The bridge to possible

Сетевой марафон Cisco:Классика LAN

Обеспечение высокой доступности в сетях кампусов. Часть 1.

Михаил Окунев
Системный Архитектор
23 марта 2021

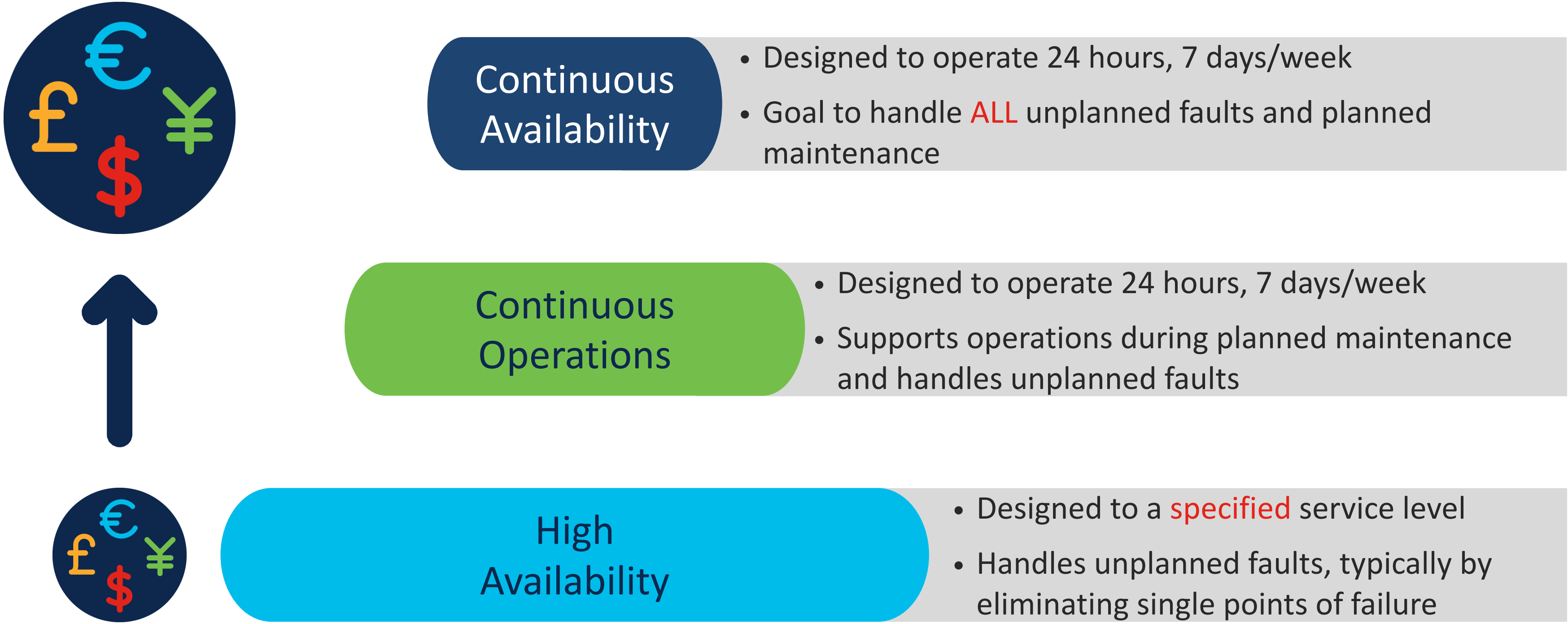
Высокая доступность:

Что это такое?

Что такое высокая доступность?

Уровни доступности

Фактическая отраслевая терминология



Continuous Availability

- Designed to operate 24 hours, 7 days/week
- Goal to handle ALL unplanned faults and planned maintenance

Continuous Operations

- Designed to operate 24 hours, 7 days/week
- Supports operations during planned maintenance and handles unplanned faults

High Availability

- Designed to a specified service level
- Handles unplanned faults, typically by eliminating single points of failure

Сложная задача...

“I need to design and deploy a network.”

Future ready

On time

Within budget

Manageable

Best practices

Design options

Platform choices

“Девятки” – Доступность сети и время простоя

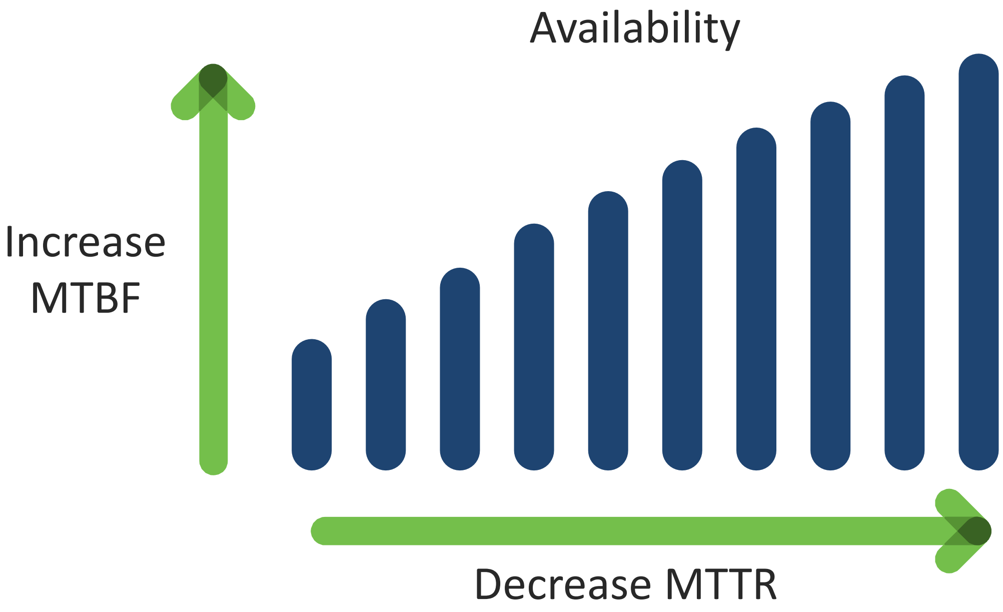
Network availability: amount of uptime of a network system over a specific time interval, measured as a percentage.

Availability	Downtime per year
90%	36 ½ days
99%	3 days, 16 hours
99.9%	8 hours, 46 minutes
99.99%	52 minutes
99.999%	5 minutes

Как мы можем измерить прогнозируемую доступность?

It’s function of:

Mean Time Between Failures (MTBF) and Mean Time To Repair (MTTR)



Базовое уравнение прогнозируемой доступности
(прогнозируемый рейтинг доступности)

Predicted Availability Equation

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

MTBF: Mean Time Between Failures

MTTR: Mean Time To Repair

Пример расчетов прогнозируемой доступности

$$\text{Availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

- Component with MTBF=87,600 hours

24 hour depot replacement

2 hour 24 minutes predicted annual downtime

$$\text{Availability} = \frac{87{,}600}{87{,}600 + 24} = .9997\ (99.9\%)$$

4 hour depot replacement

24 minutes predicted annual downtime

$$\text{Availability} = \frac{87{,}600}{87{,}600 + 4} = .99995\ (99.99\%)$$

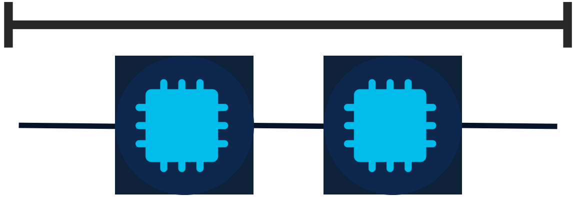
Warm spare (10 minute restore)

1 minute predicted annual downtime

$$\text{Availability} = \frac{87{,}600}{87{,}600 + .16666} = .999998\ (99.999\%)$$

Эффект резервирования для системы

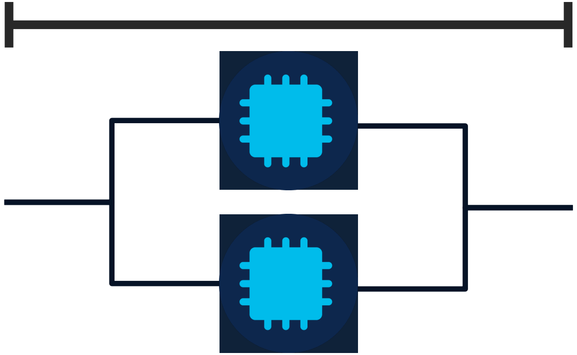
- Single components functioning in series
- System predicted availability:
  99.98%
  (~10 min./year predicted downtime)



Supervisor
99.999%
Availability
(~5 min/yr.
downtime)

Linecard
99.999%
Availability
(~5 min/yr.
downtime)

- Redundant components functioning in parallel
- System predicted availability:
  99.999999%
  (~½ second/year predicted downtime)



Linecards
99.999% Availability (~5 min/yr. downtime)

Пример прогнозируемой доступности Catalyst 6800XL (без резервирования)

Catalyst 6800XL

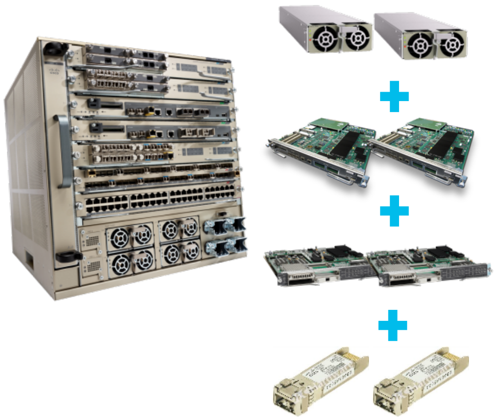
Part	MTBF (hours)	MTTR	Combined MTBF Hrs.	Combined Availability	Predicted Annual Downtime
Chassis C6807-XL	638,440	4 hrs.	638,440	99.99937348%	--
C6807-XL-FAN	3,077,880	4 hrs.	3,077,880	99.99987004%	--
SFP-10GSR	2,294,776	4 hrs.	2,294,776	99.99982569%	--
Supervisor VS-S2T-10G	231,910	4 hrs.	231,910	99.99827522%	--
WS-X6904-40G-2T	256,490	4 hrs.	256,490	99.99844051%	--
C6800-XL-3KW-AC	3,000,000	4 hrs.	3,000,000	99.99986667%	--
System MTBF			**91,987**	**99.99565168%**	**22.87 min.**

Components combined in **series** calculation

Chassis X Fan Tray X Power Supply X Line Card X Supervisor Module X SFP Uplink = System MTBF

Пример прогнозируемой доступности Catalyst 6800XL (с резервированием)

Catalyst 6800XL with Redundancy



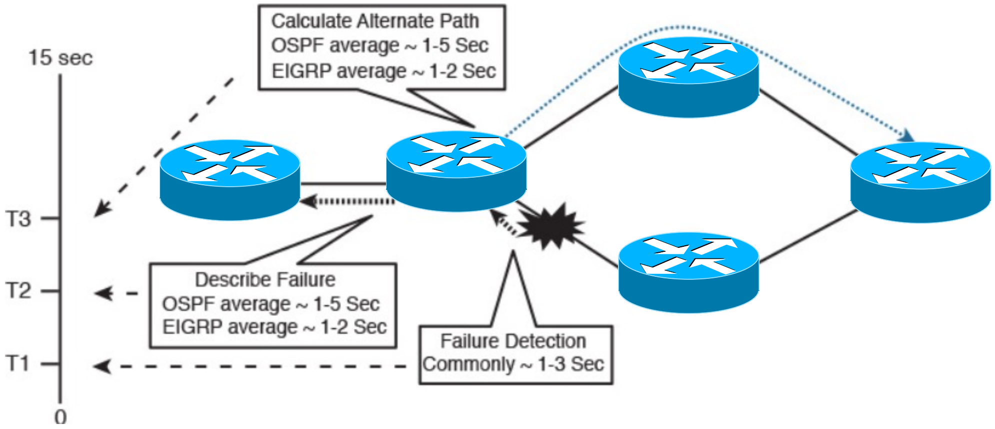
Part	MTBF Hrs.	MTTR Hrs.	Switchover time (seconds)	Combined MTBF Hrs.	Combined Availability	Predicted Annual Downtime
Chassis C6807-XL	638,444	4 Hrs.	--	638,440	99.99937348%	--
C6807-XL-FAN=	3,077,880	4 Hrs.	--	3,077,880	99.99987004%	--
SFP-10GSR	451,610	4Hrs.	.500	2,633,000,739,868	100.00000000%	--
Supervisor VS-S2T-10G	2,294,776	4 Hrs.	.500	26,891,355,961	99.99999997%	--
WS-X6904-40G-2T	402,386	4 Hrs.	.500	32,893,816,541	99.99999998%	--
C6800-XL-3KW-AC	3,000,000	4 Hrs.	**0**	4,500,003,000,001	100.00000000%	--
System MTBF				**528,687**	**99.99924347%**	**3.98min.**

Redundant components combined in **parallel** calculation

Chassis X Combined Power Supply X Combined Line Card X Combined Supervisor Module X Combined SFP Uplink = System MTBF

Сходимость

- Time to restore connectivity after a disruptive network event
- How quickly and reliably a network convergence can occur depends on several elements:
  - Event detection
  - Event propagation
  - Event processing
  - Update routing and forwarding tables



Системный подход к доступности сети кампуса

- System-level resiliency
- Network-level redundancy
- Enhanced management
- Human ear notices the difference in voice within 150–200 msec
(10 consecutive G.711 packet loss)
- Video loss is even more noticeable
- 200 msec typical end-to-end campus convergence target

Ultimate goal – 100% availability

Examples:

- Next-generation applications, video conferencing, unified messaging, e-business, wireless
- Mission-critical applications, databases, order entry, CRM, ERP
- Desktop applications, e-mail, file, print

An organization’s applications drive requirements for high availability networking

What if video delivery is key to your organization?

1920 lines of Vertical Resolution (Widescreen Aspect Ratio is 16:9)

1080 lines of Horizontal Resolution

1080p60

1080 x 1920 lines =

2,073,600 pixels per frame

x 24 bits of color per pixel

x 60 frames per second

= 2,985,984,000 bps

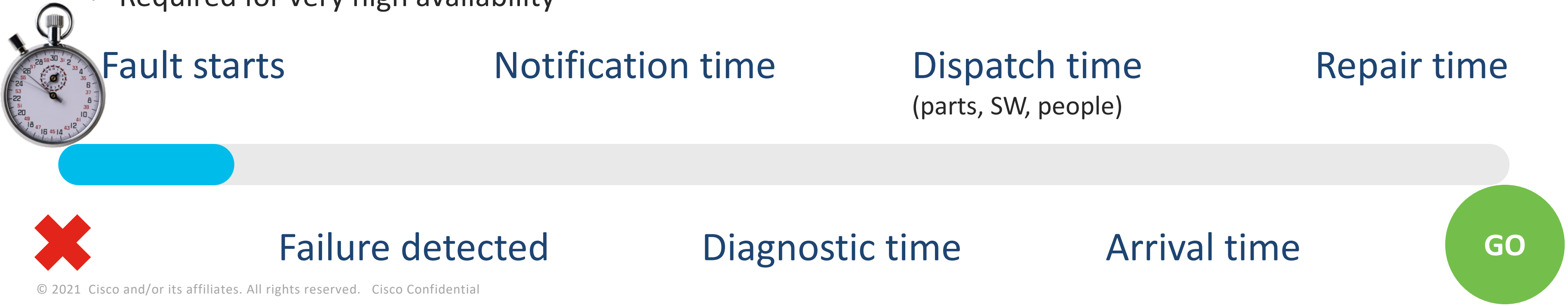
or 3 Gbps Uncompressed!

Cisco (H264/H.265) codecs transmit 3-5 Mbps per 1080p60 video stream (99.8%+ *compression, ~1000:1).*
Packet loss is proportionally magnified by compression ratios. Users can notice a single packet lost in 10,000.

HD video is ***one hundred times more sensitive to packet loss than VoIP!***

Measure and analyze event total service downtime

- Measure all previous events
  - Note each in trouble tickets
  - Analyze trends
- Automation
  - Trouble ticketing
  - Technology/database
- Redundant network design and resiliency features
  - Required for very high availability



Fault starts

Notification time

Dispatch time
(parts, SW, people)

Repair time

Failure detected

Diagnostic time

Arrival time

Примеры: измерение доступности сети

OSI model layers	Visibility / measurements
Application layer	Custom application scripts, HTML, TCL, Python, many others
Presentation layer	
Session layer	
Transport layer	ICMP ping, IP traceroute, Bidirectional Forwarding Detection, IP SLA
Network layer	
Data link layer	UDLD, BPDU, CDP, LLDP
Physical layer	Cable testers, power meters, OTDR

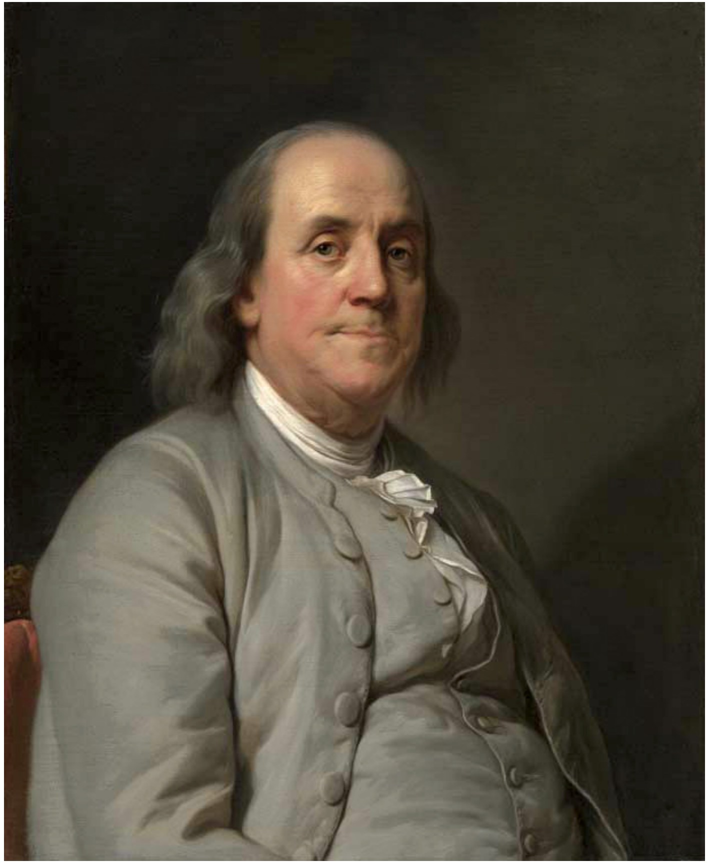
Высокая доступность:

Долго запрягаем, быстро едем ☺

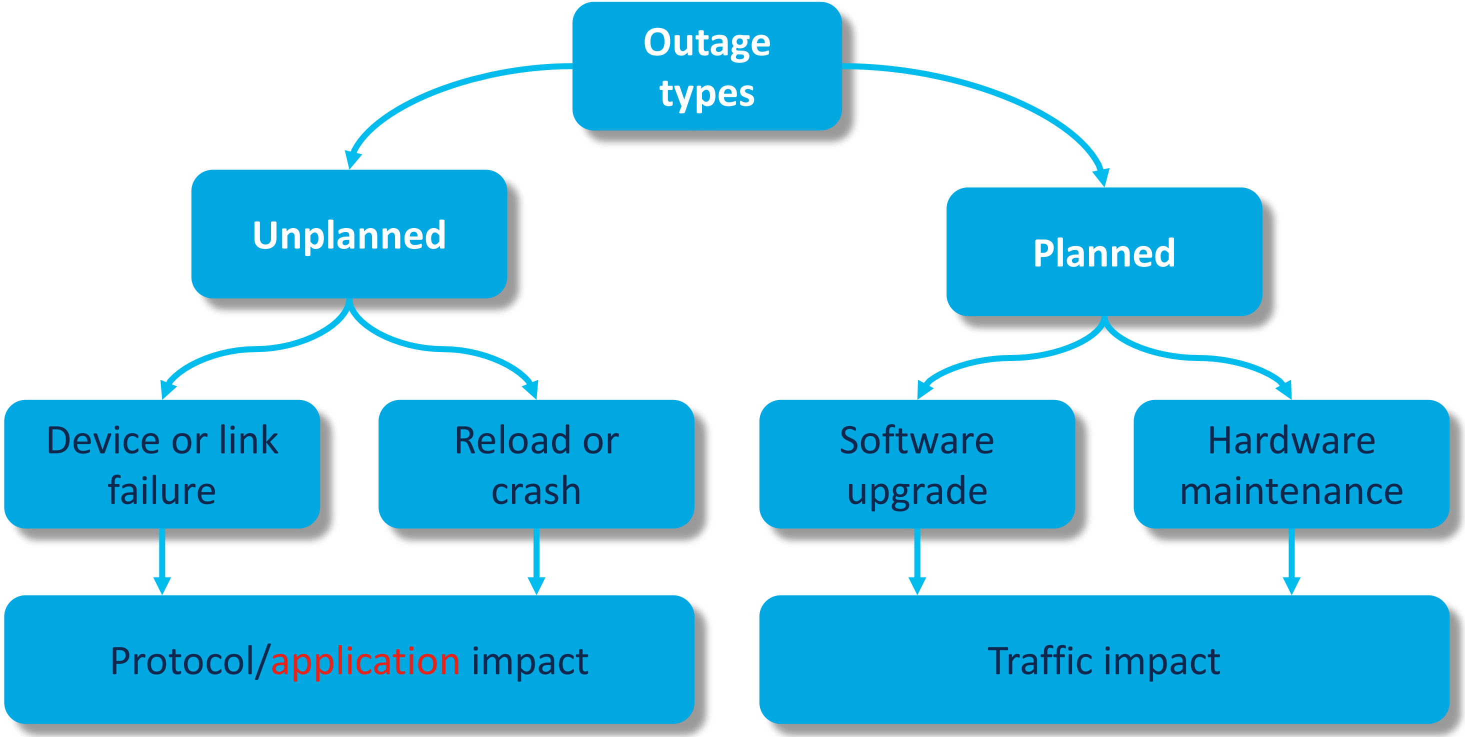
Какой русский не любит быстрой езды?

“By failing to prepare, you are preparing to fail.”

- Ben Franklin



Запланированные и незапланированные отказы



Outage types

Unplanned

Planned

Device or link failure

Reload or crash

Software upgrade

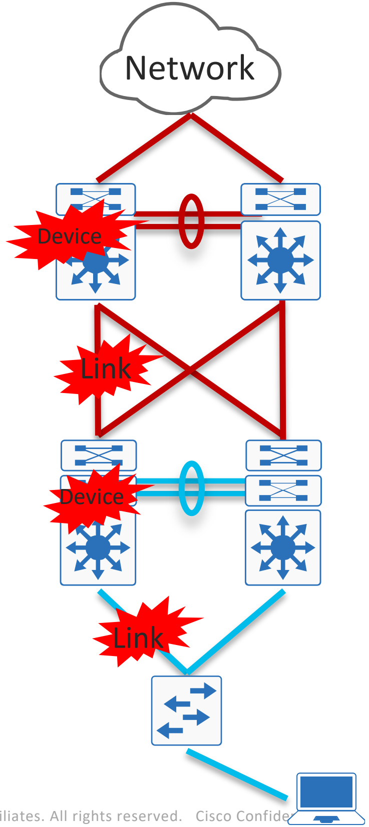
Hardware maintenance

Protocol/application impact

Traffic impact

Где могут случиться отказы?

Незапланированные отказы



Unplanned Outages]	
Link failure	Device failure
L2/L3 protocol failures	
Application failures	

Some platforms also support Process Restart

L2 Link

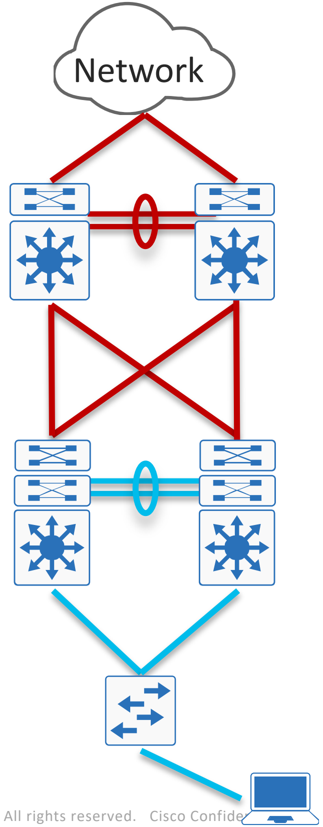
L3 Link

Failure

Solutions

Где могут случиться отказы?

Запланированные отказы



Planned Outages	
Software Upgrade	Hardware maintenance
All traffic impact	

Patching can be also used, depending on the upgrade

L2 Link

L3 Link

Solutions

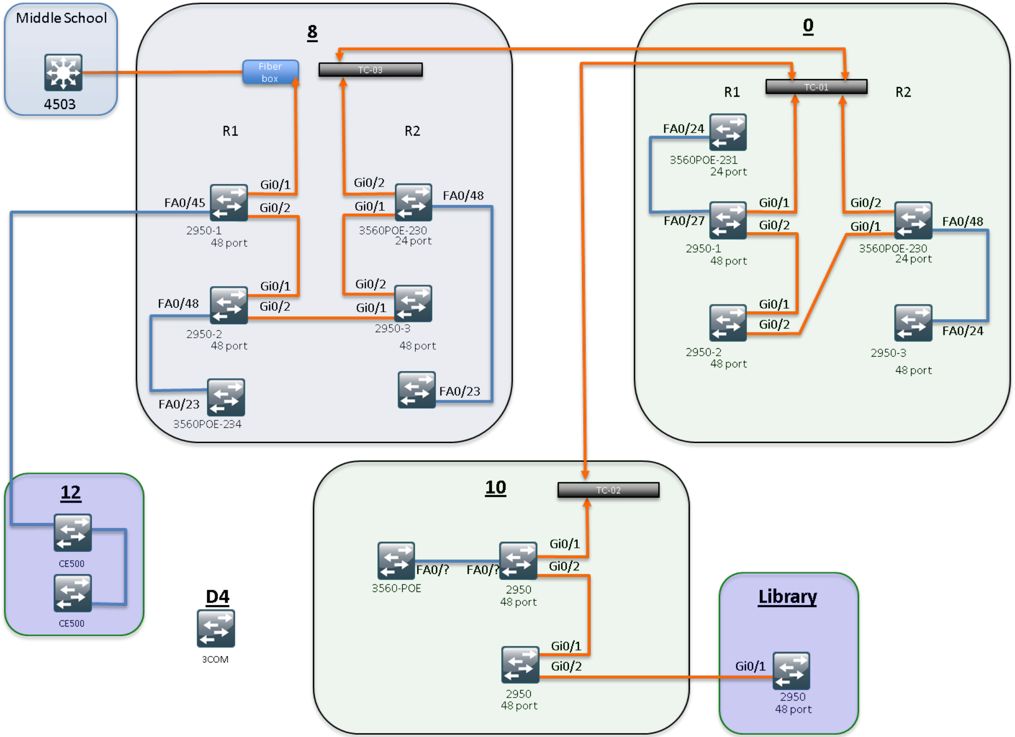
Высокая доступность:

Структурированный дизайн сети – основа высокой доступности

Чего мы стараемся избегать!

No hierarchy

Multiple single points of failure



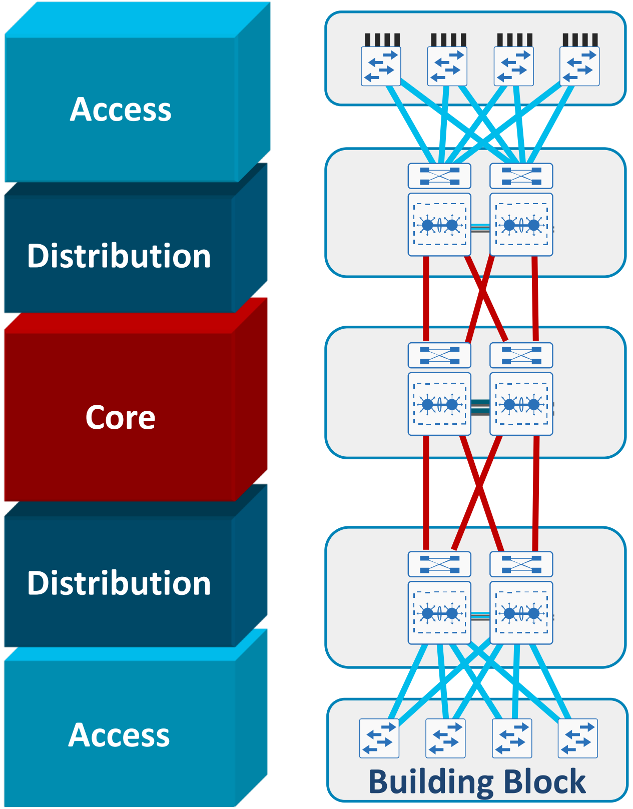
Hard to troubleshoot

Poor performance

Иерархический дизайн сети

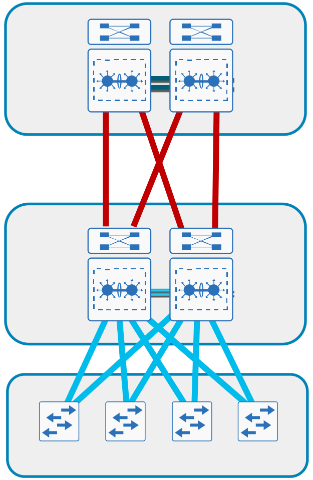
Высокая доступность за счет иерархии, модульности и структуры

- Hierarchical Design
  Each layer in hierarchy has a specific role
- Modular Design
  Modularity makes it easy to grow, understand, and troubleshoot
- Structured Design
  Creates small fault domains and predictable network behavior
  —clear demarcations and isolation
- Promotes load balancing and resilience



Иерархическая структура сети: проводная локальная сеть кампуса

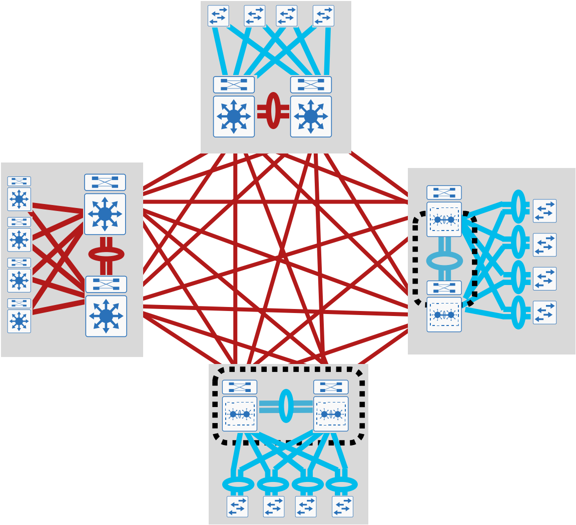
- Core
  - Connectivity, availability and scalability
- Distribution
  - Aggregation for wiring and traffic flows
  - Policy and network control point (FHRP, L3 summarization)
- Access
  - **Physical** – Ethernet wired 10/100/1000(802.3z)/mGig(802.3bz); 802.3af(PoE), 802.3at(PoE+), and Cisco Universal POE (UPOE)
  - **Policy enforcement** – security: 802.1x, port security, DAI, IPSG, DHCP snooping; identification: CDP/LLDP; QoS: policing, marking, queuing
  - **Traffic control** – IGMP snooping, broadcast control



Иерархическая структура сети: проводная сеть

Нужен ли мне уровень ядра?

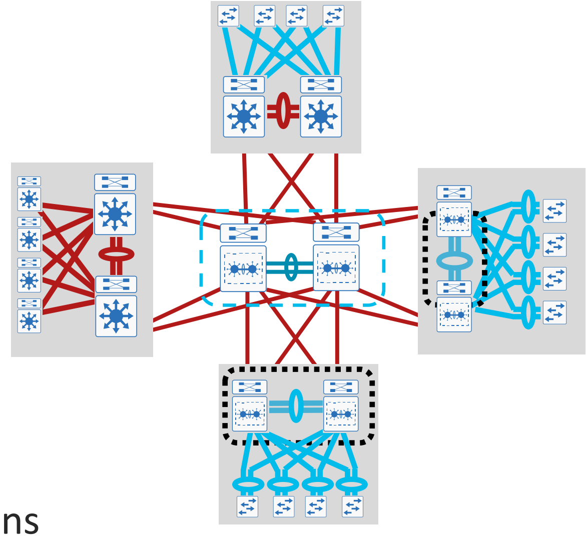
- It is a question of operational complexity and a question of scale
  - n x (n-1) scaling
  - Routing peers
  - Fiber, line cards, and port counts ($,€,₽)



Иерархическая структура сети: проводная сеть

Нужен ли мне уровень ядра?

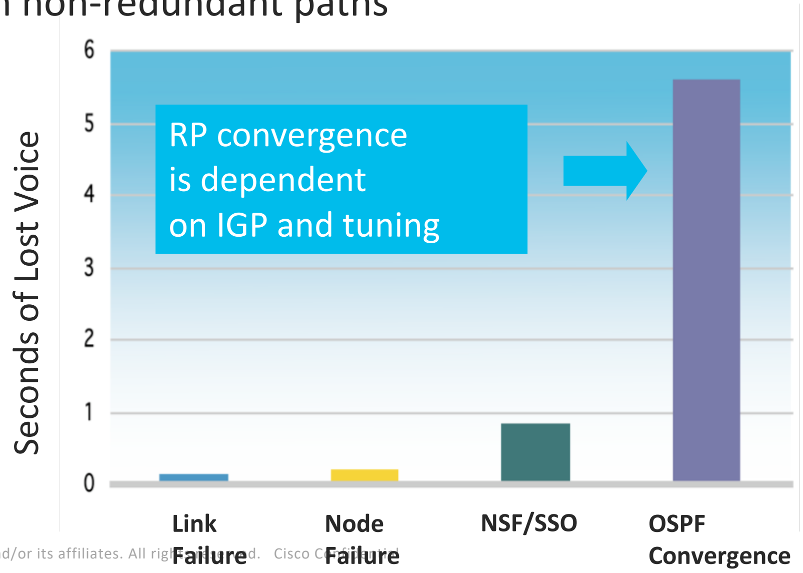
- It is a question of operational complexity and a question of scale
  - n x (n-1) scaling
  - Routing peers
  - Fiber, line cards, and port counts ($,€,£)
- Capacity planning considerations
  - Easier to track traffic flows from a block to the common core than to ‘n’ other blocks
- Geographic factors may also influence the design
  - Multi-building interconnections may have fiber limitations



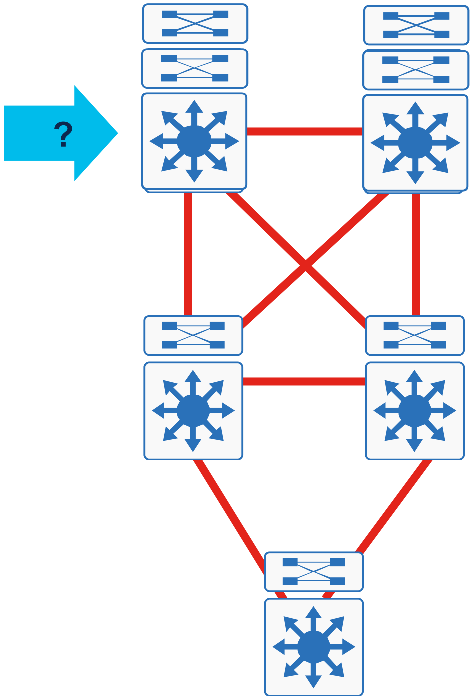
Резервирование шасси в ядре

Зависит от топологии

- Redundant topologies with equal cost multi-paths (ECMP) provide sub-second convergence
- NSF/SSO provides superior availability in environments with non-redundant paths



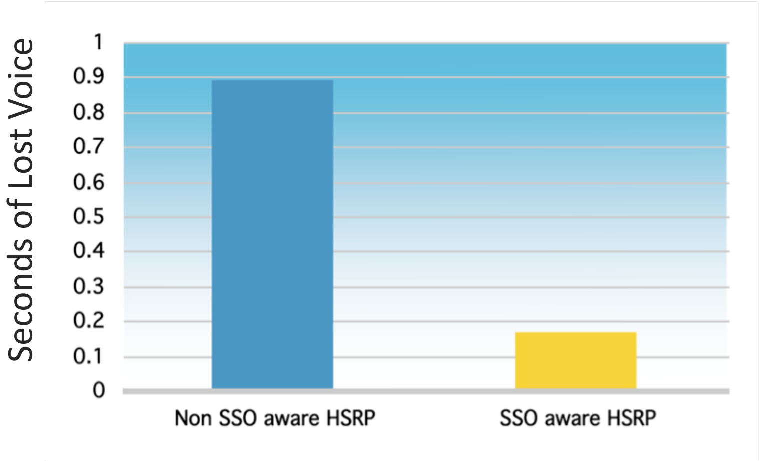
RP convergence is dependent on IGP and tuning

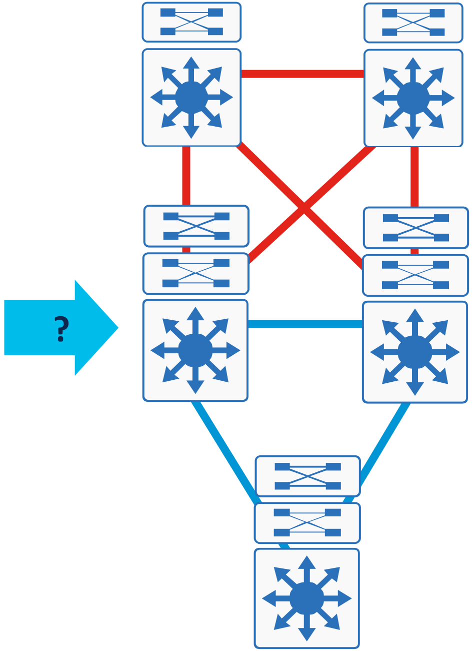


Резервирование шасси на уровне распределении

Рекомендуется

- HSRP doesn’t flap on Supervisor SSO switchover
- Reduces the need for sub-second HSRP timers

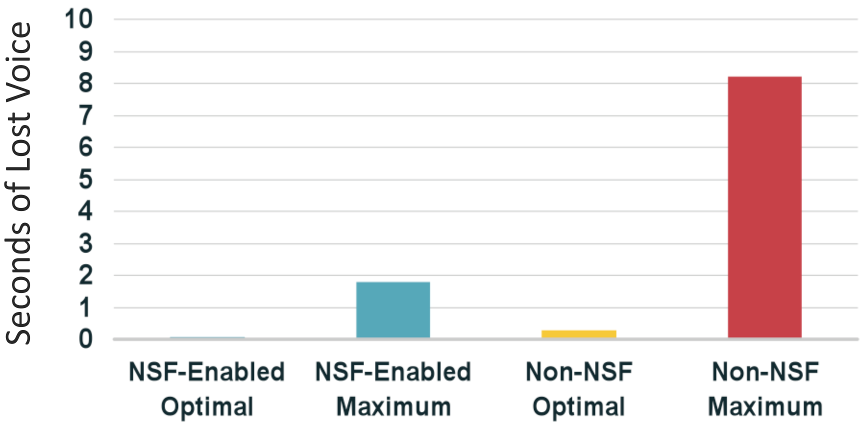


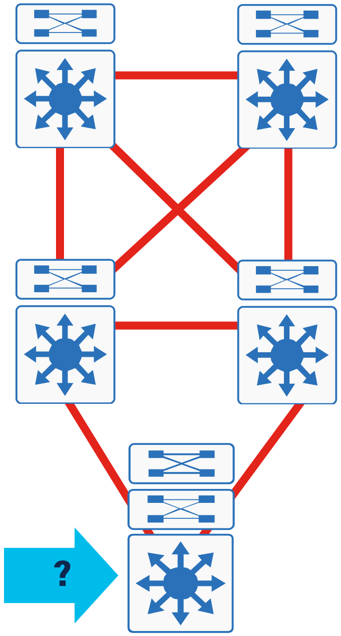


Резервирование шасси на уровне доступа

Рекомендуется для обеспечения наивысшей доступности

- Access switch is the single point of failure in best practices HA design
- Supervisor failure is most common cause of access switch service outages

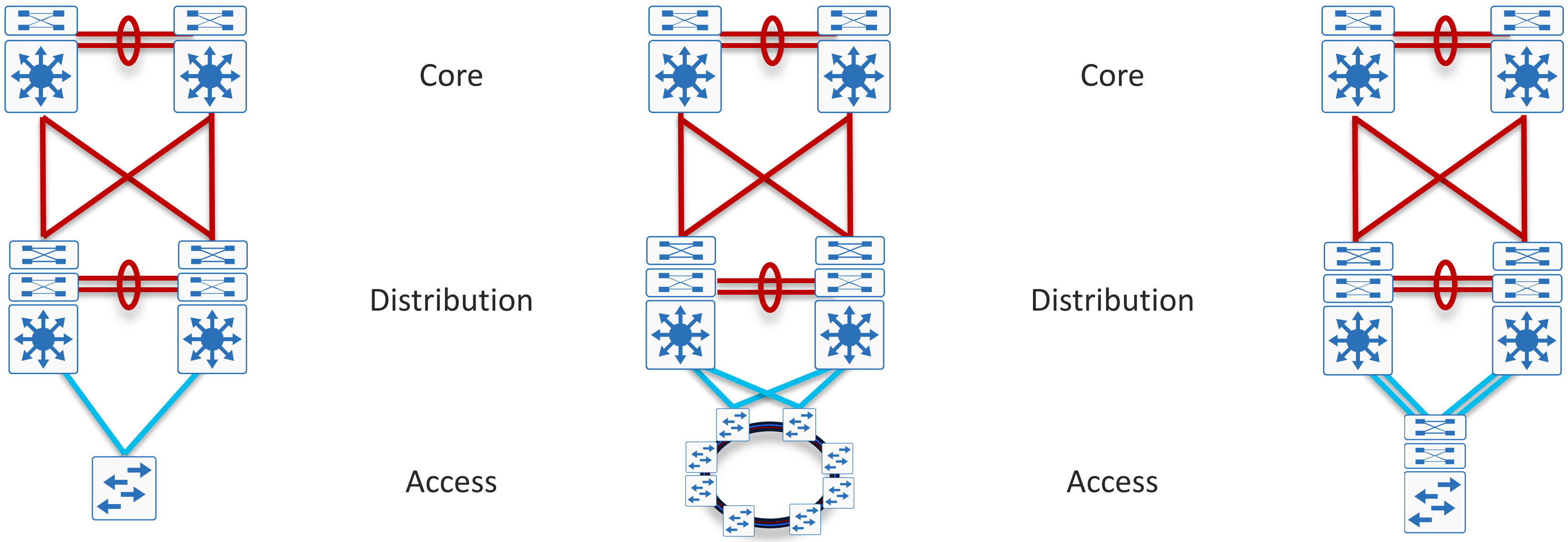




Высокая доступность:

Дизайн проводных сетей кампусов

Структурированный дизайн сети кампуса



Core

Distribution

Access

Core

Distribution

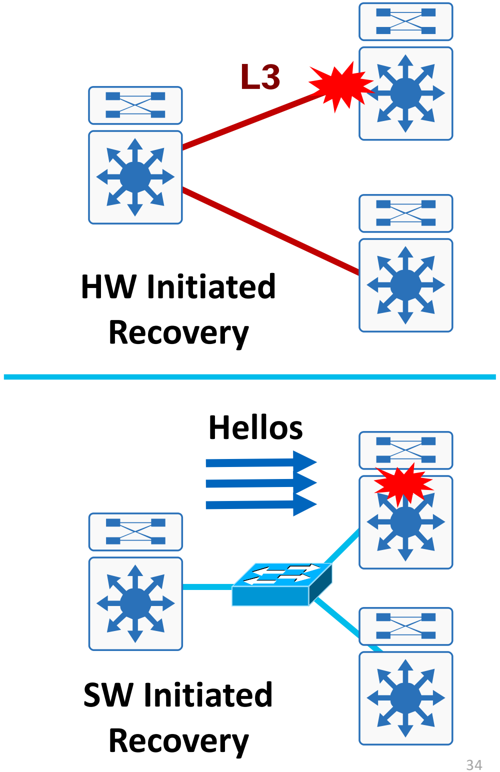
Access

- Optimize data load-sharing, redundancy design for best application performance
  - Diversify uplink network paths with cross-stack and dual-sup access-layer switches
  - Build distributed and full-mesh network paths between Distribution and Access-layer switches

Оптимизация сетевой конвергенции

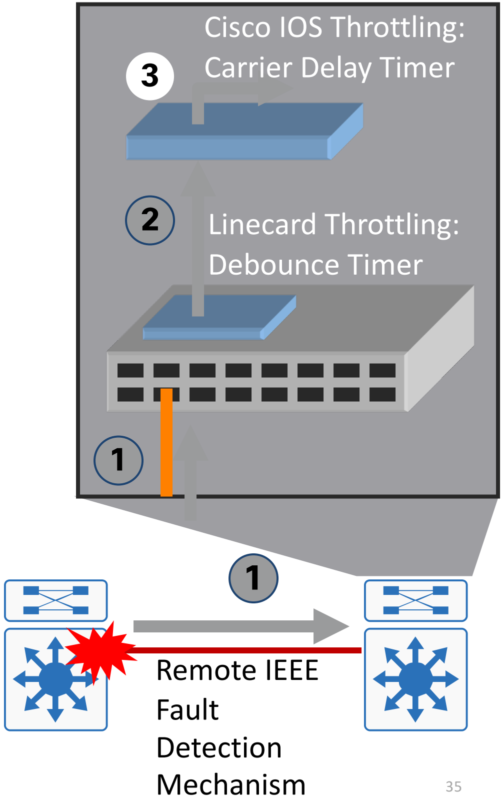
Обнаружение сбоев и восстановление сервиса

- Optimal high availability network design attempts to leverage ‘local’ switch fault detection and recovery
- Design should leverage the hardware capabilities of the switches to detect and recover traffic flows based on these ‘local’ events
- Design principle – Hardware failure detection and recovery is both faster and more deterministic
- Design principle – Software failure detection mechanisms provide a secondary, not primary, fault detection and recovery mechanism in the optimal design



Оптимизация сетевой сходимости
Обнаружение отказа соединения на L1

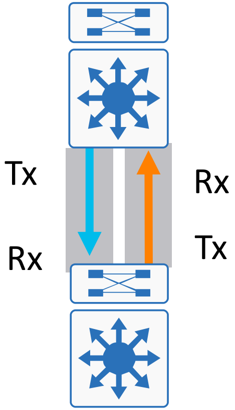
- Do not disable auto-negotiation on GigE /10GigE ports
- IEEE 802.3z and 802.3ae link negotiation define Remote Fault Indicator & Link Fault Signaling mechanisms
- IOS debounce –
  - GigE/10GigE fiber ports is 10 msec.; copper min. 300 msec.
  - NX-OS debounce – Currently 100 msec. by default
  - All 1G and 10G SFP / SFP+ based interfaces (MM, SM, CX-1) changing to a default of 10 msec.
  - RJ45 based Copper interfaces on NX-OS remains 100 msec.
- Design principle: Understand how hardware choices and tuning impact



Оптимизация сетевой сходимости

Программное обнаружение отказа соединения на L2 (e. g. UDLD)

- While 802.3z and 802.3ae link negotiation provide for L1 fault detection, hardware ASIC failures can still occur
- UDLD – L2 based keep-alive mechanism confirms bi-directional L2 connectivity
- Switch ports with UDLD send UDLD protocol packets (at L2) containing:
  port's own device / port ID
  neighbor's device / port IDs seen by UDLD on that port
- If port does not see its own device / port ID echoed by incoming UDLD packets, the link is considered unidirectional and is shutdown
- Design principle –
  Redundant fault detection mechanisms required
  (SW as a backup to HW as possible)



UDLD Keepalive

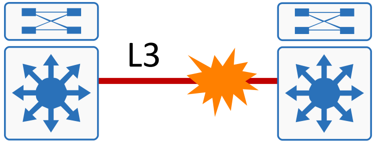
Оптимизация сетевой сходимости

L2 и 3 – Зачем использовать Routed интерфейс?

L3 routed interfaces allow faster convergence than L2 switchport with an associated L3 SVI

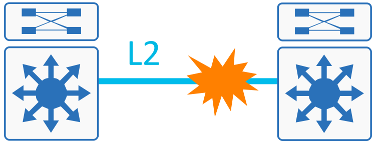
~ 8 msec loss

1. Link Down
2. Interface Down
3. Routing Update



```
21:38:37.042 UTC: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet3/1, changed state to down
21:38:37.050 UTC: %LINK-3-UPDOWN: Interface GigabitEthernet3/1, changed state to down
21:38:37.050 UTC: IP-EIGRP(Default-IP-Routing-Table:100): Callback: route_adjust GigabitEthernet3/1
```

1. Link Down
2. Interface Down
3. Autostate
4. SVI Down
5. Routing Update



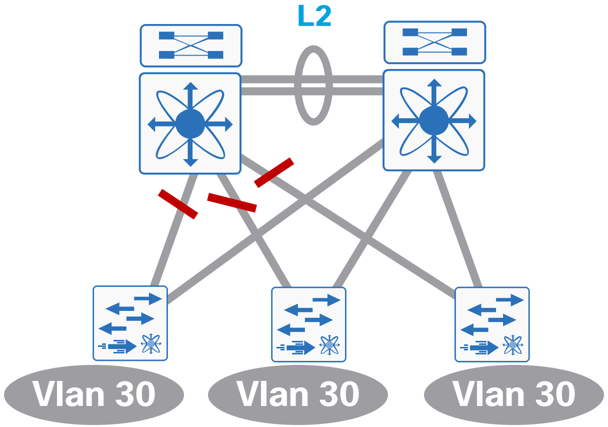
~ 200-250 msec. loss

```
21:32:47.813 UTC: %LINEPROTO-5-UPDOWN: Line protocol on Interface GigabitEthernet2/1, changed state to down
21:32:47.821 UTC: %LINK-3-UPDOWN: Interface GigabitEthernet2/1, changed state to down
21:32:48.069 UTC: %LINK-3-UPDOWN: Interface Vlan301, changed state to down
21:32:48.069 UTC: IP-EIGRP(Default-IP-Routing-Table:100): Callback: route, adjust Vlan301
```

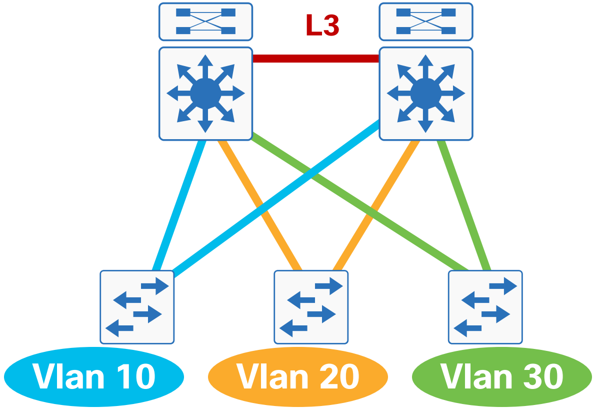
Высокая доступность:

Традиционный многоуровневый дизайн кампуса

Оптимизация Layer 2 дизайна – Spanning Tree



- At least some VLANs span multiple access switches
- Layer 2 loops
- Layer 2 and 3 running over link between distr.
- Blocked links
- More typical of a "classic" data center design

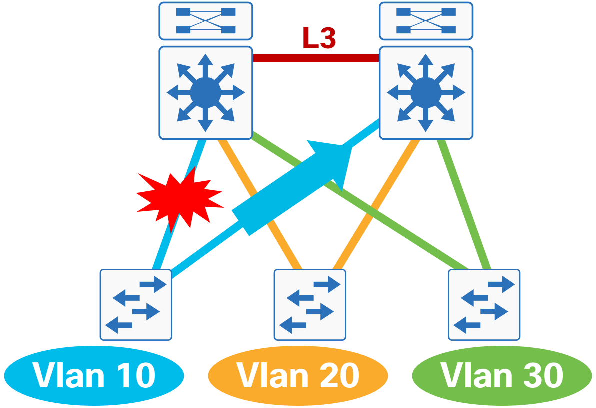


- Each access switch has unique VLANs
- No Layer 2 loops
- Layer 3 link between distribution
- No blocked links
- More typical of a campus LAN design

Оптимизация Layer 2 дизайна

Топологии STP без блокировок сходятся быстрее всего

- When STP is not blocking uplinks, recovery of access to distribution link failures is accomplished **based on L2 CAM updates** not on the Spanning Tree protocol recovery
- Time to restore traffic flows is based on: Time to detect link failure + Time to purge the HW CAM table and begin to flood the traffic
- No dependence on external events (no need to wait for Spanning Tree convergence)
- Behavior is **deterministic**



- All links forwarding – In an environment with all Links active, traffic is restored based on **HW recovery**

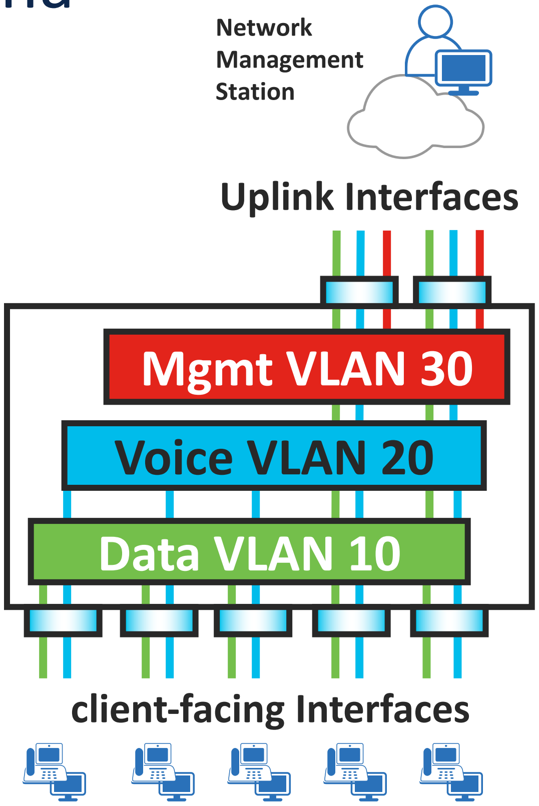
Виртуальные LANs на уровне доступа

Конфигурация коммутатора доступа

- **Data VLAN** provides access to the network for all attached devices other than IP Phones
- **Voice VLAN** for IP Phone network access
- **Management VLAN** for in-band access to the network for the switches management interface

```
vlan 10
  name Data
vlan 20
  name Voice
vlan 30
  name Management
```

⚠ Note: The management VLAN is never configured on user facing interfaces

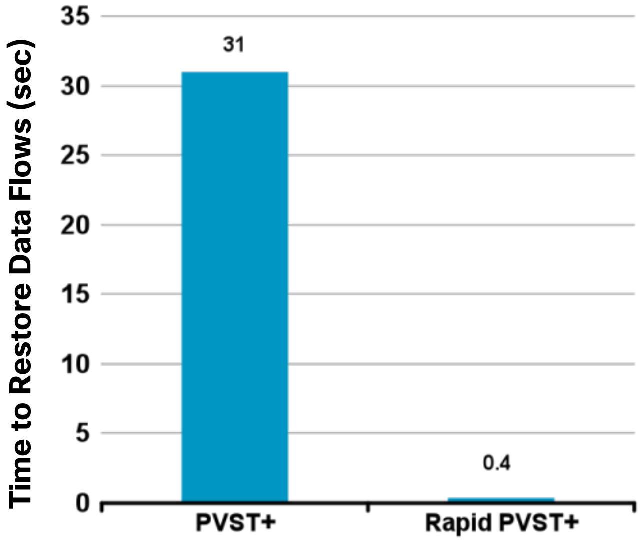


Предотвращение петель

- STP tuning (loopguard, rootguard, bpduguard, etc...)
- UDLD – Mitigates one way physical connection
- Bridge Assurance – Immediate blocking if BPDU is not received
- Flex Link – Backup/monitoring link with no STP
- Resilient Ethernet Protocol – Ring topology with fast failover

Оптимизация Layer 2 дизайна PVST+, Rapid PVST+, MST

- PVST+ (pre 802.1D-2004) - traditional spanning tree
- Rapid-PVST+ (802.1w)
  greatly improves the restoration times for any VLAN that requires a topology convergence due to link UP
- Rapid-PVST+ also greatly improves convergence time over BackboneFast for any indirect link failures
- Rapid PVST+
  Scales to large size (up to 16,000 logical ports)
  Easy to implement, proven, scales
- MST (802.1s)
  Permits very large scale STP implementations
  (up to 75,000 logical ports)

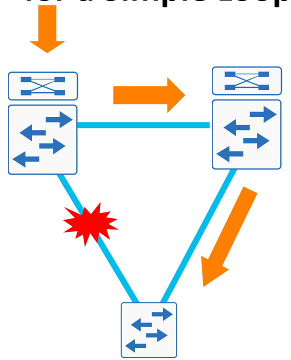


Оптимизация Layer 2 дизайна

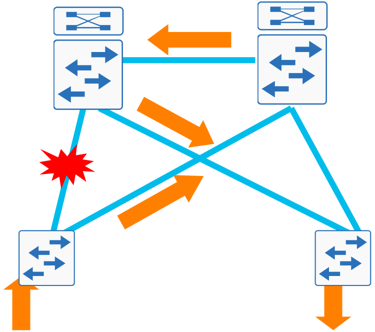
Сложные топологии сходятся дольше

- Time to converge is dependent on the protocol implemented: 802.1D, 802.1s, or 802.1w
- It is also dependent on:
  - Size and shape of the L2 topology (how deep is the tree)
  - Number of VLANs being trunked across each link
  - Number of logical ports in the VLAN on each switch
- Non-congruent topologies take longer to converge. Restricting the topology to reduce convergence times
- Prune all unnecessary VLANs from trunk configuration

400 msec Convergence for a Simple Loop



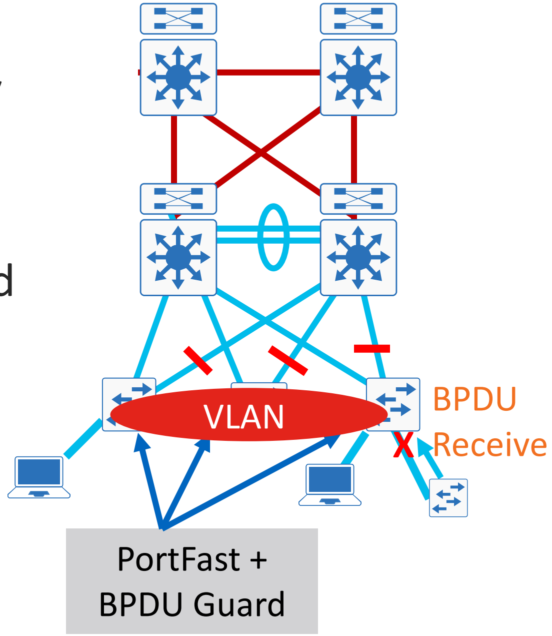
900 msec Convergence for a More Complex Loop



Оптимизация Layer 2 дизайна

Инструментарий STP – PortFast и BPDU guard

- PortFast is configured on edge ports to allow them to quickly move to forwarding bypassing listening and learning and avoids TCN (Topology Change Notification) messages
- BPDU guard can prevent loops by moving PortFast configured interfaces that receive BPDUs to errdisable state
- BPDU guard prevents ports configured with PortFast from being incorrectly connected to another switch
- When enabled globally, BPDU guard applies to all interfaces that are in an operational PortFast state



```
Switch(config-if)#spanning-tree portfast
Switch(config-if)#spanning-tree bpduguard enable
```

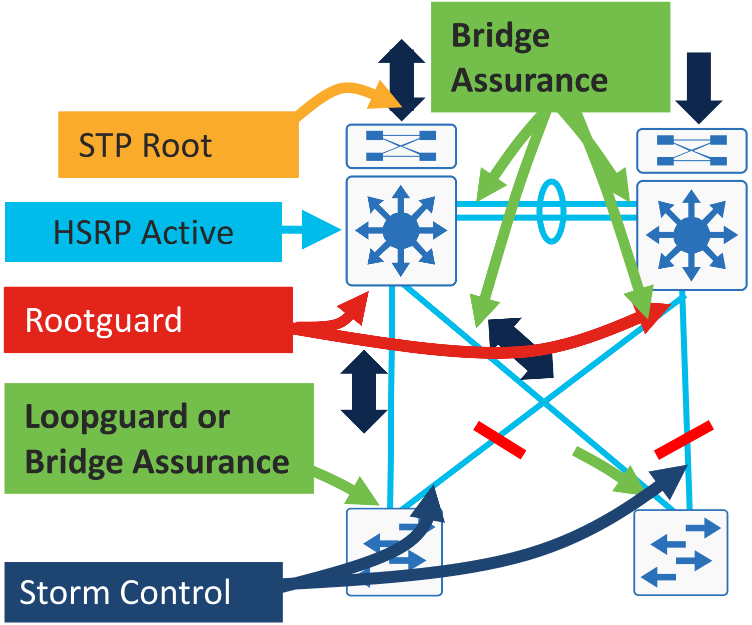
```
1w2d: %SPANTREE-2-BLOCK_BPDUGUARD: Received BPDU on port FastEthernet3/1 with BPDU Guard enabled. Disabling port.
1w2d: %PM-4-ERR_DISABLE: bpduguard error detected on Fa3/1, putting Fa3/1 in err-disable state
```

Оптимизация Layer 2 дизайна

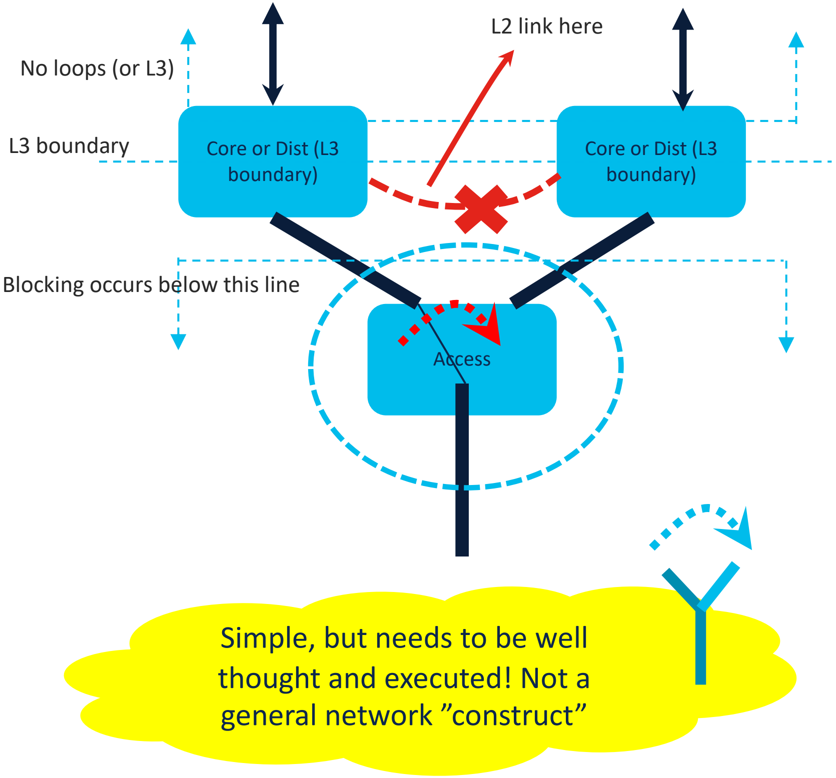
Лучшие практики STP для кампусов

Reference

- The root bridge should stay where you put it
  - Define the STP primary (and backup) root
  - Rootguard
  - Loopguard or bridge assurance
  - UDLD
- There is a reasonable limit to broadcast and multicast traffic volumes
- Configure storm control on backup links to aggressively rate limit broadcast and multicast



Технология FlexLink


No loops (or L3)
L3 boundary
L2 link here
Core or Dist (L3 boundary)
Core or Dist (L3 boundary)
Blocking occurs below this line
Access

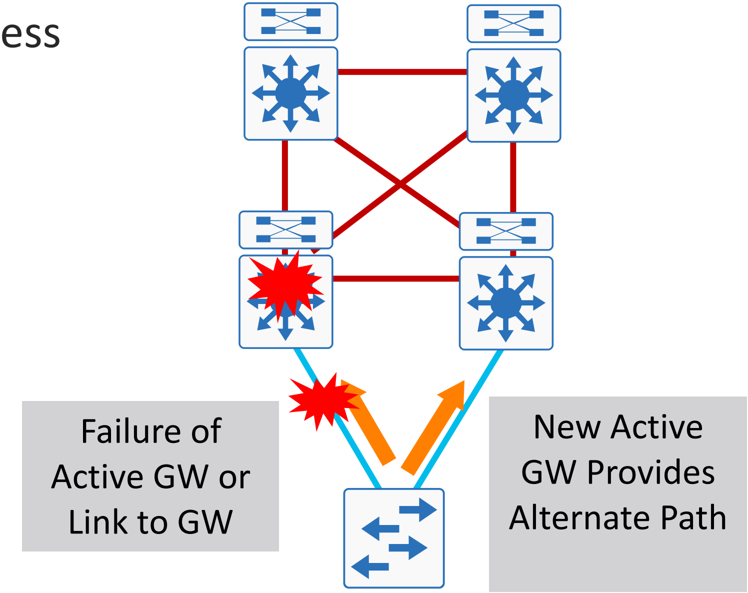
Simple, but needs to be well thought and executed! Not a general network ”construct”

- Very basic and simple construct - more like the old serial line backup interface feature.
  - **Detect** link down → force backup interface to **go fowarding**
- Relies on link down (there are several cases where there is a failure but the physical link does not go down or link down detection is too slow…)
- The topology must be such that the “blocking” always happens on the “access side”
- No box redundancy (failure at dist/core must force link facing access to go down (requires core/dist to support interface or other tracking mechanism
- No L2 between core/dist boxes (otherwise flow from the core would go back)
- No spanning tree towards the access – so loop avoidance has to be done via config/design - ”user error” can become fatal

Оптимизация Layer 2 дизайна

Протоколы резервирования шлюза (FHRP)

- HSRP, GLBP, and VRRP:
  provide a resilient default gateway / first hop address to end stations
- A group of routers act as a single logical router providing first hop router redundancy
- Protect against multiple failures
  - Distribution switch failure
  - Uplink failure
- Default recovery is ~10 Seconds



Failure of Active GW or Link to GW

New Active GW Provides Alternate Path

Резервирования шлюза по умолчанию

Subsecond timers improve convergence

HSRP Config

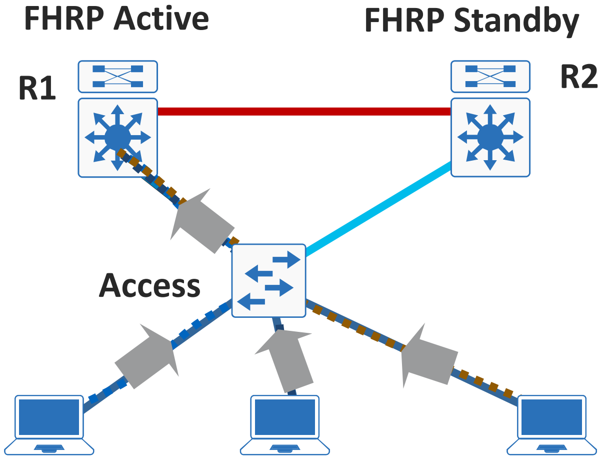
```
interface Vlan4
 ip address 10.120.4.2 255.255.255.0
 standby 1 ip 10.120.4.1
 standby 1 timers msec 250 msec 750
 standby 1 priority 150
 standby 1 preempt
 standby 1 preempt delay minimum 180
```

GLBP Config

```
interface Vlan4
 ip address 10.120.4.2 255.255.255.0
 glbp 1 ip 10.120.4.1
 glbp 1 timers msec 250 msec 750
 glbp 1 priority 150
 glbp 1 preempt
 glbp 1 preempt delay minimum 180
```

VRRP Config

```
interface Vlan4
 ip address 10.120.4.1 255.255.255.0
 vrrp 1 description Master VRRP
 vrrp 1 ip 10.120.4.1
 vrrp 1 timers advertise msec 250
 vrrp 1 preempt delay minimum 180
```



HSRP is widely used with Its rich feature set

GLBP facilitates uplink load balancing – not optimal for L2 looped topology

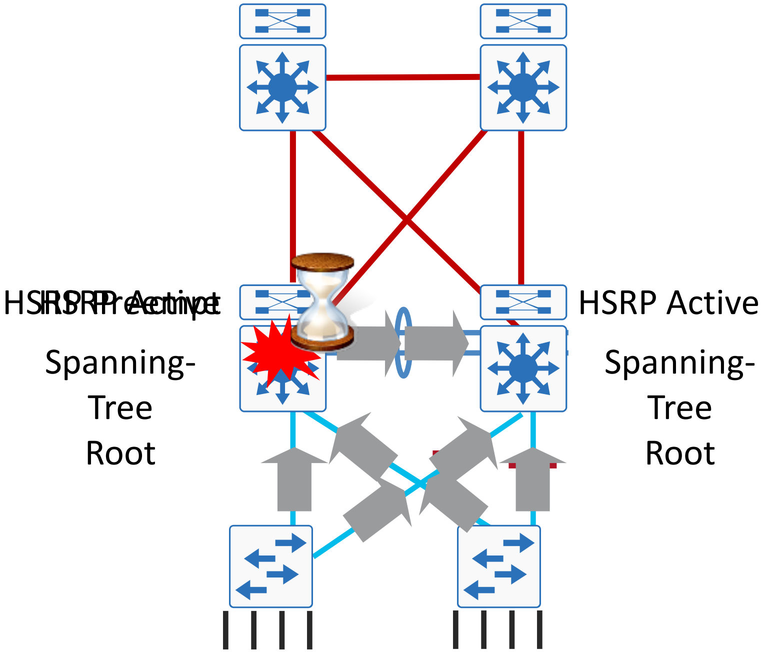
VRRP for multi-vendor interoperability

HSRP, **GLBP** and **VRRP** provide millisecond timers and excellent convergence performance

Critical for VoIP and video recovery in < 1 second

HSRP preemption—почему это желательно

- Spanning tree root and HSRP primary are aligned
- When spanning tree root is re-introduced,
  traffic takes a two-hop path to HSRP active
- **HSRP preemption** allows HSRP
  to follow the spanning tree topology



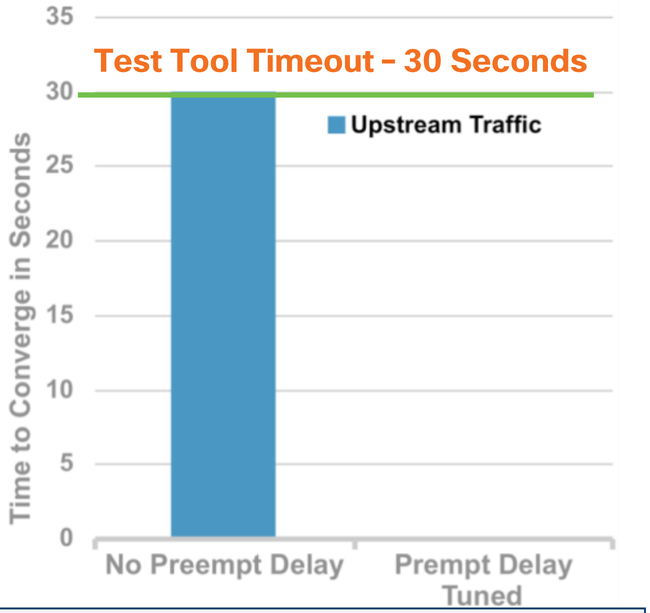
Without Preempt Delay, HSRP Can Go Active Before the Switch Is Completely Ready to Forward Traffic – L1 (Linecards), L2 (STP), L3 (IGP Convergence)

Рассмотрение FHRP дизайна

Preempt задержка должна быть дольше чем время загрузки коммутатора

- HSRP is not always aware of the status of the entire switch and network
- Ensure that you provide enough time for the diagnostics (full or partial), L1 (line cards), L2 L3 (IGP convergence)
- Tune delay and preempt delay conservatively orwarding data

```
interface Vlan402
. . .
 standby delay minimum 60 reload 600
 standby 1 ip 10.147.102.1
 standby 1 timers msec 250 msec 750
 standby 1 priority 110
 standby 1 preempt delay minimum 60 reload 600
 standby 1 authentication ese
 standby 1 name HSRP-Voice
 hold-queue 2048 in
```

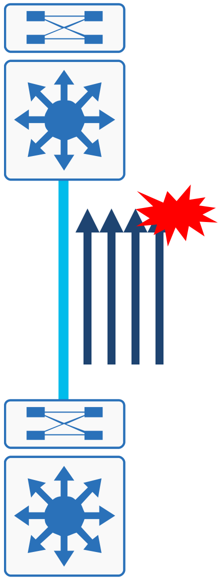


`standby delay`: Controls time interface needs to be up before HSRP starts.
`preempt delay`: Controls time to wait after HSRP establishes a neighbour relationship.
Configure both.

Рассмотрение Sub-second таймеров

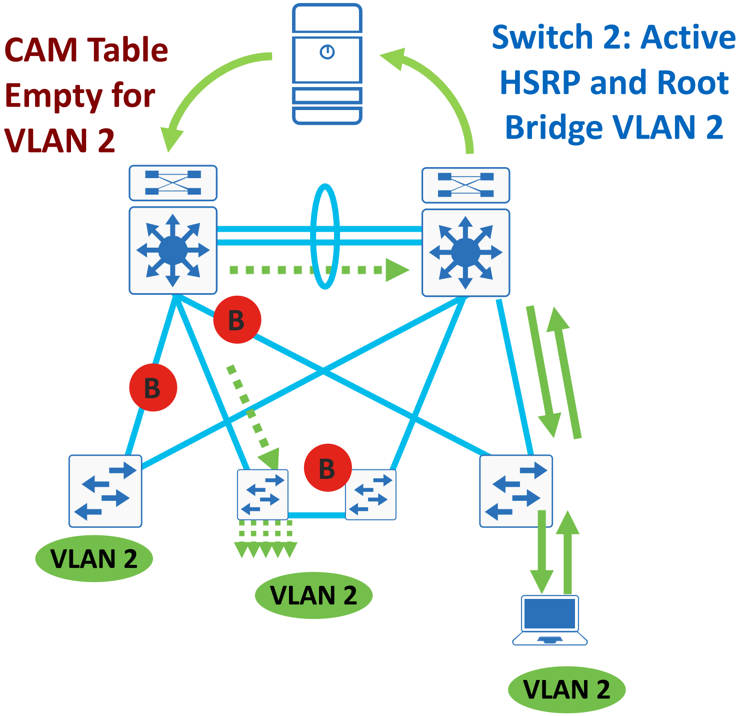
HSRP, GLBP, OSPF, PIM

- Evaluate your network before implementing any sub-second timers
- Certain events can impact the ability of the switch to process sub-second timers
  - Application of large ACL
  - OIR of line cards in Catalyst 6500/6800
- Control plane traffic volume also impacts ability to process
  - 250 / 750 msec GLBP & HSRP timers are only valid in designs with less than 150 VLAN instances (Catalyst 6x00 in the distribution)
  - Spanning Tree size



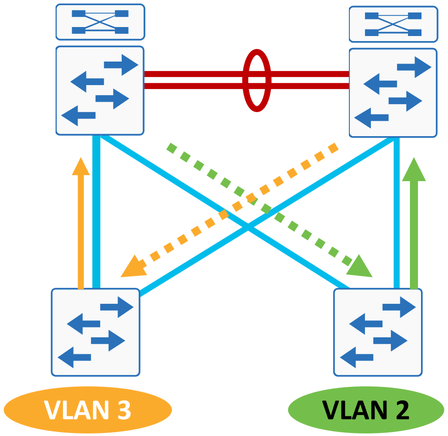
Рассмотрение FHRP дизайна — *asymmetric routing (unicast flooding)*

- Alternating HSRP Active between distribution switches can be used for upstream load balancing
- This can cause a problem with unicast flooding
- **ARP timer** defaults **to four hours** and **CAM timer** defaults to **five minutes**
- ARP entry is valid, but no matching L2 CAM table exists
- In many cases when the HSRP standby needs to forward a frame, it will have to unicast flood the frame since its CAM table is empty


CAM Table Empty for VLAN 2
Switch 2: Active HSRP and Root Bridge VLAN 2
B
B
B
VLAN 2
VLAN 2
VLAN 2

Рассмотрение FHRP дизайна —
asymmetric routing (unicast flooding) solutions

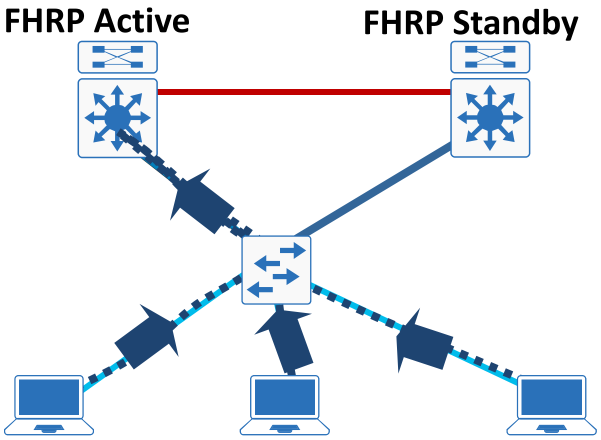
- Using ‘V’ based design with unique voice and data VLANs per access switch, this problem has no user impact
- Don’t deploy stacking switches (ie. daisy-chained switches) that depend on spanning tree for managing stack interconnects
- Tune ARP timer to 270 seconds and leave CAM timer to default, unless ARP > 10,000, change CAM timers
- Deploy MultiChassis EtherChannel with StackWise Virtual (SWV), Virtual Switching System (VSS), or Virtual Port Channel (vPC) in the distribution block



CAM timers traditionally default to 5 minutes to allow for MAC addresses (devices) to move in the network. It is safe to increase the CAM timers if the client devices will generate unicast or multicast traffic to refresh the CAM table.

Even with faster convergence from RPVST+ we still have to wait for FHRP convergence

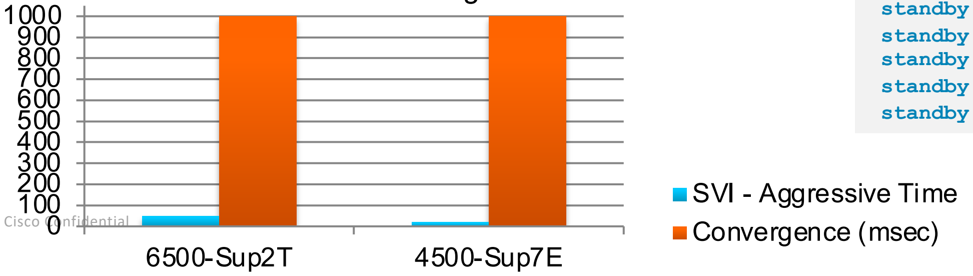
- FHRP protocol based forwarding topologies
  - Load balancing based on Per-Port or Per-VLAN
- Protocol-based fault detection and recovery –
  - Configure per-VLAN aggressive timers to protect user experience impact within <1 second boundary
- Limited network scale for system reliability
- Sub-second protocol timers must be avoided on SSO capable networks



HSRP Config

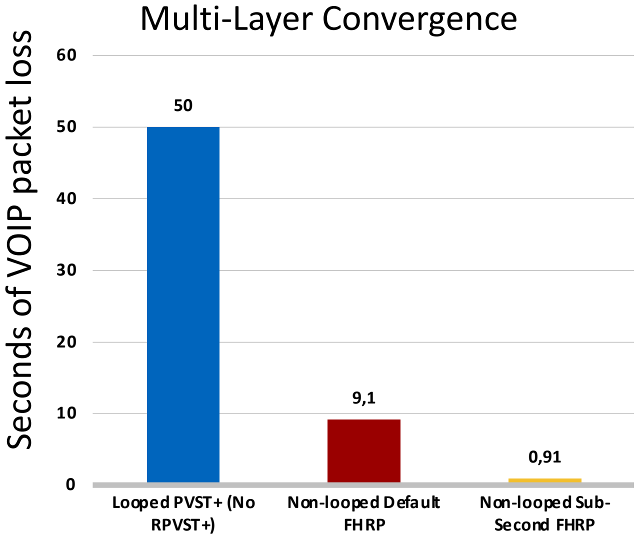
```
interface Vlan2
 ip address 10.120.2.2 255.255.255.0
 standby 1 ip 10.120.2.1
 standby 1 timers msec 250 msec 750
 standby 1 priority 150
 standby 1 preempt
 standby 1 preempt delay minimum 180
```

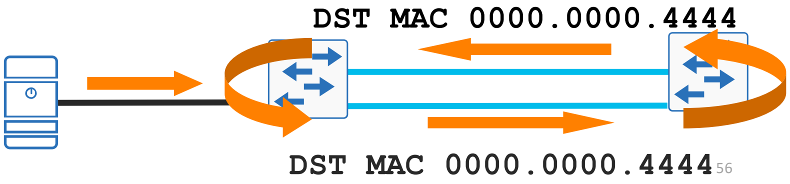
Multilayer Standalone
Network Scale And Convergence



Multilayer campus network design—It is a good solid design, but…

- Utilizes multiple control protocols
  - Spanning tree (802.1w), HSRP / GLBP, EIGRP, OSPF
- Convergence is dependent on multiple factors –
  - FHRP – 900msec to 9 seconds
  - Spanning tree – Up to 50 seconds
- Load balancing –
  - Asymmetric forwarding
  - HSRP / VRRP – per subnet
  - GLBP – per host
- Unicast flooding in looped design
- STP, if it breaks badly, has no inherent mechanism to stop the loop

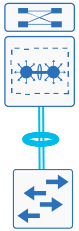




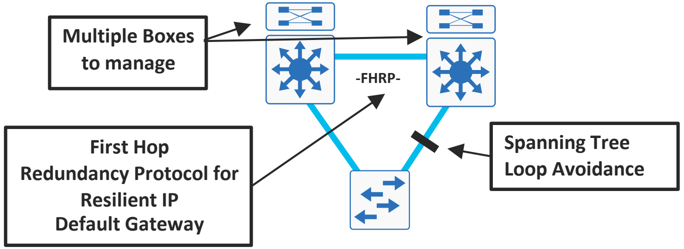
Упрощенный дизайн уровня распределения

Уровень распределения

- Traditional two box distribution layer has many points to manage
- Preferred distribution layer uses a “single box design”
  - Two switches acting as a single logical switch (StackWise Virtual or Virtual Switching System)
  - A multiple member switch stack acting as a single logical switch
- Simplified design benefits
  - Fewer boxes to manage
  - Simplified configuration
  - Logical hub-and-spoke topology



Traditional two box design



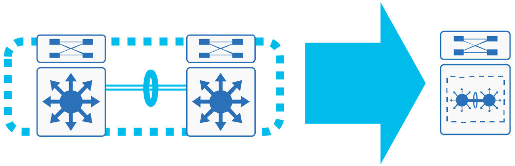
Multiple Boxes to manage

-FHRP-

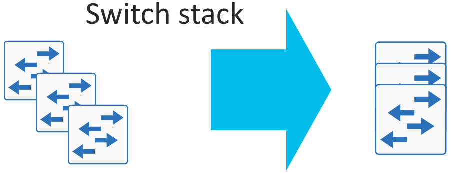
First Hop Redundancy Protocol for Resilient IP Default Gateway

Spanning Tree Loop Avoidance

SWV – StackWise Virtual



Switch stack

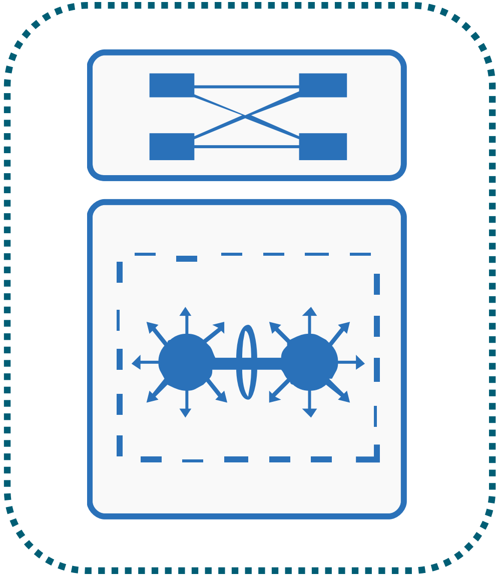


Унифицированная архитектура

StackWise Virtual (SWV)

Simplified Control-Plane

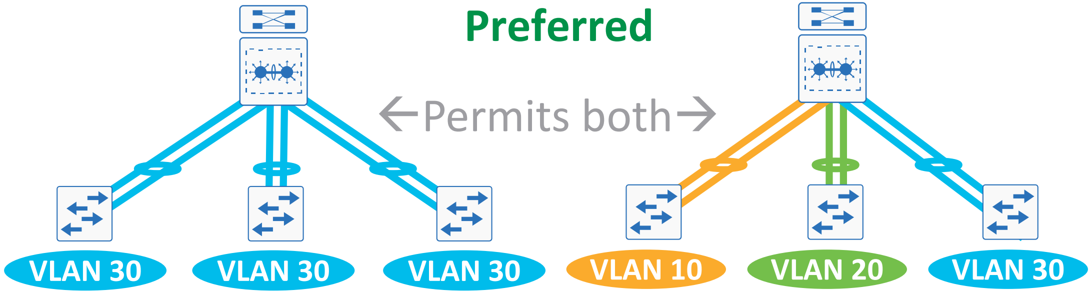
- Single control-plane to manage two physical systems
- Consistent IOS software feature parity as Standalone
- Centralized programming for distributed forwarding



Common Management

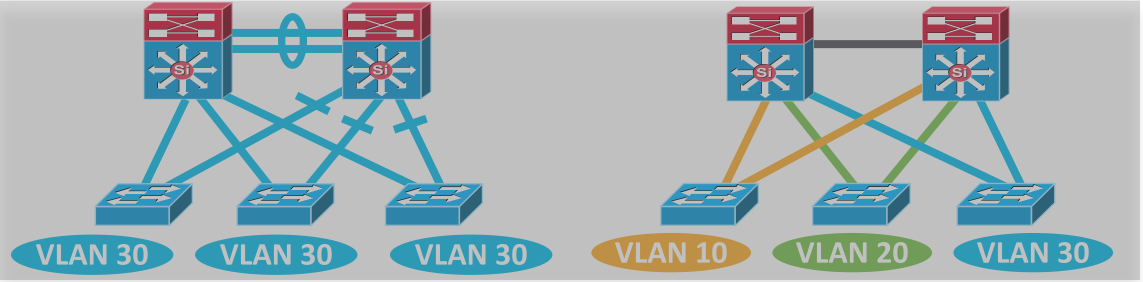
- Single virtual system for OOB/in-band management of two physical systems
- Common SNMP MIBs, traps with advanced MIBS
- Single troubleshooting point

Традиционный дизайн в сравнении с упрощенным
Уровень распределения



Traditional designs:

- Looped design with spanned VLANs
  - Relies on STP to block loops
  - Reduces available bandwidth
- Loop free design
  - Can increase bandwidth
  - Still relies on FHRP
  - Multiple distribution layer boxes to configure



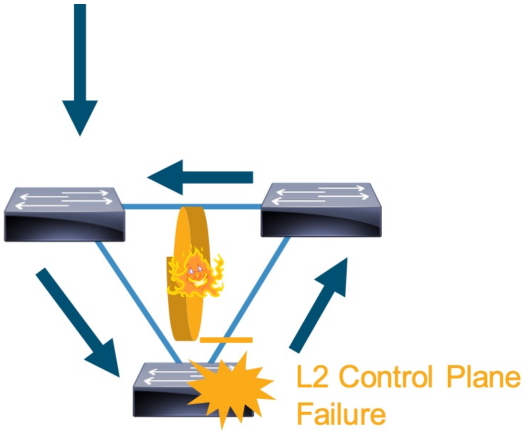
Preferred—simplified design:

- EtherChannel - resilient links, all links forwarding
- No FHRP - single default IP gateway
- Works with VLAN per closet or few VLANs spanned designs
- Logical hub-and-spoke topology
- Reduced dependence on spanning tree - keep RPVST+ for edge protection

Высокая доступность: *Маршрутизируемый уровень доступа (Routed Access)*

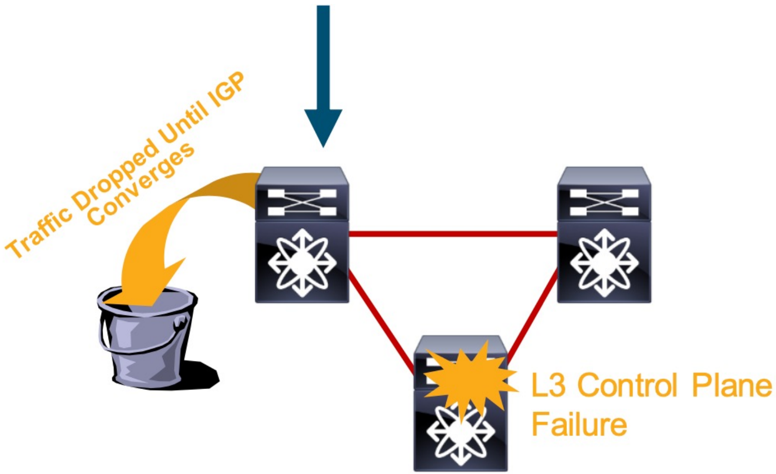
Почему L3 лучше чем L2?

- L2 Fails Open: Broadcast and Unknowns flooded



... as loop happens and network melts

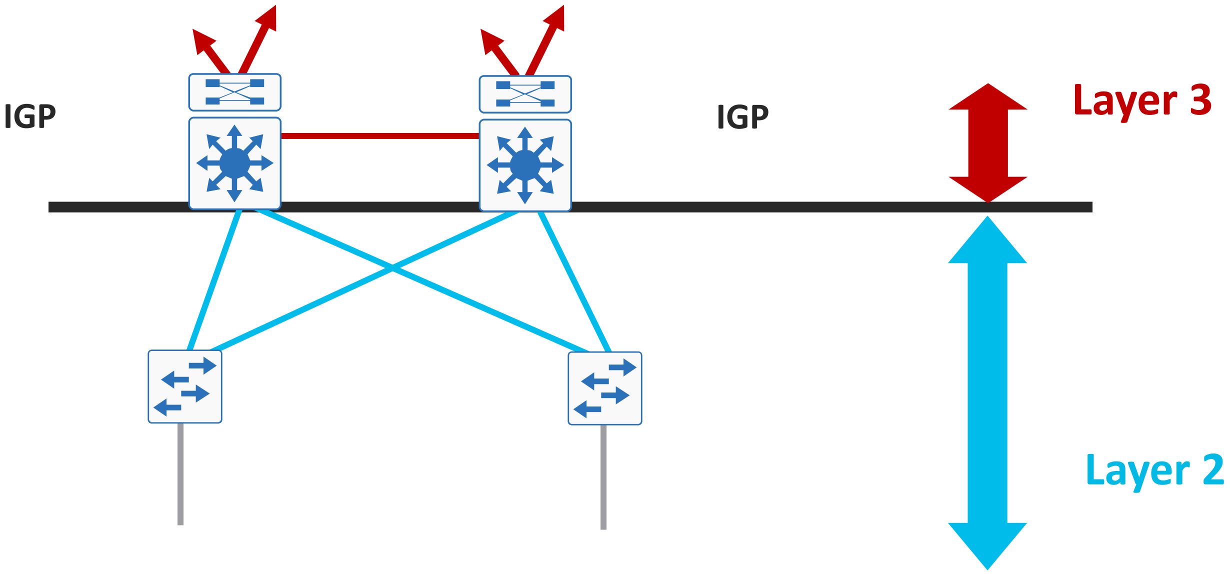
- L3 Fails Closed: As neighbor is lost



... Destination traffic blackholed

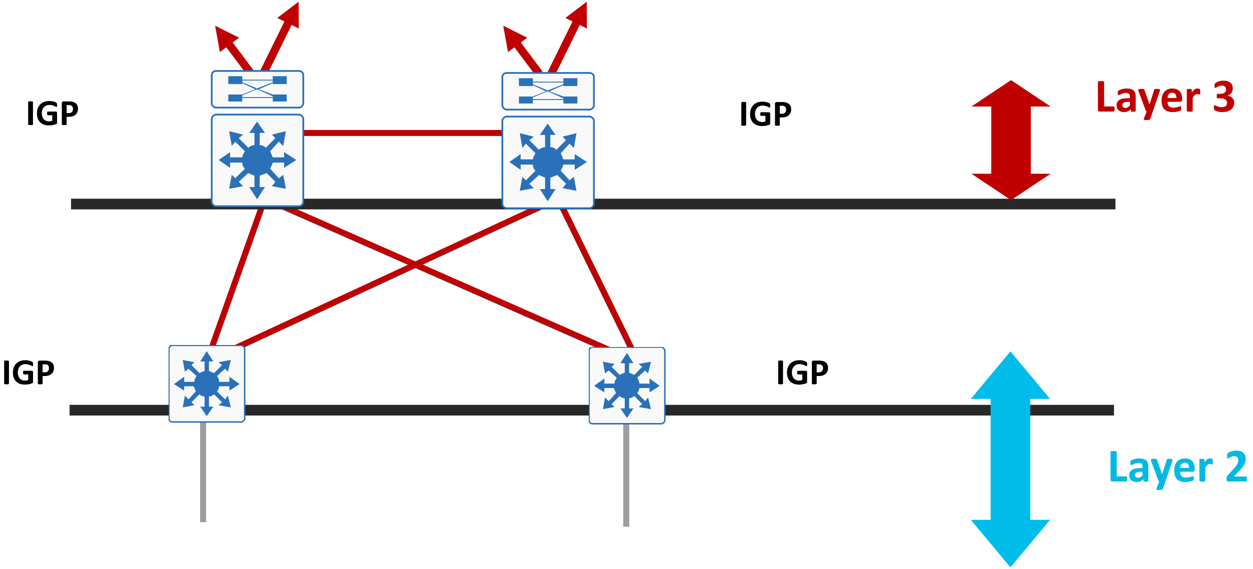
Трансформация многоуровневого кампуса

До: распределение Layer 3, доступ Layer 2



Simplification with routed access design

После: Layer 3 на уровнях распределения и доступа



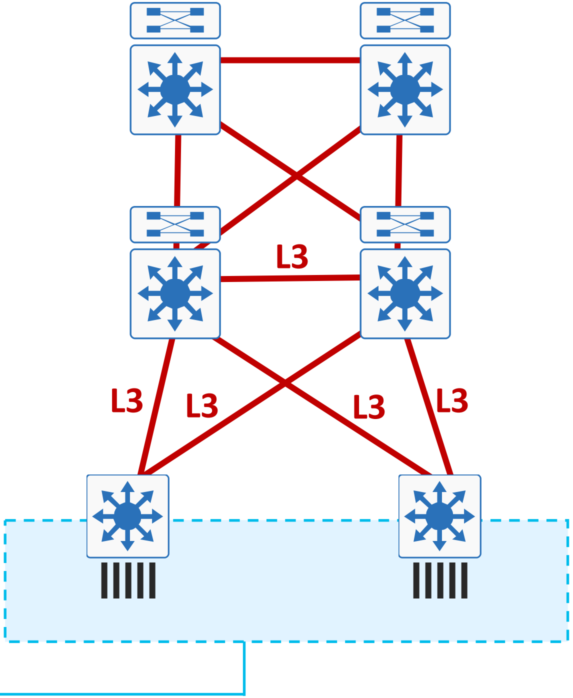
- Move the Layer 2 / 3 demarcation to the network edge
- Leverages Layer 2 only on the access ports, but builds a Layer 2 loop-free network
- **Design motivations** – Simplified control plane, ease of troubleshooting, highest availability

Преимущества Routed access

Упрощенный Control Plane

- Simplified Control Plane
  - **No STP** feature placement (root bridge, loopguard, …)
  - **No default gateway** redundancy setup/tuning (HSRP, VRRP, GLBP …)
  - **No matching of STP/HSRP priority**
  - **No asymmetric flooding**
  - **No** L2/L3 **multicast** topology **inconsistencies**
  - **No Trunking** Configuration Required

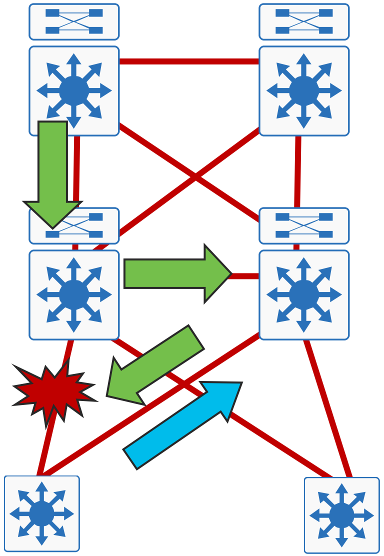
- L2 Port Edge features still apply:
  - Spanning Tree Portfast
  - Spanning Tree BPDU Guard
  - Port Security, DHCP Snooping, DAI, IPSG
  - Storm Control



Преимущества Routed access

Упрощенное восстановление сети

- Routed access network recovery is dependent on L3 re-route
- **Upstream** traffic restoration: ECMP re-route
  - Detect link failure
  - Process SW RIB update
  - Update HW FIB
- **Downstream** traffic restoration: routing protocol re-route
  - Detect link failure
  - Determine new route
  - Process SW RIB update
  - Update HW FIB



Upstream Recovery: ECMP

Downstream Recovery: Routing Protocol

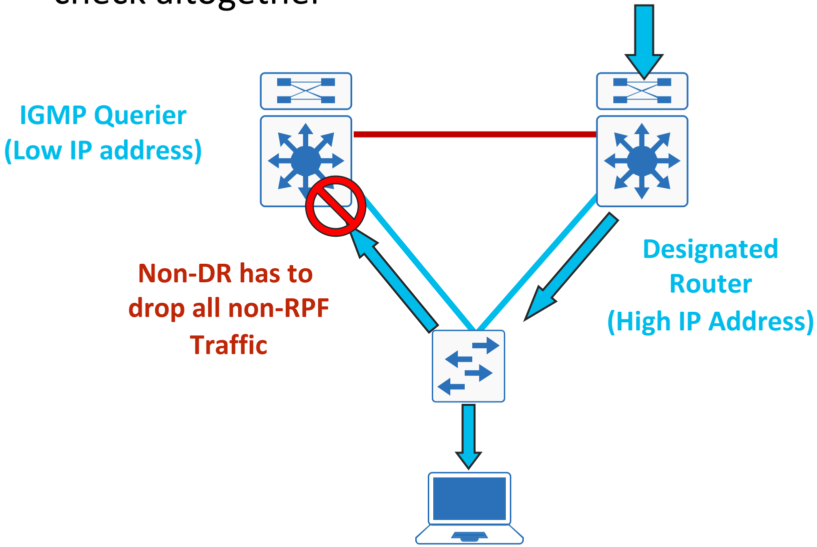
Compare to…

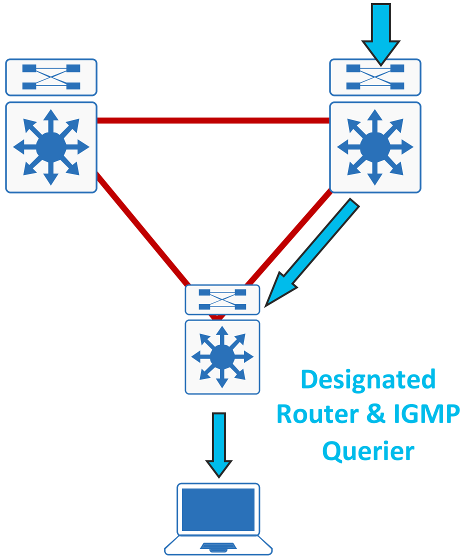
- RPVST+ convergence times dependent on FHRP tuning
- Proper FHRP design and tuning can achieve sub-second times
- EIGRP converges <200 msec
- OSPF converges <200 msec with LSA and SPF tuning

Преимущества routed access

Один маршрутизатор на подсеть: более простой multicast

- Layer 2 access has two multicast routers per access subnet, RPF checks and split roles between routers
- Routed access has a single multicast router which simplifies multicast topology and avoids RPF check altogether

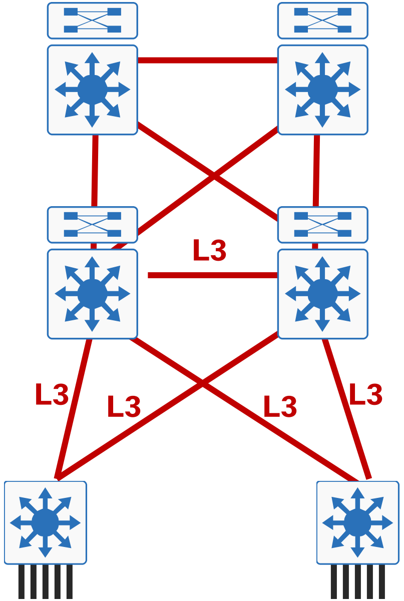




Преимущества routed access

Легче поиск и устранение неисправностей

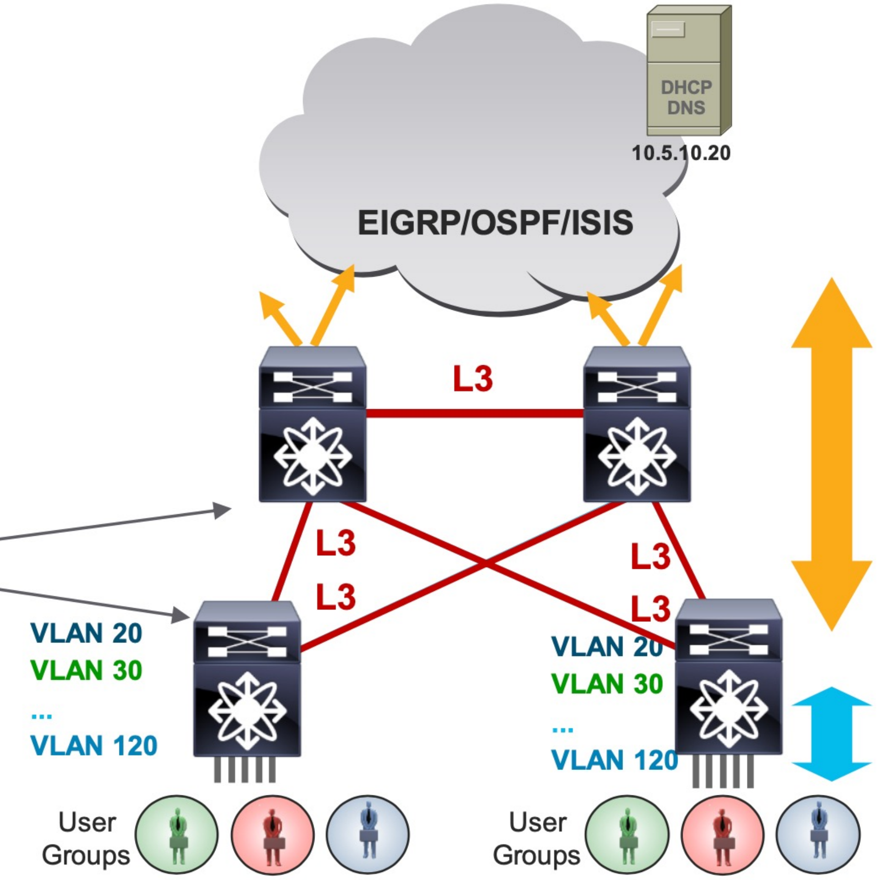
- Routing troubleshooting tools
  - **Consistent troubleshooting: access, dist, core**
  - show ip route / show ip cef
  - Traceroute
  - Ping and extended pings
  - Extensive protocol debugs
  - IP SLA from the Access Layer
- Failure differences
  - Routed topologies fail closed—i.e. neighbor loss
  - Layer 2 topologies fail open—i.e. broadcast and unknowns flooded



```
switch#sh ip cef 192.168.0.0
192.168.0.0/24
  nexthop 192.168.1.6 TenGigabitEthernet9/4
```

Дизайн Routed Access

```
interface GigabitEthernet1/1
 description Distribution Downlink
 ip address 10.120.0.196 255.255.255.254
```



- As the routing is moved to the access layer, trunking is no longer required
- /31 addressing can be used on p2p links to optimize ip space utilization

Выбор платформы для Routed Access

- **Catalyst Requirements**
  - Cisco Catalyst 2960 XR (IP Lite)
  - Cisco Catalyst 3K/4K/6K
  - Cisco Catalyst 9000
- **IP Base/Network Essential minimum license**
  - EIGRP-Stub – Edge Router
  - PIM Stub – Edge Router
  - OSPF for Routed Access
  - 1000 OSPF Routes
- **IP Services/Network Advantage license**
  - IS-IS

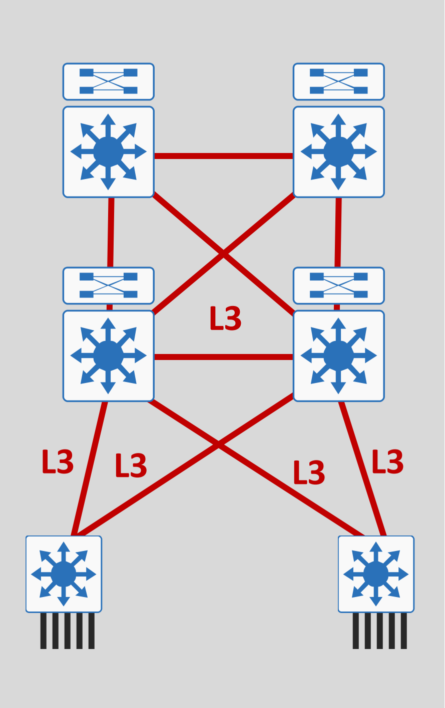
Почему бы не использовать routed access везде?

Ограничения Routed Access

- VLANs don’t span across multiple wiring closet switches/switch stacks

  Does this impact your requirements?

- IP addressing changes: more DHCP scopes and subnets of smaller sizes increase management and operational complexity
- Deployed access platforms must be able to support routing features
- Segmentation / Virtualization

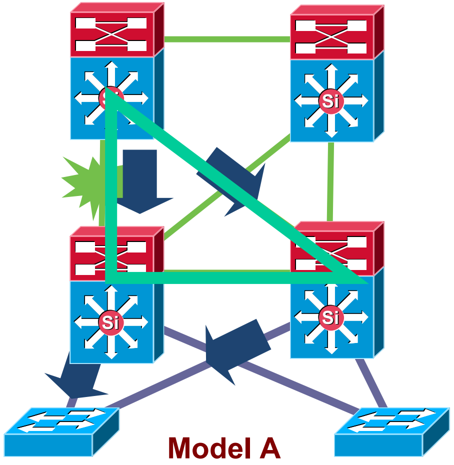


Высокая доступность: *Протоколы динамической маршрутизации*

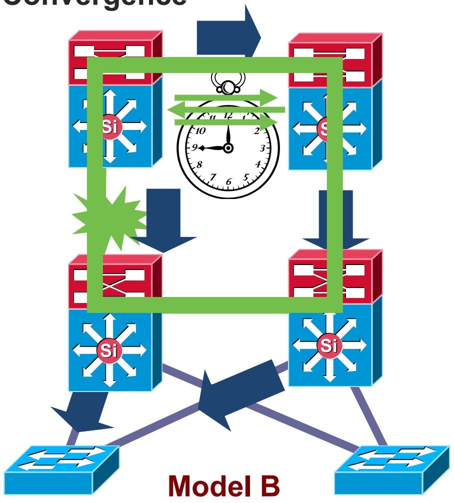
Стром треугольники, а не квадраты

Детерминированный и Недетерминированный подходы

Triangles: Link/Box Failure Does not Require Routing Protocol Convergence



Squares: Link/Box Failure Requires Routing Protocol Convergence



- Layer 3 redundant equal cost links support fast convergence
- Hardware based—fast recovery to remaining path
- Convergence is extremely fast (dual equal-cost paths: no need for OSPF or EIGRP to recalculate a new path)

Распределение нагрузки Cisco Express Forwarding (CEF)

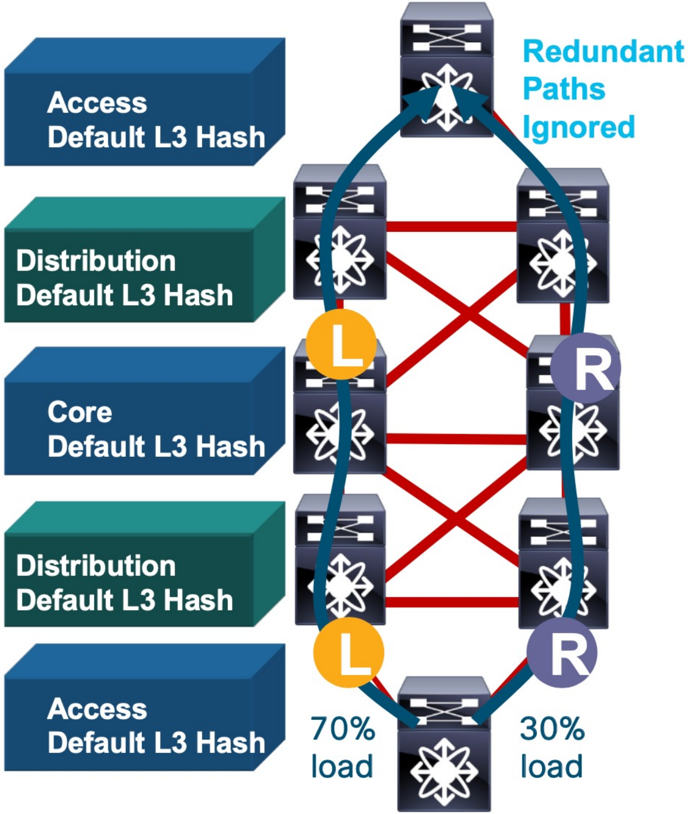
Недогруженные резервные пути Layer 3

- With defaults, CEF could select the same left/left or right/right paths and ignore some redundant paths
- Two solutions to achieve better redundant path utilization:
  - CEF Hash Tuning

```
ip cef load-sharing algorithm {original|include-ports}
```

  - CEF Universal ID (default on newer platforms; includes L4 info and random 32bit ID at each router)

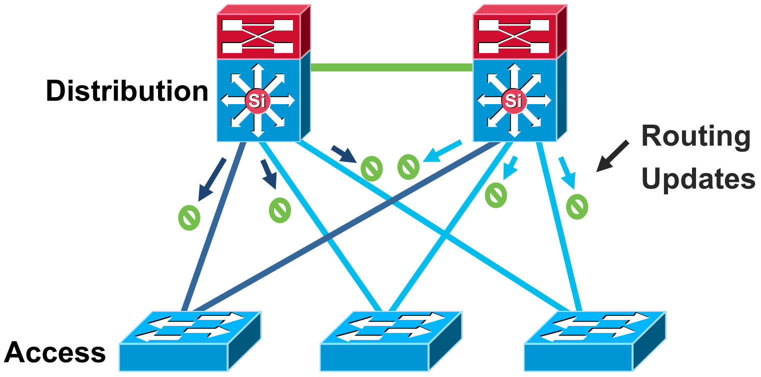
```
ip cef load-sharing algorithm universal
```



Не забываем про пассивные интерфейсы для IGP

Ограничиваем пиринг IGP на уровнях доступа и распределения

- Limit unnecessary peering using passive interface:
  - Four VLANs per wiring closet
  - 12 adjacencies total
  - Memory and CPU requirements increase with no real benefit
  - Creates overhead for IGP



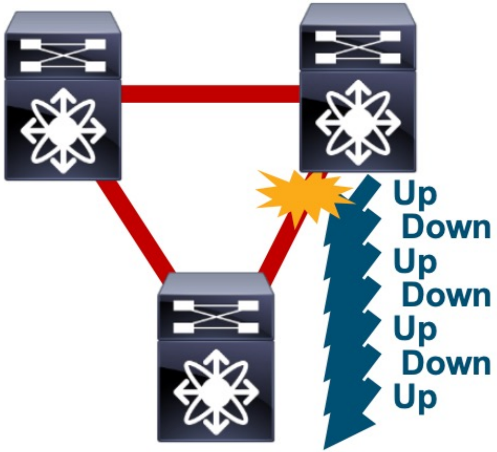
```
OSPF Example:

Router(config)#routerospf 1
Router(config-router)#passive-interfaceVlan 99

Router(config)#routerospf 1
Router(config-router)#passive-interface default
Router(config-router)#no passive-interface Vlan 99
```

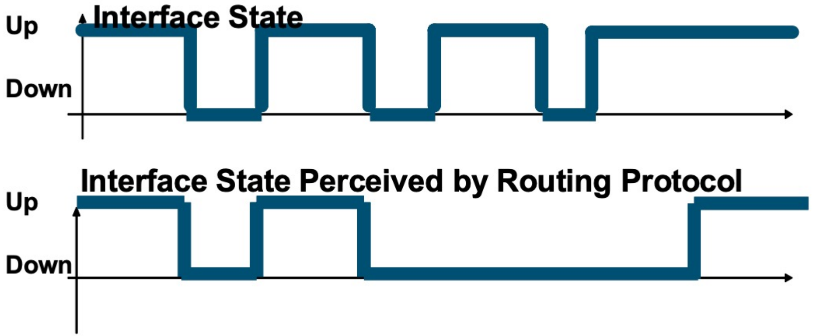
```
EIGRP Example:

Router(config)#routereigrp 1
Router(config-router)#passive-interfaceVlan 99

Router(config)#routereigrp 1
Router(config-router)#passive-interface default
Router(config-router)#no passive-interface Vlan 99
```

Демпфирование событий IP для снижения негативного воздействия на маршрутизацию

- Prevents routing protocol churn caused by constant interface state changes
- Dampening is applied on a system: nothing is exchanged between routing protocols
- Supports all IP routing protocols
  - Static routing, RIP, EIGRP, OSPF, IS-IS, BGP
  - Also supports HSRP and CLNS routing
  - Applies on physical interfaces and can't be applied on sub-interfaces individually

```
interface GigabitEthernet1/1
 description Uplink to Distribution 1
 dampening
 ip address 10.120.0.205 255.255.255.254
```





Высокая доступность:
EIGRP

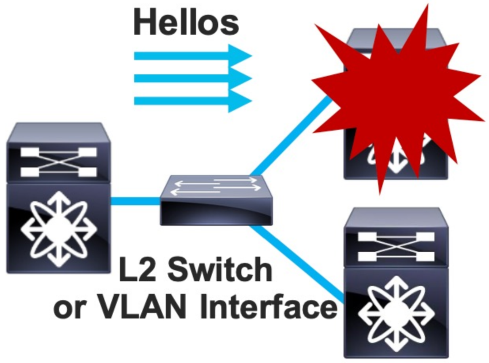
Обеспечение стабильной и быстрой сходимости EIGRP в сетях кампусов

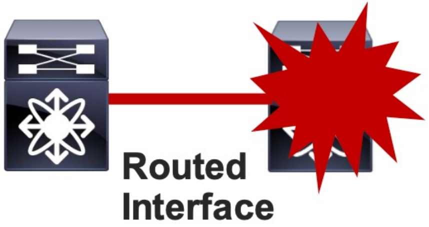
The key aspects to consider are:

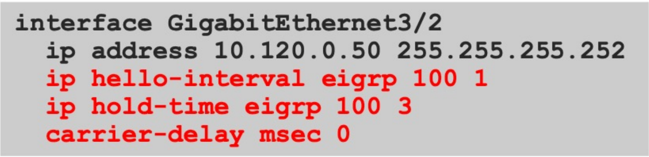
- Consider Hello and Hold Timer tuning
- Using EIGRP Stub at the access layer
- Route Summarization at the distribution layer

Обнаружение событий в EIGRP

- EIGRP neighbour relationship forms when link and routing adjacency are established
- Tune carrier delay to immediately notify routing process
- Use routed interfaces, not SVIs
- Decrease EIGRP timers
    - Hello = 1s (default is 5s for LAN)
    - Hold-down = 3







```
interface GigabitEthernet3/2
  ip address 10.120.0.50 255.255.255.252
  ip hello-interval eigrp 100 1
  ip hold-time eigrp 100 3
  carrier-delay msec 0
```

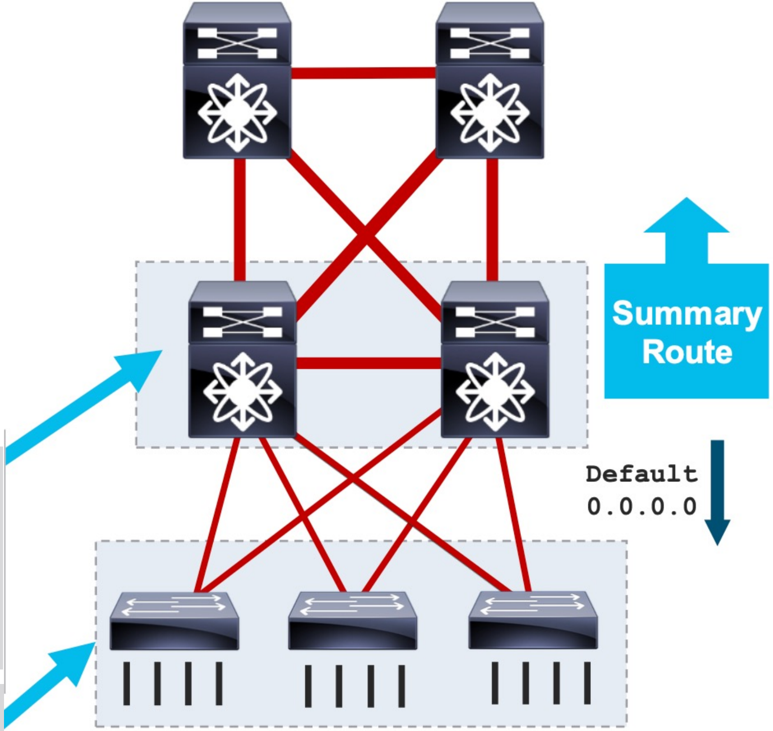
Правила дизайна EIGRP для HA кампуса

Limit Query Range to Maximize Performance

- EIGRP convergence is dependent on query response times
- Minimize the number of queries to speed up convergence
- Summarize distribution block routes to limit how far queries propagate across the campus
  - Upstream queries are returned immediately with infinite cost
- Configure access switches as EIGRP stub routers
  - No downstream queries are ever sent

```
interface TenGigabitEthernet 4/1
 ip summary-address eigrp 100 10.120.0.0 255.255.0.0 5
router eigrp 100
 network 10.0.0.0
 distribute-list Default out <mod/port>
 ip access-list standard Default permit 0.0.0.0
```

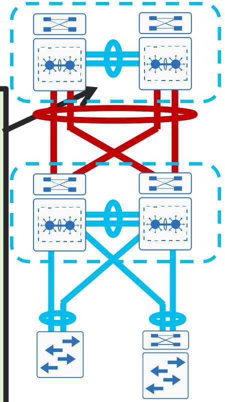
```
router eigrp 100
 network 10.0.0.0
 eigrp stub connected
```



Не забываем о защите IP routing - EIGRP

- Enable EIGRP for address space in use for core
  – just as was done in the distribution
- However…
  - No passive interfaces in core
    – route to everything from the co
- Remember to…
  - Enable authentication of neighbo
    routing protocol communication
- Enable NSF

```
key chain EIGRP-KEY
 key 1
  key-string [key]
router eigrp LAN
address-family ipv4 unicast autonomous-system 100
    network [network] [inverse mask]
    eigrp router-id [ip address of loopback 0]
    nsf
  exit-address-family
  af-interface default
    authentication mode md5
    authentication key-chain EIGRP-KEY
   exit-af-interface
```



Высокая доступность: *OSPF*

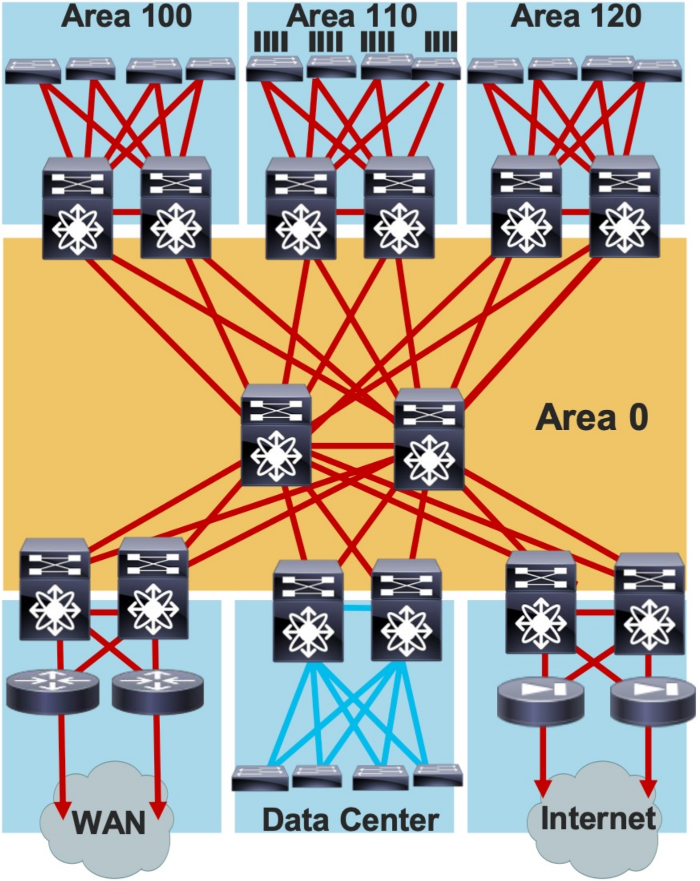
Обеспечение стабильной и быстрой сходимости OSPF в сетях кампусов

Key Objectives of the OSPF Campus Design

- Map area boundaries to the hierarchical design
- Enforce hierarchical traffic patterns
- Minimize convergence times
- Maximize stability of the network

Правила дизайна OSPF для HA кампуса

- Area design also based on address summarization
- Area boundaries should define flooding/fault domains
  - All routers within an area have same topology view of the network
  - Limit Area size to contain query range and SPF calculation
- Area 0 for core infrastructure- do not extend to the access routers



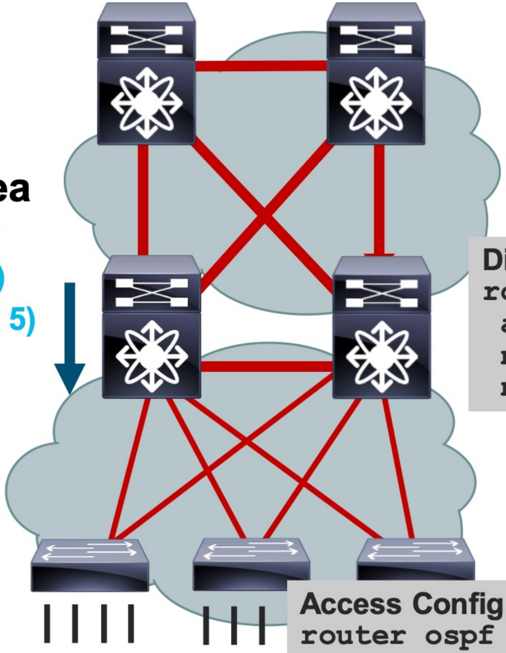
Обычная зона OSPF

ABRs Forward All LSAs from Backbone

External Routes/LSA Present in Area 120

ABR Forwards the Following into an Area

- **Summary LSAs (Type 3)**
- **ASBR Summary (Type 4)**
- **Specific Externals (Type 5)**



Distribution Config

```
router ospf 100
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
```

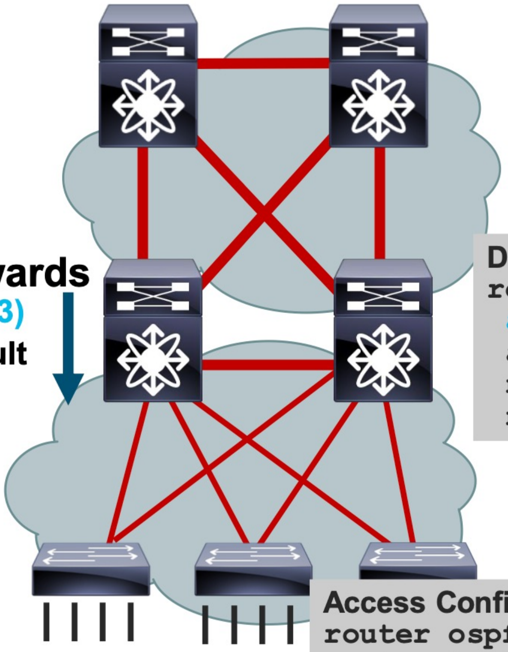
Access Config:

```
router ospf 100
 network 10.120.0.0 0.0.255.255 area 120
```

Stub зона OSPF

Consolidates Specific External Links—Default 0.0.0.0

Eliminates External Routes/LSA Present in Area (Type 5)



Backbone
Area 0

Stub Area ABR Forwards

Summary LSAs (Type 3)

Summary 0.0.0.0 Default

Area 120

Distribution Config

```
router ospf 100
 area 120 stub
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
```

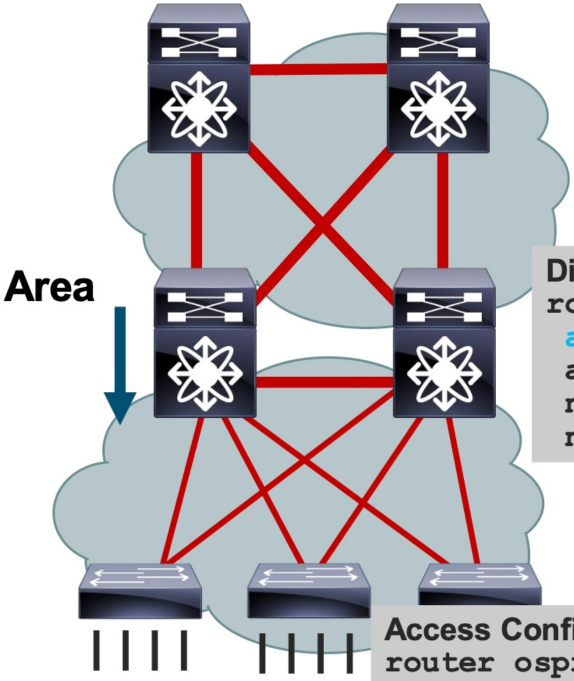
Access Config:

```
router ospf 100
 network 10.120.0.0 0.0.255.255 area 120
```

Использование OSPF Totally Stubby зон

Рекомендуемый для Routed Access на уровне доступа

Minimize the Number of LSAs and the Need for Any External Area SPF Calculations



Backbone
Area 0

A Totally Stubby Area ABR Forwards

Summary Default

Area 120

Distribution Config

```
router ospf 100
 area 120 stub no-summary
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
```

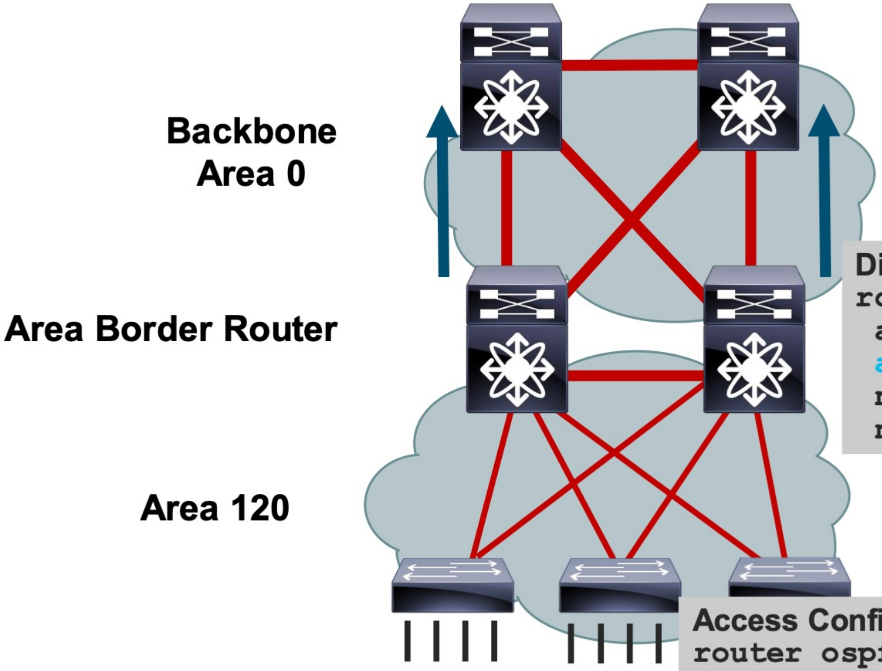
Access Config:

```
router ospf 100
 network 10.120.0.0 0.0.255.255 area 120
```

Суммаризация на уровне распределения к ядру

Уменьшение SPF и LSA нагрузки в Area 0

Minimize the Number of LSAs and the Need for Any SPF Recalculations at the Core



ABRs Forward

Summary 10.120.0.0/16

Distribution Config

```
router ospf 100
 area 120 stub no-summary
 area 120 range 10.120.0.0 255.255.0.0 cost 10
 network 10.120.0.0 0.0.255.255 area 120
 network 10.122.0.0 0.0.255.255 area 0
```

Access Config:

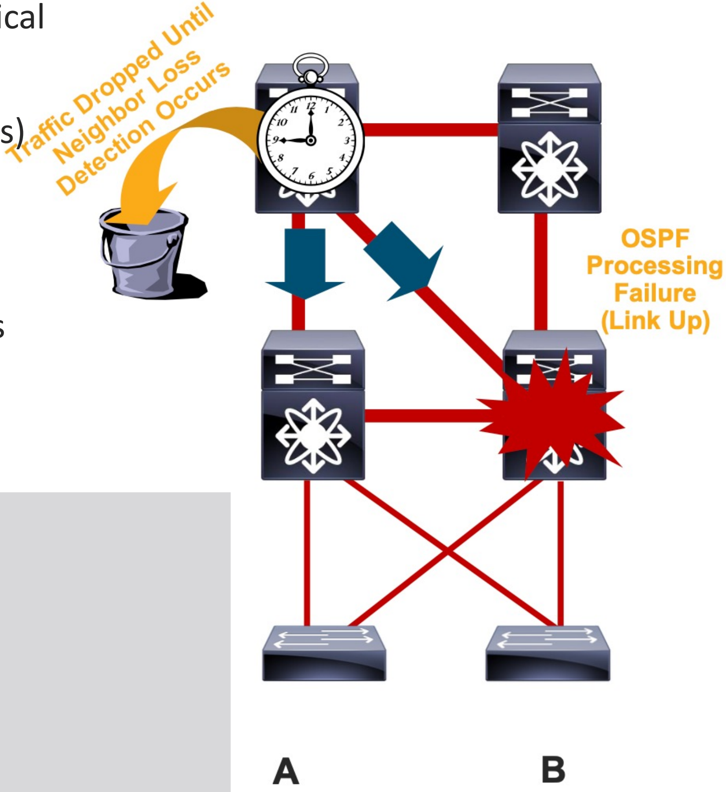
```
router ospf 100
 network 10.120.0.0 0.0.255.255 area 120
```

Subsecond Hellos

Neighbor Loss Detection—Physical Link Up

- OSPF hello/dead timers detect neighbor loss in the absence of physical link loss
- Useful where an L2 device separates L3 devices (Layer 2 core designs)
- Fast timers quickly detect neighbor failure
  - Not recommended with NSF/SSO
- Interface dampening is recommended with sub- second hello timers
- OSPF point-to-point network type to avoid designated router (DR) negotiation.

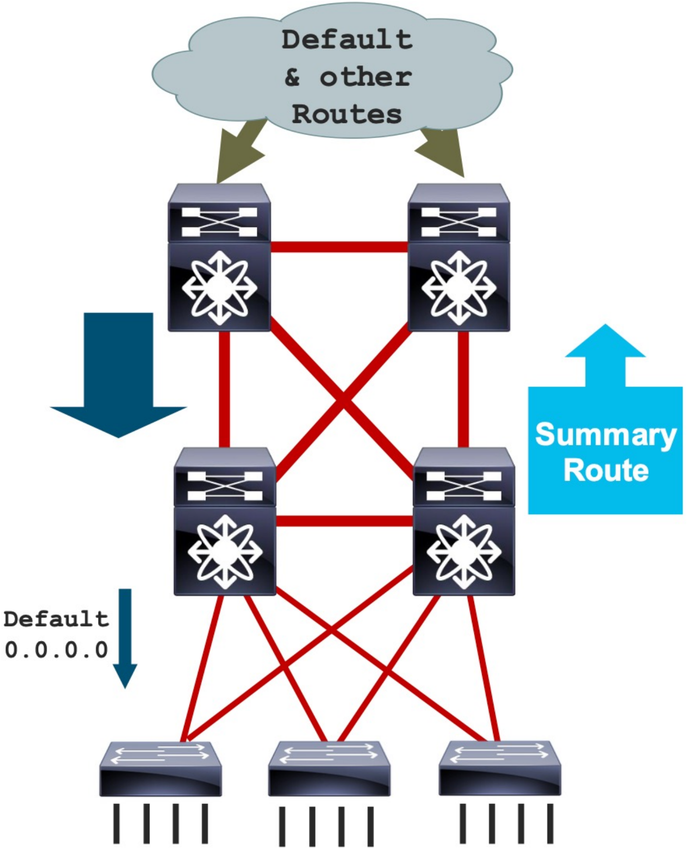
```
Access Config:
interface GigabitEthernet1/1
 dampening
 ip ospf dead-interval minimal hello- multiplier 4
 ip ospf network point-to-point

router ospf 100
 timers throttle spf 10 100 5000
 timers throttle lsa all  10 100 5000
 timers lsa arrival 80
```


Traffic Dropped Until Neighbor Loss Detection Occurs
OSPF Processing Failure (Link Up)
A
B

Дизайн OSPF Routed Access для кампусов

Подводя итог

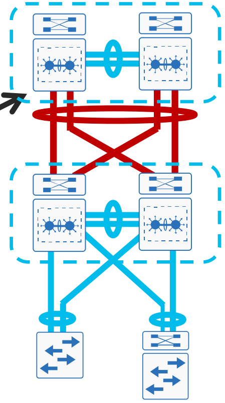
- Fast Convergence:
  - Set hello-interval = 250 milliseconds and Dead-time =
  - 1 seconds to detect soft neighbor failures *
  - Tune LSA/SPF timers
  - Set carrier-delay = 0, interface debounce "disable"
- Propagate the event:
  - All access layer switches—stub or totally stubby to limit queries from the distribution layer
  - Summarize the routes from the distribution to the core to limit queries across the campus
- Process the event:
  - Summarize and filter routes to minimize calculating new successors for the RIB and FIB



Не забываем о защите IP routing - OSPF

- Enable OSPF for address space in use for core
  – just as was done in the distribution
  - Core is OSPF Area 0
- However…
  - No passive interfaces in core
    – route to everything from the core
- Remember to…
  - Enable authentication of neighbor routing protocol communication
  - Enable NSF

```
interface [interface]
 ip ospf message-digest-key [key id] md5 [key]
router ospf 100
 router-id [ip address of loopback 0]
 nsf
 area 0 authentication message-digest
 network [network] [inverse mask] area 0
```

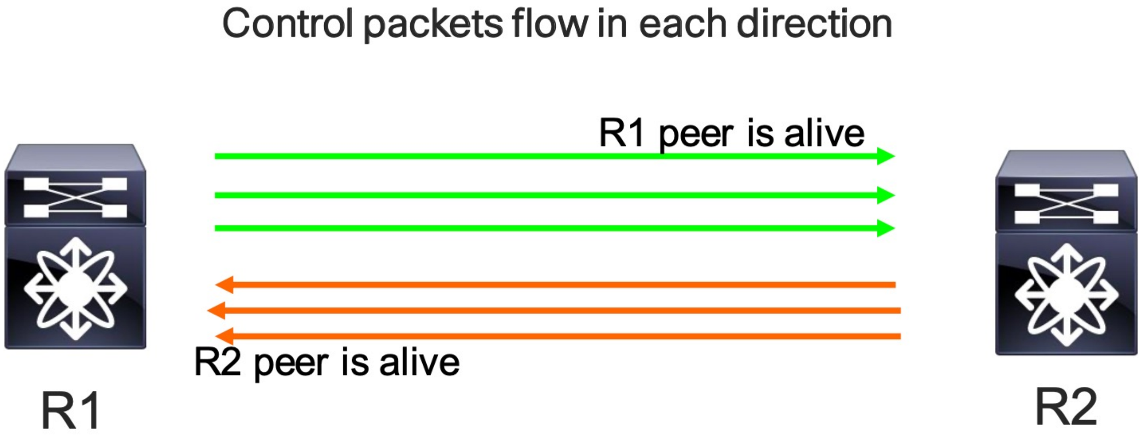


Высокая доступность: *Bi-directional Forwarding Detection (BFD)*

BFD – Что это такое?

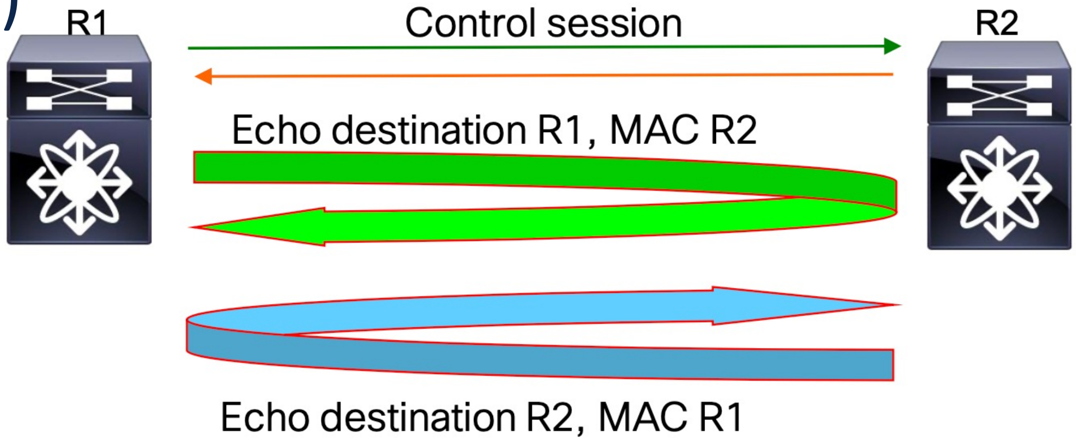
- Simple L2 hello protocol (RFC 5880, 5881)
- Control packet UDP-based for extremely fast L3 next-hop failure detection
- Independent of L3/routing protocols but invoked by interested protocols
- Session establishment between two peers via 3-way handshake
- Control packets sent with destination UDP 3784 and peer’s MAC
- Supports OSPF, EIGRP,BGP, IS-IS, HSRP, VRRP, static routes, tunnels etc.
- Nearly ubiquitous platform support

BFD Mode (no Echo)



- Routers periodically send BFD Control packets to each other
- If no packet is received for the peer during the duration of the negotiated
- detect time the session is declared to be down
- One-way nature limits testing of roundtrip forwarding path

BFD Mode (Echo)



- Echo packets sent with destination IP address as self, while destination MAC is peer
- Echo packets loop through the remote system
- Better test of bi-directional forwarding path due to loop
- Less CPU interrupt
- Not supported by some protocols (e.g. IS-IS)

BFD: Программный или аппаратный?

```
Switch#show bfd neighbor detail

IPv6 Sessions
NeighAddr                              LD/RD          RH/RS     State    Int
FE80::D68C:B5FF:FEE8:9E7F 257/257      Up         Up  Vl200

Session state is UP and not using echo function.

Session Host: Software

OurAddr: FE80::32F7:DFF:FE4E:217F

Rx Count: 111, Rx Interval (ms) min/max/avg: 20/108/90 last: 4 ms ago Tx
Count: 110, Tx Interval (ms) min/max/avg: 1/104/90 last: 76 ms ago Elapsed
time watermarks: 0 0 (last: 0)
Registered protocols: ISIS CEF
```

Настройка BFD на интерфейсе

```
interface ...
 bfd interval 100 min_rx 100 multiplier 3
 [no] bfd echo
```

- Enables BFD feature on the interface
- *interval msec*: Transmission interval to peer
- *min_rx msec*: Expected receive interval from peer
- Enables/disables echo mode
- *multiplier interval*: Missed packets before peer declared down

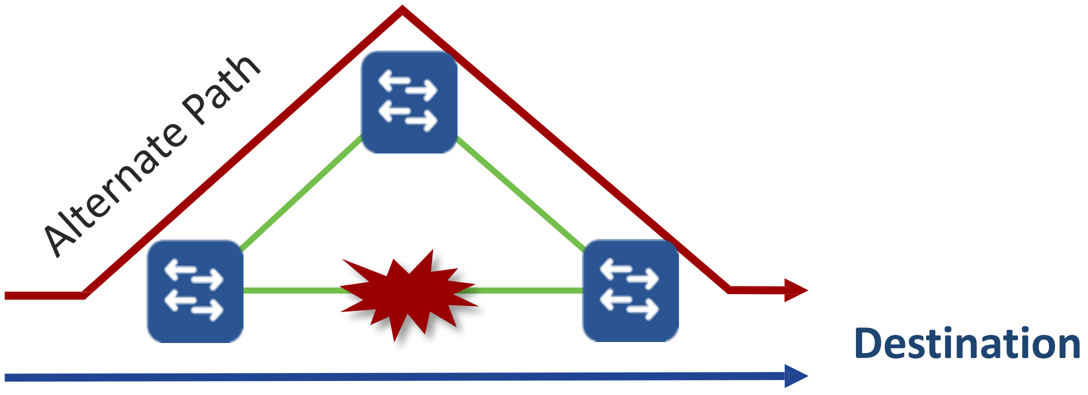
Включение BFD для протоколов динамической маршрутизации

```
router ...
[no] bfd all-interfaces
```

- Enables the corresponding routing protocols to rely on BFD for peer-failure detection
- BFD will be enabled on corresponding interface if routing protocol and bfd interface configuration is present
- “no” form of the above command will detach the association of BFD from that particular protocol

Высокая доступность:
Fast Reroute (FRR)

Benefits of IP Fast Reroute

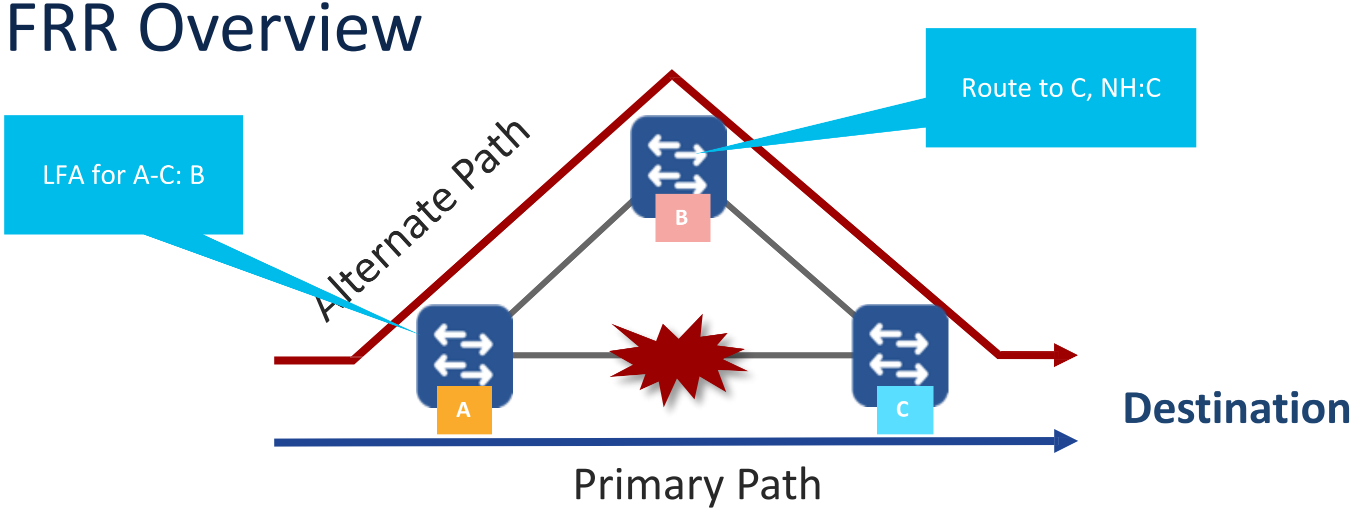


- Provides **50 ms** restoration of traffic flow in case of IP network failure
- Requires minimal configuration without the need of MPLS
- Protects against
  - Link Failures
  - Node Failures

How is it done?

- Have one pre-calculated backup next-hop that is loop free
- Have one backup next-hop for each primary next-hop in cef table
- Link State Routing Protocol
  - Uses the ability of link state routing protocol to understand the full topology
  - Performs SPF with neighboring node as root node
  - Neighbor SPF is run only after primary SPF is completed

IP FRR Overview



- Switch A computes the primary path to destination over Link A-C
- Switch A also computes the alternate path with 2 main properties
  - It should not use Link A-C
  - Next hop device will deliver traffic to destination without returning traffic to A
- This path is designated as Loop Free Alternate (LFA)

Platform Support and Limitations

- IP FRR LFA is supported only on Catalyst 9400, 9500H and 9600
- IPv6 FRR is not supported
- MPLS with FRR is not supported
- Remote LFA is not supported
- OSPF and EIGRP Protocols are supported only

Высокая доступность:

Подводя итоги

Подводя итоги первой части:

- Высокая доступность сети:
  - Счастливое руководство и пользователи
  - Расслабленные архитекторы, инженеры и операторы IT
  - Необязательно дорого
- Структурированный и модульный дизайн – основа высокой доступности
- L2 Access – тоже может обладать высокой доступностью
- Routed Access – оптимальный вариант со своей спецификой

The bridge to possible