Interconnection of DC Sites Over an SR-Enabled Core

draft-drake-bess-datacenter-gateway-02

John Drake (jdrake@juniper.net)
Adrian Farrel (adrian@olddog.co.uk)
Eric Rosen (erosen@juniper.net)
Keyur Patel (keyur@arrcus.com)
Luay Jalil (luay.jalil@verizon.com)

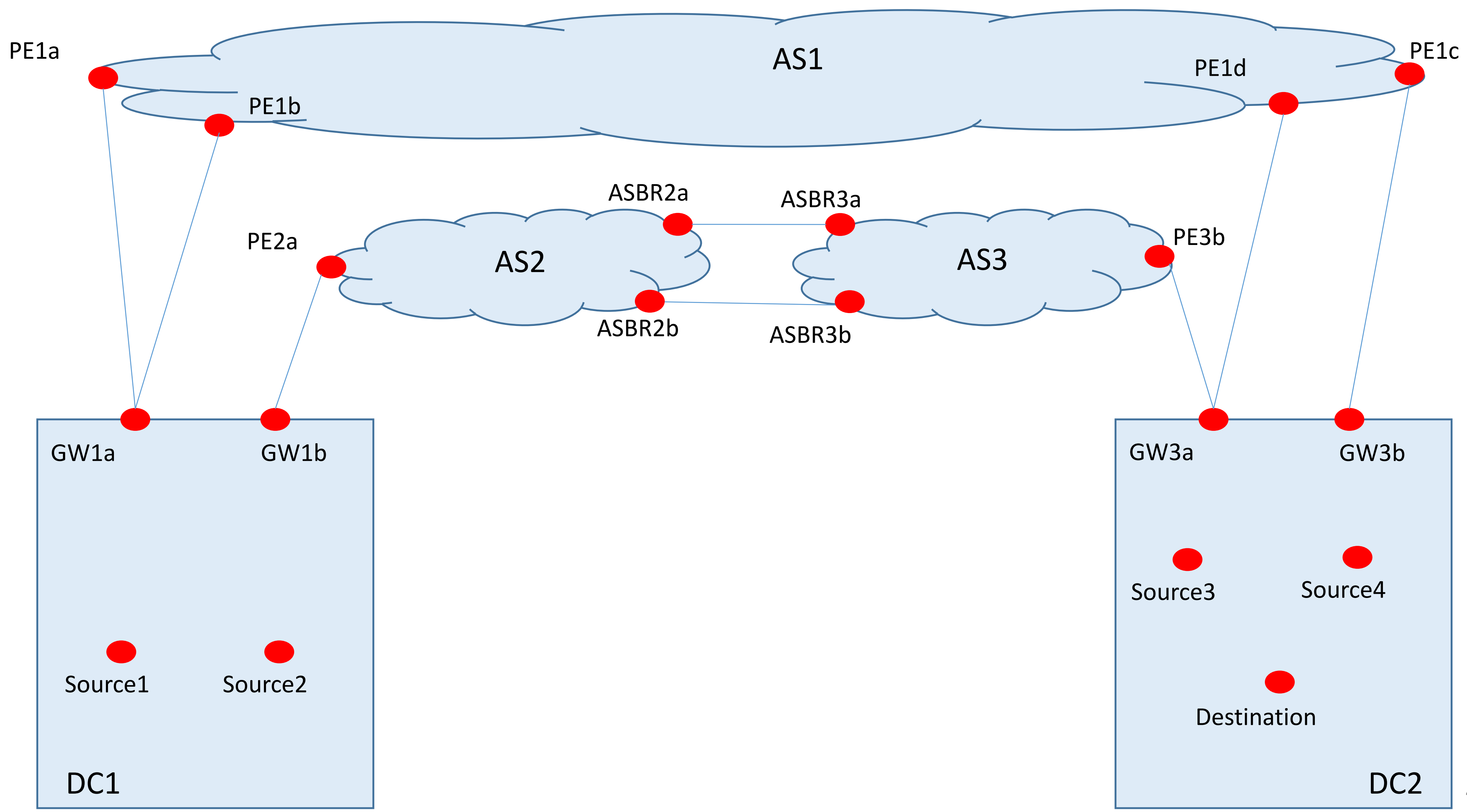
History

- It’s only a little draft
  - Just 9 pages
  - Go on, read it, you know it makes sense
- May 2016 -00
  - First version
- June 2016 -01
  - Clarify route withdrawal
- IETF-96 (July 2016)
  - Didn’t request an agenda slot (very full agenda)
- October 2016 -02
  - Additional co-authors helped with…
    - Fix typos
    - Various clarifications
    - Fix references

Background

- DC networks are usually well-planned:
  - But need traffic steering to cope with changes in load
- DC sites are interconnected by backbone networks
  - Operated by the DC owner or by a third party
  - Usually with multi-homing of sites
- Many different protocols proposed for use within the DC
- Several protocols proposed for use between DC sites
- Our objective…
  - Provide a control plane solution for end-to-end inter-site DC connectivity
  - Support the different data plane options
  - Re-use existing technologies
  - Scalable and flexible solution
  - Easy to operate
- What we ended up with is a sort of Applicability Statement
  - It's Informational

Reference Architecture



The Problem Space

- Three types of traffic flow
  - Intra-DC traffic
    - Source and destination in same site
    - TE needed to:
      - Achieve load-balancing across network resources
      - Avoid degraded or out-of-service resources
      - Achieve different qualities of service
    - But this is out of scope for us
  - Inter-DC traffic
    - Traffic between DC sites
    - TE is usually used in the backbone
      - Includes choice of DC site entry and exit points (Gateways)
  - End-to-end DC traffic
    - TE in source DC site
    - TE across backbone
    - TE in destination DC site
    - Choice of DC Gateways and backbone interconnections
      - Load balance traffic across multiple gateways <u>to</u> an egress data center
      - Traffic engineer the end-to-end path

Decomposing the Problem

- The following decisions must be made (reference the architecture picture)
  - In which DC the Destination lies
  - Which exit point from DC1 to use
  - Which entry point to DC2 to use
  - How to reach the exit point of DC1 from Source1
  - How to reach the entry point to DC2 from the exit point of DC1
  - How to reach the Destination from the entry point to DC2
- These decisions may be inter-related so break problem into 3 steps:
  - Get the packet from Source1 to the exit point of DC1
  - Get the packet from exit point of DC1 to entry point of DC2
  - Get the packet from entry point of DC2 to Destination

Requirement on Solutions

- Must support several functions and mixtures of those functions:
  - If the DC uses MPLS SR
    - Source routes may be computed by a controller or by a head-end router
    - Label stacks may be populated using BGP-LU, IGP, or PCEP
  - The DC may use non-MPLS forwarding
  - If the DCs use SR, the source and destination DCs may or may not be in the same SR domain:
    - The backbone may be a single private network under the control of the owner of the Data Centers or may be a network operated by one or more third parties
    - Backbone network may utilize MPLS TE tunnels combined with MPLS Segment Routing
  - A single controller may be used to handle the source and destination Data Centers as well as the backbone network, or there may be a different controller for each.
    - The controllers may cooperate and share information to different degrees.

Components of a Solution – Auto Discovery

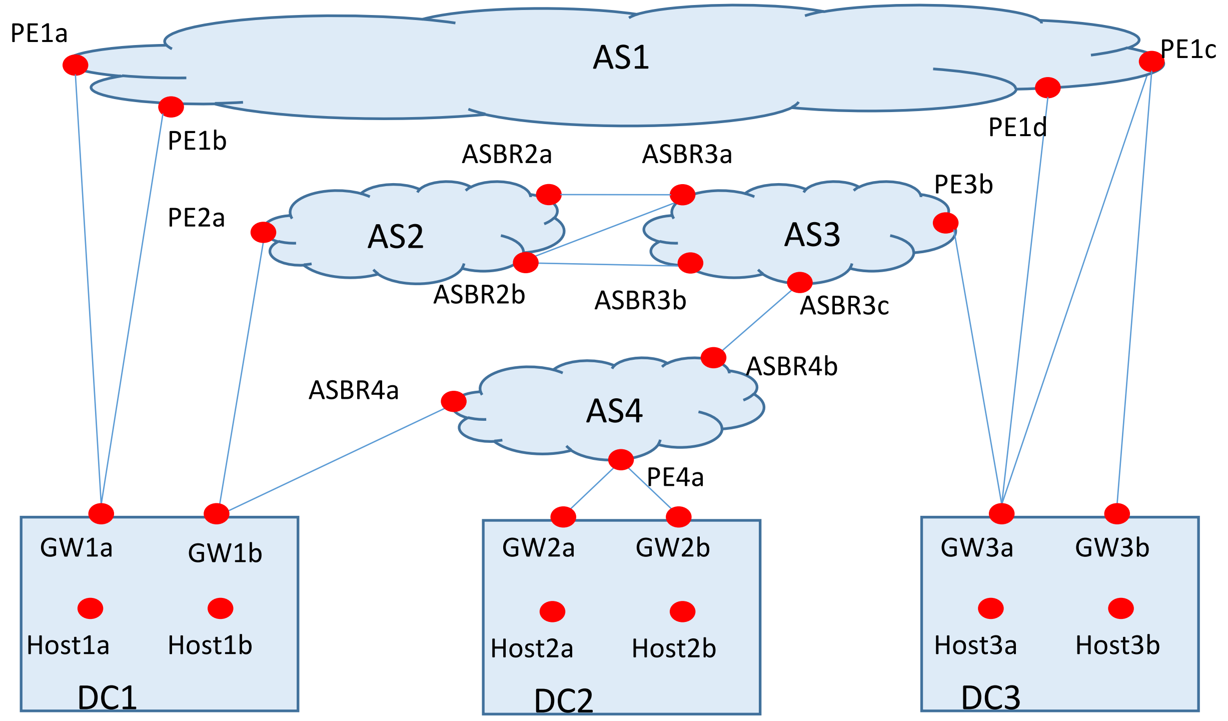
- Discover which Gateways give access to a destination prefix
  - Like an L3VPN problem
  - Existing proposed solution (assumed by EPE draft)
    - Uses BGP ADD_PATH and doesn't scale!
  - Use BGP Tunnel Encapsulation attribute
    - New tunnel type "SR tunnel"
    - GWs to a DC advertise a route to a prefix X
      - Each GW to a DC advertises all GWs for that DC
  - Also define "MPLS Label Stack" sub-TLV for the tunnel encapsulation attribute, and put this in the "SR tunnel" TLV
    - Allows the destination GW to specify a label stack that it wants packets destined for prefix X to have. This label stack represents a source route through the destination DC.

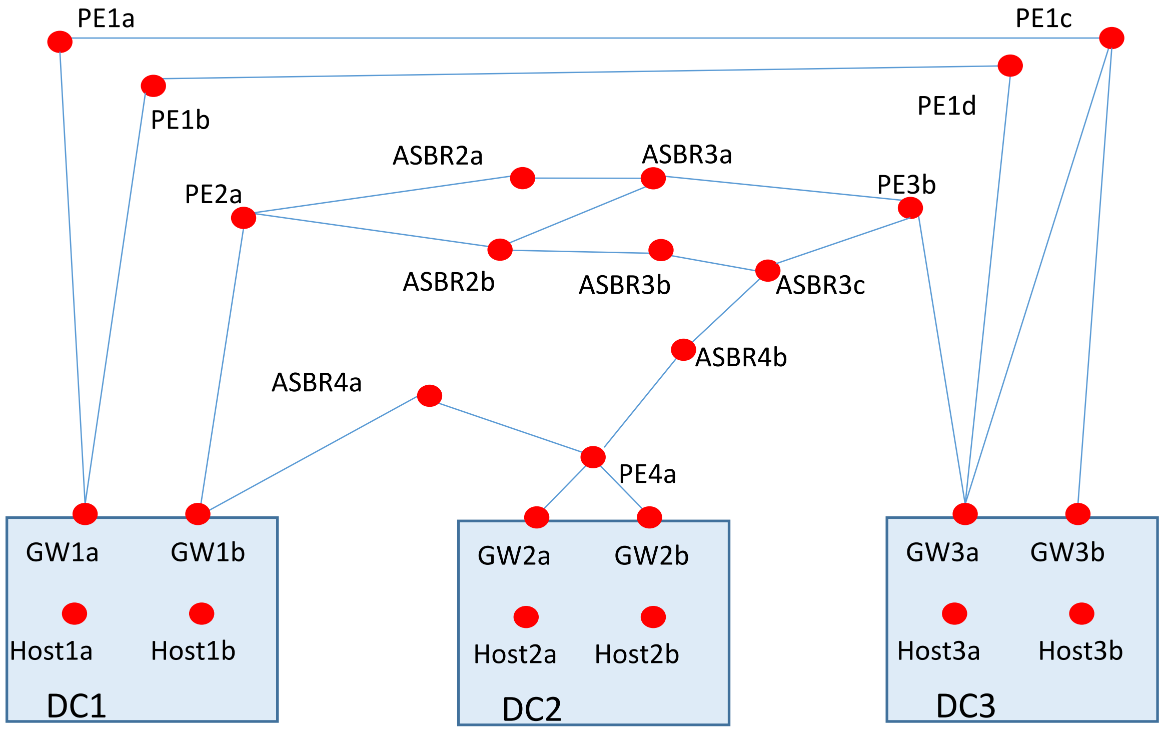
Components of a Solution - Backbone TE

- Need to figure out the egress ASBRs that are attached to a given GW at the destination DC
  - Otherwise cannot correctly traffic engineer
  - The backbone egress ASBRs distribute the information to the source controller using the EPE extensions of BGP-LS.
    - “Here is a list of my EBGP neighbors, and here is a (locally significant) adjacency-sid for each one. ”
- RSVP-TE LSP to be set up by using PCEP to talk to the LSP headend
  - PCRpt message indicates that the LSP has been set up
  - TE-PATH-BINDING TLV in PCRpt allows the headend to assign a “binding SID” to the LSP

BGP-LS Considerations

- BGP-LS can export full network details
  - But it may be more helpful to export a summary
  - Links that are advertised may be physical links, links realized by LSP tunnels, or abstract links
  - Intra-AS links are real links, RSVP-TE LSPs with allocated bandwidth, or the aggregate ECMP b/w within the AS to reach an egress node for that AS





Binding SIDs and Hierarchical LSPs

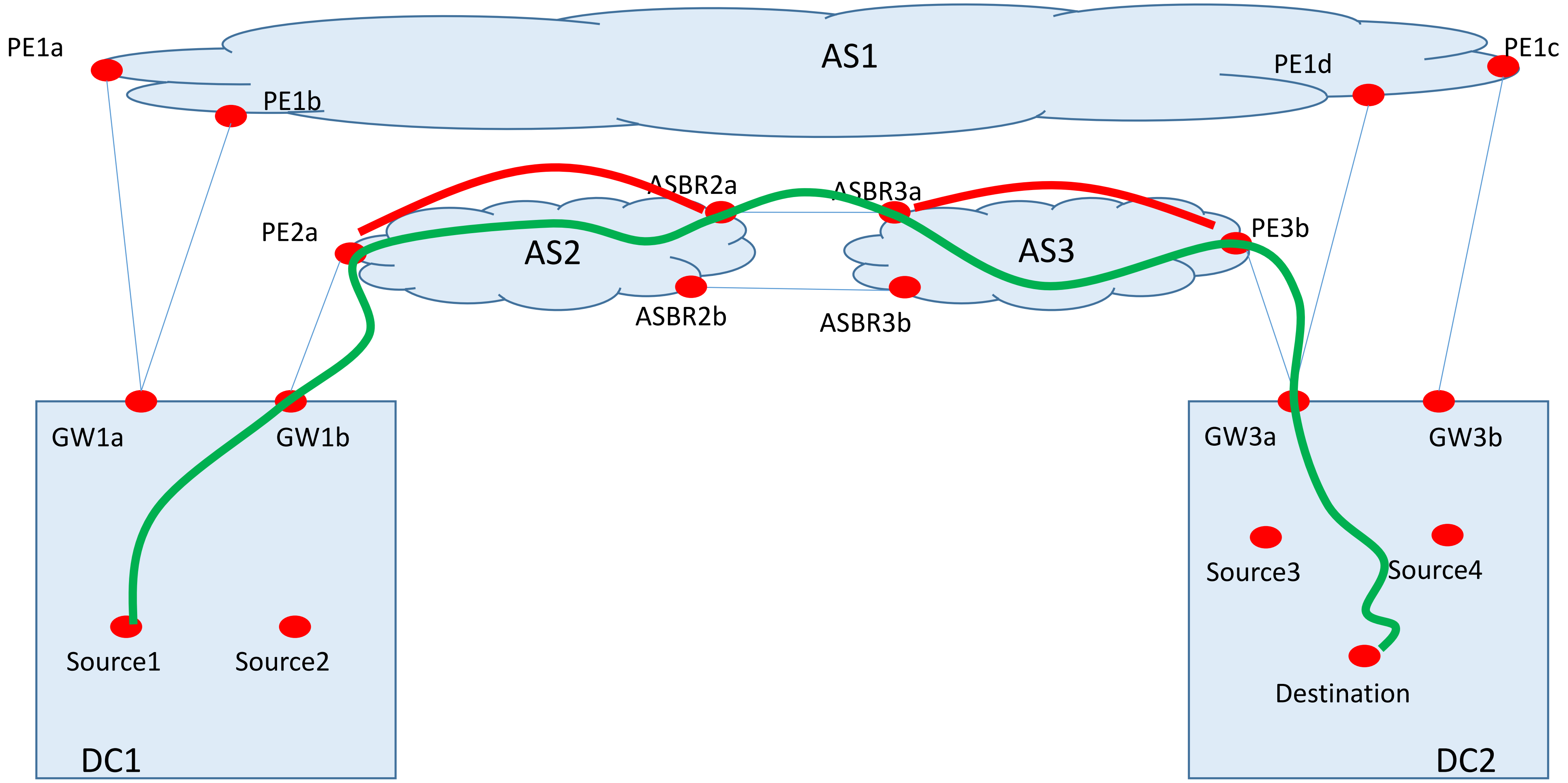
- A binding SID represents a hierarchical LSP
  - LSP may be instantiated by RSVP-TE or SR
  - SID may be assigned by controller or by headend
- The controller can include this label in the label stack that it tells the source (or the GW at the source DC) to put on the data packets being sent to prefix X
  - When the headend receives a packet with this label at the top of the stack it will send the packet onwards on the LSP

Packet Processing

- When a node processes a packet:
  - The label at the top of the label stack indicates the link (or RSVP-TE LSP) on which that node is to transmit the packet
  - The node pops that label off the label stack before transmitting the packet on the link
  - If the top label is a node-SID, the node processing the packet is expected to transmit the packet using weighted ECMP

What Does the Data Plane Look Like?

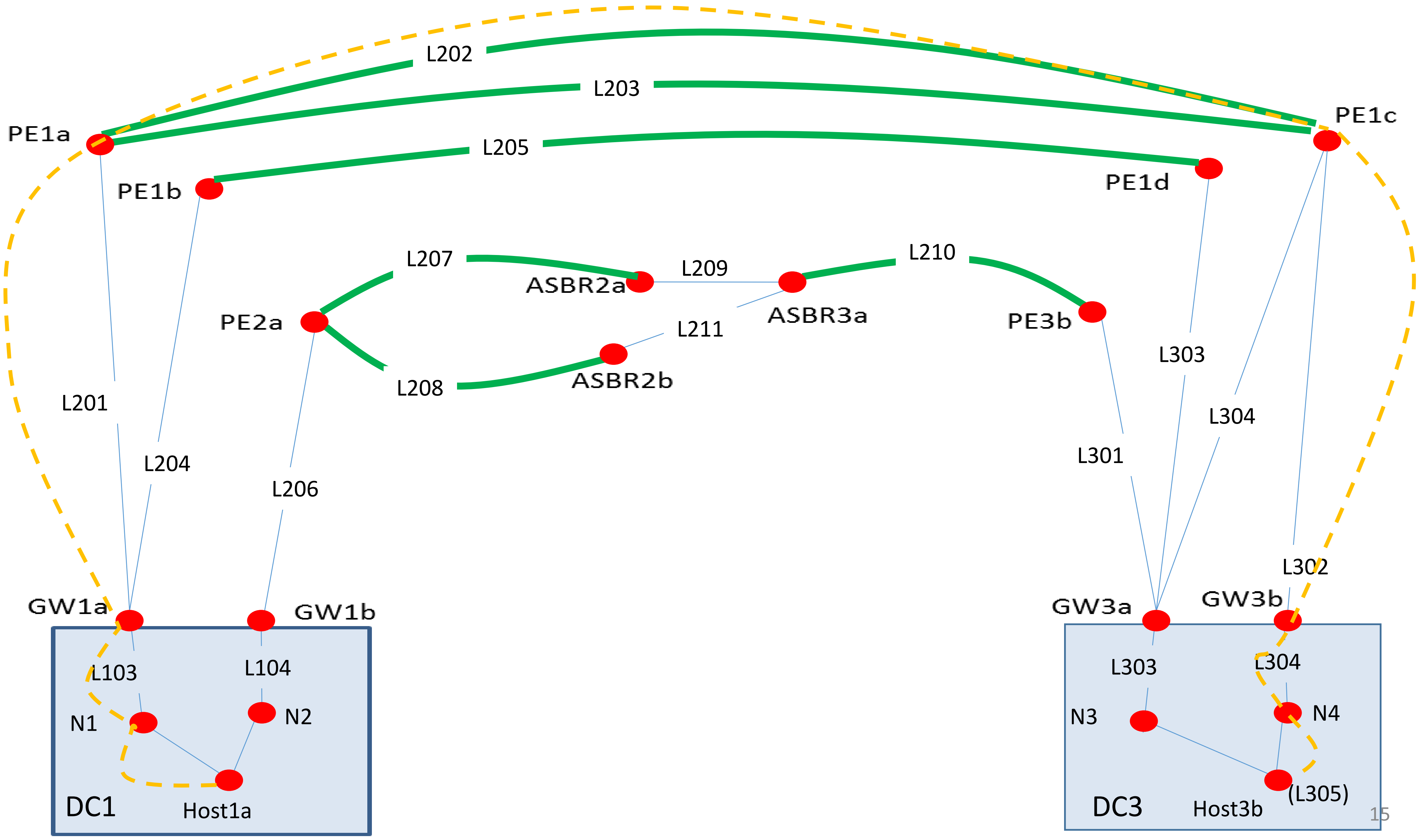
- Path: Source1—GW1b—PE2a—ASBR2a—ASBR3a—PE3b—GW3a—Destination
- Tunnels: PE2a—ASBR2a and ASBR3a—PE3b



Label Stack at Source

- Top Label:
  - **Peer-SID or adjacency-SID identifying link or links to PE2a**. These SIDs are distributed from GW1b to the controller via the EPE extensions of BGP-LS. (This label will get popped by GW1b, which will then send the packet to PE2a.)
- Second Label:
  - **Binding SID for the RSVP-TE LSP to ASBR12 as** advertised by PE2a to the controller. This binding SID is advertised via the PCEP extensions discussed above. (This label will get swapped by PE2a for the label that the LSP's next hop has assigned to the LSP.)
- Third Label:
  - **Peer-SID or adjacency-SID identifying link or links to ASBR3a**, as advertised to the controller by ASBR2a using the BGP-LS EPE extensions. (This label gets popped by ASBR2a, which then sends the packet to ASBR3a.)
- Fourth Label:
  - **Binding SID advertised by ASBR3a for the RSVP-TE LSP to PE3b**. This binding SID is advertised via the PCEP extensions discussed above. ASBR3a treats this label just like PE2a treated the second label above.
- Fifth label:
  - **Peer-SID or adjacency-SID identifying link or links to GW3a**, as advertised to the controller by ASBR3a using the BGP-LS EPE extensions. ASBR3a pops this label and sends the packet to GW3a.
- Sixth Label:
  - **Prefix-SID or other label identifying the Destination** advertised in a tunnel encaps attribute by GW3a. (This can be omitted if GW3a is happy to accept IP packets, or prefers a VXLAN tunnel for example. That would be indicated through the tunnel encapsulation attribute of course.)

Worked Example



Full Source Route Imposed at Source

- Suppose it is desired for a packet from Host1a to travel to Host 3b via the following source route:
  - Host1a→N1→GW1a→PE1a→(RSVP-TE LSP)→ PE1c→GW3b→N4→Host3b
- Host1a would impose the following label stack (with the first label representing the top of stack), and then send the packet to N1:
  - L103, L201, L202, L302, L304, L305

But What If Source DC Does Not Have Full Visibility?

- Suppose the destination DC does not export its topology
- The source host or the path computation entity will still know:
  - The GW(s) through which the destination can be reached
  - The SID to use for the destination prefix.
- Suppose we want a packet to follow the source route:

  Host1a→N1→GW1a→PE1a→(RSVP-TE LSP)→PE1c→GW3b→…→Host3b
- The label stack imposed at the source host would be:

  L103, L201, L202, L302, L305
- When the packet reaches the GW of the destination DC, it can either simply forward the packet using weighted ECMP to Host 3b, or it can insert additional labels to steer the packet

Maybe DC1 Only has Reachability Information

- The source DC (or the path computation entity)
  - May know the location of the destination in the destination DC
  - May know the GWs to the destination DC that provide reachability to the destination
  - But may have no view of the backbone network
- The packet is forwarded like 'per-domain path computation'
- At the source the label stack is:

  L101, L103, L302, L305
- The source GW determines the PE to use to enter the backbone using the BGP preferred route to the destination GW
- The first PE it has a label stack just identifying the destination GW and host (L302, L305)
  - It uses information it has about the backbone network topology and available LSPs to select an LSP tunnel, impose the tunnel label, and forward the packet
- When the packet reaches the end of the LSP tunnel, it is processed as before

Things To Do

- Is this work in the right place?
  - We think so – it shows how to use BGP and BGP-LS
- Should other WGs care?
  - Yes
    - IDR oversight, but no change to protocol
    - TEAS to be aware of traffic engineering
    - PCE for conveying binding SID
    - SPRING because this is segment routing
  - But we wanted to get BESS opinions first
- So, reviews and comments please
  - Decide whether BESS wants to pursue this