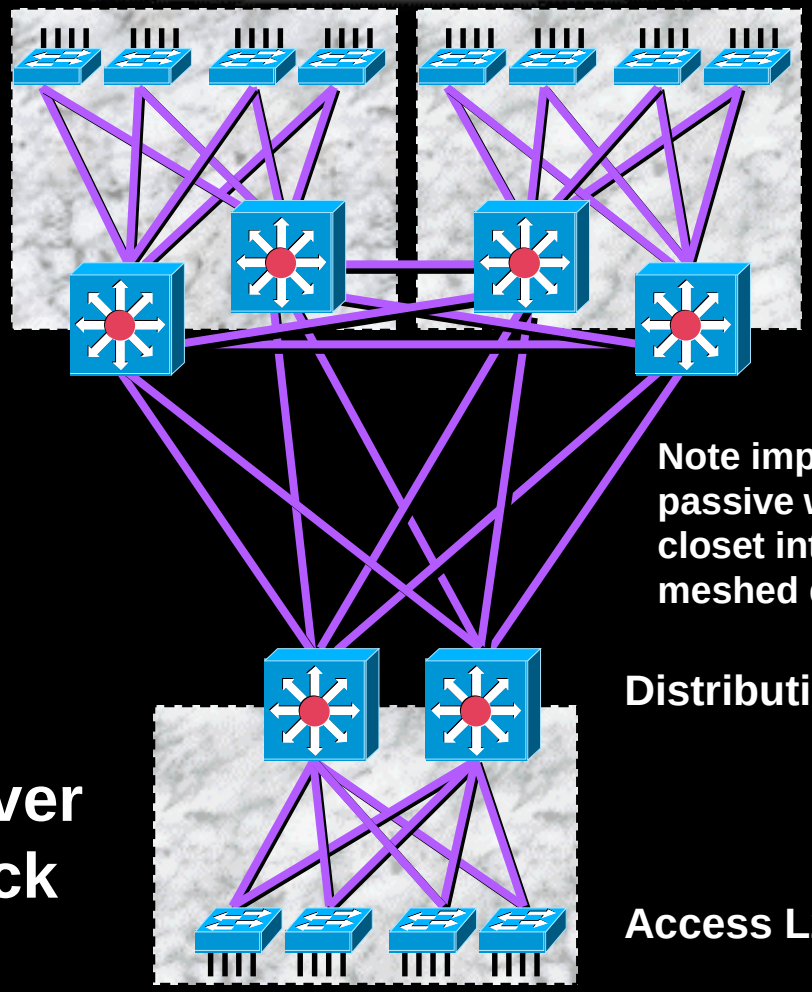


Scale depends on physical plant and policy more than performance

10/100 BaseT

GE or GEC

2) Full Mesh Backbone
small campus - n squared limitation



Client Blocks

Server Block

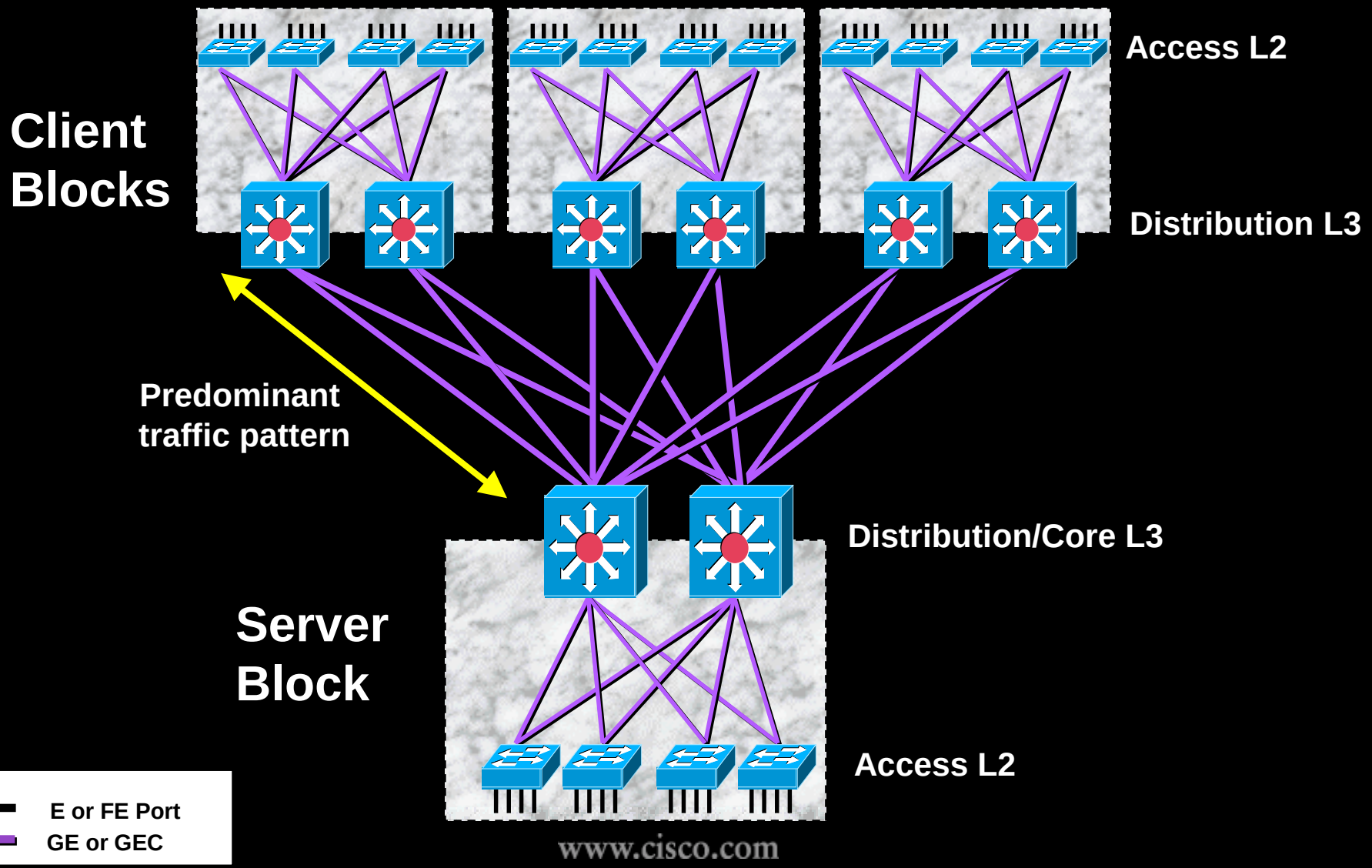
2 blocks - 6 peerings
3 blocks - 15 peerings
4 blocks - 28 peerings
5 blocks - 45 peerings

Note importance of passive wiring closet interfaces in meshed designs!

E or FE Port
GE or GEC

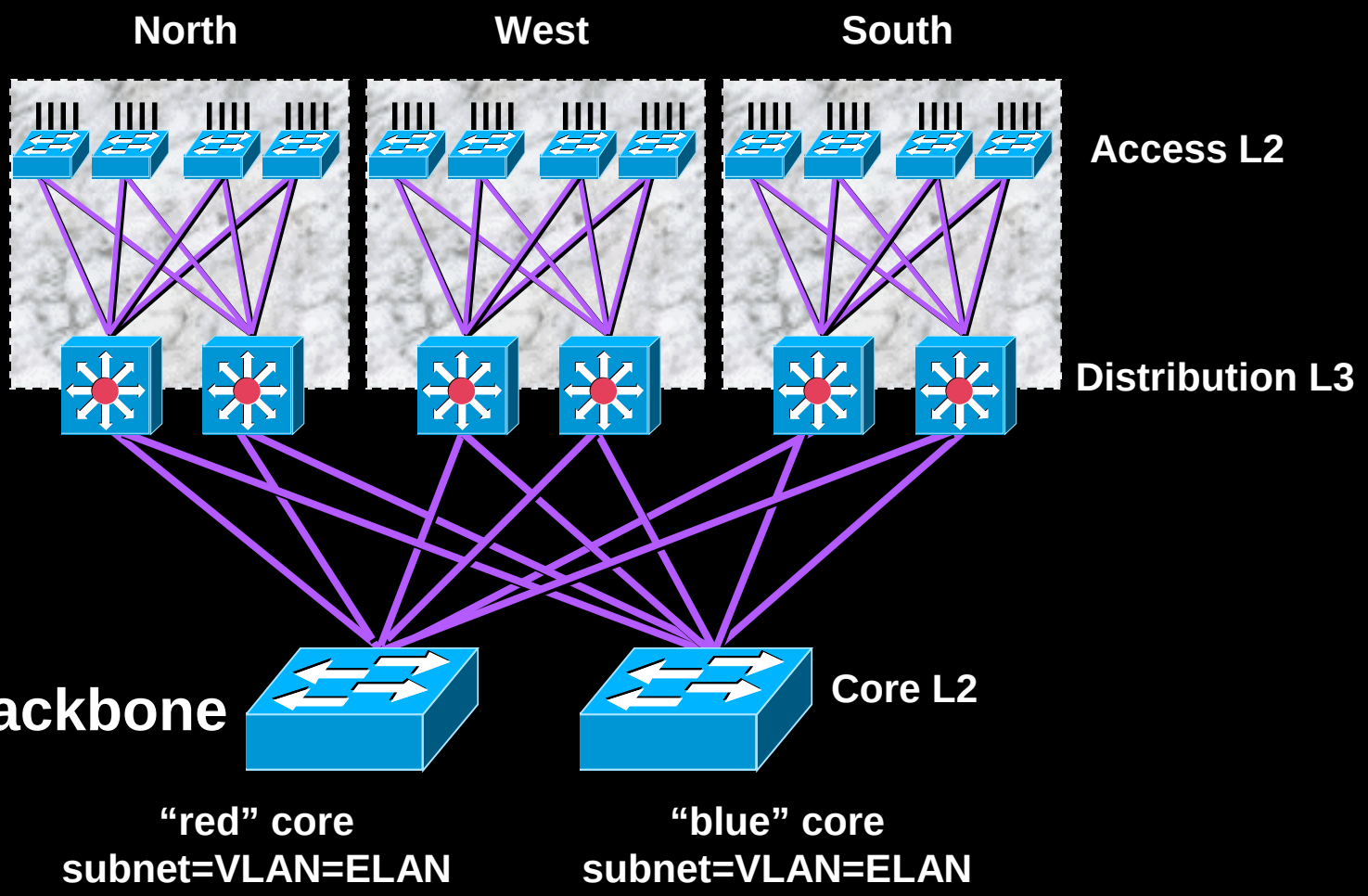
3) Partial Mesh Backbone

medium campus - traffic flow to server farm



4) Dual-Path L2 Switched Backbone

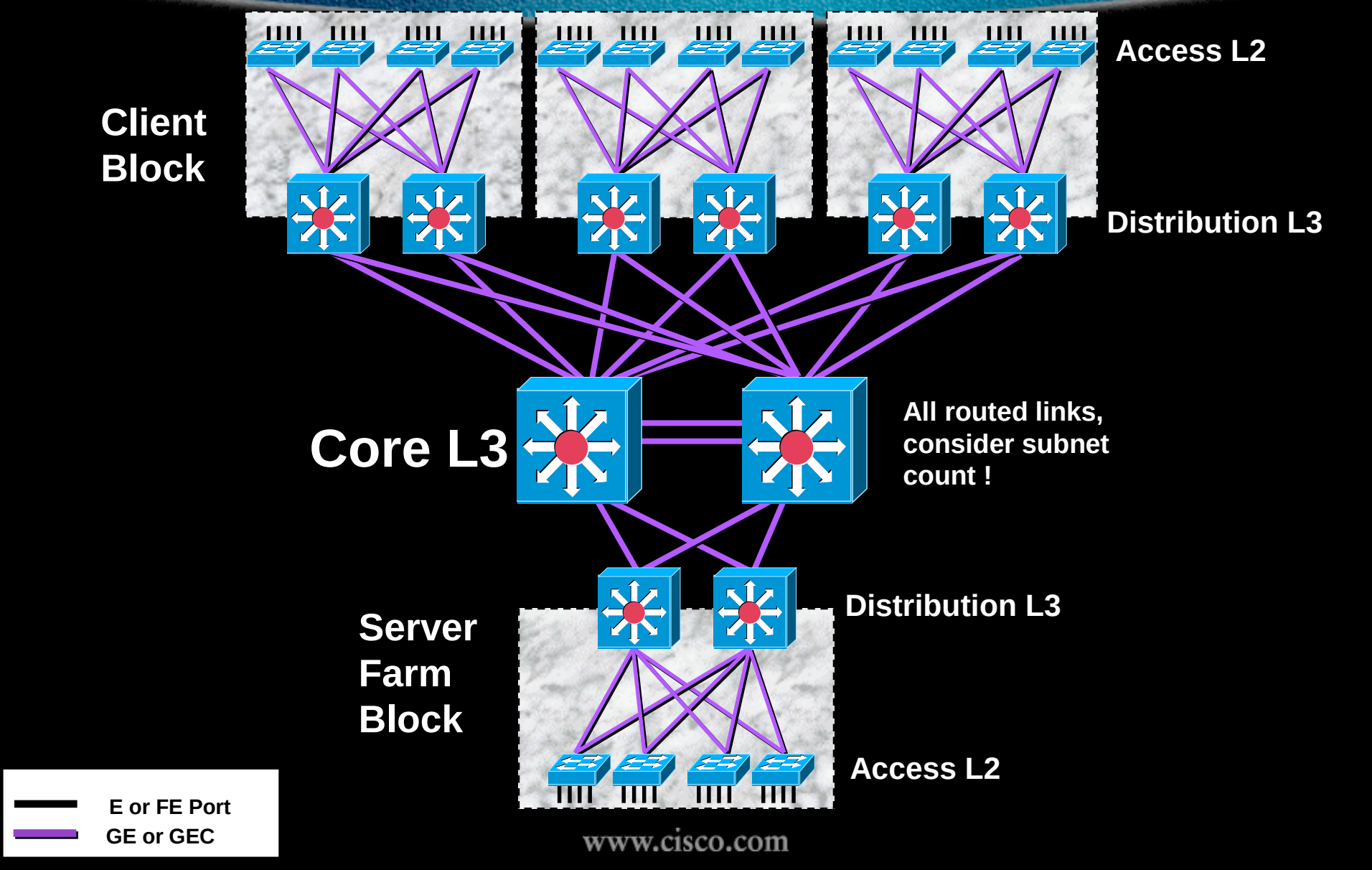
no STP loops or VLAN trunks in core



5a) Benefits of a L3 Backbone

- ✔ Multicast PIM routing control
- ✔ Load balancing
- ✔ No blocked links
- ✔ Fast convergence EIGRP/OSPF
- ✔ Greater scalability overall
- ✔ Router peering reduced
- ✔ IOS features in the backbone

5b) Dual-Path L3 Backbone
largest scale, intelligent multicast



Restore Considerations

- ✔ Restoring can take longer in some cases - more complex - schedule
- ✔ On power up L1 may come up before L3 builds routing table - temporary black hole for HSRP
- ✔ Use “preempt delay” for HSRP

Campus Failover Layer 2
Recovery & Tuning

STP

- Tune ‘diameter’ on root switch
- Improves recovery time maxage

UplinkFast

- No tuning, 2 seconds, wiring closet only
- Only applies with forwarding & blocking link

PortFast

- Server or desktop ports only 1 s
- Move directly from linkup into forwarding

Backbonefast

- Converges 2 sec + 2xFwd_delay for indirect link failures
- Eliminates maxage timeout

Campus Failover Layer 3
Recovery & Tuning

Caution with aggressive tuning

Good when network is stable, highly summarized

OSPF (fast LAN links)

Tune hello timer 1 sec, dead timer 3 sec

<4s to recognize problem, then converge

HSRP (fast LAN links)

Tune hello timer 1 sec, dead timer 3 sec

<4s to converge

EIGRP (fast LAN links)

Tune hello timer 1 sec, hold timer 3 sec

<4s to recognize problem, then converge

Keeping Networks Available!

- ✔KISS - eliminate complex L2
- ✔ASU - building blocks
- ✔Redundant backbone
- ✔Redundant L3 paths
- ✔L3 segments failure domain

EMPOWERING THE
INTERNET GENERATION℠

www.cisco.com