Microsoft

Unified Forwarding with Segment Routing

Mohan Nanduri

August 3, 2014

Some Challenges…

- Control plane scale and complexity
- Many protocol and features, also many bugs
- Forwarding table size and FIB capacity
- Programmatic control over BGP policy
- Hyper scale datacenters
  - Many encapsulations
  - Lot of links
  - Commodity hardware
  - IP address mobility
- Inter-DC vs External workloads
  - Latency-sensitive vs bulk
  - Scheduled vs unscheduled

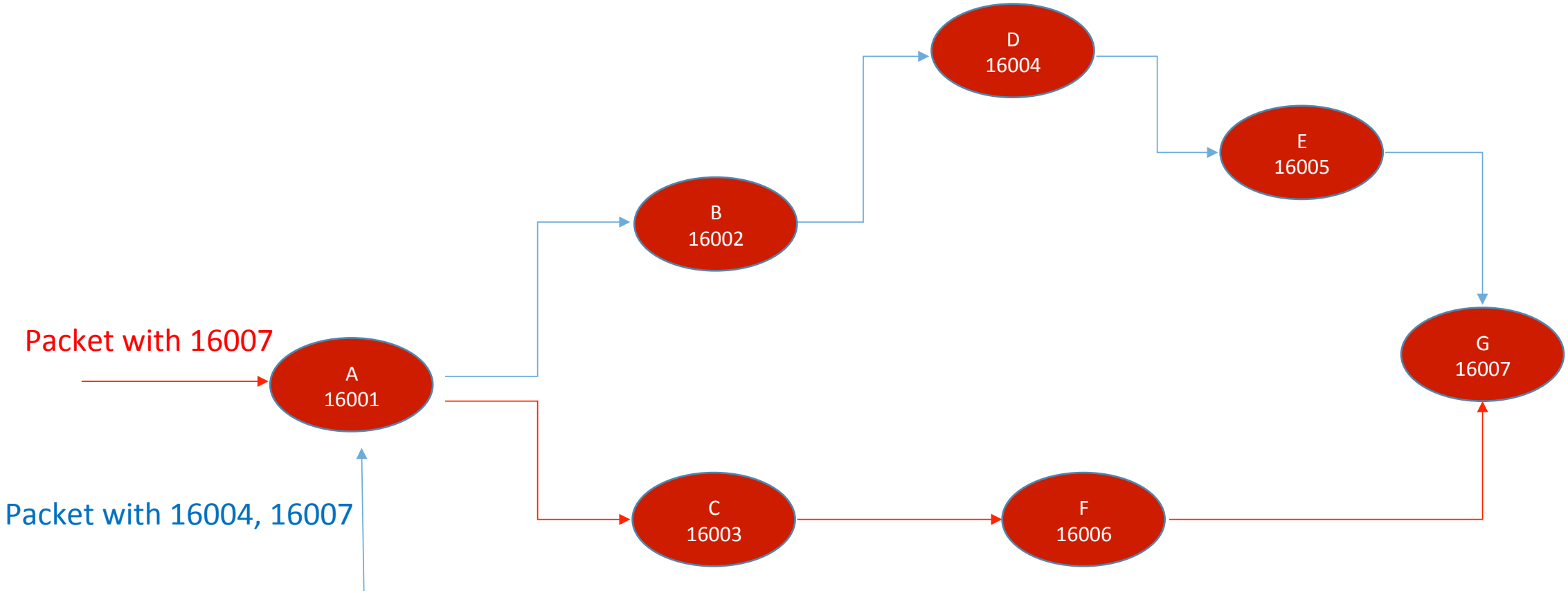
What's the goal?

- Reduce complexity and state in the network
  - Number of protocols
  - Control plane state
- Unified forwarding plane for core, edge and datacenter
- Programmatic interface to the network
- Reduce feature dependencies
- Leverage commodity hardware across network layers
- Reduce FIB table size

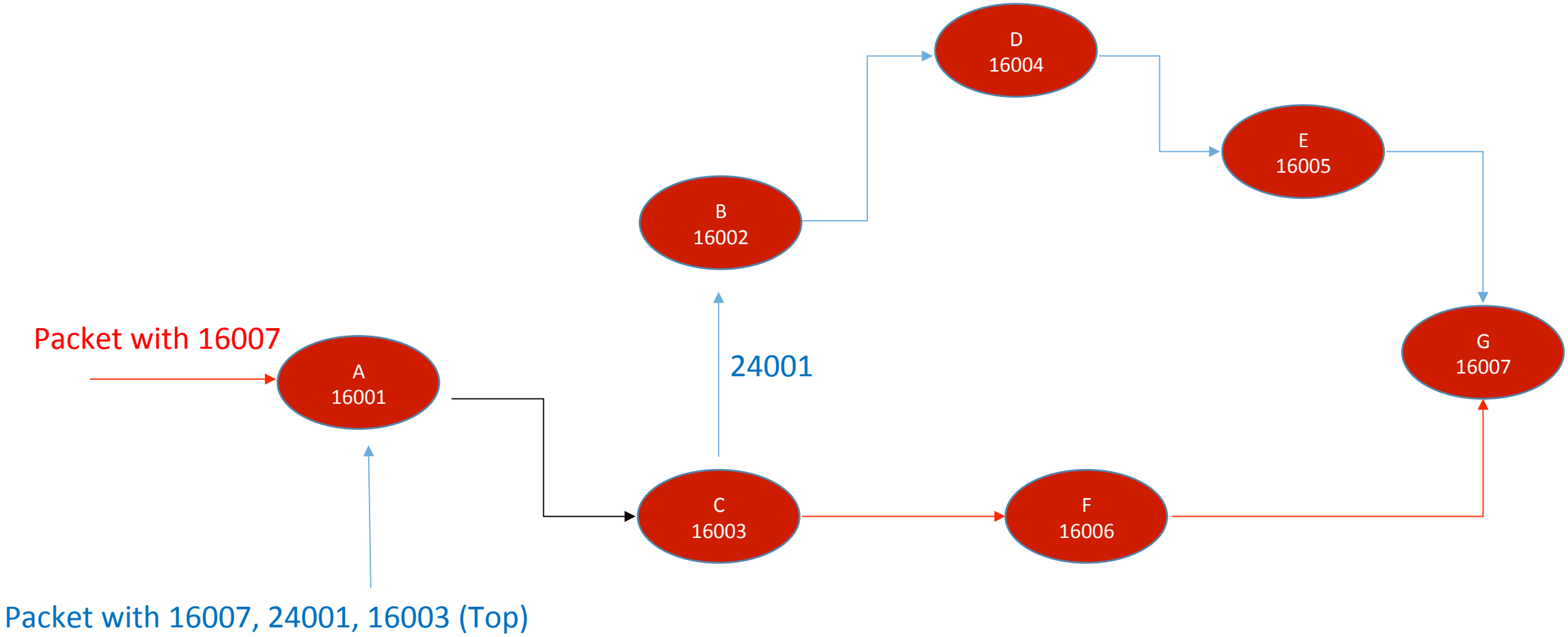
Segment Routing Primer

- Label-based source routing aka SPRING
- Uses existing MPLS data plane
- IGP floods labels throughout the SR domain
- Node-SIDs – Devices are configured with globally significant labels
  - Installed on all devices in the domain
- Adjacency-SIDs - Each SR router generates locally significant link labels
  - Installed only locally
  - They are still flooded to the entire domain

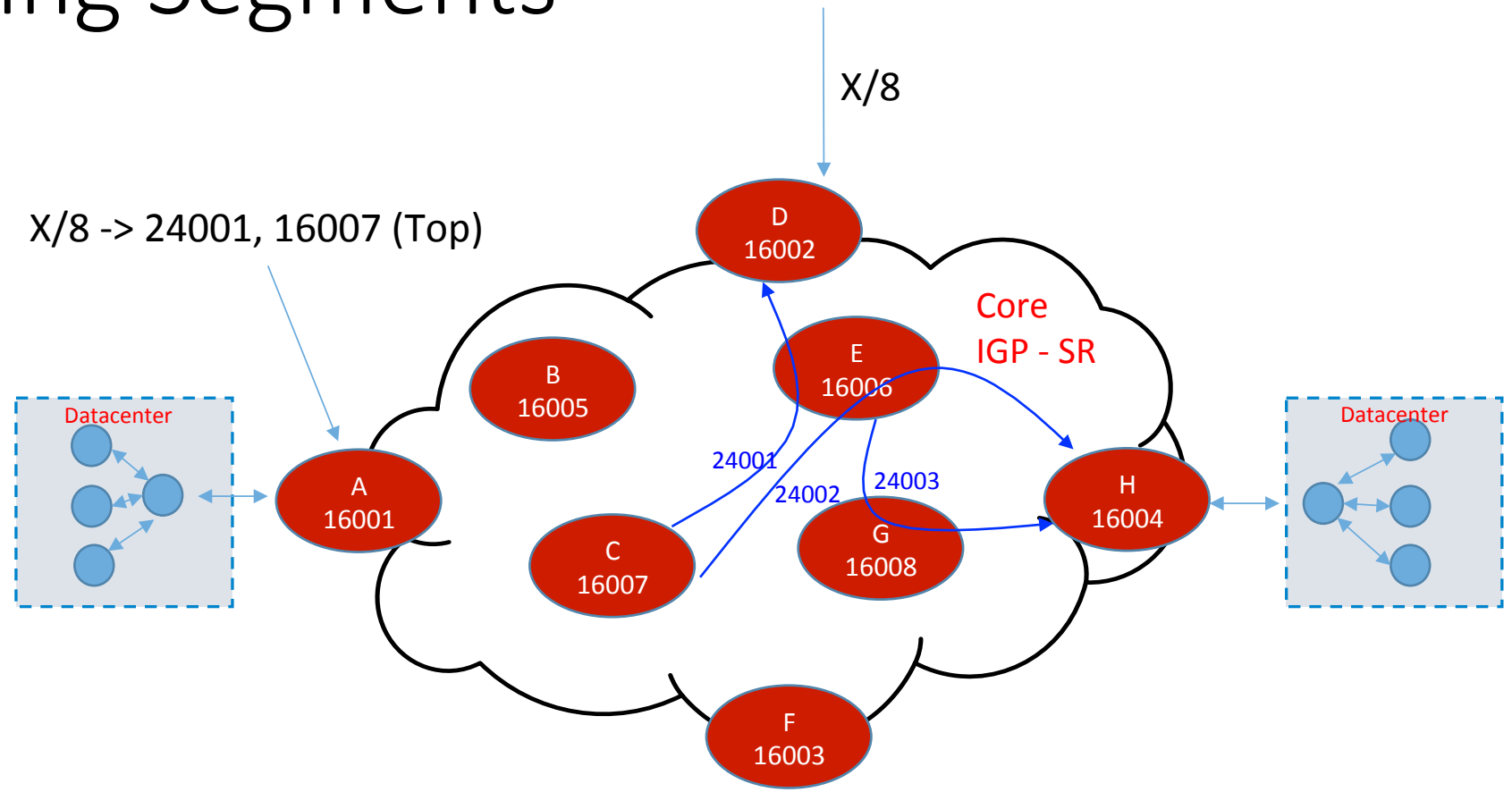
Use of Node SIDs, label distribution via IGP



Use of Adjacency SIDs and Shortest Path Override

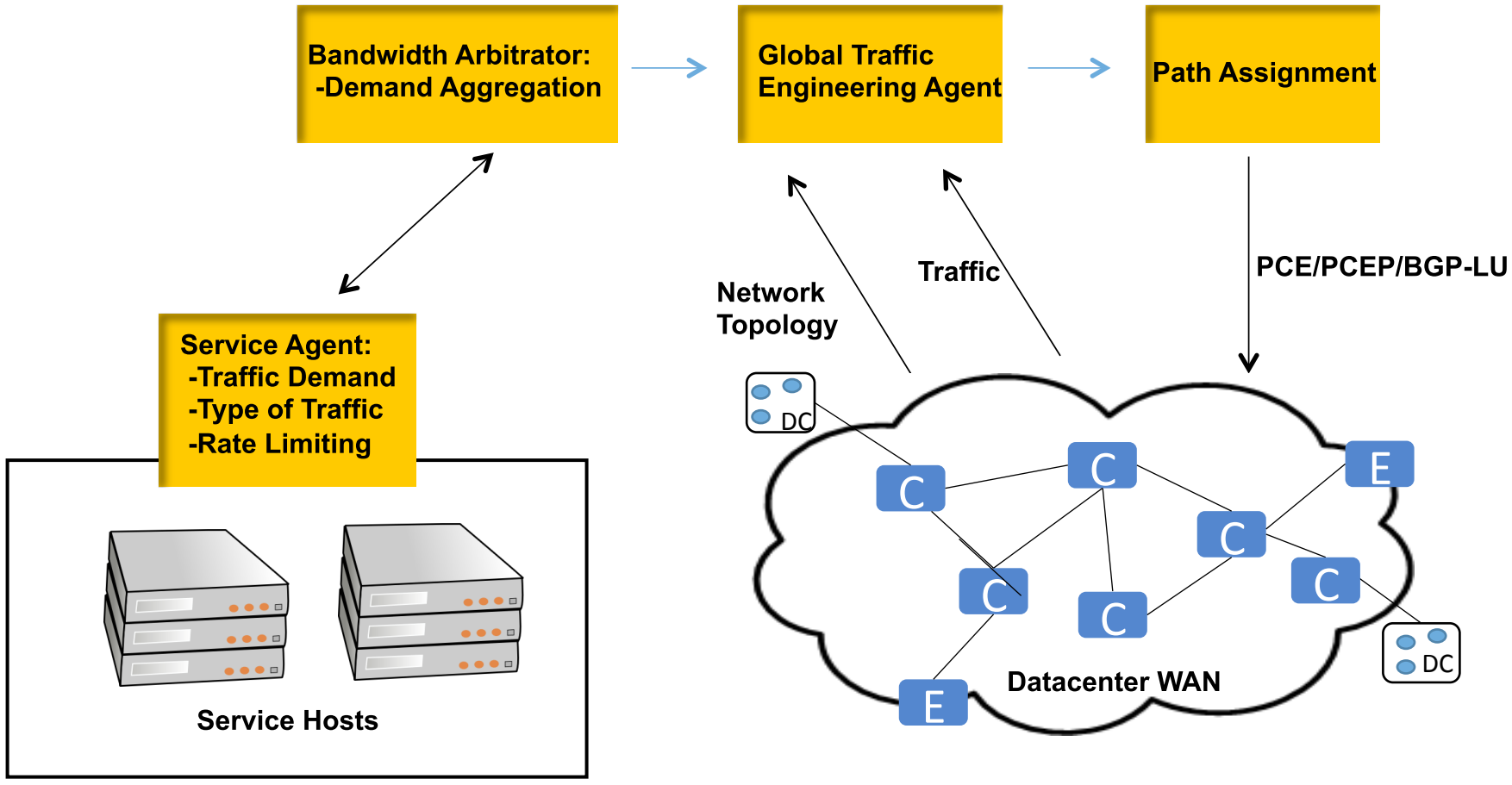


Binding Segments

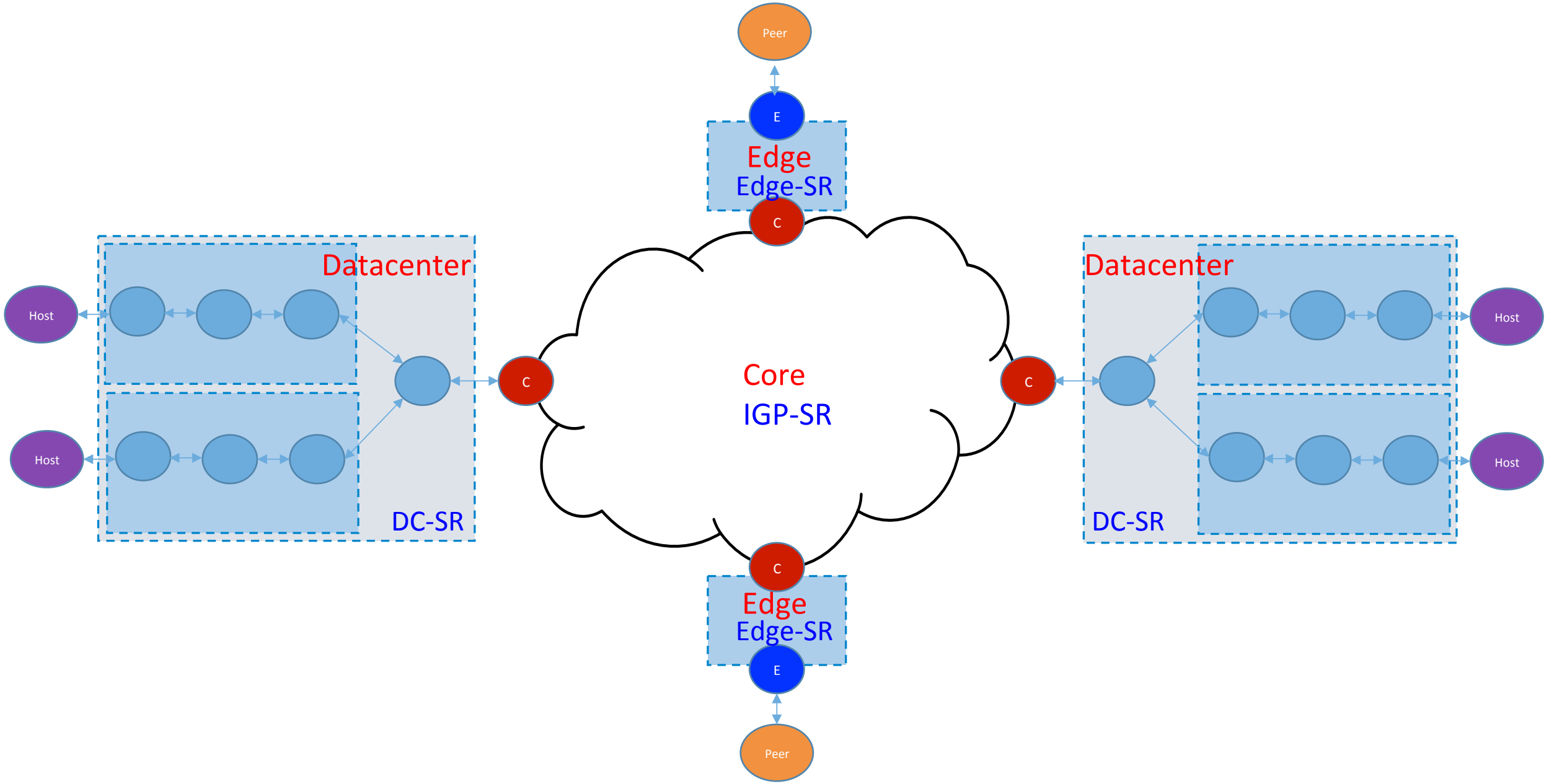


- Binding SID represents a kind of a tunnel, pop Binding SID, push one or more labels

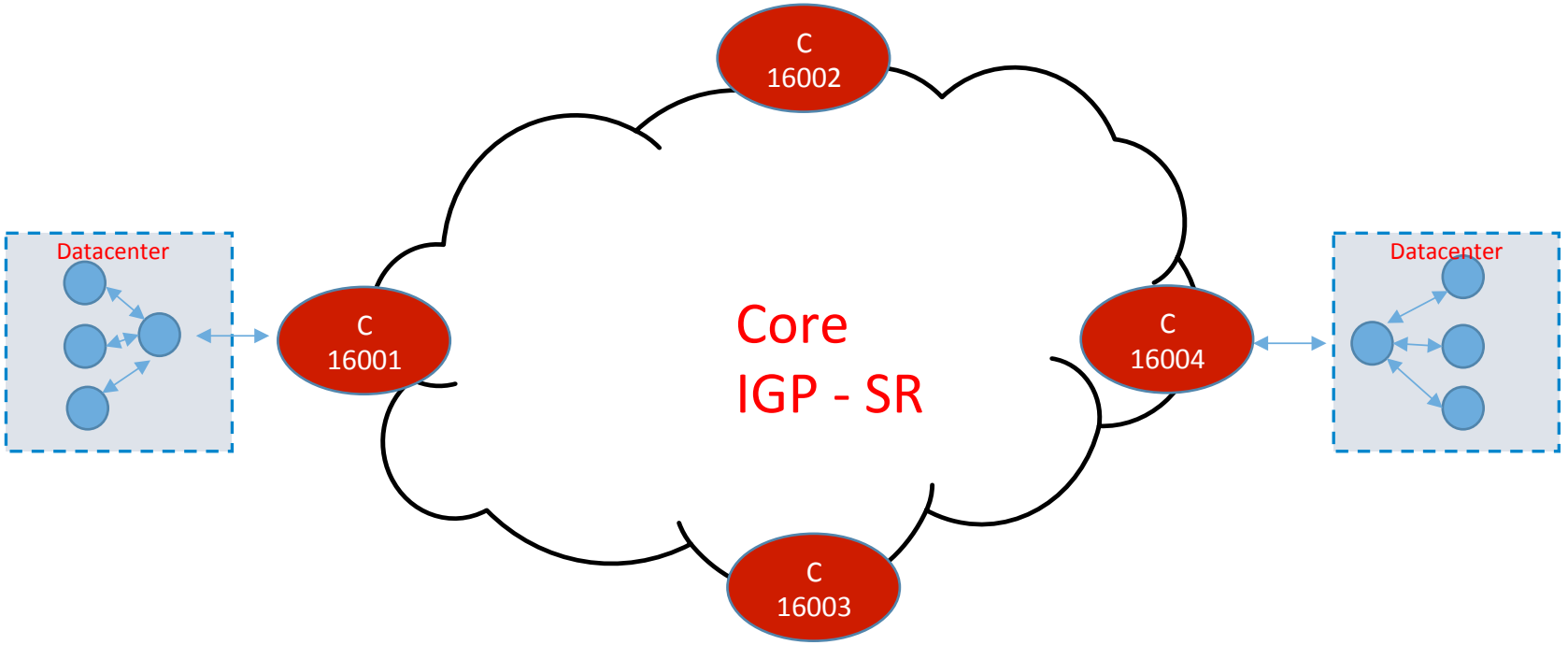
Software Driven Network



Unified Forwarding – Core, Edge and Datacenter



Unified Forwarding – Core



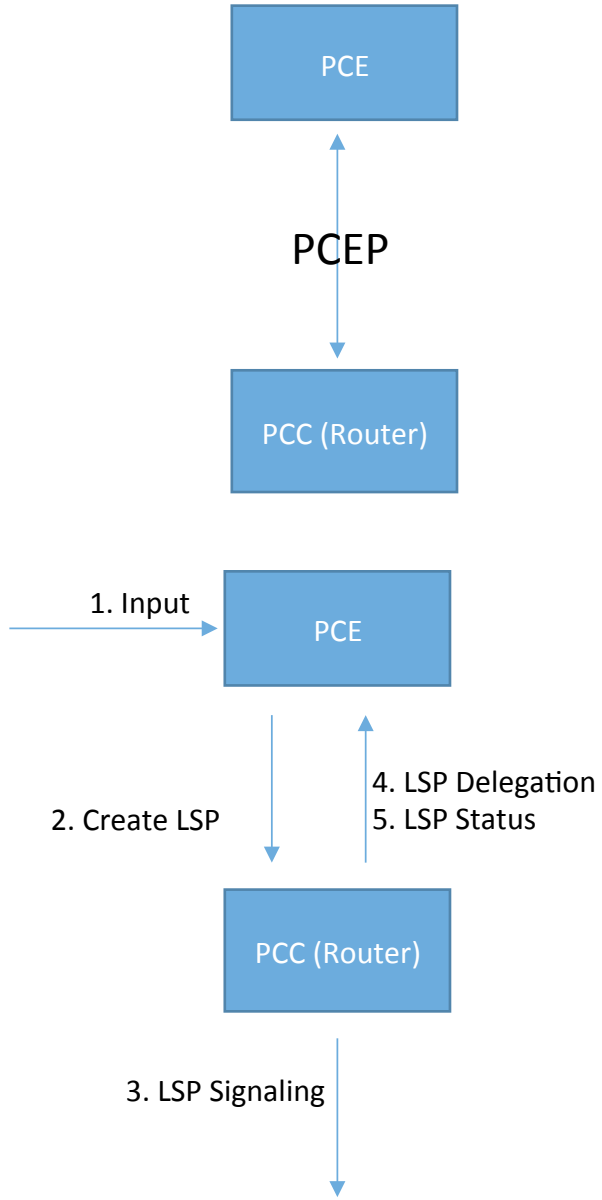
- SR Extensions for IGP
- PCE/PCEP with SR-TE
- BGP-LU

Core Traffic Engineering

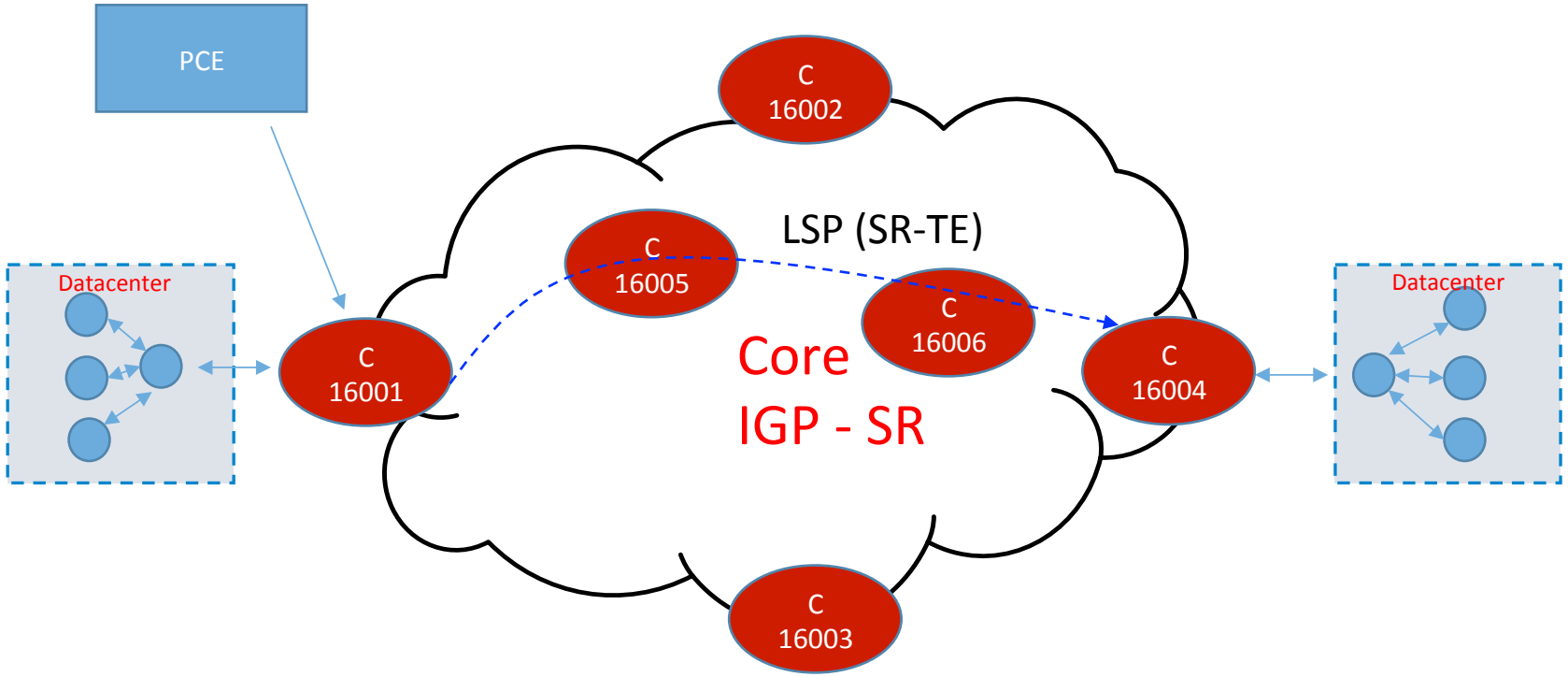
- PCE/PCEP
  - SR-TE
- BGP-TE
  - BGP-LU

PCE/PCEP with SR-TE

- Stateful PCE components
  - Path computation element (PCE)
  - Path computation client (PCC)
  - Path computation element communication protocol (PCEP)
- Offline computation and program path using PCEP



PCE/PCEP with SR-TE



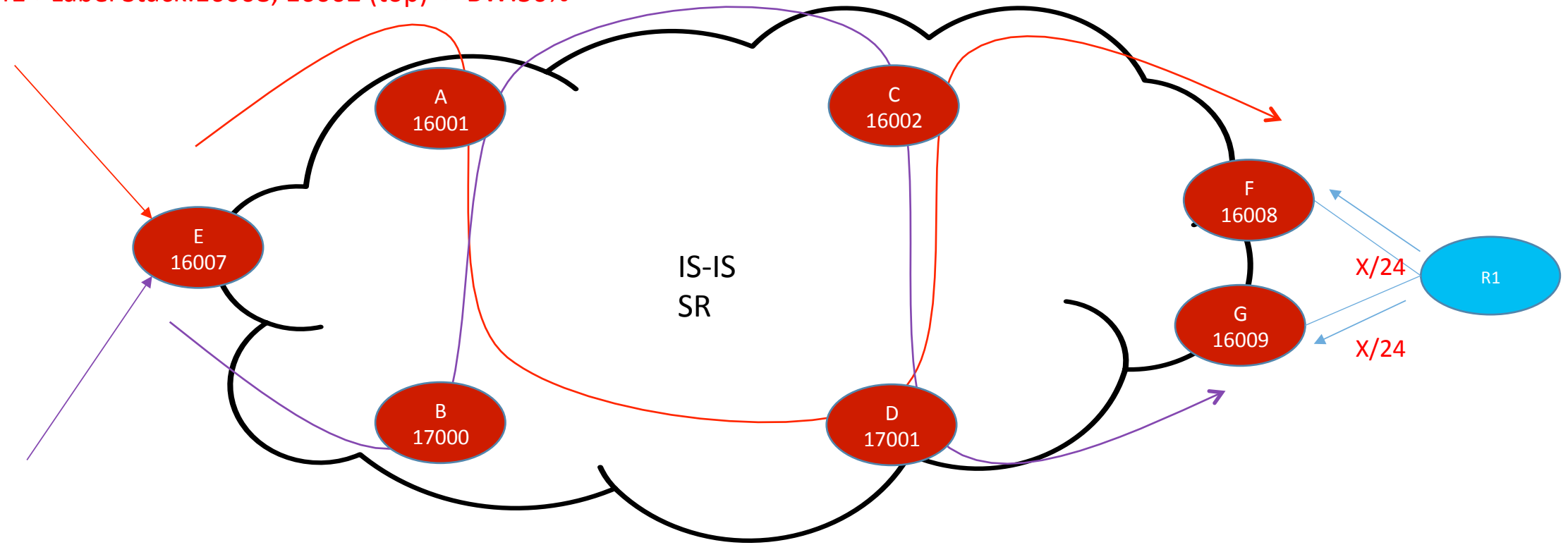
- Offline computation and PCE/Controller will program LSPs

BGP Traffic Engineering (BGP-TE)

- X/8 → nhop address, label stack 1, Link Bandwidth 1
  nhop address, label stack 2, Link Bandwidth 2
- With Binding SID, any label in the stack can be a binding segment label

BGP-TE Load Balancing

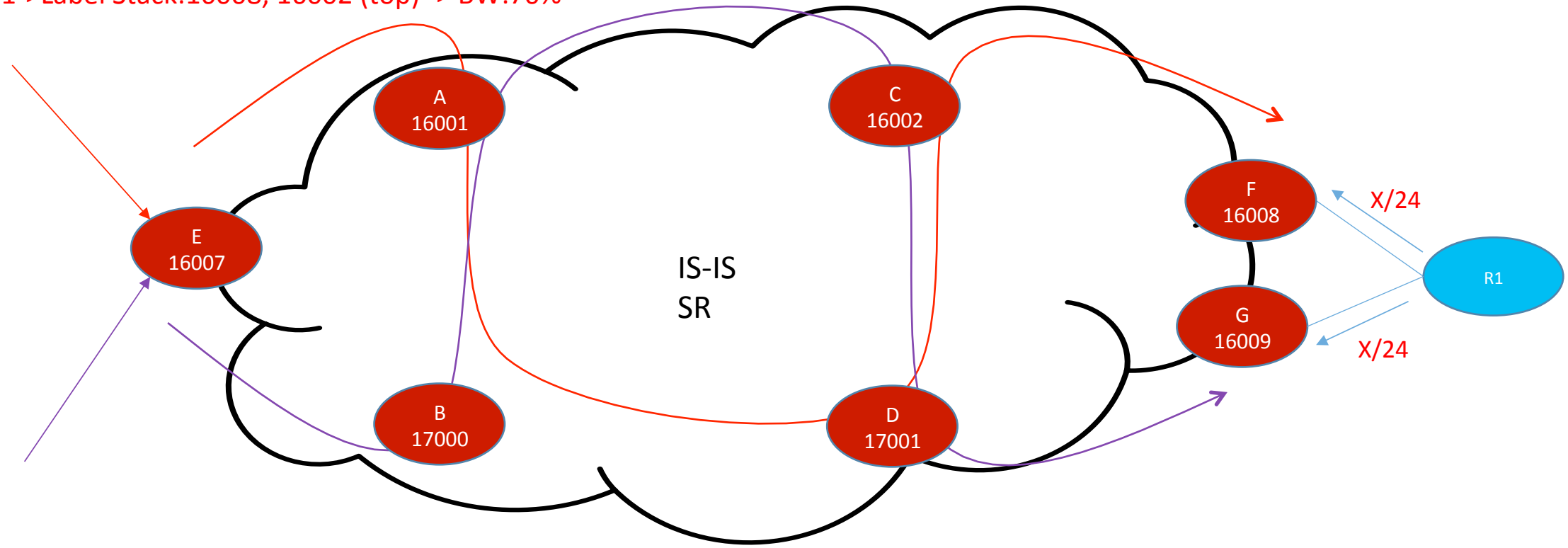
X/24 -> nh:if1->Label Stack:16008, 16002 (top) -> BW:50%



X/24 ->nh:if2->Label Stack:16009,17001 (top) -> BW:50%

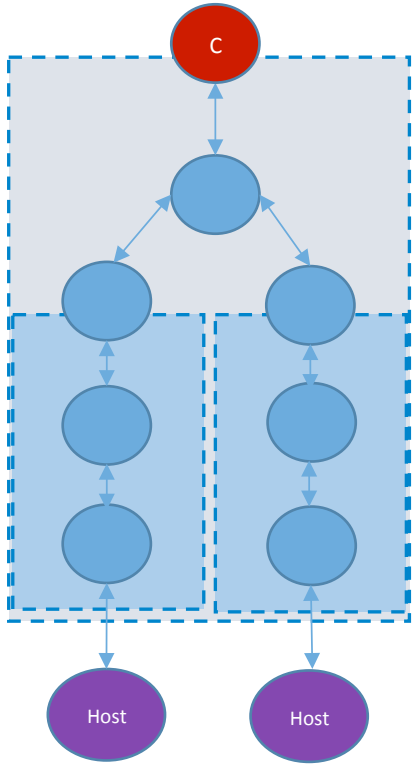
BGP-TE Unequal Cost Load Balancing

X/24 -> nh:if1->Label Stack:16008, 16002 (top) -> BW:70%

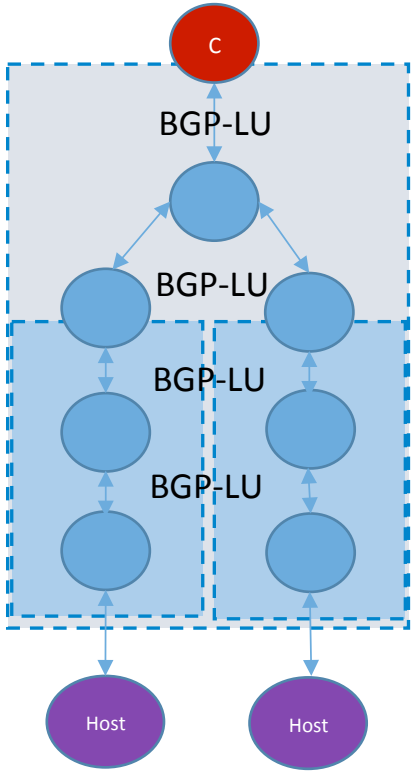


X/24 ->nh:if2->Label Stack:16009,17001 (top) -> BW:30%

Unified Forwarding – Datacenter



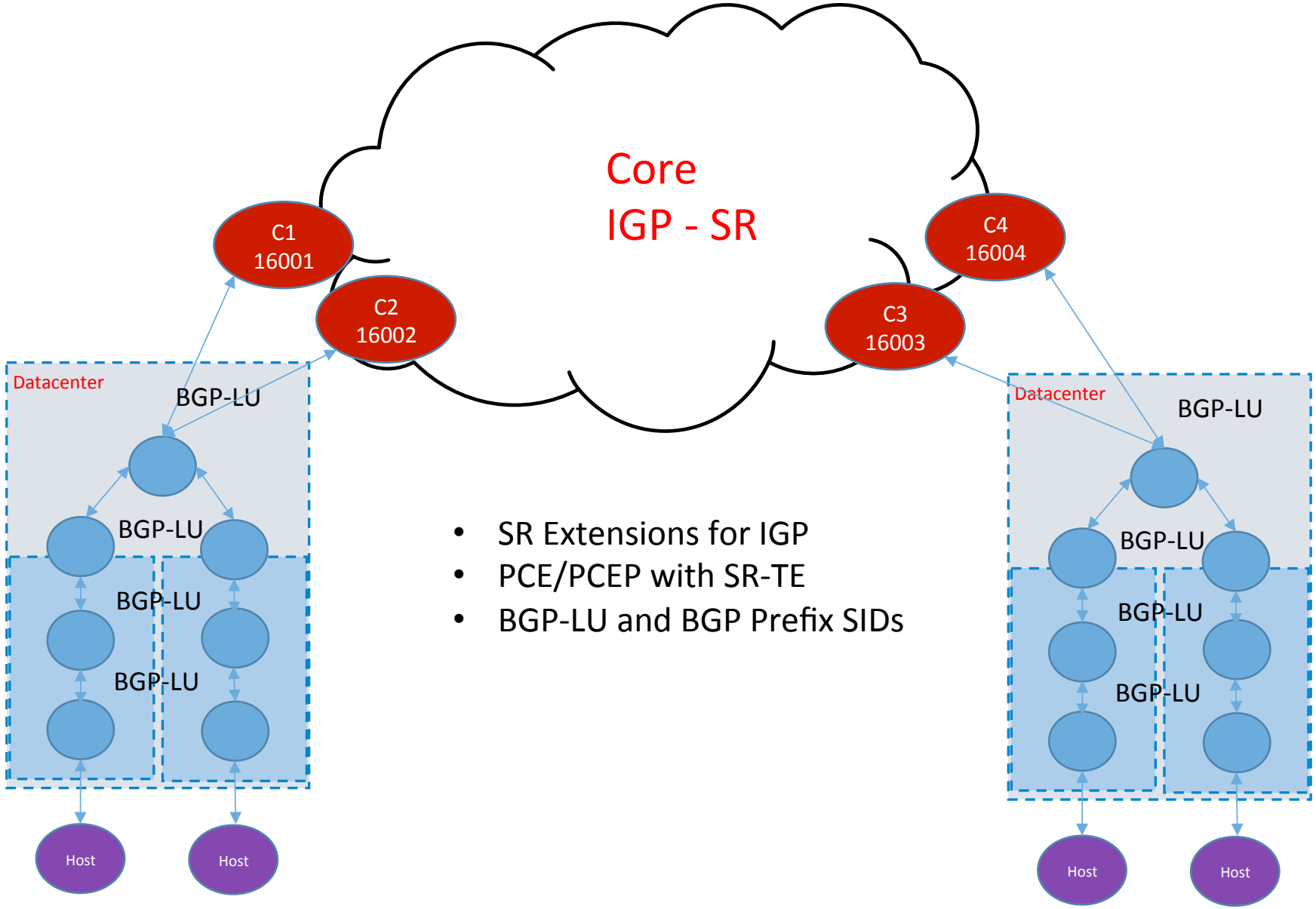
Datacenter CLOS



Datacenter CLOS with SR

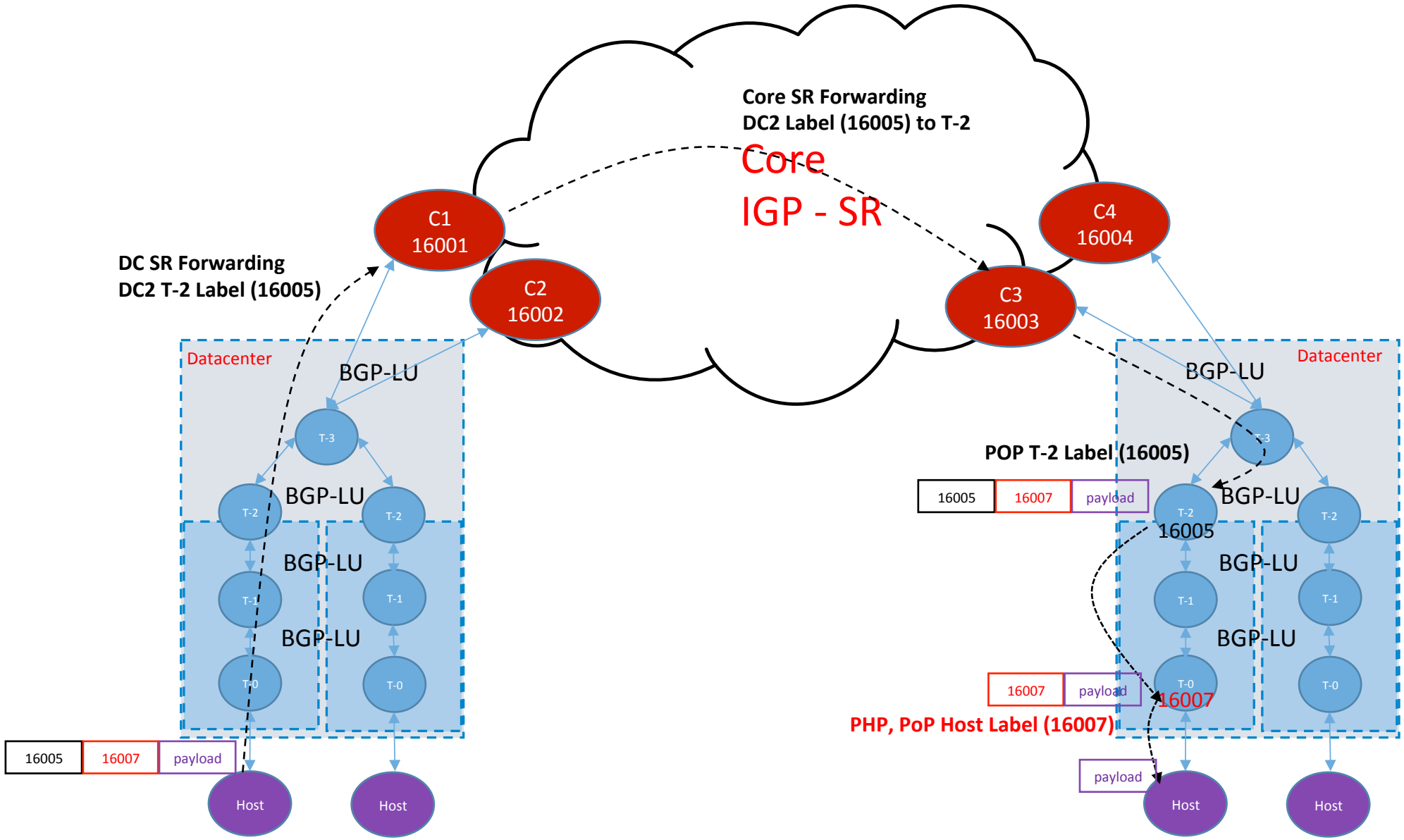
- BGP Prefix-SIDs
- BGP-LU

Unified Forwarding – Core and Datacenter

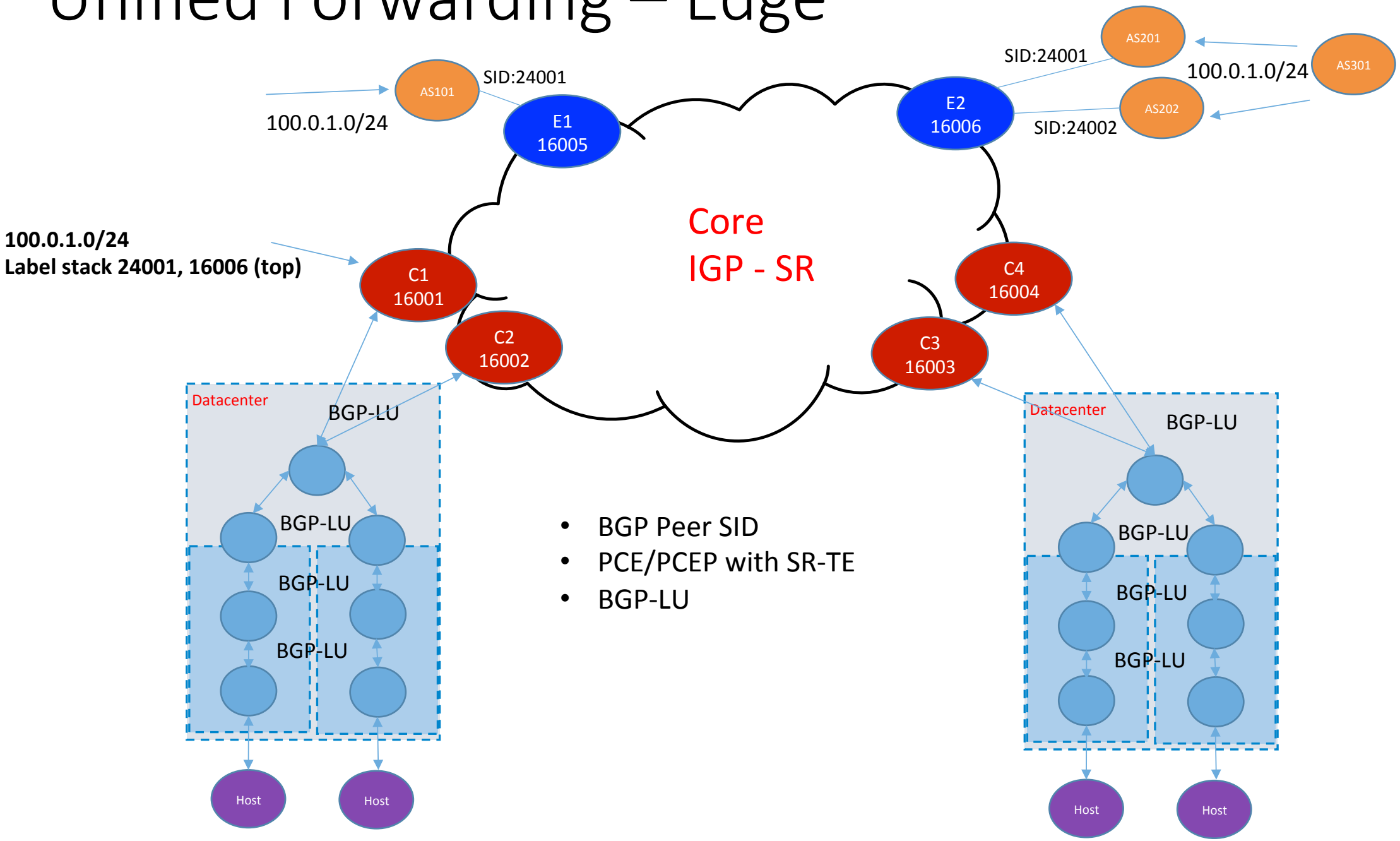


- SR Extensions for IGP
- PCE/PCEP with SR-TE
- BGP-LU and BGP Prefix SIDs

Unified Forwarding – Core and Datacenter



Unified Forwarding – Edge



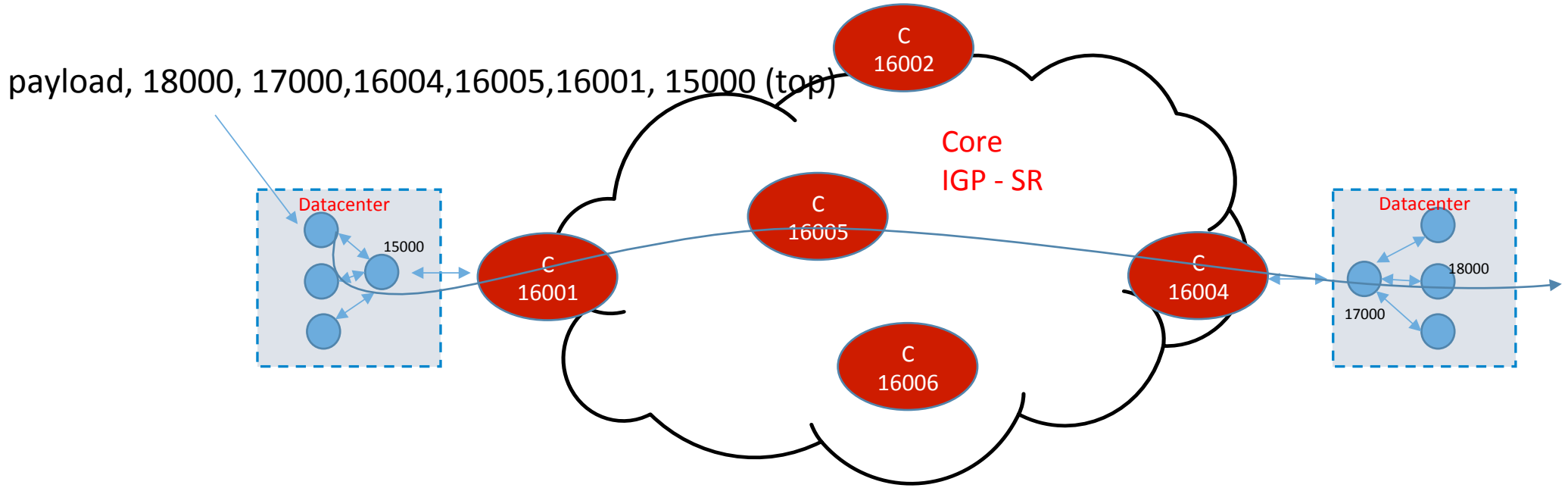
- BGP Peer SID
- PCE/PCEP with SR-TE
- BGP-LU

Unified Forwarding – Edge

- Granular traffic engineering
  - Prefix based
  - AS based
  - Overriding egress next-hops
  - Performance based routing, diverting traffic based on performance and load
- Centralized control
  - peering egress control
  - Policy engine, reduces custom configuration and standardize peering policies
- Security
  - Remote black hole triggering
  - Inject flowspec rules

Data plane Monitoring

- Simplifies end-to-end monitoring
- Construct and signal probe packets for data plane health check
- Construct paths without creating state in the network



Useful URLs

- http://www.segment-routing.net/
- http://datatracker.ietf.org/wg/spring/documents/
- http://datatracker.ietf.org/wg/idr/documents/
- http://datatracker.ietf.org/wg/pce/documents/


Microsoft