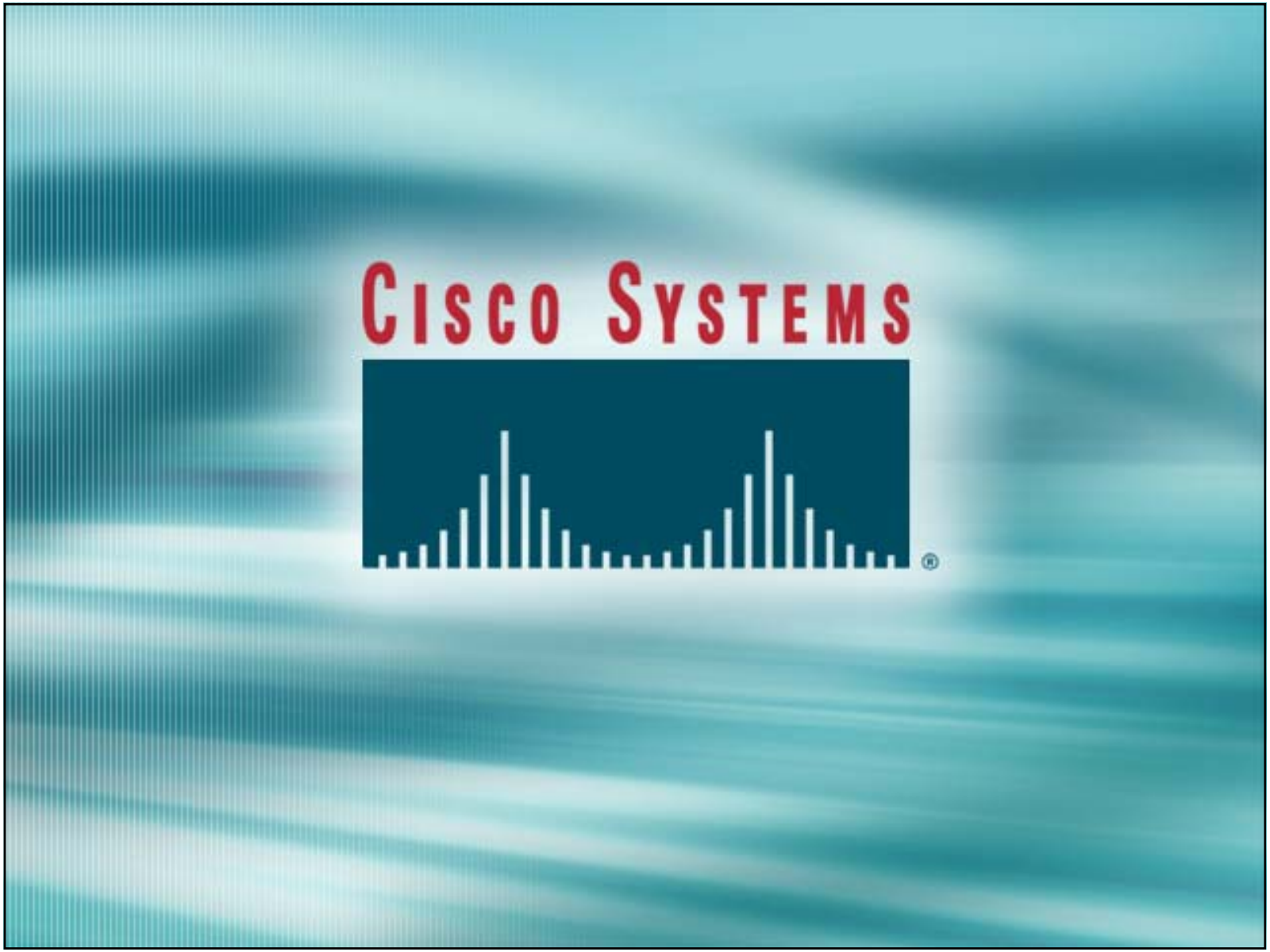
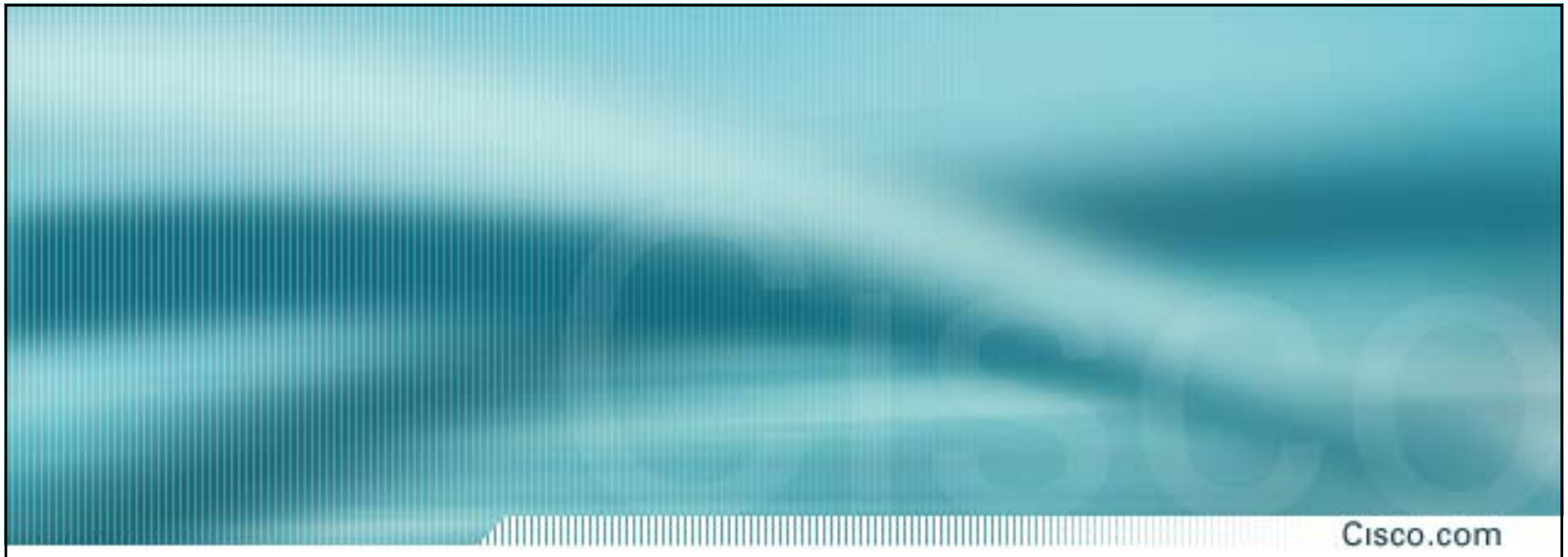

CISCO SYSTEMS

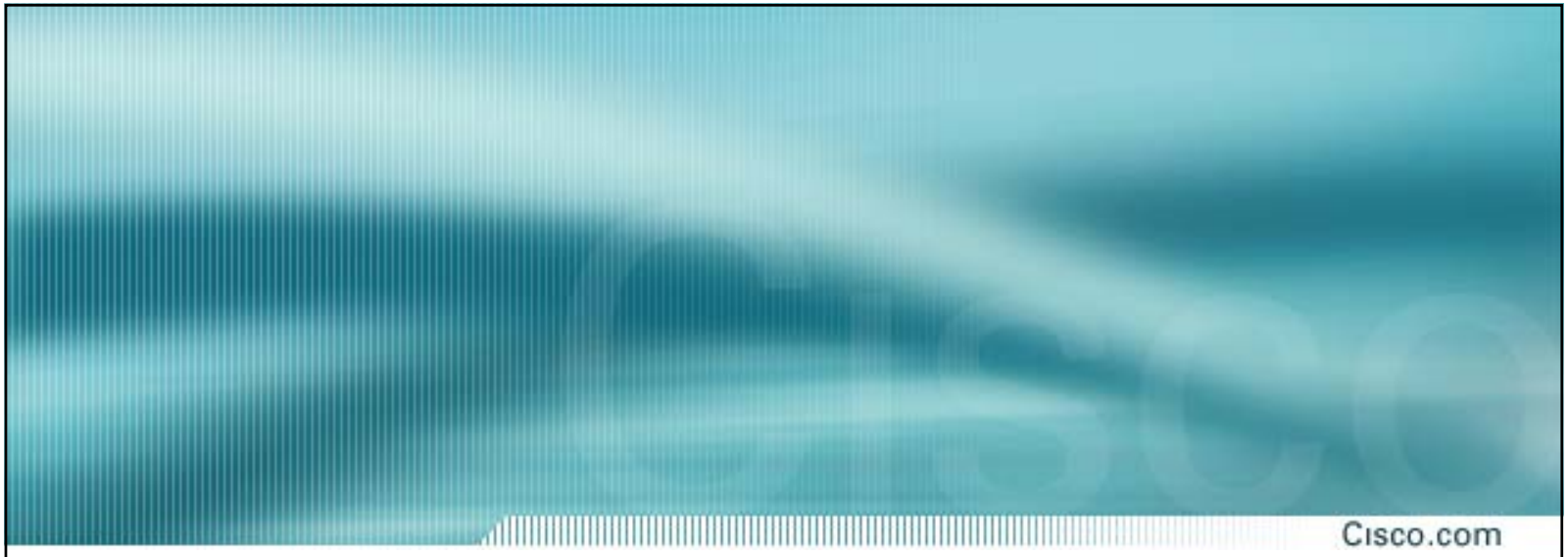


Deploying Scalable Core Network for IP and MPLS Services and Best Practices

Agenda

Cisco.com

- ISP Routing and Common Issues – Quick Overview
- Scaling IGP in Service Provider Networks
- Scaling BGP in Service Provider Networks
- Deploying Scalable MPLS VPN



ISP Routing—Quick Review

What Is an IGP?

- Interior Gateway Protocol
- Within an autonomous system
- Carries information about internal infrastructure prefixes
- Examples—OSPF, ISIS, RIPv2…

Why Do We Need an IGP?

- **ISP backbone scaling**
  - **Hierarchy**
  - **Modular infrastructure construction**
  - **Limiting scope of failure**
  - **Healing of infrastructure faults using dynamic routing with fast convergence**

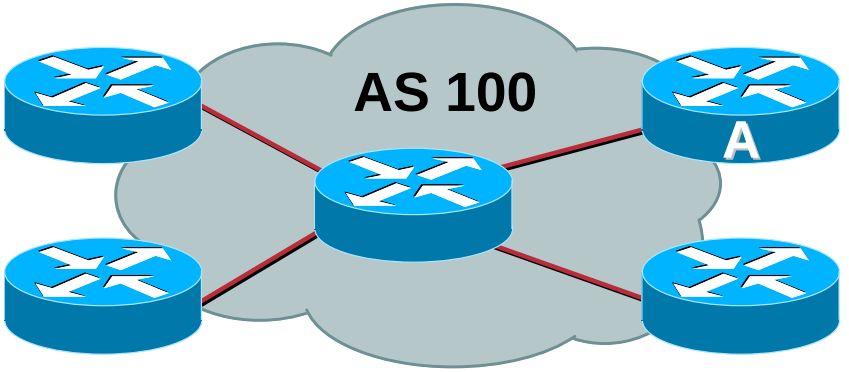
What Is an EGP?

- **Exterior Gateway Protocol**
- **Used to convey routing information between autonomous systems**
- **De-coupled from the IGP**
- **Current EGP is BGPv4**

Why Do We Need an EGP?

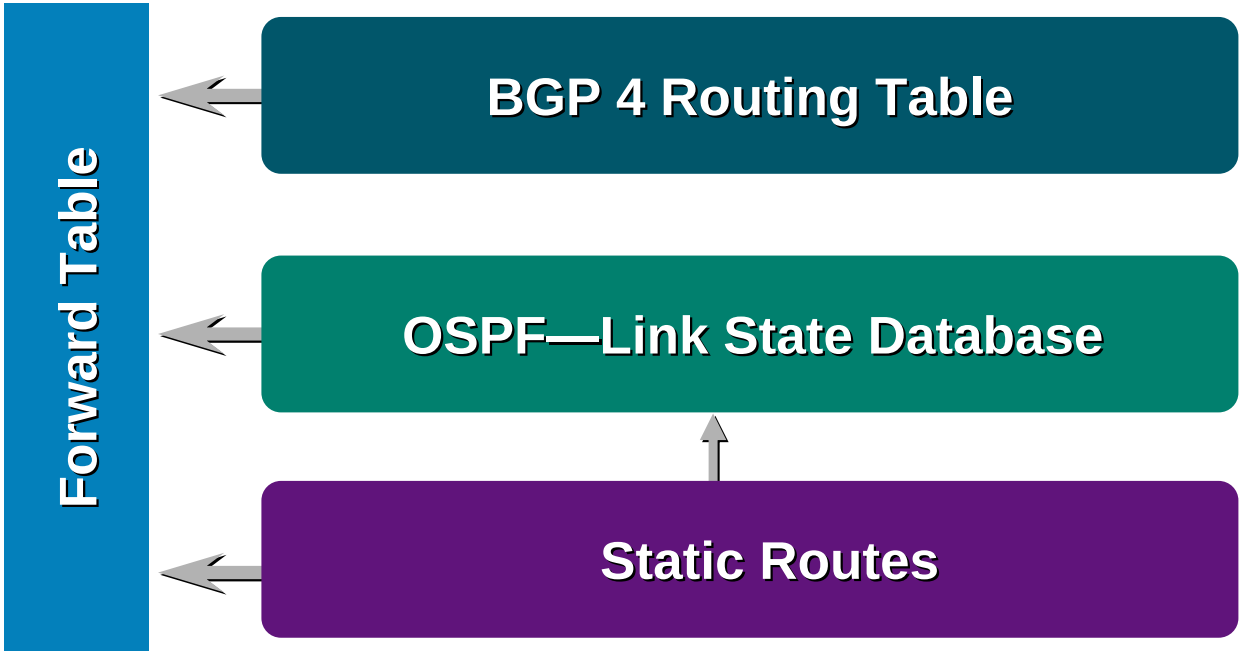
- **Scaling to large network**
  - **Hierarchy**
  - **Limit scope of failure**
- **Policy**
  - **Control reachability to prefixes**
  - **Merge separate organizations**
  - **Connect multiple IGPs**

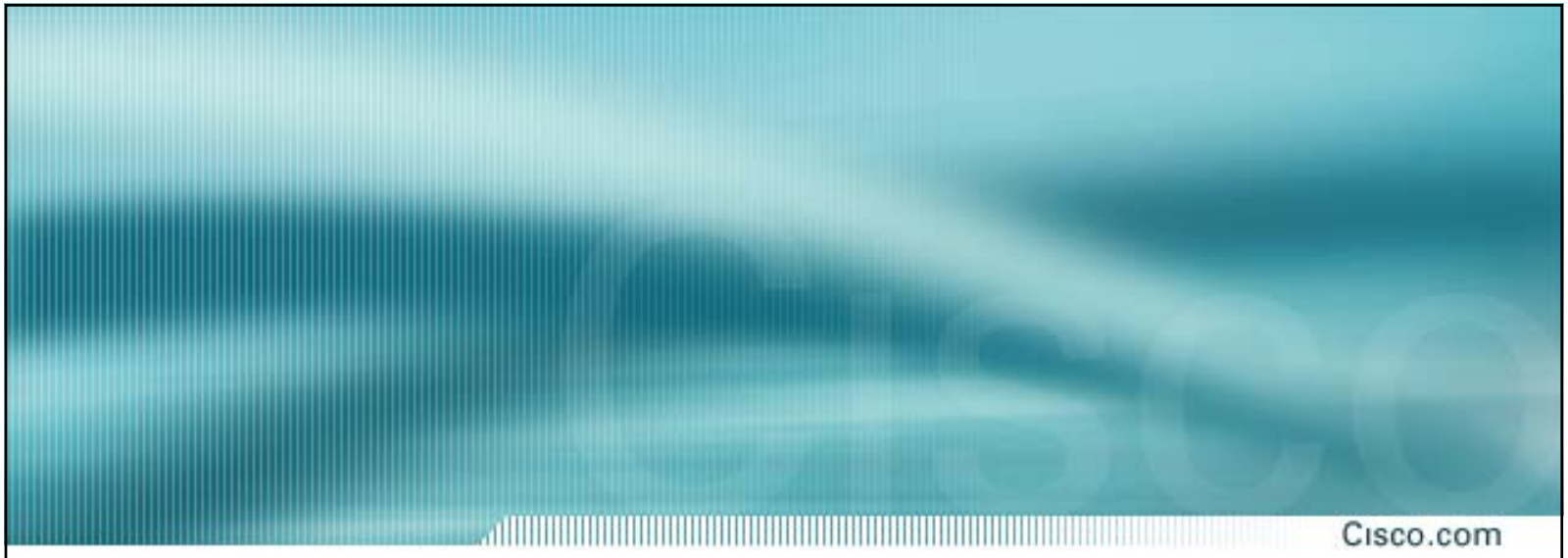
Autonomous System (AS)



- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control

Routing Tables Feed the Forwarding Table



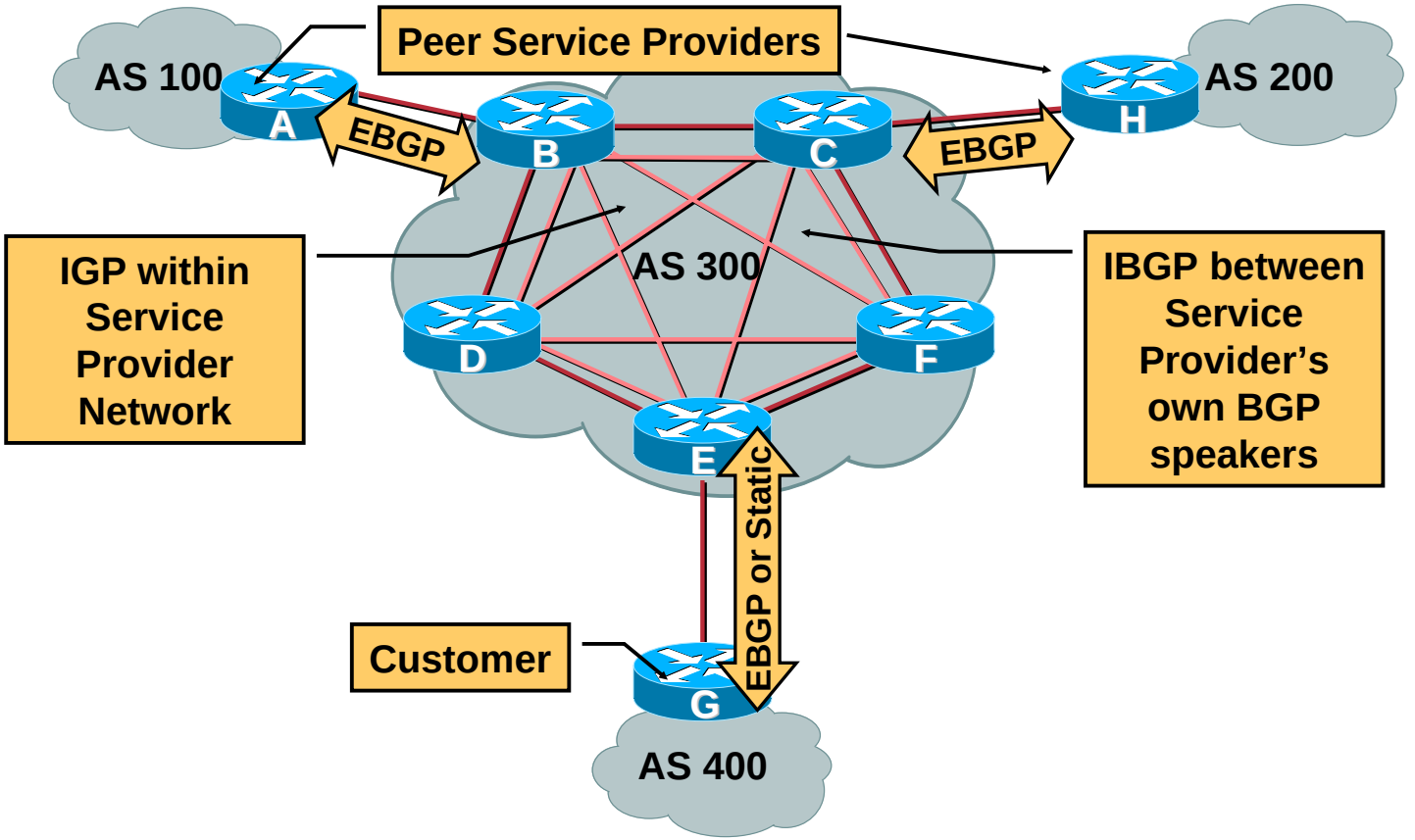


Common Routing Issues in Service Provider Networks

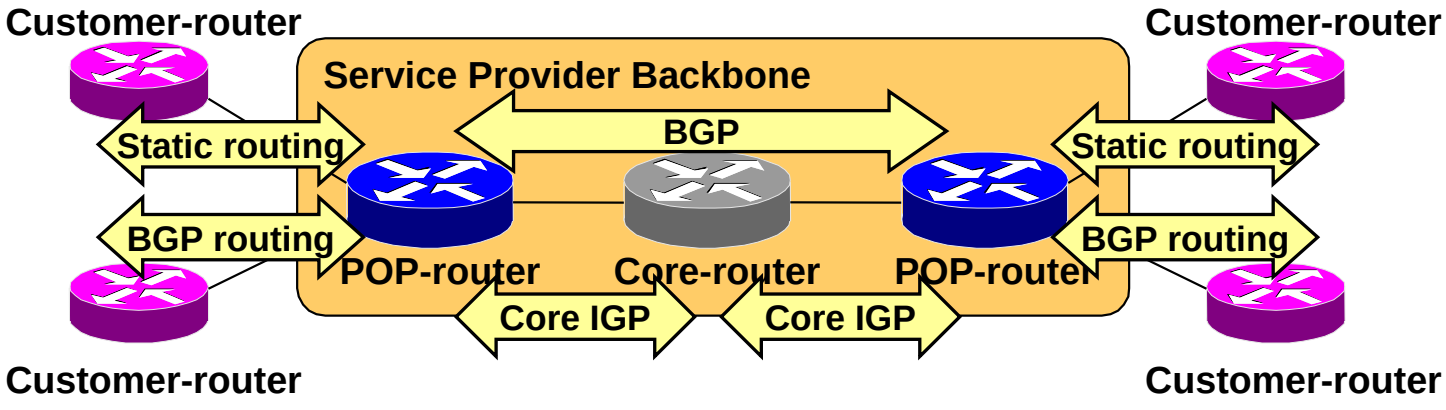
Common Service Provider Network Structure

- Networks are divided into Points-of-Presence (POPs)
- Different types of media are concentrated at the POP
- Optimal routing between POPs is desired
- Runs BGP or static routing with customer
- Exchanges routes with other Service Providers via EBGP
- Runs IBGP among own BGP speakers
- Runs one instance of IGP (OSPF or ISIS)
  - IGP used for internal routes only

Common Service Provider Network Structure



Common Service Provider Routing



- POP-routers use BGP or static routing with Customer-routers
- Core IGP is a single instance of IS-IS or OSPF
- Core IGP is used only within the Service Provider Backbone

Internet Service Providers

- **BGP Scaling**
  - **BGP carries customer routes**
  - **BGP carries other provider routes**
- **Scaling IGP**
  - **IGP only responsible for next hop**

Do NOT Redistribute BGP into IGP

Cisco.com

- IGP performance and convergence time suffer if large number of routes are carried
- No IGP is capable of carrying several 10s of thousands of routes
- Full Internet routing table has exceeded 120000+ routes

BGP Responsibilities

Cisco.com

- BGP update generation
- Scaling BGP policies
- Scaling IBGP mesh
- Reduce impact of flapping routes

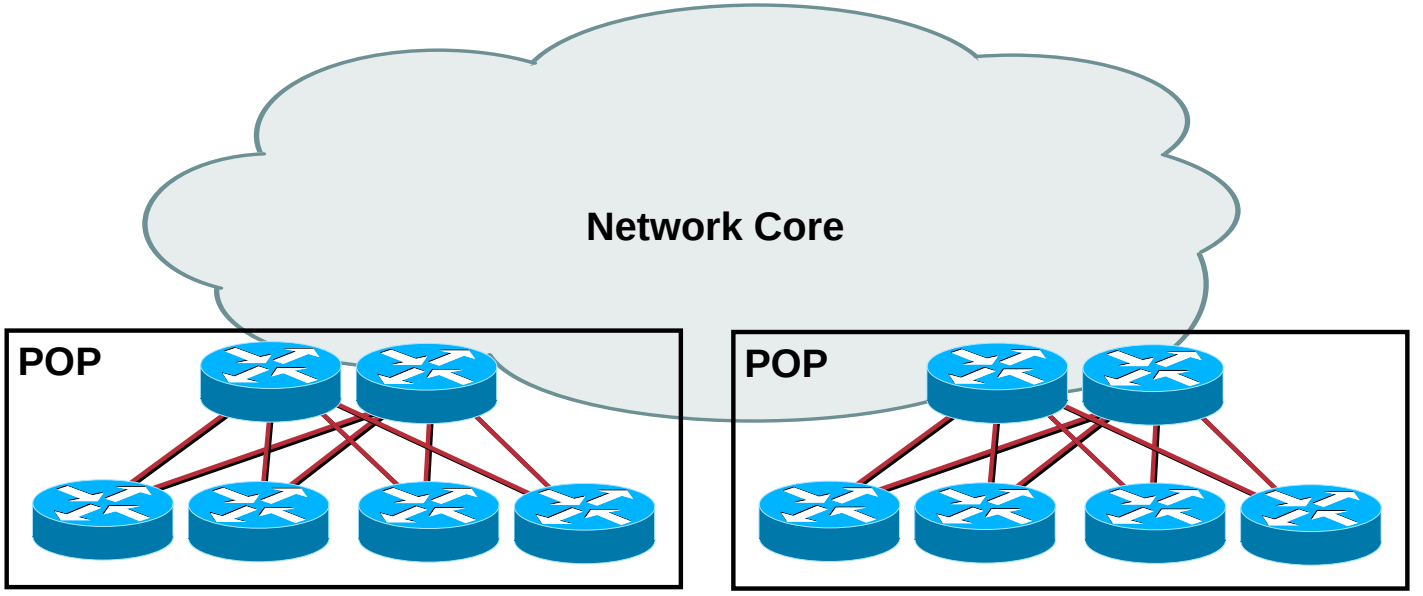
IGP Responsibilities

- Carry route to BGP next-hop
- Provide optimal path to the next-hop
- Converge to alternate path so that the BGP peering is maintained

Scaling IGP

- Loopbacks and internal links carried only
- Good addressing structure within the POP required
- Summarization of internal link addresses on POP-level
- Optimal routes to loopbacks needed only (with proper summarization)

Assign Addresses to Allow Summarization



Link address from range 210.1.1.0/24
Loopbacks from range 173.16.1.16/28

Link address from range 210.1.2.0/24
Loopbacks from range 173.16.1.32/28

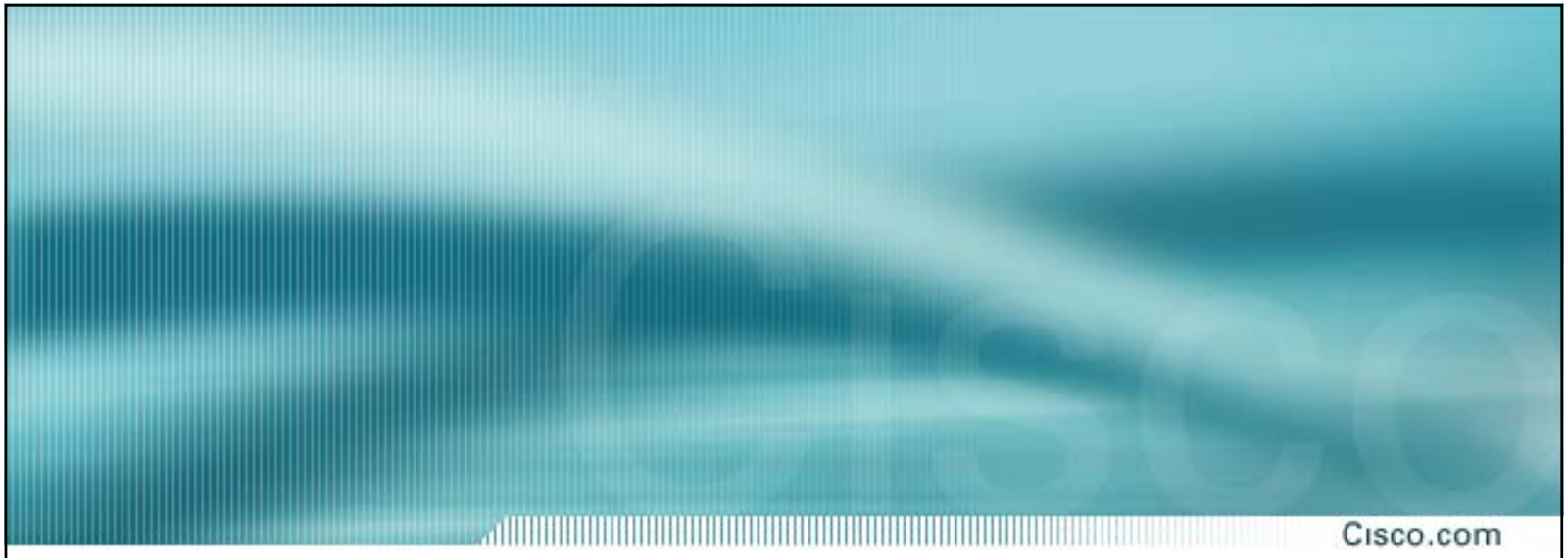
Digression—Loopback Interface

- Most ISPs make use of the router loopback interface
- IP address configured is a host address
- Configuration example:

```
interface loopback 0
 description Loopback Interface of CORE-GW3
 ip address 215.18.3.34 255.255.255.255
 no ip redirects
```

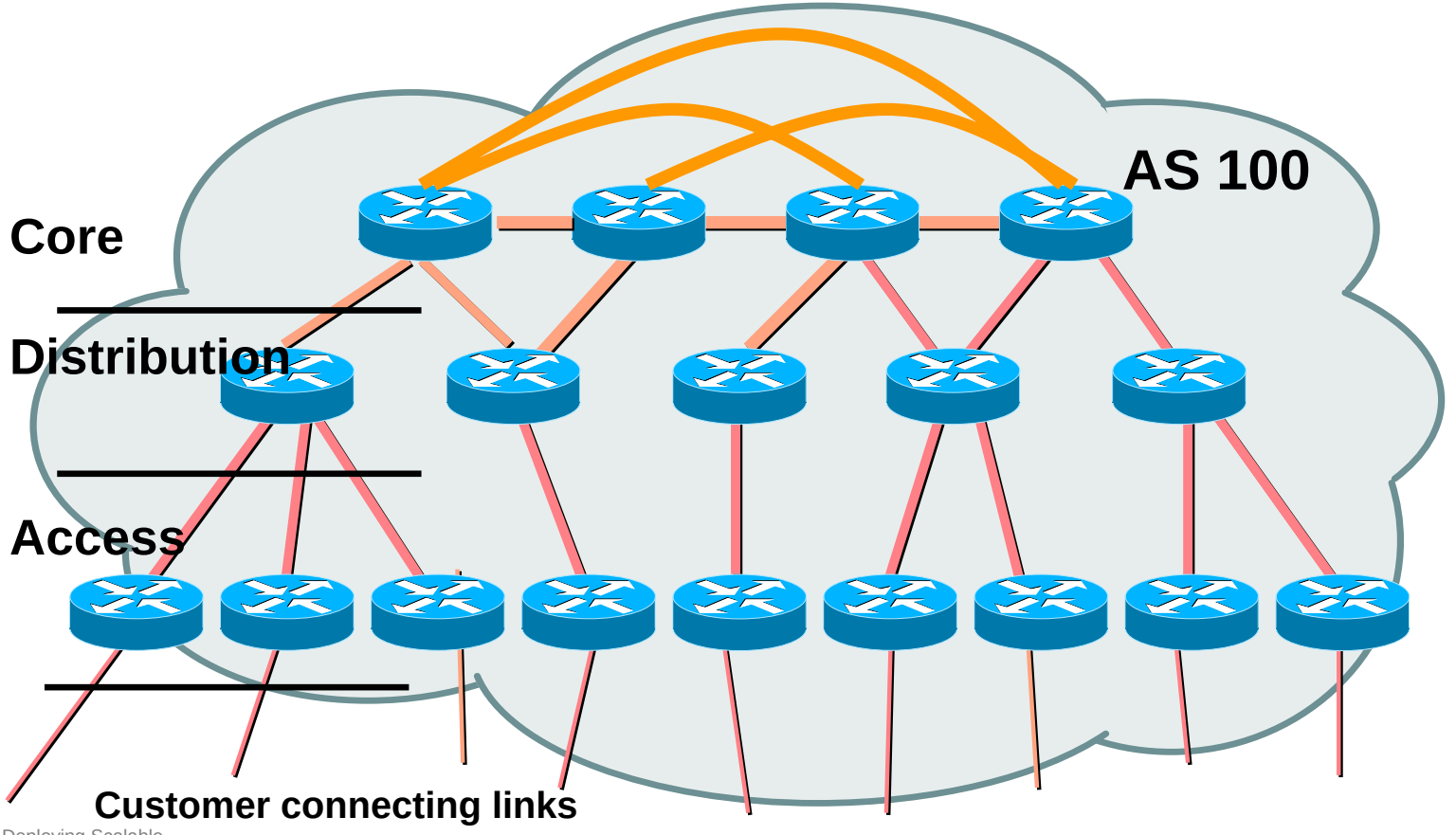
Digression—Loopback Interface

- Loopback addresses taken out of a different address space and not summarized
- Loopback interfaces on ISP backbone usually numbered:
  - Out of one contiguous block, or
  - Using a geographical scheme, or
  - Using a per PoP scheme
- Aim is to increase stability, aid administration, and improve security



Scaling IGP in Service Provider Networks

General Topology

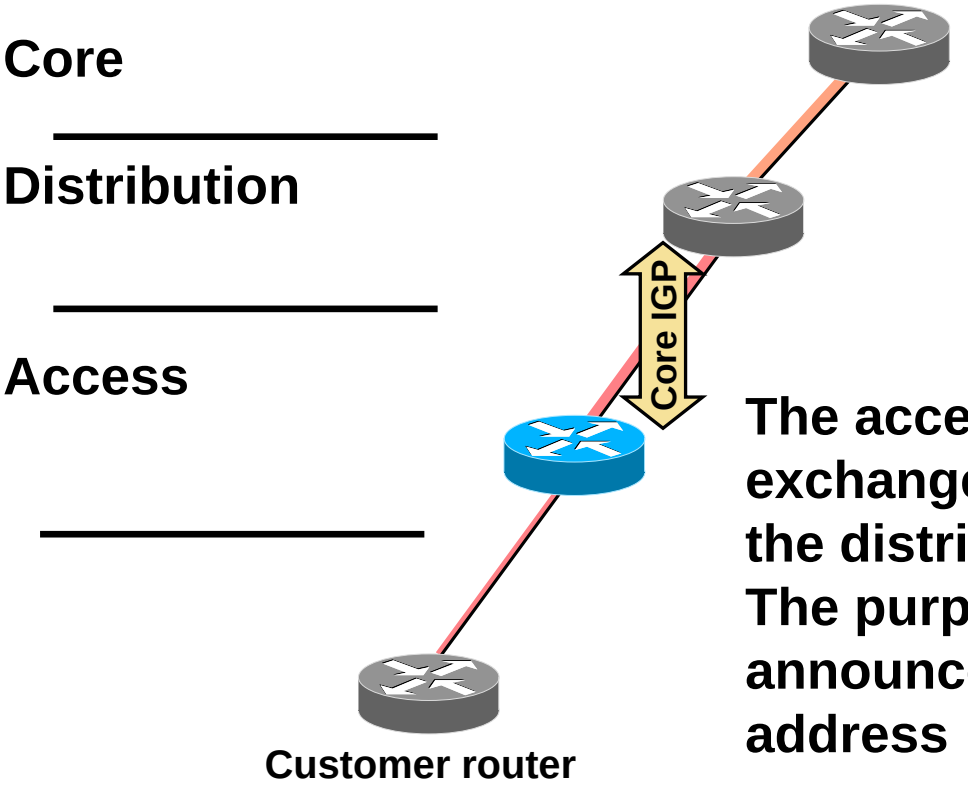


Access Routers in IGP

- **Access Routers**
  - **Access routers connect to customer routers**
  - **No IGP exchange with customer routers**
  - **No customer routes in IGP**
  - **Avoid advertising link to customer router**

Access Router IGP Exchange

Cisco.com



The access router exchanges IGP with the distribution router. The purpose is to announce its loopback address

Distribution Router in IGP

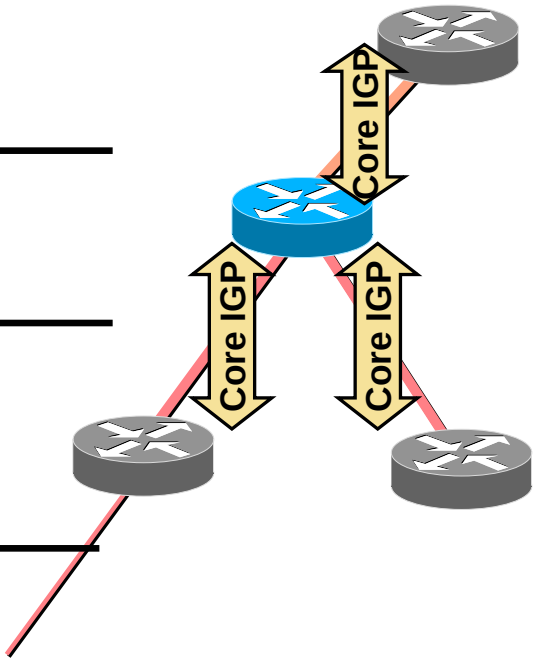
- **Distribution router**
  - **Acts as concentrator of access routers**
  - **Peers with both core and access routers**
  - **High speed customer may connect to distribution layer**
  - **Minimize routing update propagation to the core if an individual access line is flapping**

Distribution Router IGP Exchange

Core

Distribution

Access



The distribution router exchanges IGP with both the access and core routers. Summarize access-link addresses.

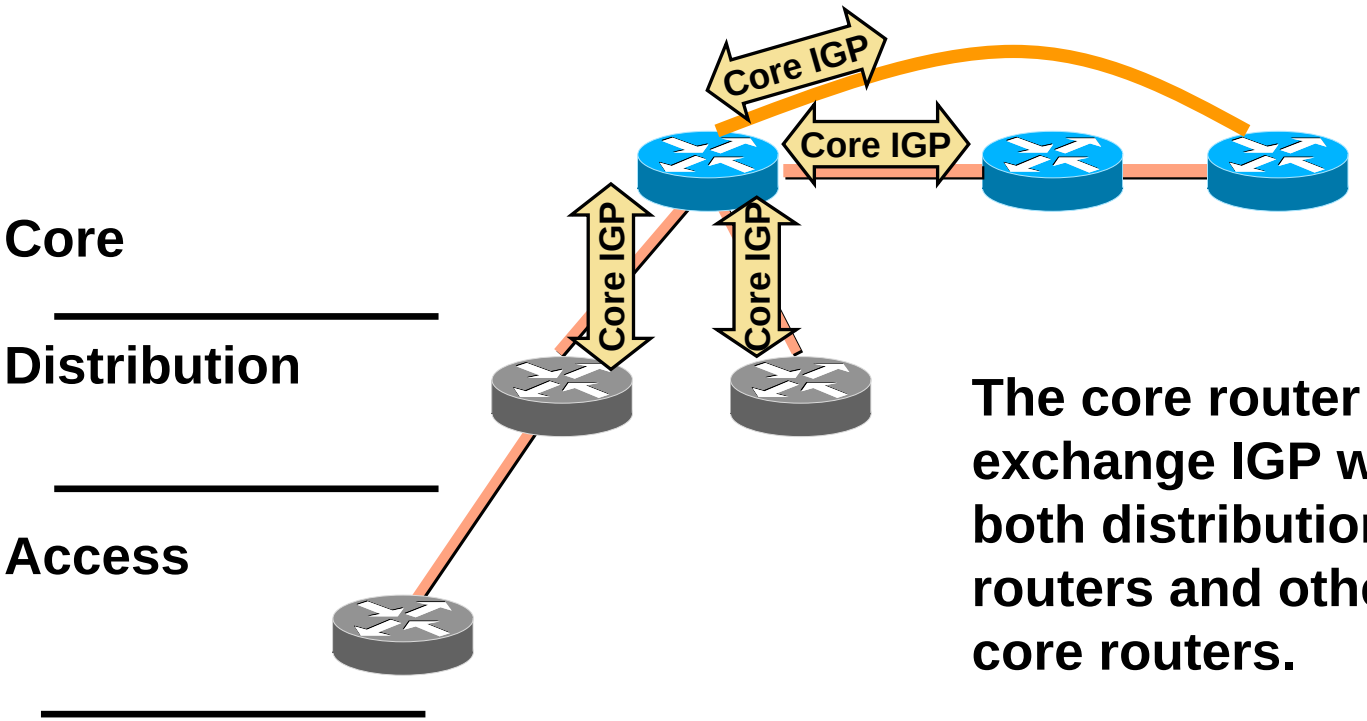
Summarization in Distribution Router

- **Distribution router**
  - **Links to access routers should be out of a contiguous block**
  - **Good practice to make it an ABR/L1L2 router**
  - **Summarize the intra POP links, do not summarize loopbacks**

Core Router in IGP

- **Core Router**
  - **Well connected, possibly fully-meshed, with other core routers**
  - **Provides high speed accesses between regions**
  - **Full routing information for all of the regions**
  - **Optimum routing to explicit loopback addresses**
  - **Peers with distribution routers and other core routers**
  - **Should be protected by distribution routers from unnecessary updates**

Core Router IGP Exchange



The core router exchange IGP with both distribution routers and other core routers.

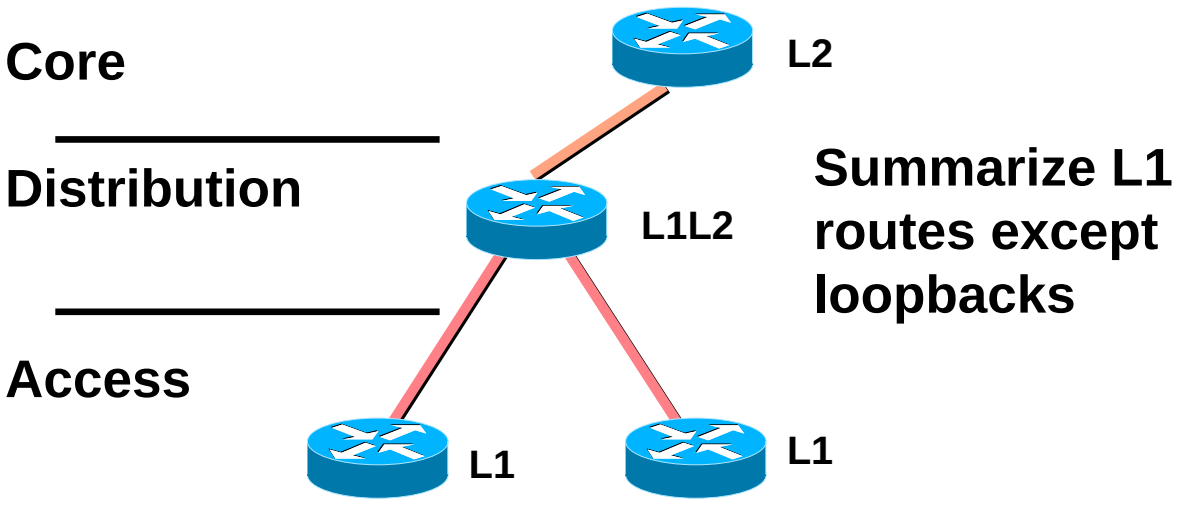
IGP Selection in SP Network

- Use either ISIS or OSPF
- Both are link state protocols
- Consider the appropriate design guideline (OSPF has more strict design rules)

ISIS—Intermediate System to Intermediate System

- Link state routing protocol
- OSI development now continued in IETF
- Supports VLSM
- Low bandwidth requirements
- Supports two levels
  - The backbone (Level 2) and areas (Level 1)
- Define intra POP router as L1
- Define the distribution router as L1/L2 router
- Summarize the intra POP links at the L1/L2 border

ISIS Routing Hierarchy



ISIS

- ISIS L1 router needs optimal path to all other POP router
- Use route leaking feature for ISIS, leak only the loopback routes into L1 areas

ISIS

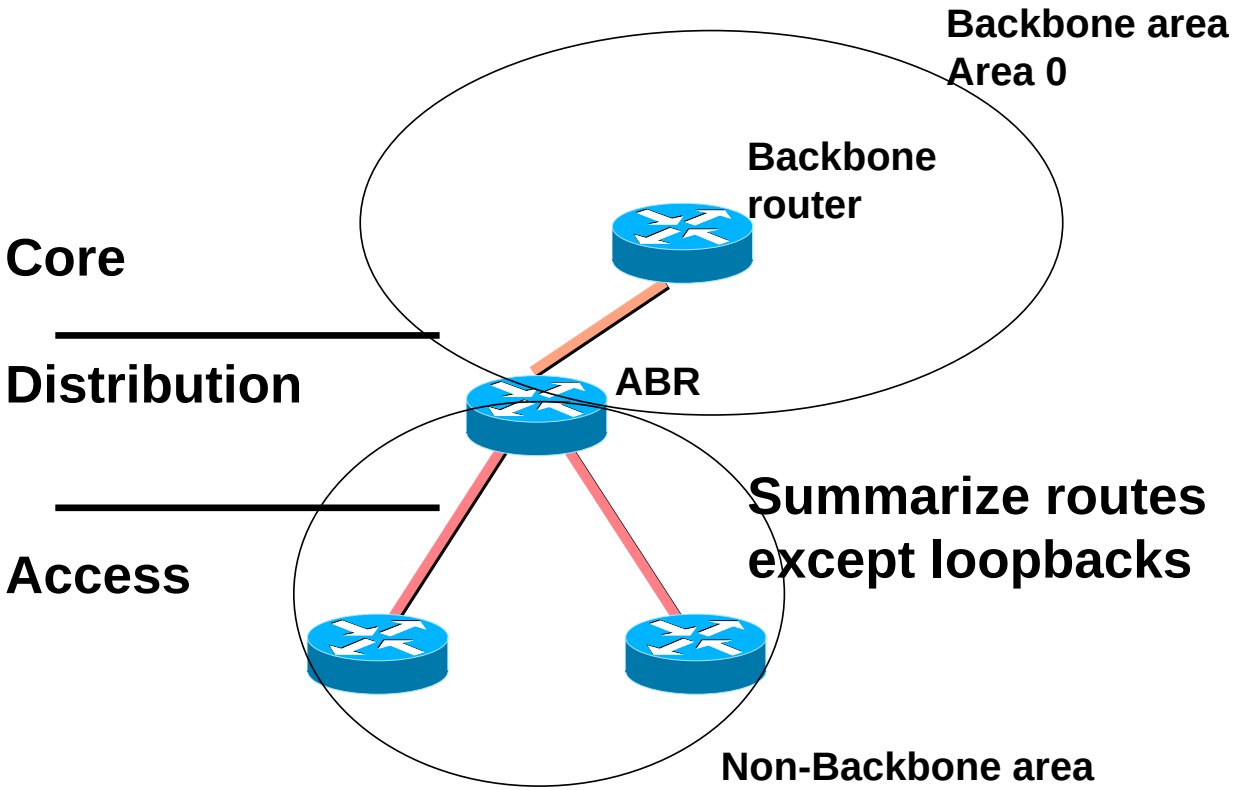
If you are building from scratch and network is not big, build the backbone that is L2 only and then expand

Do not build L1 only network

OSPF—Open Shortest Path First

- Link state routing protocol
- Designed by IETF for TCP/IP—RFC2328
- Supports VLSM
- Low bandwidth requirements
- Supports different types of areas
- Define intra POP router as internal router
- Define the distribution router as ABR
- Summarize the intra POP links at the ABR
- Route summarisation and authentication

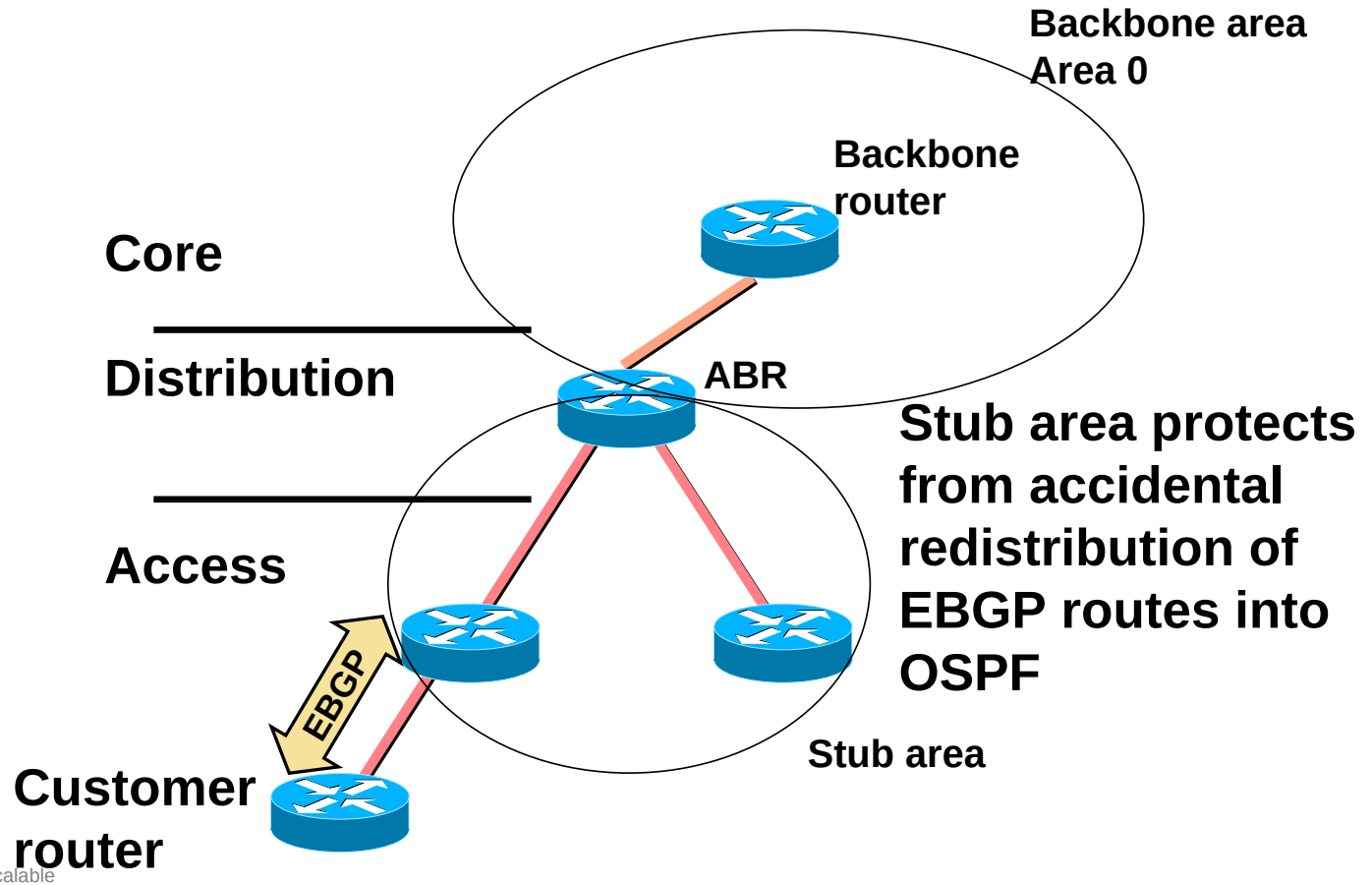
OSPF Routing Hierarchy



OSPF

- Send loopbacks as specific
- Summarize all intra POP links at ABR
- If you make your POP areas to be Stub and keep all your EBGP connections within the stub area, you can avoid the accidental redistribution of BGP into OSPF

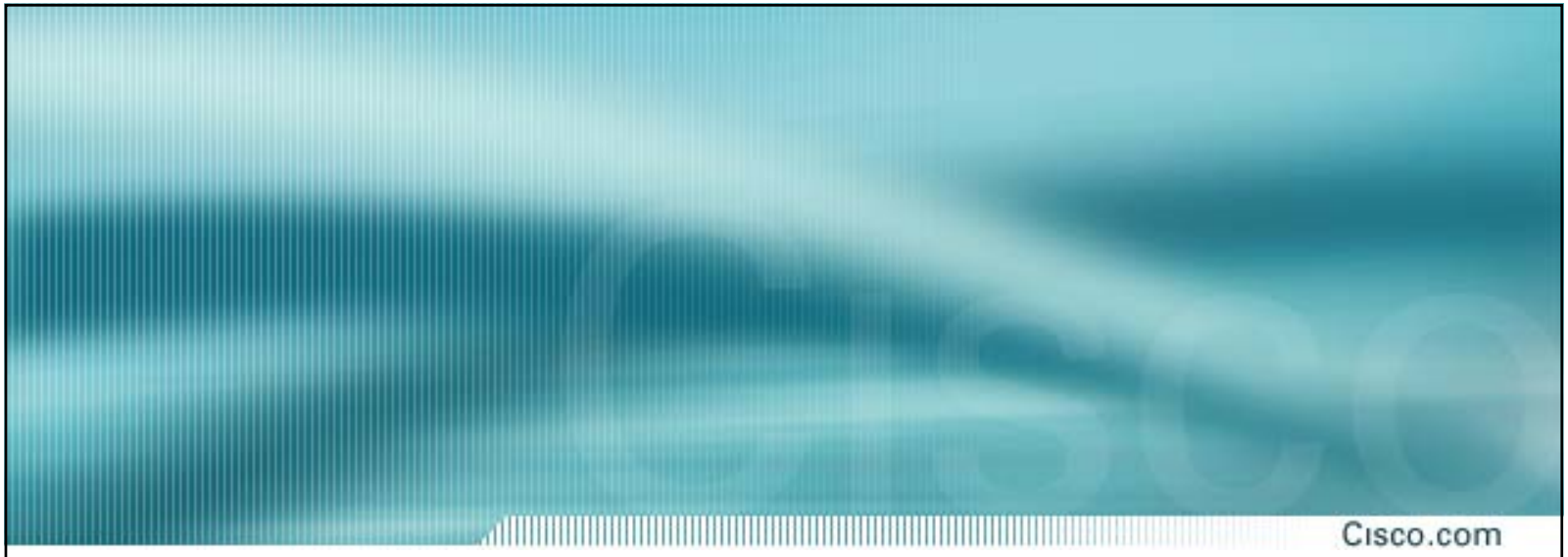
Use of Stub Area



Stub area protects from accidental redistribution of EBGP routes into OSPF

Scalable Network Design

- **ISIS**
  - **Implement Level 1—Level 2/Level 1 hierarchy for large networks only**
  - **Internet friendly enhanced features**
- **OSPF**
  - **Implement area hierarchy**
  - **Enforces good network design**
- **Requires addressing plan**
- **Implement route summarization**
- **Redistributed Connected interface**

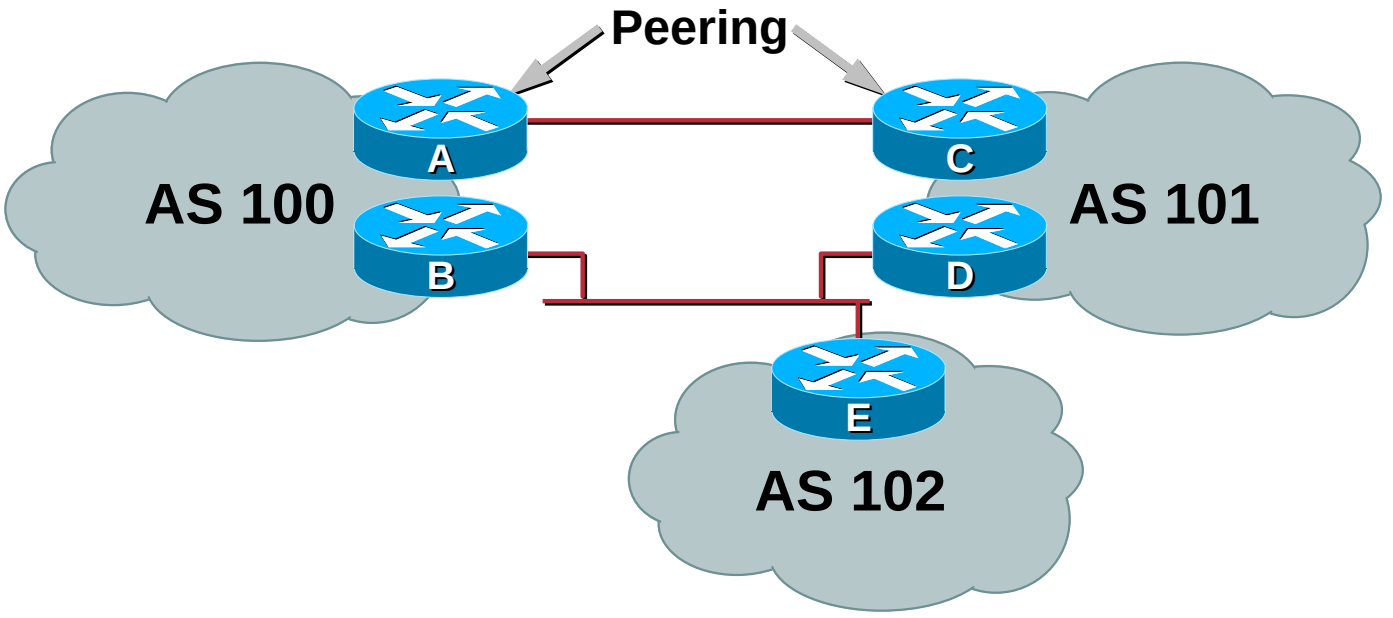


BGP Quick Review

BGP

- RFC 1771
- Border Gateway Protocol
- Version 4 is current
- Exterior routing protocol (vs. interior)
- Uses TCP for transport
- Many options for policy enforcement
- Classless Inter-Domain Routing (CIDR)
- Widely used for Internet backbone
- Autonomous systems

BGP Basics

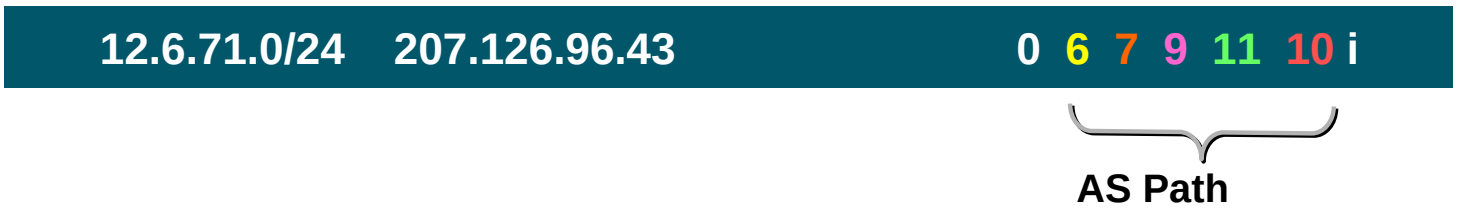


- Runs over TCP
- Path vector protocol
- Incremental update

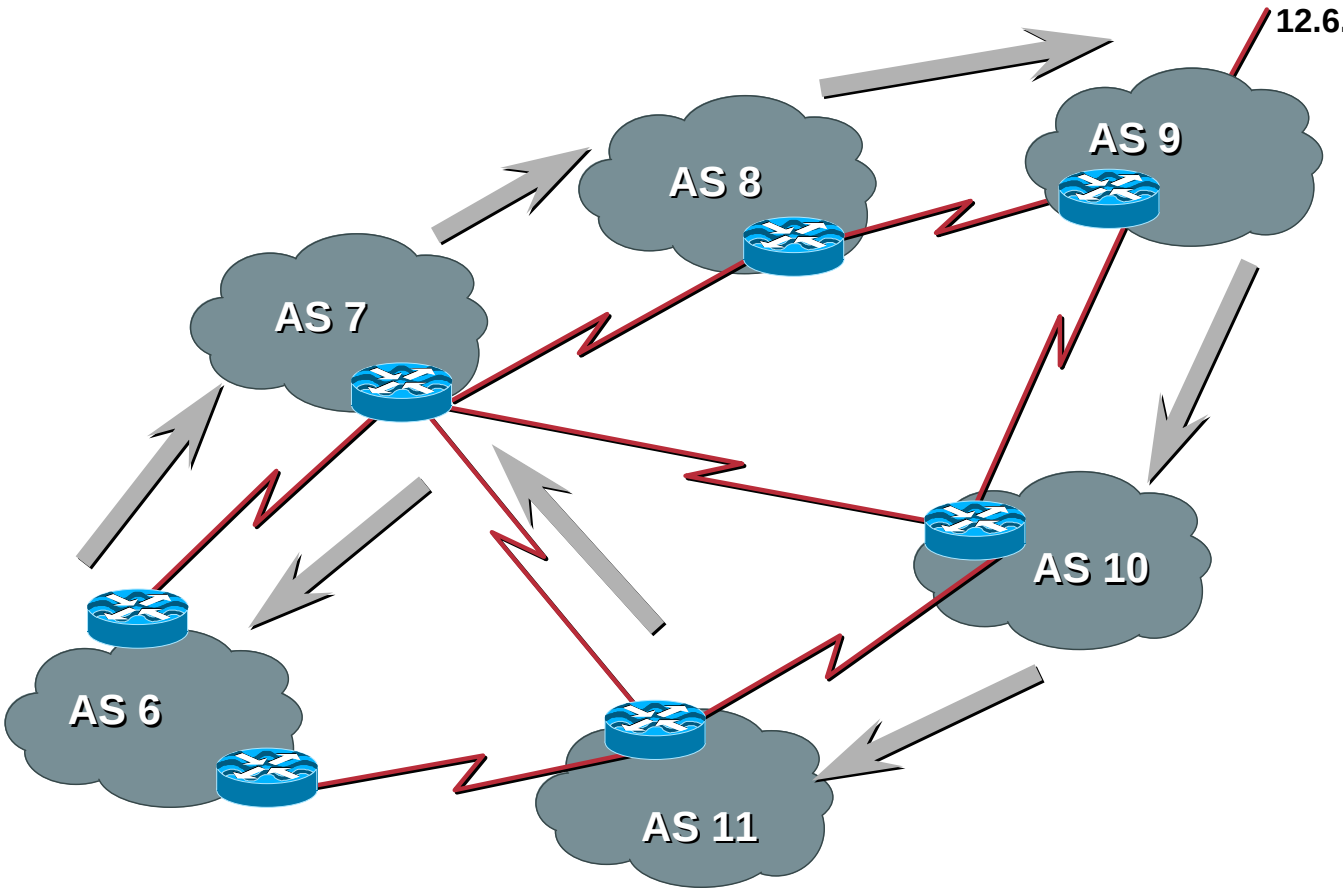
Path Vector Protocol

- **BGP is classified as a path vector routing protocol (see RFC 1322)**

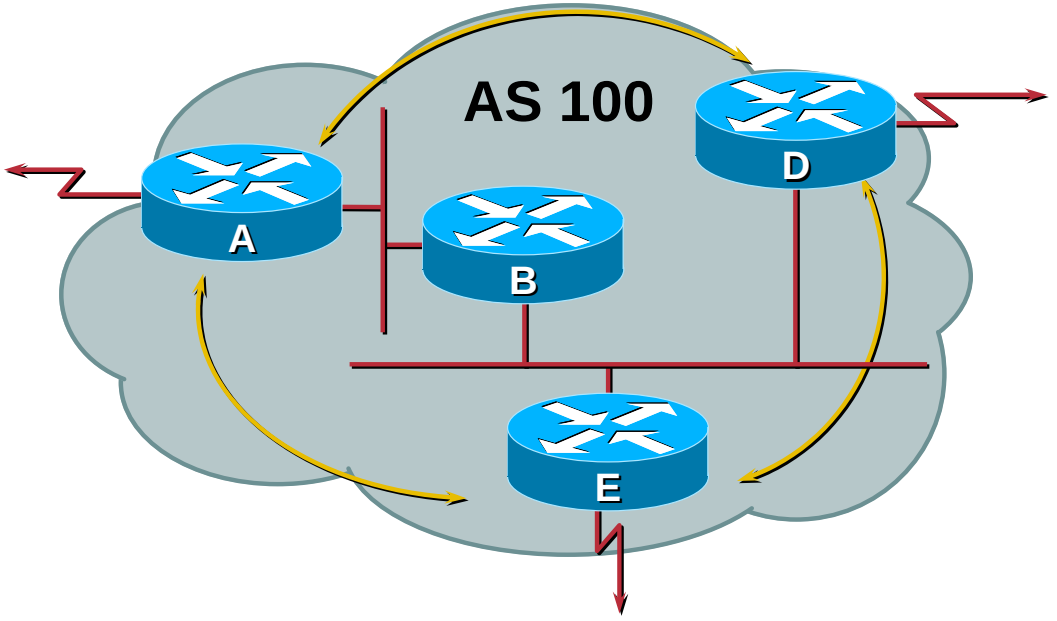
  A path vector protocol defines a route as a pairing between a destination and the attributes of the path to that destination



Path Vector Protocol

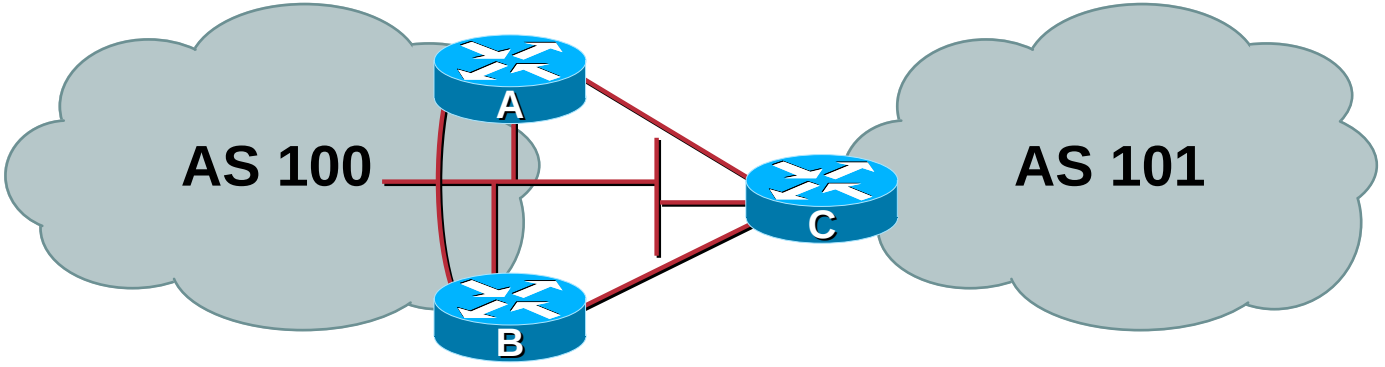


Internal BGP (iBGP) Peering



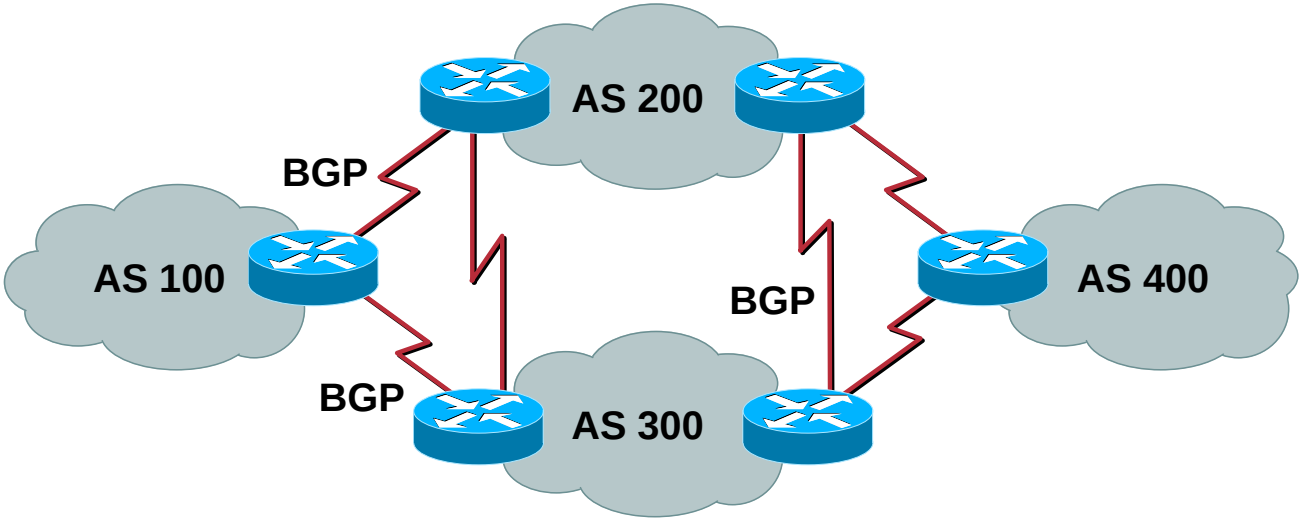
- BGP peer within the same AS
- Not required to be directly connected
- iBGP neighbors should be fully meshed
- Few BGP speakers in corporate network

External BGP (eBGP) Peering



- Between BGP speakers in different AS
- Should be directly connected
- Don’t run an IGP between eBGP peers

Policy Drives BGP Requirements



- **Policy for AS 100: Always use AS 300 path to reach AS 400**

BGP Versus OSPF/ISIS

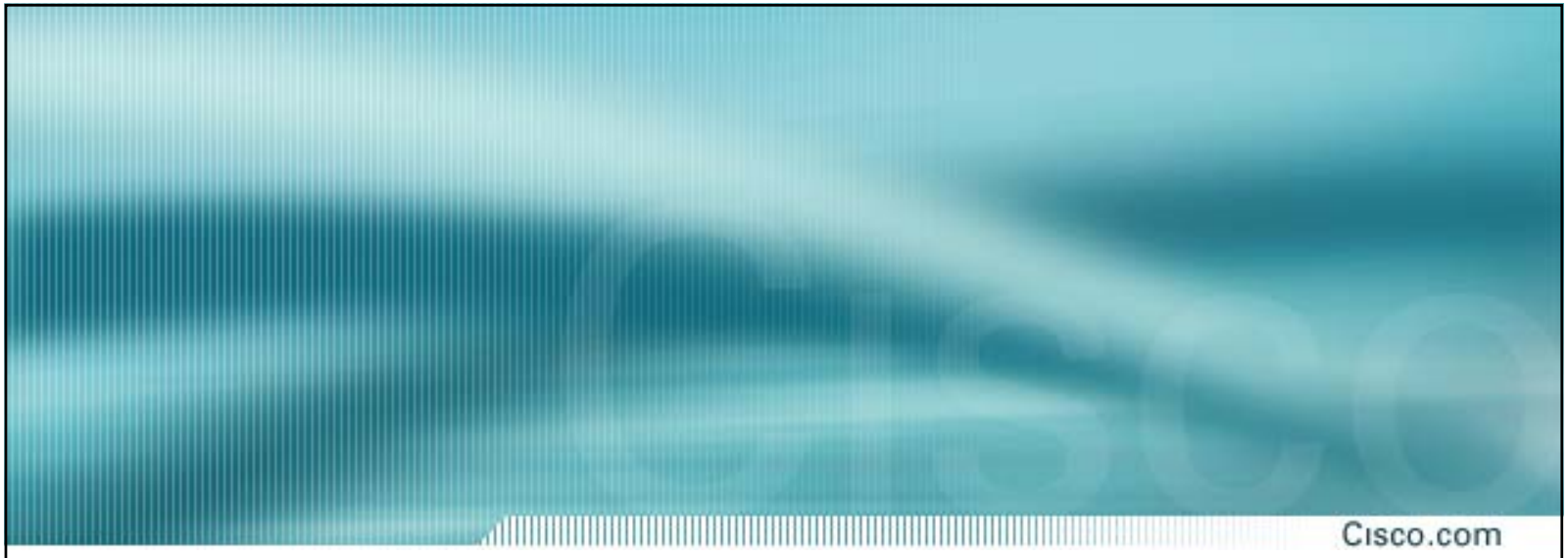
- **Internal routing protocols (IGPs)**
  - **Examples are ISIS and OSPF**
  - **Used for carrying infrastructure addresses**
  - **NOT used for carrying Internet prefixes or customer prefixes**

BGP vs. OSPF/ISIS

- BGP used internally (iBGP) and externally (eBGP)
- iBGP used to carry
  - some/all Internet prefixes across backbone
  - customer prefixes
- eBGP used to
  - exchange prefixes with other ASes
  - implement routing policy

BGP vs. OSPF/ISIS

- **DO NOT:**
  - **distribute BGP prefixes into an IGP**
  - **distribute IGP routes into BGP**
  - **use an IGP to carry customer prefixes**
- **YOUR NETWORK WILL NOT SCALE**



Scaling BGP in Service Provider Networks

BGP Scaling

- **Policy Scaling**
  - **The AS routing policy should be unitary and easy to maintain**
  - **This is achieved by re-using the same configuration in all EBGP speaking routers**
- **IBGP mesh scaling**
  - **Avoid unnecessary duplicate updates over a physical link**
- **Updates and table size**
  - **Route summarization is the key to scalability**

Scaling Policies

- Prefix-based policies are expensive
- Community and AS path policies scale better
- Assign community to customer prefix to scale policies

Policy Scaling – Using Communities

Cisco.com

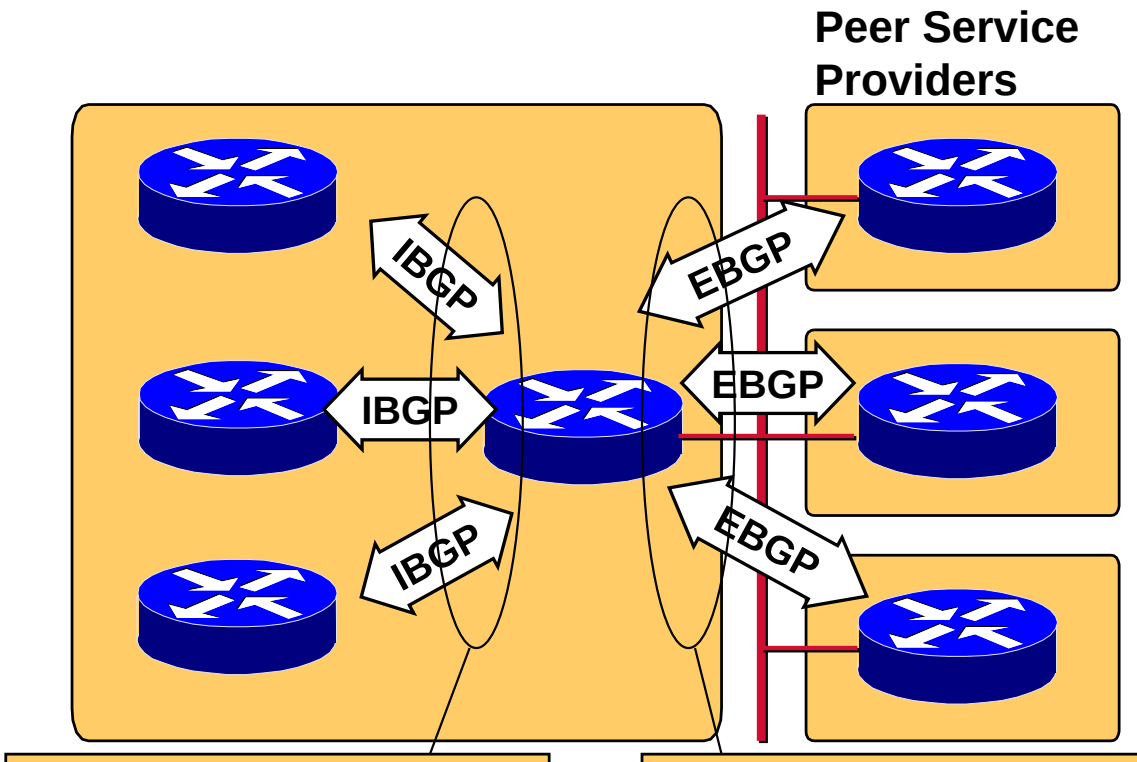
- **Communities - scale policy configuration**
  - **Used to signal a policy**
  - **Assigned to a route in a router close to the subnet**
  - **Acted upon by the EBGP speaker implementing the AS routing policy**
  - **Used to group destinations**
  - **Community attribute carried across ASs**

Policy Scaling – Using Peer Groups

Cisco.com

- **Peer groups - saves CPU cycles**
  - **Run through the route-maps implementing the policy once, update many neighbors**
  - **Also saves configuration lines**

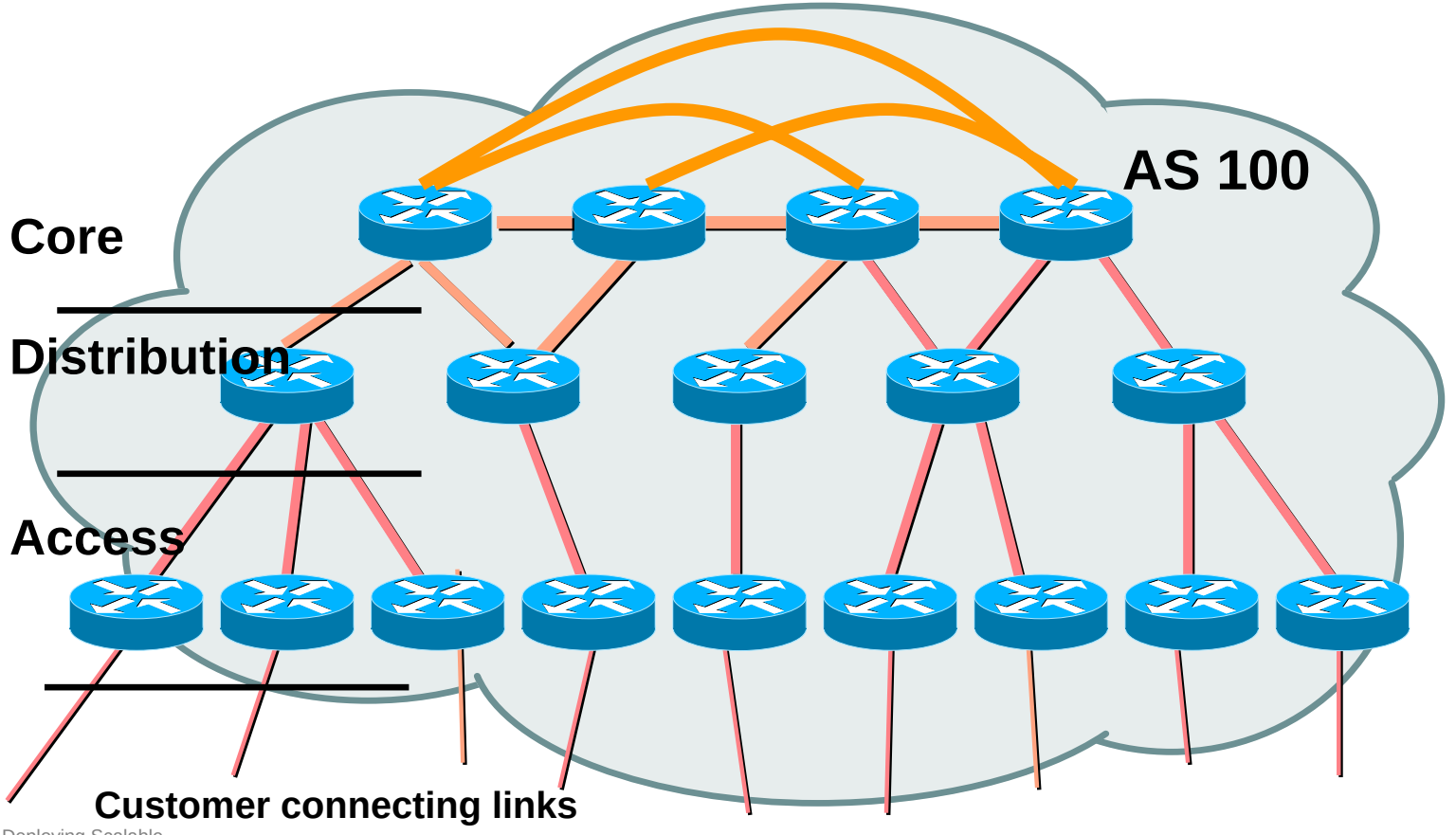
Example: Peer-groups



IBGP Mesh

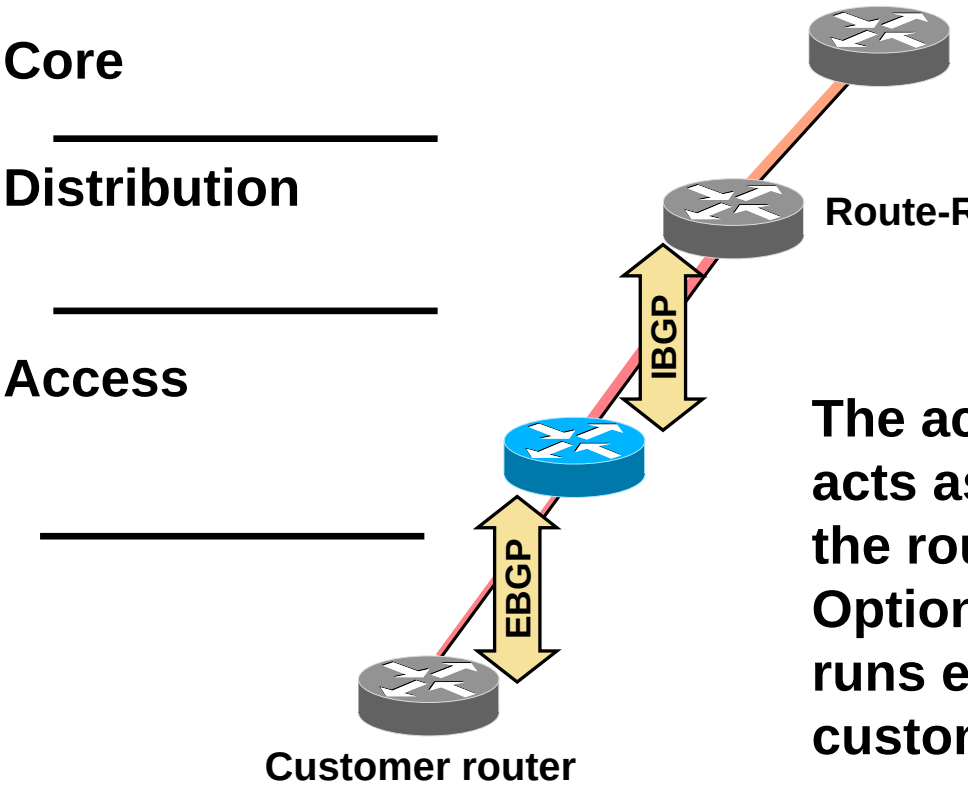
- Avoids routing information loop
- Scaling IBGP mesh
- Solution should not change the current behavior
- Two solutions:
  - Route reflectors (better and widely deployed)
  - Confederation

General Topology



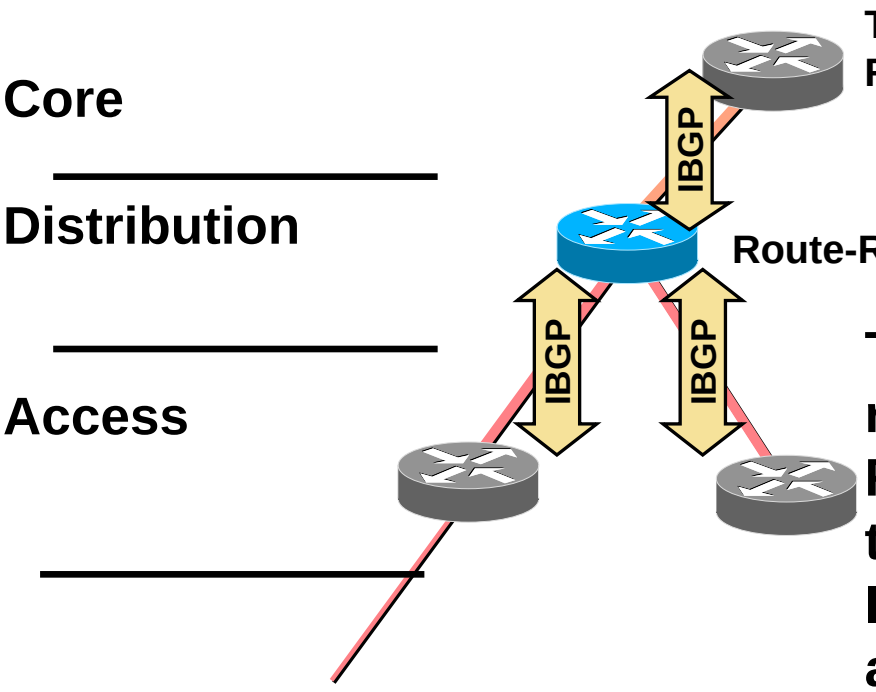
Access router and BGP Route-reflectors

Cisco.com



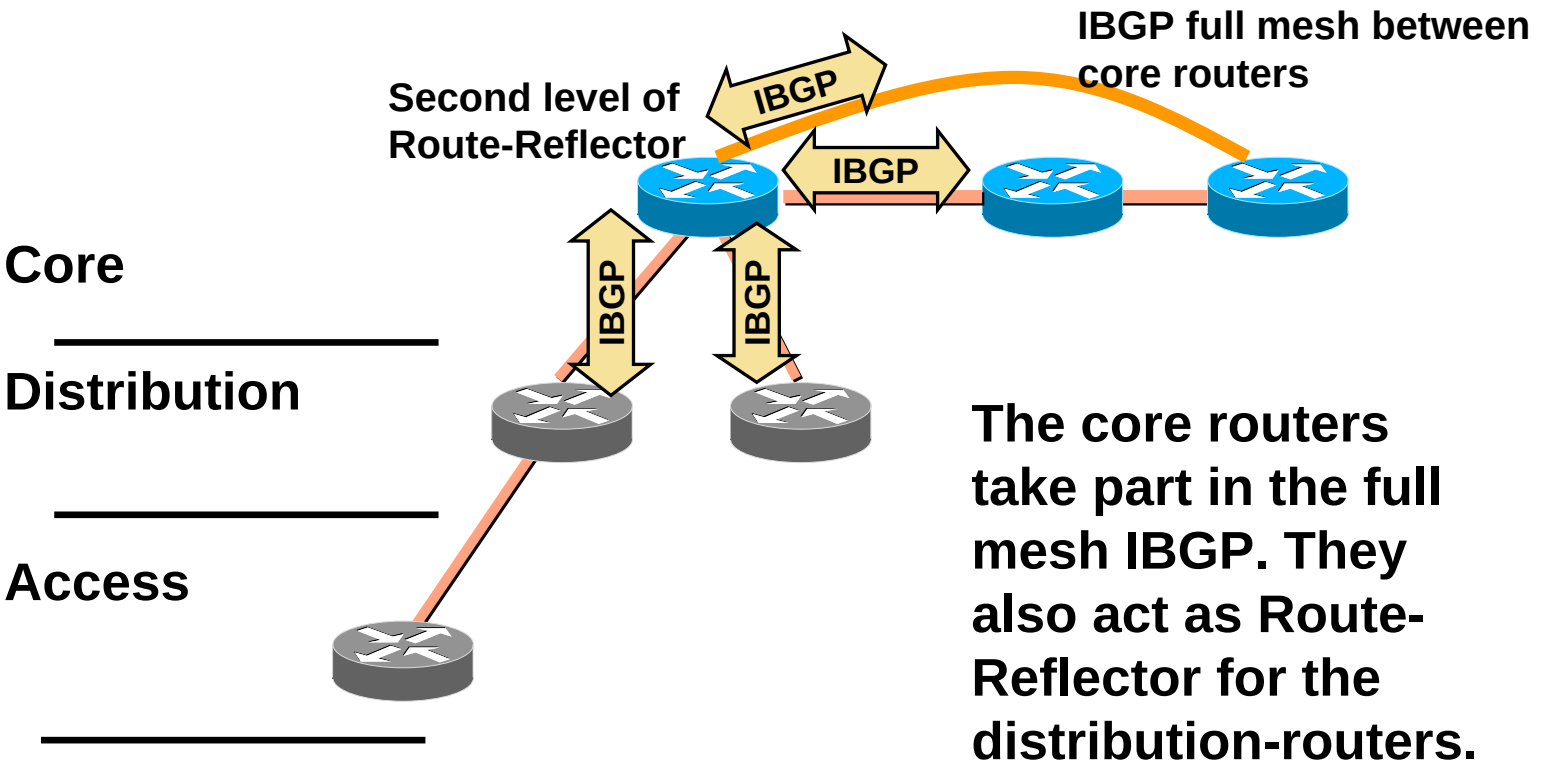
The access router acts as client to the route-reflector. Optionally, it also runs eBGP with customers.

Distribution Router and BGP Route-reflectors



The distribution router acts as Route-Reflector for the access-router. It simultaneously acts as client to the next level of Route-Reflectors

Core Router and BGP Route-Reflectors



The core routers take part in the full mesh IBGP. They also act as Route-Reflector for the distribution-routers.

Route Reflector: Benefits

Cisco.com

- Solves IBGP mesh problem
- Packet forwarding is not affected
- Normal BGP speakers co-exist
- Multiple reflectors for redundancy
- Easy migration
- Multiple levels of route reflectors

Route Reflector—Caveats

Cisco.com

- Single IGP
- Cannot change attribute of reflected routes
- In some IOS releases a RR with next-hop-self configuration actually changes the next-hop attribute on reflected routes – be careful especially in MPLS world

Confederations

- Divide the AS into sub-AS
  - eBGP between sub-AS, but some iBGP information is kept
    - Preserve NEXT_HOP across the sub-AS (IGP carries this information)
    - Preserve LOCAL_PREF and MED
- Usually a single IGP
- Visible to outside world as single AS—"Confederation Indentifier"
- Confederation is not popular in ISP World, Route-reflector deployed widely

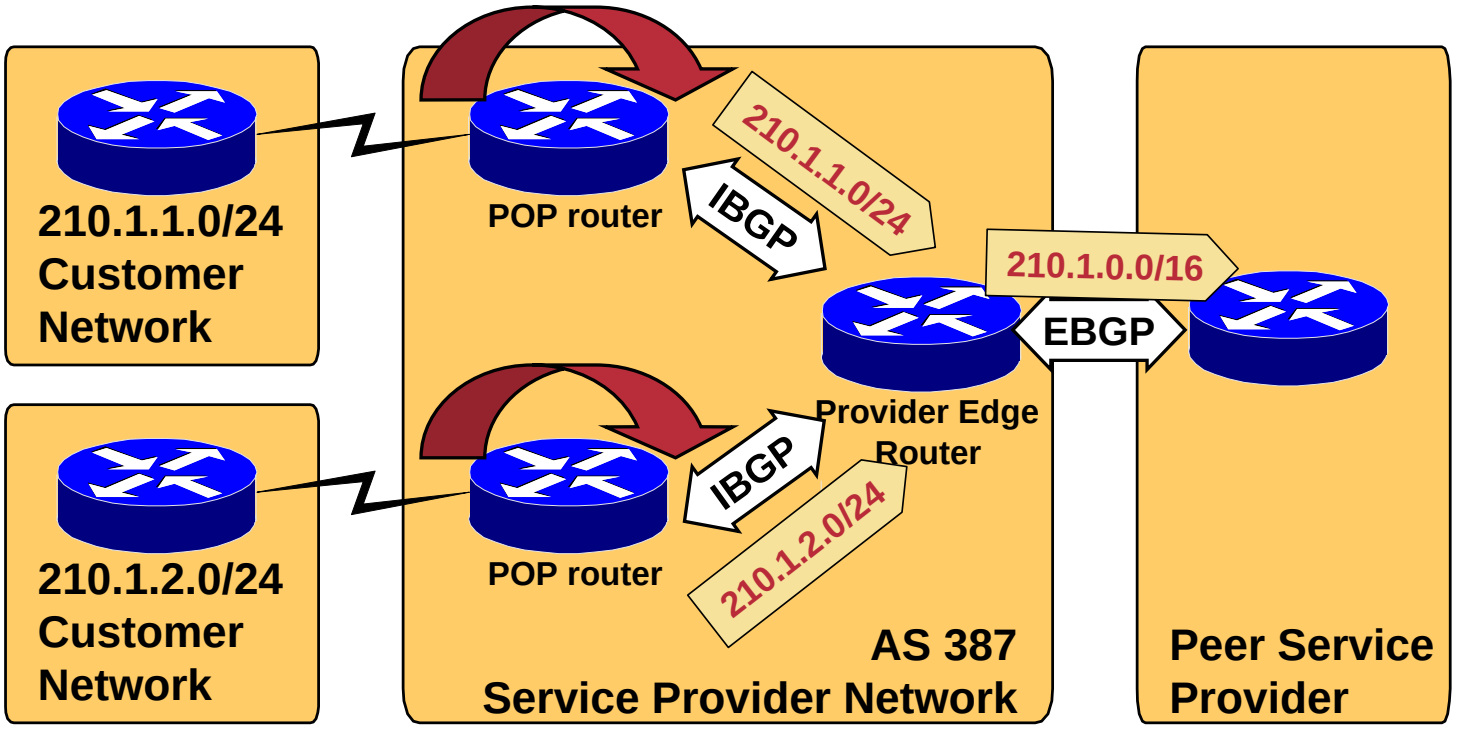
BGP - Route Summarization

- Route summarization is necessary in large networks
- The Internet would not survive without aggregation
- Only the IP prefix assigned to the ISP`s AS is announced via EBGP sessions

Single-Homed Customers

- Most customers are connected to a single service provider
- Single-homed customers should be assigned IP addresses by the ISP
- Routes to individual customers are kept within the ISP AS
- All customer routes are announced within the prefix assigned to the ISP AS

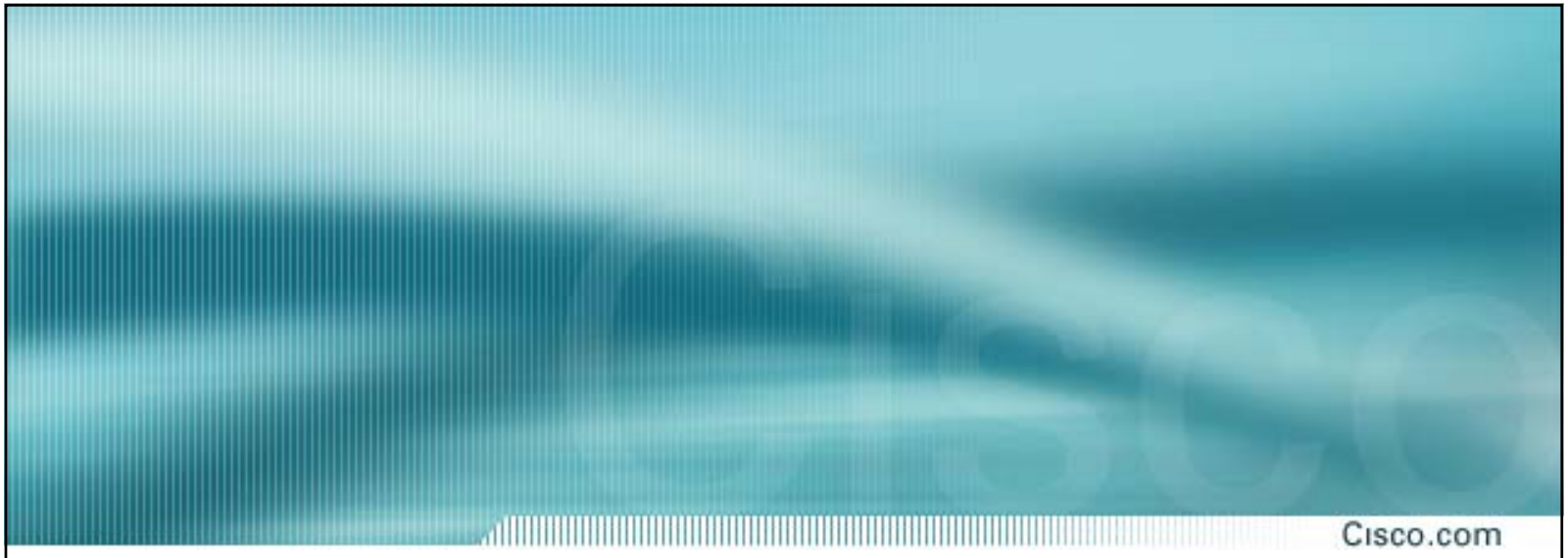
Route Summarization Example



Route Summarization: Benefits

Cisco.com

- **Reduced routing table**
  - **Conserves memory**
- **Less flaps**
  - **Conserves CPU and bandwidth**

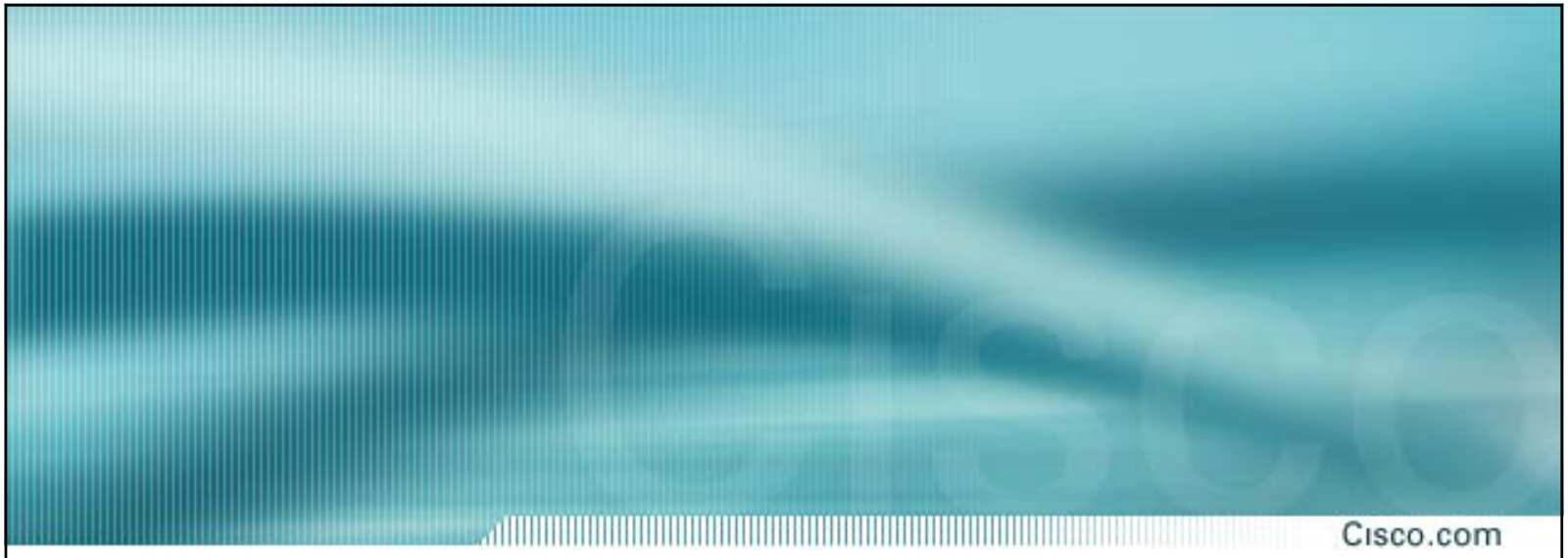


Deploying Scalable MPLS VPN

MPLS VPNs

Cisco.com

- Most popular MPLS application
- Deployed by majority of Cisco MPLS customers
- Offer QoS-based services
- Most common—Single private network
  - Many have also deployed it in a multi-AS environment
  - Also overlaid are Internet and VPN on the same network
  - 200–400 PEs
  - 200–500 VPNs average with as many as 1000+ VPNs
  - 4K sites per VPN
- Few deploying advanced features such as CsC



Quick MPLS VPN Overview

MPLS VPN Terminology

- **Provider network (P-network)**

  The backbone under control of a service provider

- **Customer network (C-network)**

  Network under customer control

- **CE router**

  Customer Edge router; part of the C-network and interfaces to a PE router

MPLS VPN Terminology

- **Site**

  Set of (sub)networks part of the C-network and co-located

  A site is connected to the VPN backbone through one or more PE/CE links

- **PE router**

  Provider edge router; part of the P-network and interfaces to CE routers

- **P router**

  Provider (core) router, without knowledge of VPN

MPLS VPN Terminology

- **Route-Target**

  64 bits identifying routers that should receive the route

- **Route Distinguisher**

  Attributes of each route used to uniquely identify prefixes among VPNs (64 bits)

  VRF-based (not VPN-based)

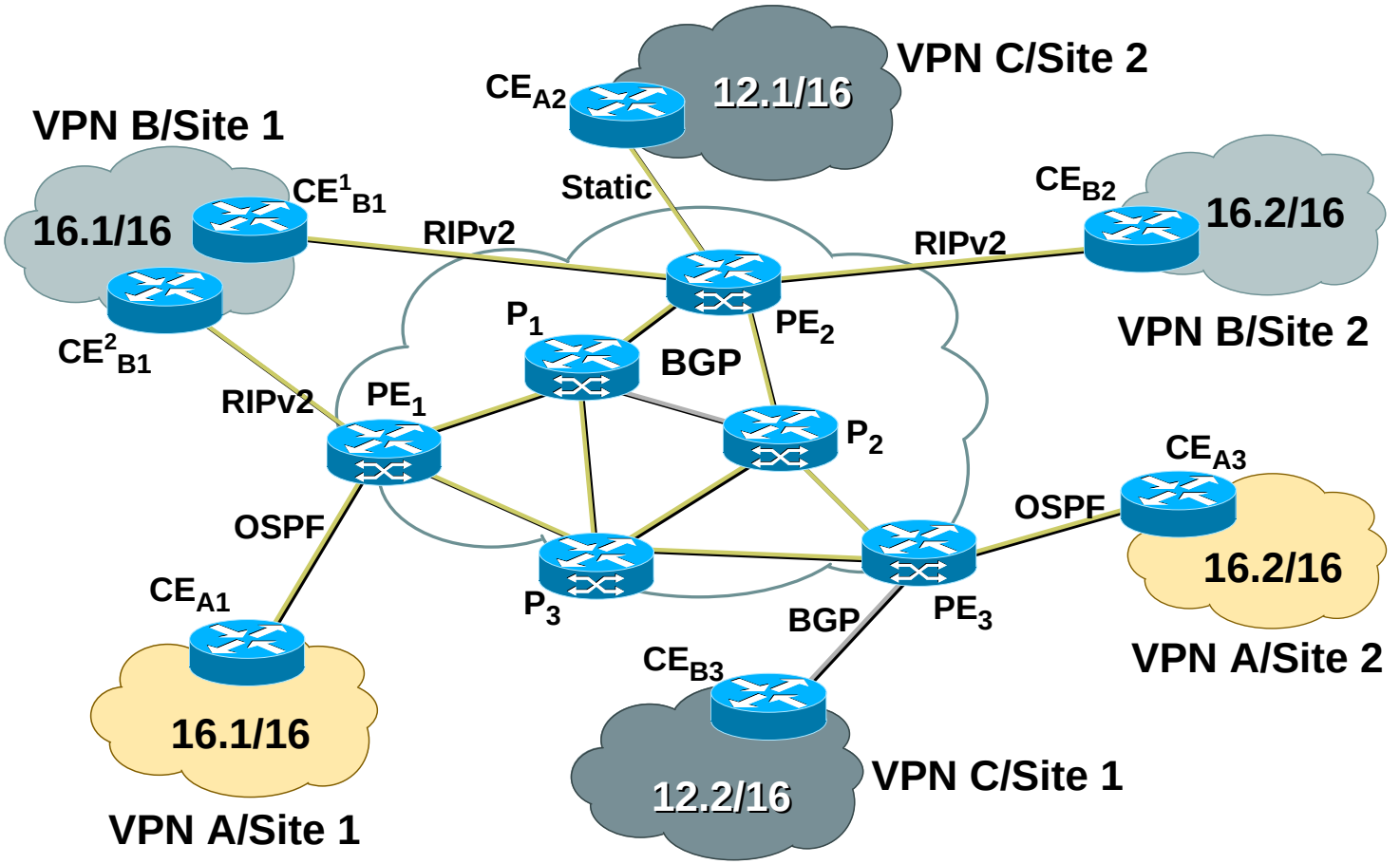
- **VPN-IPv4 addresses**

  Address including the 64 bits Route Distinguisher and the 32 bits IP address

MPLS VPN Terminology

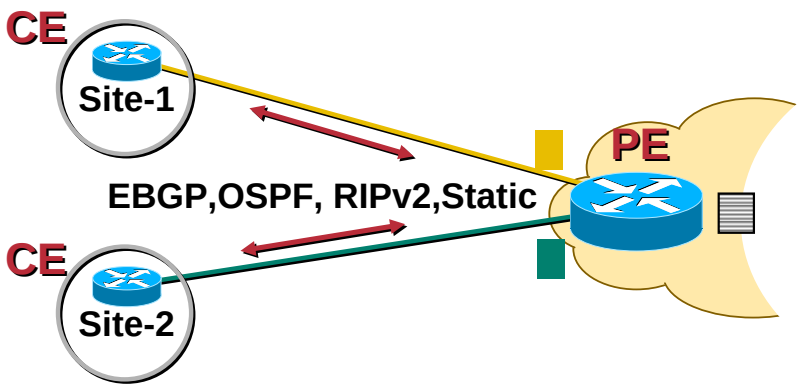
- **MP-BGP**
  - Multi-protocol extensions to BGP
- **VRF**
  - VPN routing and forwarding instance
  - Routing table and FIB table
  - Populated by routing protocol contexts
- **VPN-aware network**
  - A provider backbone where MPLS-VPN is deployed
- **VPN-aware application**
  - Apps aware of VRF context: vrf-ping, vrf-trace…

MPLS VPN Terminology



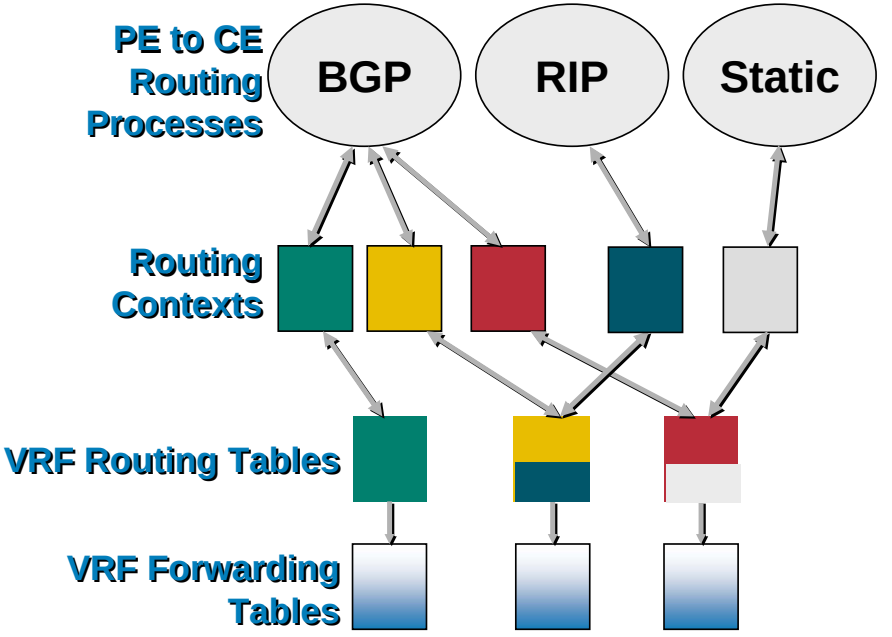
VRF Route Population

- **VRF is populated locally through PE and CE routing protocol exchange**
  - **RIP version 2, OSPF, BGP-4 and static routing**
- **Separate routing context for each VRF**
  - **Routing protocol context (BGP-4 and RIP v2)**
  - **Separate process (OSPF)**

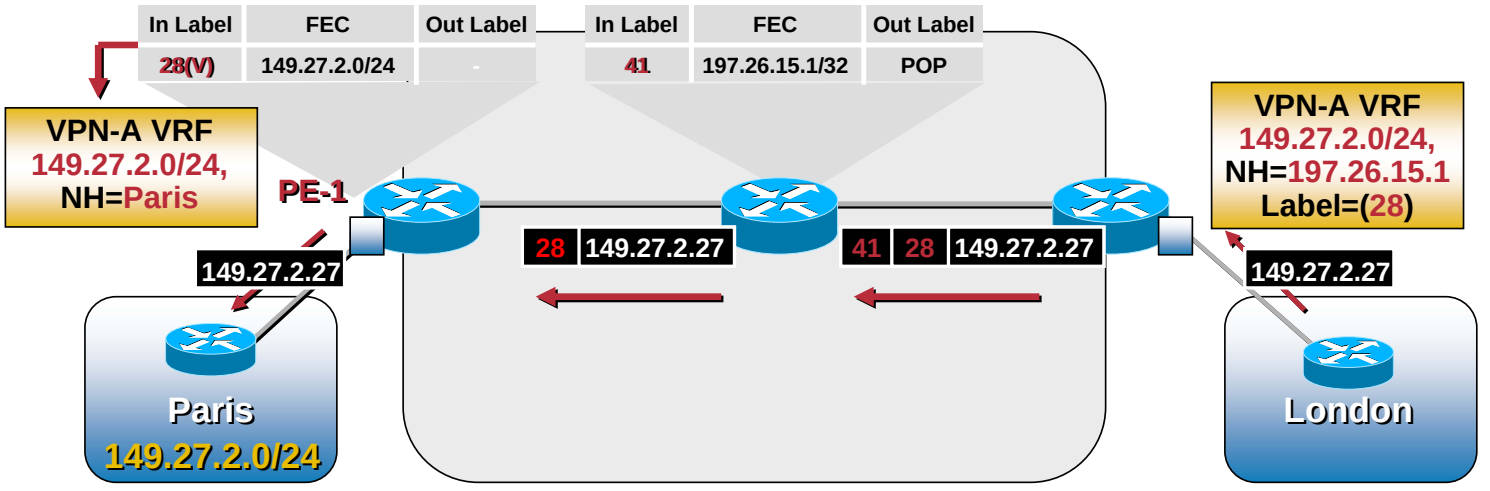


VRF and Multiple Routing Instances

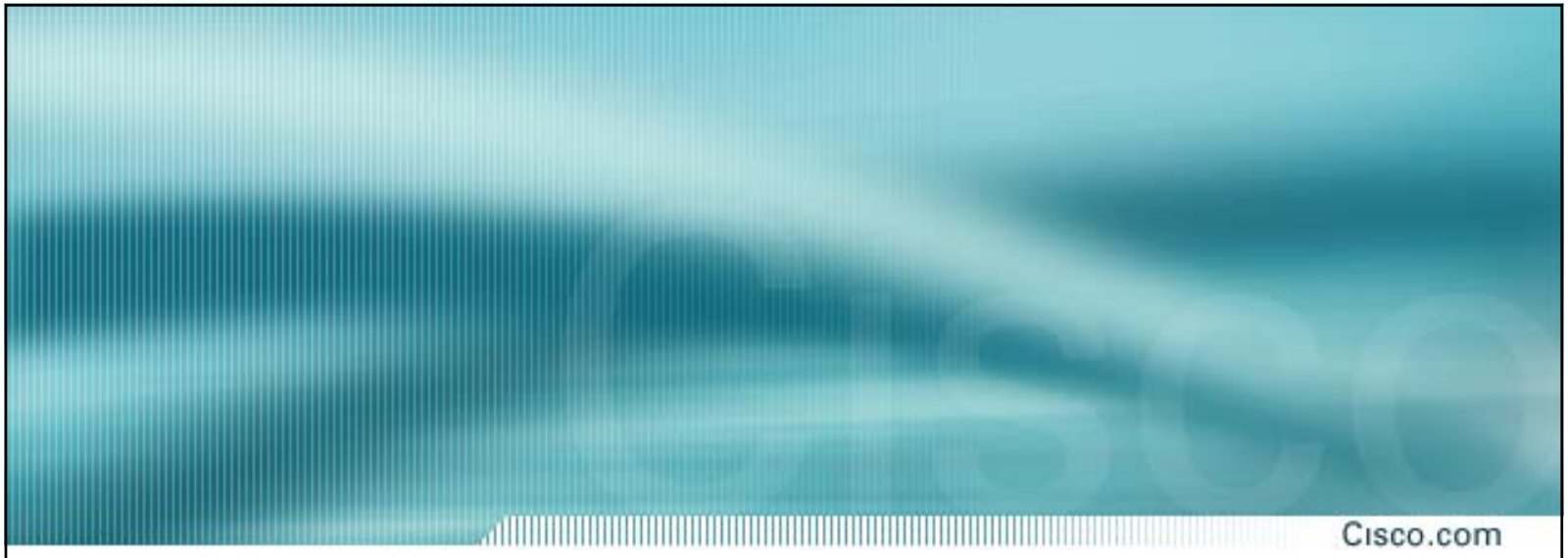
- Routing processes run within specific routing contexts
- Populate specific VPN routing table and FIBs (VRF)
- Interfaces are assigned to VRFs



MPLS/VPN Label Stacking and Packet Forwarding



Packet Forwarding Based on Stack of Labels

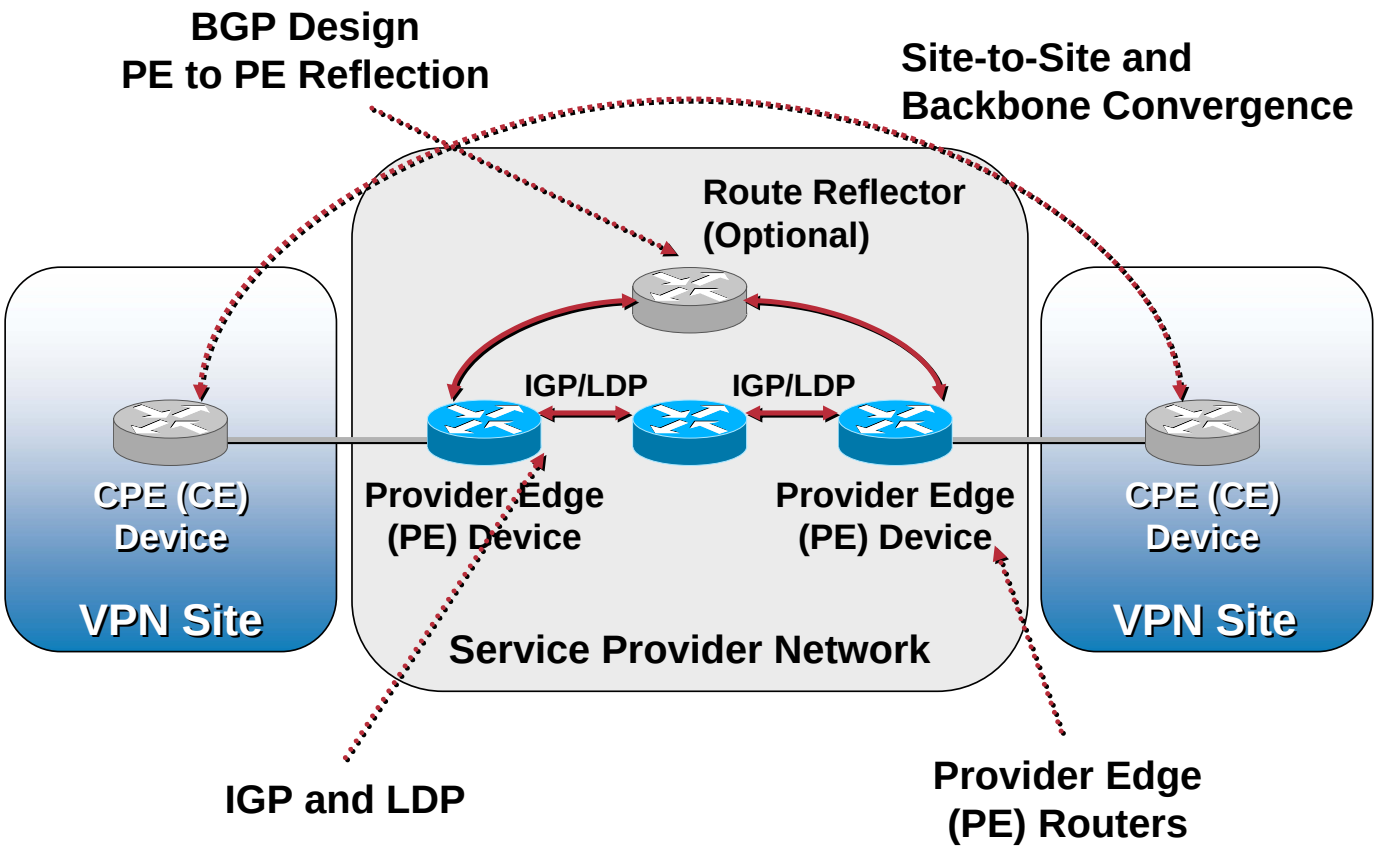


MPLS VPN

Scalability Elements

Elements in Sizing an MPLS/VPN Network

Cisco.com



Elements in Sizing an MPLS/VPN Network

Cisco.com

- **Provider Edge (PE) routers**

  How many do I need, how many can I have, what will affect these numbers, when should I add more?

- **Interior gateway and label distribution protocols**

  How do they interact, how large can they grow, what factors will increase the network deployment size?

- **iBGP**

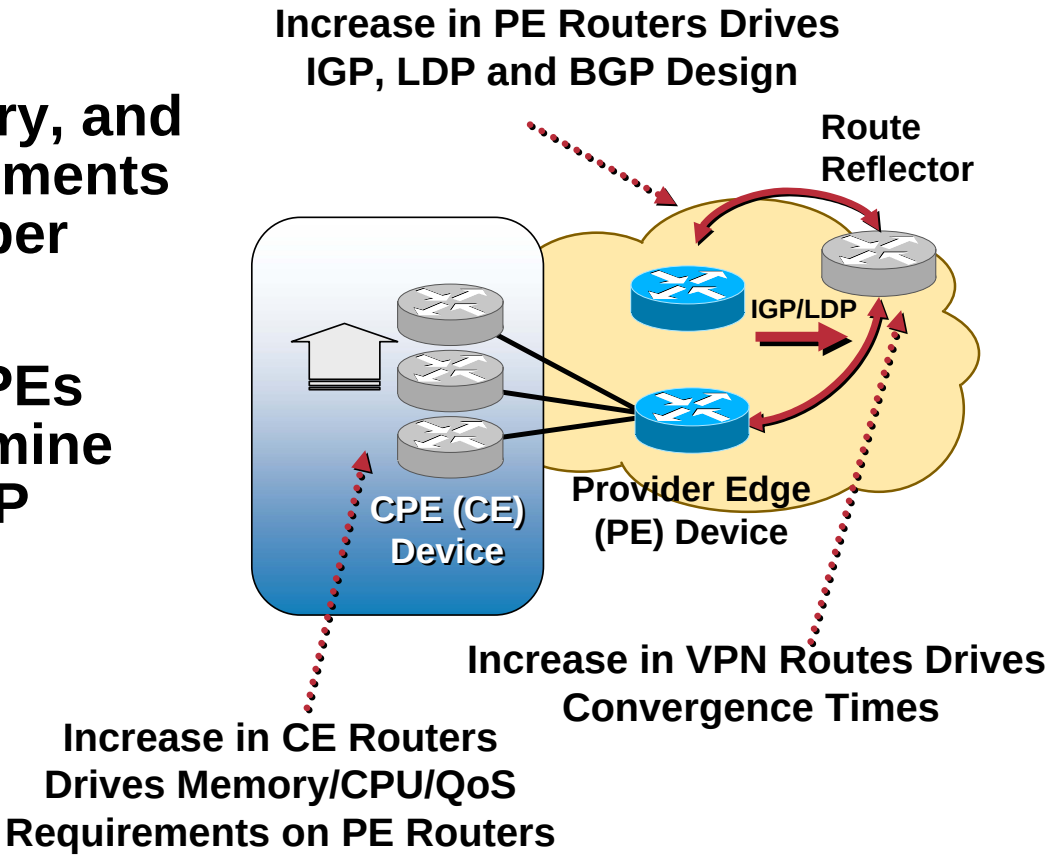
  How many sessions can I have, do I need RRs, Should I confederate?

- **Convergence**

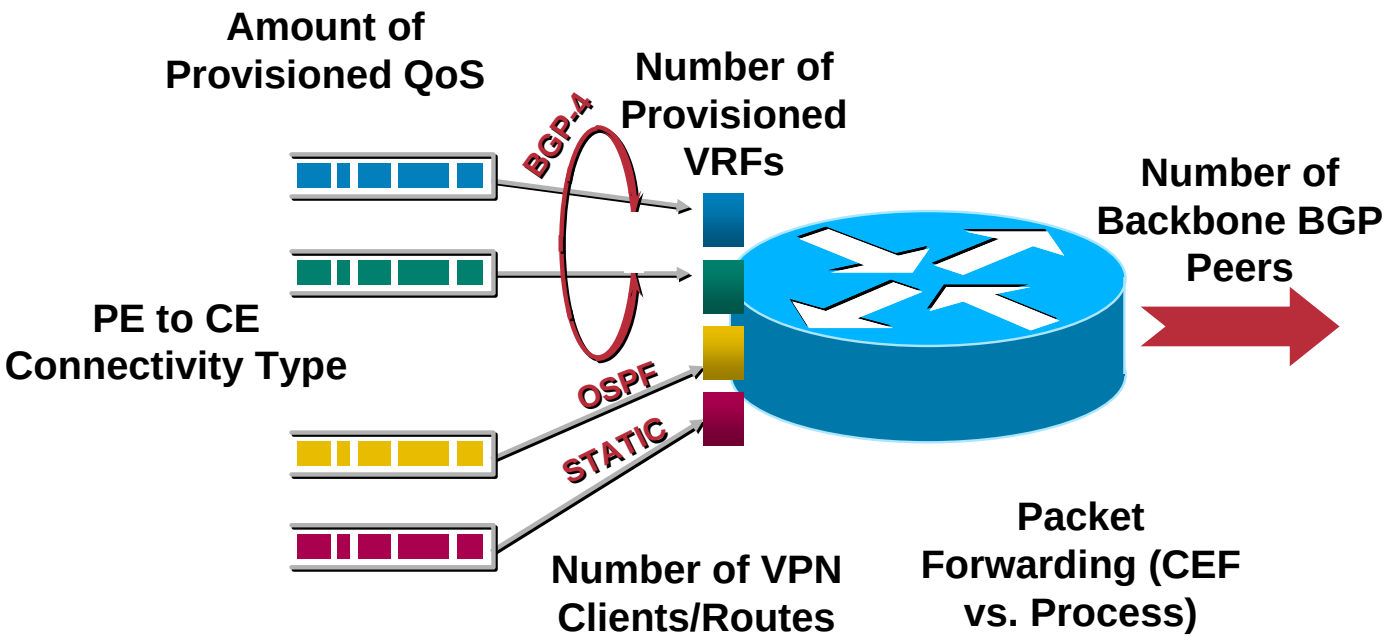
  How quickly can I detect failure, how quickly can I detect new routes?

Elements in Sizing an MPLS/VPN Network

- CPU, memory, and QoS requirements drives number of PEs
- Number of PEs helps determine IGP and BGP design

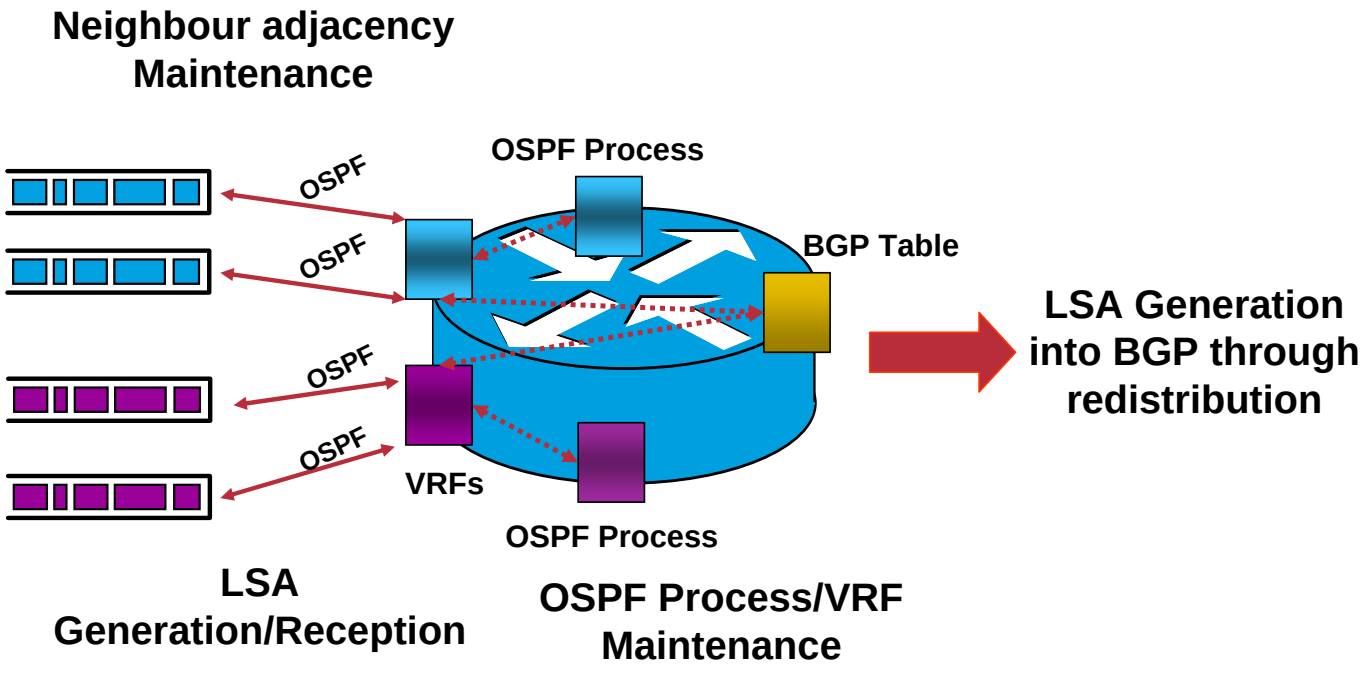


Sizing Provider Edge (PE) CPU Considerations

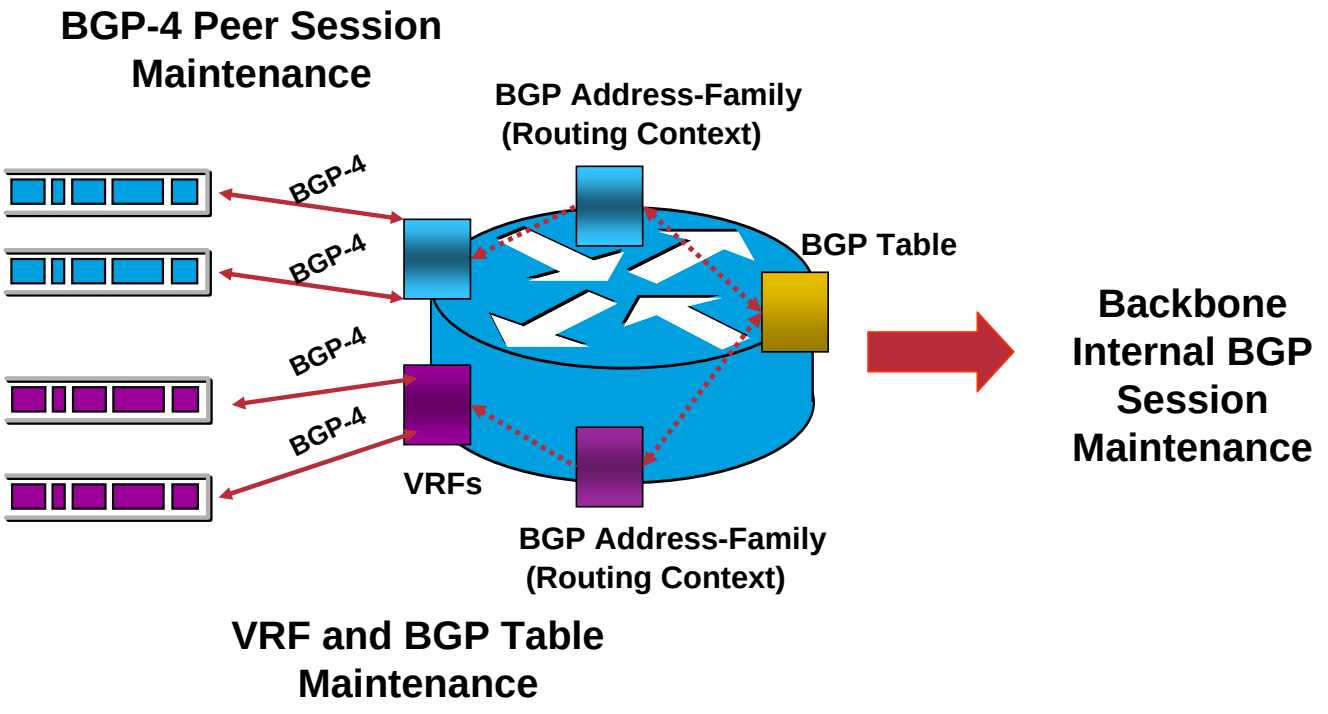


Several Factors Determine CPU Usage

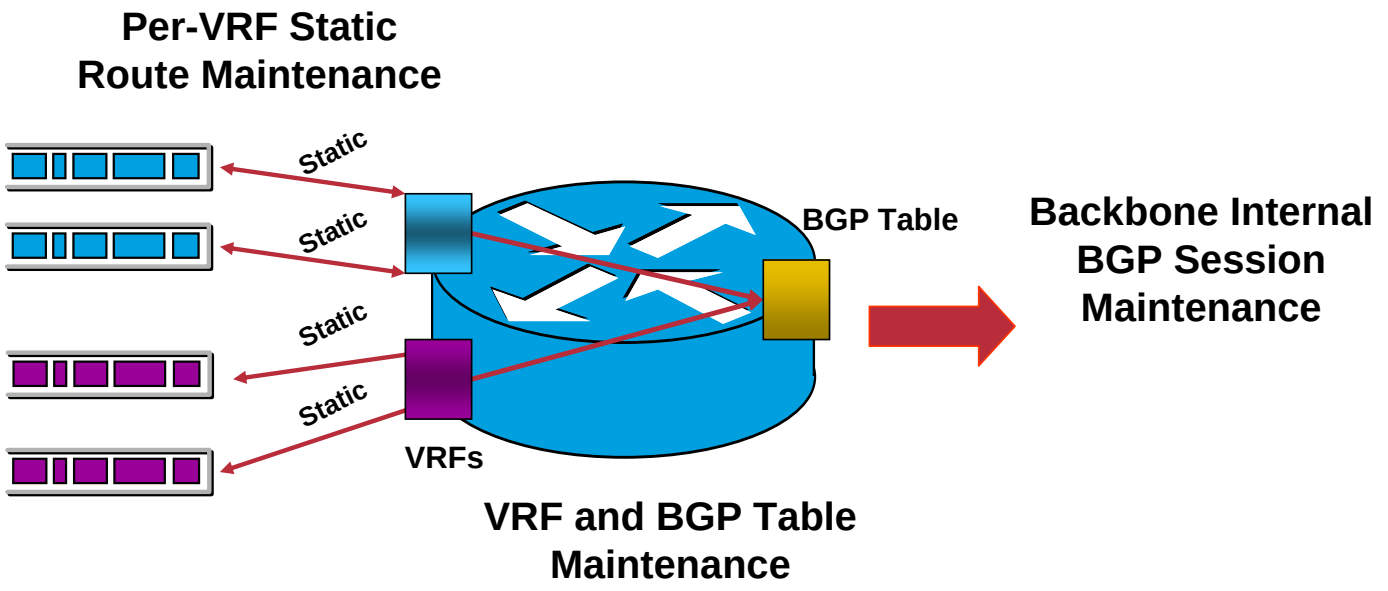
Sizing Provider Edge (PE) OSPF Considerations



Sizing Provider Edge (PE) BGP-4 Considerations

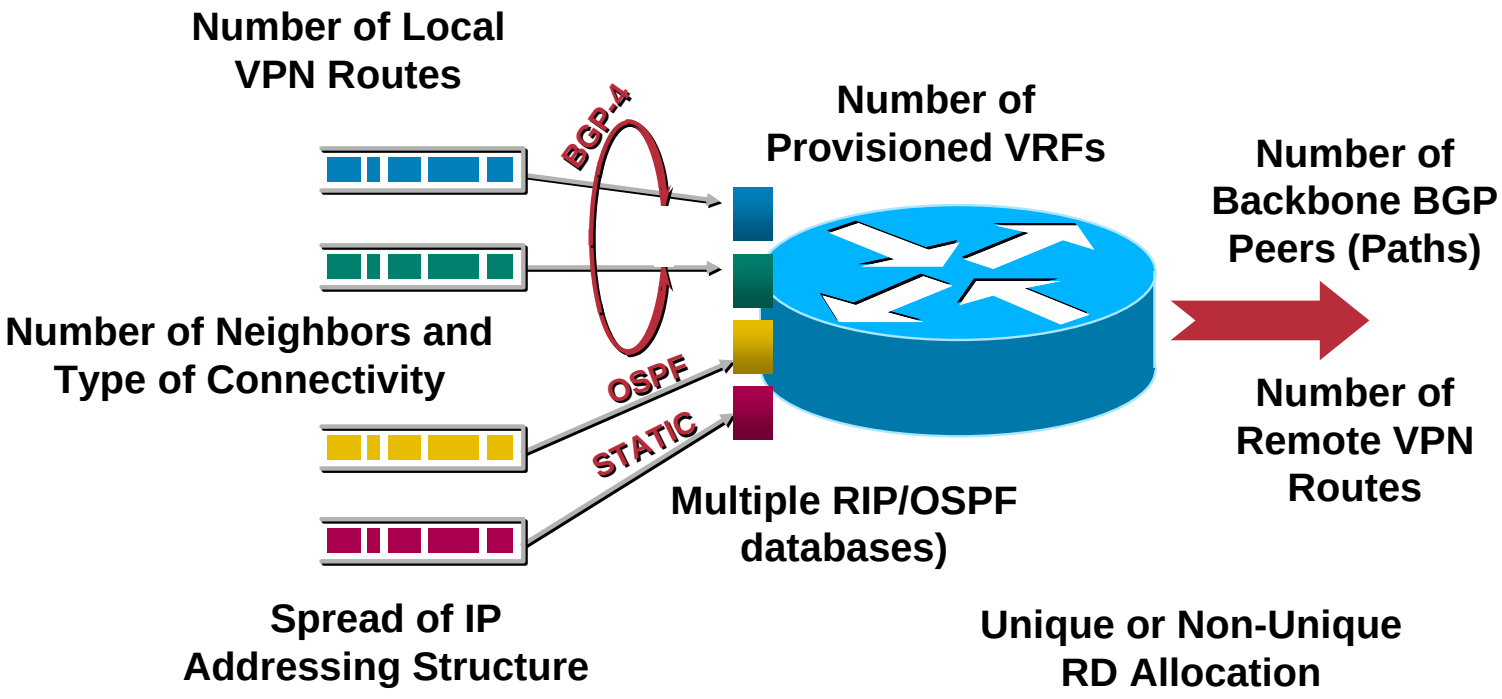


Sizing Provider Edge (PE) Static Routing Considerations



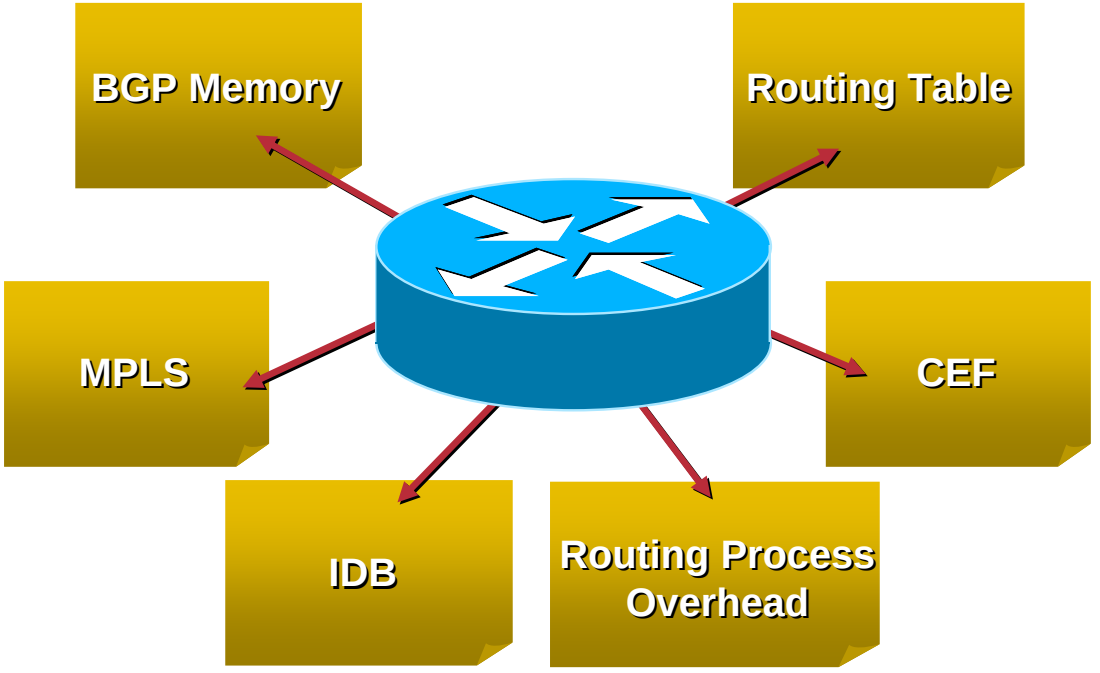
Sizing Provider Edge (PE) Memory Considerations

Cisco.com



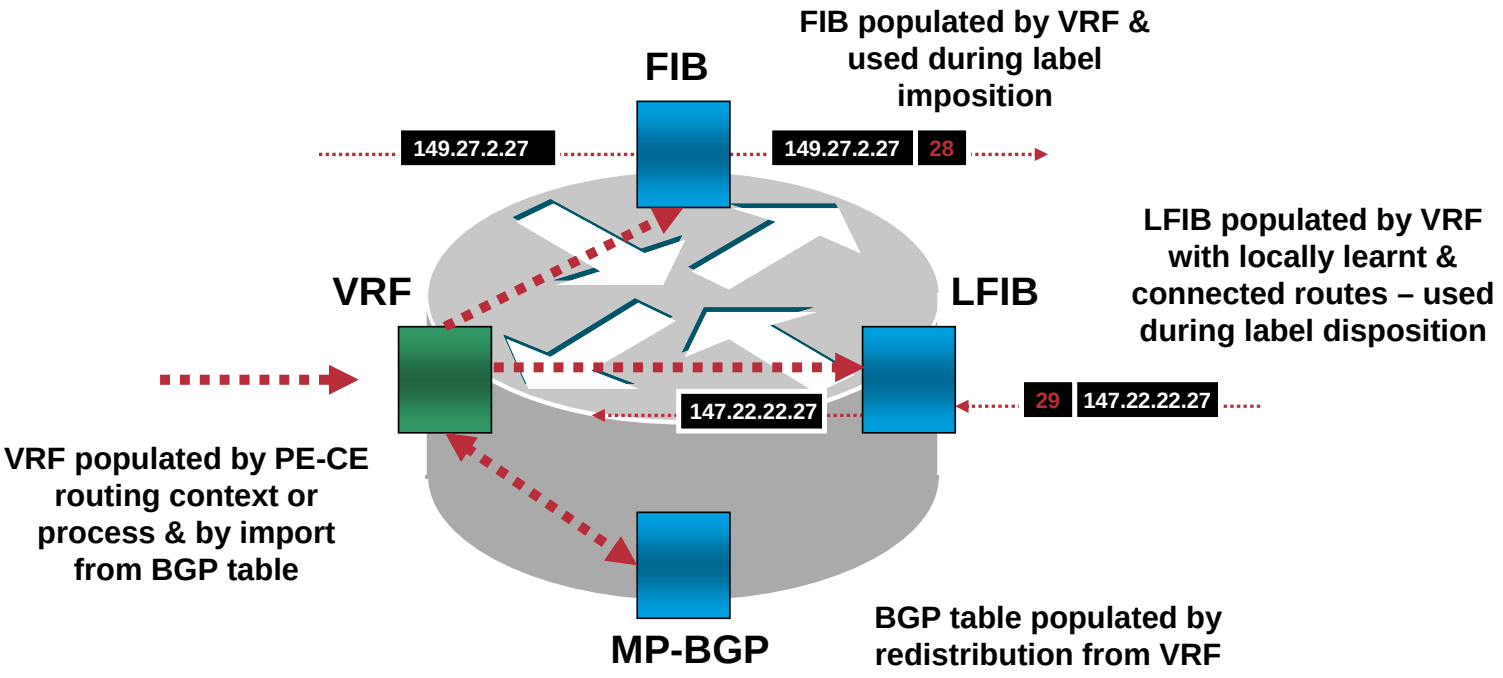
Several Factors Determine Memory Usage

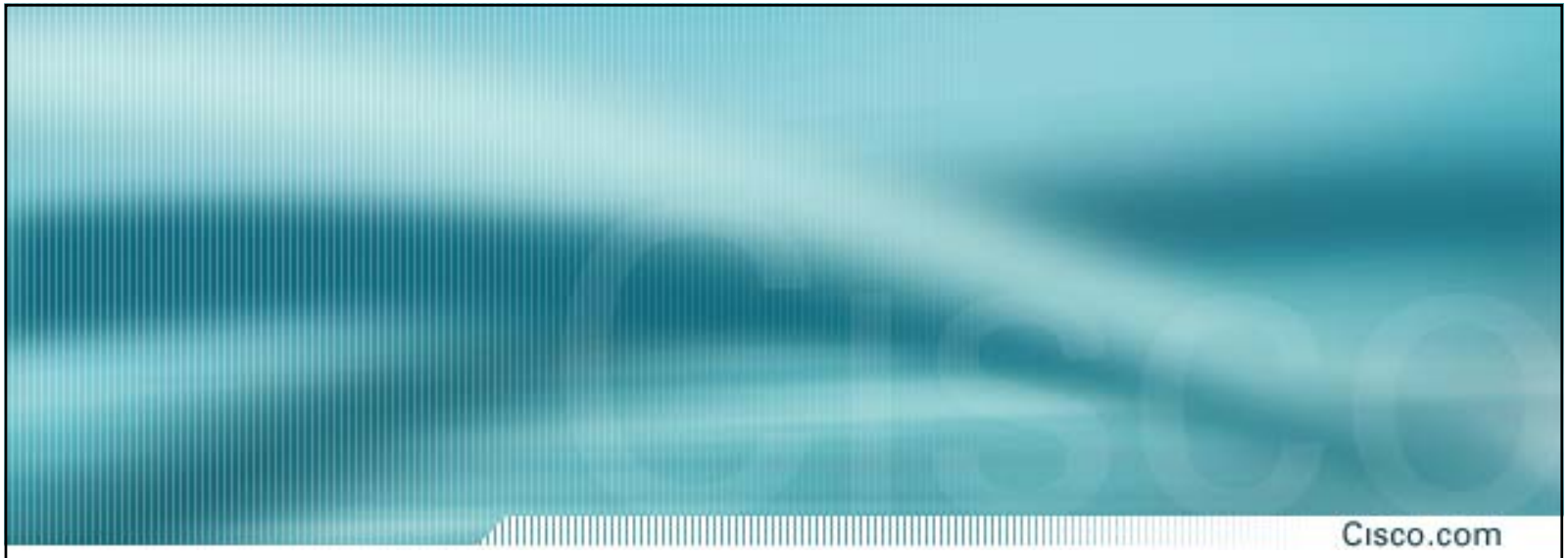
Sizing Provider Edge (PE) Memory Considerations



Several Areas of Memory Usage

Sizing Provider Edge (PE) Table Population



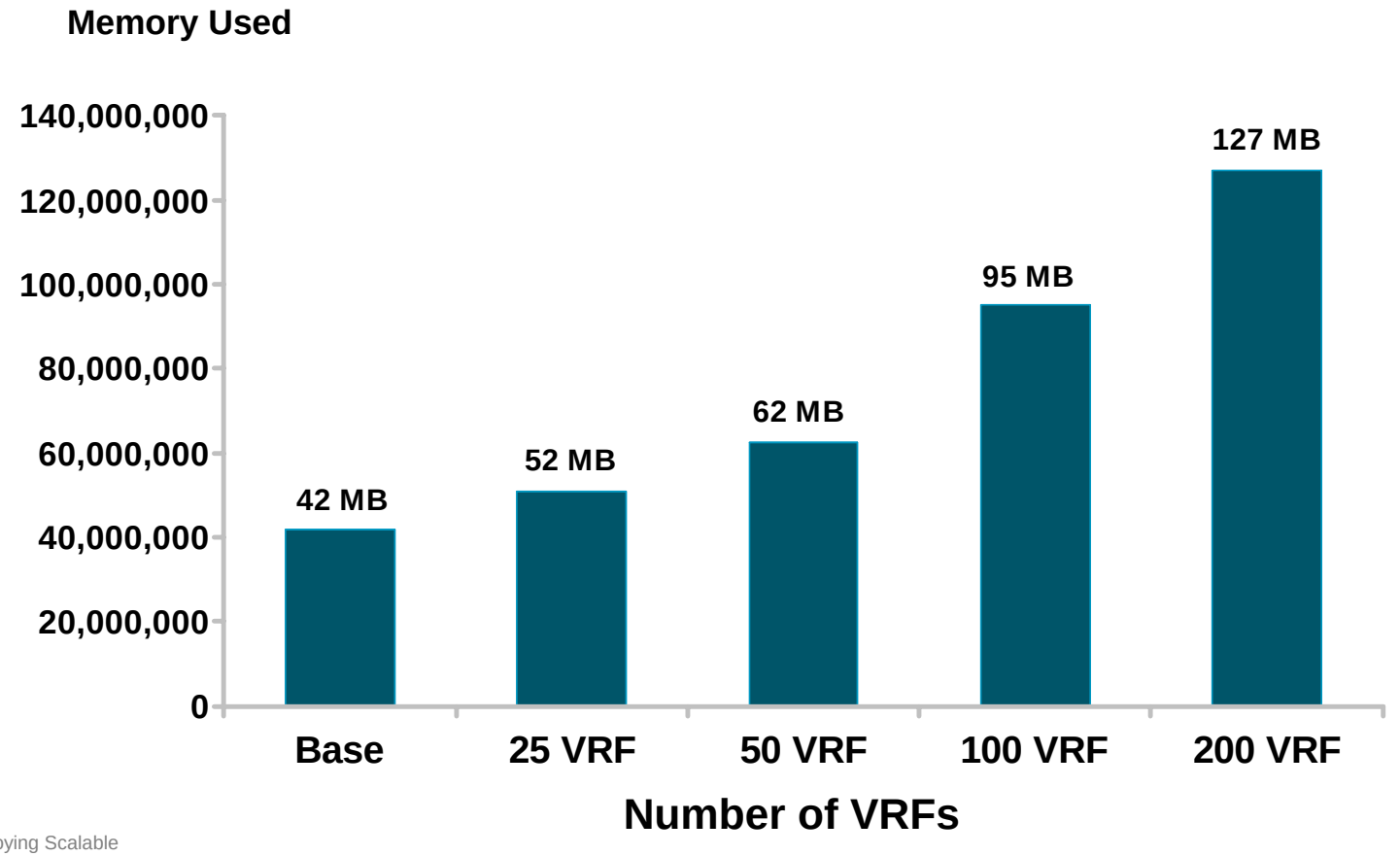


MPLS VPN

Memory Sizing Design Guidelines

VPN Memory Consumption Comparison

Cisco.com



PE Memory Sizing Design Rules

- **~60–70K per VRF**

  32K for base VRF control block, other memory such as CEF, TFIB overhead, IDBs and so on

- **~800–900 bytes per route**

  (Includes CEF, TFIB, and RIB memory in BGP)

- **Remember Cisco IOS® uses memory**
- **Remember Routing Protocol memory**
- **Remember Internet routes!**
- **Remember to leave transient memory**

  Recommended to leave ~30MB free

PE Memory Sizing Design Observations

- **128 Mb platforms are limited**

  (3600 may NOT be suitable for full Internet table and VPNs!!!)

- **256 Mb minimum recommended on PE devices (if possible)**

- **Limit the number of Route Distinguishers per VRF in the same VPN**

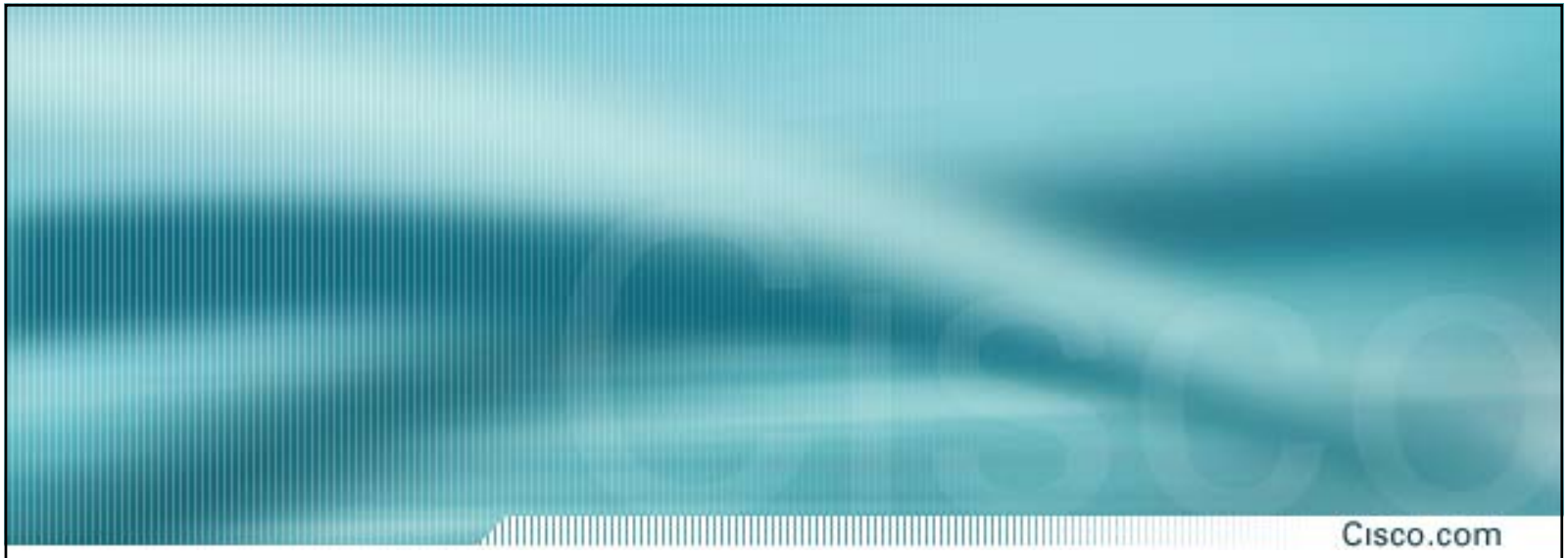
  Unless you require iBGP load balancing with RRs

VRF and Route Limits Summary

- **VRF limits**
  - **Constrained mainly by CPU**
    - **Between 500 and 1000 VRFs for static routing (depending on platform—10 routes per VRF)**
    - **Between 250 and 500 VRFs if using EBGP or RIPv2 (depending on platform—500 routes per VRF)**
- **VPN and global route limits**
  - **Constrained mainly by available memory**
    - **With 256 Mb, 200,000 routes total (IPv4 and VPNv4)**
    - **If Internet table is present, this reduces the memory available for VPNs (current calculations are near 65 Mb for 100K Internet routes—with tightly packed attributes)**

Memory Requirements in One Slide

- PE Memory Requirement:
- ~60-70k per VRF
- ~1K bytes per route
- 10,000 routes across 100 VRFs = 17MB
- 100 VRFs = 70k x 100 = 7 MB
- + 10,000 routes x 1k = 10 MB
- Total = 17MB

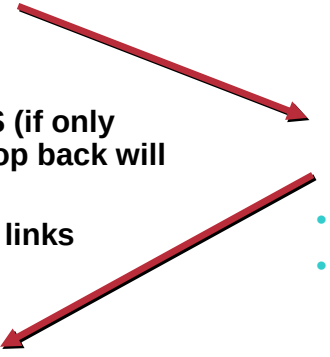


Internal and External Routing Protocol Sizing

IGP Scaling Evolution

Flat Topology

- Link flaps are topology wide
- Full SPF (OSPF), partial IS-IS (if only prefix/metric change—PE loop back will ALWAYS be full SPF!)
- Good for TE—Good if stable links



Area Topology

- Link flaps restricted to area
- Change in type-3 (or type-5) -> partial SPF for all type-3's which announce the prefix
- Requires area filtering (OSPF) or route leaking (IS-IS) for /32 loop backs
- More complex for TE

Confederated AS

- Different IGP per sub-region
- PE addresses restricted to sub-region
- Next-hop re-written at sub-region boundary

IGP Scalability Task

- **Convergence is not the ONLY consideration**

  Flooding, CPU usage, link flapping and so on

- **Task: Design an IGP backbone to support one million VPN sites**

- **Quick math:**

  2000 PE devices * 500 VRF forwarding interfaces

  This breaks the (theoretical) IGP rules for number of routers in a single area

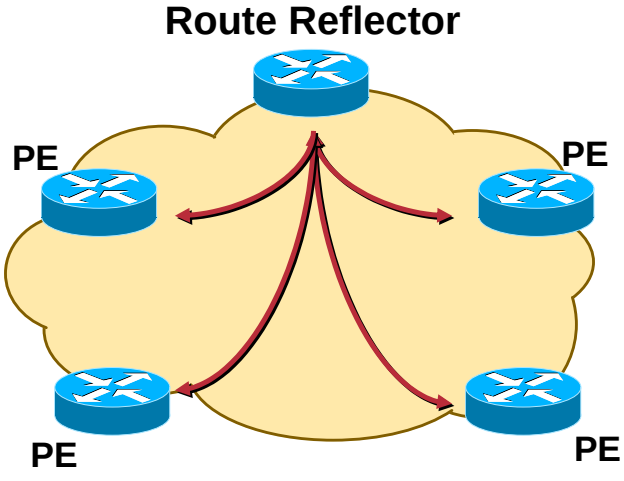
  Assumes multiple areas or even confederations

Sizing the IGP Area Topology

- **Backbone area sizing issues**
  - **Number of routers**
  - **Number of links and topology**
  - **Leaking of PE loop back addresses**
    - **Large MPLS networks = lots of PEs potentially!**
- **Maximum allocated labels**
  - **Defaults to 100K labels—including IGP and VPN**
  - **Can be changed ‘mpls label range’ command**

MP-BGP Deployment Requirements

- Full iBGP mesh requirement between PE routers that require the same VPN information
- Could partition to break up the mesh—Mini iBGP clusters
- Easier to use route reflectors and/or confederations
- Confederations require Inter-AS mechanism, RRs much easier
- Could partition to break up the topology and reduce processing overhead on PE routers



Remember—PE Only Needs Routes for Attached VPNs

MP-BGP Used to Distribute VPN Prefix Information

Peer-Group Concepts for RR Clients

Cisco.com

- Same set of updates sent to group of neighbors using peer-group concept
- Results in lower CPU and I/O memory consumption
- Dramatic improvement observed in convergence using peer-group over non-peer group
- 250K prefixes with 500K paths—100 core and 32 PE clients

  With peer-group → 12 minutes, 29 seconds

  Without peer-group → 56 minutes, 6 seconds

Use Redundant Route Reflectors with Peer-Groups

RR Tuning and Recommendations

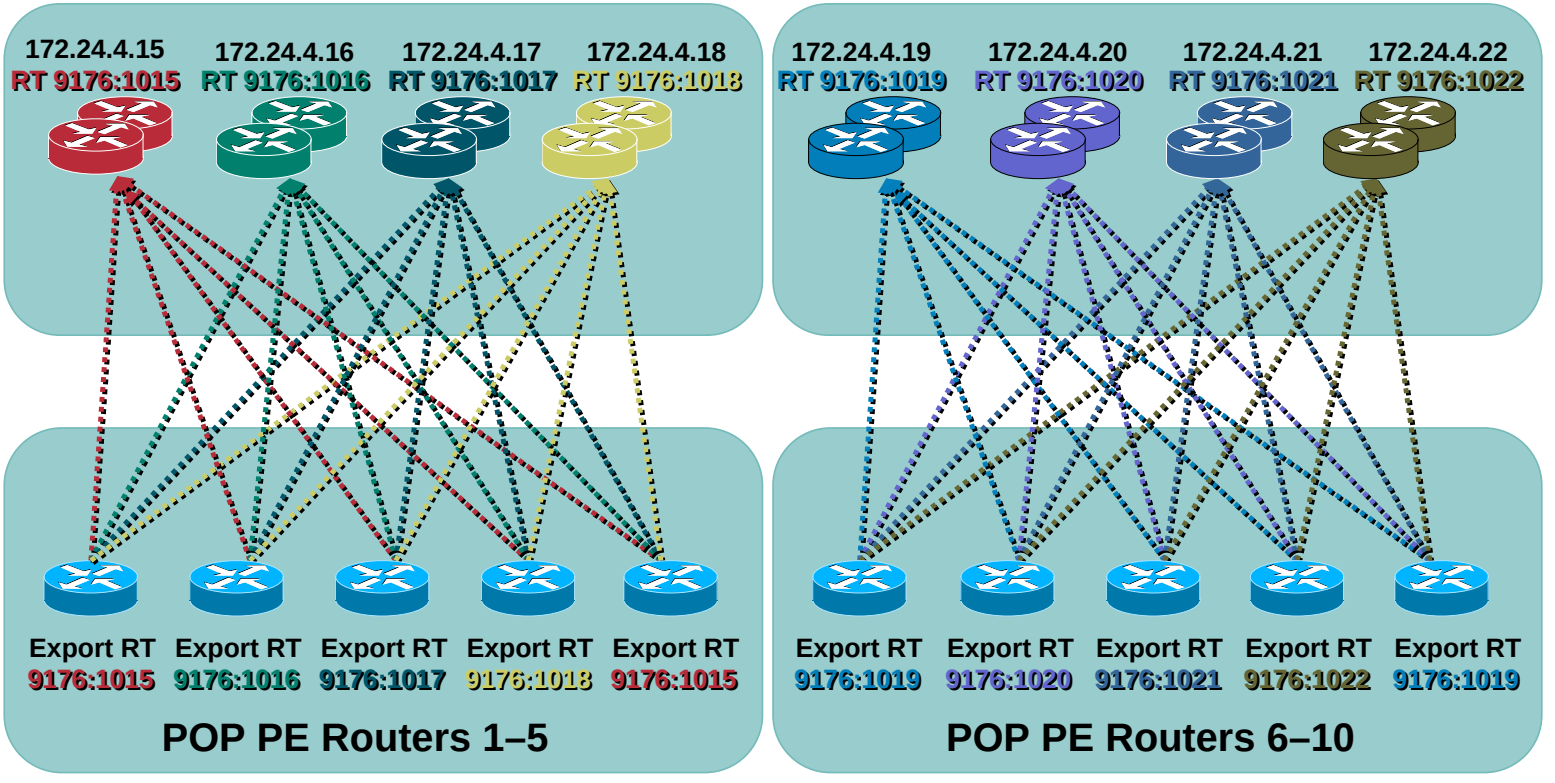
Cisco.com

- **Keep RRs separate for IPv4 and VPNv4**
  - **Better stability, faster convergence**
  - **Meets SLA requirements (with high number of routes)**
- **Use route reflector server model**
  - **Dedicated for RR function—Not in forwarding path**
  - **Conserve CPU and memory for faster convergence**
- **Recommended RR for VPN is 7400**
  - **1RU, 512MB, 375MHz R7K**
  - **NPE-400 also a good choice—3RU (7204), 512MB, 350MHz R7K**
  - **Good CPU Power**

RR Core Design for Large Scale VPN Deployment

- **Scale for 250 thousand VPN sites**
- **500 VRF interfaces per PE device**
  - **500 PE routers in total**
- **50 POP topology**
  - **10 PE routers per POP**
- **Centralized route reflection topology**
  - **Split PE routers into groups and use specific route reflector RT values**

Centralized Route Reflection Design for Large Scale VPN Deployment



RT Value Assignment for Each Core Route Reflector