BGP-Prefix Segment in large-scale data centers

draft-filsfils-spring-segment-routing-msdc-00

Clarence Filsfils – Cisco Systems
Stefano Previdi – Cisco Systems
Jon Mitchell – Microsoft Corporation
Benjamin Black – Microsoft Corporation
Dmitry Afanasiev – Yandex
Saikat Ray – Cisco Systems
Keyur Patel – Cisco Systems

IETF91, November 2013, Honolulu, US

Purpose

- Segment routing use case in BGP+MPLS based MSDC
  - No new requirements on Spring. BGP extensions are presented in IDR
- Illustration of

  Prefix-SID

  Egress Peer Engineering

  Capacity optimization

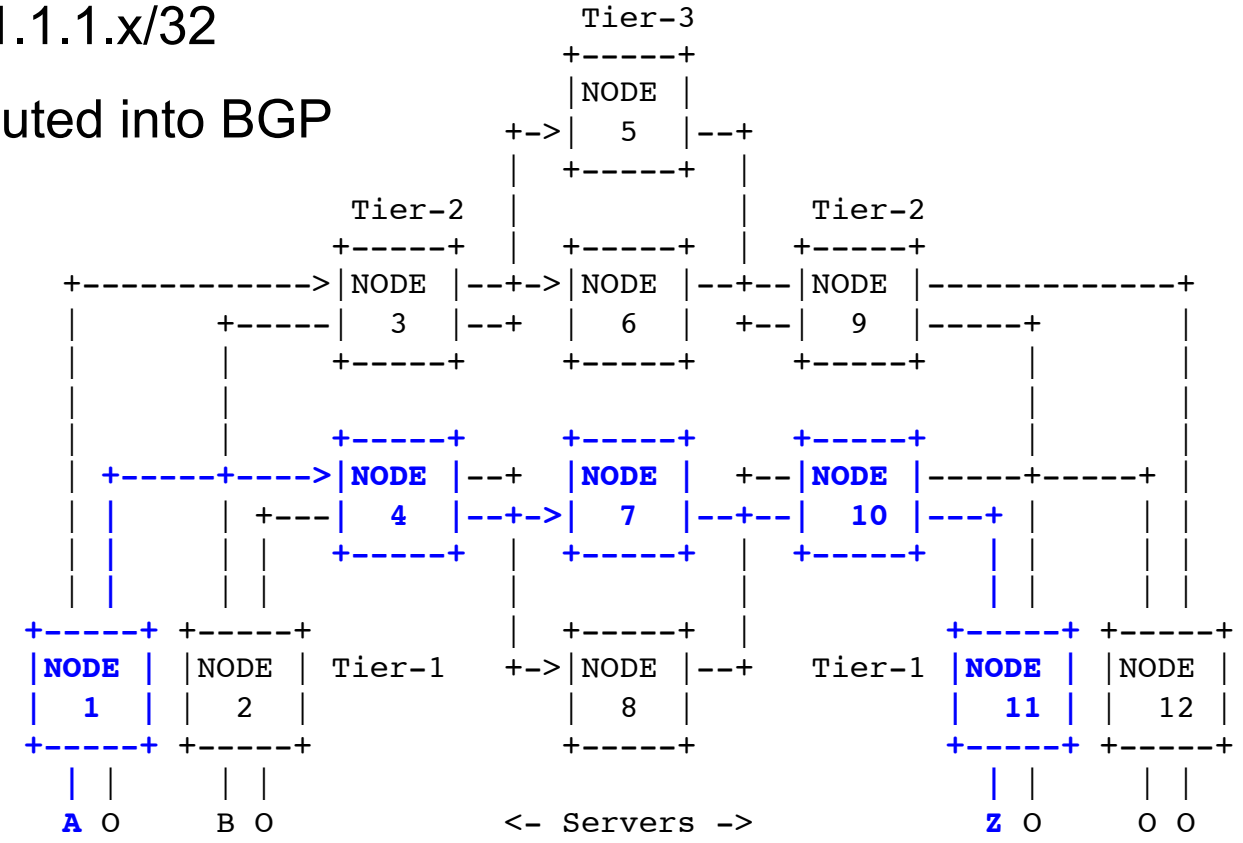
  Incremental deployment

  Anycast

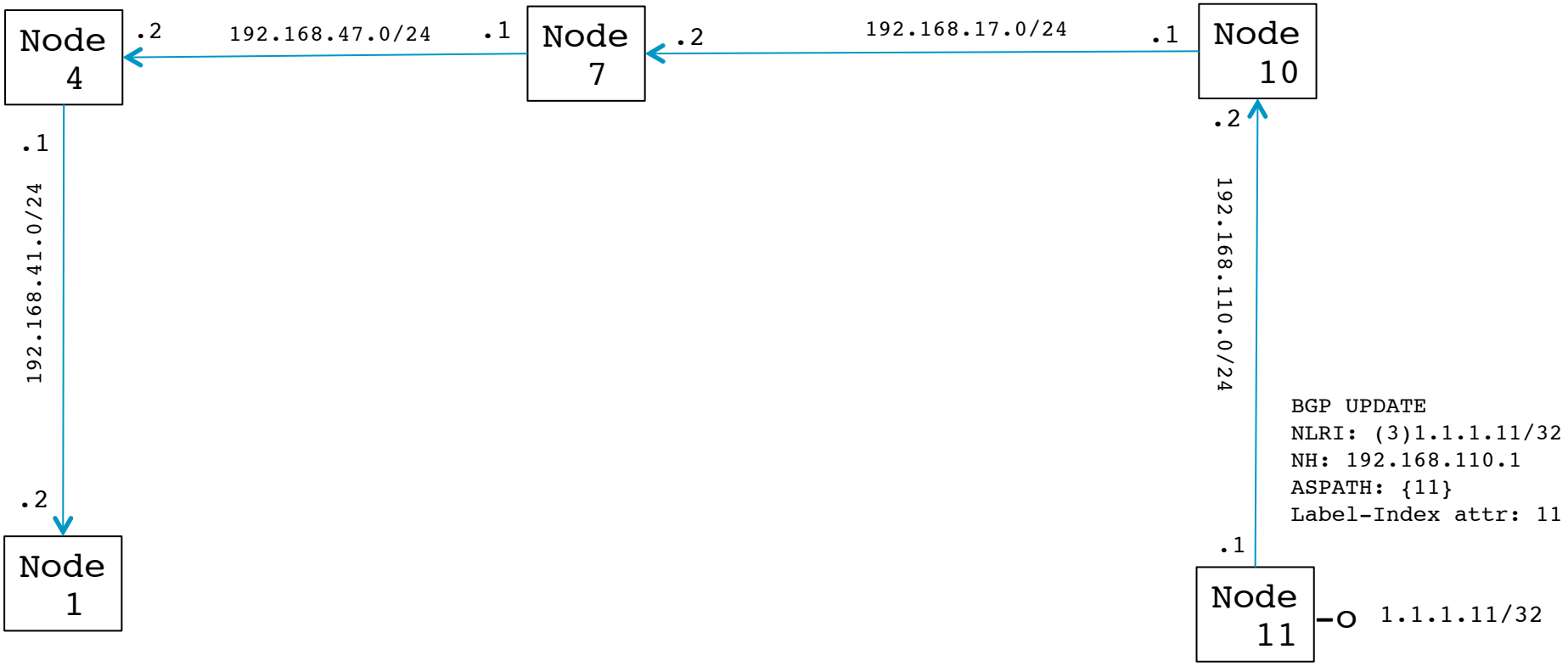
Reference topology

- Node ‘x’ has ASN ‘x’
- BGP IPvX labeled-unicast sessions (3107) between directly connected nodes
- Node ‘x’ has loopback 1.1.1.x/32
- Loopbacks are redistributed into BGP and advertised
- SRGB: [16000, 23999]
- Label index for 1.1.1.x/32 is ‘x’

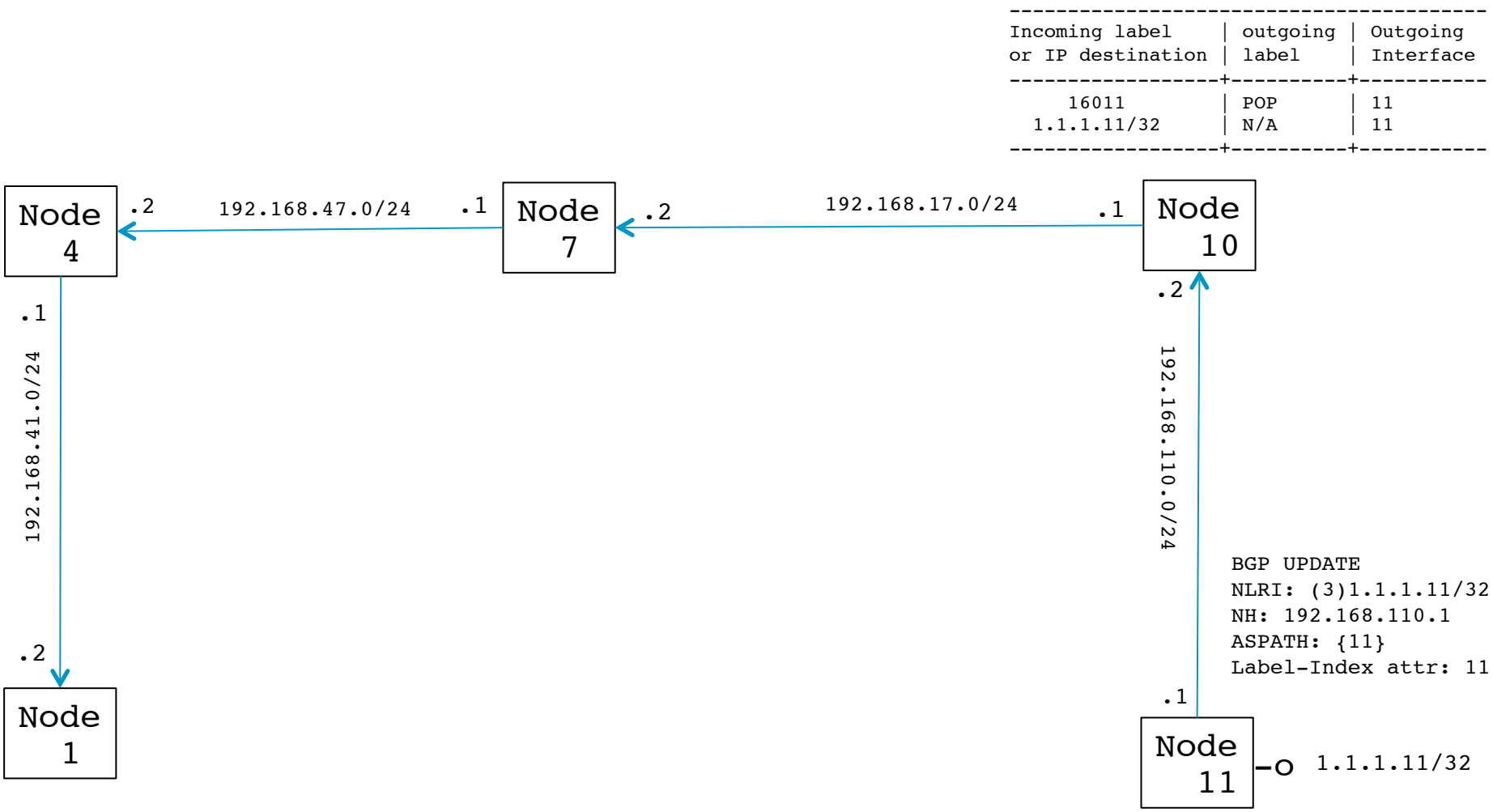
- Tier-2 and Tier-3 nodes: MPLS forwarding
- Tier-1 nodes: IP2MPLS or MPLS2MPLS forwarding



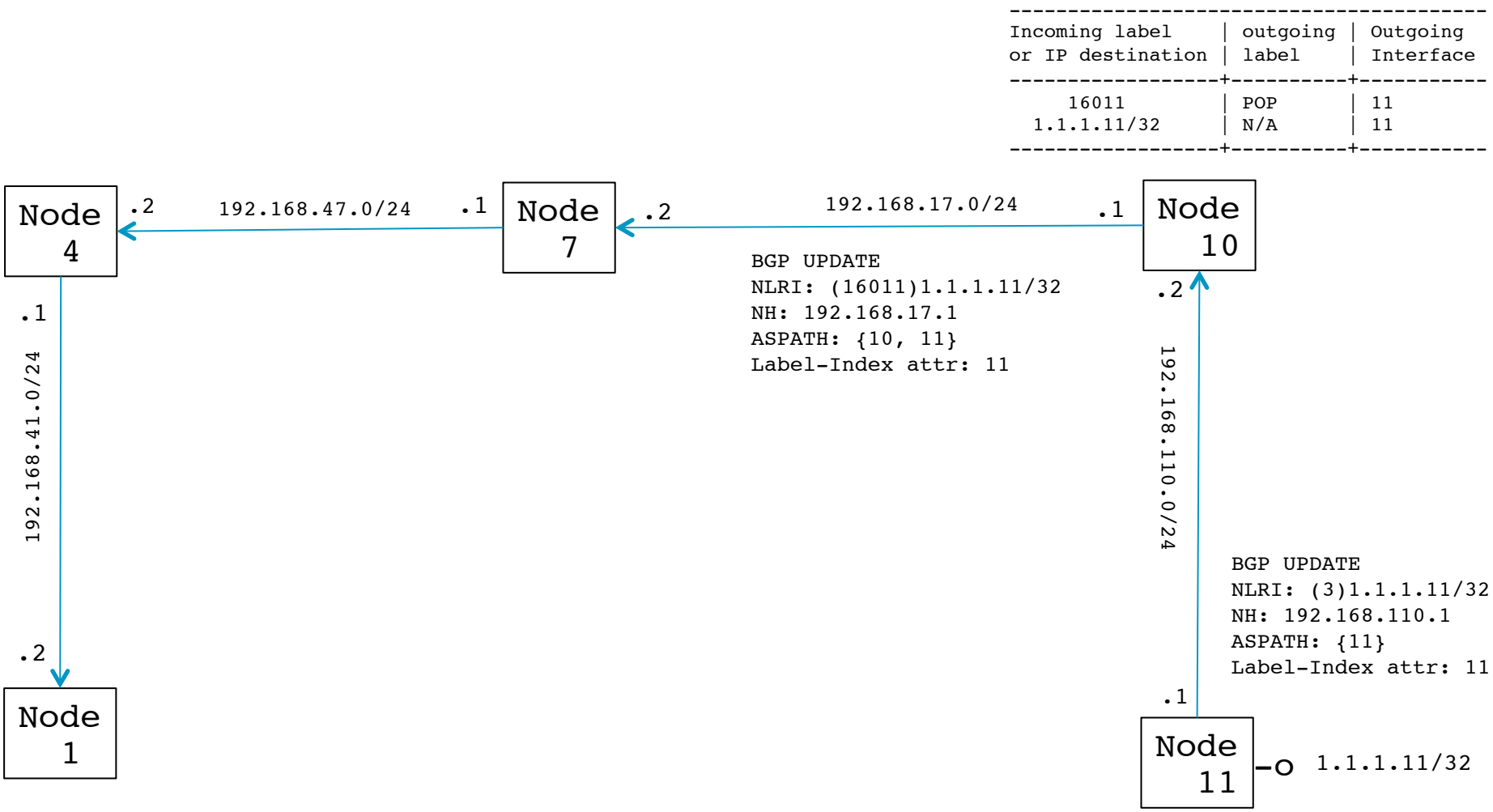
BGP Prefix SID: Control and dataplane



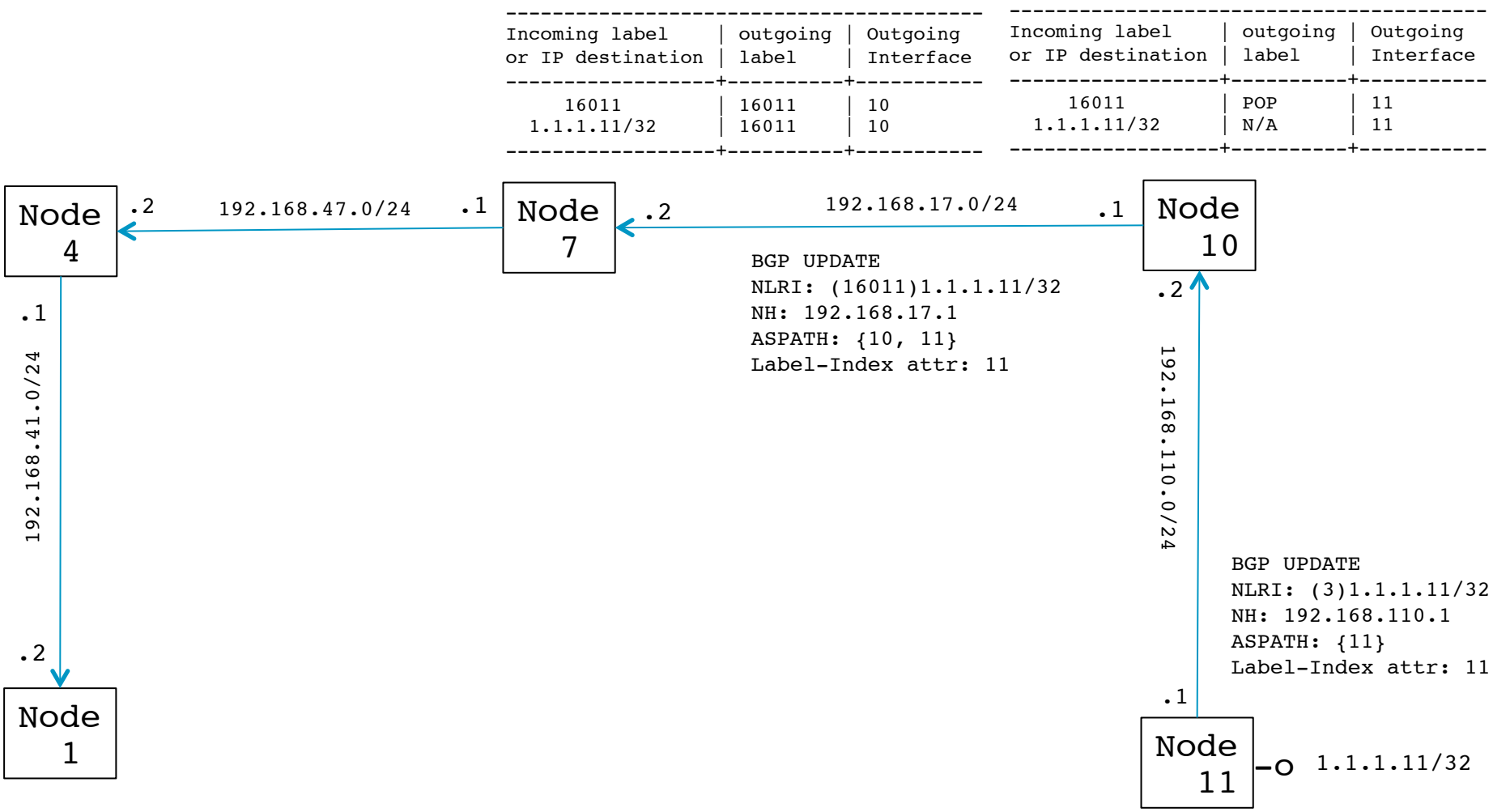
BGP Prefix SID: Control and dataplane



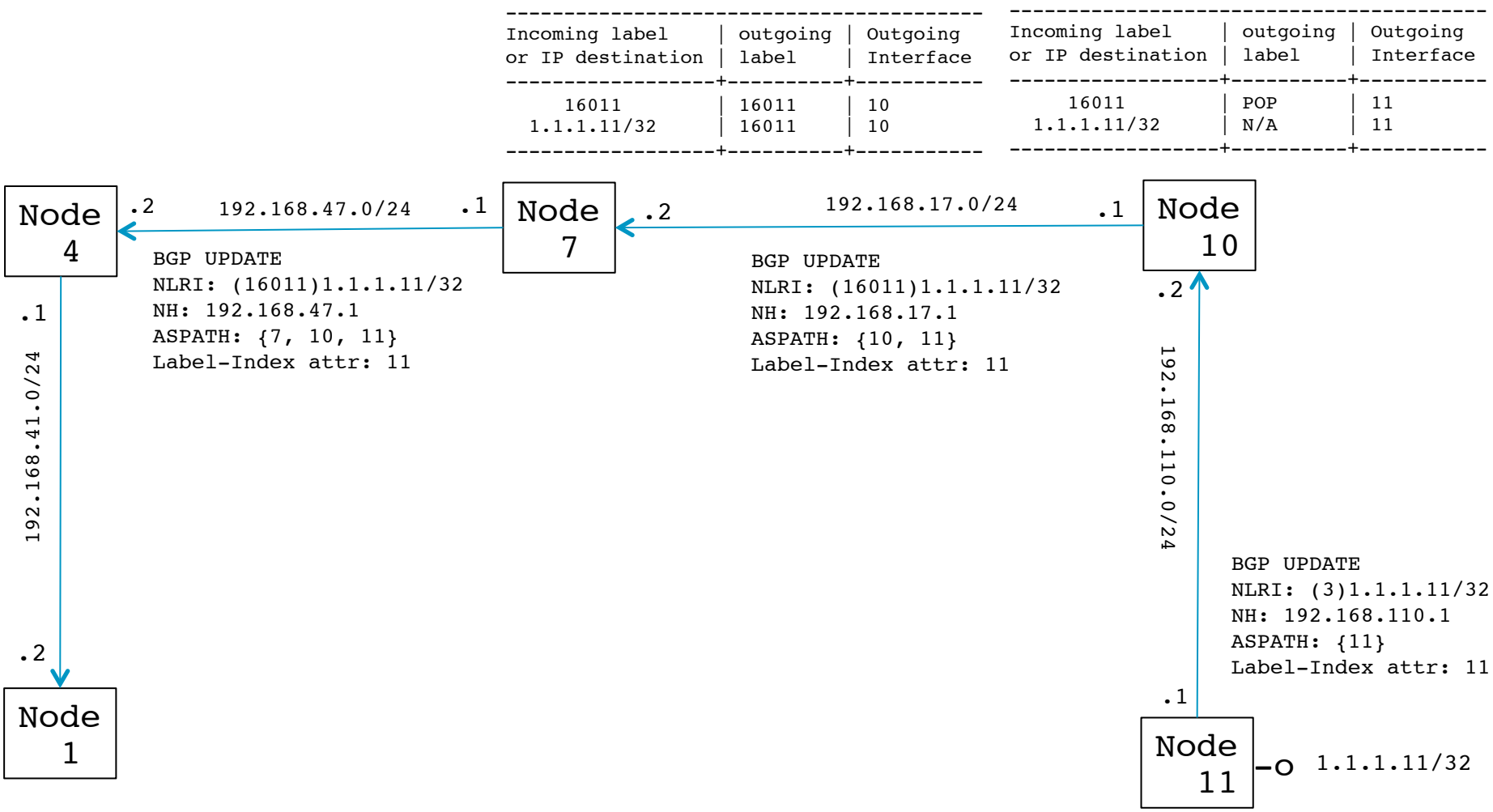
BGP Prefix SID: Control and dataplane



BGP Prefix SID: Control and dataplane



BGP Prefix SID: Control and dataplane

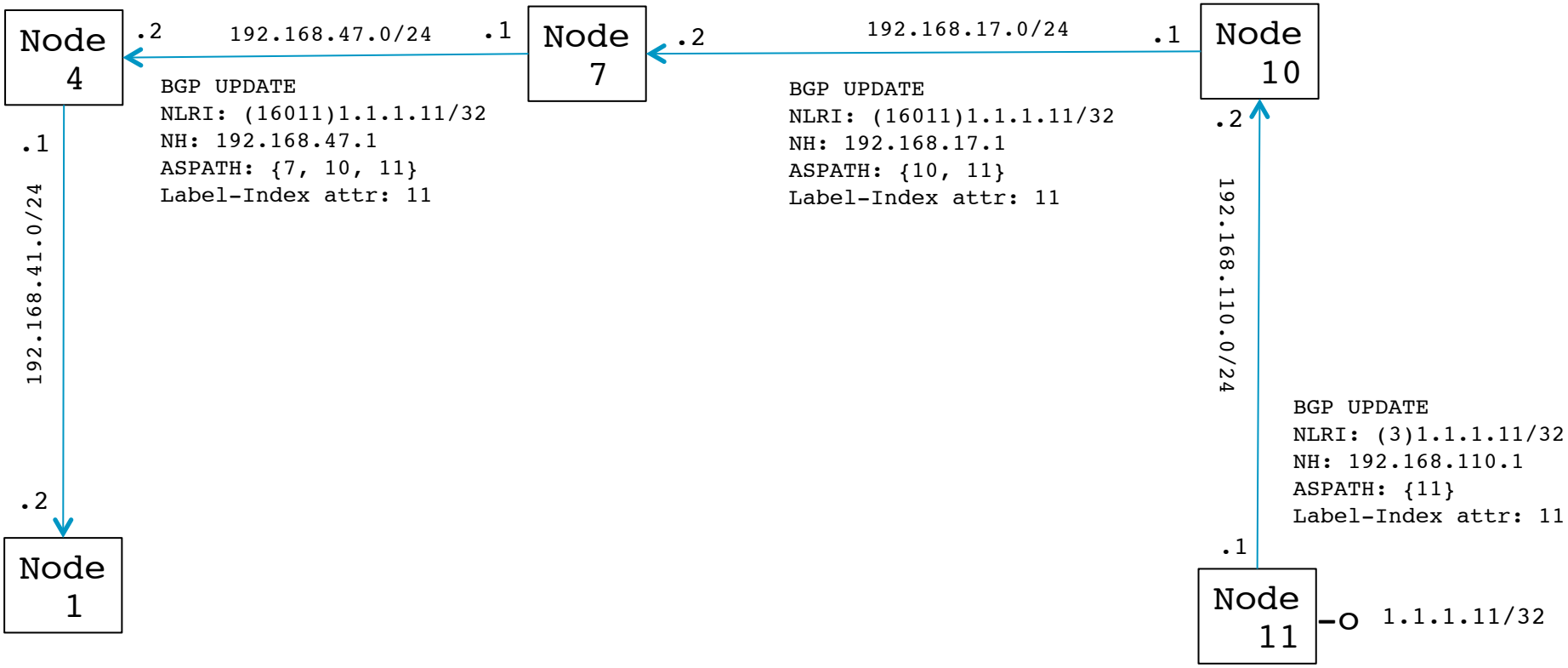


BGP Prefix SID: Control and dataplane

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	ECMP{7, 8} ECMP{7, 8}

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	10 10

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	POP N/A	11 11

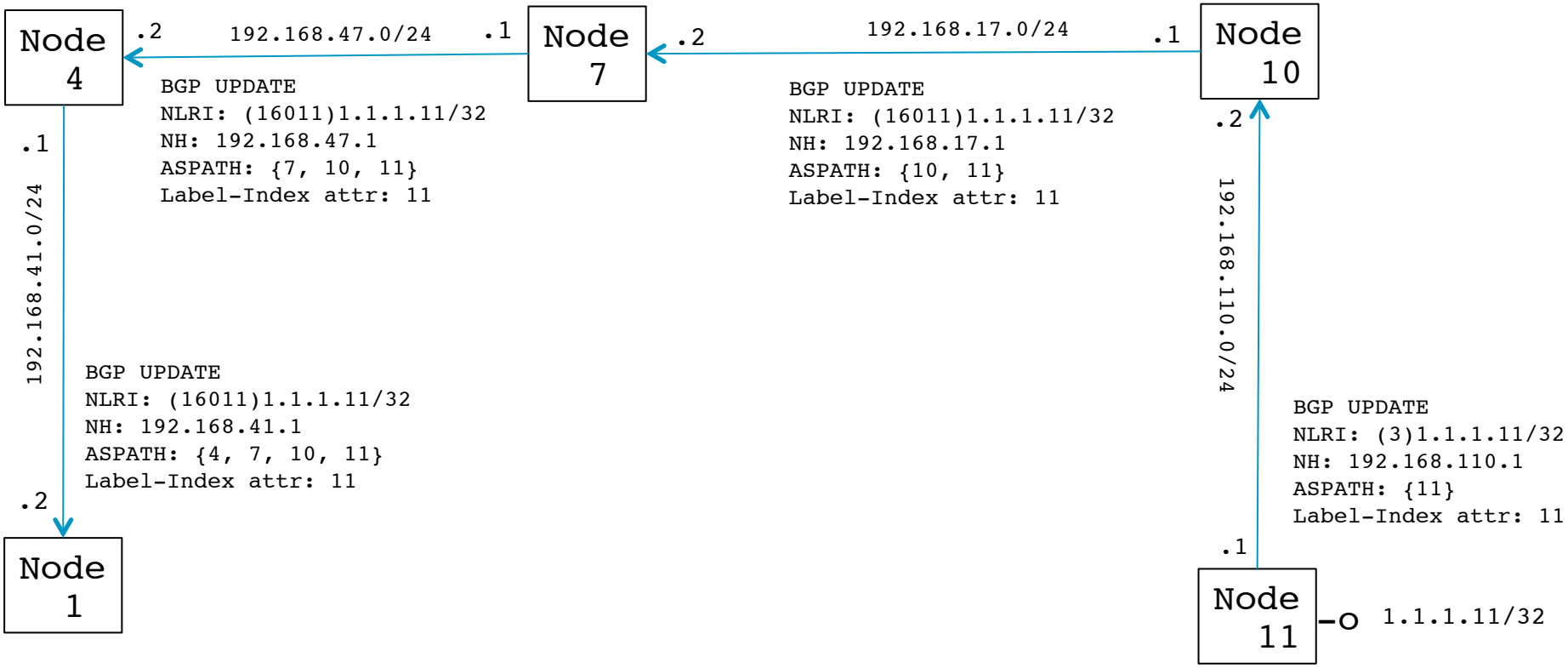


BGP Prefix SID: Control and dataplane

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	ECMP{7, 8} ECMP{7, 8}

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	10 10

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	POP N/A	11 11

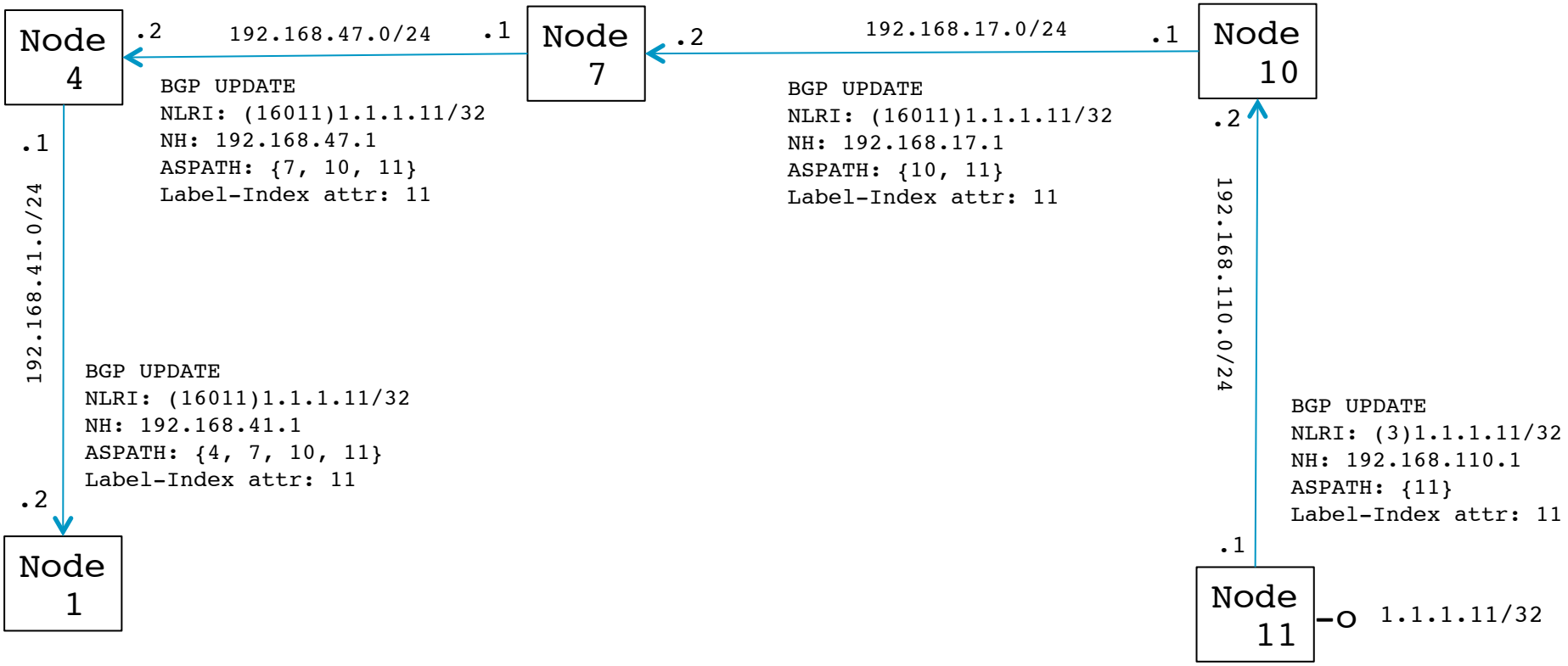


BGP Prefix SID: Control and dataplane

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	ECMP{7, 8} ECMP{7, 8}

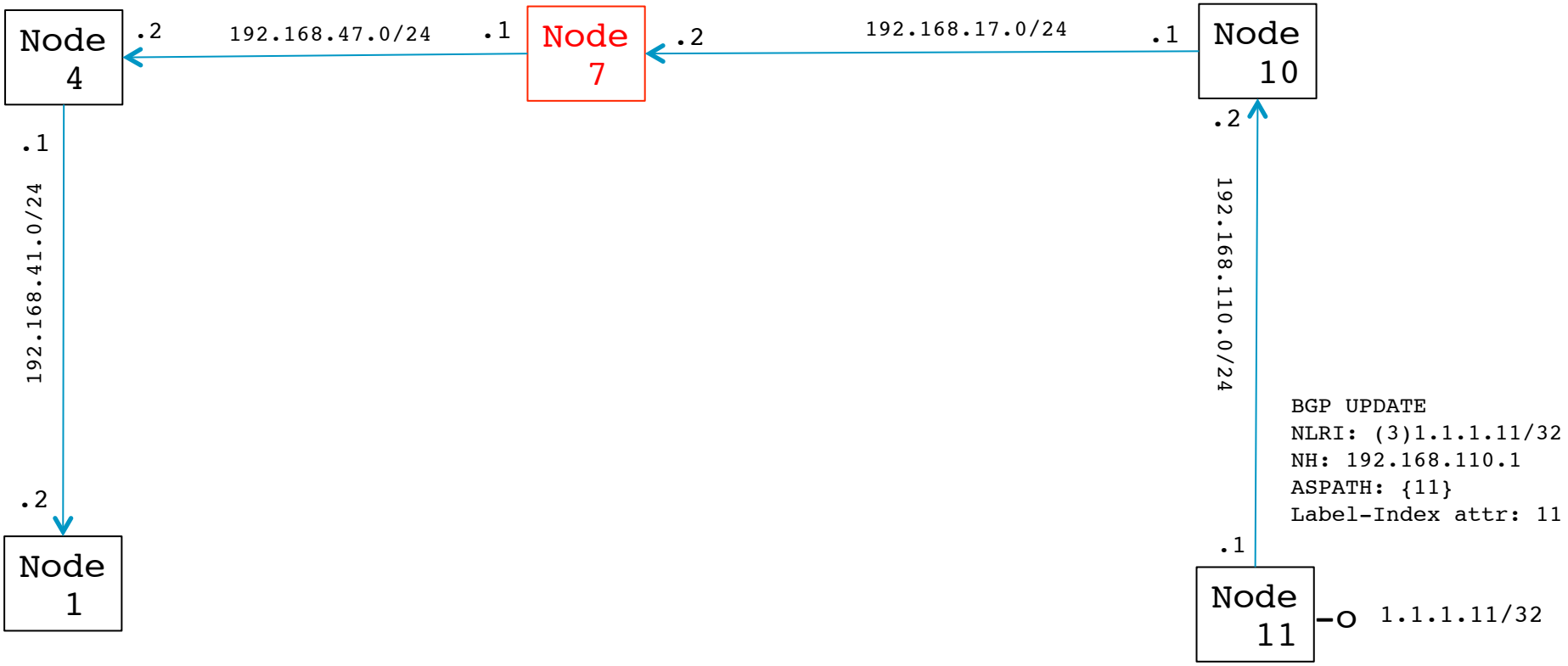
Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	10 10

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	POP N/A	11 11

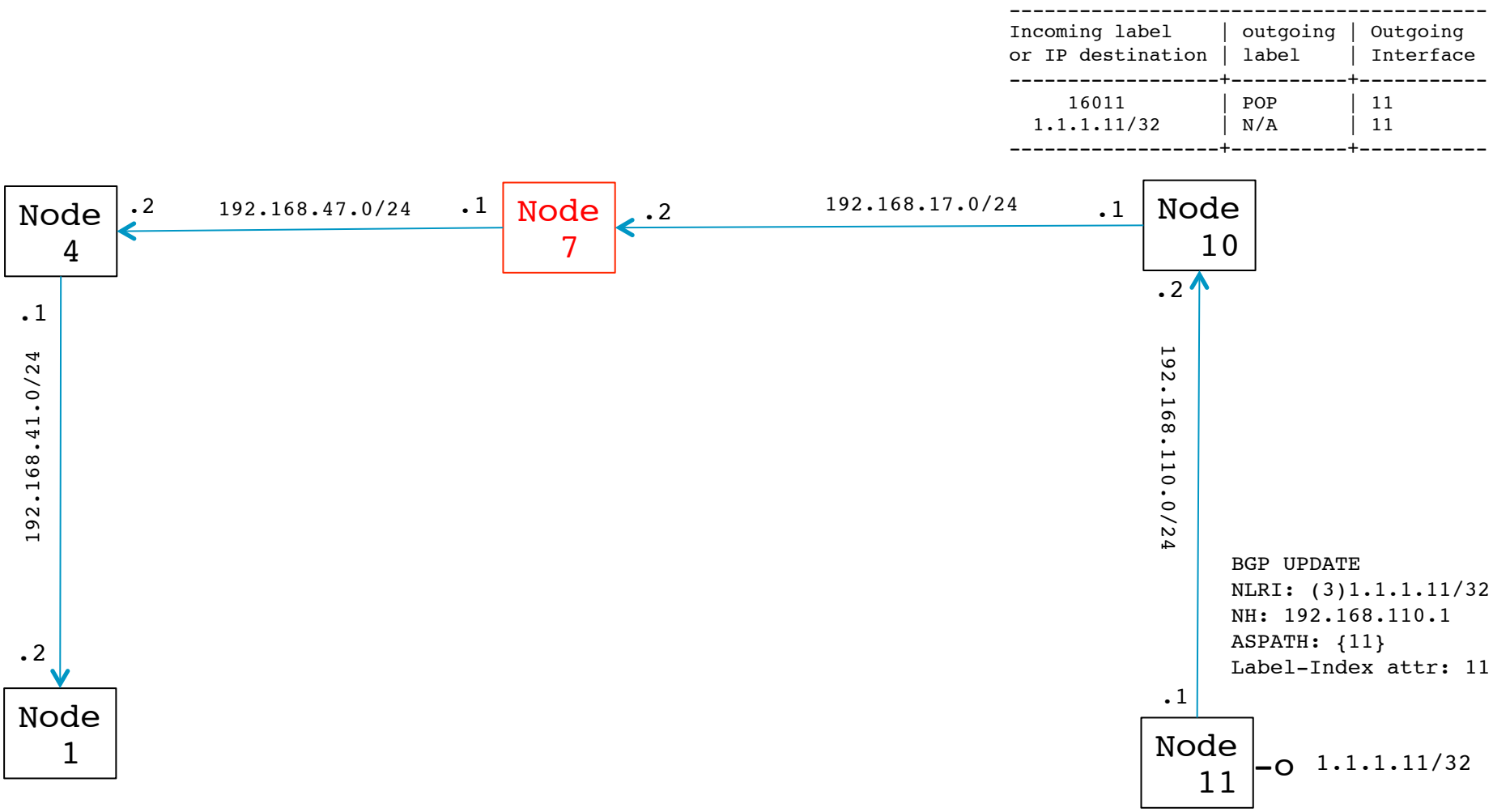


Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	ECMP{3, 4} ECMP{3, 4}

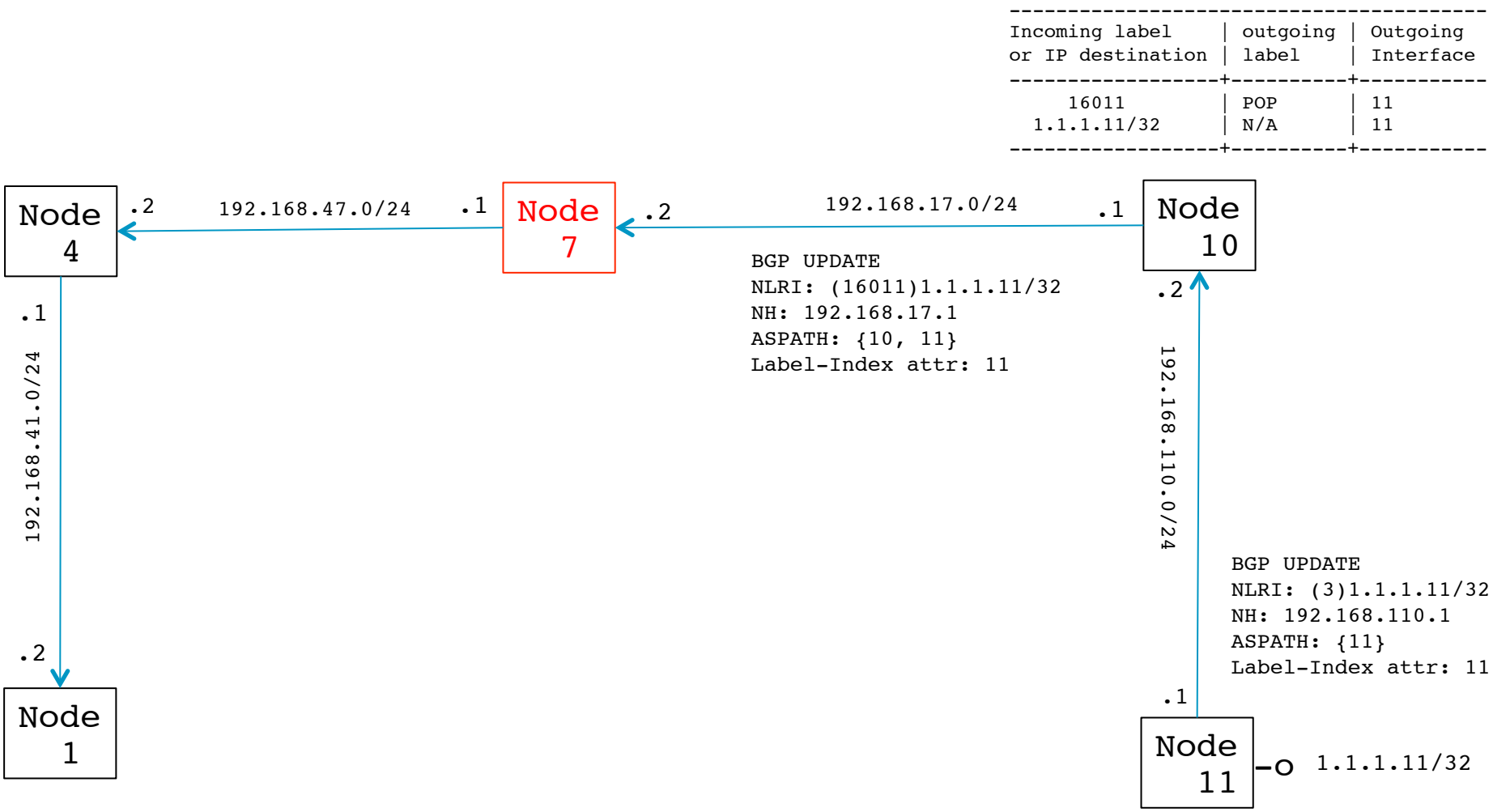
BGP Prefix SID: Non-SR node in the middle



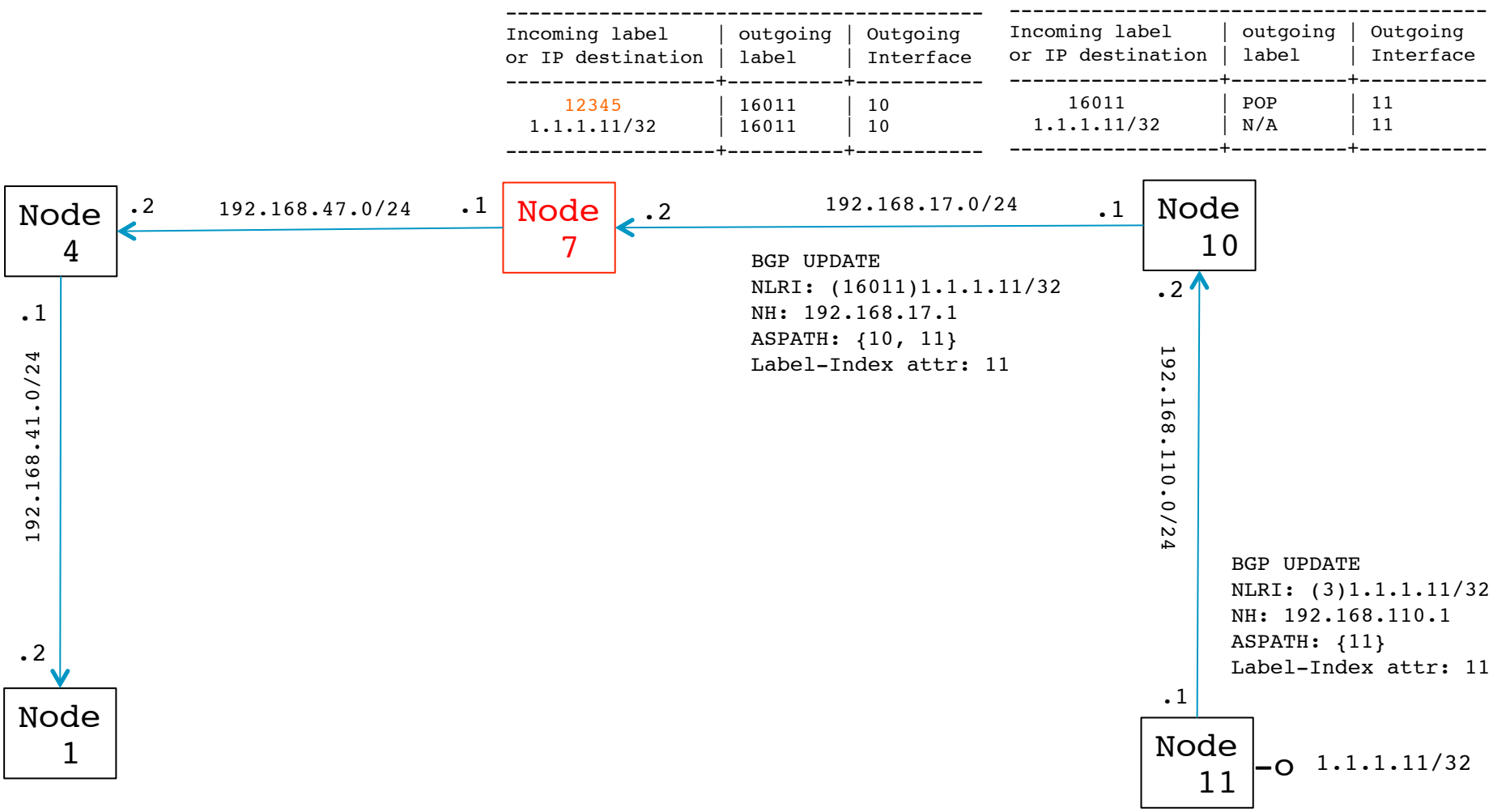
BGP Prefix SID: Non-SR node in the middle



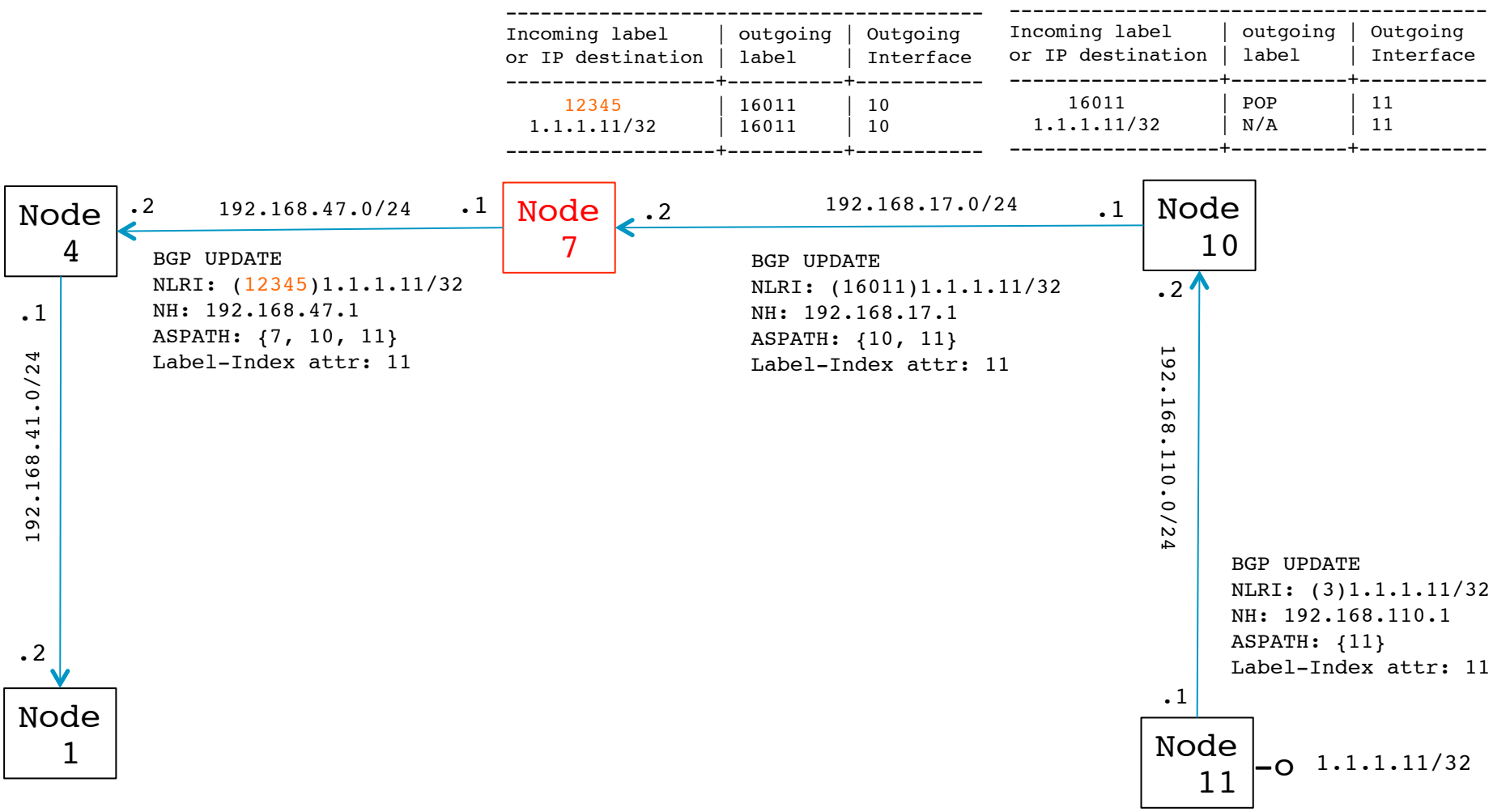
BGP Prefix SID: Non-SR node in the middle



BGP Prefix SID: Non-SR node in the middle



BGP Prefix SID: Non-SR node in the middle

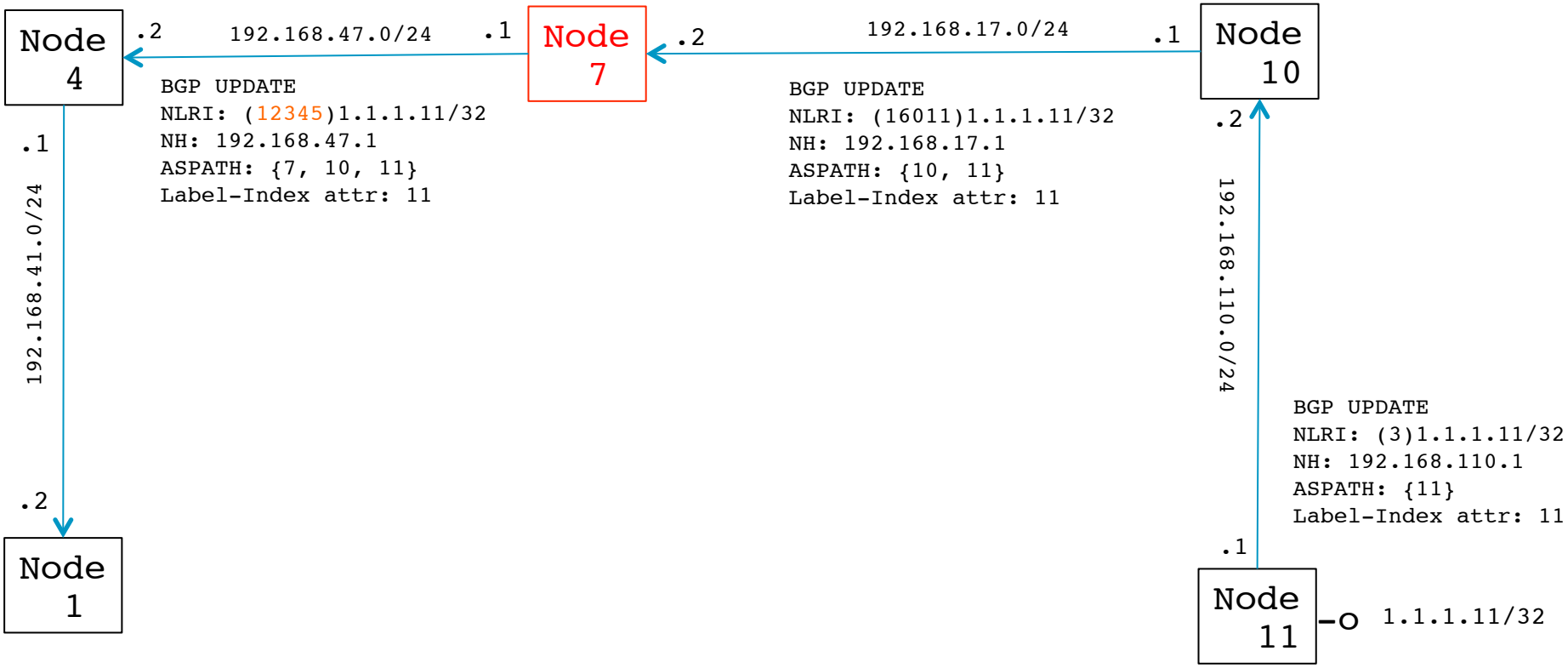


BGP Prefix SID: Non-SR node in the middle

Incoming label or IP destination	outgoing label	Outgoing Interface
16011	12345	ECMP{7, 8}
1.1.1.11/32	12345	ECMP{7, 8}

Incoming label or IP destination	outgoing label	Outgoing Interface
12345	16011	10
1.1.1.11/32	16011	10

Incoming label or IP destination	outgoing label	Outgoing Interface
16011	POP	11
1.1.1.11/32	N/A	11

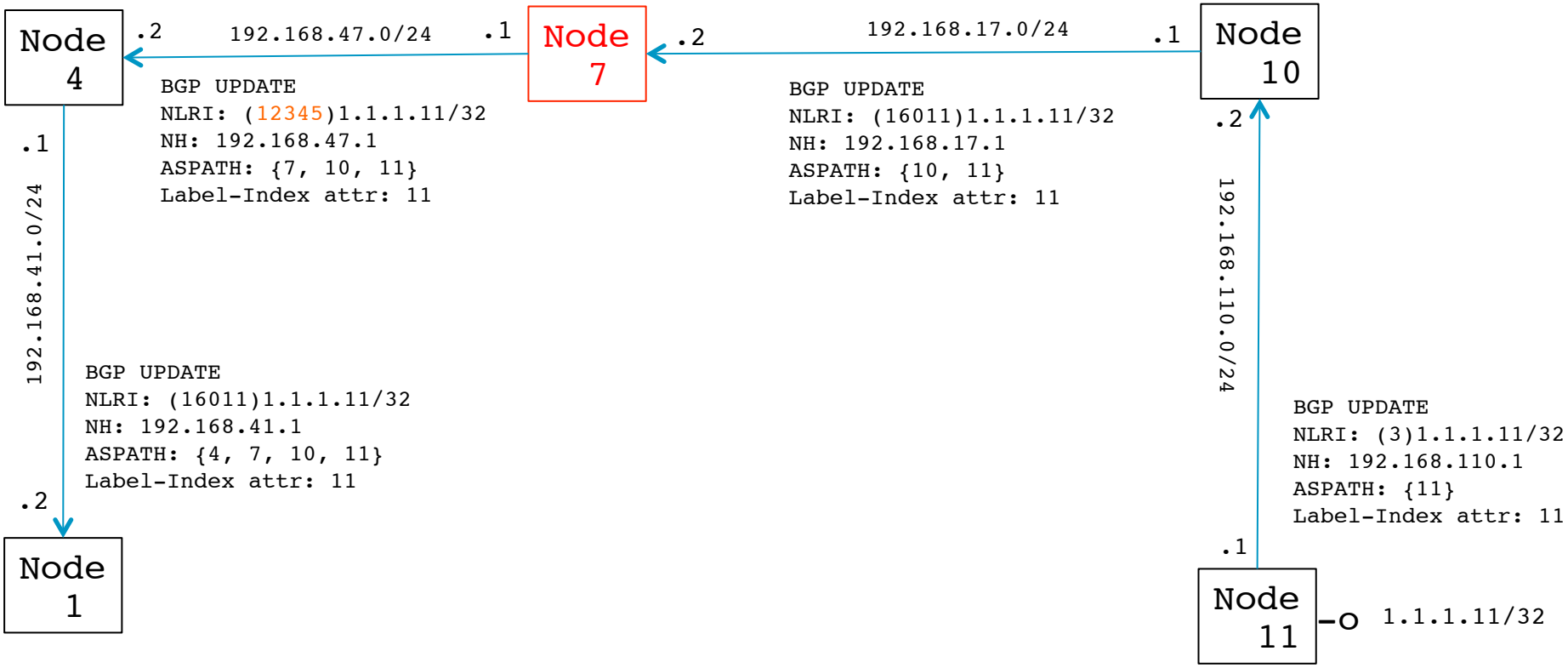


BGP Prefix SID: Non-SR node in the middle

Incoming label or IP destination	outgoing label	Outgoing Interface
16011	12345	ECMP{7, 8}
1.1.1.11/32	12345	ECMP{7, 8}

Incoming label or IP destination	outgoing label	Outgoing Interface
12345	16011	10
1.1.1.11/32	16011	10

Incoming label or IP destination	outgoing label	Outgoing Interface
16011	POP	11
1.1.1.11/32	N/A	11

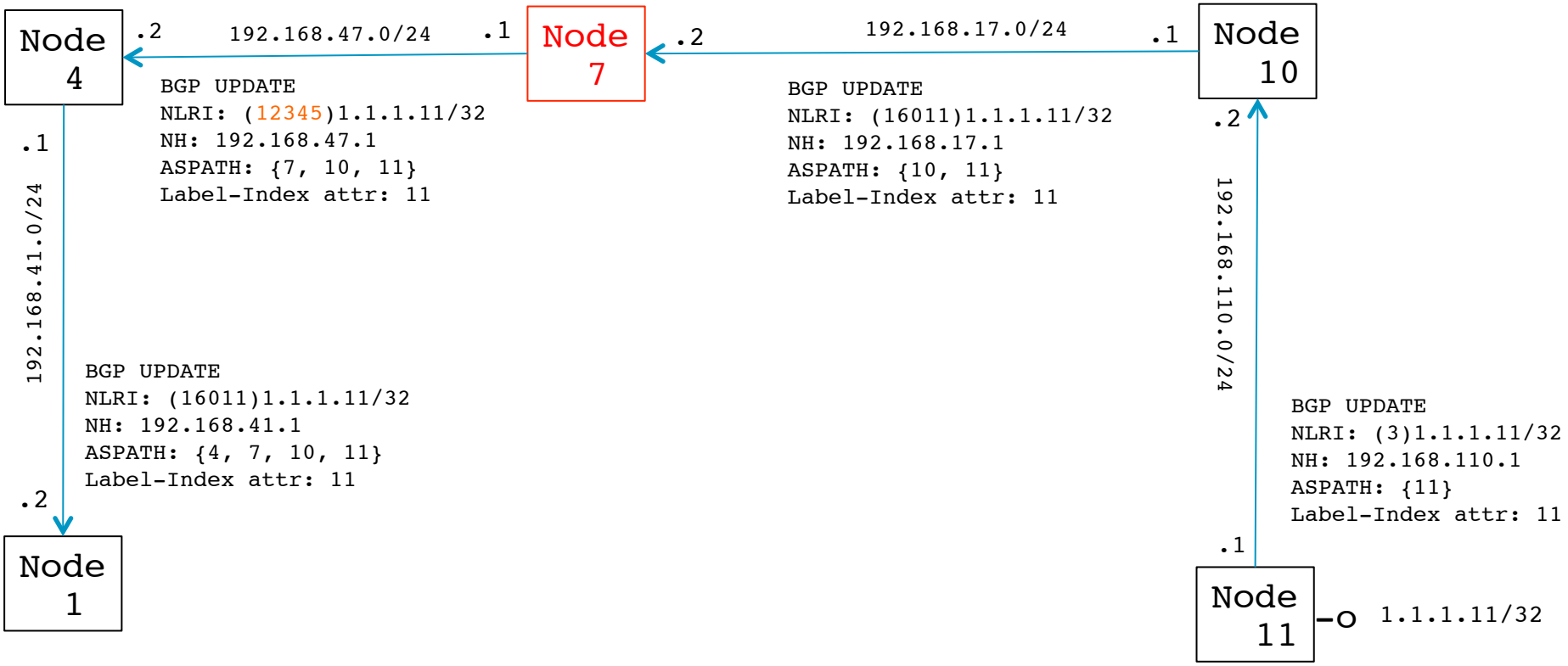


BGP Prefix SID: Non-SR node in the middle

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	12345 12345	ECMP{7, 8} ECMP{7, 8}

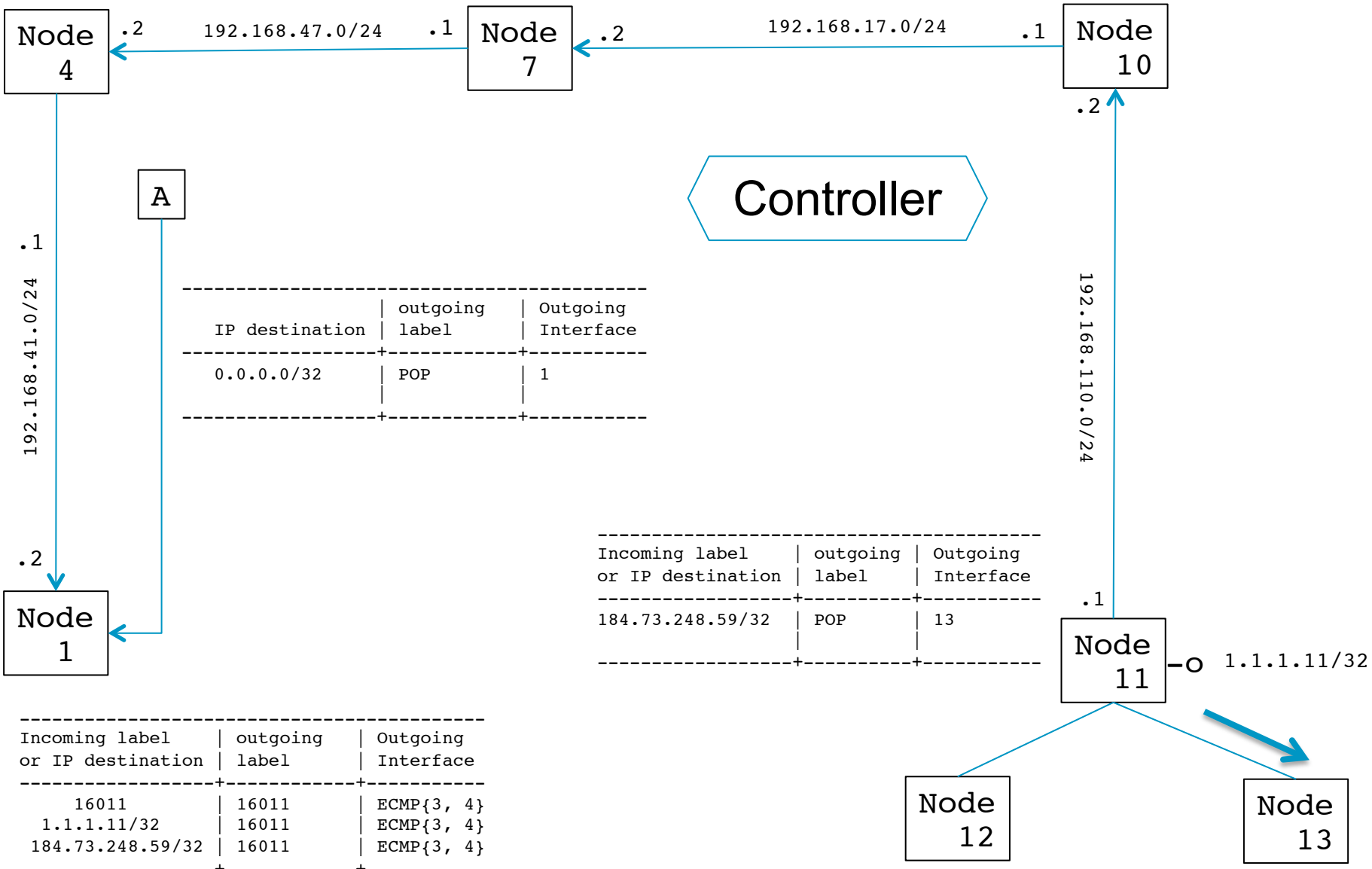
Incoming label or IP destination	outgoing label	Outgoing Interface
12345 1.1.1.11/32	16011 16011	10 10

Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	POP N/A	11 11

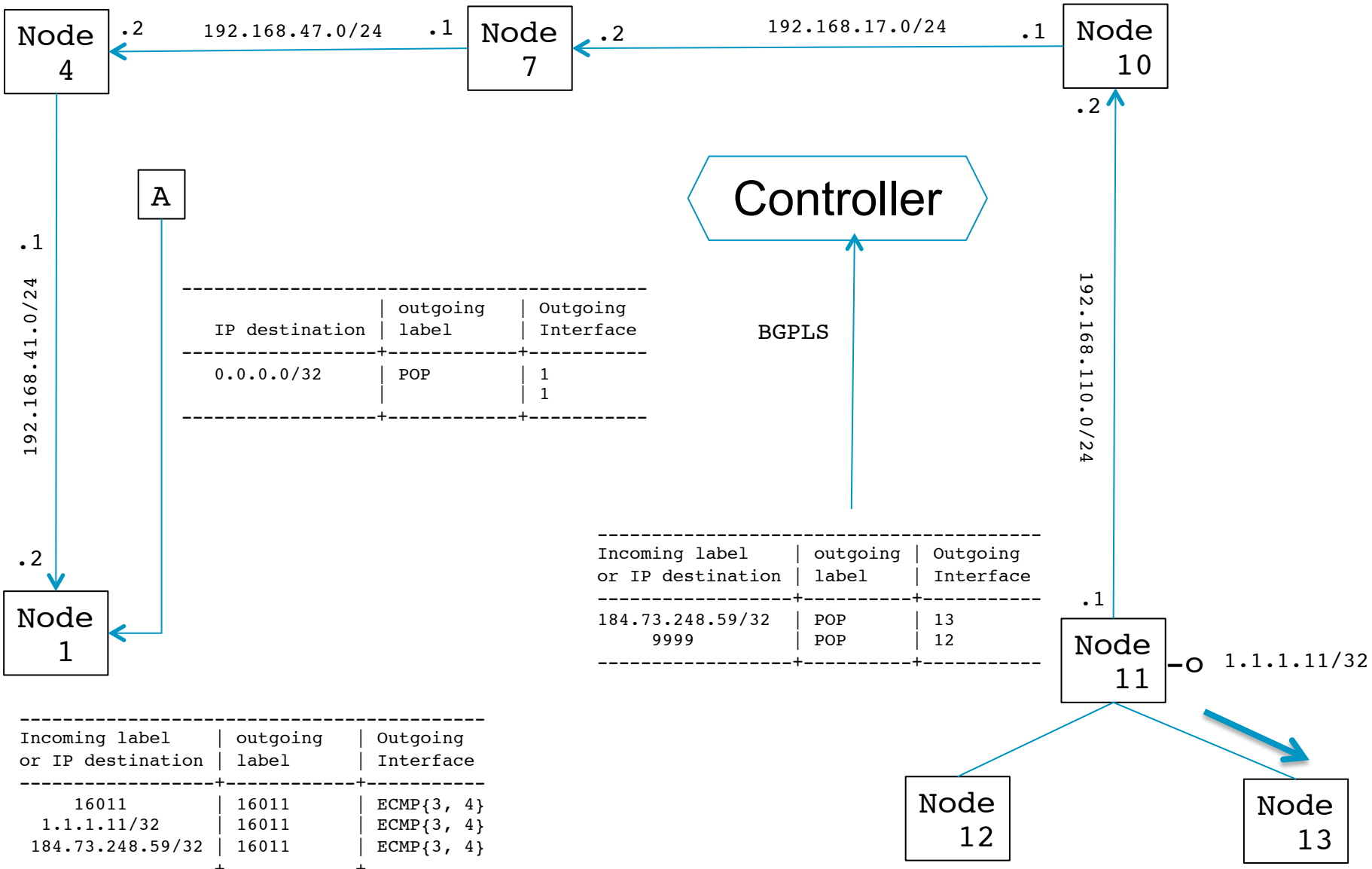


Incoming label or IP destination	outgoing label	Outgoing Interface
16011 1.1.1.11/32	16011 16011	ECMP{3, 4} ECMP{3, 4}

BGP Egress Peer Engineering



BGP Egress Peer Engineering

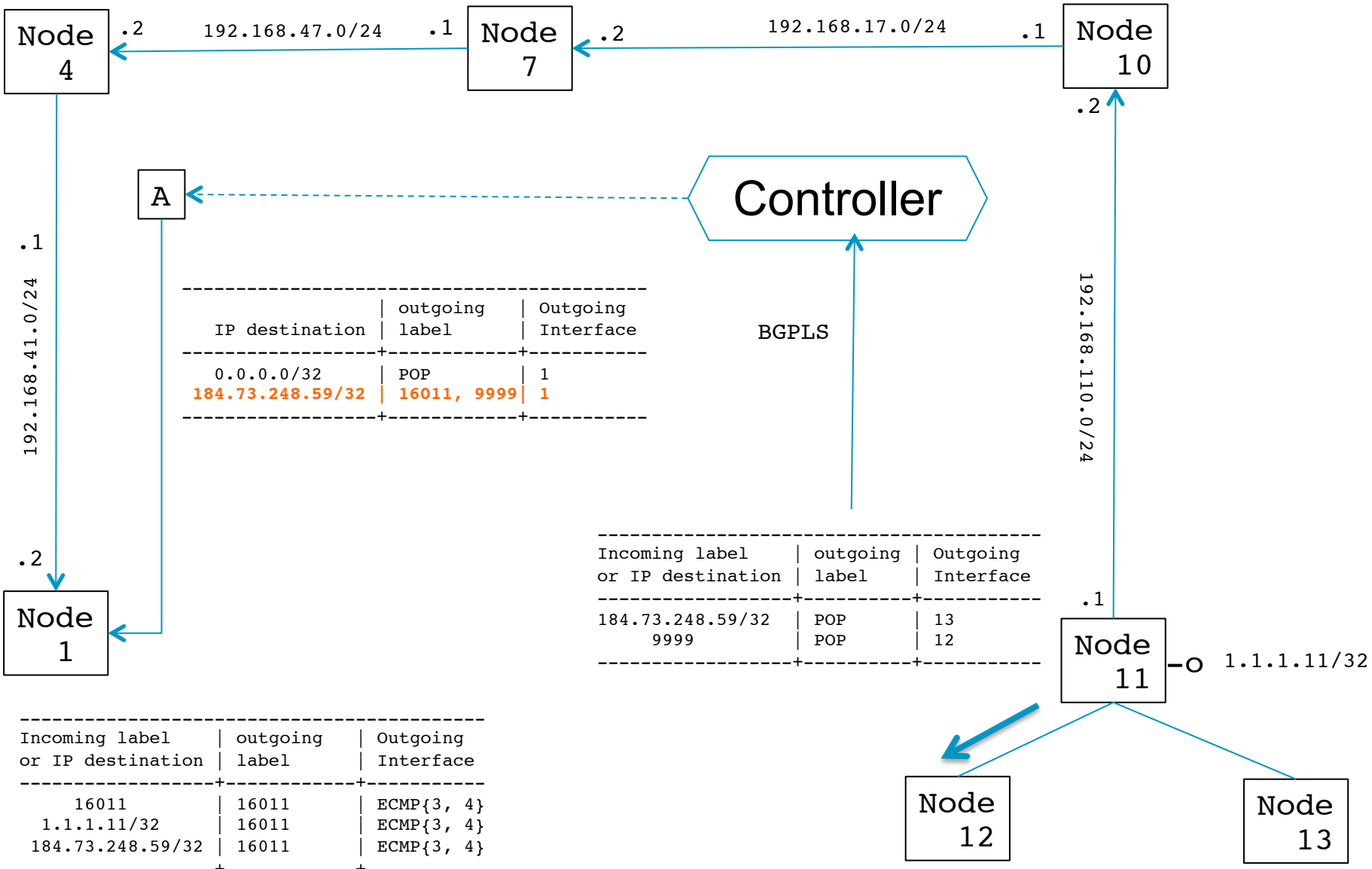


IP destination	outgoing label	Outgoing Interface
0.0.0.0/32	POP	1 1

Incoming label or IP destination	outgoing label	Outgoing Interface
184.73.248.59/32	POP	13
9999	POP	12

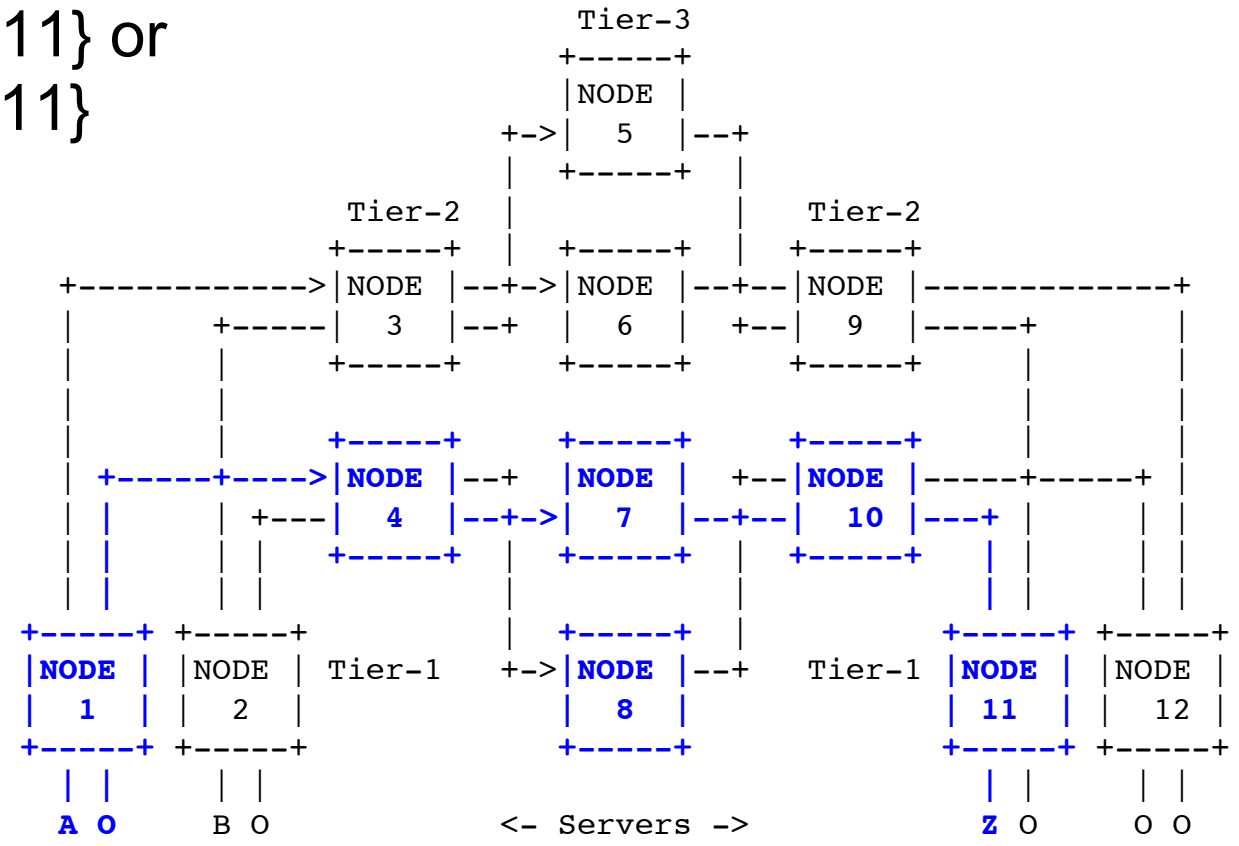
Incoming label or IP destination	outgoing label	Outgoing Interface
16011	16011	ECMP{3, 4}
1.1.1.11/32	16011	ECMP{3, 4}
184.73.248.59/32	16011	ECMP{3, 4}

BGP Egress Peer Engineering



Capacity optimization

- Steer packets using label stack at ingress
- Normal (min cost) path from A to Z is via node 5
- Node 5 is congested
- A uses {16004, 16011} or
  {16008, 16011}
- Packets traverse via lower paths



Other considerations

- Anycast: Assigning the same label-index to an anycast loopback interface address (residing on multiple nodes) achieves load-balanced reachability (to any of the nodes).
- Minimizing the FIB table: The user can only use MPLS forwarding in Tier-2 and Tier-3 nodes. In such cases, the IP part of the FIB need not be programmed.
- Operational simplicity:

  RSVP/LDP is no longer needed

  Use of same SRGB on all nodes leads to the same label value for a given prefix on all nodes. This drastically simplifies troubleshooting.

  A controller device can uniformly program label stack to hosts.

Summary

- This document presents an use case for segment routing in BGP+MPLS based MSDC
- The design illustrated retains multiple benefits

  Bandwidth and traffic patterns

  Capex/opex minimization

  Traffic engineering

  Fast routing convergence

  Anycast/load-balancing

  Operationally simplified MPLS dataplane

  Egress peer engineering

  Capacity optimization

  Incremental deployment

Questions/Comments?