TOMORROW
starts here.

Cisco live!

Routed Fast Convergence

Agenda

- Thinking About Fast Convergence
- Reactive Convergence
- Proactive Convergence
- Closing Remarks

Agenda

- Thinking About Fast Convergence
  - Fast Convergence Mindset
  - Measuring Fast Convergence
- Reactive Convergence
- Proactive Convergence
- Closing Remarks

Fast Convergence Mindset

- How Fast?
  - 200ms (or less)
  - 50ms – SONET APS
- Do I Need It?
  - Complexity vs. Return
  - Business Drivers
  - Risks
- More than timers
  - Processes
  - Monitoring
  - Applications
  - **Everything Matters!**

Fast Convergence Mindset

- **Not** the same thing, but faster
- **Not** just about routing protocols
- **Not** just about failure recovery
- **Not** just about one node

Agenda

- ➢ Thinking About Fast Convergence
  - ○ Fast Convergence Mindset
  - ➢ Measuring Fast Convergence
- Reactive Convergence
- Proactive Convergence
- Other Convergence Tools
- Closing Remarks

Measuring Convergence

Convergence =

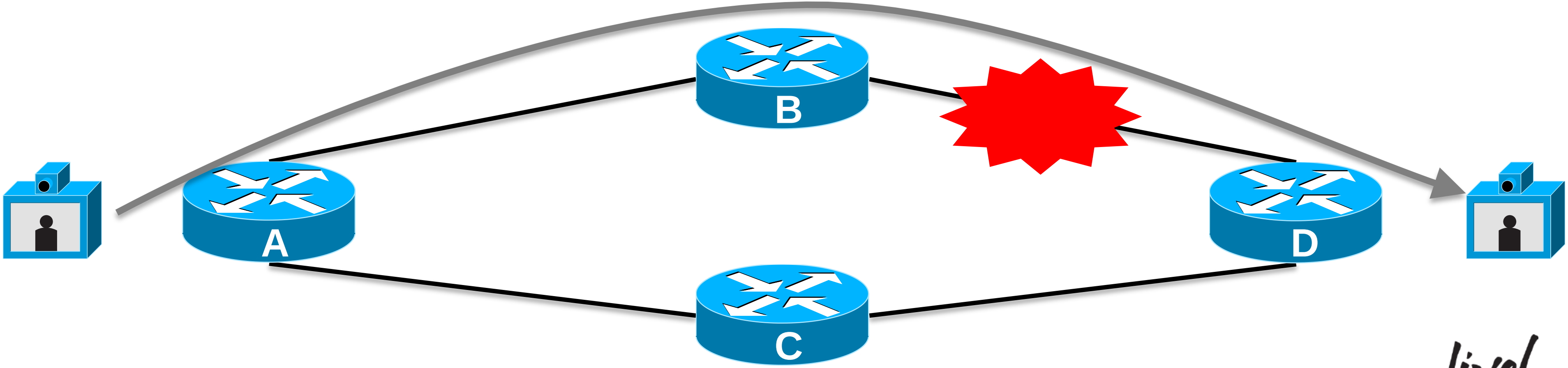
Failure Detection + Event Propagation + Routing Process + FIB Update

Neighbor Down

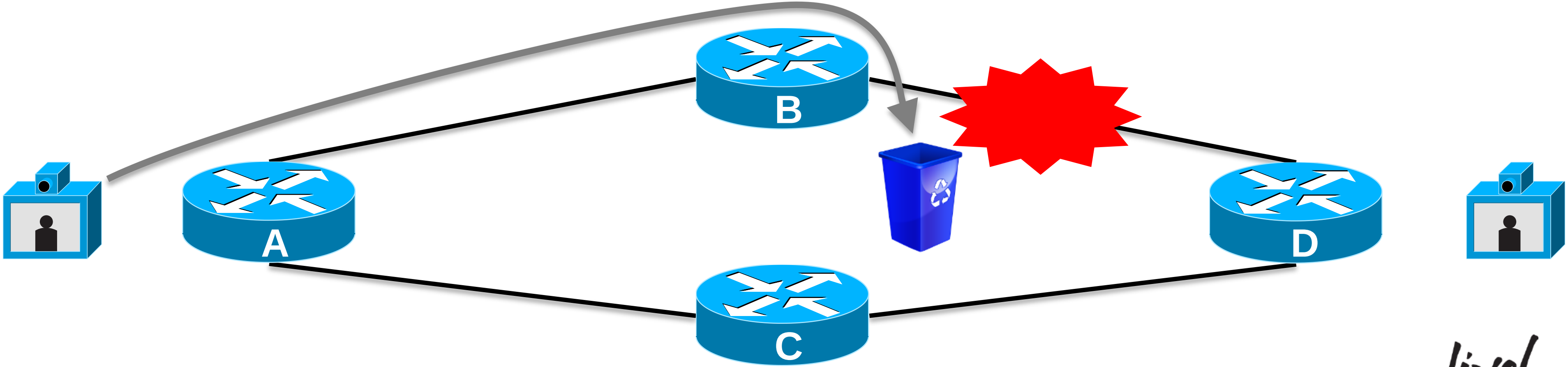
Tell Neighbors

RIB + CEF + Hardware

Measuring Fast Convergence

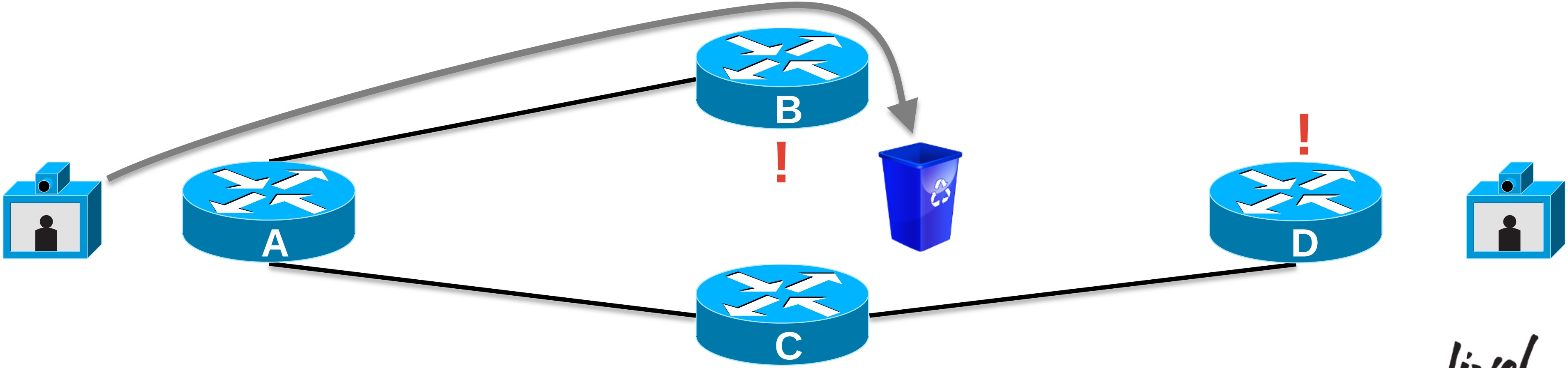


Measuring Fast Convergence



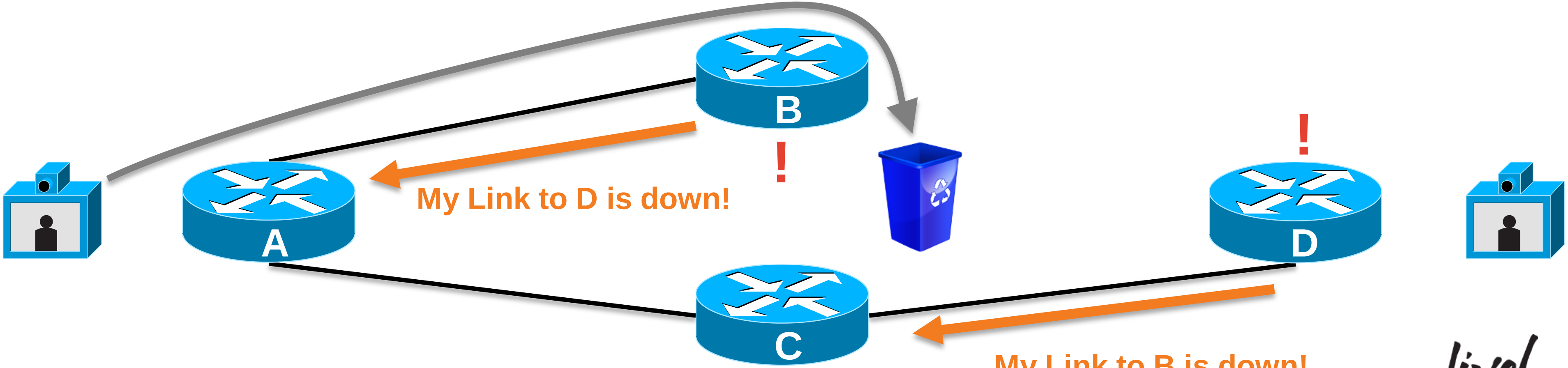
Measuring Fast Convergence

- Failure Detection
  - What happened?



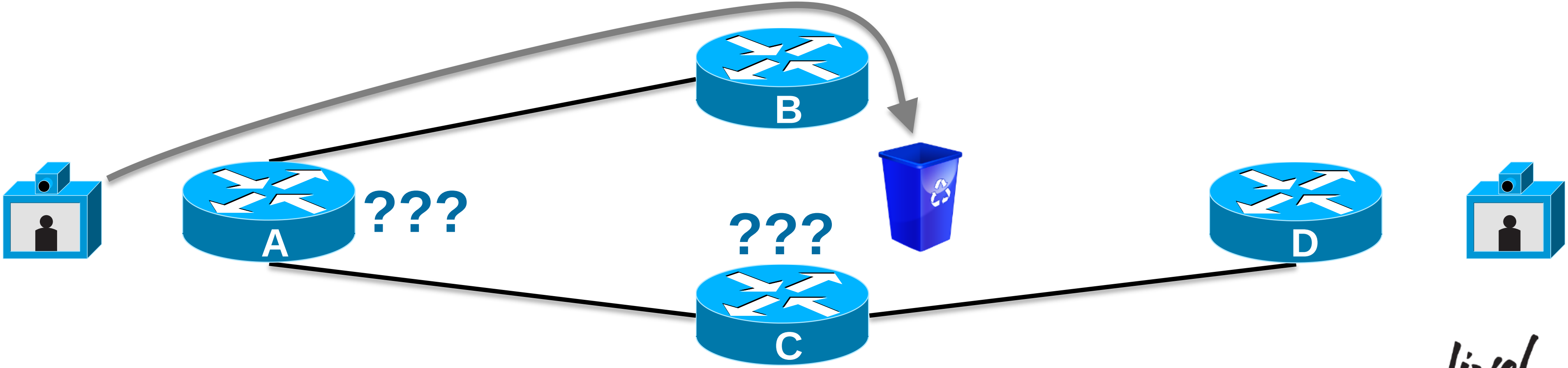
Measuring Fast Convergence

- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word



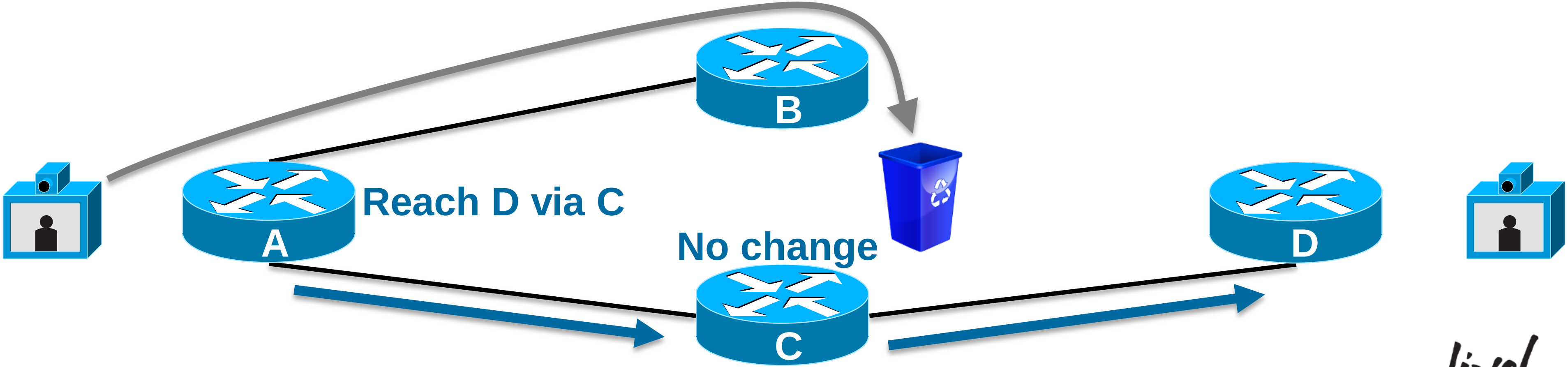
Measuring Fast Convergence

- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?



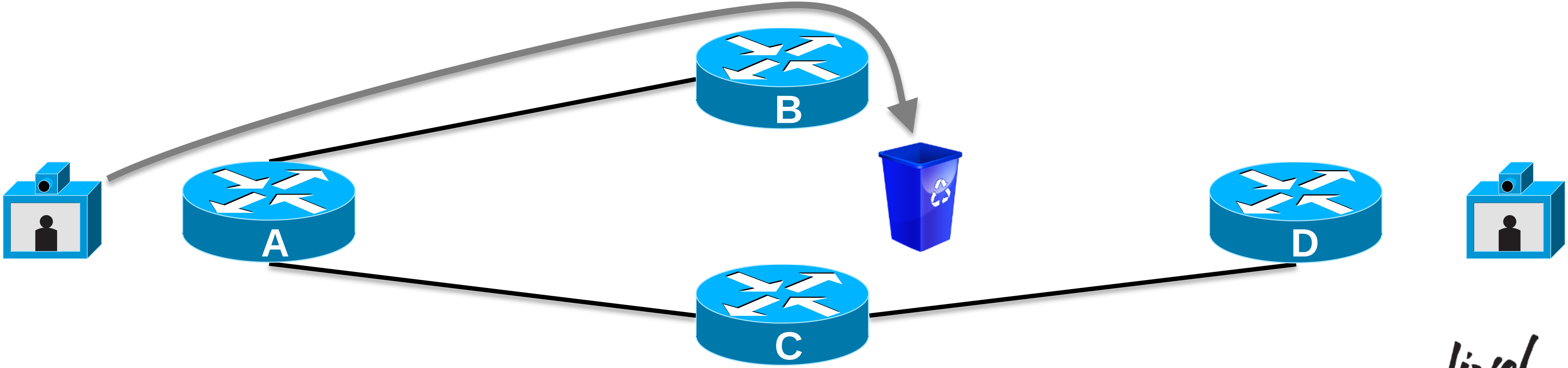
Measuring Fast Convergence

- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?



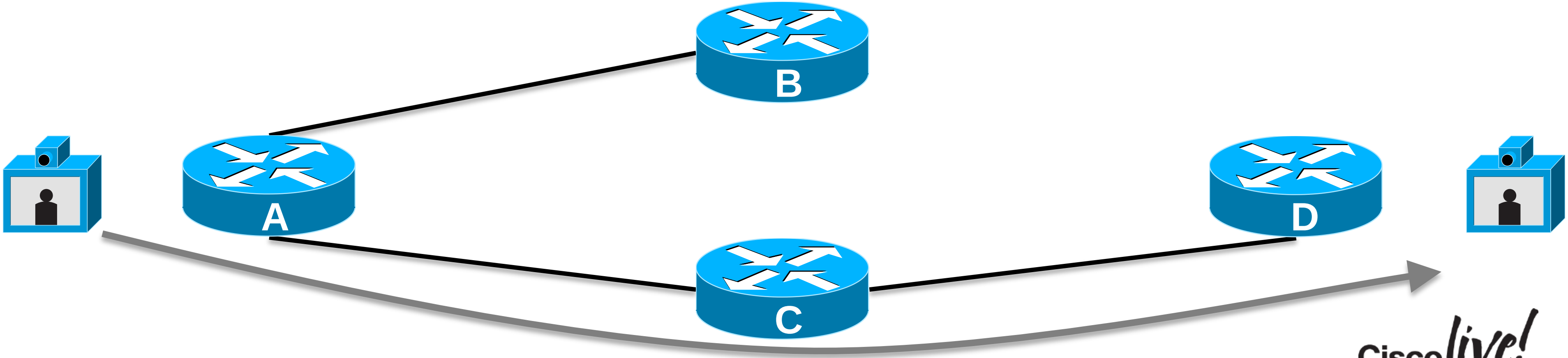
Measuring Fast Convergence

- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so



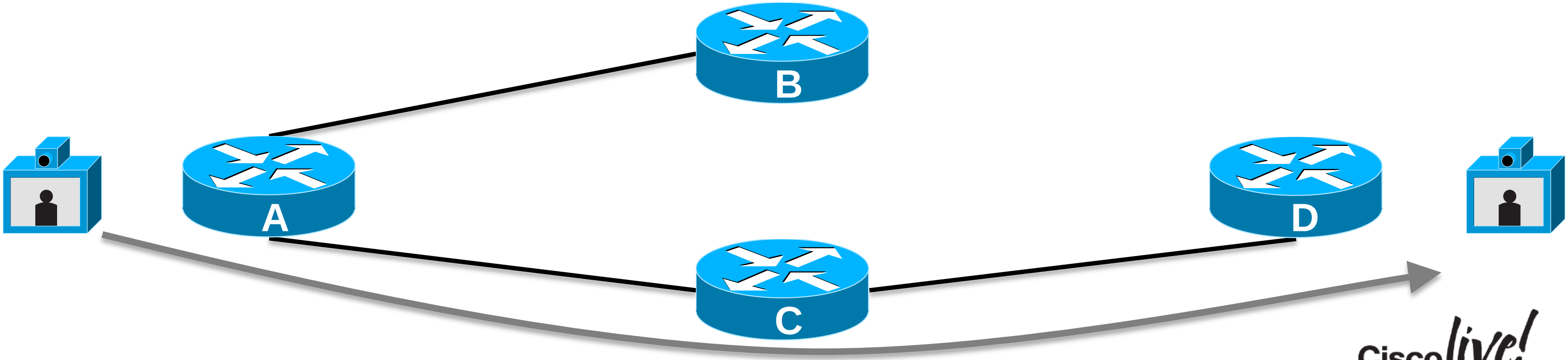
Measuring Fast Convergence

- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so



Measuring Fast Convergence

- Failure Detection 0 to 150 ms
  - What happened?
- Routing Process 10+ ms
  - Now where do we go?
- Event Propagation 0 to 10 ms
  - Spread the word
- FIB Update 0 ms to 5+ minutes
  - Make it so



Agenda

- ○ Thinking About Fast Convergence
- ➤ Reactive Convergence
  - ➤ Failure Detection
    - ➤ Detecting Link Failures
    - Fast Hellos and BFD
  - – Event Propagation
  - – Routing Update
  - – Forwarding Table Update
  - – BGP Convergence
- Proactive Convergence
- Closing Remarks

Measuring Fast Convergence

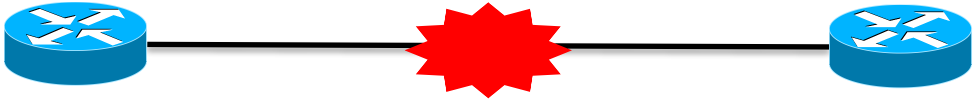
- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

Measuring Fast Convergence

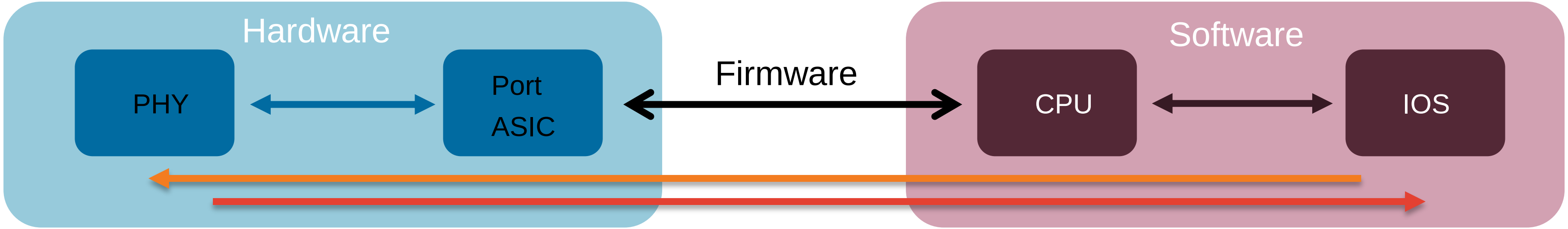
- Failure Detection
  - What happened?
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

Failure Detection

Detecting Link Failure



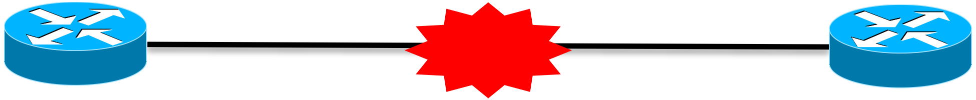
- Link Failure -> Interface Down, Easy?



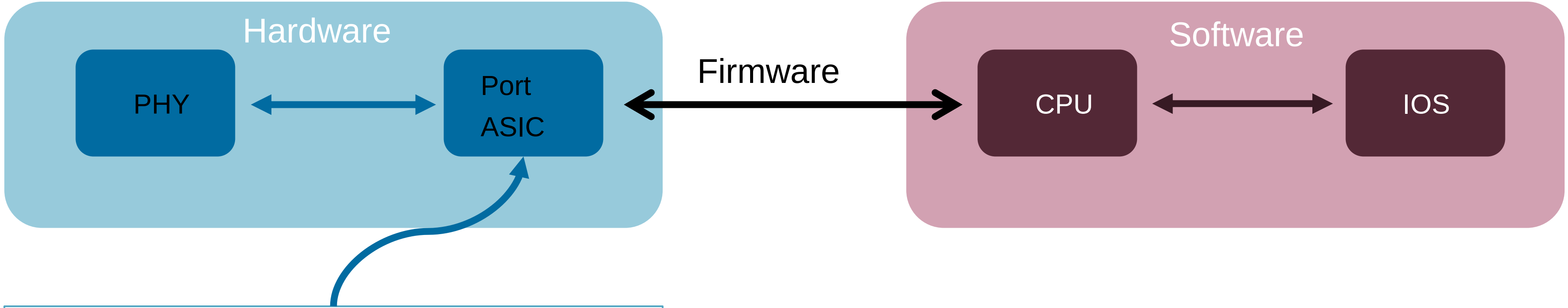
- Hardware Dependent
  - Polling vs Interrupt
    - 6748-GE-TX: 20ms/port * 48 ports = 960ms (polled)
    - Nexus 7k, ASR9k, 6708-10GE/ES/ES+: <10ms (interrupt)

Failure Detection

Detecting Link Failure



- Link Failure -> Interface Down, Easy?



Hardware

PHY

Port ASIC

Firmware

Software

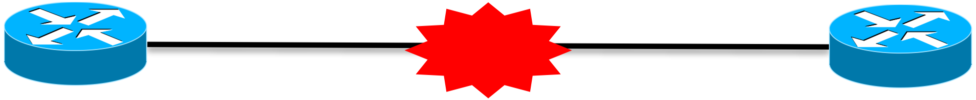
CPU

IOS

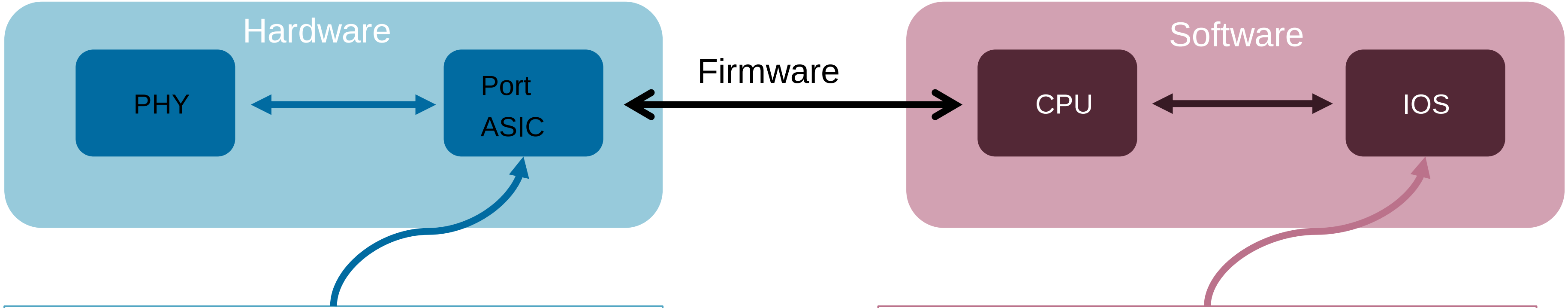
- Debounce Timer
  - Throttles *down* notification
  - Switches only

Failure Detection

Detecting Link Failure



- Link Failure -> Interface Down, Easy?



- Debounce Timer
  - Throttles *down* notification
  - Switches only

- Carrier Delay
  - Throttles up + down
  - Routers only

Detecting Link Failure

Debounce Timer

- Not Always Configurable
- Platform/Linecard/Media Dependent
  - 7600
    - 10ms on Fiber (10Gig)
    - 300ms Copper
  - NX-OS
    - 100ms
  - ASR9k
    - 0ms
- Recommendation: Leave unchanged

```
7600(config)# interface ...
7600(config-if)# link debounce time ...
```

Carrier Delay

- Generally Configurable
- Software Dependent
  - IOS/IOS-XE
    - 2 Seconds
  - NX-OS
    - 100ms (SVI Only)
  - XR/ASR9k
    - 0 ms
- Recommendation: 0 down, 2sec Up

```
7600(config)# interface ...
7600(config-if)# carrier-delay msec 0
7600(config-if)# carrier-delay up 2
```

Agenda

- Thinking About Fast Convergence
- Reactive Convergence
  - Failure Detection
    - Detecting Link Failures
    - Fast Hellos and BFD
  - Event Propagation
  - Routing Update
  - BGP Convergence
  - Forwarding Table Update
- Proactive Convergence
- Closing Remarks

Detecting the Event

Fast Hellos

- Normal Hellos…but fast!
  - ~1 second detection
- Process Driven
- 1 Hello/Protocol
  - PIM, LDP, BGP, OSPF
- Handled by Central CPU
- 50+ Bytes

BFD

- Even Faster
  - 50ms x 3 = 150ms detection
- Interrupt Driven (like CEF)
- 1 Hello to Rule Them All
- Hardware Offload Possible
  - Nexus 7k, ASR 1k/9k, me3600-CX, 7600 ES+
- ~24 bytes

Detecting the Event

~~Fa Hellos~~

- ~~No l Hellos. . fast!~~
  - ~~~1 s nd de ion~~
- ~~Proces n~~
- ~~1 Hello/P ol~~
  - ~~PIM P, B OSPF~~
- ~~Ha d by Ce l CPU~~
- ~~Bytes~~

BFD

- Even Faster
  - 50ms x 3 = 150ms detection
- Interrupt Driven (like CEF)
- 1 Hello to Rule Them All
- Hardware Offload Possible
  - Nexus 7k, ASR 1k/9k, me3600-CX, 7600 ES+
- ~24 bytes

Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

Agenda

- Thinking About Fast Convergence
- Reactive Convergence
  - Failure Detection
  - Event Propagation
  - Routing Update
  - BGP Convergence
  - Forwarding Table Update
- Proactive Convergence
- Closing Remarks

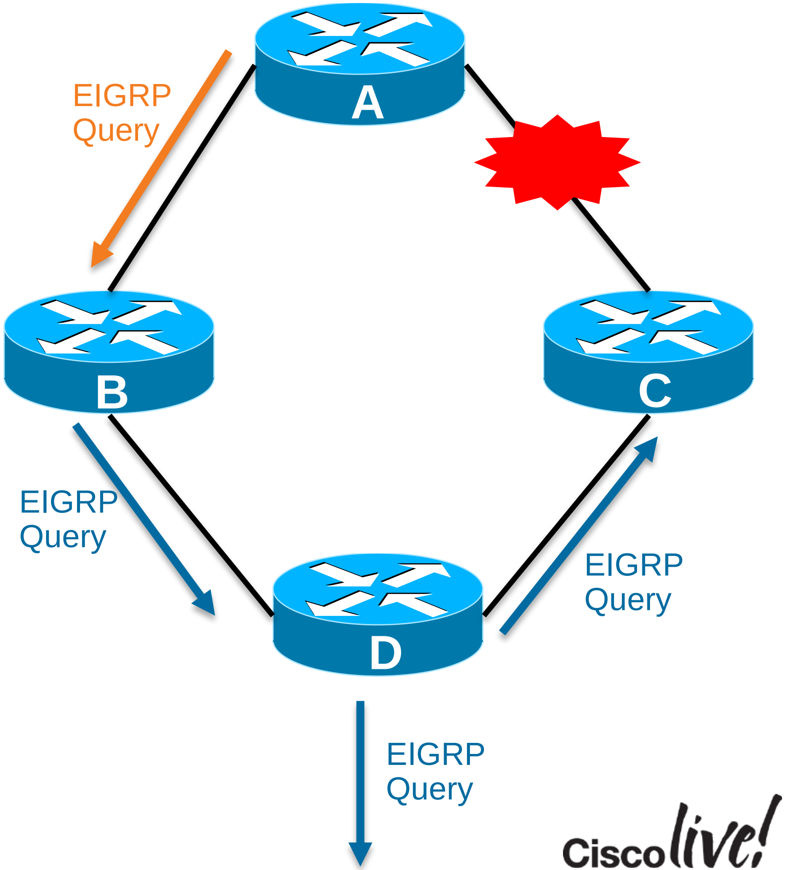
Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)
- Event Propagation
  - Spread the word
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

Event Propagation in EIGRP

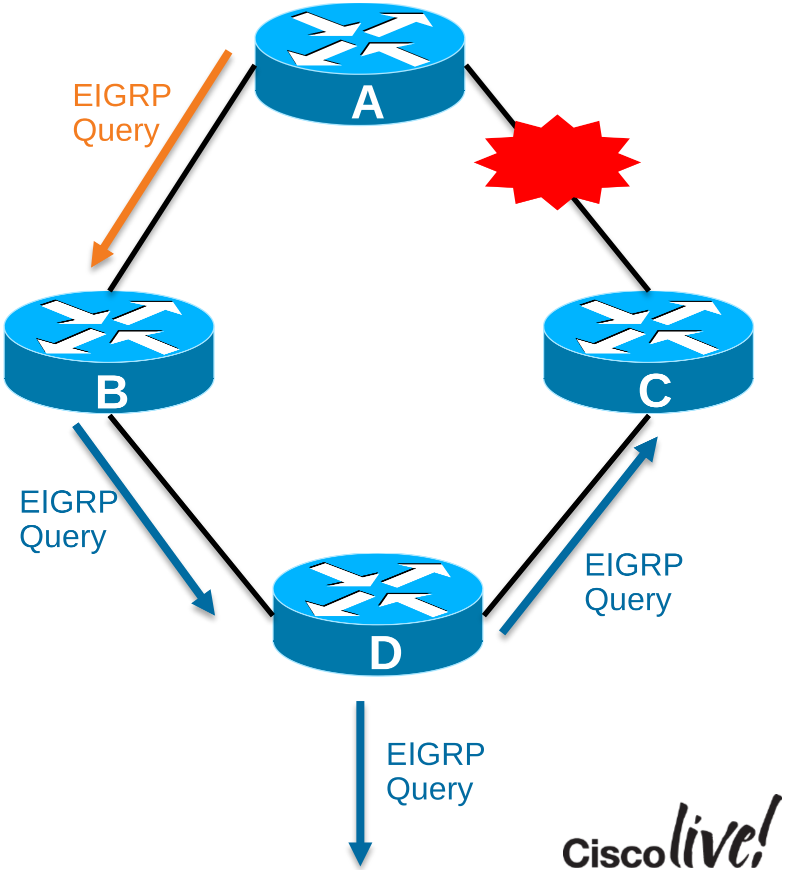
- The Good
  - Immediate event notification

- The Bad
  - Query Domain Size



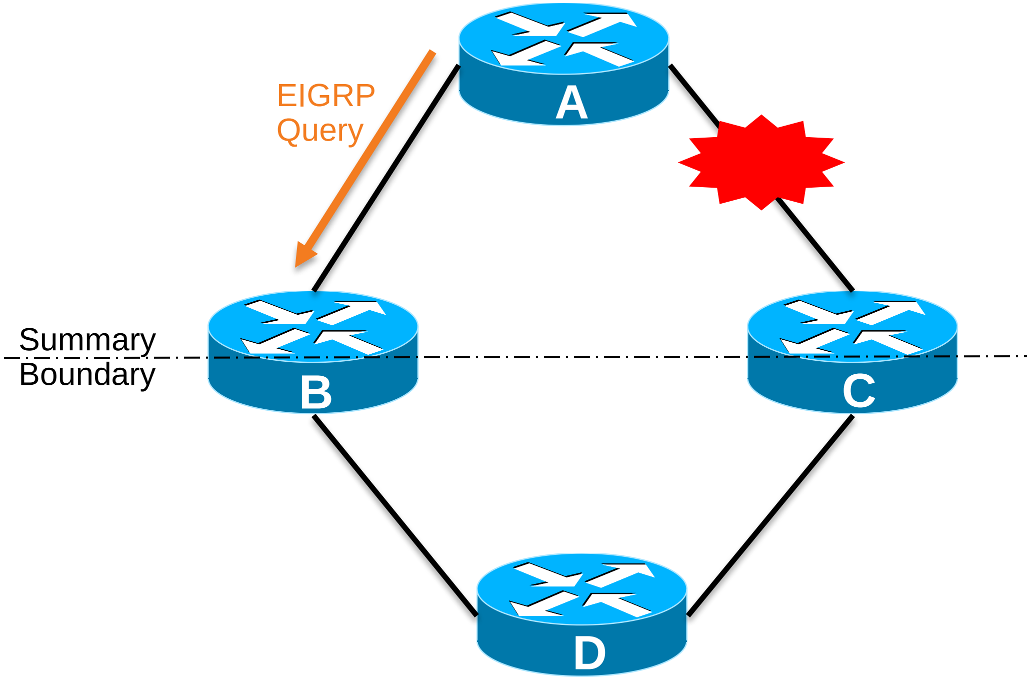
Event Propagation in EIGRP

- The Good
  - Immediate event notification
- The Bad
  - Query Domain Size
- The Ugly
  - Stuck In Active



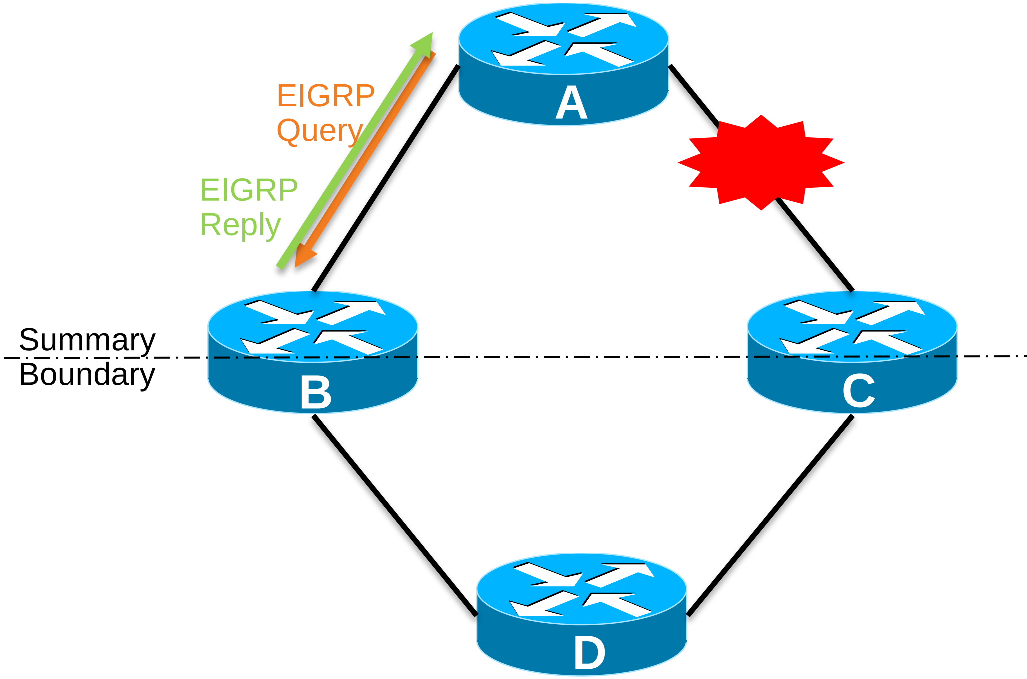
Improving EIGRP Event Propagation

- Reduce Query Domains
  - Summary
  - Stub
  - Filters



Improving EIGRP Event Propagation

- Reduce Query Domains
  - Summary
  - Stub
  - Filters

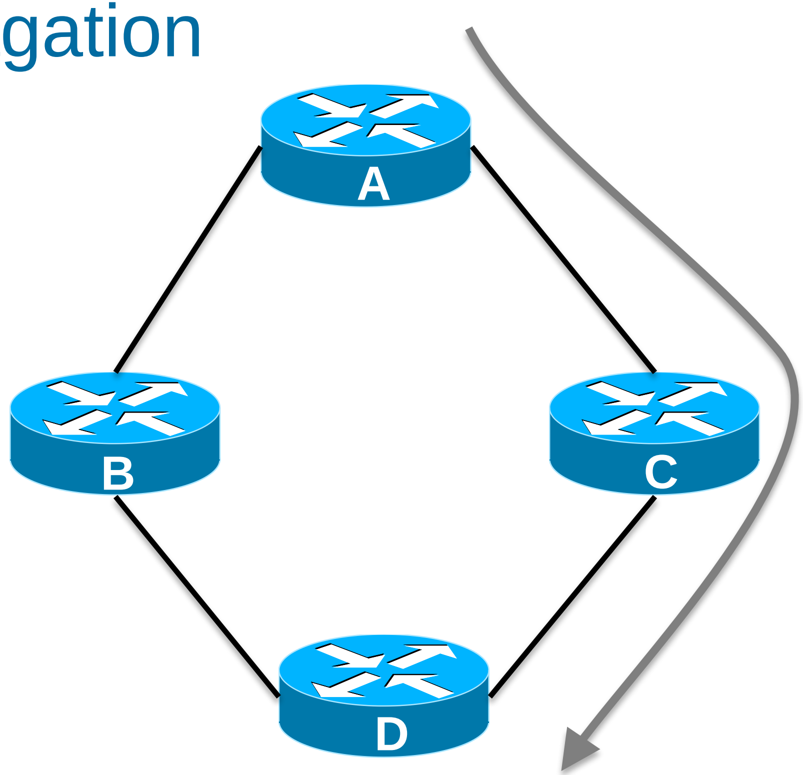


Improving EIGRP Event Propagation

- Reduce Query Domains
  - Summary
  - Stub
  - Filters

<10s ms

- Feasible Successors
  - Don't even ask!



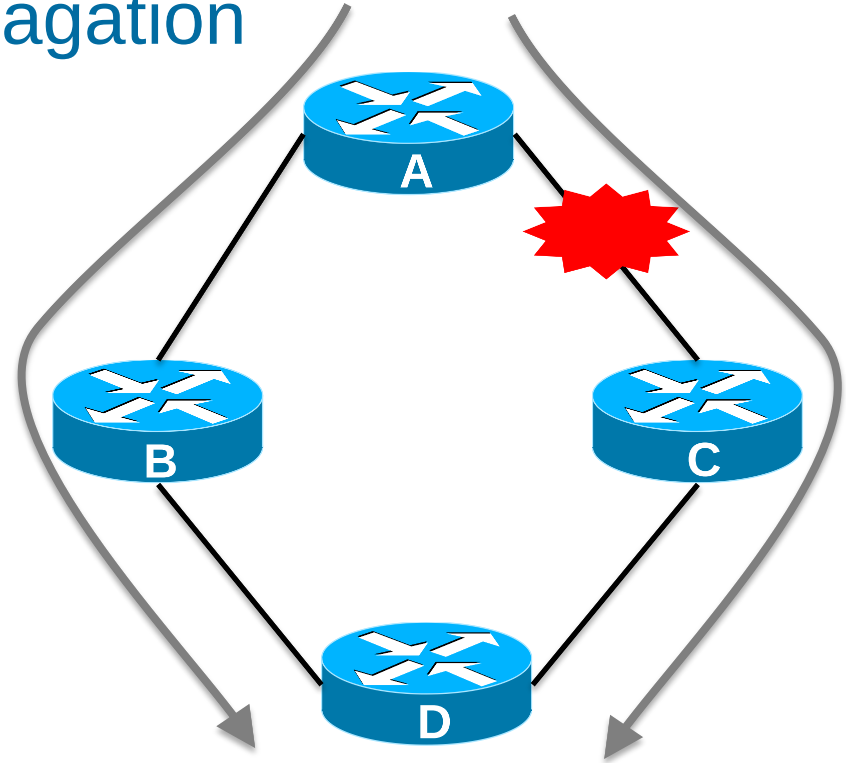
Improving EIGRP Event Propagation

- Reduce Query Domains
  - Summary
  - Stub
  - Filters

<10s ms

- Feasible Successors
  - Don’t even ask!
  - No Query/Reply

~0 ms



Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)
- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

Improving OSPF Event Propagation

- LSAs Sent After Change
- Delay for Router/Network LSAs
  - XE: 5000ms
  - NX-OS: 200ms
  - XR: 50ms

```
XE-XR(config)#router ospf 10
XE-XR(config-router)#timers throttle lsa [all] <start> <hold> <max>
XE-XR(config-router)#timers lsa arrival <timer>
```

- Start: First LSA
- Hold: Repeat LSA (flap) (*2)
- Max: Maximum Wait Time

Improving OSPF Event Convergence

Wild Side

- Start: 0ms
- Hold: 20ms
- Max: 5000ms

Nice and Easy

- Start: 5ms
- Hold: 40ms
- Max: 10000ms

General theory for timer tuning

React immediately the first time, then wait significant periods of time for subsequent events

Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)
- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
  - OSPF LSA Throttling (~0-5 ms)
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

Improving ISIS Event Propagation

- Default LSP Generation 50ms (XE/XR/NX-OS)
- SPF runs on change
  - Can beat LSP Propagation

```
XE-NX(config)#router isis CLUS
XE-NX(config-router)#lsp-gen-interval <max> <initial>
XE-NX(config-router)#fast-flood
```

```
RP/0/RSP0/CPU0:XR# configure
RP/0/RSP0/CPU0:XR(config)# router isis CLUS
RP/0/RSP0/CPU0:XR(config-isis)# lsp-gen-interval initial-wait <time>
RP/0/RSP0/CPU0:XR(config-isis)# interface g0/3/0/0
RP/0/RSP0/CPU0:XR(config-isis-if)# lsp fast-flood threshold <num of LSPs>
```

Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)

- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
  - OSPF LSA Throttling (0-5 ms)
  - ISIS LSP Fast Flooding (1 ms)

- Routing Process
  - Now where do we go?

- FIB Update
  - Make it so

Agenda

- Thinking About Fast Convergence
- Reactive Convergence
  - Failure Detection
  - Event Propagation
  - Routing Update
    - BGP Convergence
    - Forwarding Table Update
- Proactive Convergence
- Closing Remarks

Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)
- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
  - OSPF LSA Throttling (0-5 ms)
  - ISIS LSP Fast Flooding (1 ms)
- Routing Process
  - Now where do we go?
- FIB Update
  - Make it so

EIGRP Routing Update

- Based on DUAL Algorithm
- Runs when all Queries are answered
  - Doesn’t run with Feasible Successors (no query!)
- Only calculates changed prefixes
  - Not much work compared to link-state protocols
  - DUAL can finish in < 1ms

Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)

- Routing Process
  - Now where do we go?
  - EIGRP DUAL (<1 ms)

- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
  - OSPF LSA Throttling (0-5 ms)
  - ISIS LSP Fast Flooding (1 ms)

- FIB Update
  - Make it so

OSPF Routing Update

- SPF Run on LSA Reception
- Delayed by Default
  - XE: 5 seconds
  - NX-OS: 200ms
  - XR: 50ms

```
XE-XR(config)#router ospf 10
XE-XR(config-router)#timers throttle spf <start> <hold> <max>
```

- Start: First SPF run
- Hold: Repeat SPF run
- Max: Maximum Wait Time

ISIS Routing Update

- SPF Run on LSP Reception
- Delayed by Default
  - XE: 10 seconds
  - NX-OS: 50ms
  - XR: 50ms

```
XE-XR(config)#router isis CLUS
XE-XR(config-router)#spf-interval <max> <start> <hold>

XE(config-router)#prc-interval <max> <start> <hold>
```

- Start: First SPF run
- Hold: Repeat SPF run
- Max: Maximum Wait Time

PRC and iSPF

- PRC – Partial Route Calculation
  - Route change without topology change
  - No SPF run
  - Default in OSPF (Type 4/5)
  - ISIS
    - XE: extra configurable timer
    - NX-OS/ XR: baked in

- iSPF – incremental SPF
  - Runs SPF shortcut
  - Only relevant to some network changes
  - Minor difference on modern platforms
  - Disabled by default
  - Not recommended*

Measuring Fast Convergence

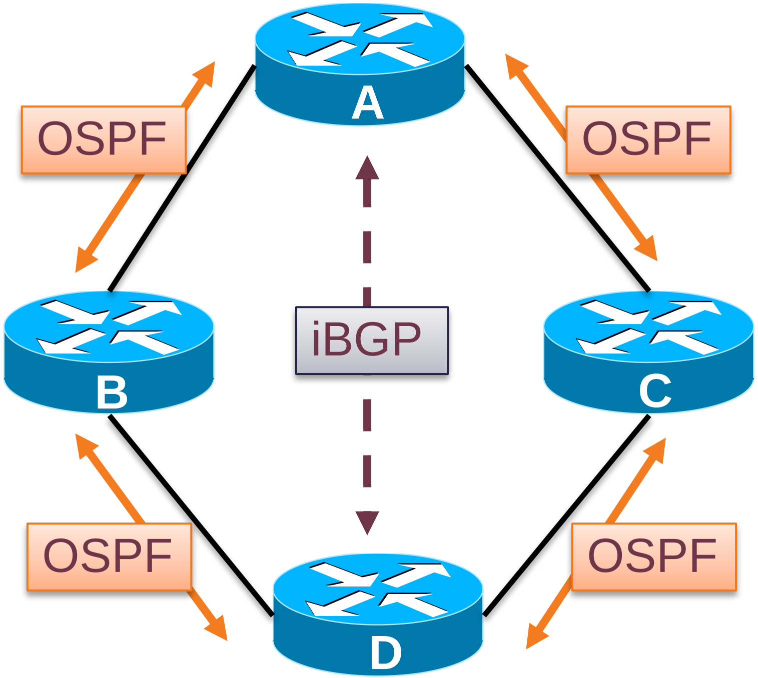
- Failure Detection
  - What happened?
  - BFD (150 ms)
- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
  - OSPF LSA Throttling (0-5 ms)
  - ISIS LSP Fast Flooding (1 ms)
- Routing Process
  - Now where do we go?
  - EIGRP DUAL (<1 ms)
  - ISIS/OSPF SPF (5ms)
- FIB Update
  - Make it so

Agenda

- Thinking About Fast Convergence
- Reactive Convergence
  - Failure Detection
  - Event Propagation
  - Routing Update
  - BGP Convergence
  - Forwarding Table Update
- Proactive Convergence
- Closing Remarks

BGP Fast Convergence Primer

- BGP != IGP
- Different Goals
- Lots of Data….
  - ….means lots of CPU
  - ….means lots of memory
  - ….means lots of packets
- BGP generally relies on IGP
- Little Events vs. Big Events
  - Route Flap vs. `clear ip bgp *`



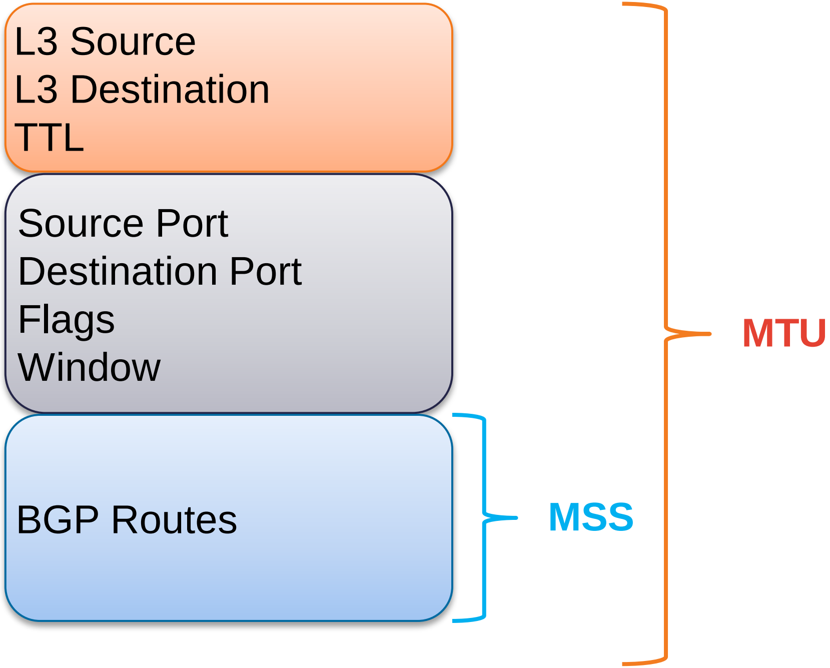
Think about data plane over control plane

BGP Failure Detection

- Keepalives
  - 60/180s default
  - Don’t tune (at least not aggressively)
- BFD
  - `neighbor <> fall-over bfd`
- Interface Tracking
  - Notifies BGP if interface/route down
  - Enabled by default

BGP Event Propagation

- MTU
  - Bigger packets
- BGP Based On TCP
  - MSS
    - Maximum amount of TCP data
  - Window Size
    - Local TCP buffer
    - ACKs reduce window as it fills
- Update Groups
  - Single policy update per group
  - More groups = more work



BGP Routing Update

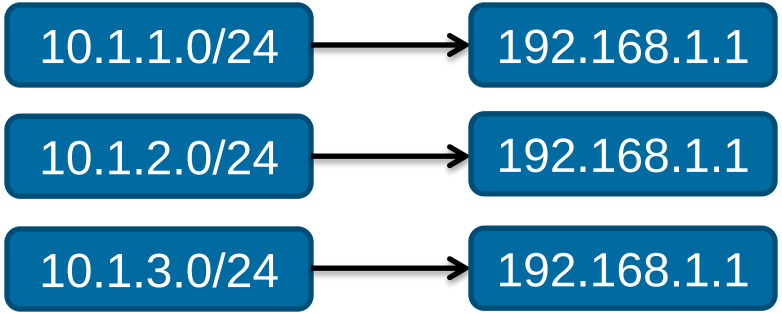
- BGP Scanner
  - Old and Busted
  - The janitor of BGP
  - Runs every 60 seconds
- Next Hop Tracking
  - New Hotness
  - Event driven (3-5 sec delay)
  - IGP metric or path change

```
XE-NX(config)# router bgp 65535
XE-NX(config-router)# bgp nexthop trigger-delay <>
```

```
RP/0/RSP0/CPU0:XR# configure
RP/0/RSP0/CPU0:XR(config)# router bgp 65535
RP/0/RSP0/CPU0:XR(config-bgp)# address-family ipv4 unicast
RP/0/RSP0/CPU0:XR(config-bgp-af)# nexthop trigger-delay critical <> non-critical <>
```

BGP Routing Update – PIC Core

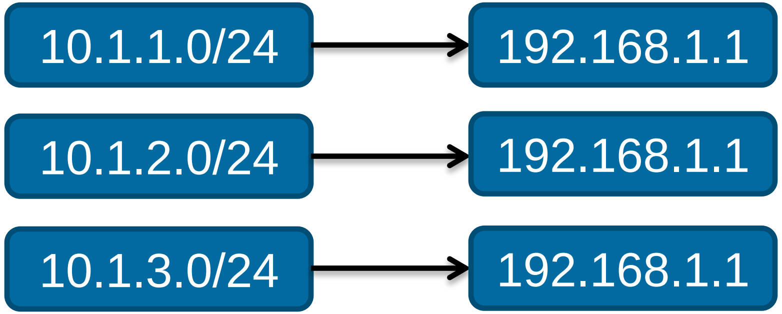
- Flat RIB = slow convergence



- Before PIC
  - Update per route
  - Convergence dependent on BGP RIB size

BGP Routing Update – PIC Core

- Instead of flat FIB, Hierarchical



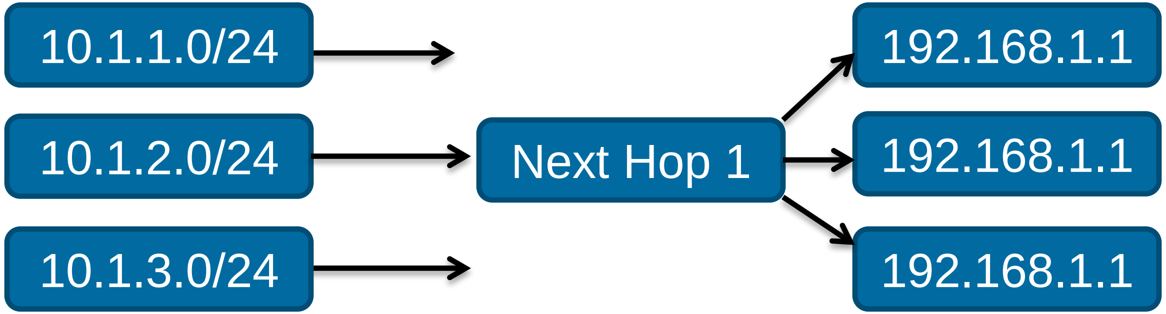
10.1.1.0/24 → 192.168.1.1

10.1.2.0/24 → 192.168.1.1

10.1.3.0/24 → 192.168.1.1

BGP Routing Update – PIC Core

- Instead of flat FIB, Hierarchical



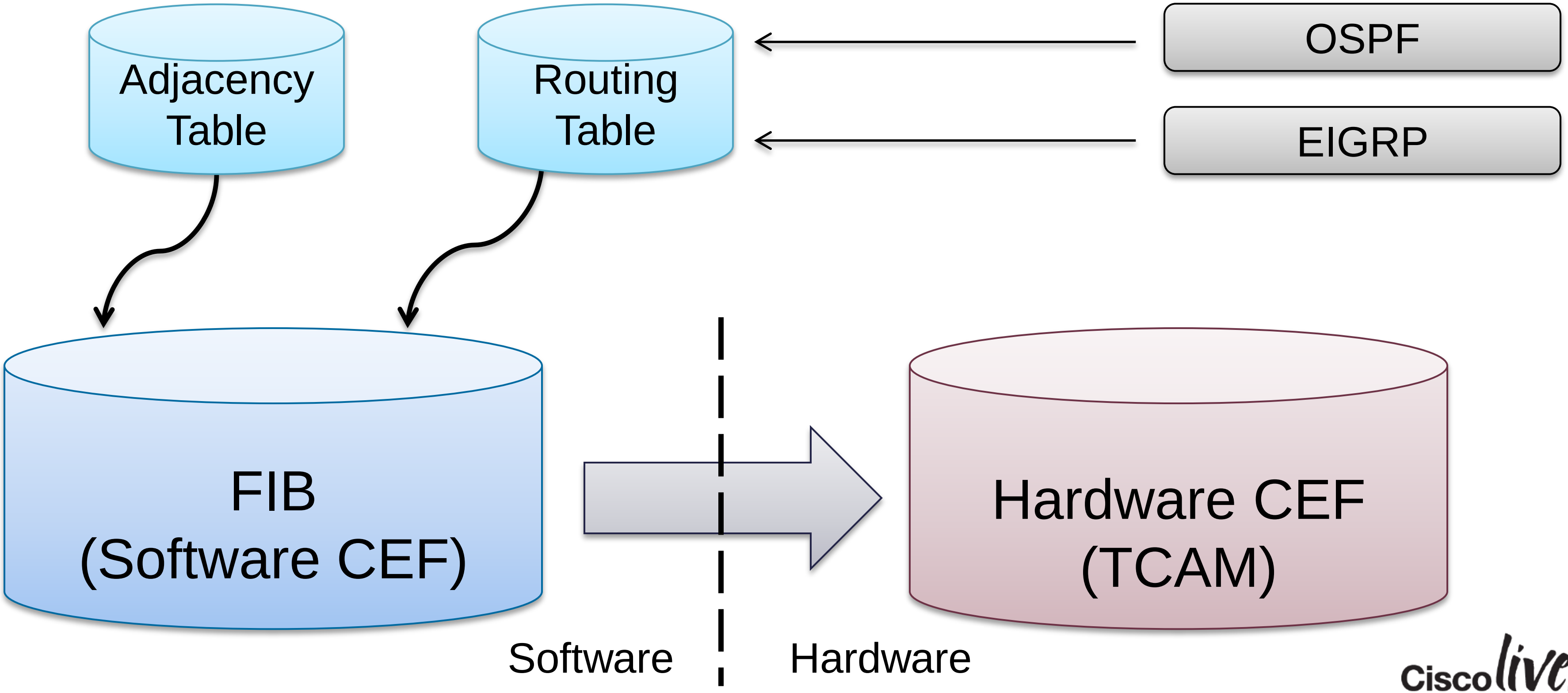
- Single change updates multiple entries
- Convergence time independent from prefix count

```
7600(config)# cef table output-chain build favor convergence-speed
```

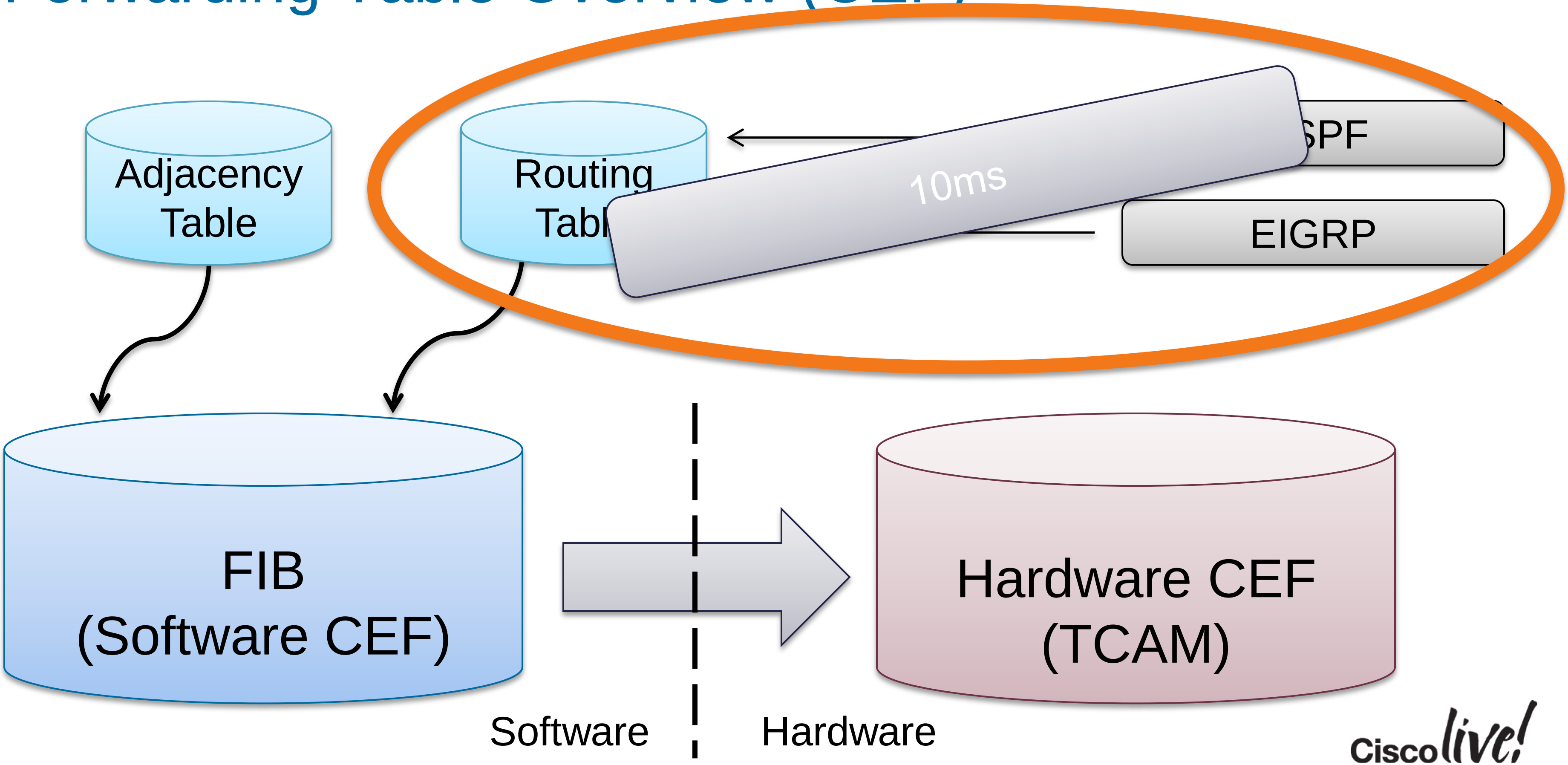
Agenda

- Thinking About Fast Convergence
- ➤ Reactive Convergence
  - Failure Detection
  - Event Propagation
  - Routing Update
  - BGP Convergence
  - ➤ Forwarding Table Update
- Proactive Convergence
- Closing Remarks

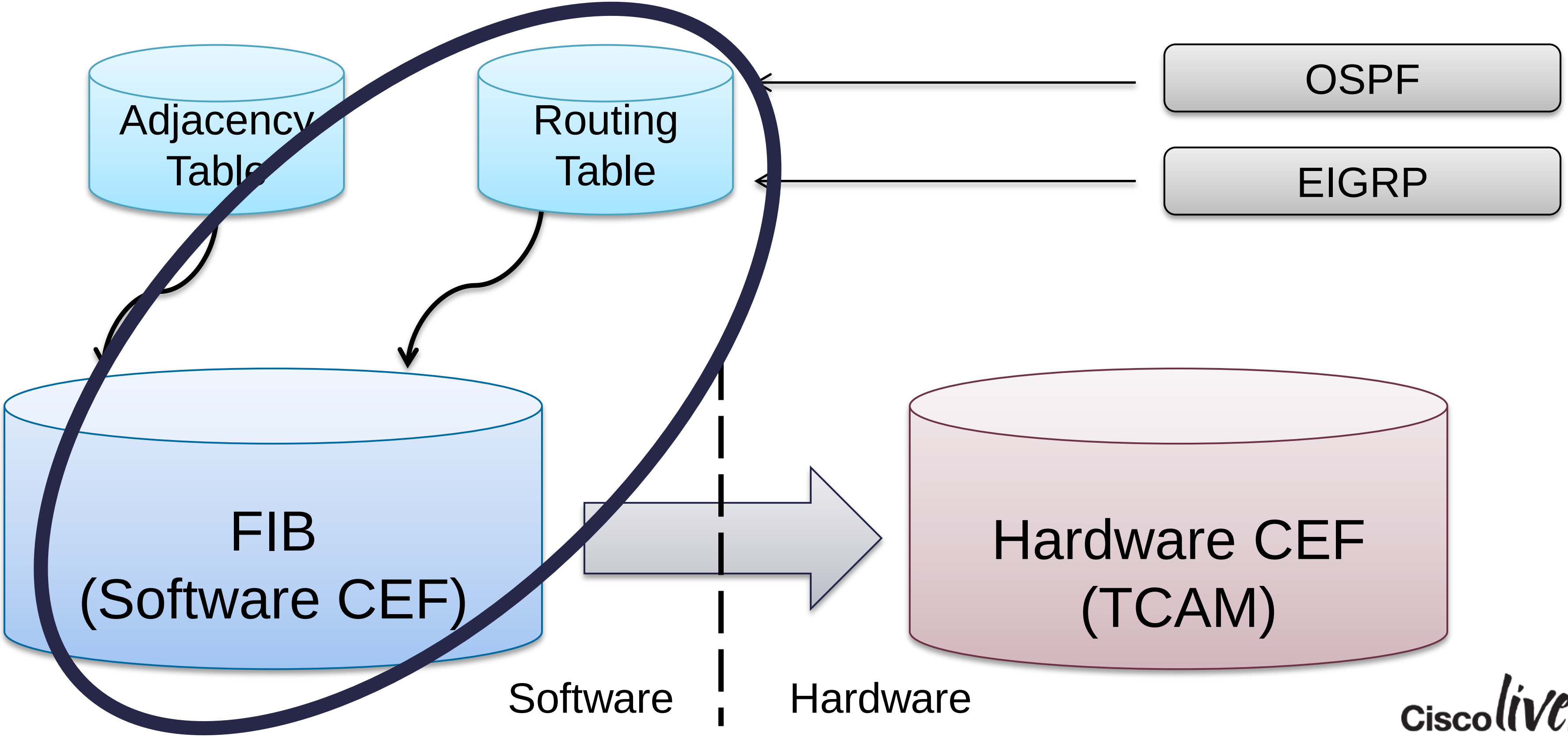
Forwarding Table Overview (CEF)



Forwarding Table Overview (CEF)

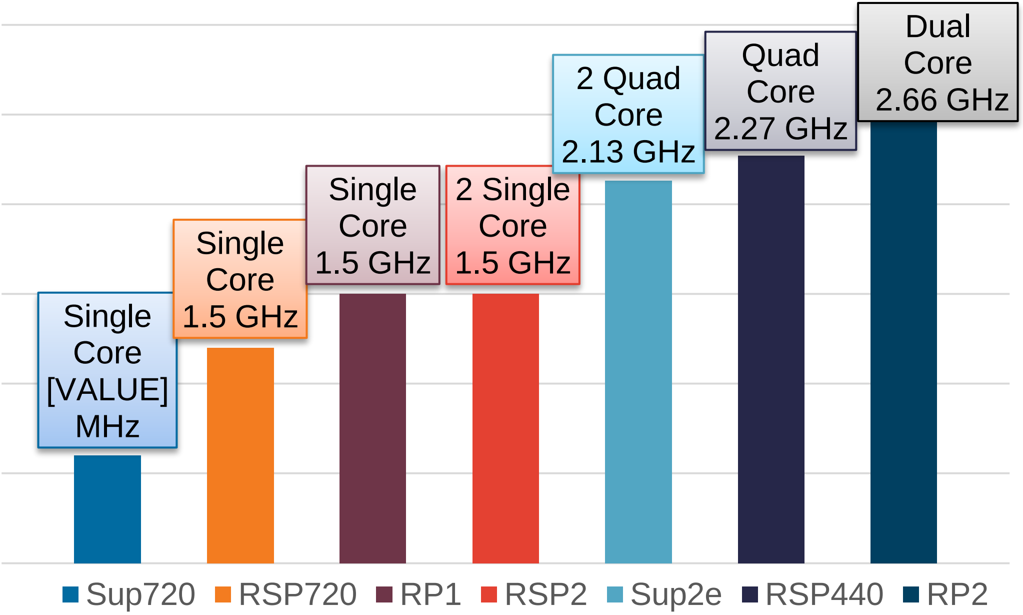


Forwarding Table Overview (CEF)

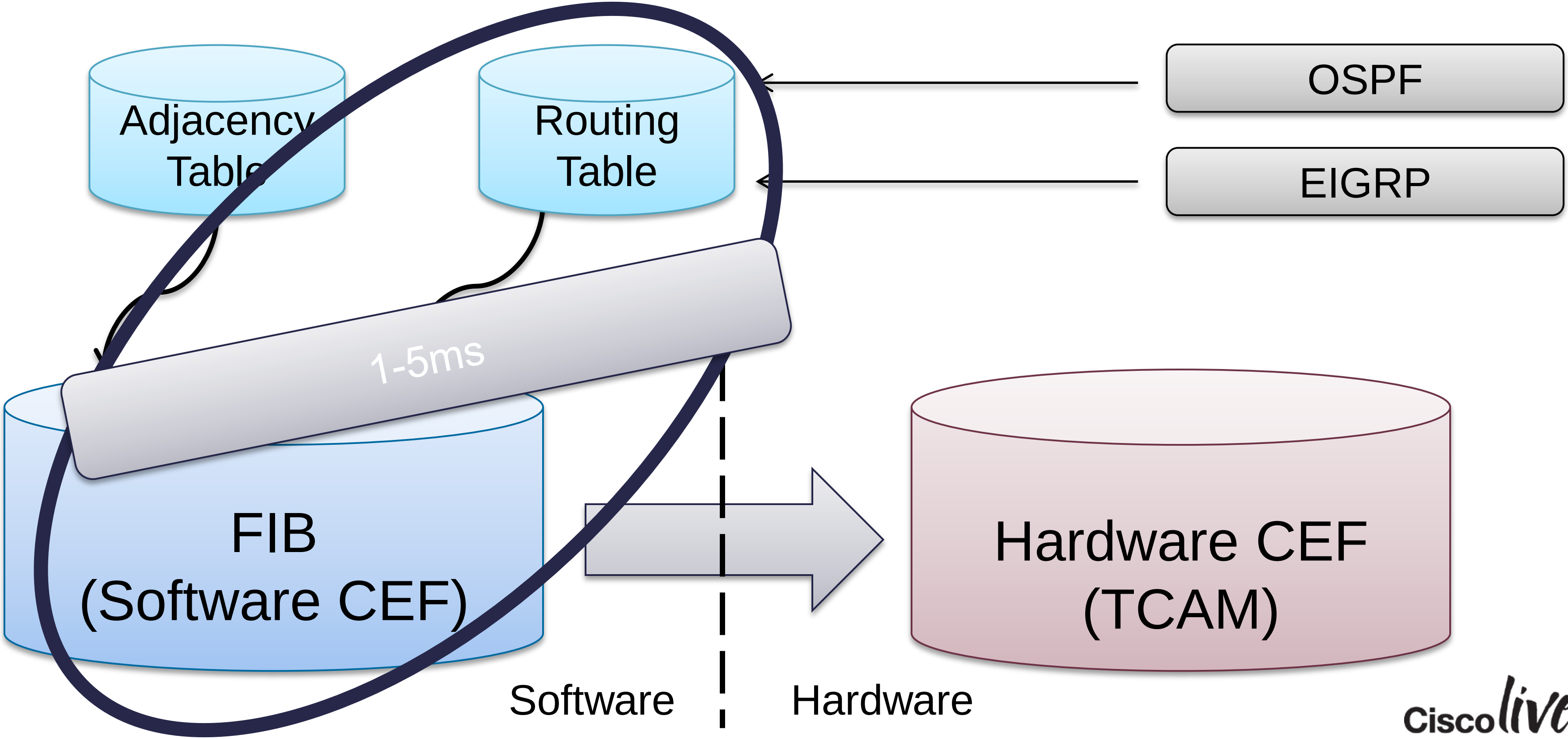


Software CEF Updates

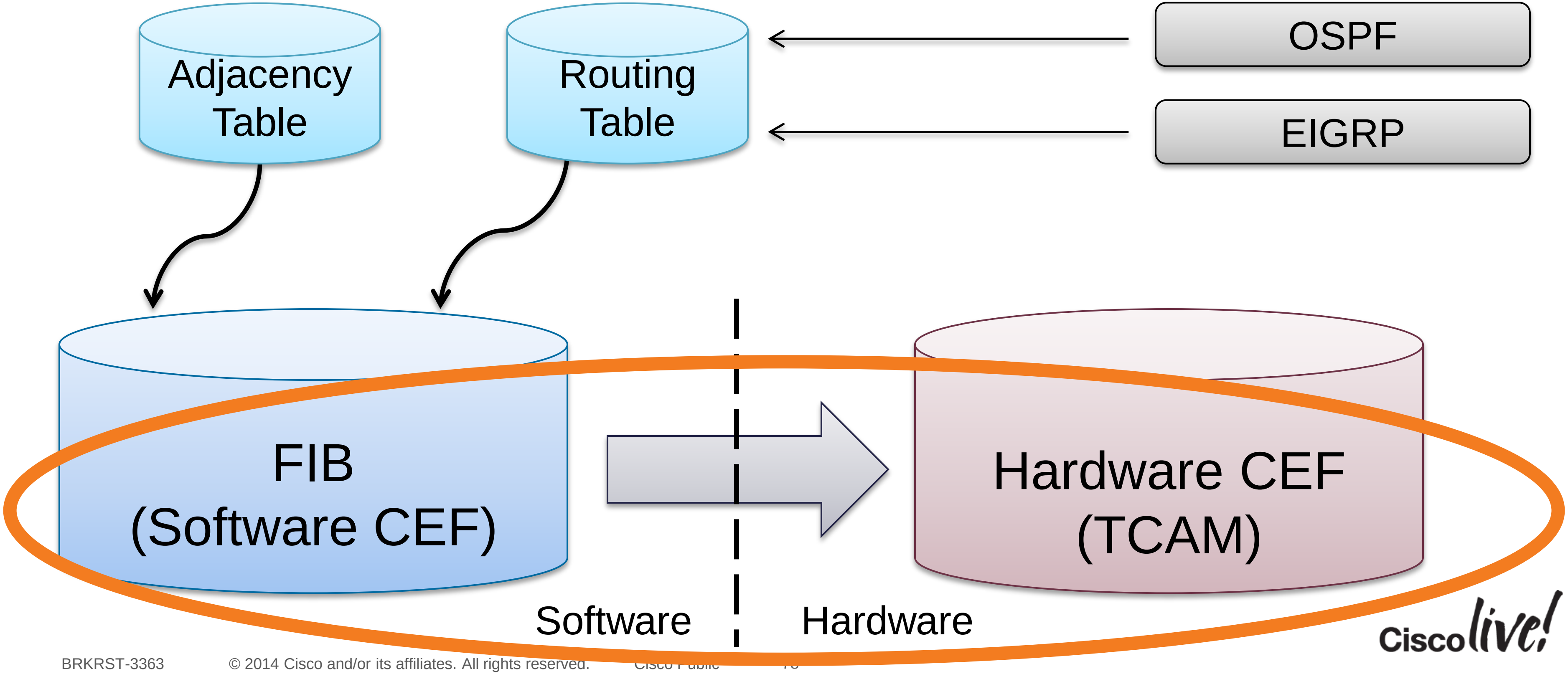
- Controlled by CPU + OS
- Supervisors Matter
- RIB Size Matters
- Summarize and Filter
  - XE: OSPF prefix suppression
  - XR/XE: ISIS advertise passive-only
- Process quantum
  - XE only
- Prefix Prioritization
  - Install /32s first



Forwarding Table Overview (CEF)



Forwarding Table Overview (CEF)



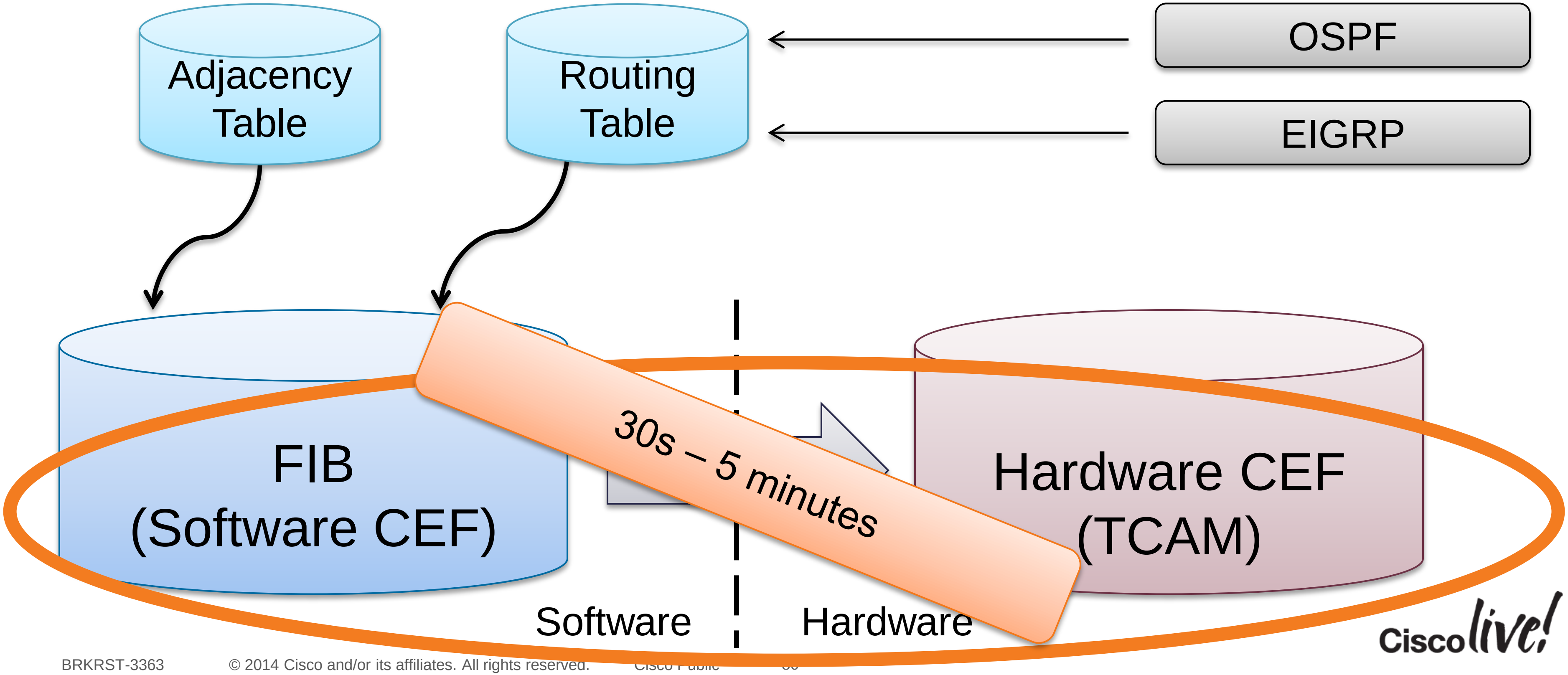
Hardware CEF Updates

- TCAM/SRAM Based Platforms
  - Fast Reads (linerate)
  - Slooooow Writes
- Can be slowest piece to converge
  - 350k Routes ~27 seconds
  - 700k Routes ~360 seconds
- Hardware Matters!
- 7600:

```
7600(config)#hw-module slot <mod> process-max-time 50
7600(config)#hw-module slot <sup> sp process-max-time 50
```

- Work Smarter, Not Harder!

Forwarding Table Overview (CEF)



Measuring Fast Convergence

- Failure Detection
  - What happened?
  - BFD (150 ms)

- Routing Process
  - Now where do we go?
  - EIGRP DUAL (<1 ms)
  - ISIS/OSPF SPF (5ms)

- Event Propagation
  - Spread the word
  - EIGRP Feasible Successors (~0 ms)
  - OSPF LSA Throttling (0-5 ms)
  - ISIS LSP Fast Flooding (1 ms)

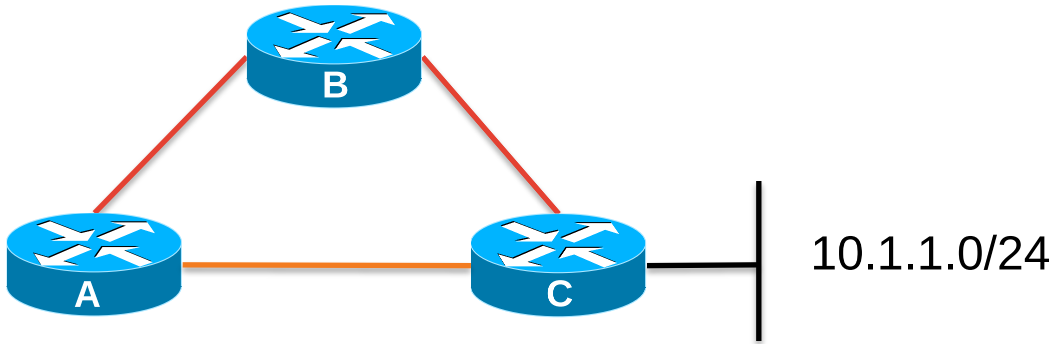
- FIB Update
  - Make it so (1ms-5 min)

153ms – 5+ minutes

Agenda

- o Thinking About Fast Convergence
- o Reactive Convergence
- ➢ Proactive Convergence
  - ➢ Loop Free Alternate (IP FRR)
  - – BGP PIC Edge
  - – LDP Session Protection
- Closing Remarks

OSPF Loop Free Alternate



- A has a primary (A-C) and secondary (A-B-C) path to 10.1.1.0/24
- Link State allows A to know entire topology
- A should know that B is an alternative path
- Loop Free Alternate (LFA)

OSPF Loop Free Alternate

- OSPF presents a primary and backup to CEF
  - Backup calculated from secondary SPF run

```
RouterA# show ip route 10.1.1.0
Routing Descriptor Blocks:
 * 172.16.0.1, from 192.168.255.1, 00:01:57 ago, via Ethernet4/1/0
      Route metric is 2, traffic share count is 1
      Repair Path: 192.168.0.2, via Ethernet4/2/0

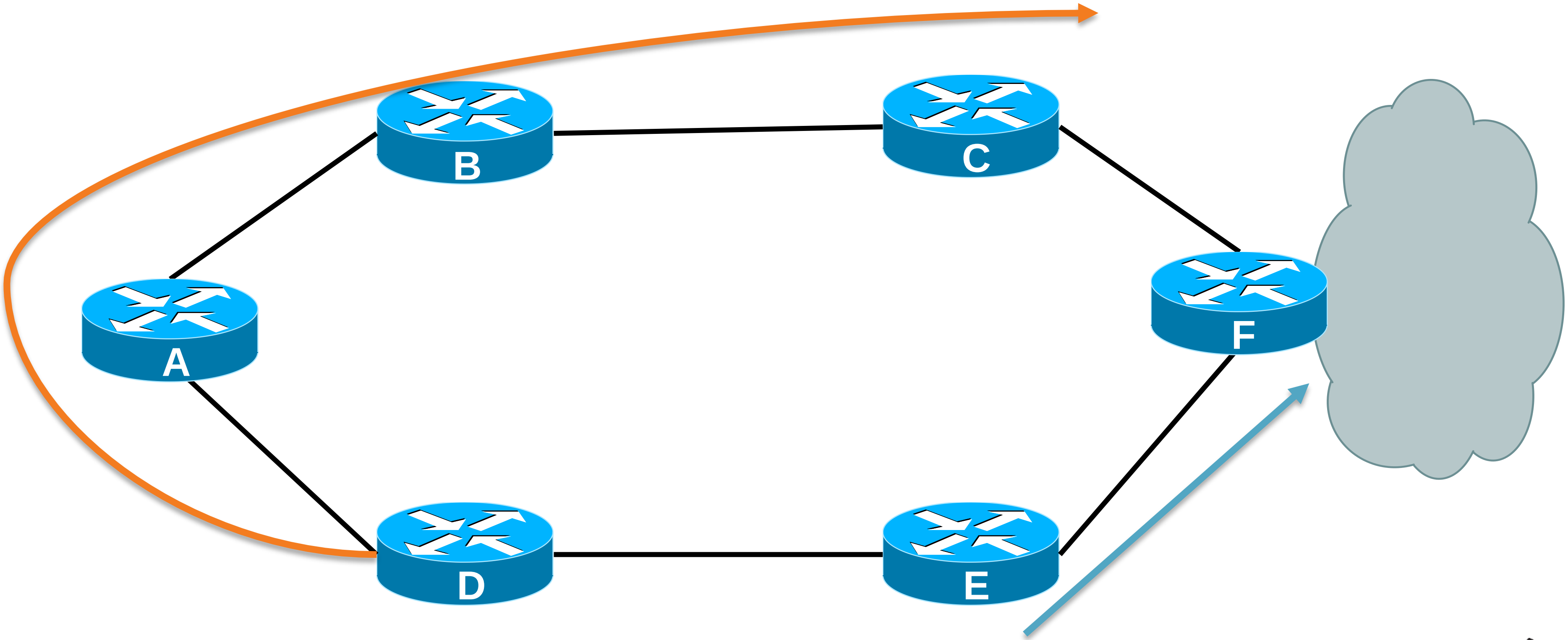
RouterA#show ip CEF 10.1.1.0
10.1.1.0/24
  nexthop 172.16.0.1 Ethernet4/1/0
    repair: attached-nexthop 192.168.0.2 Ethernet4/2/0
```

EIGRP LFA

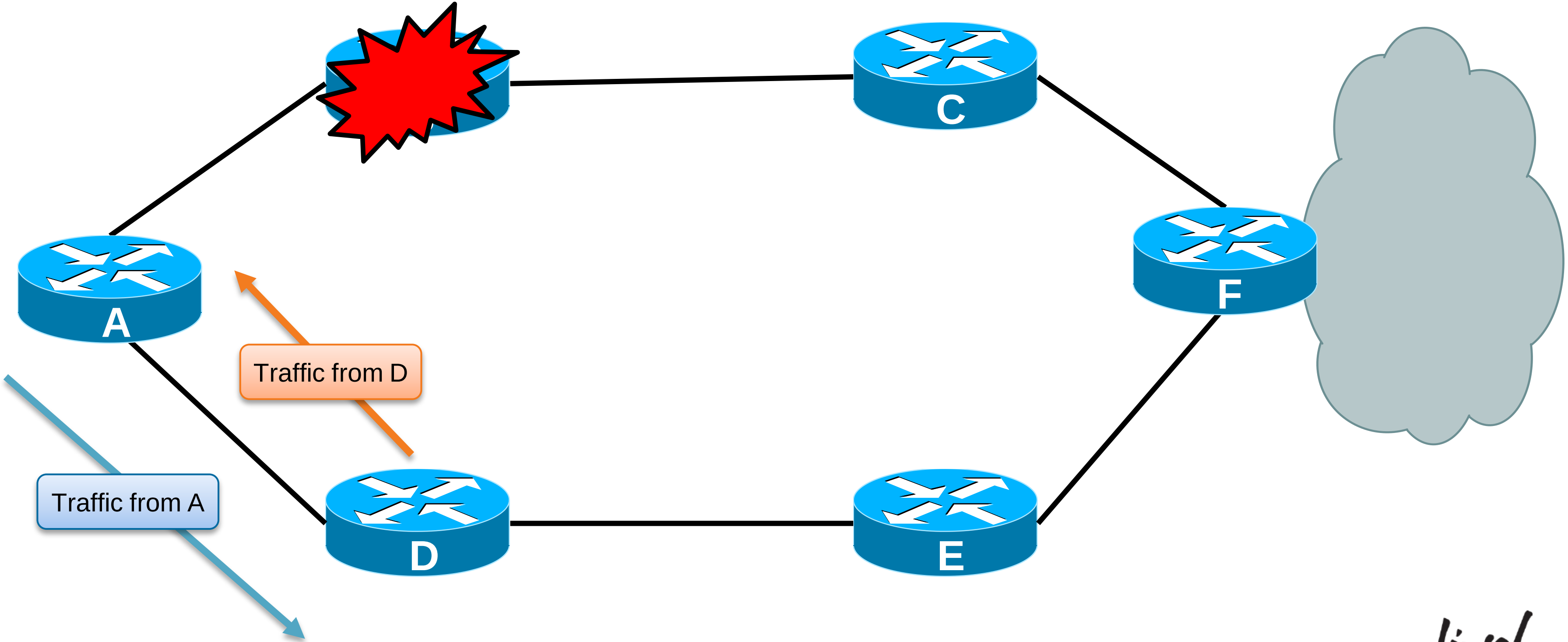
```
RouterB#show ip route 172.16.2.0
 Known via "eigrp 10", distance 90, metric 1100800, type
internal
   * 172.16.1.2, from 172.16.1.2, 00:00:17 ago, via Ethernet0/1
         Route metric is 281600, traffic share count is 1
         Repair Path: 192.168.1.1, via Ethernet0/0


RouterB#show ip cef 172.16.2.0
172.16.2.0/24
  nexthop 172.16.1.2 Ethernet0/1
    repair: attached-nexthop 192.168.1.1 Ethernet0/0
```

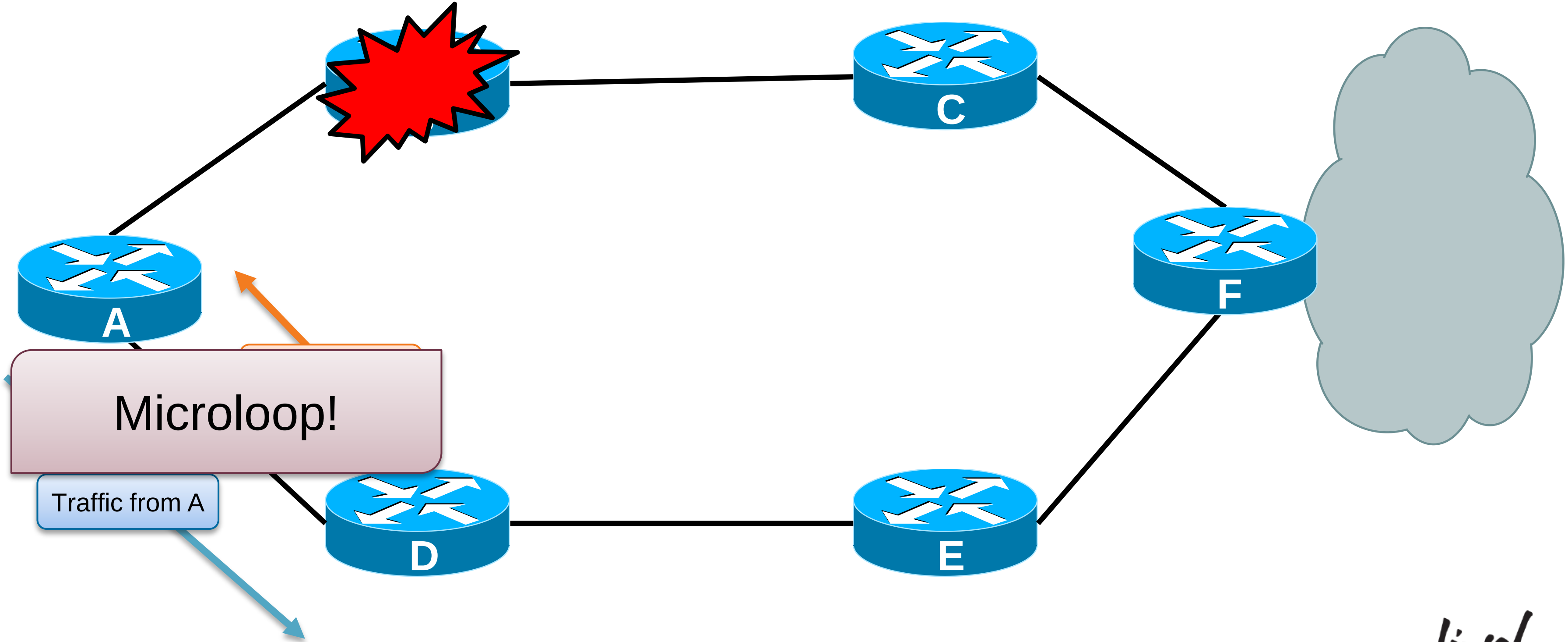
Remote LFA



Remote LFA



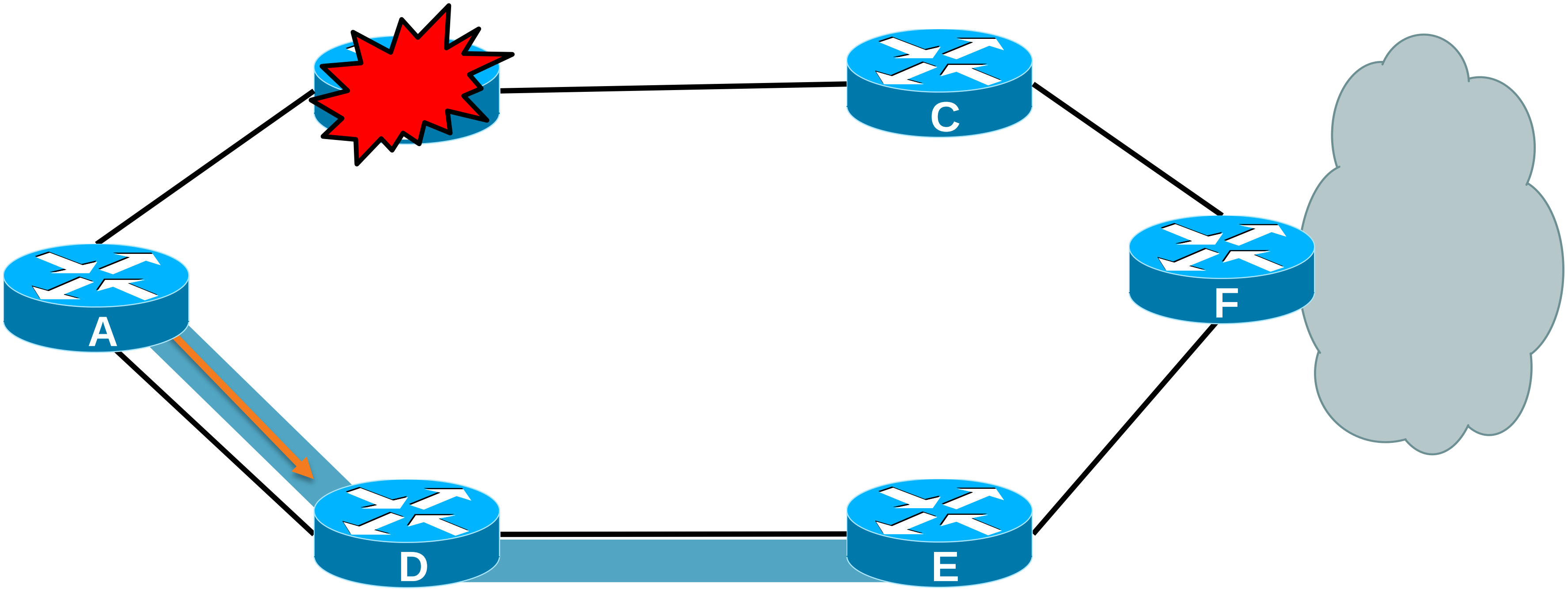
Remote LFA



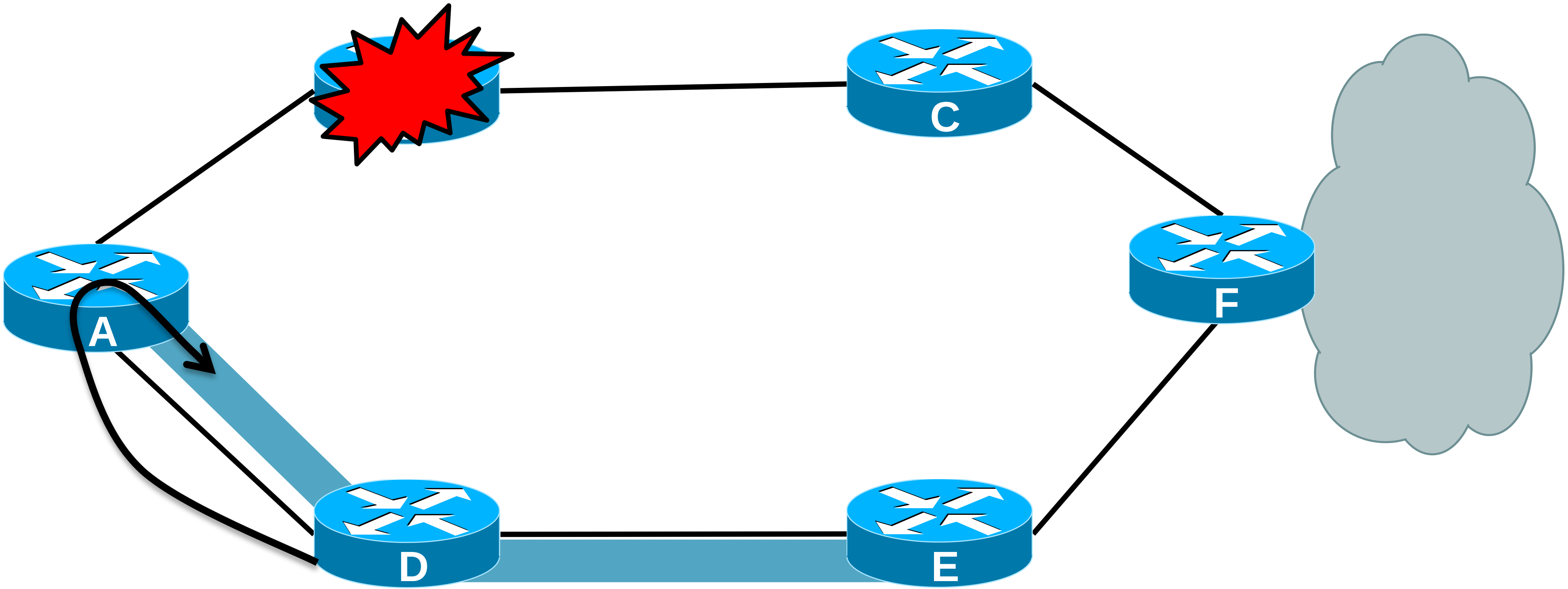
Remote LFA

- Local node runs Secondary SPF from the point of view of the remote node
- Automatic MPLS TE Fast Reroute
- Use TE Tunnel to get between local and remote nodes
- Requires LDP in the ring

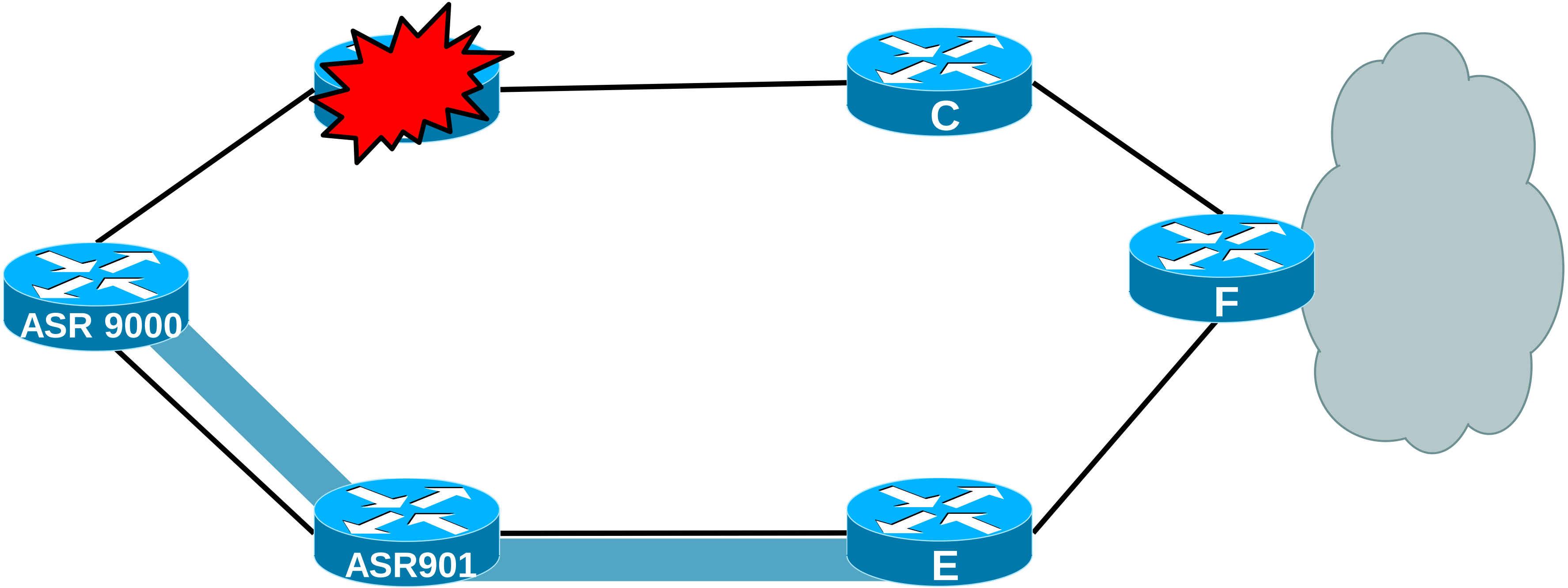
Remote LFA



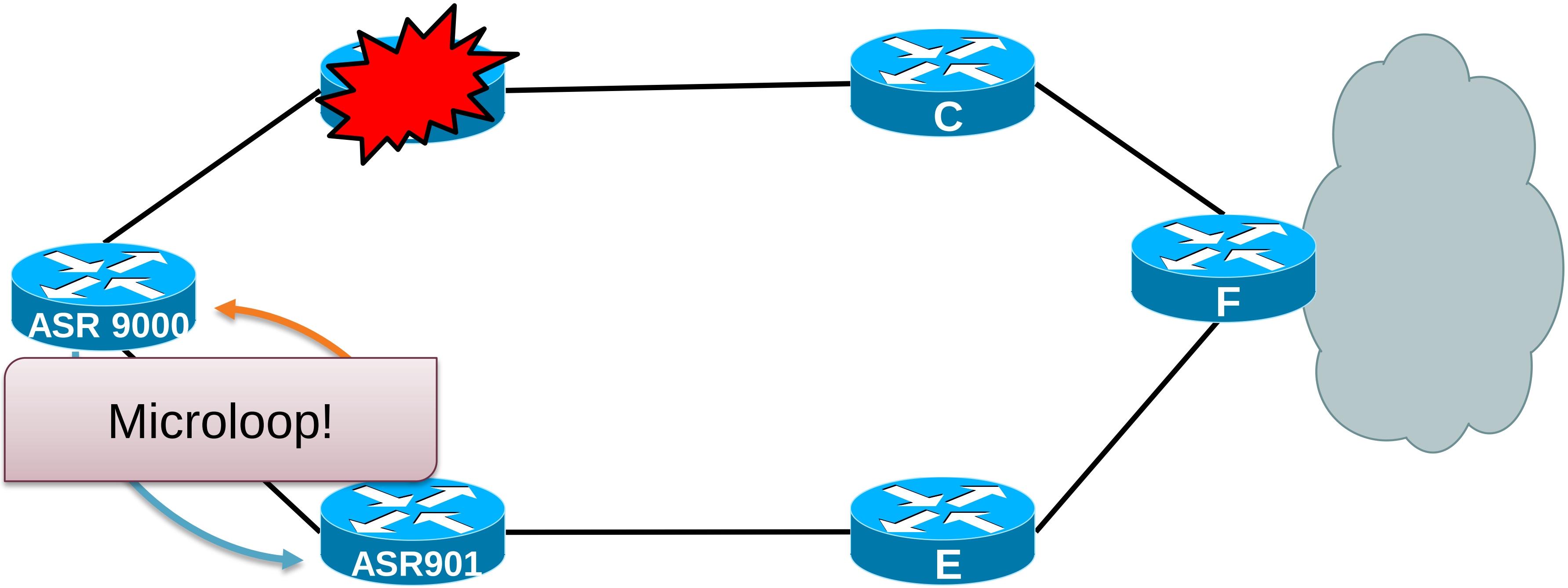
Remote LFA



Remote LFA Microloop Avoidance

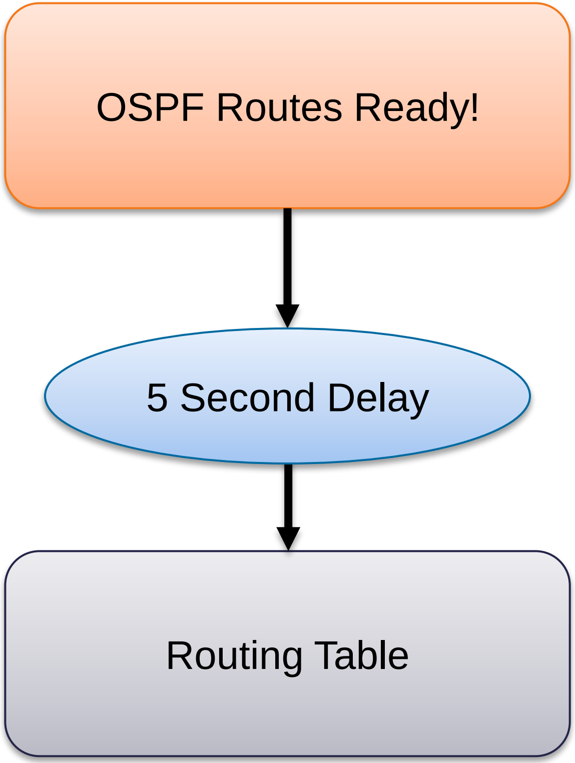


Remote LFA Microloop Avoidance



Remote LFA Microloop Avoidance

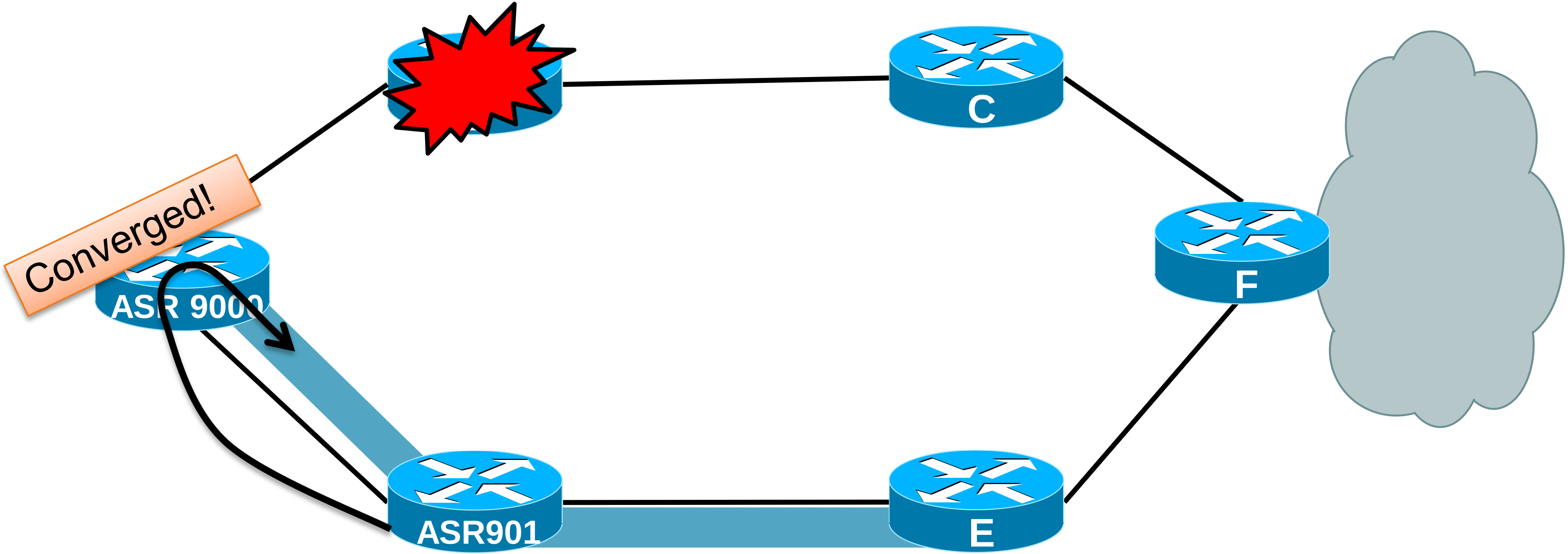
- rLFA Head End removes tunnel when converged
- Could converge faster than other nodes
- Keep Tunnel
- Delay install of routes to RIB
- Allows slower node to catch up
- Enabled automatically with rLFA


OSPF Routes Ready!

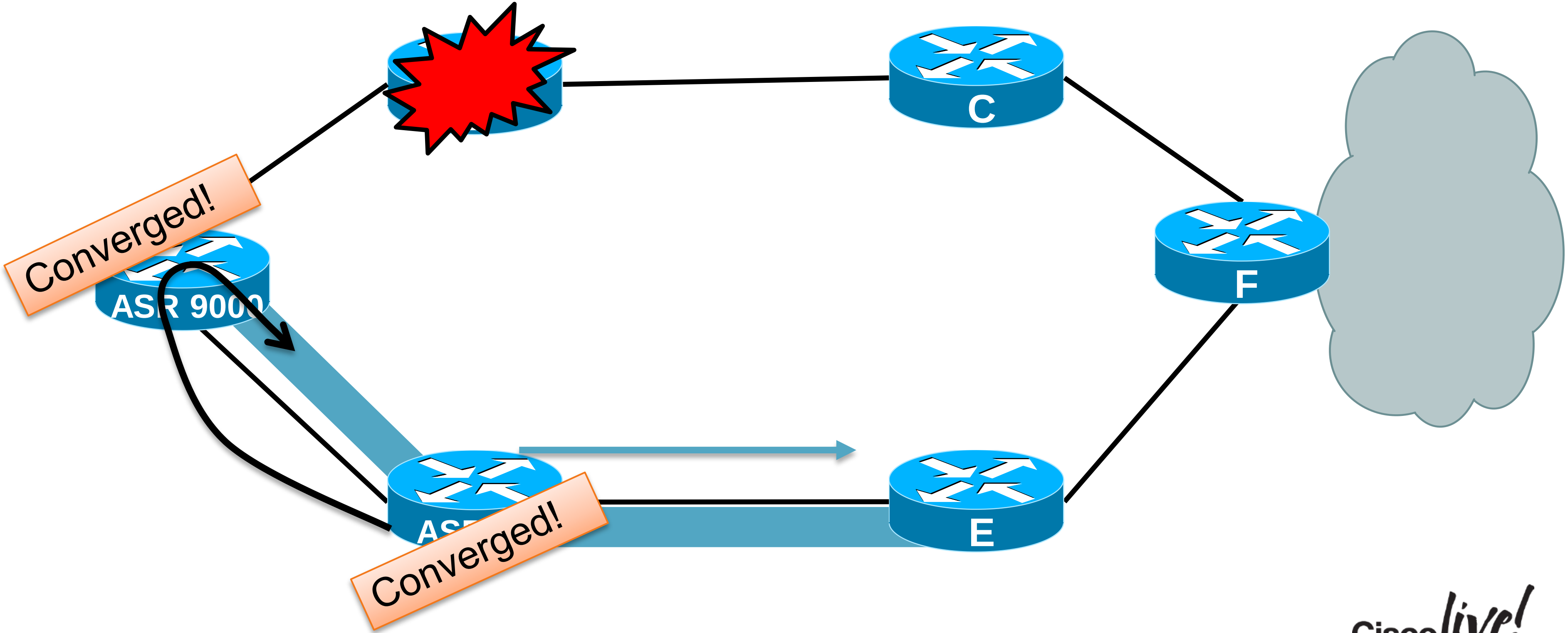
5 Second Delay

Routing Table

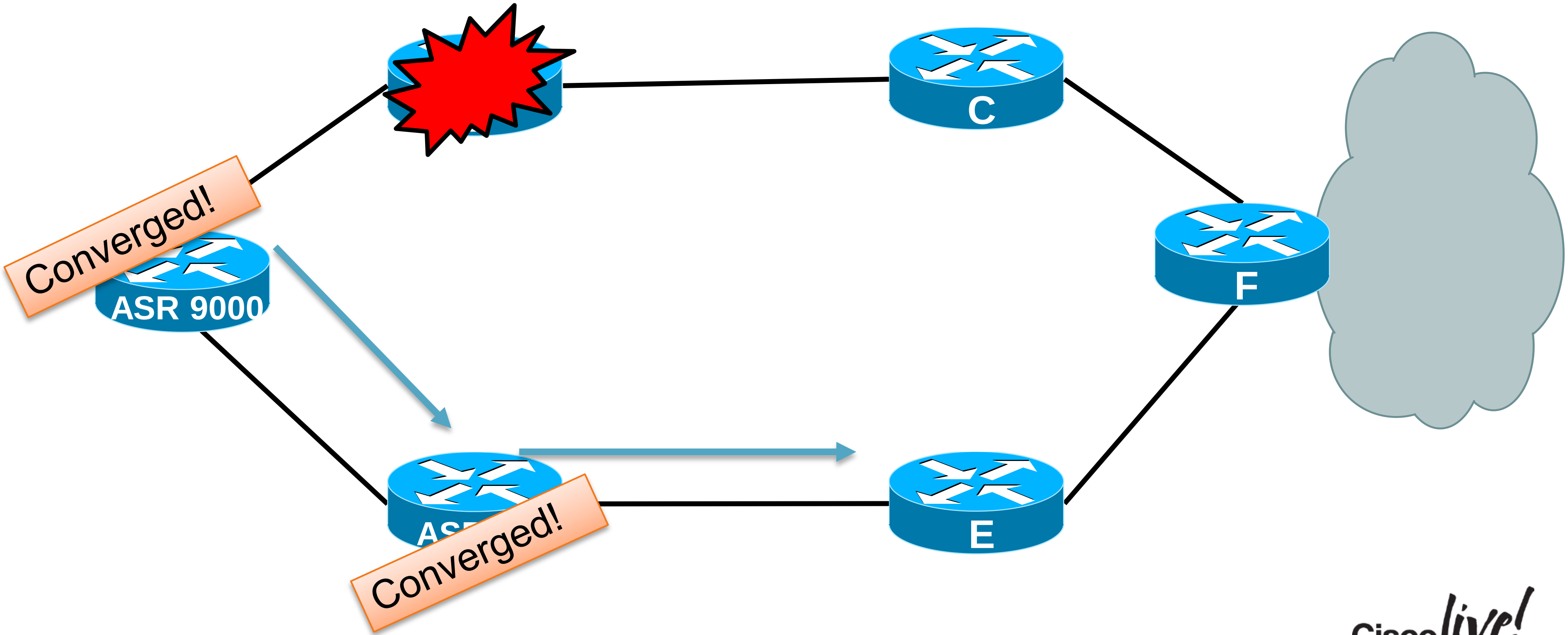
Remote LFA Microloop Avoidance



Remote LFA Microloop Avoidance



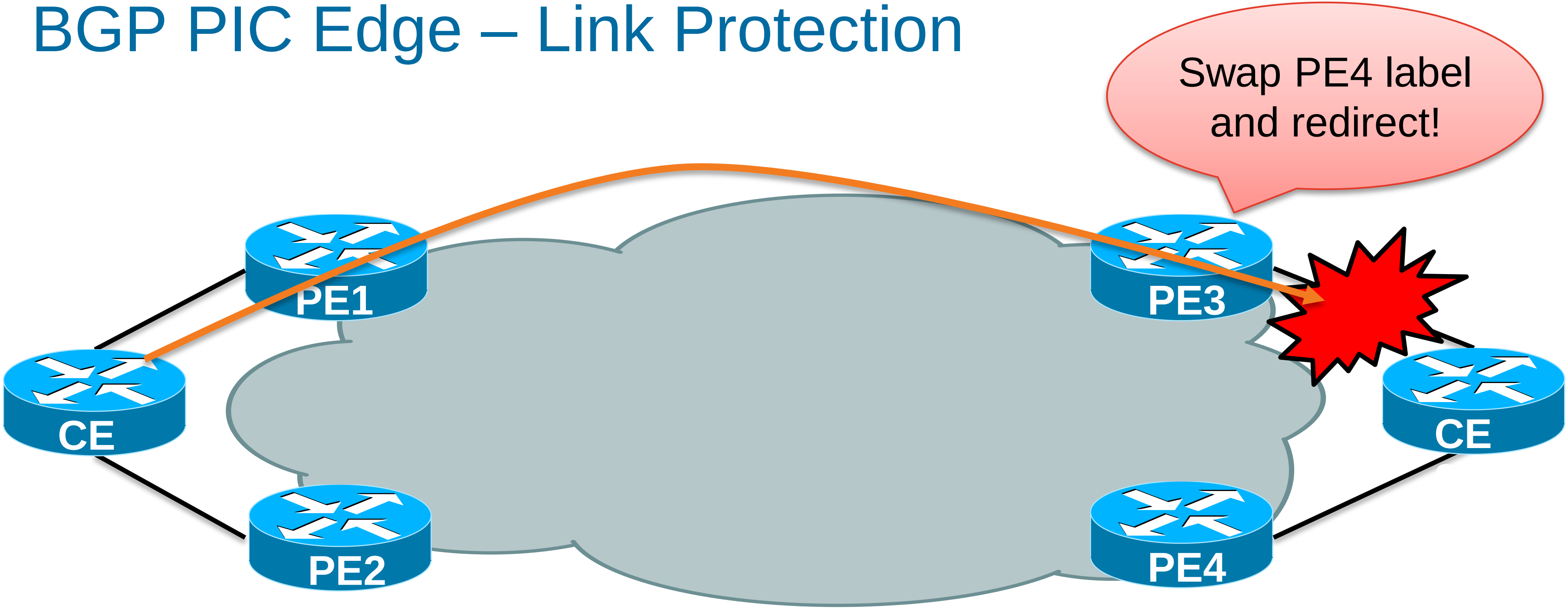
Remote LFA Microloop Avoidance



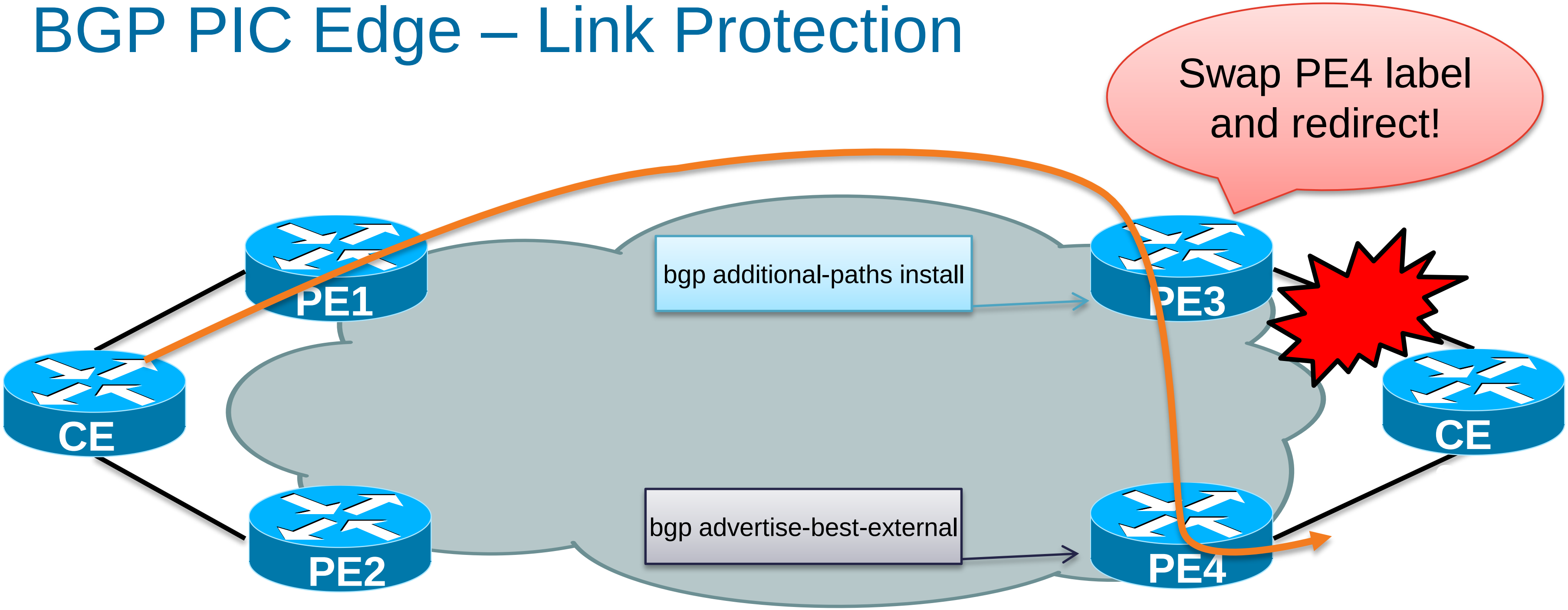
Agenda

- Thinking About Fast Convergence
- Reactive Convergence
- Proactive Convergence
  - Loop Free Alternate
  - BGP PIC Edge
  - LDP Session Protection
- Closing Remarks

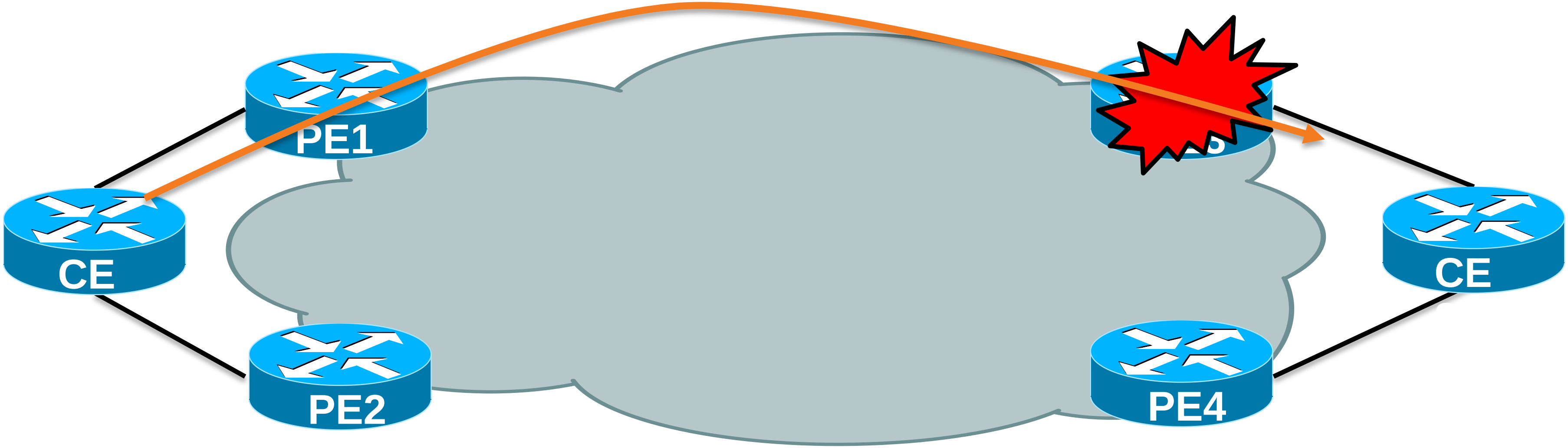
BGP PIC Edge – Link Protection



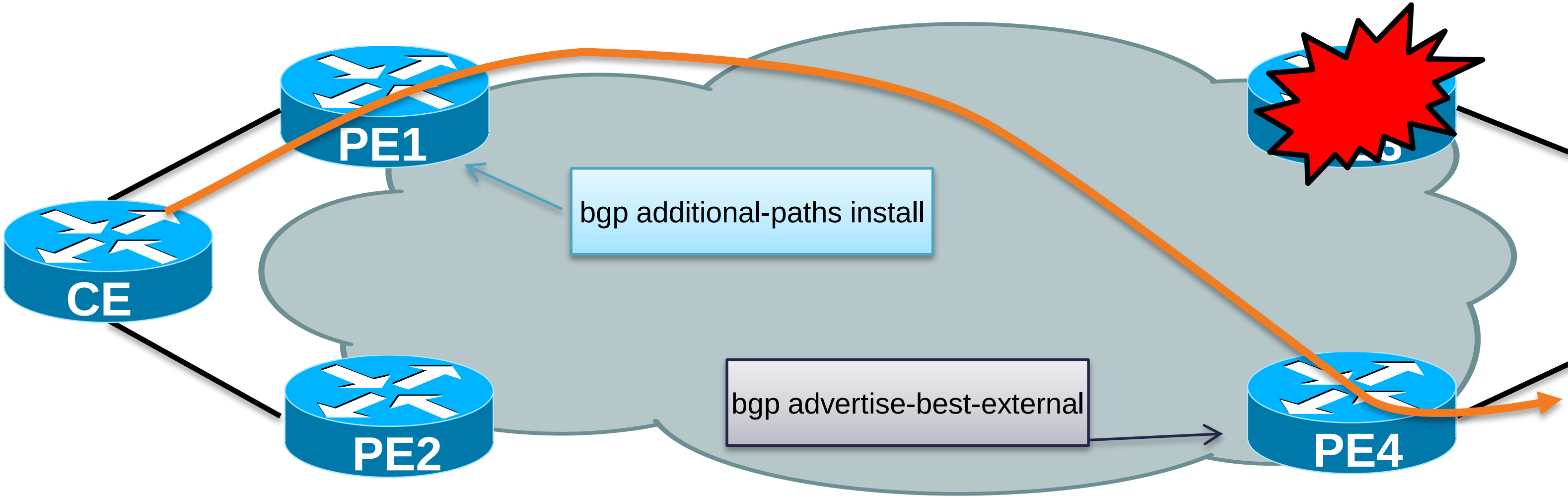
BGP PIC Edge – Link Protection


Swap PE4 label and redirect!
PE1
PE3
CE
CE
PE2
PE4
bgp additional-paths install
bgp advertise-best-external

BGP PIC Edge – Node Protection



BGP PIC Edge – Link Protection



bgp additional-paths install

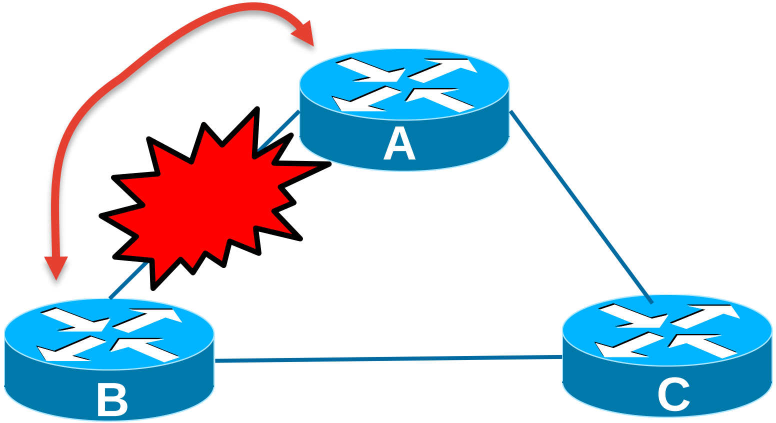
bgp advertise-best-external

Agenda

- Thinking About Fast Convergence
- Reactive Convergence
- Proactive Convergence
  - Loop Free Alternate
  - BGP PIC Edge
  - LDP Session Protection
- Closing Remarks

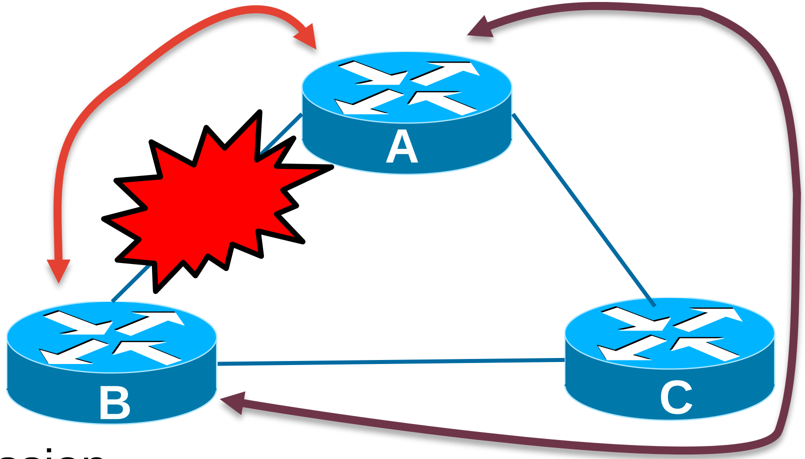
LDP Session Protection

- LDP is based on TCP
- IGP peers = LDP peers*
- Exchange Labels after IGP Convergence
  - Label per global prefix



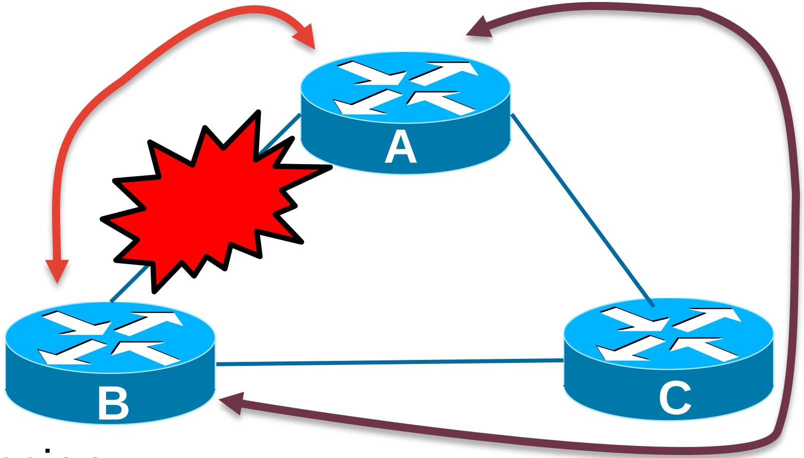
LDP Session Protection

- LDP is based on TCP
- IGP peers = LDP peers*
- Exchange Labels after IGP Convergence
  - Label per global prefix
- Link Failure requires label re-exchange
- No MPLS traffic without labels
- Session Protection creates targeted LDP Session



LDP Session Protection

- LDP is based on TCP
- IGP peers = LDP peers*
- Exchange Labels after IGP Convergence
  - Label per global prefix
- Link Failure requires label re-exchange
- No MPLS traffic without labels
- Session Protection creates targeted LDP Session
- Keep labels after failure if peer is still alive
- Immediately forward on IGP convergence



Agenda

- Thinking About Fast Convergence
- Reactive Convergence
- Proactive Convergence
- ➢ Closing Remarks

Other Considerations

- Punt Path
  - Path between interface and CPU
  - CoPP
  - Input Queue (IOS/IOS-XE)
  - General Packet Handling
    - ASR1k issues with jumbo MTU
- Neighbor Establishment Delays
  - OSPF DR / ISIS DIS
  - Use point-to-point interface
- Control Plane QoS
  - DSCP markings on egress control traffic
  - Does ingress QoS accommodate?

Final Thoughts

- Timers are just the beginning
- Everything matters
  - CPU, Hardware, Software, Latency, Operating System
- Fast Convergence is a tradeoff
- Think about both proactive and reactive convergence
- Consider network relationships and dependencies
  - Physical -> IGP -> BGP
- Culture of Engineering
  - Tolerance for false positive
  - Willing and able to work on hard problems

Recommended Sessions

- BRKARC-2350 – IOS Routing Internals
- BRKDCT-2333 – Data Center Network Failure Detection
- BRKRST-3371 – Advances in BGP
- BRKRST-3007 – Advanced Topics and Directions in Routing Protocols
- BRKARC-3472 - NX-OS Routing Architecture and Best Practices
- BRKRST-2336 (EIGRP), 2337 (OSPF), 2338 (ISIS) – Deployment in Modern Networks
- BRKRST-2041 - WAN Architectures and Design Principles
- BRKRST-2042 – Highly Available Wide Area Network Design
- BRKCRS-2031 – Enterprise Campus Design: Multilayer Architectures and Design Principles
- BRKNMS-2518 – Secrets to Achieving High Availability

Thank you.

cisco