Cisco Evolved Programmable Network Transport Design Guide, Release 5.0

First Published: 2017-06-23

Americas Headquarters
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
http://www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 527-0883

CONTENTS

CHAPTER 1 **Overview** 1

Cisco Evolved Programmable Network Overview 1

Programmable Transport Network Overview 1

Transport Models 2

Access Node Roles 2

Size of the Network 3

Transport Programmability 4

System Test Topology 4

CHAPTER 2 **Transport Architecture Design** 7

Use Cases 8

Small Network - End to End Segment Routing 8

Large Network - End to End Segment Routing 9

Large Network - End to End Programmable Segment Routing 10

Large Network - Inter-AS End to End Segment Routing 11

Large Network - Segment Routing and LDP Interworking 12

Large Network - Layer2 Access to Segment Routing Transport 14

Benefits 15

Greenfield Deployment 15

Migrating from EPN 4.0 or Traditional Unified MPLS Transport 15

Control Plane and Data Plane Flow 15

Traditional Unified MPLS - Control Plane 16

Traditional Unified MPLS - Data Plane 16

BGP LU with BGP Prefix SID - Control Plane 17

BGP LU with BGP Prefix SID - Data Plane 17

Programmable Transport - Control Plane 18

Programmable Transport - Data Plane 18

CHAPTER 3 Functional Components Design 19

Redundancy and High Availability 19

Loop-Free Alternate with Bidirectional Forwarding 19

Microloop Avoidance in Remote Loop-Free Alternate Fast Reroute 20

Segment Routing Topology Independent Loop-Free Alternate with Bidirectional Forwarding 20

Border Gateway Protocol — Fast Reroute — BGP Prefix Independent Convergence Core/Edge 21

XR Transport Controller 21

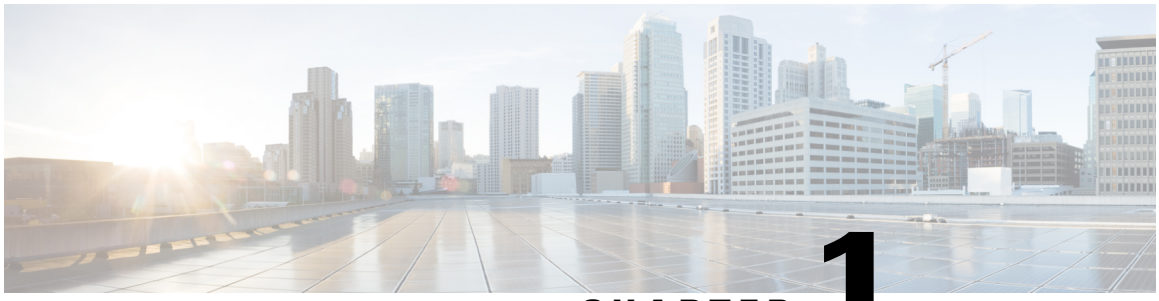
Border Gateway Protocol Route Reflector Architecture 22

Transport Route Reflector Design 23

Service Route Reflector Design 24

Quality of Service 24

CHAPTER 4 Conclusion 31



CHAPTER 1

Overview

This chapter contains the following section:

- Cisco Evolved Programmable Network Overview, page 1

Cisco Evolved Programmable Network Overview

Cisco Evolved Programmable Network (EPN) is built on the successful EPN architecture framework to bring greater programmability and automation. This effort is part of a multi-year ongoing development program that is built towards a flexible, programmable, and cost-optimized network infrastructure that Cisco targets to deliver in-demand fixed and mobile network services.

The Cisco EPN system design follows a layered design aimed to simplify the end-to-end transport and service architecture. By decoupling the transport and service infrastructure layers of the network, it allows these two distinct entities to be provisioned and managed independently. At the same time, enhancements in the transport and orchestration layers now allow for a tighter coupling of transport and services, enabling transport to become raw clay that services can be molded based on their needs and requirements. The Cisco EPN allows programmatic interaction between the service and transport layers.

This guide focuses on the design aspect of the transport layer. The transport layer provides the framework to achieve connectivity among the nodes in the network, and enables all the consumer and enterprise services that the Cisco EPN system promotes.

Programmable Transport Network Overview

The Cisco EPN system incorporates a network architecture designed to consolidate multiples services on a single Multi-Protocol Label Switching (MPLS) transport network, which is primarily Application Engineered Routing (AER) transport optionally co-existing with traditional MPLS technology.

The continuous growth in consumer and enterprise services, combined with ubiquitous LTE-driven mobile broadband adoption and the anticipation of explosive 5G adoptions, has introduced unprecedented levels of scale in terms of the access network. This factor, combined with services requiring ubiquitous connectivity from the access domain into and across the core network, has led to scale challenges in the MPLS network. Moreover, the need to reduce operating cost and the demand to deliver variety of services in a short time has put tremendous challenge in continuing to use traditional MPLS transport. This necessitates the transition to Segment Routing (SR) based MPLS transport, which is also referred as AER in this chapter.

In MPLS, the Service Edge (SE) node must be identified by a /32 IP address, thus precluding route summarization from being performed among access, aggregation, and core regions of the network. To address the resulting route scale problem, the Cisco EPN system promotes a network design that leverages a unified MPLS-based hierarchical approach.

The unified MPLS adopts a divide-and-conquer strategy where the core, aggregation, and access networks are partitioned in different MPLS/IP domains that are isolated from Interior Gateway Protocol (IGP) perspective. The adoption of divide-and-conquer strategy reduces the size of routing and forwarding tables within each domain, which in turn leads to better stability and faster convergence. The SR enabled IGP is used for label distribution to build Label Switched Path (LSP) within each independent IGP domain. This enables a device inside an access, an aggregation, or a core domain to have reachability through intra-domain SR LSPs to any other device in the same region. For migration purposes in brownfield customers, the Cisco EPN also retains few domains running traditional MPLS (LDP).

Within a domain, both Intermediate System to Intermediate System (IS-IS) and Open Shortest Path First (OSPF) are suitable choices of IGP protocols, but the protocol selection is based on operator's preference.

Reachability across domains is achieved either using RFC 3107 procedures whereby BGP-Labeled Unicast (BGP LU) is used to build inter-domain hierarchical LSPs across domains, or by a controller that pushes the segment list (SR label stack) for a particular service. This allows the link state database of the IGP in each isolated domain to remain as small as possible, leaving all external reachability information to be carried through BGP, which is designed to scale to the order of millions of routes.

- In Single-AS multi-area designs, the interior Border Gateway Protocol (iBGP)-labeled unicast is used to build inter-domain LSPs.
- In Inter-AS designs, the iBGP-labeled unicast is used to build inter-domain LSPs inside the AS, while exterior Border Gateway Protocol (eBGP)-labeled unicast is used to extend the end-to-end LSP across the AS border.

In both the cases, the unified MPLS transport across domains uses hierarchical LSPs that rely on a BGP-distributed label to transit across the isolated MPLS domains, and on SR segment within the domain to reach the inter-domain Border Router (BR) or Autonomous System Border Router (ASBR) corresponding to the labeled BGP next hop. If an IGP domain is not SR capable, the domain can run LDP instead of SR.

Alternatively, a controller with visibility on the entire network can provide the end to end path for interdomain reachability. Such visibility is attained through BGP Link State (BGP-LS) feeds from each domain providing topology and state information.

Transport Models

The transport architecture considers the role of the access node, the size of the network, and the mechanism by which the transport LSP is built.

Access Node Roles

Depending on the capability of the access node and the desired type of service, the access node can have one of the following roles:

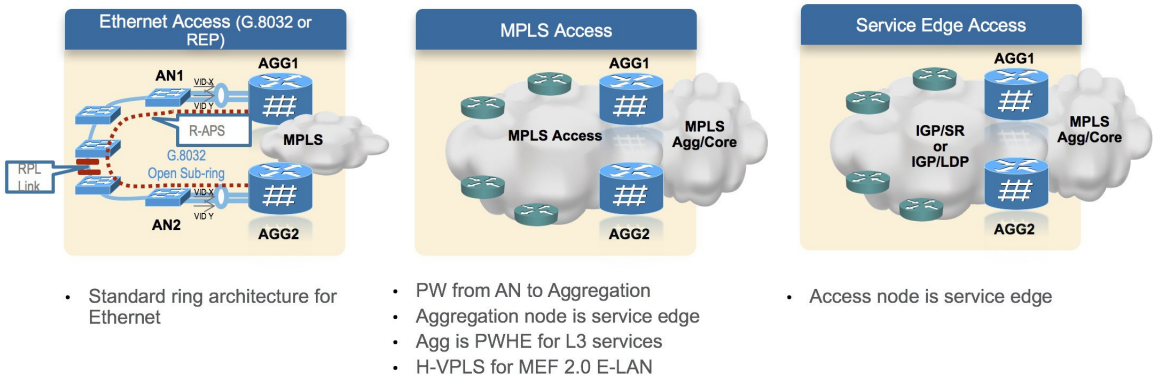
1 **Ethernet Access Node:** In this role, the access node is part of a Layer-2(L2) ring network, for example Resilient Ethernet Protocol (REP) or G.8032. It provides L2 access to the service edge node, which is either a pre-aggregation or an aggregation layer depending on the size of the network. The service edge node is responsible for associating the L2 service identification such as a VLAN tag, to the end-to-end

service. The service provisioning system enables to configure both the access node and the service edge node to complete service provisioning task. The access node in this role is required only to support corresponding L2 protocol (REP or G.8032).

2 **MPLS Access Node:** In this role, the access node is part of an MPLS access network, which could be SR or traditional LDP based MPLS access network. The access node provides Pseudo-Wire (PW) access to the service edge node, which is either a pre-aggregation or aggregation layer depending on the size of the network. The service edge node is responsible for associating the PW to the end-to-end service. This is accomplished through Pseudo Wire Head-Eend (PWHE) or Hierarchical Virtual Private LAN service (H-VPLS) capability. Similar to Ethernet Access, the service provisioning system allows to configure both the access node and service edge node to complete the service-provisioning task. The access node has to carry only routes and segments (or labels) in the corresponding access domain. It does not have to carry any service routes or reachability information for remote access nodes. Therefore, the scale and feature requirement for this access node is very limited.

3 **Service Edge Access Node:** The access node is a full-fledged service edge node. It participates in service mapping, service route exchange, IGP/SR in the access layer, and BGP LU. The route scale requirement is also much higher compared to the MPLS or Ethernet access node roles. The service provisioning is simplified as it only involves configuring the access node. The route scale requirement for the transport routes can be mitigated using Software Defined Networking (SDN) controlled transport route filtering. However, this allows the SDN controller which is a service provisioning system, to configure both the access node as well as the next layer (pre-aggregation or aggregation) node. For some services such as Layer-3 VPN, the service route scale requirement can be mitigated using the Route Target Constraint (RTC).

Figure 1: Access Node Roles



- Standard ring architecture for Ethernet
- PW from AN to Aggregation
- Aggregation node is service edge
- Agg is PWHE for L3 services
- H-VPLS for MEF 2.0 E-LAN
- Access node is service edge

Size of the Network

The size of the network is an important factor in dictating the network architecture. For some small networks, hierarchical LSP (BGP LU) may not be required. On the other hand, for large networks, multiple layers including access, pre-aggregation, and aggregation within the single AS are required. Some large networks require gluing of multiple AS networks together. For large networks, it is important to apply appropriate route filtering and domain isolation, to reduce the scale requirements for access devices.

Transport Programmability

BGP LU (RFC 3107), which is also referred to as unified MPLS, has been the central element for Cisco EPN transport architecture in the previous releases. In the unified MPLS model, BGP LU protocol is used to build the end-to-end LSP for the transport. In Cisco EPN, Release 5.0, unified MPLS with additional capabilities is also one of the transport options. In addition, Cisco EPN, Release 5.0 includes programmable transport using AER. In general, there are three transport options in Cisco EPN, Release 5.0:

1 **Traditional BGP LU:**

   This is a protocol driven transport that was also supported in the previous releases. It relies on BGP to build the end-to-end hierarchical LSP. Fast convergence across IGP boundaries is achieved using BGP Prefix Independent Convergence (BGP PIC) feature. This model is suitable for networks such as traditional MPLS network.

2 **BGP LU with BGP Prefix SID:**

   This is also a protocol driven transport. However, the BGP prefix SID feature enables to uniquely identify each IGP border node and service node in the network using a unique prefix segment ID (BGP prefix SID). This transport option requires end-to-end segment routing capability as well as BGP prefix SID capability at IGP border routers.

3 **Programmable Transport:**

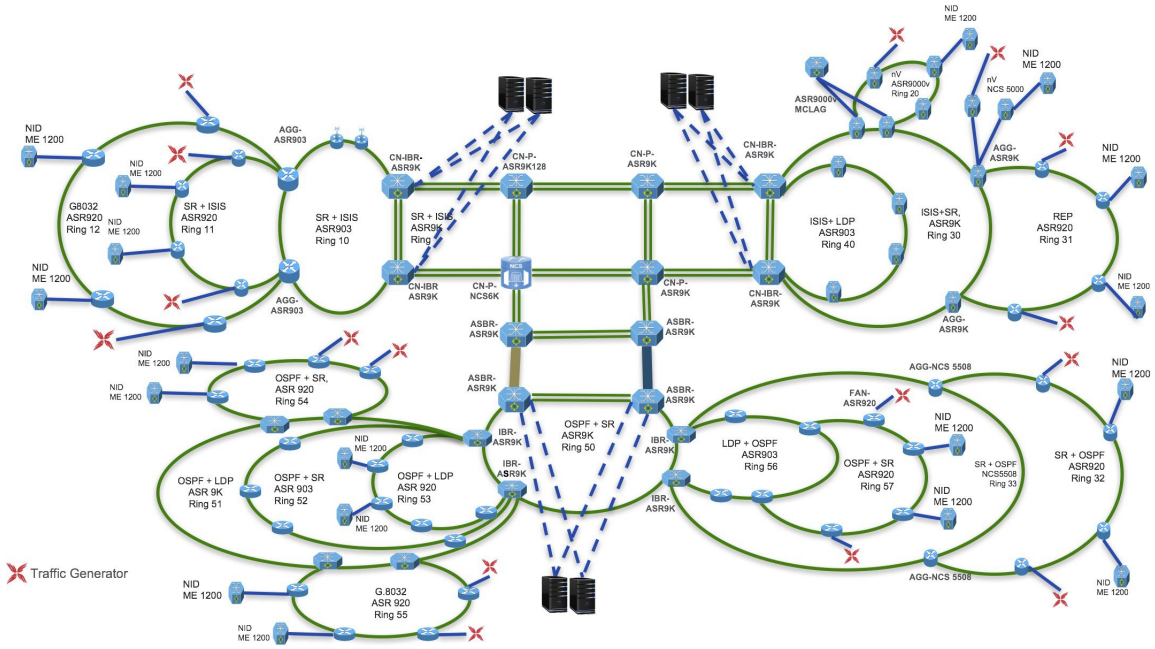
   This transport option is SDN driven. Each domain runs IGP/SR across the domain, and two IGP border routers in each domain uses BGP LS to feed topology, bandwidth, reliability, latency, SRLG and other transport states of the IGP domain to the SDN controller. The SDN controller uses the topology data and current state of the network, to build the best path and alternate disjoint path that satisfies a given service requirement and pushes the corresponding segment list to the service edge router.

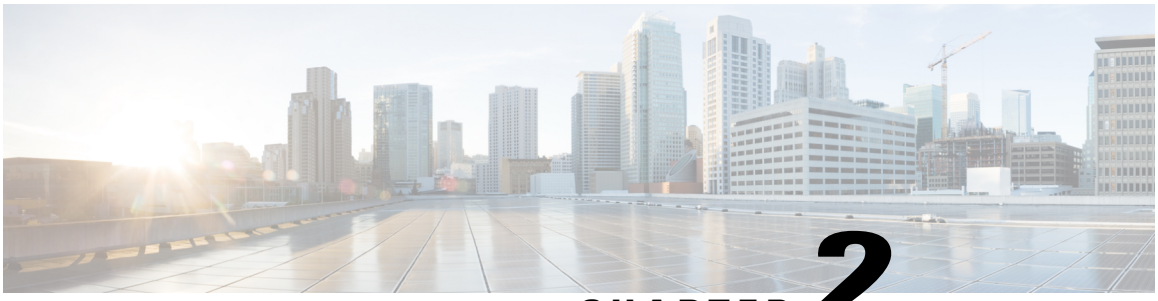
System Test Topology

The development process for the Cisco EPN 5.0 system provides extensive validation of various functional aspects of the system design. This validation is conducted on a test bed that is designed to emulate the

characteristics of a converged operator's production network environment. The details of the system test bed are illustrated in the *System Test Topology Reference Guide*.

Figure 2: System Test Topology





CHAPTER 2

Transport Architecture Design

This chapter contains the following sections:

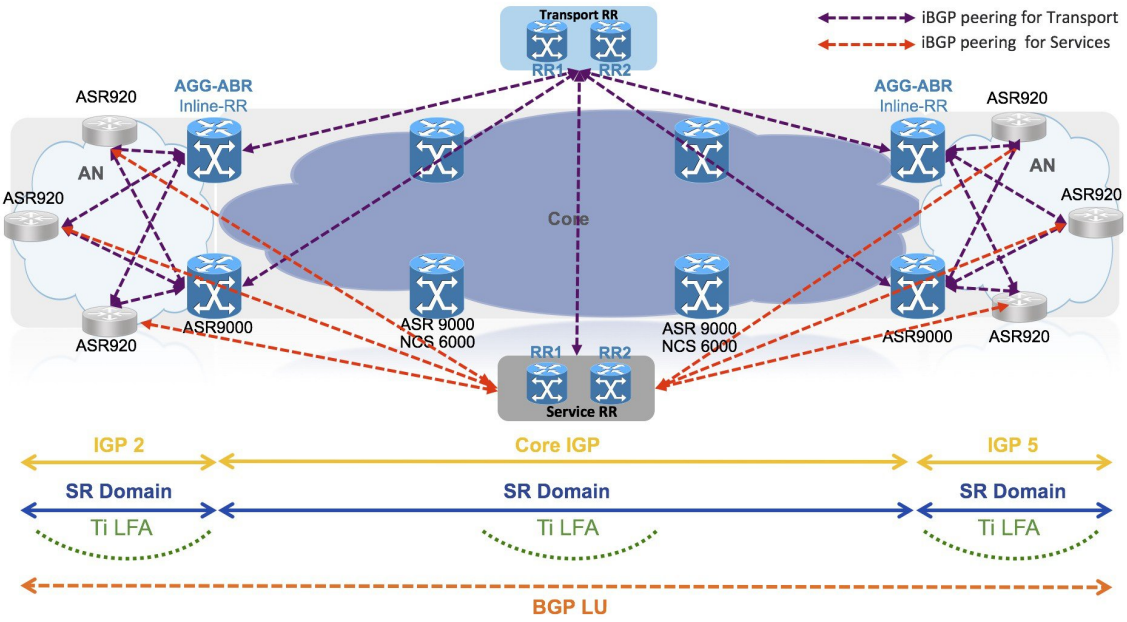
- Use Cases, page 8
- Benefits, page 15
- Control Plane and Data Plane Flow, page 15

Use Cases

Small Network - End to End Segment Routing

This section focuses on the network design of a small network with two layers namely access and core layers. The access node is also a service edge node. The end-to-end transfer is established using BGP LU with BGP prefix-SID extension as shown in the below figure.

Figure 3: Small Network - End to End Segment Routing



The characteristics of this network are given below:

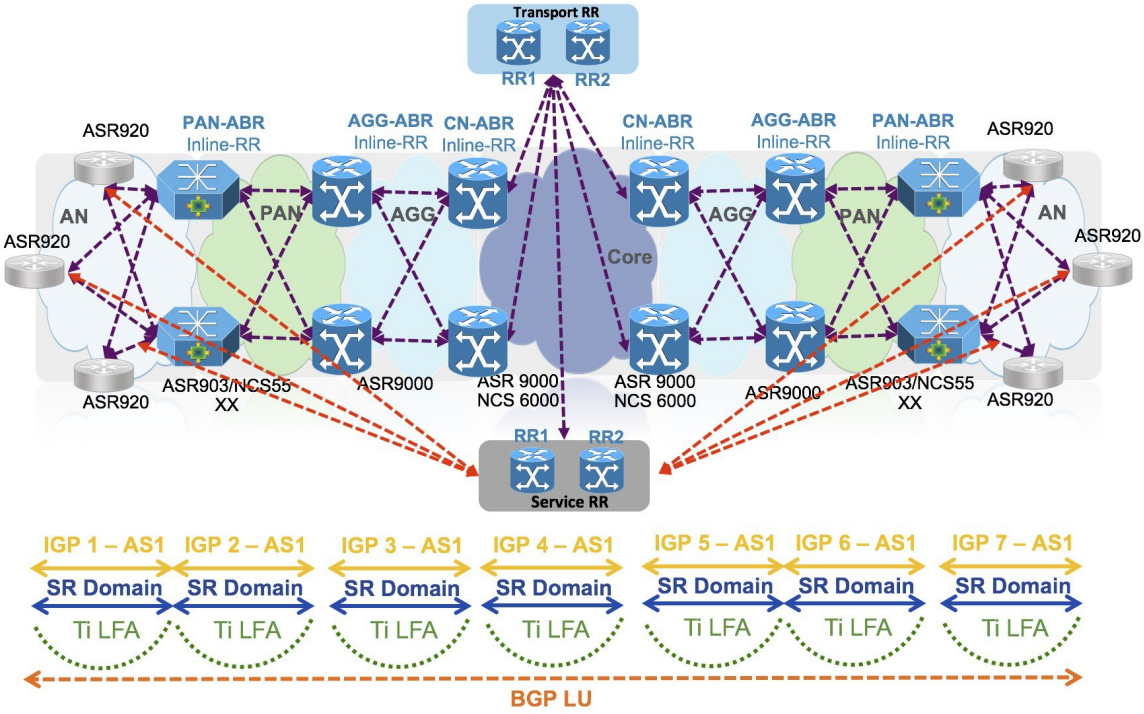
- Each of the domains, like core or access, is running independent IGP with SR.
- Fast convergence within each domain is achieved with Topology Independent Loop Free Alternate (TI-LFA) technology using SR/IGP capabilities.
- The IGP in each domain could be OSPF single area (preferably area 0) or ISIS single level (preferably level-2).
- Each access domain is connected to the core domain using two IGP Aggregation Area Border Routers (AGG-ABRs).
- End-to-end transport is achieved using BGP LU with BGP prefix SID.
- Each access node and Aggregation Area Border Routers (AGG-ABRs) participate in BGP LU.
- Each AGG-ABR has iBGP peering (nexthop-self) to two dedicated transport route reflectors (RR).
- Each AGG-ABR is an inline RR for all corresponding access nodes in its region.

- BGP-PIC is used for Fast Reroute (FRR).

Large Network - End to End Segment Routing

This section focuses on the network design of a large network with four domains namely access, pre-aggregation, aggregation, and core. The end-to-end transfer is established using BGP LU with BGP prefix-SID extension as shown in the below figure. With the exception of the size and the depth of the hierarchy, this architecture is similar to that of small network explained in chapter Small Network - End to End Segment Routing

Figure 4: Large Network - End to End Segment Routing



The characteristics of this network are given below:

- Each of the domains namely core, aggregation, pre-aggregation, and access is running on independent IGP with SR.
- Fast convergence within each domain is achieved with the Topology Independent Loop Free Alternate (TI-LFA) technology using SR/IGP capabilities.
- The IGP in each domain could be OSPF single area (preferably area 0) or ISIS single level (preferably level-2).
- Each access domain is connected to the pre-aggregation domain using two IGP Pre-Aggregation Area Border Routers (PAN-ABRs).
- Each pre-aggregation domain is connected to corresponding aggregation layer.

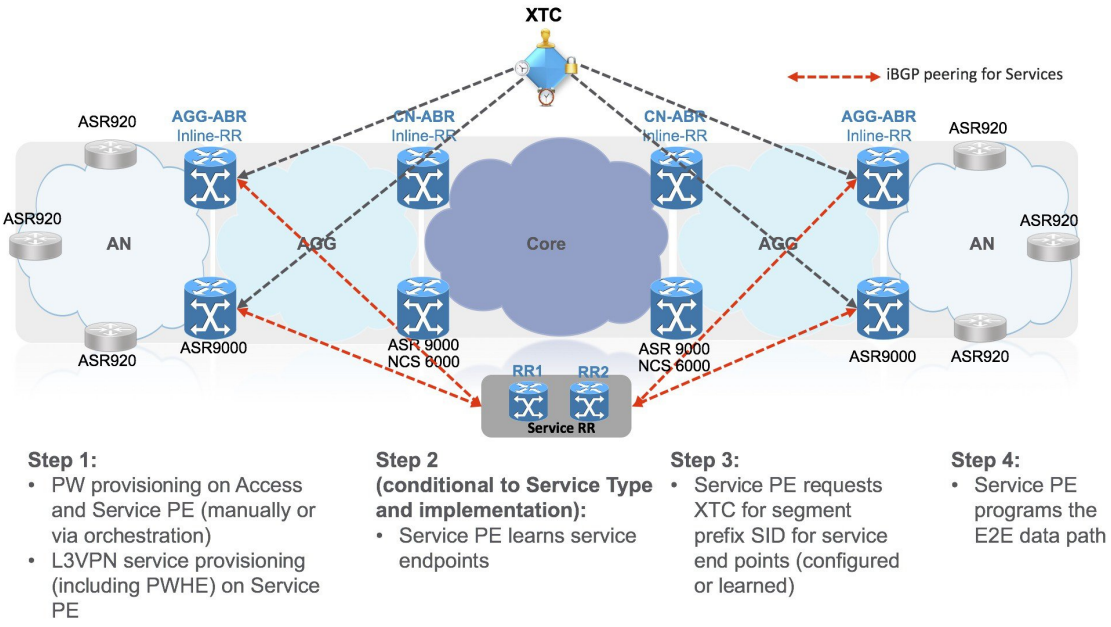
- Each aggregation domain is connected to the core layer.
- End-to-end transfer is achieved using BGP LU with BGP prefix SID in the same way as mentioned in section Small Network - End to End Segment Routing , on page 8.
- Each access node and all the Area Border Rrouters (ABRs) participate in BGP LU.
- Each core ABR that connects to aggregation ABR has iBGP peering (nexthop-self) to two dedicated transport Route Reflectors (RR).
- The core ABRs are inline RR for the corresponding aggregation ABRs.
- The aggregation ABRs are inline RR for the corresponding pre-aggregation ABRs.
- Each pre-aggregation ABR is an inline RR for all corresponding access nodes in its region.
- BGP-PIC is used for Fast Reroute (FRR).

Large Network - End to End Programmable Segment Routing

This section focuses on the network design of a programmable transport network. The main differentiating factor with respect to use cases namely Small Network - End to End Segment Routing , on page 8 and Large Network - End to End Segment Routing , on page 9, is that the programmable transport uses SDN technologies to build end-to-end LSP paths instead of using BGP LU (BGP 3107) to glue together multiple IGP domains.

IGP border routers provide the IGP topology (or link state) information to a path computation engine, implemented by Cisco XR Transport Controller (XTC) using BGP LS. As a result, XTC has full view of the topology between the service edge routers, which are also Pseudo-Wire Head End (PWHE) capable devices. XTC builds SRTE Policy between service edge routers by Segment Routing On-Demand Next-hop (SR-ODN).

Figure 5: Large Network - End to End Programmable Segment Routing



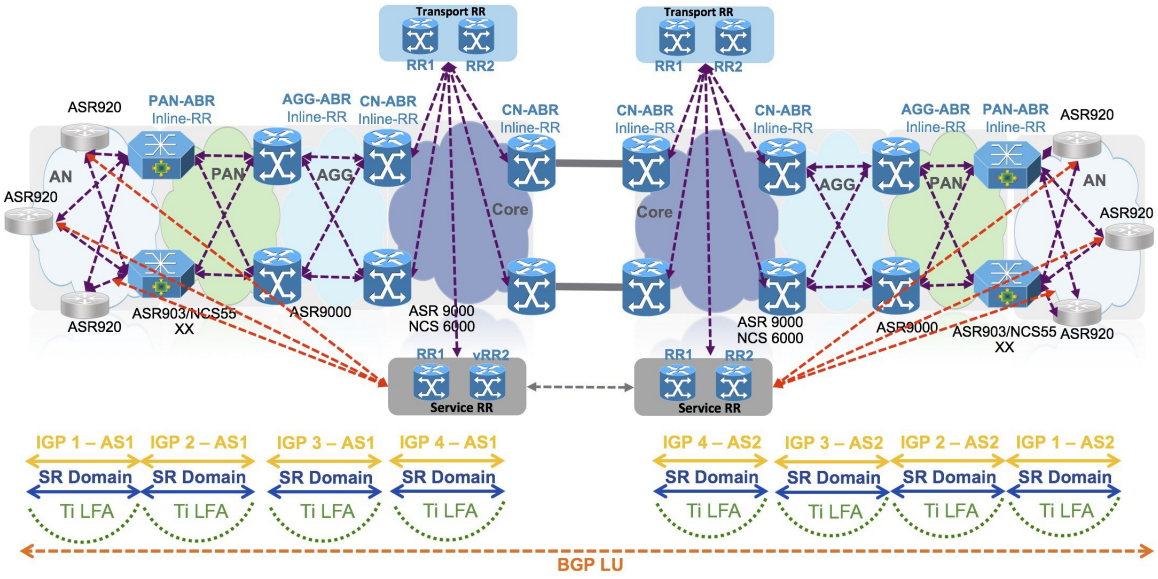
The characteristics of the network are described below:

- Each of the domains, including core, aggregation, and access, is running independent IGP with SR.
- Fast convergence within each domain is achieved with Topology Independent Loop Free Alternate (TI-LFA) technology using segment routing/IGP capabilities.
- The IGP in each domain could be OSPF single area (preferably area 0) or ISIS single level (preferably level-2).
- Each access domain is connected to the aggregation domain using two IGP Aggregation Area Border Routers (AGG-ABRs). AGG-ABRs are service edge routers and use PWHE feature to provide the service.
- Each aggregation domain is connected to the core layer.
- End-to-end transport is achieved using SRTE Policy (SR-ODN).
- The ABR pairs advertise the IGP topology information (link state information) to XTC through BGP LS.

Large Network - Inter-AS End to End Segment Routing

This section focuses on the network design of a large network with Inter-Autonomous System (Inter-AS). It is similar to Large Network - End to End Segment Routing with the main difference being that there two ASs interconnected through their core networks. The end-to-end transport is established using BGP LU with BGP prefix-SID extension as shown in the below figure.

Figure 6: Large Network - Inter-AS End to End Segment Routing



The characteristics of this network are listed below:

- There are multiple BGP autonomous systems.

- Each of the domains in a given AS - core, aggregation, pre-aggregation, and access is running on independent IGP with SR.
- Fast convergence within each domain is achieved with Topology Independent Loop Free Alternate (TI-LFA) technology using segment routing/IGP capabilities.
- The IGP in each domain could be OSPF single area (preferably area 0) or ISIS single level (preferably level-2).
- Each access domain is connected to the pre-aggregation domain using two IGP Area Border Routers (ABRs).
- Each pre-aggregation domain is connected to the corresponding aggregation layer.
- Each aggregation domain is connected to the core layer.
- An AS is interconnected to another AS using AS Border Routers (ASBRs), that establish eBGP sessions among each other, allowing transport routes to be exchanged through BGP-LU.
- End-to-end transport is achieved using BGP LU with BGP prefix SID as done for Small Network - End to End Segment Routing , on page 8 and Large Network - End to End Segment Routing , on page 9.
- Each access node and all the Area Border Routers (ABRs) participate in BGP LU.
- Each core ABR that connects core to aggregation ABR has iBGP peering to two dedicated transport Route Reflectors.
- Each ASBR, which is part of the core IGP, has also iBGP peering to two central T-RRs in its AS.
- The core ABRs are inline RR for the corresponding aggregation ABRs.
- The aggregation ABRs are inline RR for the corresponding pre-aggregation ABRs.
- Each pre-aggregation ABR is an inline RR for all corresponding access nodes in its region.
- BGP-PIC is used for Fast Reroute (FRR).

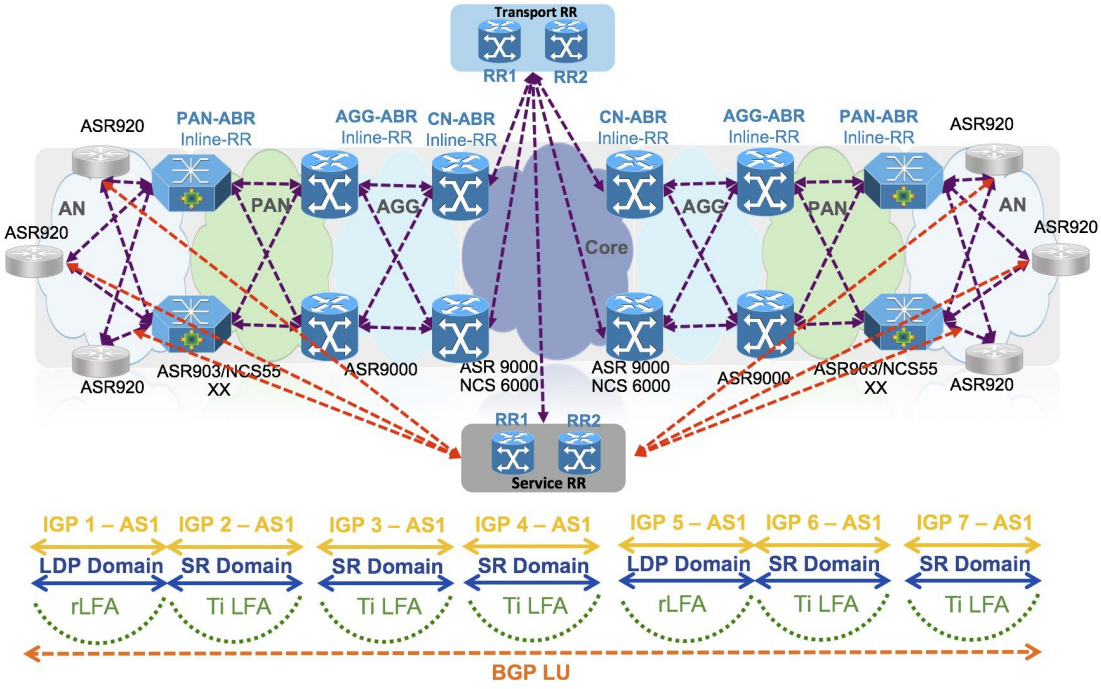
Large Network - Segment Routing and LDP Interworking

This section focuses on the design of the large network with some domains running traditional LDP, while others have been migrated to SR. It is similar to Large Network - End to End Segment Routing , on page 9 except with the following main diffrences:

- Some IGP domains use LDP, while other IGP domains use SR.
- BGP-LU without BGP prefix-SID (BGP prefix-SID is not end-to-end).

• Used BGP edge/core PIC for FRR across IGP domains.

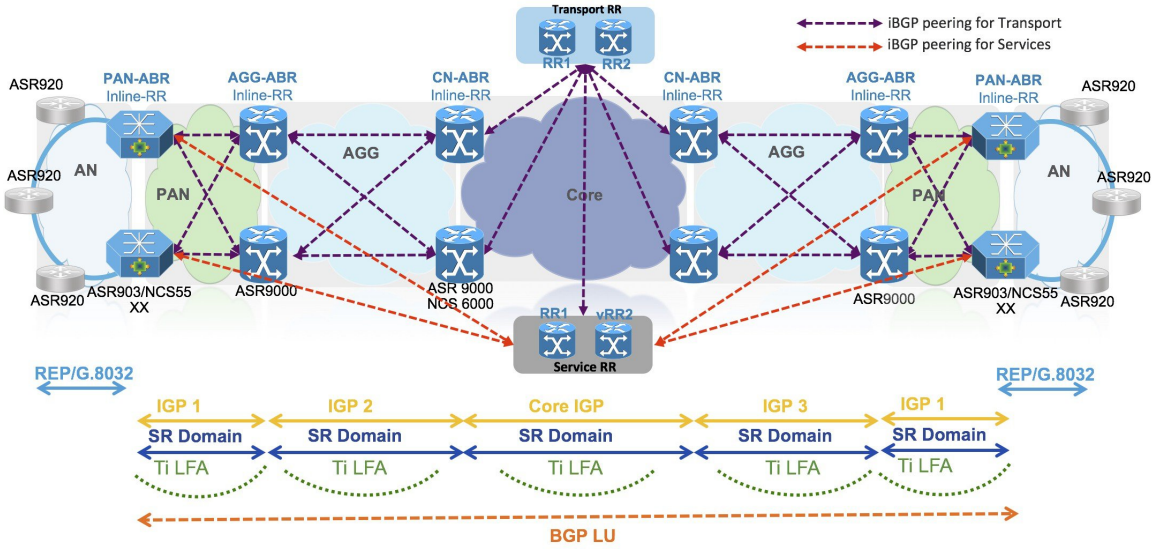
Figure 7: Large Network - Segment Routing and LDP Interworking



Large Network - Layer2 Access to Segment Routing Transport

This section focuses on the network design of a large network with four layers namely access, pre-aggregation, aggregation and core layers. The access layer uses legacy L2 protocol REP/G.8032, while the other domains use segment routing as transport protocol.

Figure 8: Large Network - Layer2 Access to Segment Routing Transport



The characteristics of this network are listed below:

- Domains such as core, aggregation and pre-aggregation are running independent IGP with segment routing.
- Access domain is running legacy L2 protocol REP/G.8032.
- Fast convergence within each domain is achieved with Topology Independent Loop Free Alternate (TI-LFA) technology using segment routing/IGP capabilities.
- The IGP in each domain could be OSPF single area (preferably area 0) or ISIS single level (preferably level-2).
- Each pre-aggregation domain is connected to the corresponding aggregation layer.
- Each aggregation domain is in turn connected to the core layer.
- End-to-end transport is achieved using BGP LU with BGP prefix SID.
- All the Area Border Routers (ABRs) participate in BGP LU.
- Each core ABR that connects core to aggregation ABR has iBGP peering to two dedicated transport route reflectors.
- The core ABRs are inline RR for the corresponding aggregation ABRs.
- The aggregation ABRs are inline RR for the corresponding pre-aggregation ABRs.
- Each pre-aggregation ABR is an inline RR for all the corresponding access nodes in their region.

- BGP-PIC is used for Fast Reroute (FRR).

Benefits

The use cases covered in this chapter are representative of possible deployment scenarios. The possible deployment scenarios are much more than the seven use cases covered here. This section discusses the benefits and applicability of each use case.

Greenfield Deployment

Operators deploying a new end-to-end network without the need to accommodate an existing deployment, have the option to select the hardware, software and orchestration system that provide the best value. In this scenario, the third use case - Large Network - End to End Programmable Segment Routing describes the best architecture. While the use case tested in Cisco EPN focuses on a large network, it is applicable for all the sizes of network. Programmable transport enables full SDN capability for the service provisioning. The service can request the underlying transport to meet the service requirements.

Alternatively, for operators who are not yet ready for a SDN enabled deployment, the first use case Small Network - End to End Segment Routing , second use case Large Network - End to End Segment Routing , or fourth use case Large Network - Inter-AS End to End Segment Routing is applicable depending on the size of the network.

Migrating from EPN 4.0 or Traditional Unified MPLS Transport

The Cisco EPN 4.0 transport uses BGP LU with LDP for interdomain connectivity and BGP PIC for inter-IGP convergence. There are multiple scenarios to migrate from Cisco EPN 4.0 to Cisco EPN 5.0 depending on the capabilities of the existing network devices:

1. Equipment is capable of supporting Cisco EPN 5.0 programmable transport with only a software upgrade.
2. There is a mix of equipments with and without Cisco EPN 5.0 programmable transport capability, but routers that do not support programmable transport are due for replacement.
3. There is a mix of equipments with and without Cisco EPN 5.0 programmable transport capability.

In the first and second scenarios, the network can be migrated to programmable transport to take advantage of full SDN capability.

In case of the third scenario, there is a mixed environment where some IGP domains are running LDP while IGP domains with capable hardware are running Segment Routing (SR). This scenario takes advantage of SDN capability only in SR domains and should be used when slower migration is needed.

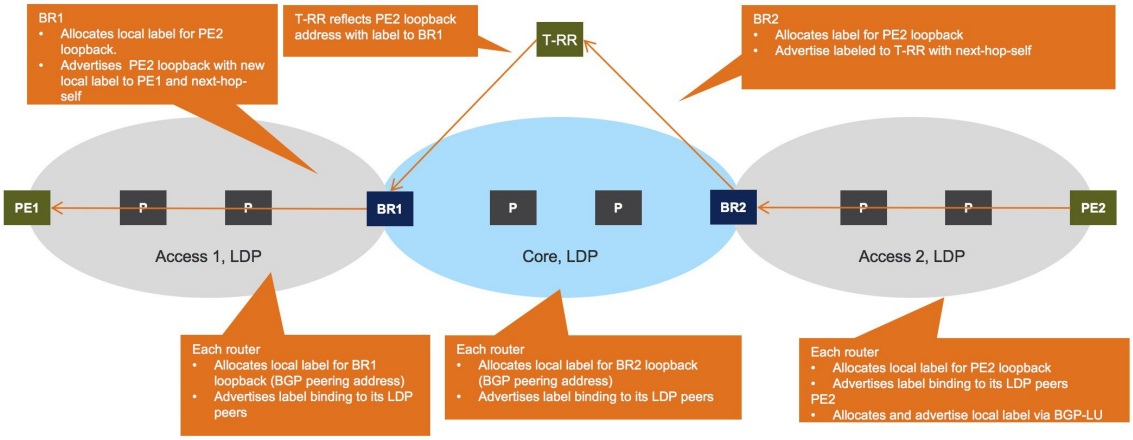
Control Plane and Data Plane Flow

This section describes the control plane and data plane flows for traffic going from PE1 (left) to PE2 (right). The control plane flow is described from right to left, while the corresponding data plane flow is described from left to right.

Traditional Unified MPLS - Control Plane

The control plane flow of the traditional unified MPLS is described from right (PE2) to left (PE1) as depicted in below figure. LDP is used in each IGP domain for label distribution.

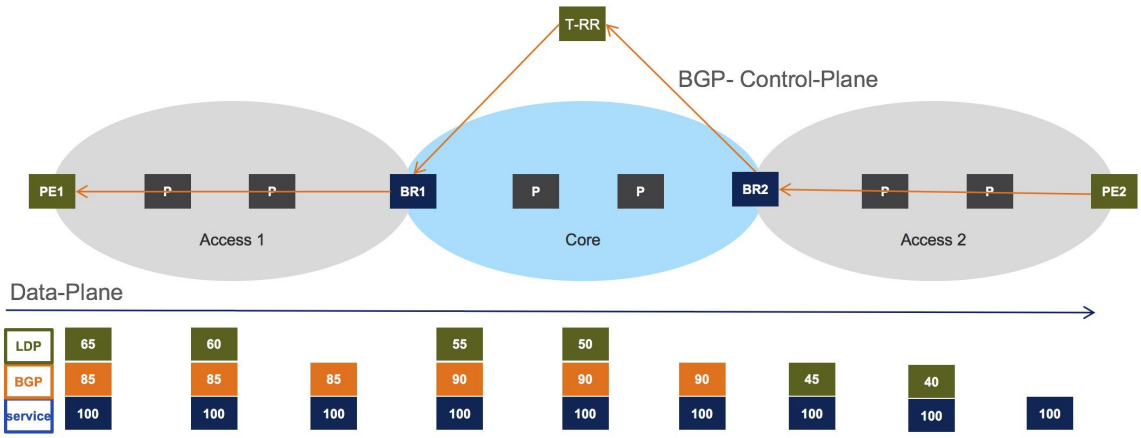
Figure 9: Traditional Unified MPLS - Control Plane



Traditional Unified MPLS - Data Plane

This section describes the data plane (traffic) flow and label imposition of traditional unified MPLS from left (PE1) to right (PE2) as depicted in the below figure.

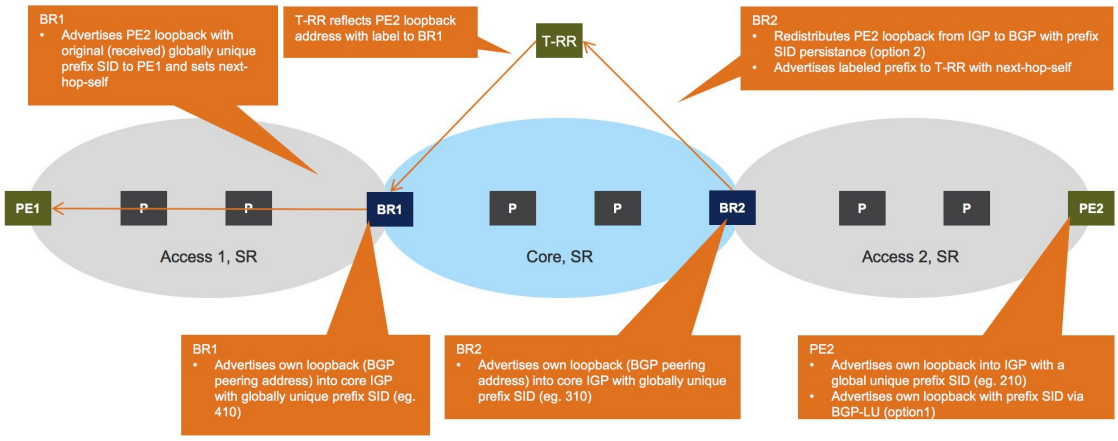
Figure 10: Traditional Unified MPLS - Data Plane



BGP LU with BGP Prefix SID - Control Plane

This section describes the control plane flow of BGP labeled unicast with BGP prefix SID from right (PE2) to left (PE1) as depicted in the below figure. Segment routing is enabled in each IGP domain and is used for label distribution.

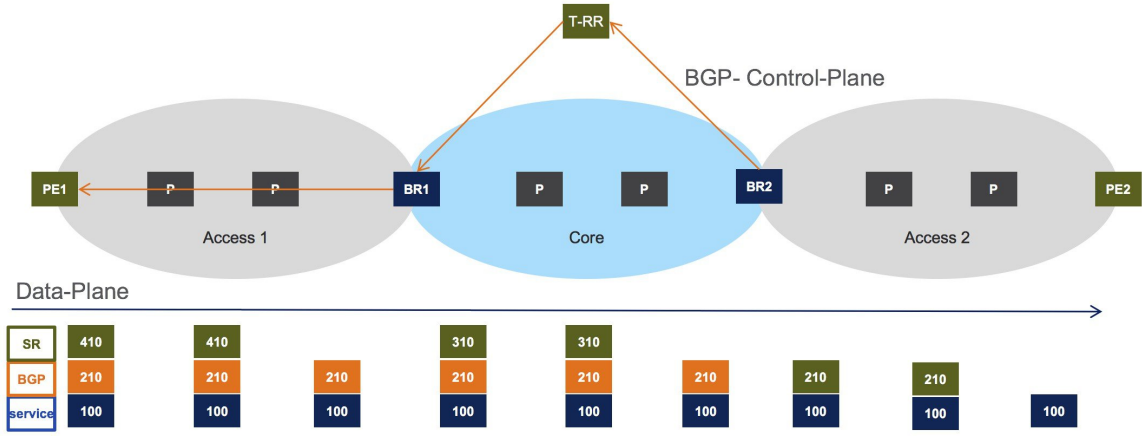
Figure 11: BGP LU with BGP Prefix SID - Control Plane



BGP LU with BGP Prefix SID - Data Plane

This section describes the data plane (traffic) flow and label imposition of BGP labeled unicast with BGP prefix SID from left (PE1) to right (PE2) as depicted in the below figure.

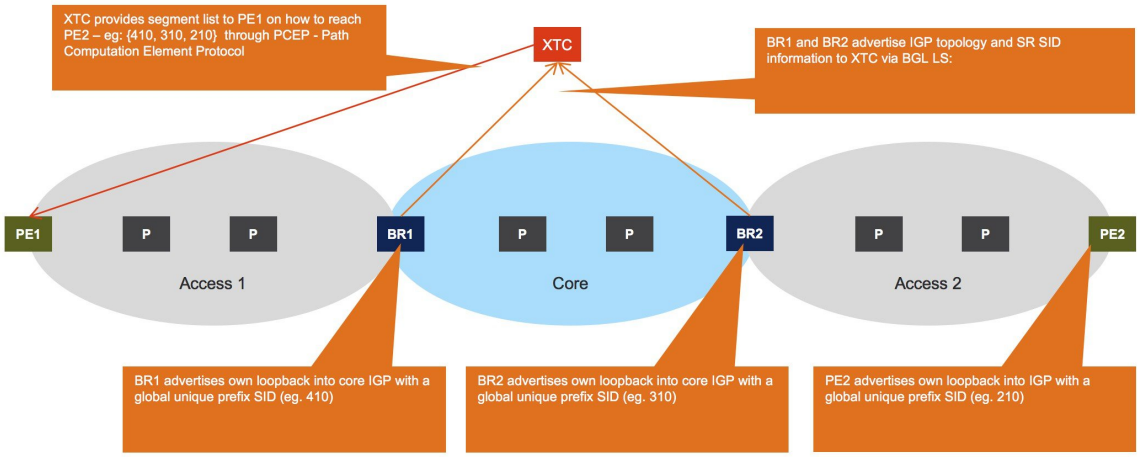
Figure 12: BGP LU with BGP Prefix SID - Data Plane



Programmable Transport - Control Plane

This section describes the control plane of programmable transport as depicted in the below figure. Segment routing is enabled in each IGP domain and is used for label distribution. More information about XR Transport Controller (XTC) and Path Computation Element Protocol (PCEP) can be found in section XR Transport Controller, on page 21.

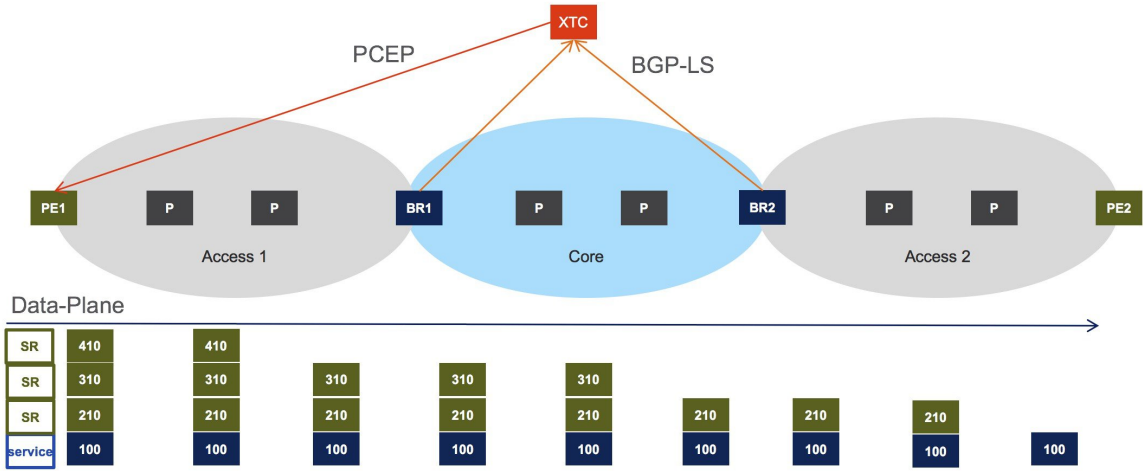
Figure 13: Programmable Transport - Control Plane

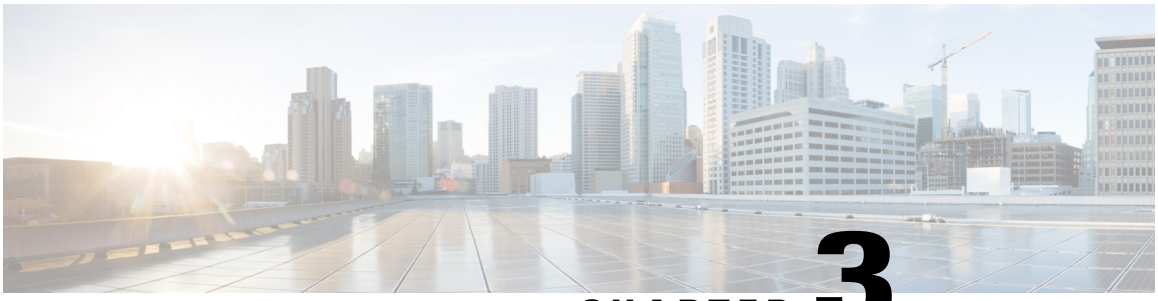


Programmable Transport - Data Plane

This section describes the data plane (traffic) flow and label imposition of the programmable transport from left (PE1) to right (PE2) as depicted in the below figure.

Figure 14: Programmable Transport - Data Plane





CHAPTER 3

Functional Components Design

This chapter contains the following sections:

- Redundancy and High Availability, page 19
- XR Transport Controller, page 21
- Border Gateway Protocol Route Reflector Architecture, page 22
- Quality of Service, page 24

Redundancy and High Availability

Cisco EPN system architecture implements high availability at both the transport network level and service level. By utilizing a combination of several technologies throughout the network, the Cisco EPN design is capable of meeting stringent availability Service Level Agreements (SLAs).

This chapter covers the implementation of high availability technologies for the transport level. It focuses on fast recovery from failure using the transport layer. The service level resiliency implementations are covered in the respective service implementation guides.

High availability at the transport network layer is provided through the combination of several technologies including:

- Loop Free Alternate (LFA) and remote-LFA (rLFA) Fast Rerouting with BFD.
- Topology Independent LFA (TI-LFA).
- BGP edge and core Prefix Independent Convergence (PIC).
- Bi-directional Forward Detection (BFD) at the IGP.

Loop-Free Alternate with Bidirectional Forwarding

Loop-free Alternate Fast Rerouting (LFA FRR) pre-calculates a backup path for every prefix in the IGP routing table, and allows the node to rapidly switch to the backup path with recovery times on the order of 50 msec, when a failure is encountered. Remote LFA FRR functionality extends LFA FRR functionality to ring networks and other topologies.

LFA is also integrated with Bidirectional Forwarding (BFD) rapid failure detection and Intermediate System to Isntermediate System (IS-IS)/OSPF extensions for incremental Shortest-Path First (SPF) and Link-State Advertisement (LSA)/SPF throttling.

More information regarding LFA FRR can be found in IETF RFC 5286, RFC 5714 and RFC 6571.

Microloop Avoidance in Remote Loop-Free Alternate Fast Reroute

In a network comprised of different platforms, the convergence time of one platform might be faster than that of the other platforms. This difference in the convergence time can lead to a condition where a node is forwarding traffic to the same neighboring node from which traffic was being received prior to the topology change. This is referred to as a microloop within the topology.

With remote LFA-FRR activated, the backup path is used until the computing node learns about the topology change and reinstalls new paths for the prefix. Microloops are likely to occur based on the convergence period of the computing node with respect to its neighbors. To prevent the occurrence of microloops, a microloop avoidance mechanism is provided to postpone the protected prefixes by an additional delay, to allow for convergence in its neighbors.

Segment Routing Topology Independent Loop-Free Alternate with Bidirectional Forwarding

SR based TI-LFA is capable of building backup path without establishing targeted Label Distribution Protocol (tLDP) or creating backup Resource Reservation Protocol - Traffic Engineering (RSVP-TE) tunnels. TI-LFA with BFD can provide backup path on post-convergence against the link or node failure in any Interior Gateway Protocol (IGP) network. The benefits of TI-LFA with BFD are given below:

- Eliminates the need to establish targeted LDP (tLDP) session with remote nodes.
- Uses segment list to construct the repair path.
- Eliminates the need to create state in the network to enforce an explicit FRR path, since only the Point of Local Repair (PLR) need to have the state of the FRR path.
- Automatically calculates detour path for each specific destination and for each possible failure in the network.
- The FRR path follows the post-convergence path - a single change from pre-convergence to post-convergence as opposed to three changes (from pre-failure to FRR to post-convergence), which is the case for LDP based FRR.
- TI-LFA algorithm reduces the number of label stacks for the repair path following the post-convergence path.

Also integrated are BFD rapid failure detection and IS-IS/OPSF extensions for incremental Shortest Path First (SPF) and Link State Advertisement (LSA)/SPF throttling. More information regarding LFA FRR can be found in IETF RFC 5286, RFC 5714 and RFC 6571.

Border Gateway Protocol — Fast Reroute — BGP Prefix Independent Convergence Core/Edge

The BGP FRR provides deterministic network re-convergence, even if the BGP prefix scale is encountered in the Cisco EPN System design. The BGP FRR pre-calculates a backup path similar to remote Loop-Free Alternate FRR (LFA FRR), for every prefix in the BGP forwarding table. This backup path precalculation is performed based on a hierarchical Label Forwarding Information Base (LFIB) structure, to allow multiple paths to be installed for a single BGP next-hop. The BGP FRR consists of two different functions namely core and edge, each of which is used to handle different failure scenarios within the transport network as given below:

- FRR core protection is used when the BGP next-hop is still active, but there is a failure in the path to that next hop. As soon as the IGP has re-converged, the pointer in BGP is updated to use the new IGP next hop and forwarding resumes. Thus, the re-convergence time for BGP is same as that of the IGP re-convergence, regardless of the number of BGP prefixes in the Routing Information Base (RIB).
- FRR edge protection is used for redundant BGP next-hop nodes in the case of redundant Area Border Routers (ABRs). BGP additional-path functionality is configured on the Provider Edge (PE) routers and Route Reflectors (RRs), in order to install both ABR nodes' paths in the RIB and LFIB instead of just the best path. When the primary ABR fails, BGP forwarding simply switches to the path of the backup ABR, instead of having to wait for the BGP to re-converge at the time of the failure.

XR Transport Controller

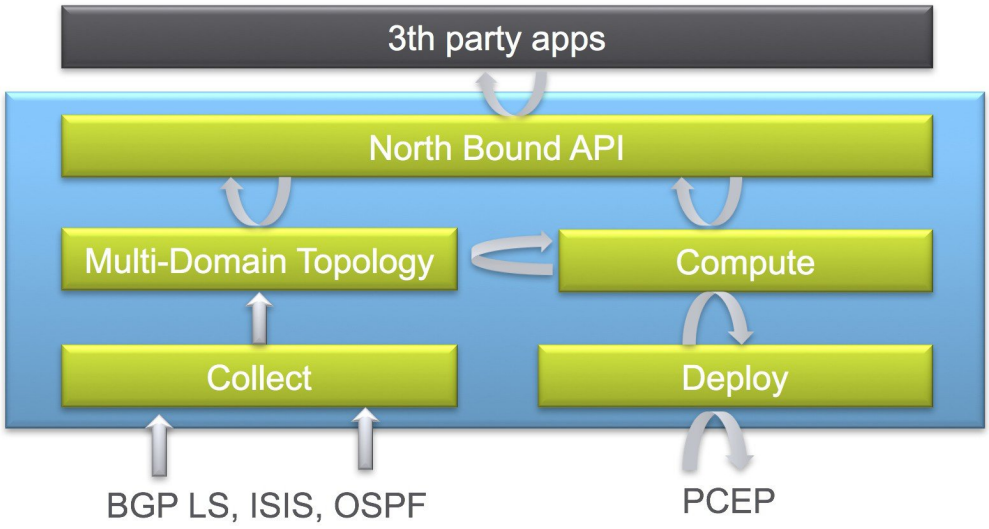
XR Transport controller, or XTC, is Cisco Path Computation Engine (PCE) and it is implemented as a feature included as part of Cisco IOS-XR operating system. The function is typically deployed on a Cisco IOS-XR cloud appliance XRv9000, as it involves control plane operations only. The XTC gains network topology awareness from BGP-LS advertisements received from the underlaying network. Such knowledge is leveraged by the embedded multi-domain computation engine to provide optimal path to Path Computation Element Clients using the Path Computation Element Protocol (PCEP). The PCC is the device where the service originates and therefore requires end-to-end path over segment routing multi-domain network.

The XTC provides a path based on PCC constraints like:

- Shortest path (IGP metrics).
- Traffic-Engineering metrics.

- Disjoint path.

Figure 15: XR Transport Controller



Border Gateway Protocol Route Reflector Architecture

The Border Gateway Protocol (BGP) is used for multiple purposes including:

- BGP labeled unicast (BGP-LU) (BGP 3107)with or without BGP Prefix SID for unified multiprotocol label switching (MPLS) architecture.
- Service route exchange including layer3 VPN (L3VPN), Virtual Private LAN Service (VPLS), Virtual Private Wire Service (VPWS), and IPv4/IPv6 Internet. In case of VPLS and VPWS, both signaling and discovery are enabled.
- BGP-LS for delivering topology or link state data to controller.

It is important to use BGP scaling methodology, and in particular route-reflector, in any small or large production size deployment.

Consider the following factors while designing route-reflector architecture:

- BGP LU (BGP 3107) requires inline route-reflection at each IGP boundary (IGP border routers) to implement unified MPLS.
- Service edge routers, IGP border routers and all BGP route-reflectors have to participate in BGP LU exchange.
- Only service edge routers require service routes. All other routers, with the exception of corresponding route-reflectors need not carry service routes such as L3VPN, VPLS, VPWS, or IPv4/IPv6 Internet.
- Only controllers require topology or link state data. IGP border routers need to export topology data towards controller through the corresponding route-reflector.

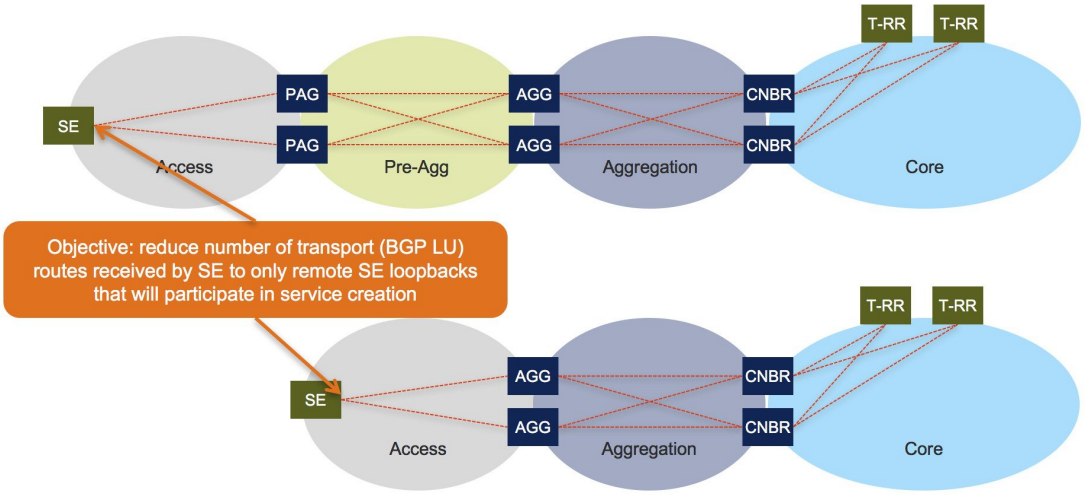
- Most access devices have limited route table size. Therefore, it is important to implement route-filtering techniques for both transport and service routes, to reduce the amount of routes received by access devices.

Transport Route Reflector Design

The Transport Route Reflector (T-RR) is designed such that:

- Each AS has a pair of transport route-reflectors (T-RR).
- The Route Reflector (RR) clients for T-RR are the IGP border routers of the core network.
- Border routers are inline router-reflectors for border router in the next lower layer or the service edge routers.

Figure 16: Transport Route Reflector Design



Transport route reduction for service edge routers is achieved by performing the following:

On SE routers

- During initial setup, create prefix-list with two loopback addresses to permit the corresponding regional service RR pairs.
- Attach prefix-list inbound on BGP peer towards SE.
- On service creation or deletion, the orchestrator (NSO).
    - Adds or removes loopback IP addresses of remote SEs involved in the service, to or from the prefix list respectively.
    - Issues BGP soft clear outbound towards SE.

Service Route Reflector Design

When a Service Route Reflector (S-RR) is on a growth trajectory and serving more than 5000 service edge or Provider Edge (PE) routers in the network, it is important to consider hierarchical RR design for scalability. The hierarchical service RR is designed such that:

- Two RRs are used in the core of the network.
- Network is split into multiple regions, each serving about 1000 service edge or provider edge routers. It is important to have sufficient room for future growth of PEs in the region while the IOS XR supports 5000 BGP peers.
- Each network region has two service RRs servicing the PEs.

Note The physical location of the RRs with respect to the PEs is not critical. However, if possible avoid placing the pair of RRs in the same location, to minimize the probability that both are failing at the same time.

- The regional S-RRs are RR clients of the two core S-RRs. It is recommended to have two RR for all RR deployments.

In case of L3VPN (IPv4 and IPv6), the service route reduction for SE routers is accomplished using Route Target Constraint (RTC) feature. RTC is a BGP capability specified in RFC4684.

- Normally, S-RR sends all the service routes to the client SE routers. But typically, only a small subset of the routes is needed by the given SE.
- The SE will install only the required routes to its forwarding table. However, it has to receive and process all routes before installation.
- With RTC feature, the SE will automatically send route filter to the S-RRs, then S-RR apply this filter before sending the routes to SE.
- To activate RTC between a client (SE) and an S-RR:
  - On IOS XR, configure: " address-family ipv4 rt-filter" for the neighbor.
  - On IOS XE, configure: "address-family rtfilter unicast" for the neighbor.

Quality of Service

Although congestion is more likely where statistical estimates of peak demand are conservative (that is, under-provisioned), such as in access and aggregation links, it can occur anywhere in a transport network. Therefore, all nodes in a transport network are required to implement congestion management techniques, which involve classification and proper scheduling functions.

The Cisco EPN System applies the Differentiated Services (DiffServ) architecture that is defined by the Internet Engineering Task Force (IETF) in RFC 2475 across all network layers, based on the classification mechanisms like MPLS Experimental (EXP) bits, IP DSCP, IEEE 802.1p and ATM CoS, for implementing the DiffServ per-hop behaviors (DiffServ PHBs). Within the aggregation and core networks, where strict control over consumer and enterprise subscriber's SLA is not required, a flat QoS policy with a single-level

scheduler is sufficient for achieving the desired DiffServ functionality among the different classes of traffic, as all links are operated at full line rate transmission.

Hierarchical QoS (H-QoS) policies are required, whenever the relative priorities across the different classes of traffic are significant within the level of service offered to a given subscriber and/or within a given service category such as consumer, enterprise or mobile. In downstream direction, H-QoS for a given subscriber should be performed at the SE node whenever possible to guarantee the most optimal usage of link bandwidth throughout the access network.

For an Ethernet-based access Network-to-Network Interface (NNI) and consumer services, the SE node acting as BNG device:

- Applies QoS at the subscriber level, with per-subscriber queuing and scheduling.
- Applies QoS at the aggregate level for all consumer subscribers sharing the same N:1 VLAN or a range of 1:1 VLANs.

Aggregated QoS at the consumer service level is beneficial to manage oversubscription of the AN from residential traffic and to control sharing of the access-facing NNI bandwidth with mobile and enterprise services. Similarly, enterprise service interfaces at the service edge implementing H-QoS for the deployment of subscriber level SLAs and sharing of the access bandwidth.

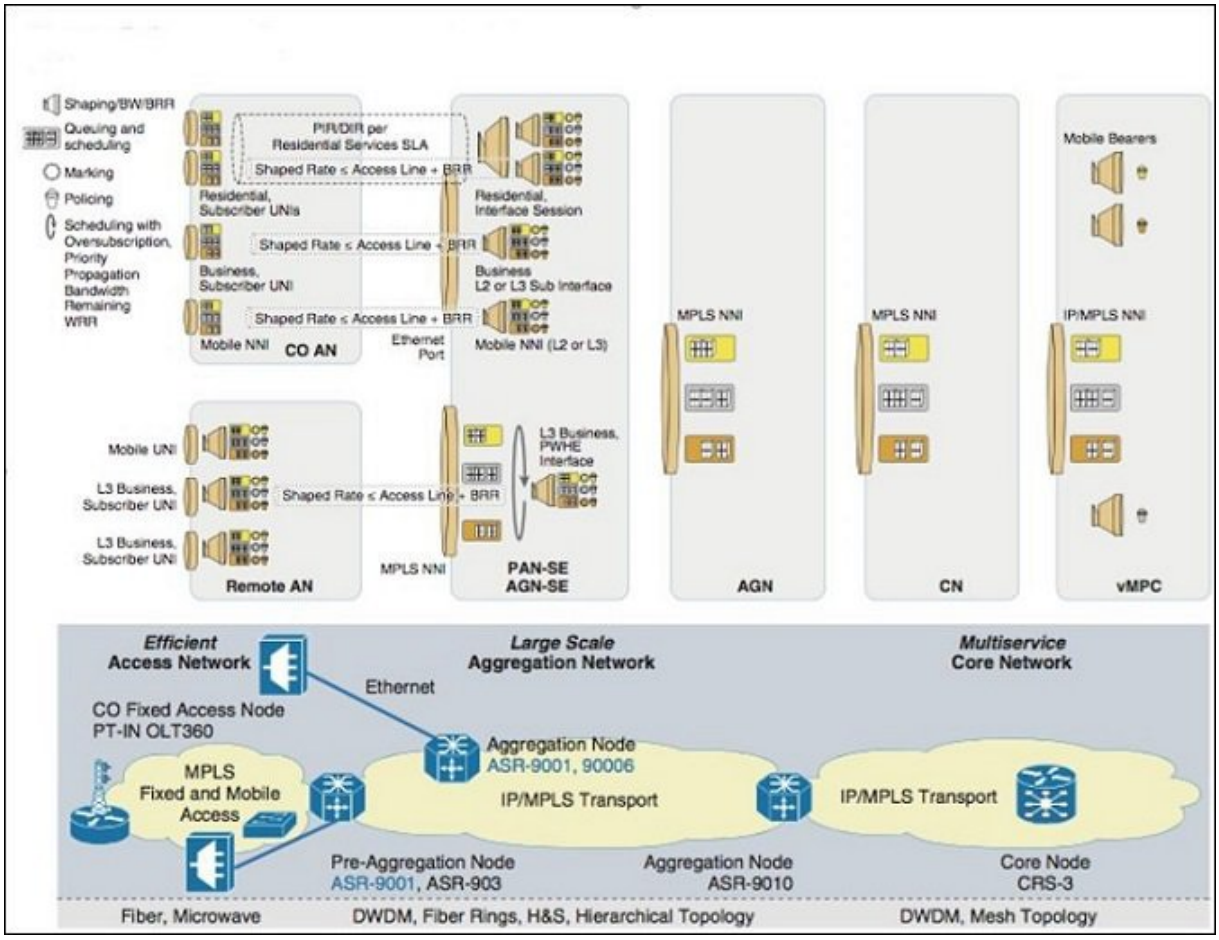
Mobile services also require the implementation of H-QoS for sharing of access bandwidth. In case of microwave links where the wireless portion of the link is only capable of sub-Gb speeds (typically 400 Mbps sustained), a parent shaper may be used to throttle transmission to the sustained microwave link speed.

Whenever subscriber SLAs are managed at the service edge and the access UNI is not multiplexed, you can apply a flat QoS policy to the AN to manage relative priority among the classes of traffic at each UNI port. Multiplexed User Network Interface (UNI) dedicated for enterprise services, require an H-QoS policy for relative prioritization among the services initially and then between classes of traffic within each service. In those scenarios, H-QoS on the SE nodes may drive peak information rate (PIR) level traffic, while the access UNI may force the Committed Information Rate (CIR) levels.

For an MPLS-based NNI, most services do not have a corresponding attachment point at the SE node. Therefore, the majority of the service level H-QoS logic happens at the AN. The exception are the L3VPN enterprise services for which the Customer-Edge to Provider-Edge (CE-PE) LSP is terminated over a pseudo wire

headend (PWHE) interface at the SE node, being the injection point for H-QoS. The figure below depicts the downstream QoS model.

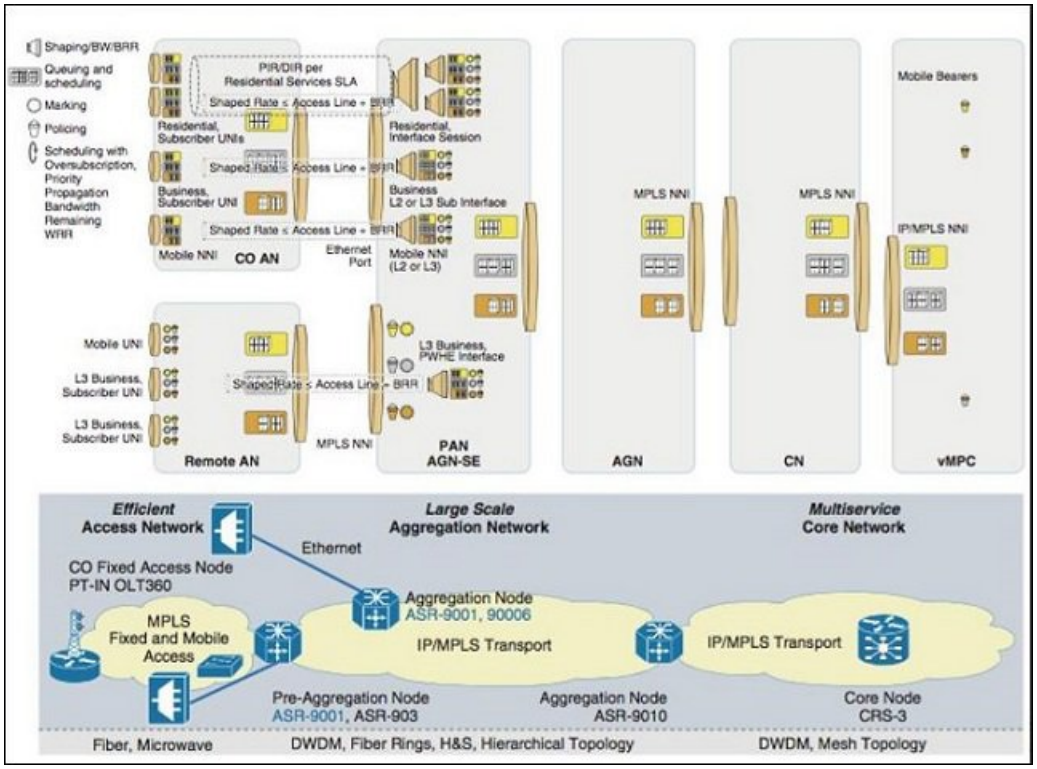
Figure 17: Downstream—QoS model



Upstream QoS mainly involves applying flat egress QoS policies to various network nodes for relative prioritization among the different classes of traffic. Additionally, ingress QoS is required at both the AN UNI and at the SE node NNI, to enforce per subscriber SLAs', when an attachment point for the policy is available. At the SE nodes, you can use ingress-coupled policers to throttle the overall subscriber transmission rate at

the committed speed, while ensuring minimum bandwidth guarantee for traffic classes with full subscriber transmission rate. The figure below depicts the upstream QoS model.

Figure 18: Upstream—QoS model



Administrative boundaries are implemented at the access nodes for the traffic received (upstream) from the subscribers and at the core nodes for the traffic sent (downstream) to the subscribers, to control traffic markings within the operator's network. Moreover, administrative boundaries may also exist across domains managed by different operators. Thus, further traffic remarking may be required to correlate the PHBs between different administrative domains.

The figure shown below depicts the DiffServ QoS model. The traffic classification, marking and DiffServ PHB behaviors in system architecture are targeted to fit the deployment of consumer, enterprise, mobile and Metro Ethernet Forum (MEF) transport services. Traffic across all the three services is divided into three main categories:

- Expedited forwarding (EF).
- Assured forwarding (AF) and Class Selector (CS).

- Best effort (BE).

Figure 19: DiffServ QoS model

Traffic Class	PHB	Unified MPLS Transport		Service Edge			Fixed/Mobile Access Ethernet/TDM/ATM UNI		
		Core, Aggregation, Access		Business PWHE		Res/Bus Ethernet	M	R, B, M	M, B
		DSCP	EXP	DSCP	EXP	802.1P	DSCP	802.1P	ATM
Network Management	CS	56	7	56	7	7	56	(7)	VBR-nrt
Network Control Protocols	CS	48	6	48	6	6	48	(6)	VBR-nrt
Residential Voice Business Realtime Network Sync (1588 PTP) Mobility & Signalling Traffic Mobile Conversation/Streaming	EF	46	5	46	5	5	46	5	CBR
Residential TV and Video Distribution	AF	32	4	32	4	4	NA	4	NA
Business Telepresence	AF	24	3	24	3	3	NA	3	NA
Business Critical In Contract Out of Contract	AF	16 8	2 1	16 8	2 1	2 1	16 8	2 1	VBR-nrt
Residential HSI Business Best Effort Mobile Background VQE Fast Channel Change, Repair	BE	0	0	0	0	0	0	0	UBR

Traffic marked as EF is grouped in a single class serviced with priority treatment, to satisfy stringent latency and delay variation requirements. The EF PHB defines a scheduling logic to guarantee an upper limit to the per hop delay variation that is caused by the packets from non-EF services.

This category includes residential voice and business real-time traffic, mobile network timing synchronization (1588 PTP) and mobile signaling and conversation traffic (GSM Abis, UMTS IuB control plane and voice user plane, LTE S1c, X2c, and LTE guaranteed bit rate (GBR) user plane).

Traffic marked as AF is divided over multiple classes. Each class is guaranteed a predefined amount of bandwidth. Thus, maintaining relative priorities among the classes and limiting the amount of latency traffic that may occur in each class.

The Cisco EPN system defines five CS/AF classes. Two classes are reserved for network traffic, control and management, while the remaining three classes are dedicated to handle the traffic from residential and business services such as residential TV and video distribution, business TelePresence and mission-critical applications.

The third category, best effort (BE), encompasses all traffic that can be transmitted only after all the other classes have been served within their fair share. This traffic is neither time nor delay sensitive. This category manages residential H.323 Signaling Interface (HSI) traffic, mobile background traffic and video quality control traffic.

For Ethernet UNI, upstream traffic classification is based on IP Differentiated Services Code Point (DSCP) or 802.1P CoS markings. The ingress QoS policy matches and maps these markings to the corresponding DSCP and/or MPLS EXP value, depending on the access NNI being Ethernet or MPLS. In the downstream direction, IP DSCP markings are preserved through the unified MPLS transport and may be used for queuing and scheduling at the UNI as well as for restoring 802.1P CoS values.

Specifically to mobile services, TDM UNI interfaces transported through CEoP pseudowires require all traffic to be classified as real-time with EF PHB. The ingress QoS policy matches all traffic inbound to the interface, and applies an MPLS EXP value of 5. No egress service policy is required for TDM UNI interfaces.

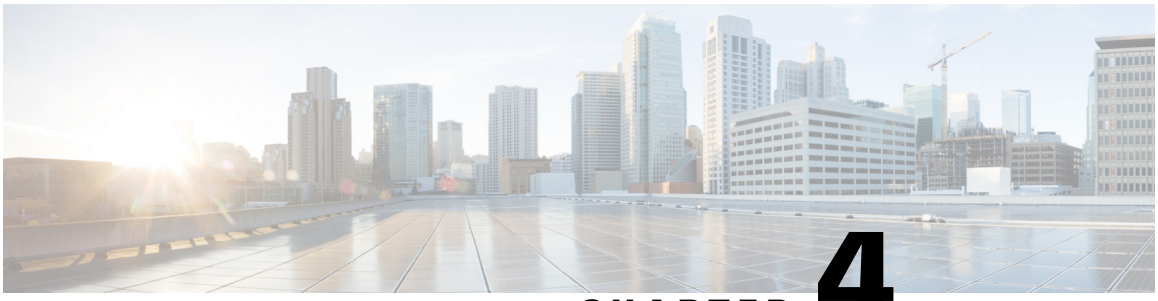
For ATM UNI interfaces to be transported through CEoP pseudowires or used for business services, the traffic is classified according to the ATM CoS on a particular VC. You can apply the ingress QoS service policy to the ATM Permanent Virtual Circuit (PVC) subinterface, so that it maps to an MPLS EXP value that corresponds

to the type of traffic carried on the VC and proper ATM CoS. For further distinction, the ingress QoS policy may have the ability to match on the cell loss priority (CLP) bit of the incoming ATM traffic, and to map to two different MPLS EXP values. For egress treatment, the PVC interface is configured with the proper ATM CoS. If the CLP-to-EXP mapping is being used, then an egress QoS policy applied to the ATM PVC subinterface can map an EXP value back to a CLP value for proper egress treatment of the ATM cells.

At the SE node, the classification performed at the access-facing NNI may use a different set of marking depending on the technology used. For an Ethernet-based access NNI in upstream direction, the classification is based on IP DSCP or 802.1P CoS markings. The ingress QoS policy matches and maps these markings to the corresponding MPLS EXP value for transport toward the core. In the downstream direction, the IP DSCP markings are preserved through the unified MPLS Transport and may be used for queuing and scheduling as well as for restoring 802.1P CoS values before forwarding.

For a MPLS-based access NNI in upstream direction, the classification is based on IP DSCP or MPLS EXP markings. The ingress QoS policy matches the markings, which are retained when forwarding towards the core. In the downstream direction, you can use the IP DSCP or MPLS EXP markings preserved through the unified MPLS Transport for queuing and scheduling toward the access NNI.

For all the core, aggregation, and access networks, the traffic classification is based on MPLS EXP or DSCP. The core network may use different traffic marking and simplified PHB behaviors, thus traffic remarking in between the aggregation and core networks may be required.



CHAPTER 4

Conclusion

The Cisco EPN 5 System gives operators a proven architecture, platforms, and solutions to address the dramatic changes in subscriber behavior and consumption of communications services and provides operational simplification, all at optimized cost points.

EPN 5 helps to migrate from well established Unified MPLS design by seamless integration with existing network.

Segment-Routing as a key transport protocol of EPN 5 brings enormous network simplification together with SDN programmability (SR TE Policy), Fast convergence/High-Availability (FRR TI-LFA) and natural integration with SDN transport controllers such as XR Transport Controller (XTC).

Simplicity, Scalability, High-Availability, Agility and Programmability are main requirements of the next generation Service Provider Networks and EPN 5 addresses all of them.