
cisco
You make possible

Multicast and Segment Routing

IJsbrand Wijnands – Distinguished Engineer

BRKIPM-2249

cisco Live!

Barcelona | January 27-31, 2020

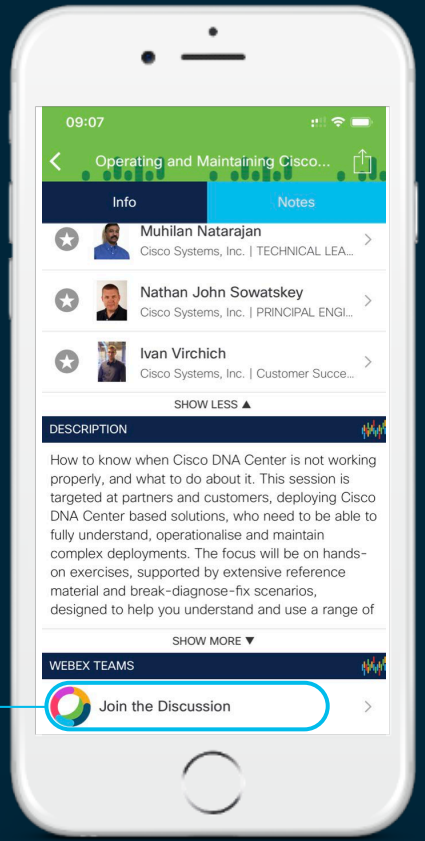
Cisco Webex Teams

Questions?

Use Cisco Webex Teams to chat with the speaker after the session

How

1. Find this session in the Cisco Events Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space



Agenda

- Introduction
- What is Segment Routing
- Traditional Multicast
- TreeSID – controller based Multicast
- BIER – Bit Indexed Explicit Replication
- Conclusion

Introduction

- Segment routing is a technology that uses Source Routing to forward packets through the network.
- A packet is forwarded from Segment to Segment based on information carried in the packet.
- Due to adding more information in the packet, less state needs to be maintained in the network and can potentially be simplified.
- What about Multicast????

SR Technology Overview

Segment Routing

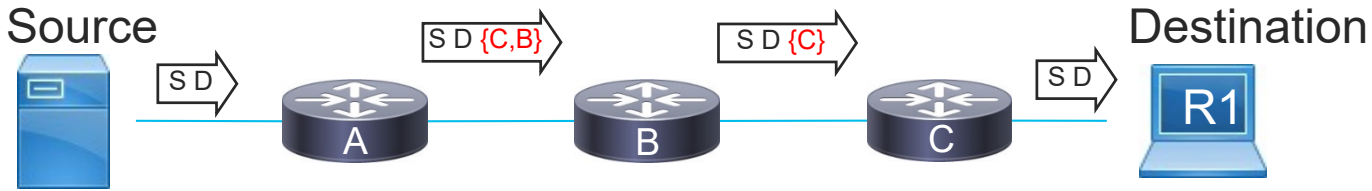
- Source Routing
  - the source chooses a path and encodes it in the packet header as an ordered list of segments
  - the rest of the network executes the encoded instructions
- Segment: an identifier for any type of instruction
  - forwarding or service

Segment Routing – Forwarding Plane

- MPLS: an ordered list of segments is represented as a stack of labels
- IPv6: an ordered list of segments is encoded in a routing extension header
- This presentation: MPLS data plane

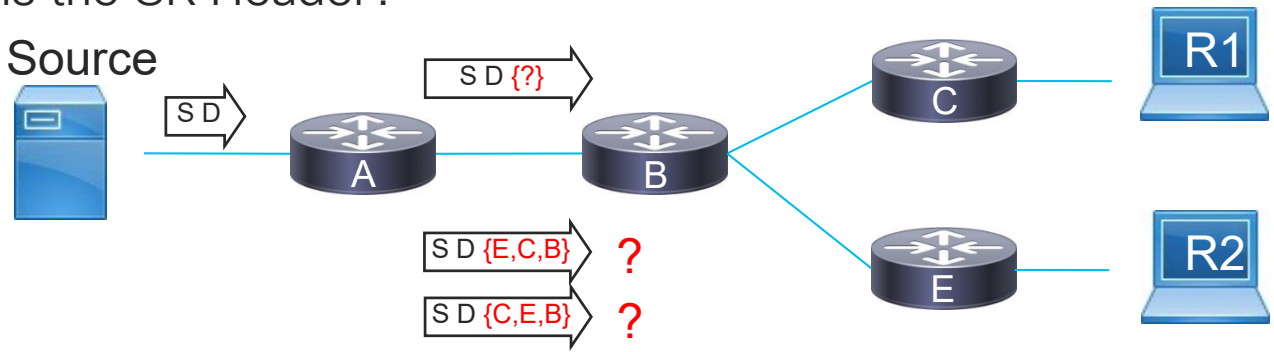
How does it work for Unicast

- The Source sends a Unicast packet to Destination R1.
- Router A determines that a SR Header has to be inserted and inserts routers B and C as the path to reach the Destination.
- The packet is Sourced routed through B and C.
- Note, the SR header { } is processed sequentially.



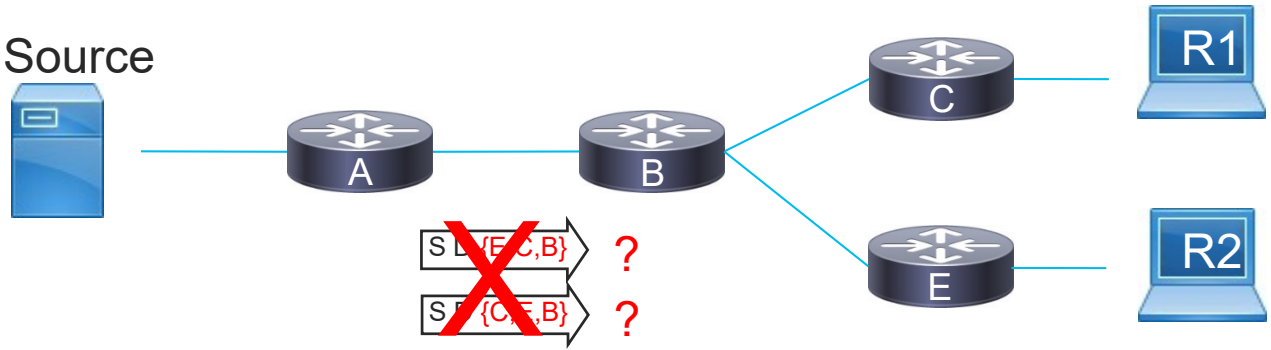
Does it work for Multicast?

- The Source sends a Multicast packet to Receivers R1 and R2.
- Router A determines that C and E are the Egress routers.
- Router A MUST only send the packet **once**, this is Multicast!...
- Router B needs to replicate the packet to C and E
- What is the SR Header?



Does it work for Multicast?

- Source Routing a packet is not possible for Multicast!
- Parsing a sequential list of hops does not allow to represent a replication.



Solutions for Multicast

- There is no exact solution for Multicast that is like Unicast SR.
- Depending on the requirements, we can choose the best fit from the following options:

1. Deploy traditional Multicast Solutions
   - PIM
   - mLDP
   - RSVP-TE
   - Ingress Replication (IR)
2. TreeSID – a controller based solution – new

   SID (System IDentifier), comes from SR lingo.
3. Bit Index Explicit replication (BIER) - new

Traditional Multicast Options

- Ingress Replication
- Multicast LDP
- P2MP RSVP-TE
- PIM

Traditional Multicast options

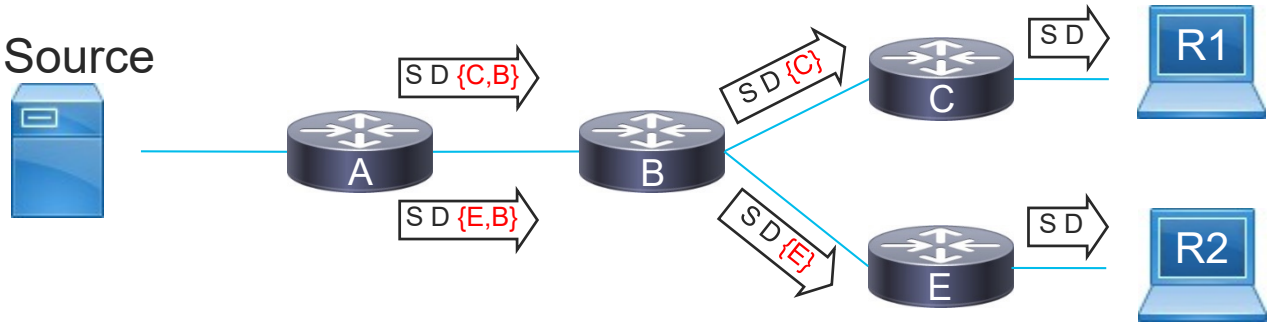
- First of all, deploying SR for unicast is **orthogonal** to what is used for Multicast.
- Nothing prevents existing protocols to continue to work, like:
  - PIM
  - mLDP
  - RSVP-TE
  - Ingress Replication (IR)
- In that sense, there is no requirement to change the Multicast deployment.
- However, if there is a technology that would benefit from being simplified and scale improved, it's Multicast ☺

Traditional Multicast Options

- Ingress Replication
- Multicast LDP
- P2MP RSVP-TE
- PIM

Ingress Replication

- The packets are replicated on the Ingress Router A for each Egress destination (C and E).
- No Multicast replication is used in the core network.
- Packets follow the Unicast SR path to each destination.



Ingress Replication

- Requires Explicit Tracking of the receivers
  - Mostly used with BGP MVPN SAFI and Leaf A-D routes
- Mostly used if the number of replications (or required bandwidth) is low.
- Scale concerns on the Ingress Router as it has to do many replications.
- For MVPN an additional VPN Label is assigned to differentiate between Unicast and Multicast traffic.
- This is a Simple supported solution, if it matches the scale requirements.

Traditional Multicast Options

- Ingress Replication
- Multicast LDP
- P2MP RSVP-TE
- PIM

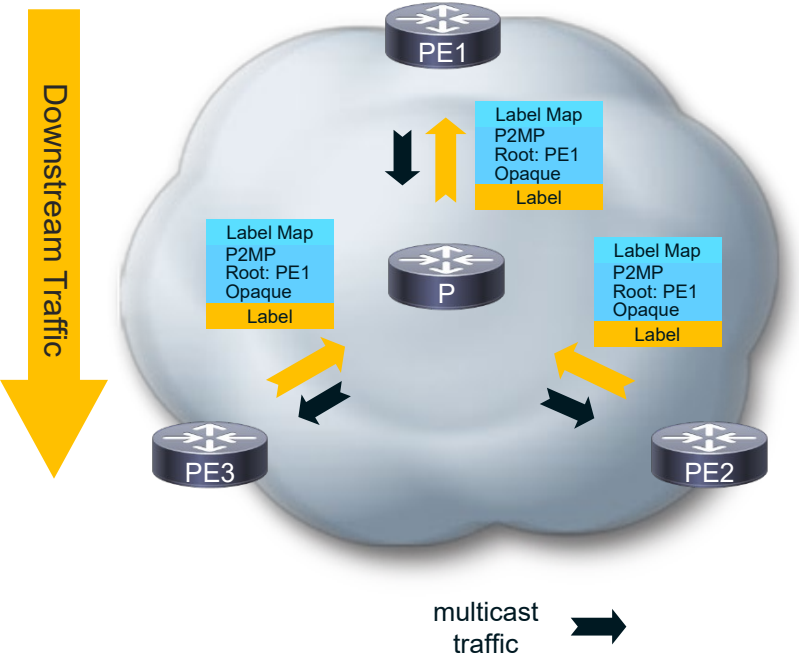
mLDP

- mLDP is a protocol that builds
  - P2MP LSPs
  - MP2MP LSPs
- Very often and **preferred** deployed for Multicast VPNs
- It's a receiver driven tree building protocol like PIM.
- mLDP uses the LDP Transport to exchange Label Mappings.
- It's much simpler compared to PIM and RSVP-TE.
  (Mostly due to being stateful and not periodic)
- Supports Link protection through SR-TE, RSVP-TE or Loop Free Alternate (LFA).

mLDP Signaling and Packet forwarding

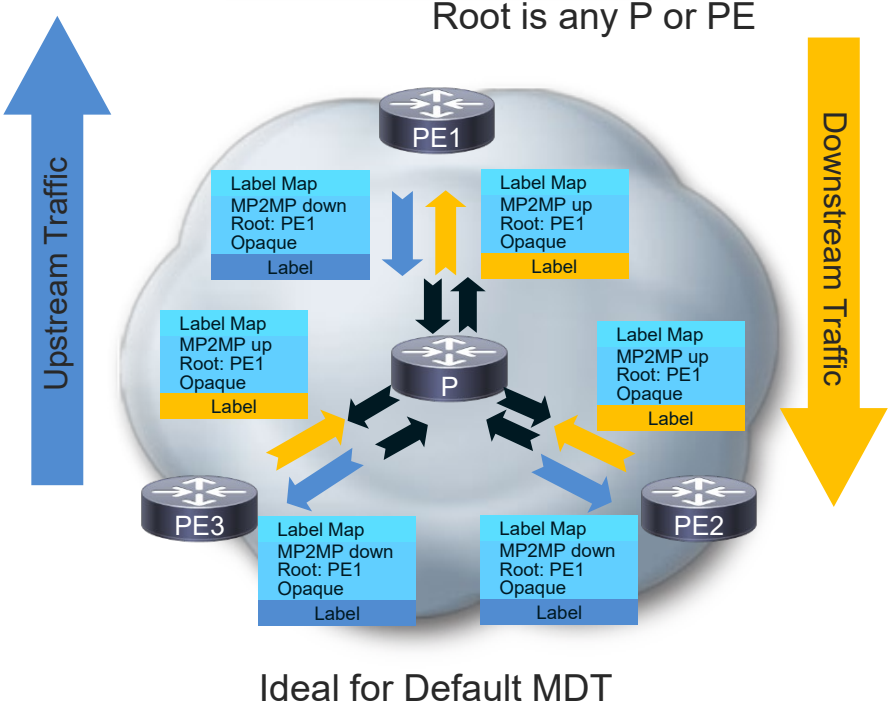
P2MP Tree

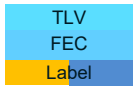
Root is ingress PE



MP2MP Tree

Root is any P or PE





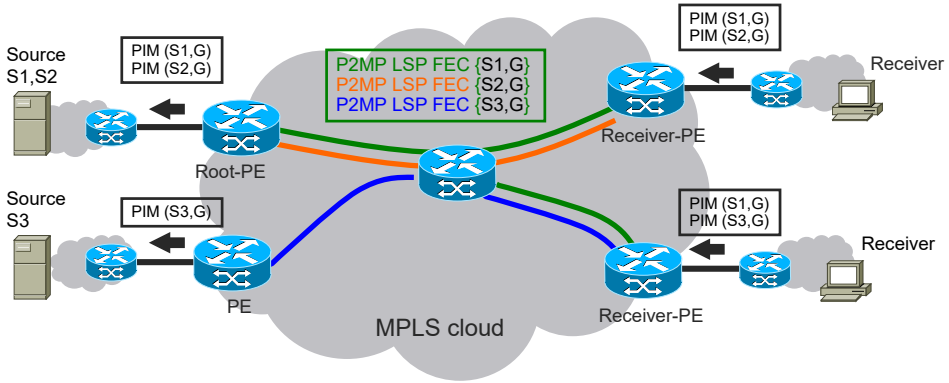
Ideal for Default MDT
Data plane is still P2MP

mLDP signaling – mapping to an LSP

- mLDP signaling is used to build a MP-LSP through the network.
- There are two mechanisms to assign an IP Mulitcast flows to the MP-LSP.

1. Overlay Signaling
   - The mapping between the flows and the MP-LSP is carried through a second protocol.
   - This can be BGP, PIM or Static.
2. In-Band signaling
   - The mapping between the flow and the MP-LSP is carried inside mLDP

mLDP In-band signaling



- PIM (S,G) tree is mapped to a mLDP P2MP LSP.
- Root PE is learned via BGP Next-Hop of the Source address.
- Receiver-PE may use SSM Mapping if Receiver is not SSM aware.

mLDP In-band signaling

- Very useful for IPTV deployments.
- Works with PIM SSM and PIM SM shared trees.
- SSM Mapping may be deployed to convert to SSM.
- One-2-One mapping between PIM tree and mLDP LSP.
- No flooding/wasting of bandwidth.
- Works well if the amount of state is bound.
- Supported in the VRF context.

mLDP and SR

- One of the benefits of SR is that LDP can be removed from the network.
- Does that mean mLDP can't run? Yes ☐ No ☑
- When rfc7473 is supported, LDP can be configured such that Labels are not exchanged for Unicast prefixes.
- mLDP will continue to use the Transport service that LDP provides.
- LDP will not participate in any Unicast forwarding.
- Its really no problem to continue to use/deploy mLDP.

Traditional Multicast Options

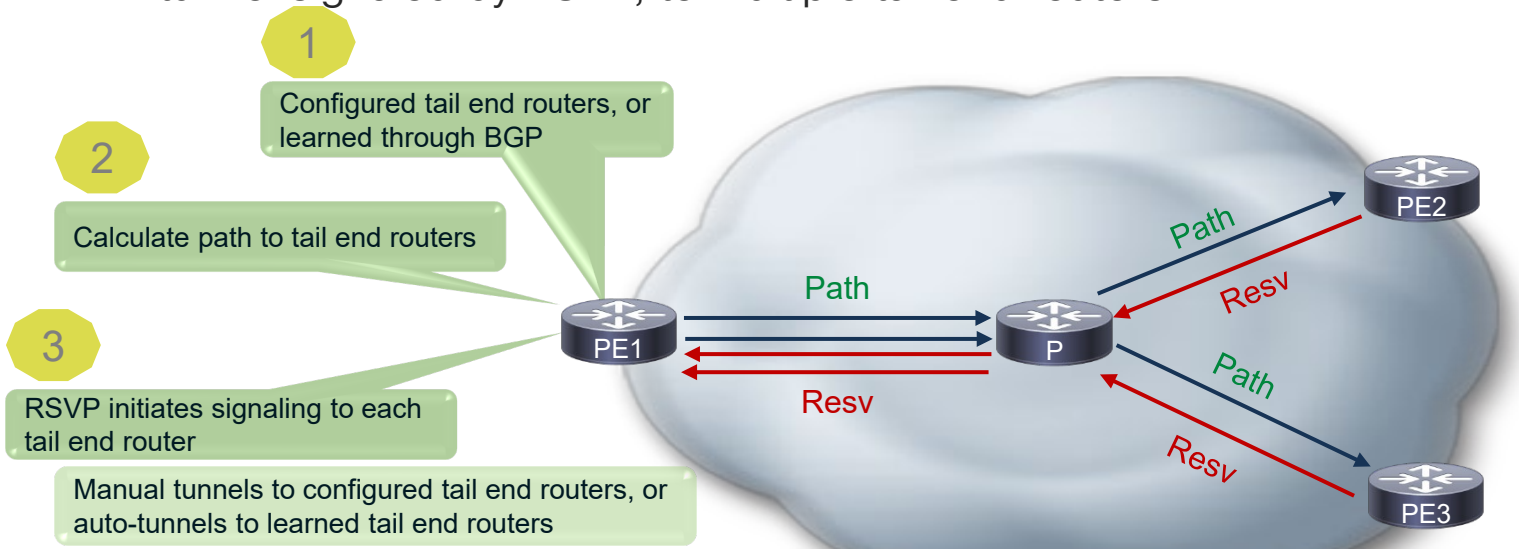
- Ingress Replication
- Multicast LDP
- P2MP RSVP-TE
- PIM

P2MP TE RSVP-TE

- Explicit (source) routing
- Bandwidth reservation
- Fast ReRoute (FRR) protection
- Uses RSVP for TE
- P2MP: extensions for RSVP-TE and IGP
- P2MP TE: looks and feels like P2P TE
- Replication of mcasticast on the core routers

P2MP TE RSVP-TE

- P2MP tunnel signaled by RSVP, to multiple tail end routers

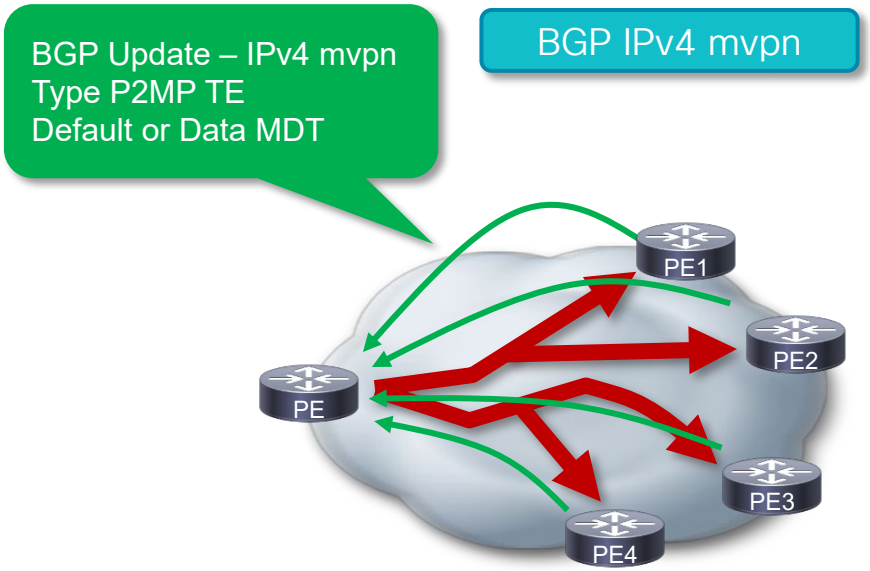


P2MP TE RSVP-TE

- Configure or discover the tail end routers of P2MP TE tunnels

configure/static

discover/auto-tunnel



P2MP TE RSVP-TE

- Traffic engineering on a larger scale is challenging with RSVP-TE.
- The number of sub-LSPs that need to be created for each destination can get very high.
  - This is especially true in MVPN deployments where there is a full-mesh requirement.
- One of the key benefits of SR is the traffic engineering scale benefits!
- Customers that are interest in using SR for traffic engineering generally want to move away from RSVP-TE.

Traditional Multicast Options

- Ingress Replication
- Multicast LDP
- P2MP RSVP-TE
- PIM

PIM

- There is really nothing to say about PIM that is specific to SR.
- The general trend is to move away from PIM.
- It's a complicated protocol to operate and troubleshoot.
- PIM is being pushed more and more to the edge.

TreeSID - new
A controller based approach to building a Tree

Tree Segment Identifier (TreeSID)

- TreeSID is an SDN controller based approach to building P2MP trees.
- Due to the controller, the tree can be built using any constraint (like P2MP RSVP-TE).
- A TreeSID identifier can be:
  - IPv4/IPv6 Source and Group (S,G)
  - A name string.
  - A numeric value.
- In this presentation we'll focus on using MPLS.

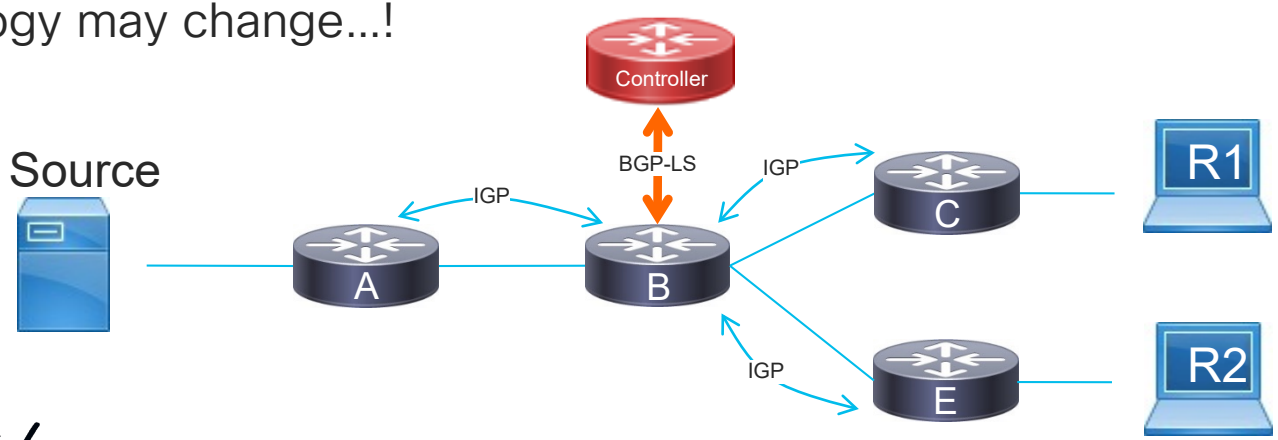
TreeSID

- In order for the controller to build the Tree it needs to:

1. Know the topology.
2. Know the Root and Leaf's of the Tree.
3. Know the MPLS Labels it can use.
4. Have a mechanism to program the Forwarding state.

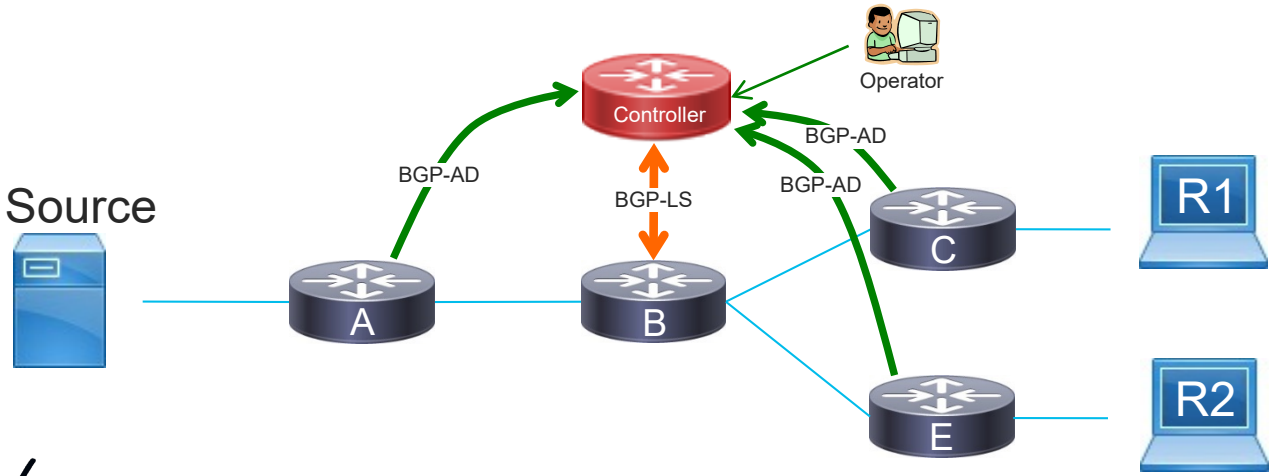
The controller – Learning the topology

- A common mechanism to learn the topology is using BGP Link State (LS).
- Through BGP-LS, the controller sucks up the Link State database.
- Through the LS database, the controller can use any sort of algorithm (like Dijkstra) to calculate paths.
- Topology may change…!



The controller – Learning the Tree

- The controller also needs to know the Tree Root and End-points.
- This can be defined by an operator.
- Dynamically through a protocol, like BGP Auto Discovery (AD)



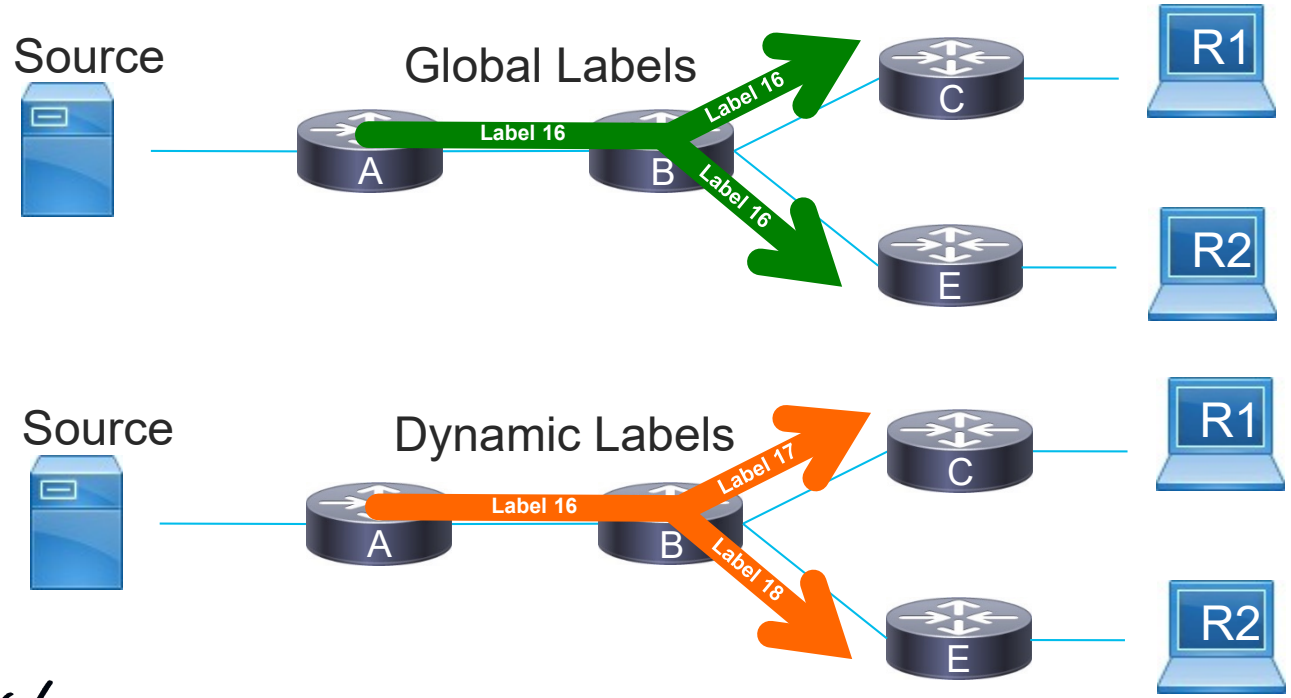
The controller – MPLS Label Allocation

- MPLS Labels are platform specific by architecture.
- That means each router needs to allocate/reserve Labels for TreeSID.
- The allocation and programming of the Label for each TreeSID is done by the Controller, that means the controller needs to know the labels that it can use for programming TreeSID on each router.
- We export a router's label range to the controller.
  - Static configuration.
  - BGP-LS.
  - Label range can come from the SR Local Block (SRLB).

The controller – Global vs Dynamic labels

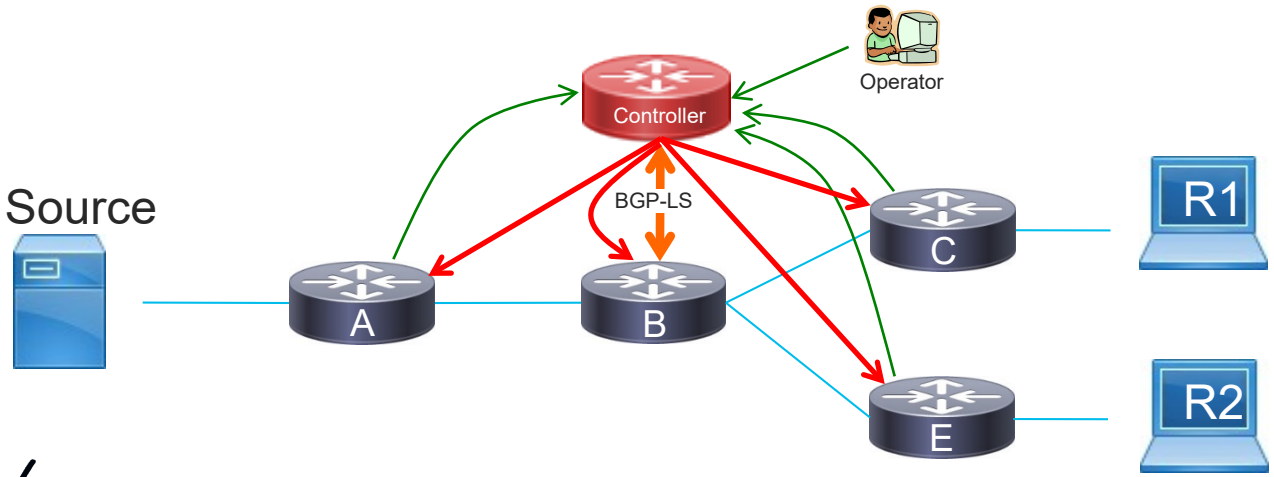
- One of the benefits of Segment Routing is that the MPLS Label assigned to a Segment can be Global.
  - This means its well known and predictable.
  - It makes it easier to manage and troubleshoot the network.
- For TreeSID, the entire Tree can be seen as a Segment.
  - All the routers in the network allocated the same Label range for TreeSID.
  - The controller assigns the same Label for a Tree on all the routers.
  - This has consequences for convergence, no local repair.
- Each path (leg) of the Tree may get a dynamic label
  - Allocated by the controller from the routers MPLS range.

The controller – MPLS Label Assignment



The controller - programming

- When a controller is responsible for creating the Tree, it needs to program forwarding state on all the routers in the path of the Tree.
- This also means the controller is responsible to determine the ingress and egress mapping (IP to/from MPLS)



The controller – programming

There are various mechanisms to have the controller program the Tree.

- Netconf/YANG
- Restconf APIs (XML, JSON)
- Path Computation Element Protocol (PCEP)
- BGP

The controller – Programming Example

A. A controller leans the SRLB Label range from all routers through BGP-LS or BGP-AD.

B. The controller learns the Topology through BGP-LS.

C. The controller learns the Root and Leafs from the operator or through BGP-AD.

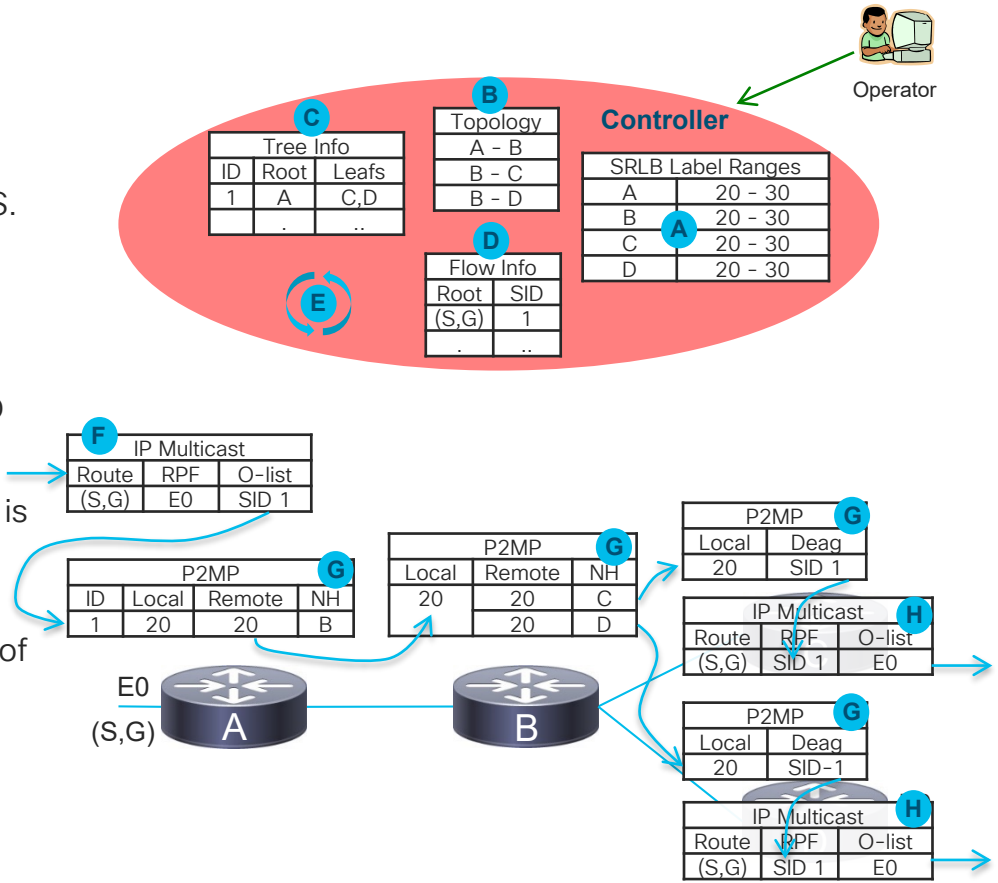
D. The controller learns the Flow to TreeSID mapping from BGP-AD or Operator. Note, we use a TreeSID to map the flow to the TreeSID.

E. With the Topology and the Tree Info, the controller is able to calculate the router forwarding tables. The controller uses Label from the TreeSID range.

F. Imposition entry, maps (S,G) to TreeSID by means of an SID (1).

G. MPLS Forwarding Table for P2MP.

H. Disposition entry, uses the SID to do RPF check.



Tree SID - Convergence

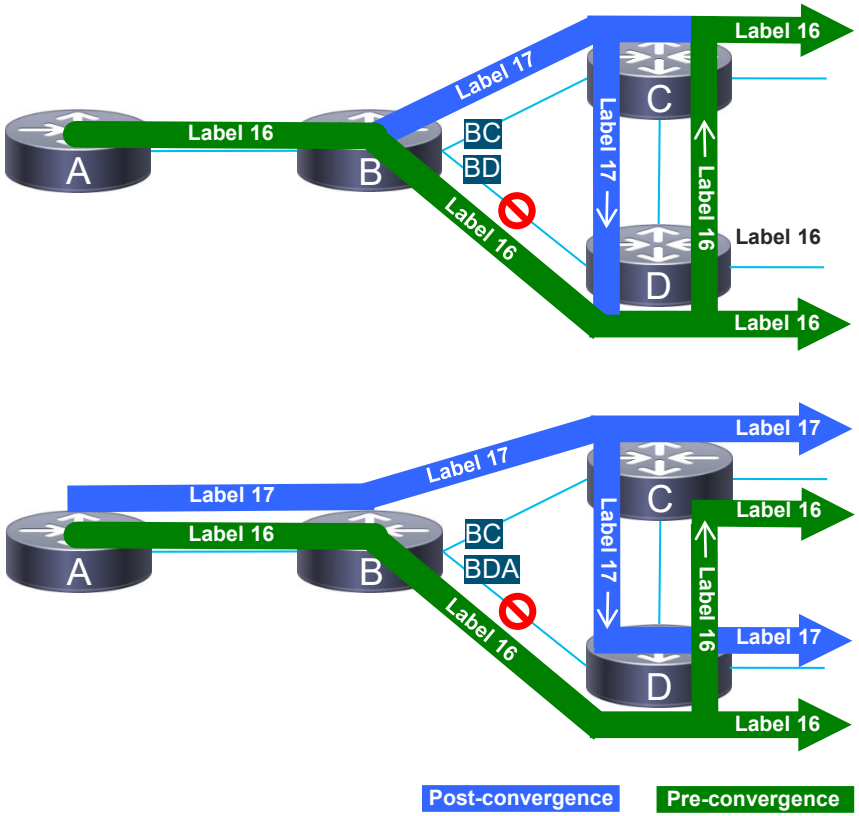
- Solutions:

Assign different Labels for the changed path!

Note, TreeSID is not identified by a single label value!

Rebuild the entire TreeSID with the new Label.

A link failure in the network causes the entire tree to be re-programmed.

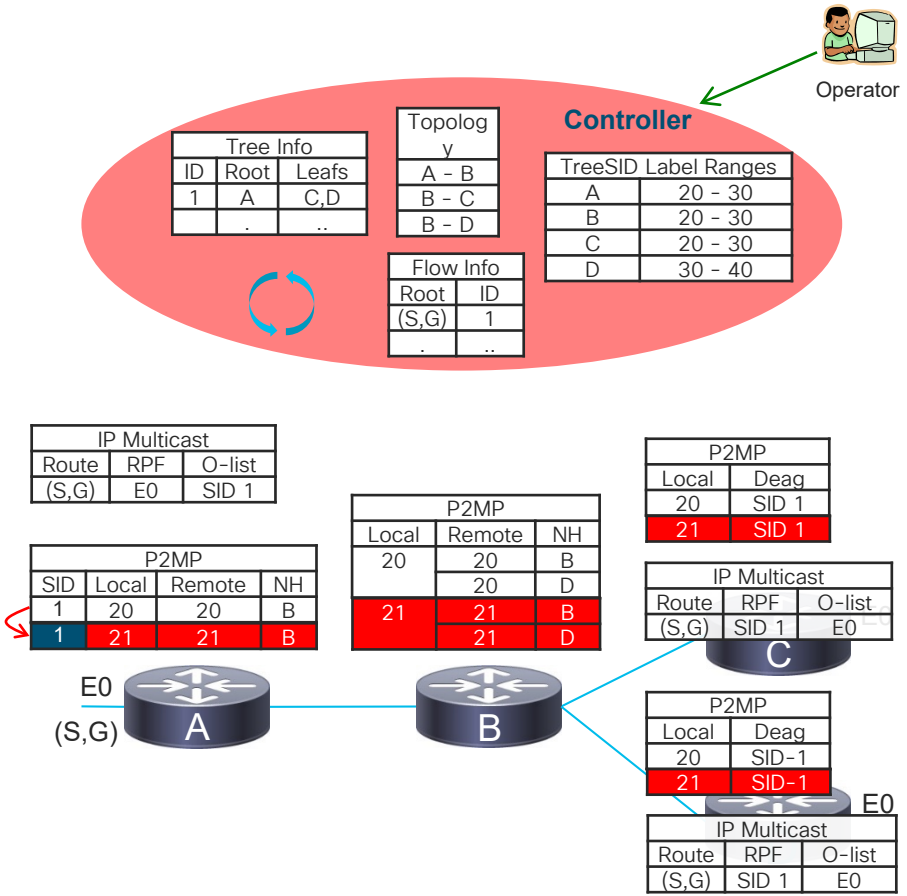


Tree SID – Fast Convergence

- Convergence of the TreeSID will be slower compared to PIM and mLDP, it might be similar to RSVP-TE.
- The controller has to be notified of topology failures and re-program the TreeSID.
  - Note, repairing the Tree locally (option 1) is faster than changing the entire Tree.
- Existing technologies can be used to support Fast ReRoute
  - Live-Live
  - MoFRR
  - Link Protection using SR-TE/RSVP-TE/LFA/TI-LFA

Make Before Break

- The controller builds a new Tree with new labels.
- The old Tree is untouched.
- The IP Multicast traffic moved to the new Tree by assigning the SID from the old to the new Label.
- The old Tree can now be removed.



Make Before Break

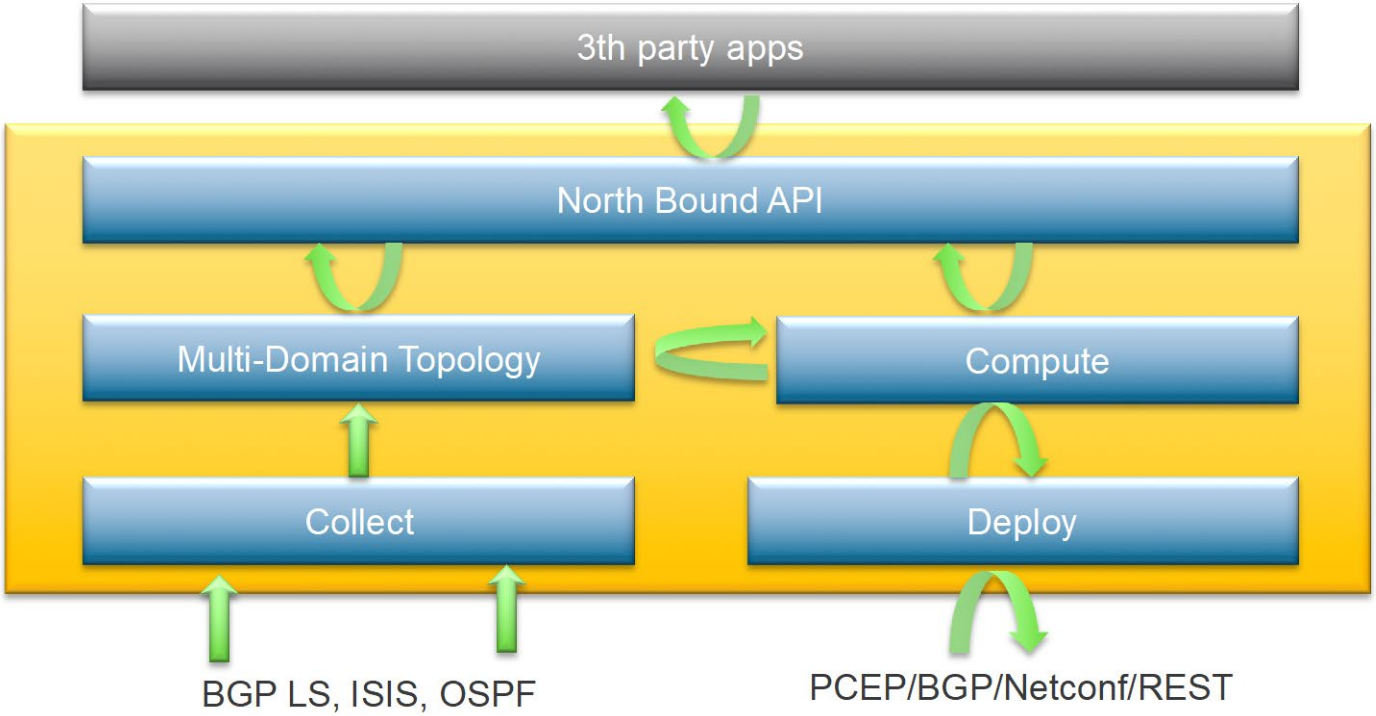
- MBB is required to minimize the impact on receivers that are still receiving traffic, while they are changed to a new Tree.
- This is relevant for:
  - Tree SID Convergence option 1.
  - Fast Convergence re-optimization

SR-PCE Controller

SR-PCE

- Calculating a Path and/or Tree is a well known Router function of the Path Computation Element (PCE).
- PCE has been an integral part of RSVP-TE, and extended for SR-TE
- What if the “Controller” is just used to calculates paths and trees, can the Controller just be a router in the network?
- Can we re-use the PCE?

SR-PCE



SR-PCE

- SR-PCE is an effort within Cisco to provide a “multi-domain” stateful PCE running in an IOS-XR virtual router. Thus acting as a standalone “Controller” element
- The code is taken from XR.
- The SR-PCE controller runs on XR, potentially on IOS, Nexus and Linux.
- SR-PCE can be deployed as a single Controller
- SR-PCE is the controller for TreeSID.

Summary

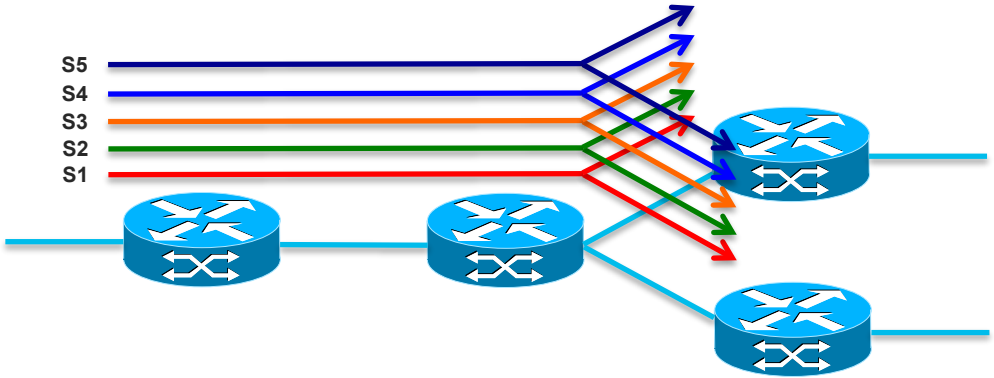
- TreeSID is a more scalable mechanism for building a constrained Tree compared to RSVP-TE P2MP.
  - No periodic signaling and no sub-LSPs
- The TreeSID can be identified by a single Label value, like SR!
- TreeSID is not Source Routed!
- TreeSID does not solve state and convergence problems we see with Multicast!
- TreeSID + SR-PCE is a mechanism to build trees without relying on PIM, mLDP and RSVP-TE.
- Focus for TreeSID is on IPTV

BIER- new
Bit Indexed Explicit Replication

Solution Overview

What is the problem with Multicast

- Each Tree has its own unique receiver population.
- To efficiently forward/replicate there is a Tree per flow!
- This means State is created in the network.



Multicast Routing State

- State is created in the network using a Multicast routing protocol like PIM, mLDP, RSVP-TE, TreeSID.
- State means resources consumed, memory/CPU.
- Convergence is impacted by the amount of State.
- In order to manage the network, the protocol needs to be understood by the network operator.
- Different levels of complexity based on protocol choice.

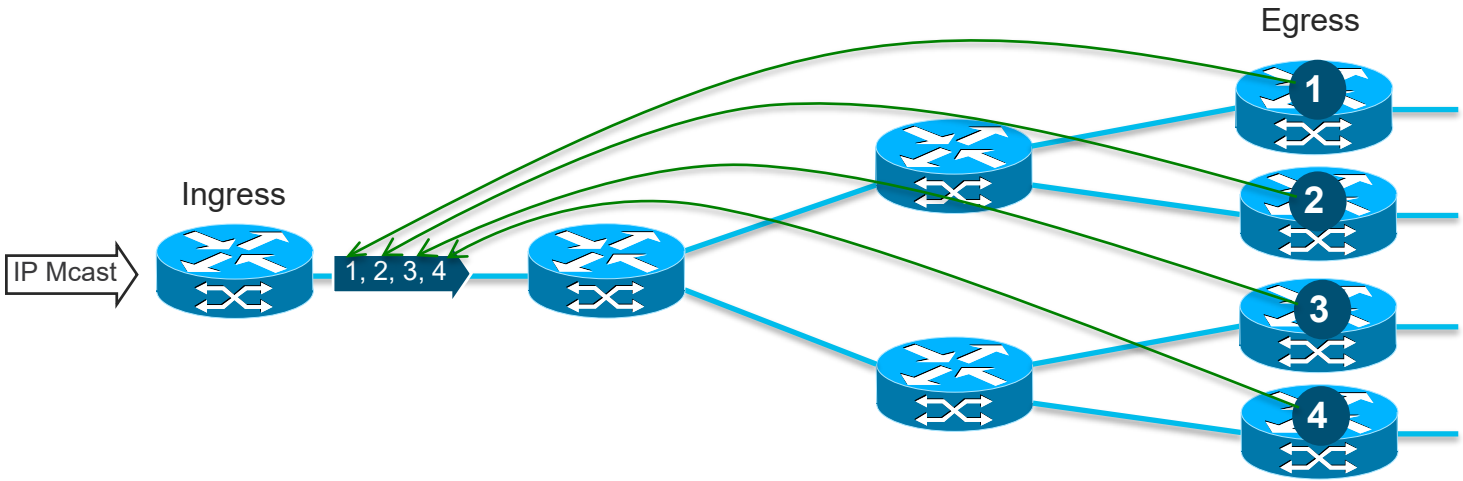
What is BIER?

What is BIER?

- BIER is a new forwarding paradigm to forward and replication multicast packets through the network.
- Packets are forwarded using a special Header that is embedded into the packet.
- Routers build a special forwarding table to forward/replicate using the BIER header.
- BIER forwarding is State Less!

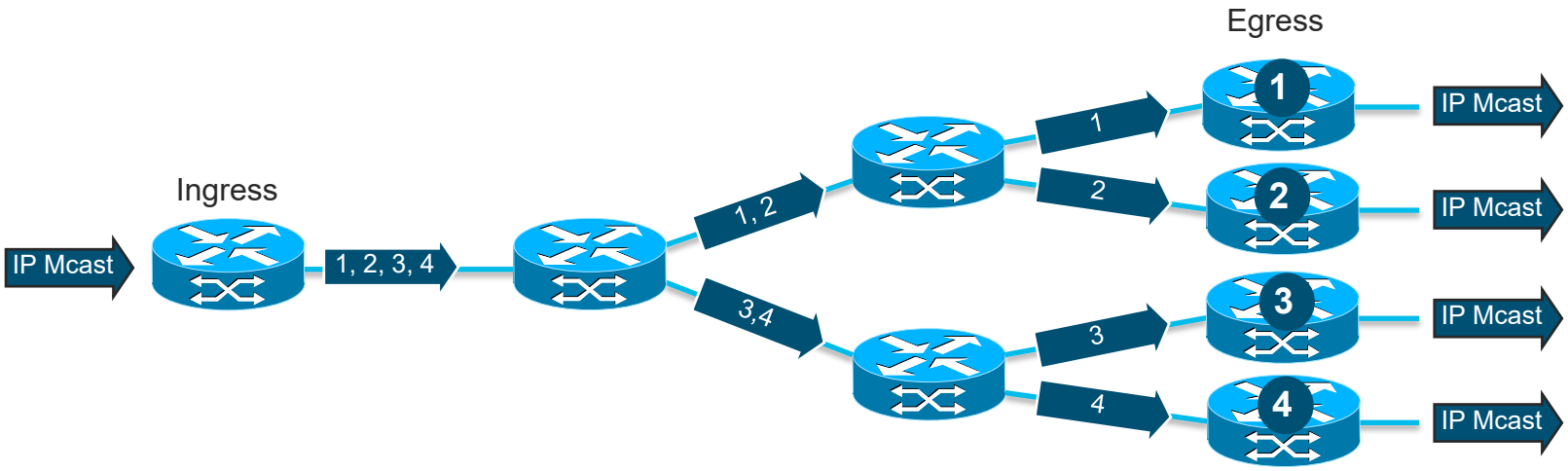
How does BIER work

- We give the Egress routers an identifier.
- The Ingress router includes the "identifiers" in the packet.



How does BIER work

- Packet is forwarded hop-by-hop using the “identifier”
- Each “identifier” is forwarded along the unicast (SPF) path.

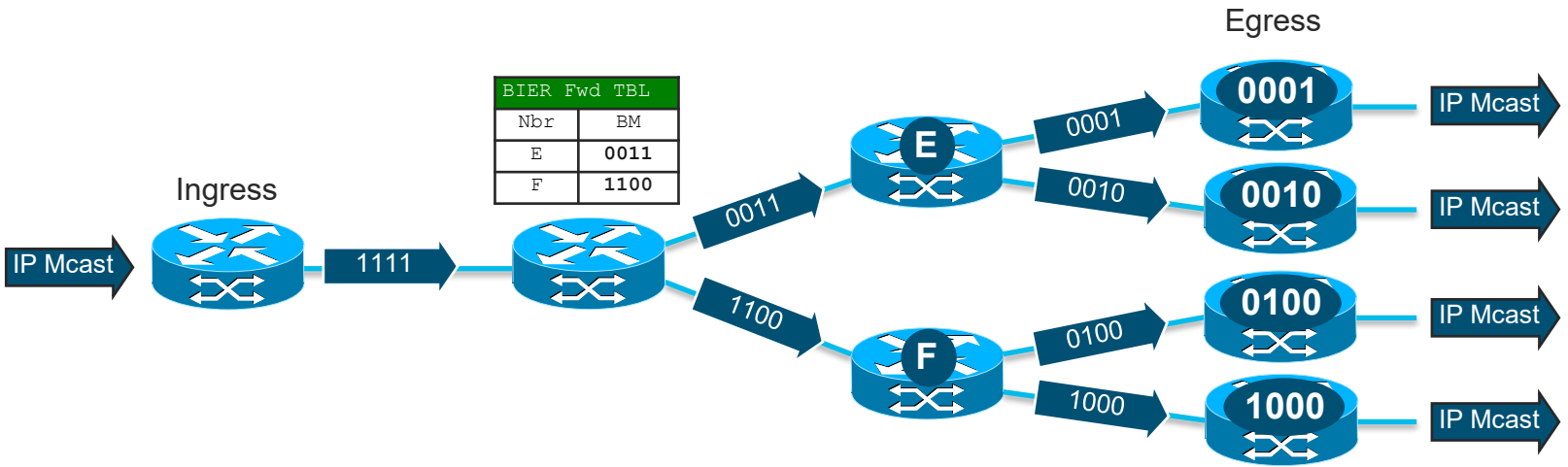


How does BIER work

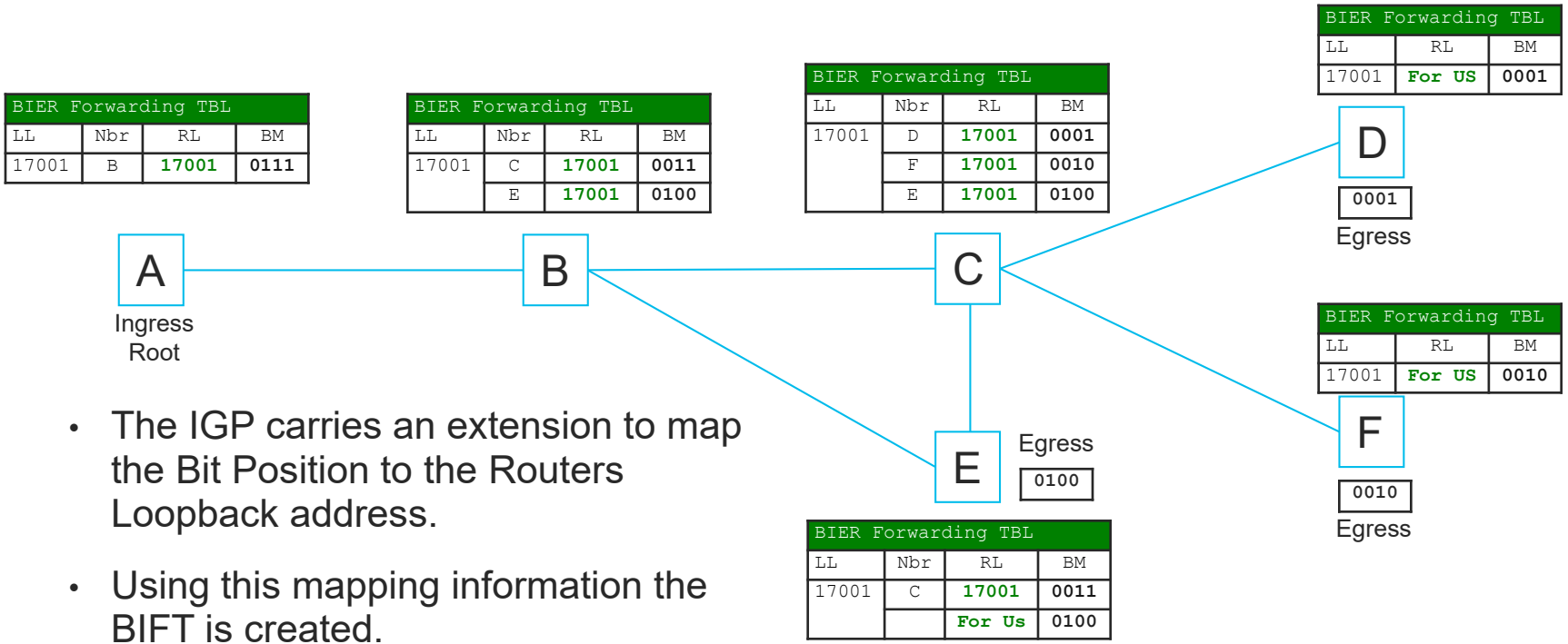
- The smaller the identifier, the more we can fit into a single packet, how small can an identifier be?
- A single Bit!!!
- With BIER the Egress Identifier is a Bit Position.
- We include a BitString of 256 bits into the packet.
- Manipulations of a BitString is much easier compared to including a list of numbers.

How does BIER work

- Each Egress BIER router has a unique Bit Position (we call BFR-id).
- Routers maintain a forwarding table of Bitmask's

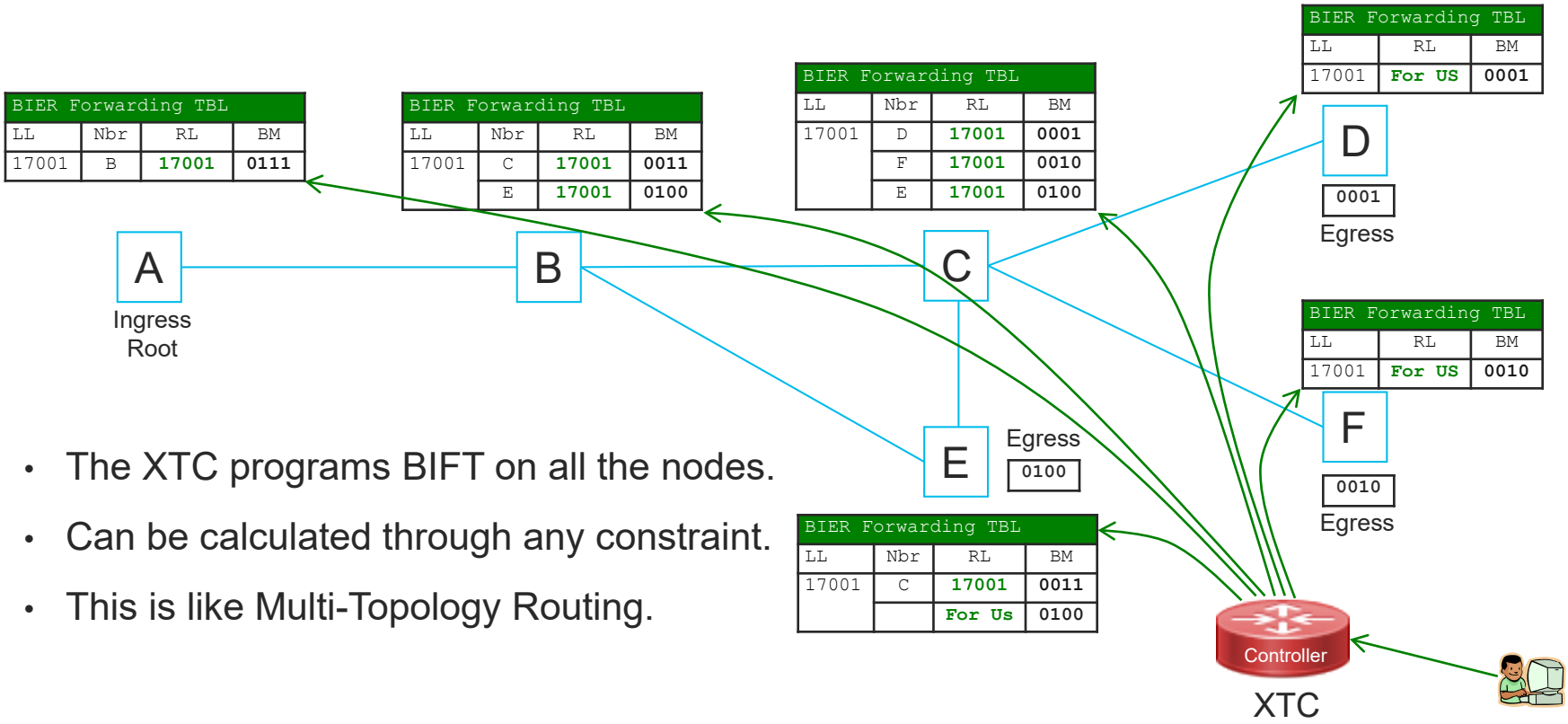


Bit Index Forwarding Table (BIFT) by IGP



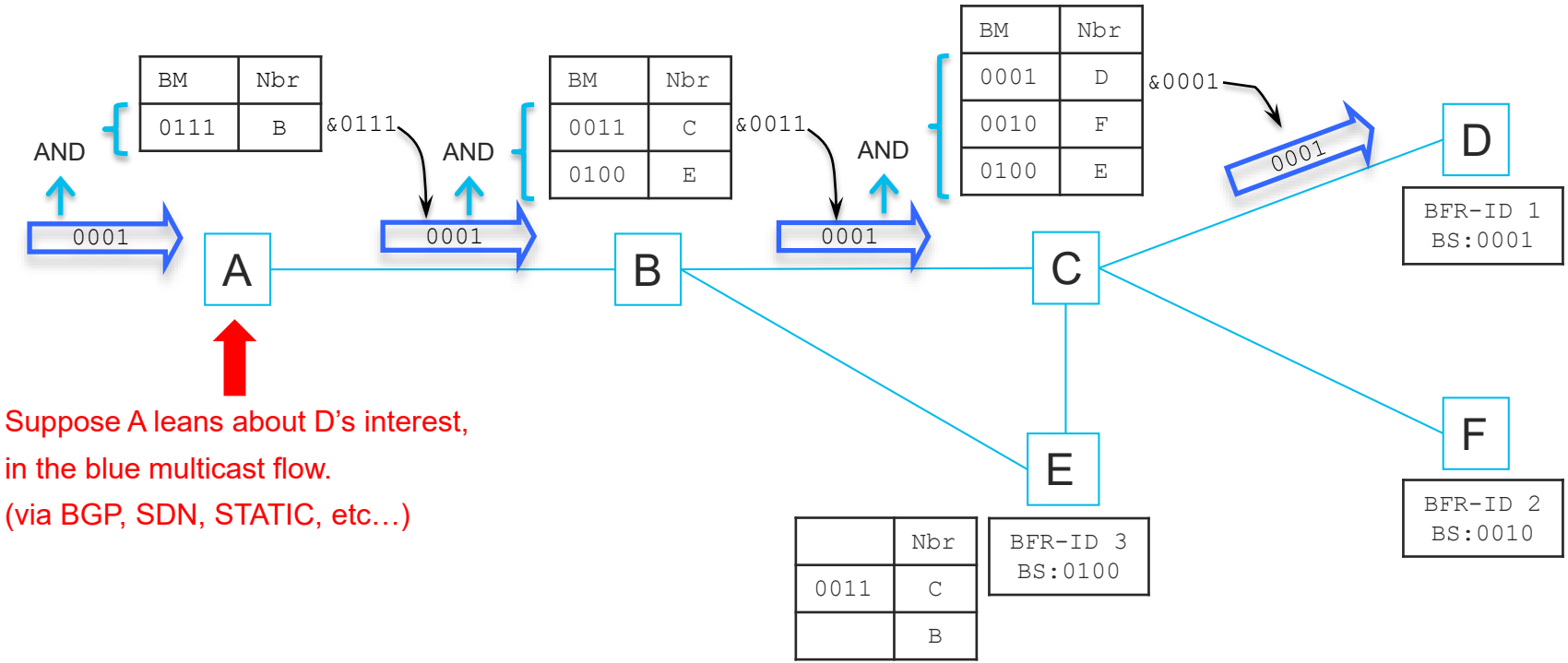
- The IGP carries an extension to map the Bit Position to the Routers Loopback address.
- Using this mapping information the BIFT is created.
- This Table follows the IGP SPF.

Bit Index Forwarding Table (BIFT) by XTC

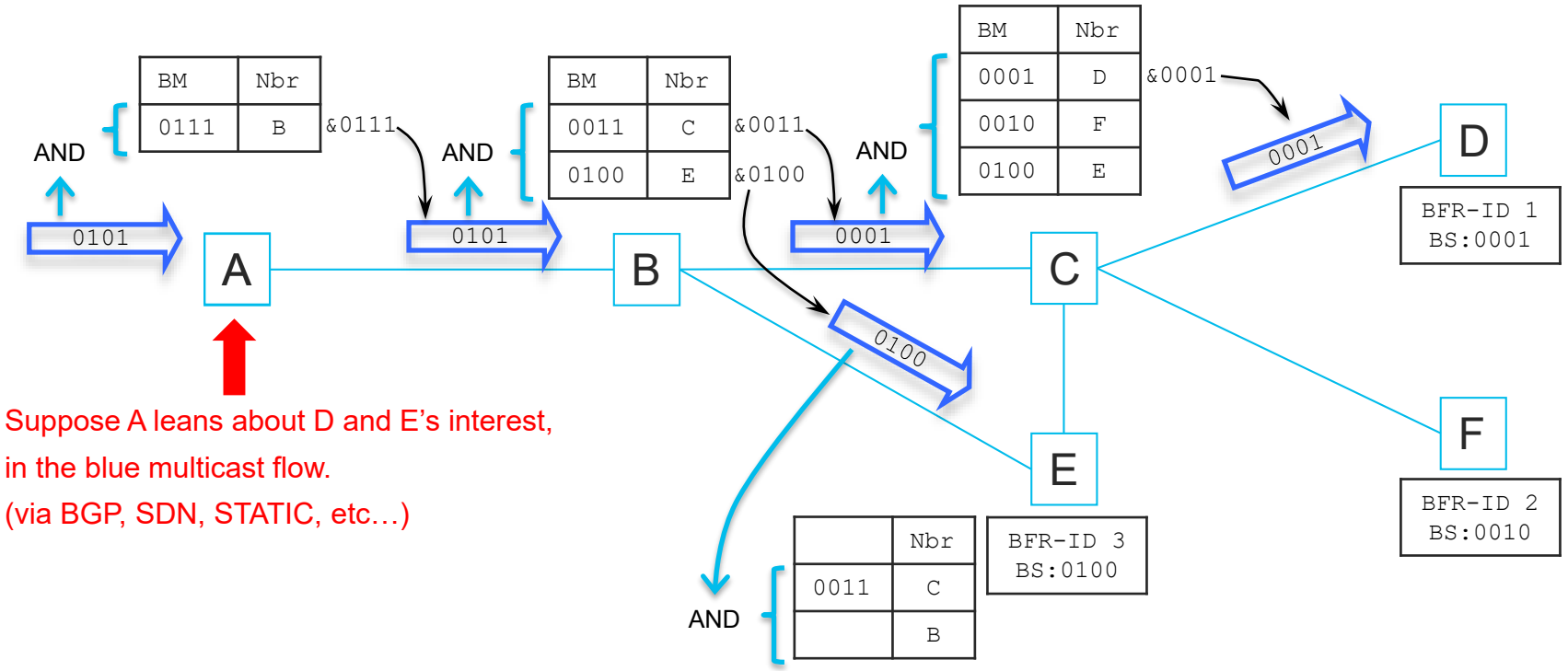


- The XTC programs BIFT on all the nodes.
- Can be calculated through any constraint.
- This is like Multi-Topology Routing.

Forwarding Packets

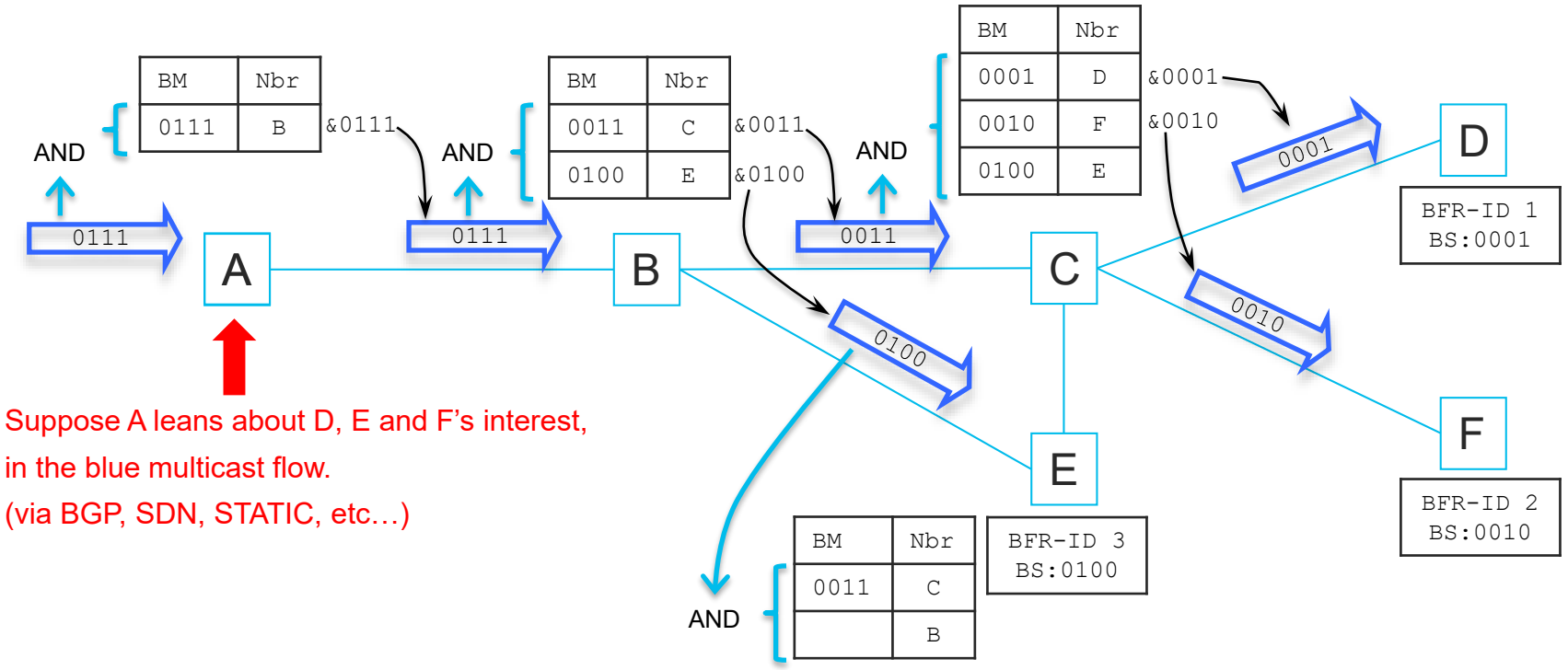


Forwarding Packets



Suppose A leans about D and E's interest,
in the blue multicast flow.
(via BGP, SDN, STATIC, etc…)

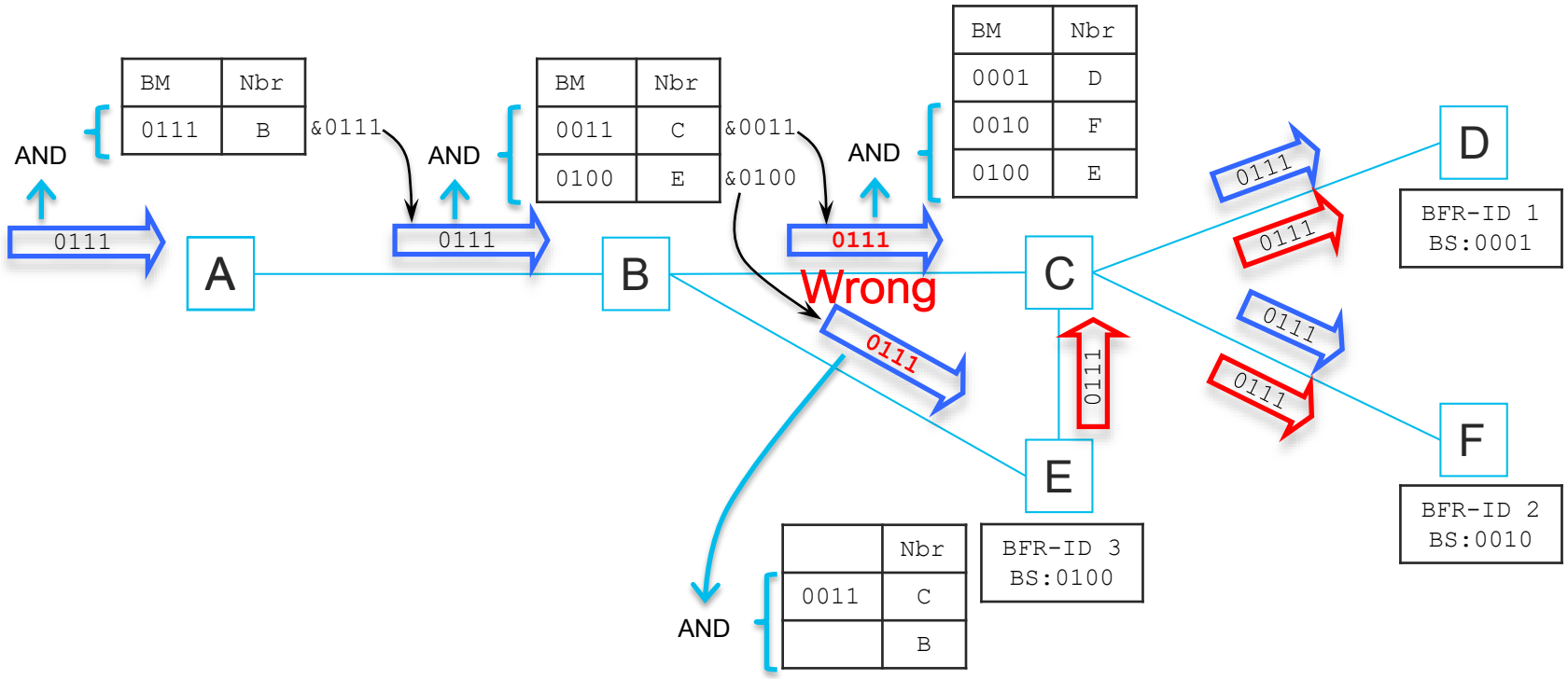
Forwarding Packets



Forwarding Packets

- As you can see from the previous slides, the result from the bitwise AND (&) between the Bit Mask in the packet and the Forwarding table is copied in the packet for each neighbor.
- This is the key mechanism to prevent duplication.
- Look at the next slide to see what happens if the bits are not reset
- If the previous bits would not have been reset, E would forward the packet to C and vice versa.

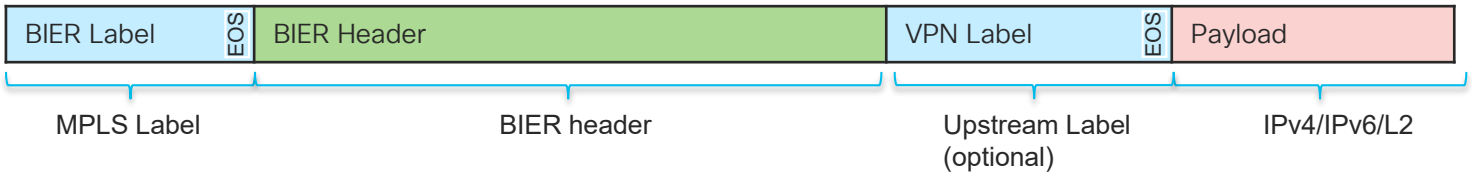
Forwarding Packets (wrong behavior)



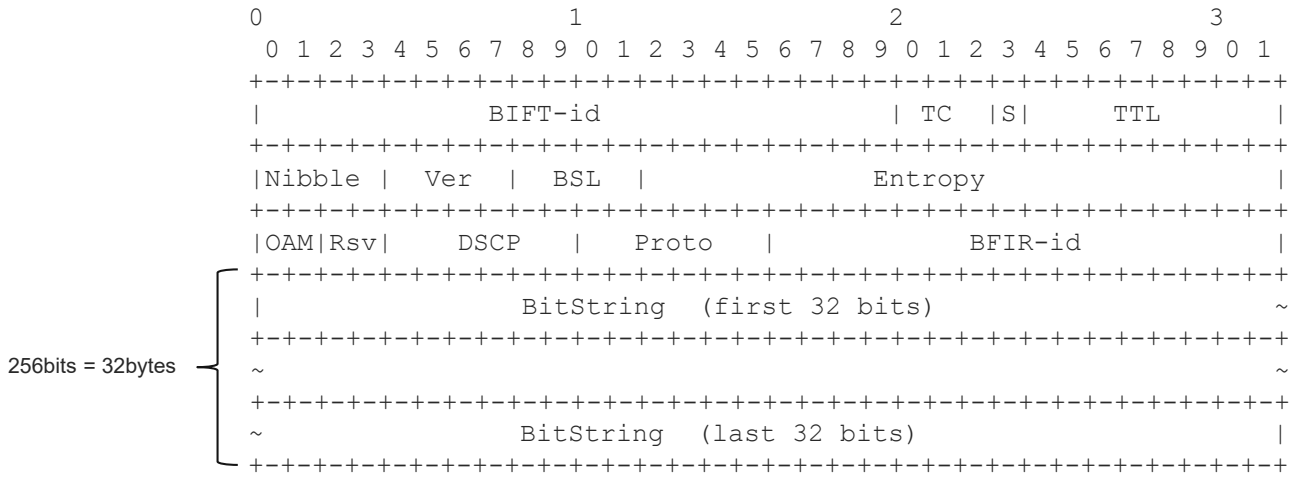
Encapsulation

MPLS encapsulation

- The Top Label is allocated by BIER from the downstream platform label space.
- The BIER Header follows directly below the BIER label.
- There is a single BIER label on top, unless the packet is re-encapsulated into a unicast MPLS tunnel.
- The VPN label is allocated from the upstream context label space (optional).



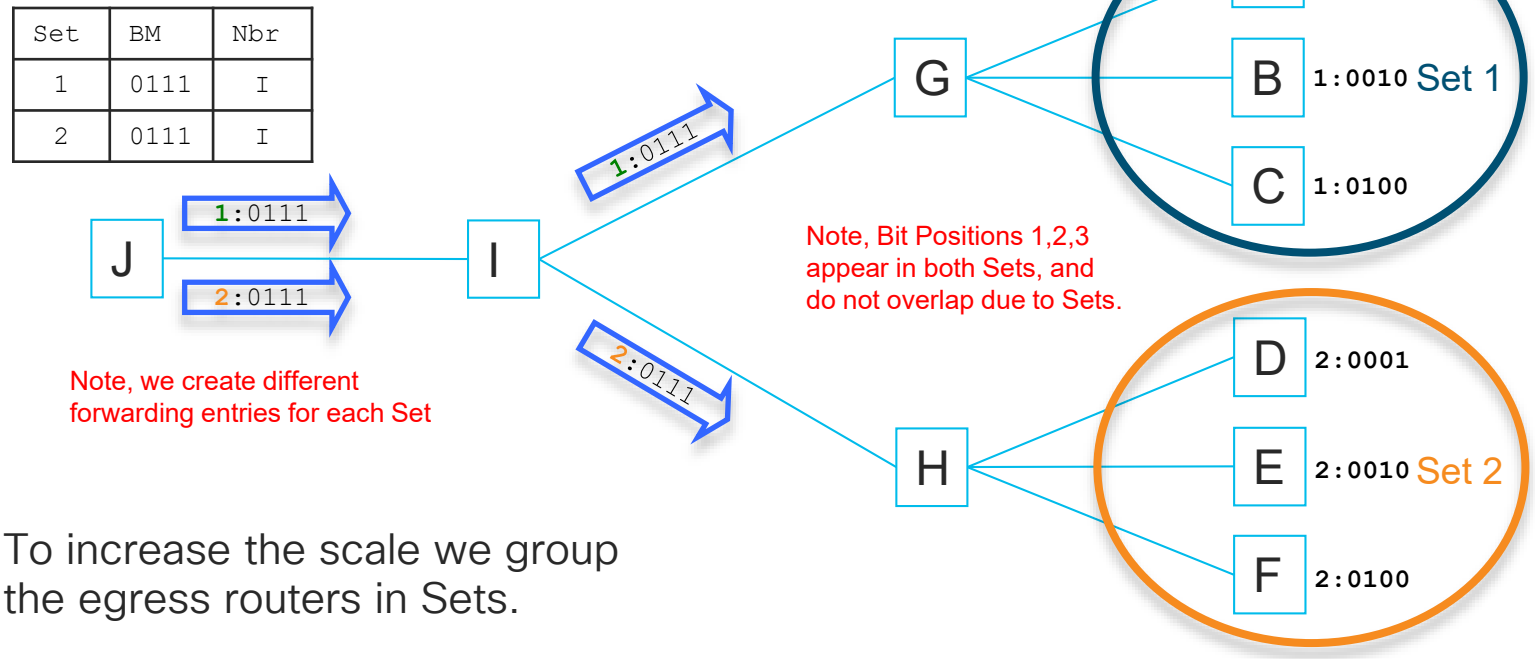
BIER Header



- https://tools.ietf.org/html/rfc8296

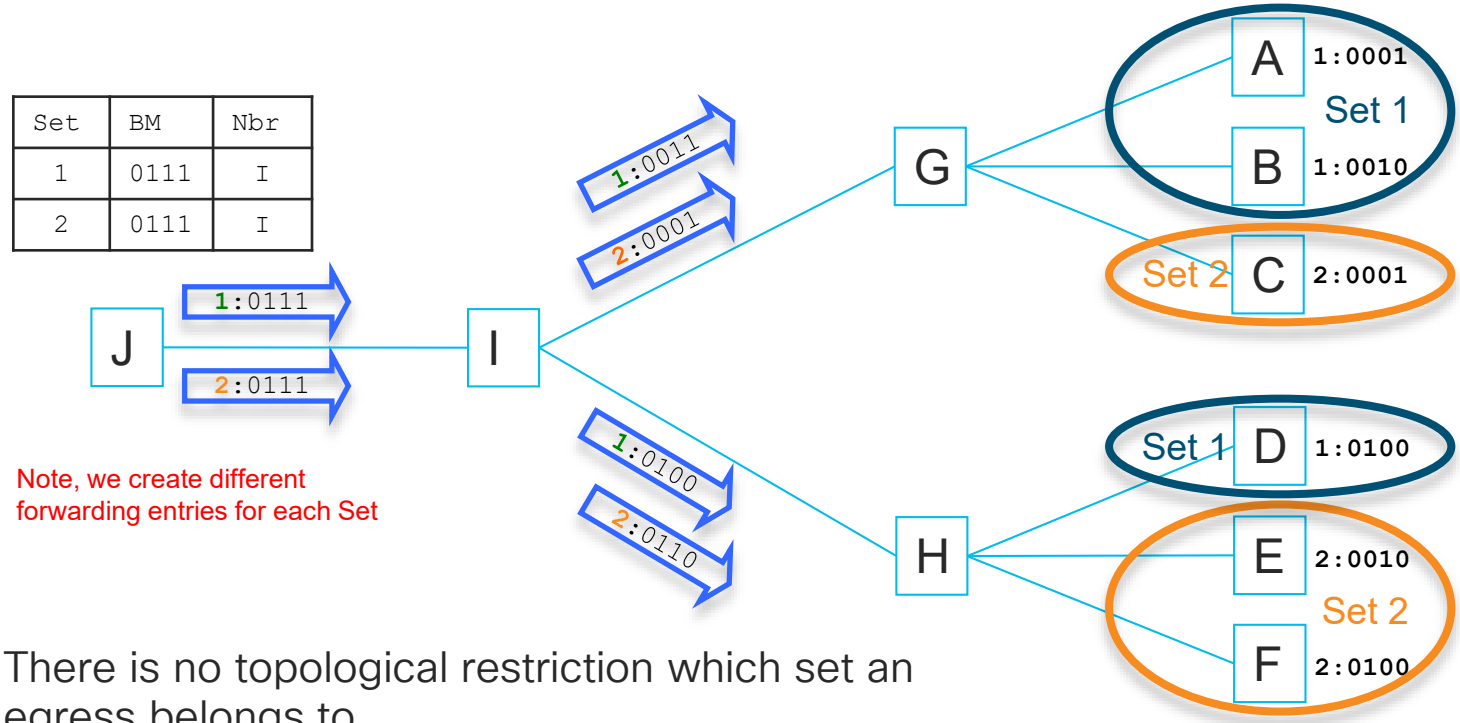
Sets

BIER Sets



- To increase the scale we group the egress routers in Sets.

BIER Sets



- There is no topological restriction which set an egress belongs to

BIER Sets

- If a multicast flow has multiple receivers in different Sets, the packet needs to be replicated multiple times by the ingress router, for each set once.
- Is that a problem? We don’t think so…
- The Set identifier is part of the packet.
- Can be implemented as MPLS label.
- The value of the Set is derived from the BFR-id, no need for signaling.

How many Bits and encoding

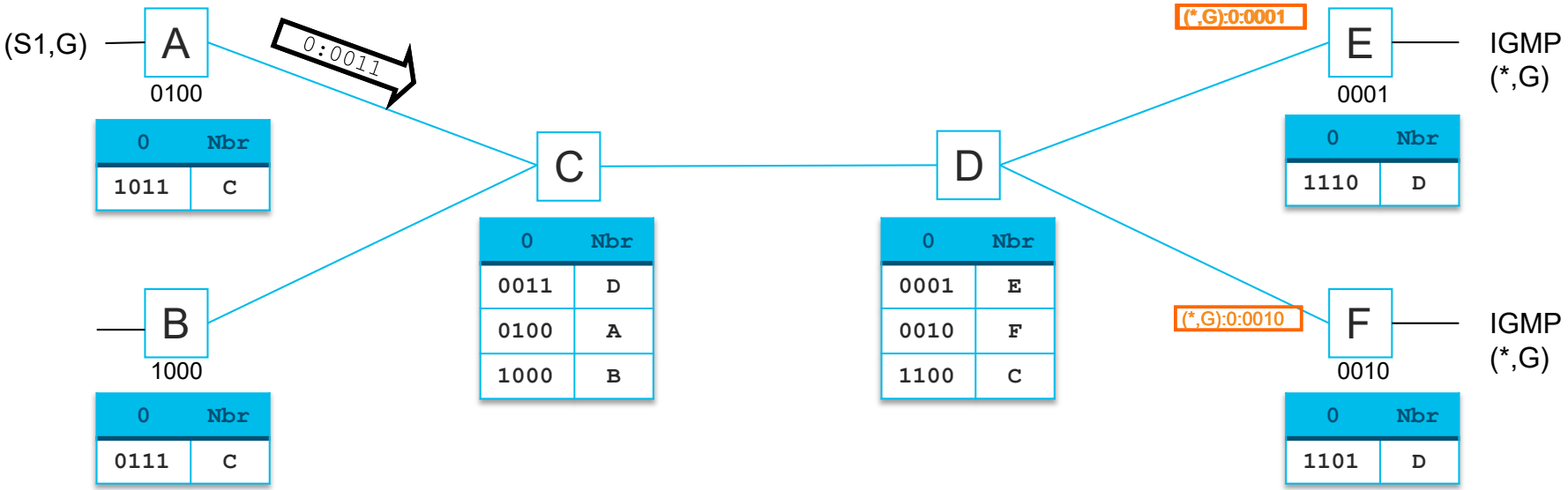
- The number of multicast egress routers that can be addressed is depending on the number of Bits that can be included in the BitString.
- The architecture RFC defined a range from 64 to 4096.
- Default is 256 bits, agreed with other vendors at IETF.
- We identified 5 different encoding options, most attractive below:
  1. MPLS, below the bottom label and before IP header.
  2. New Ethertype (0xAB37) for BIER.

Native BIER

Native BIER

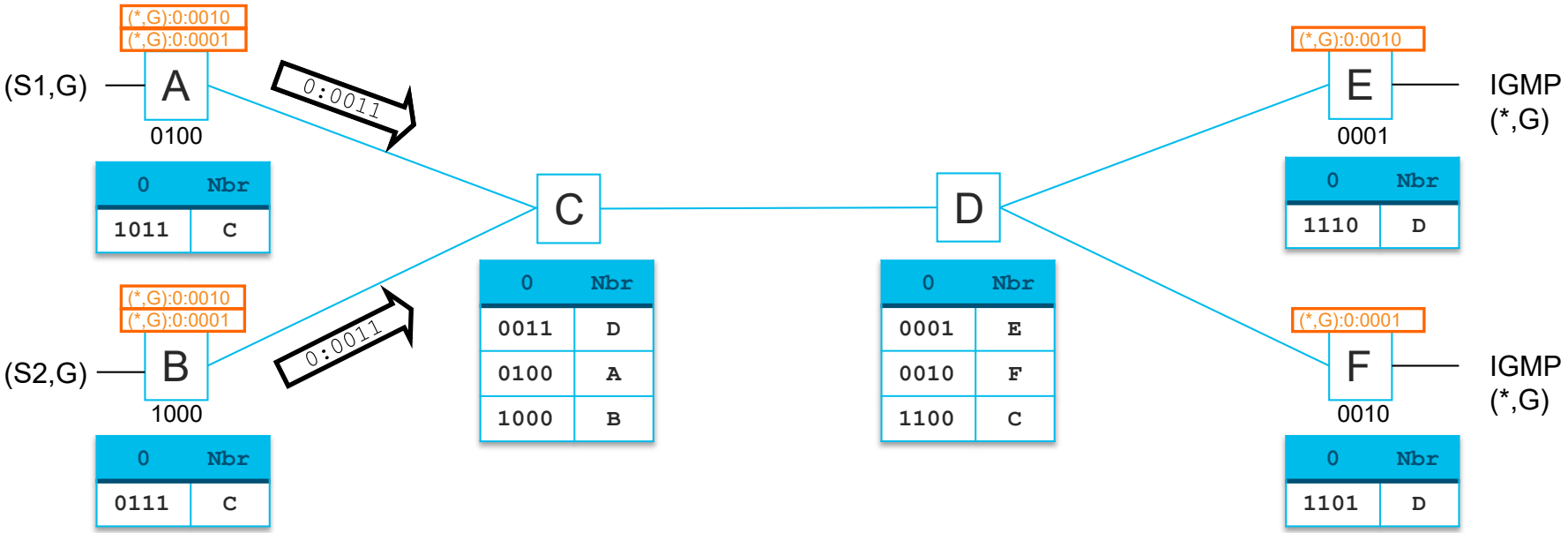
- With Native BIER there is NO PIM involved, just IGMP and BIER.
- The Source and Receiver(s) are connected to BIER router.
- There are no RP's.
- There is no equivalent of PIM modes, like sparse, ssm, bidir etc..
- We speak of 'single' sender and 'multi' sender, which is basically the same solution.
- The overlay signaling can be BGP or SDN based.

Native BIER



- E and F announce their Group membership via overlay to all other routers.
- A BIER router connected to the Source can immediately start sending.

Native BIER



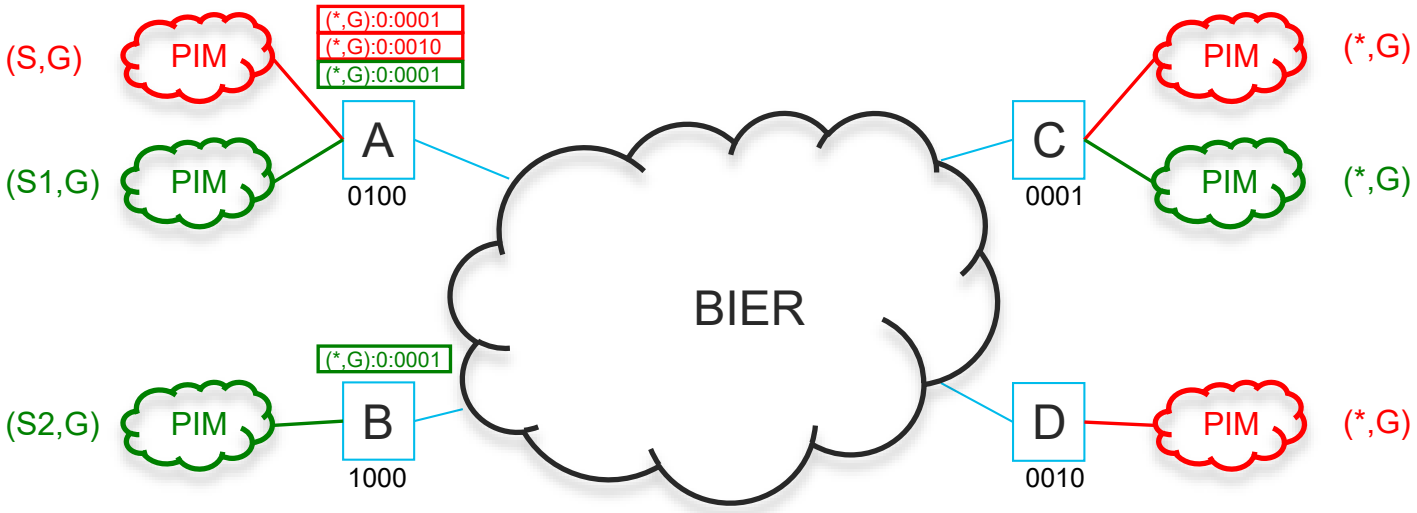
- When B learns about a new source, it can immediately start sending.

MVPN over BIER

MVPN over BIER

- BIER replaces PIM, mLDP, RSVP-TE or IR in the core.
- BIER represents a full mesh (P2MP) connectivity between all the PE's in the network.
- There is no need to explicitly signal any MDT's (or PMSI's).
- With MVPN there are many profiles,
  - This is partly due to the tradeoff between 'State' and 'Flooding'.
  - Different C-multicast signaling options.
- MVPN over BIER, there is one profile.
  - BGP for C-multicast signaling.
- No need for Data-MDTs.

MVPN over BIER



- The BGP control plane defined for MVPN can be re-used.
- Big difference, there is no Tree per VPN…!!!
- The BIER packets needs to carry Source ID and upstream VPN context label

IETF

IETF

- The BIER idea was presented in a BOF at the IETF in Hawaii.
  - November 2014.
- A new BIER Working Group has been formed (bier@ietf.org)
- BIER architecture became RFC 8279 (November 2017)
- Vendors collaborating (co-authoring) with us;

BIER Conclusion

Stateless

- There is no Multicast receiver or flow state in the core network (only edge).
  - Imposition of the BIER Header may be done by application, removes state from ingress.
- There is no tree state in the network.
- There is no tree building protocol or logic in the network.
- There is only topology state for the BFER's, derived from unicast routing.

Scale

- Since there is no flow and tree state, converges as fast as unicast.
- Compared to Ingress Replication, saves 256x (minimum)

Simplicity

- No Reverse Path Forwarding (RPF)
- No Rendezvous Points
- No shared tree / source tree switchover
- No receiver driven tree building
- BIER is like unicast
- State is in the packet (like Segment Routing)

More information and references

- bier@cisco.com
- https://dcloud-cms.cisco.com/demo/cisco-bit-indexed-explicit-replication-v2
- https://datatracker.ietf.org/wg/bier/charter/

Multicast and Segment Routing Conclusion

Conclusion

- Simplifying the network and making it Scale better is top of mind for most customers looking to change their Multicast network.
- The solution that solves scale and simplicity best is BIER.
  - BIER is not yet widely available due to HW dependency, target end 2020 (XR 7.3.1).
- Networks that have deployed SR-TE with XTC and have simple Multicast requirements (IPTV) could consider TreeSID.
  - TreeSID available in XR 7.0.1
- We recommend using mLDP for deploying Multicast in SR networks today.

Complete your online session survey

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (starting on Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Content Catalog on ciscolive.com/emea.

Cisco Live sessions will be available for viewing on demand after the event at ciscolive.com.

cisco Live!

Continue your education

Demos in the Cisco Showcase

Walk-In Labs

Meet the Engineer 1:1 meetings

Related sessions

Thank you

You make **possible**