Segment Routing
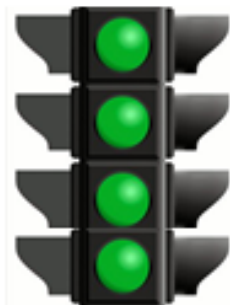
Deployment Experience and Technology Update

Clarence Filsfils

Cisco Fellow

cf@cisco.com

Deployed !

Deployment

- In CY2015, SR will be deployed in all of these markets

WEB

SP Core/Edge

SP Agg/Metro

Large Entreprise

- First deployments in 2015
- Strong start in 2016 with many new deployments

Agenda
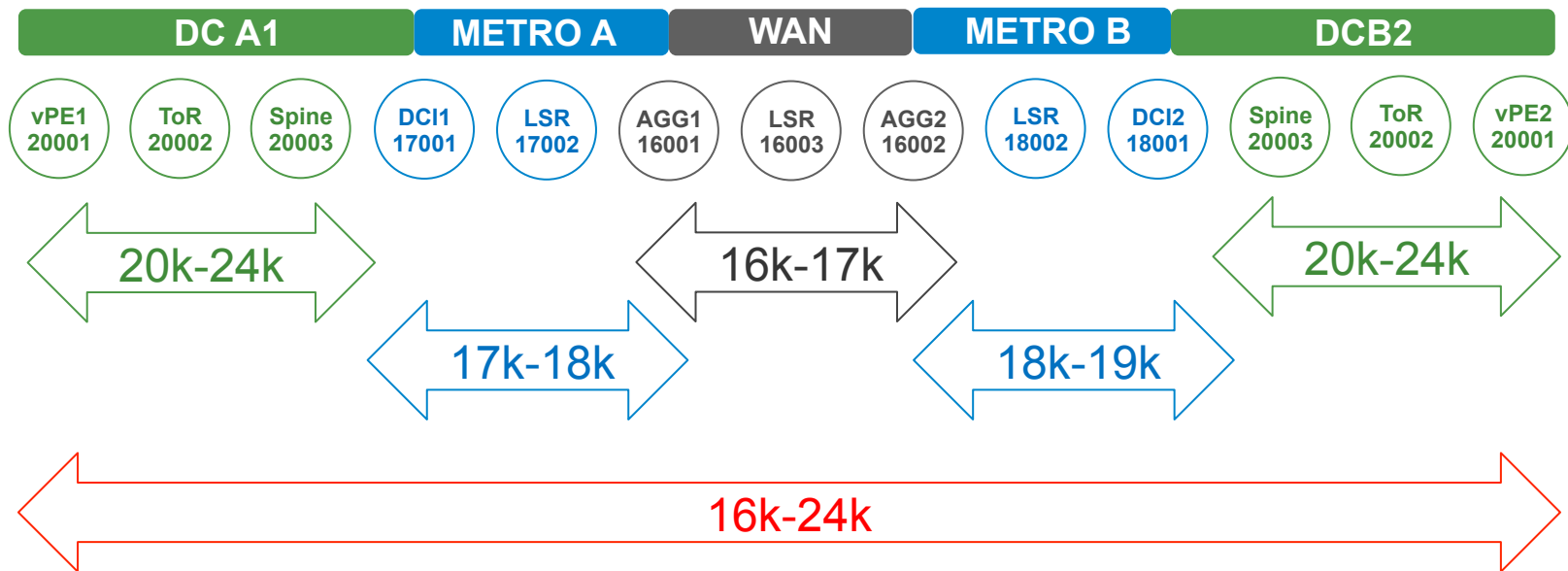
- Review advanced use-cases and related new technology

- Demo's are available at the booth

- Detailed SR tutorial
  - http://www.segment-routing.net/home/tutorial

Inter-Domain Policy at Scale

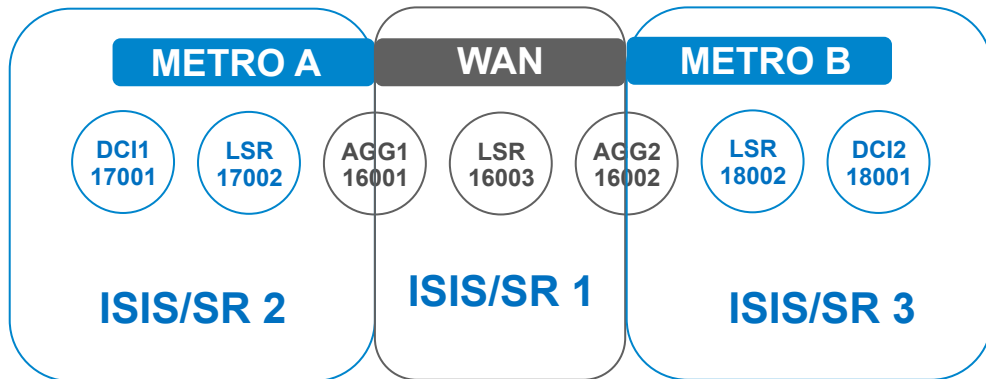draft-filsfils-spring-large-scale-interconnect
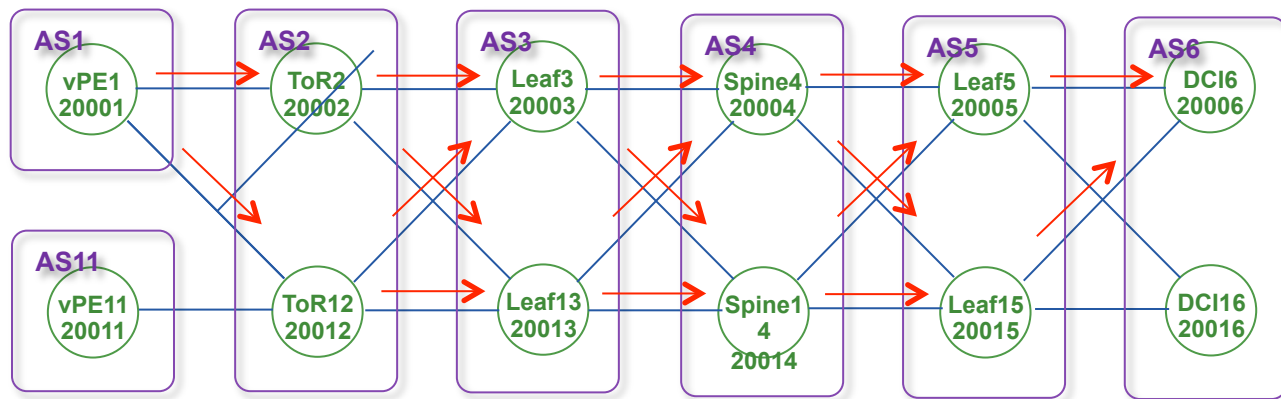
SRGB and SID allocation

• Homogenous SRGB for simplicity

IGP/SR within WAN and Metro

- Each domain runs ISIS/OSPF SR

- Technology available since June 2014

- Incremental deployment and seamless interworking with LDP

SR in the DC

- 20006 is the BGP Prefix SID to DCI6
  - ECMP, simplicity (no LDP/RSVP) and policy
- Available on Nexus/XE and NCS5k/XR

Inter-Domain Routing

DC A1			METRO A		WAN			METRO B		DCB2		
vPE1 20001	ToR 20002	Spine 20003	DCI1 17001	LSR 17002	AGG1 16001	LSR 16003	AGG2 16002	LSR 18002	DCI2 18001	Spine 20003	ToR 20002	vPE2 20001

← WAN Aggs ← WAN Aggs WAN Aggs → WAN Aggs →
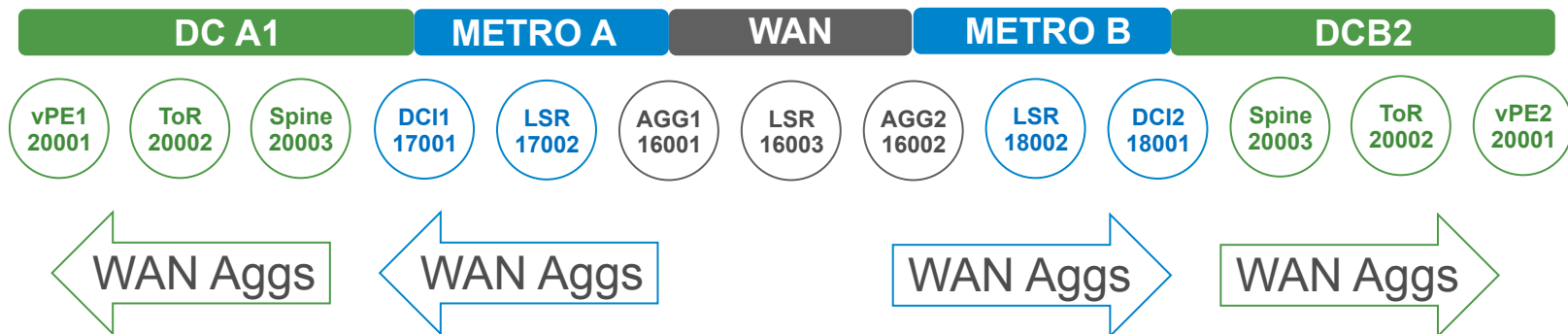
- WAN aggs are re-distributed down to Metro and DC

- Nothing is redistributed up

- How does vPE1 reaches vPE2?

SR PCE

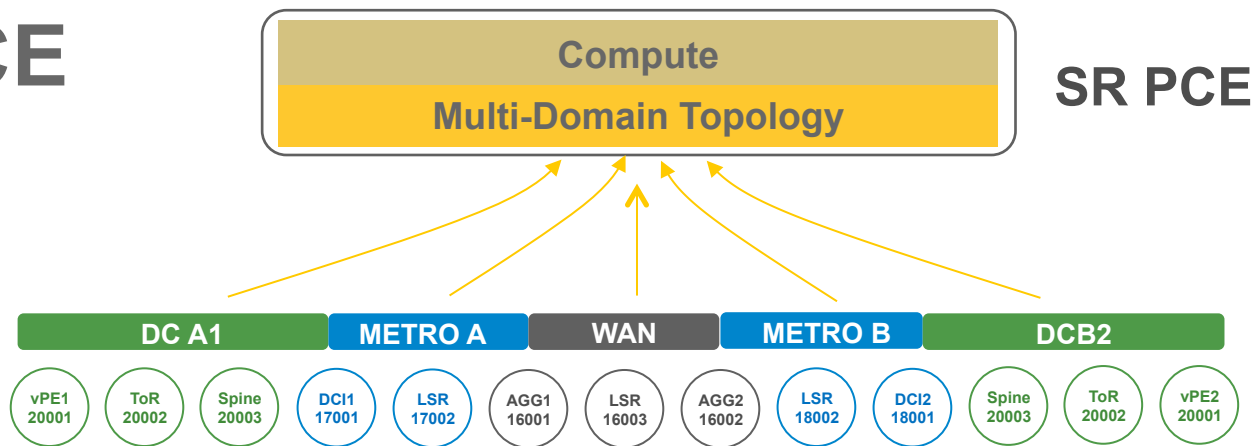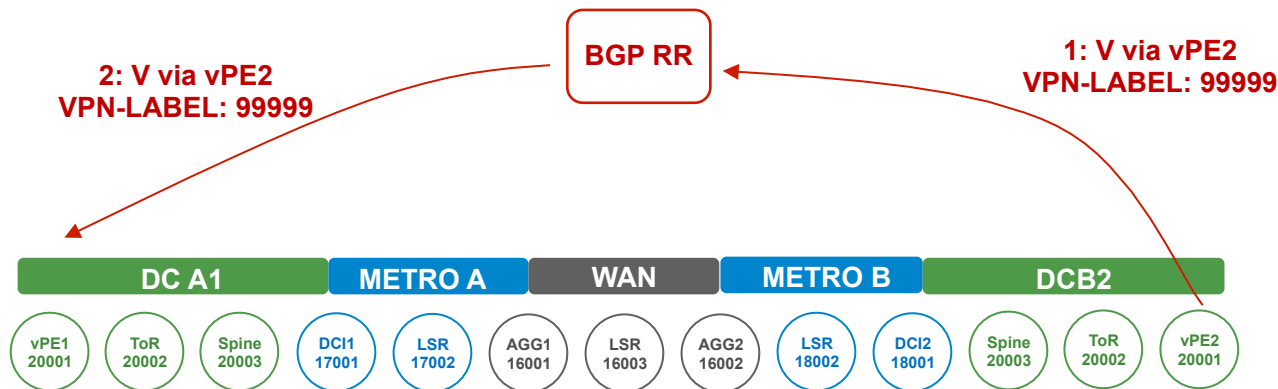- Multi-Domain topology
  - Realtime reactive feed via BGP-LS/ISIS/OSPF from multiple domains
  - Including ip address and SID
- Compute: stateful with native SRTE algorithms

Service Provisioning

2: V via vPE2
VPN-LABEL: 99999

BGP RR

1: V via vPE2
VPN-LABEL: 99999

DC A1	METRO A	WAN	METRO B	DCB2

vPE1 20001 | ToR 20002 | Spine 20003 | DCI1 17001 | LSR 17002 | AGG1 16001 | LSR 16003 | AGG2 16002 | LSR 18002 | DCI2 18001 | Spine 20003 | ToR 20002 | vPE2 20001

- vPE1 learns about a service route with nhop vPE2

- How does vPE1 reach the nhop?
  - vPE1 only has routes within DC A1 and to the AGG's of the WAN domain
  - Solution: ODN (next slide)

On-Demand SR Next-Hop
Reachability

4: {16002, 18001, 20001}

2: V via vPE2
VPN-LABEL: 99999

BGP RR

1: V via vPE2
VPN-LABEL: 99999

SR PCE

3: vPE2 ?

| DC A1 | METRO A | WAN | METRO B | DCB2 |

vPE1 20001 · ToR 20002 · Spine 20003 · DCI1 17001 · LSR 17002 · AGG1 16001 · LSR 16003 · 16002 · LSR 18002 · 18001 · Spine 20003 · ToR 20002 · vPE2 20001
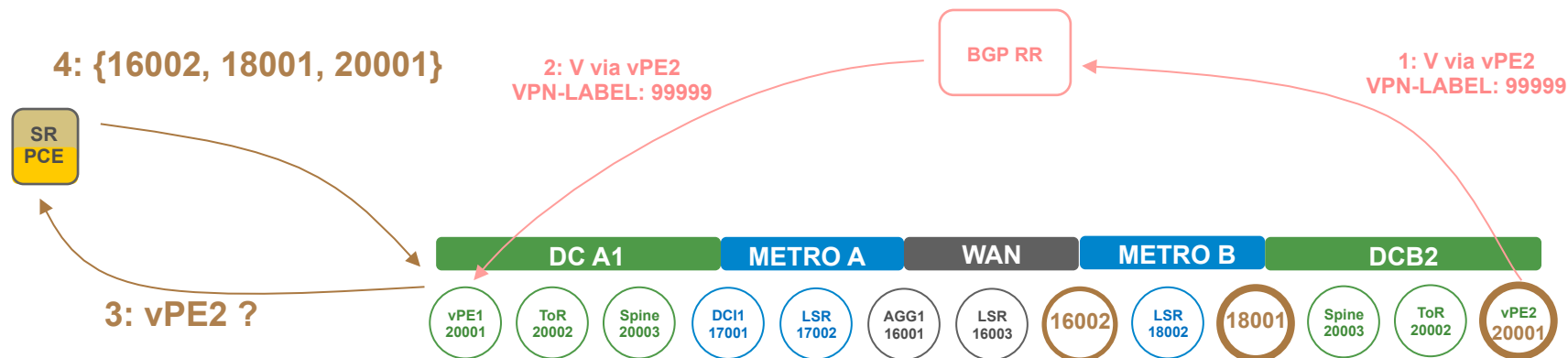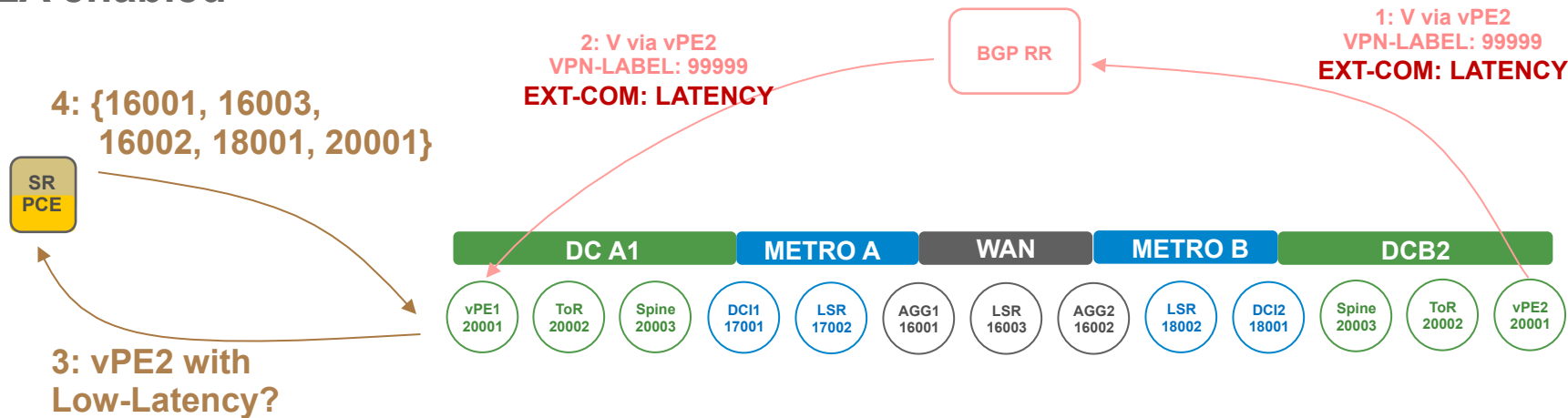
- vPE1's ODN functionality automatically request a solution from SR-PCE

- Scalable: vPE1 only gets the inter-domain paths that it needs

- Simple: no BGP3107 pushing all routes everywhere

On-Demand SR Next-Hop
SLA enabled

2: V via vPE2
VPN-LABEL: 99999
EXT-COM: LATENCY

1: V via vPE2
VPN-LABEL: 99999
EXT-COM: LATENCY

BGP RR

4: {16001, 16003,
16002, 18001, 20001}

SR PCE

3: vPE2 with
Low-Latency?

DC A1	METRO A	WAN	METRO B	DCB2

vPE1 20001 · ToR 20002 · Spine 20003 · DCI1 17001 · LSR 17002 · AGG1 16001 · LSR 16003 · AGG2 16002 · LSR 18002 · DCI2 18001 · Spine 20003 · ToR 20002 · vPE2 20001
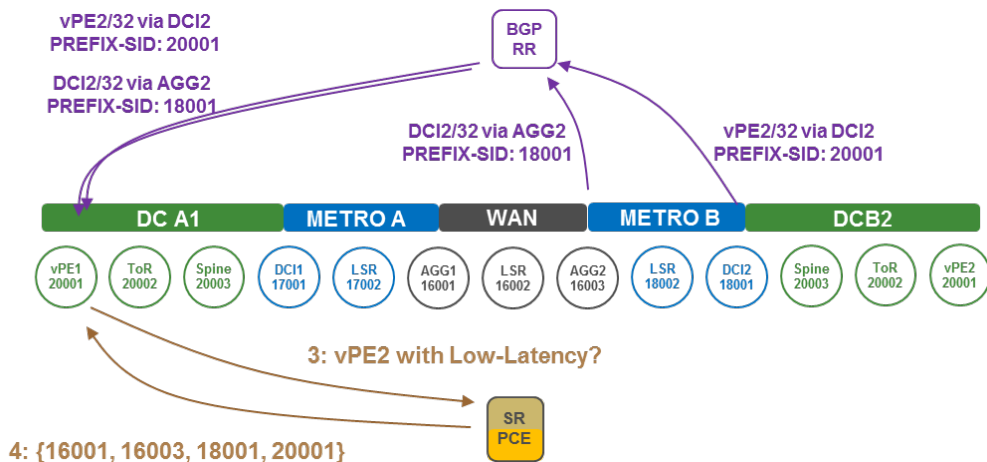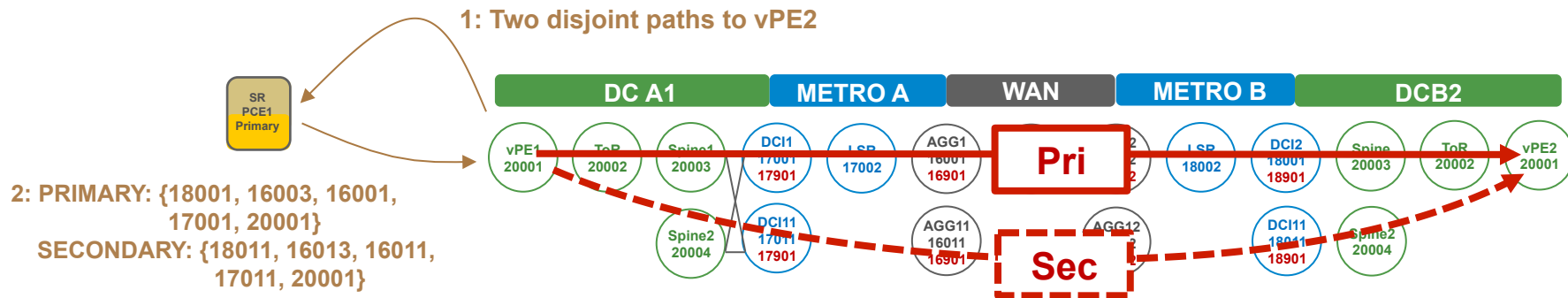
• Inter-domain SLA with scale and simplicity

– No RSVP, no midpoint state, no tunnel to configure !!
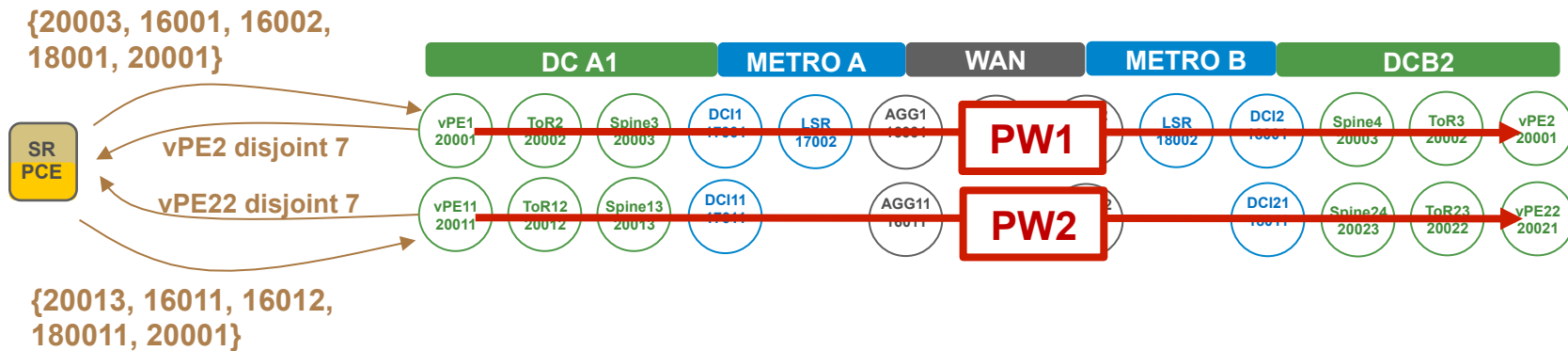
Seamless Transition

- Best-effort reachability is provided by BGP3107

- ODN and SRTE-PCE provides interdomain reachability with SLA

- Eventually, migration of more/all services over SR PCE

Inter-Domain PW - Disjoint Primary/Backup

1: Two disjoint paths to vPE2

SR PCE1 Primary

| DC A1 | METRO A | WAN | METRO B | DCB2 |

vPE1 20001 — ToR 20002 — Spine1 20003 — DCI1 17001 17901 — LSR 17002 — AGG1 16001 16901 — **Pri** — LSR 18002 — DCI2 18001 18901 — Spine 20003 — ToR 20002 — vPE2 20001

Spine2 20004 — DCI11 17011 17901 — AGG11 16011 16901 — AGG12 — DCI11 18011 18901 — Spine2 20004 — **Sec**

2: PRIMARY: {18001, 16003, 16001, 17001, 20001}
SECONDARY: {18011, 16013, 16011, 17011, 20001}

- ODN/SR-PCE automatically computes disjoint primary/sec paths for the PW

- sBFD runs at 3x50msec on each SRTE path

- Upon failure detection of the primary, the secondary SRTE Path is used

- Inter-domain SLA with scale and simplicity
  - No RSVP, no midpoint state, no tunnel to configure !!

Two Disjoint Inter-domain PW's

{20003, 16001, 16002, 18001, 20001}

vPE2 disjoint 7

vPE22 disjoint 7

{20013, 16011, 16012, 180011, 20001}

- ODN/SR-PCE automated compute disjoint paths for PW1 and PW2

- PW1 and PW2 do not share the same headend, neither the same tailend

- Inter-domain SLA with scale and simplicity
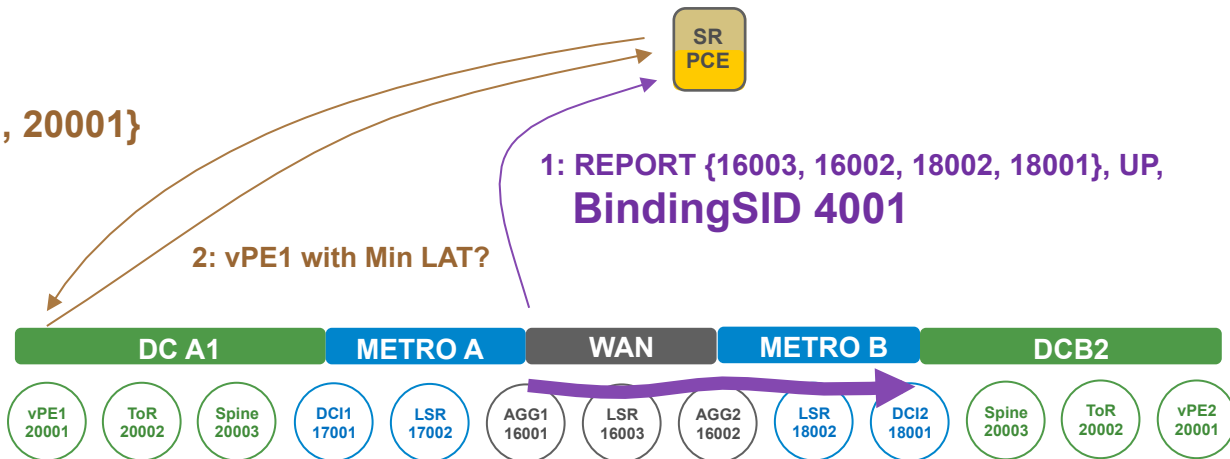 - No RSVP, no midpoint state, no tunnel to configure !!

Binding SID to stitch Policies

3: REPLY {16001, 4001, 20001}

instead of
{16001,
16003, 16002, 18002, 18001,
20001}

SR PCE

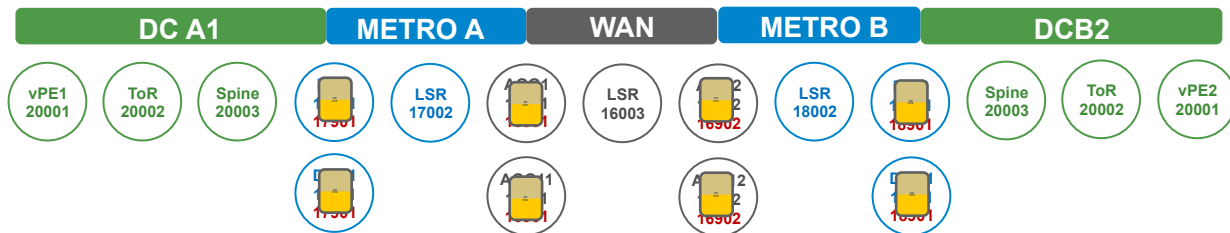**1: REPORT {16003, 16002, 18002, 18001}, UP,**
BindingSID 4001

2: vPE1 with Min LAT?

DC A1	METRO A	WAN	METRO B	DCB2

vPE1 20001 · ToR 20002 · Spine 20003 · DCI1 17001 · LSR 17002 · AGG1 16001 · LSR 16003 · AGG2 16002 · LSR 18002 · DCI2 18001 · Spine 20003 · ToR 20002 · vPE2 20001

- End-to-end policies can be composed from more basic ones
  – An SRTE policy is bound by default to a Binding SID
  – RSVP-TE tunnels can also be bound to a Binding SID and hence RSVP-TE tunnels can be used within an end-to-end SR policy

- **Shorter SID list and churn isolation between domains**
  – Even if the WAN-MetroA sub-path changes, the related Binding SID 4001 is constant
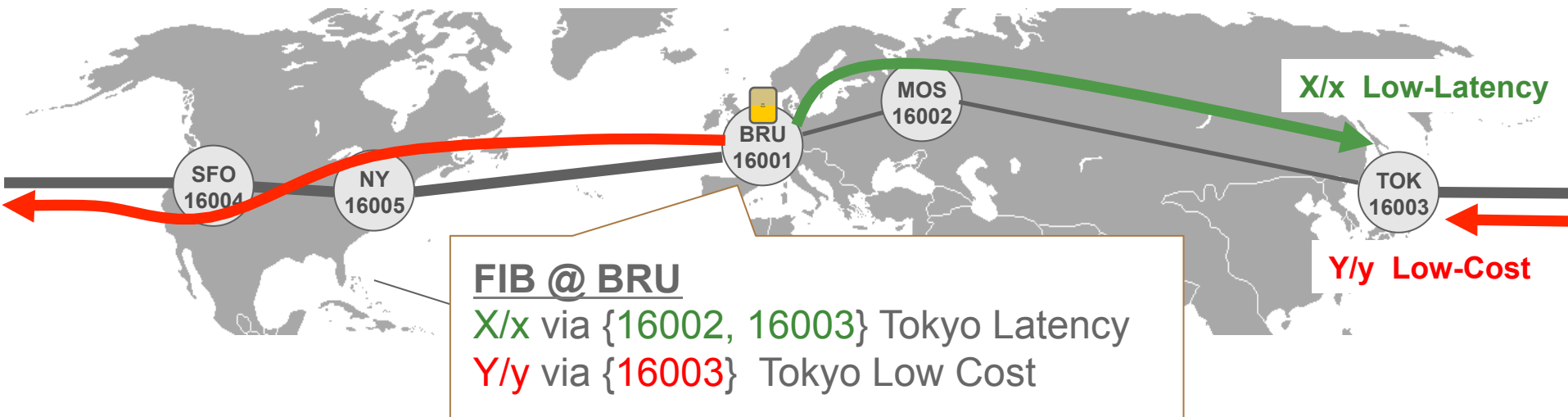
Fundamentally Distributed

DC A1	METRO A	WAN	METRO B	DCB2

- vPE1 20001
- ToR 20002
- Spine 20003
- 17901
- LSR 17002
- 1 1
- LSR 16003
- 2 16902
- LSR 18002
- 16901
- Spine 20003
- ToR 20002
- vPE2 20001

- 17901
- A 1 1
- A 2 16902
- 16901

- SR-PCE not to be considered as a single "god" box

- SR-PCE deployment model more like RR

- Different vPE's can use different pairs of SR PCE's

- SR PCE preference can either be based on proximity or service

Fully Distributed SRTE

X/x Low-Latency

FIB @ BRU
X/x via {16002, 16003} Tokyo Latency
Y/y via {16003} Tokyo Low Cost

Y/y Low-Cost

- Bru learns the routes from Tok and dynamically compute the SRTE policy to nhop
  – Elimination of RSVP signalling, midpoint states and headend tunnel configuration !
  – Commonality and Distribution: The router's SRTE functionality is same as SRTE-PCE
  – No PBR steering complexity
  – No PBR perfomance tax

Topology Independent LFA (TI-LFA)

TI-LFA - Benefits

- **50msec Protection** upon local **link**, **node** or **SRLG** failure

- **Simple** to operate and understand
  - automatically computed by the router's IGP process (ISIS and OSPF)
  - 100% coverage across any topology
  - predictable (backup = postconvergence)

- **Optimum backup path**
  - leverages the post-convergence path, planned to carry the traffic
  - avoid any intermediate flap via alternate path

- **Incremental deployment**
  - also protects LDP and IP traffic

CISCO

uLoop Avoidance

uLoop is a day-1 IP drawback
Upon link down convergence

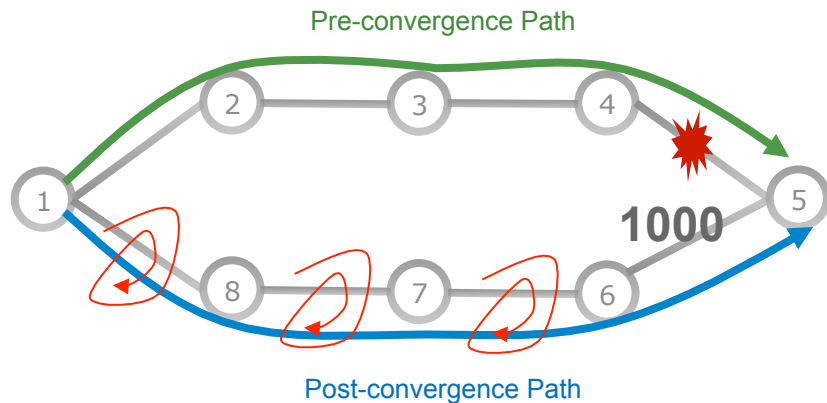Pre-convergence Path

1000

Post-convergence Path

Illustration for the post-convergence uloop impacting traffic from 1 to 5 after link45 going down. Default link metric 10

- IP hop-by-hop routing may induce uloop at any topology transition
  - Link up/down, metric up/down

SR uloop avoidance

- Prevent any uloop upon isolated convergence due to
  - link up/down event
  - metric increase/decrease event
- If multiple back-to-back convergences, fall back to native IP convergence

`microloop avoidance segment-routing`

SRv6

SRv6

- SR architecture thought since day 1 for IP

- All the control-plane benefits directly applicable
  - TILFA Link/Node/SRLG, uLoop Avoidance, SRTE, ODN, SRTE-PCE

- Data-plane requires SR extension header

- Smooth deployment plan
  - SRv6 only at a node inserting/removing/updating SRH
  - IPv6 classic at other nodes

Conclusion

Conclusion

- SR is fundamental architecture for modern IP network
- Unified Fabric with Policy through DC, Metro and WAN
- Functionality never seen before
- Simplification through Automation and protocol removal
- Strong operator endorsement
- Multi vendor consensus
- Impressive deployment and velocity

Foundation for modern IP/MPLS networking

- Simplicity

  – Unified fabric from DC to WAN including aggregation, metro and mobile backhaul

  – Set of few well-chosen building blocks

    > Lightweight extensions to core IP control-plane (BGP, ISIS, OSPF, PCEP)

    > Remove protocols (LDP, RSVP-TE)

  – Intelligent automation

    > Automated Local protection, per-prefix, any topology, sub-50msec (TI-LFA)

    > Automated uloop avoidance

    > Automated traffic matrix

    > On-Demand SR Next-Hop

Foundation for modern IP/MPLS networking

- Functionality: solves unsolved problems
  - End-to-end policies through domains internal to the SP and external: disjointness, low-latency
  - SRTE algorithm: local to router or centralized PCE
  - Local protection, per-prefix, any topology, sub-50msec (TI-LFA)
  - uLoop avoidance
  - On-Demand SR Next-Hop
  - Application controls the network without the complexity/performance-impact of PBR/DPI

- Scale
  - No SRTE midpoint state
  - Not even any SRTE headend configuration (On-Demand SR Next-Hop)
  - Binding SID for compressed SID list length

Foundation for modern IP/MPLS networking

- Seamless Deployment
  - SR/LDP interworking
  - SR/RSVP-TE interworking (binding SID)
  - Ship-in-the-night co-existence
  - SRTE policies progressively migrated from the edge to the DCI to the ToR to the vswitch to the host
  - SW upgrade (reuse existing platform)

- Decoupling data and control planes
  - Architecture natively thought to accommodate decoupled data and control planes

- Cost
  - Architecture natively thought to accommodate merchant silicon
  - Automation leads to opex saving
  - Tactical BW optimization

Industry Consensus

- Huge operator adoption
  - WEB, SP, Entreprise
  - Core, Edge, Aggregation, DC

- Standardized
  - Cisco is leading all the innovation and productization and is commited since day1 on disclosing all the elements of the architecture for IETF standardization and as well in terms of very detailed public tutorial
    > http://www.segment-routing.net/home/ietf
    > http://www.segment-routing.net/home/tutorial

- Multi-vendor consensus
  - Nokia, Ericsson, Juniper, Huawei, Arista
  - Several rounds of proven interops

- Open
  - ONOS…

Stay Informed

- http://www.segment-routing.net/

- ask-segment-routing@cisco.com

Thank you