IP Routing Protocol Scalability Theory and Examples

Alvaro Retana (aretana@cisco.com)

IP Routing Deployment and Scalability

Slides by: Scott Sturgess, Gerry Redwine and Daniel Walton.

Agenda

- **Scope of the Presentation**
- **Scalability Building Blocks**
  - **Hierarchy**
  - **Redundancy**
  - **Addressing and Summarization**
- **Link State Scalability**
  - **ISIS Scalability**
  - **OSPF Scalability**
- **BGP Scalability**

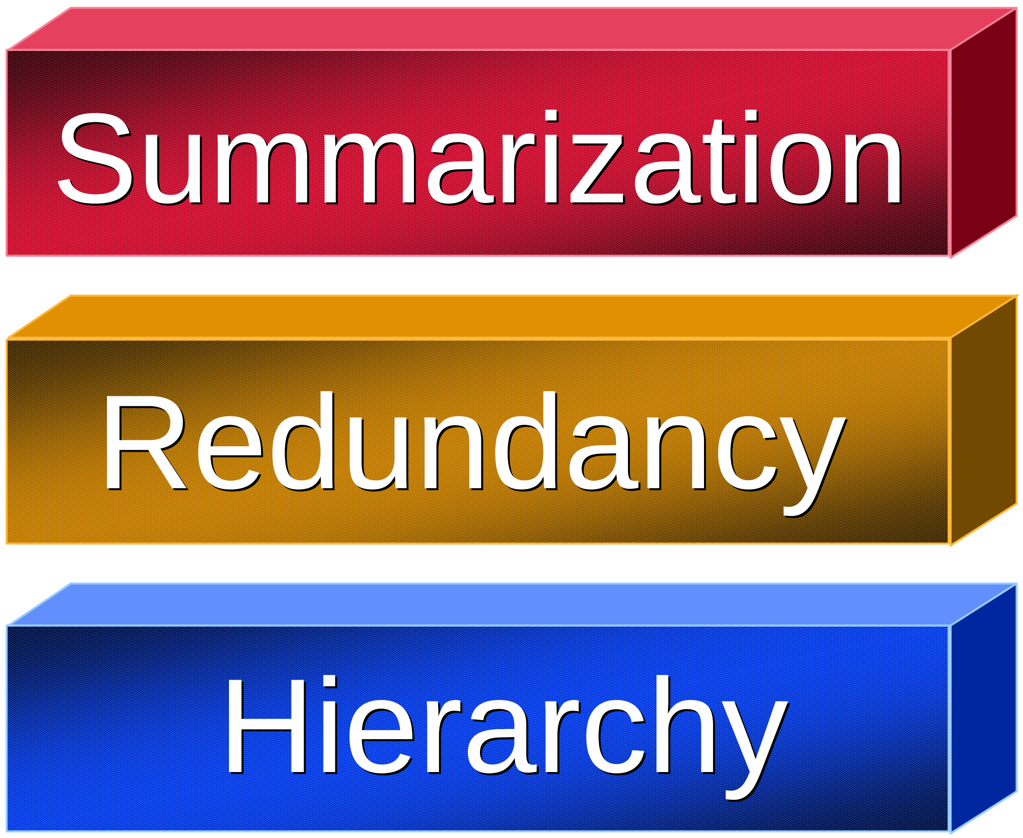
Scope of the Presentation

- Cover the building blocks of scalable IP routing networks: hierarchy, redundancy and summarization.
- Correlate these building blocks to characteristics and features in ISIS, OSPF and BGP.
- Prior knowledge of the protocols is assumed.

Agenda

- Scope of the Presentation
- **Scalability Building Blocks**
  - **Hierarchy**
  - **Redundancy**
  - **Addressing and Summarization**
- Link State Scalability
  - ISIS Scalability
  - OSPF Scalability
- BGP Scalability

Scalability Building Blocks



Agenda – Building Blocks

- Relationship between Convergence, Stability and Scalability.
- Impact/Use of Hierarchy/Redundancy/Addressing and Summarization
  - Hierarchy
  - Redundancy
  - Addressing and Summarization

Network Design Goals

Fast Convergence

Stable

Scalable

Network Design Goals

- **Often, *conflicting* design goals must be achieved when building a network.**
- **Fast Convergence usually implies taking an aggressive approach at changes in the network.**
- **Stability is related to minimizing the changes and/or their propagation in a network.**
- **A scalable network design takes into account both requirements and builds a compromise between the two.**

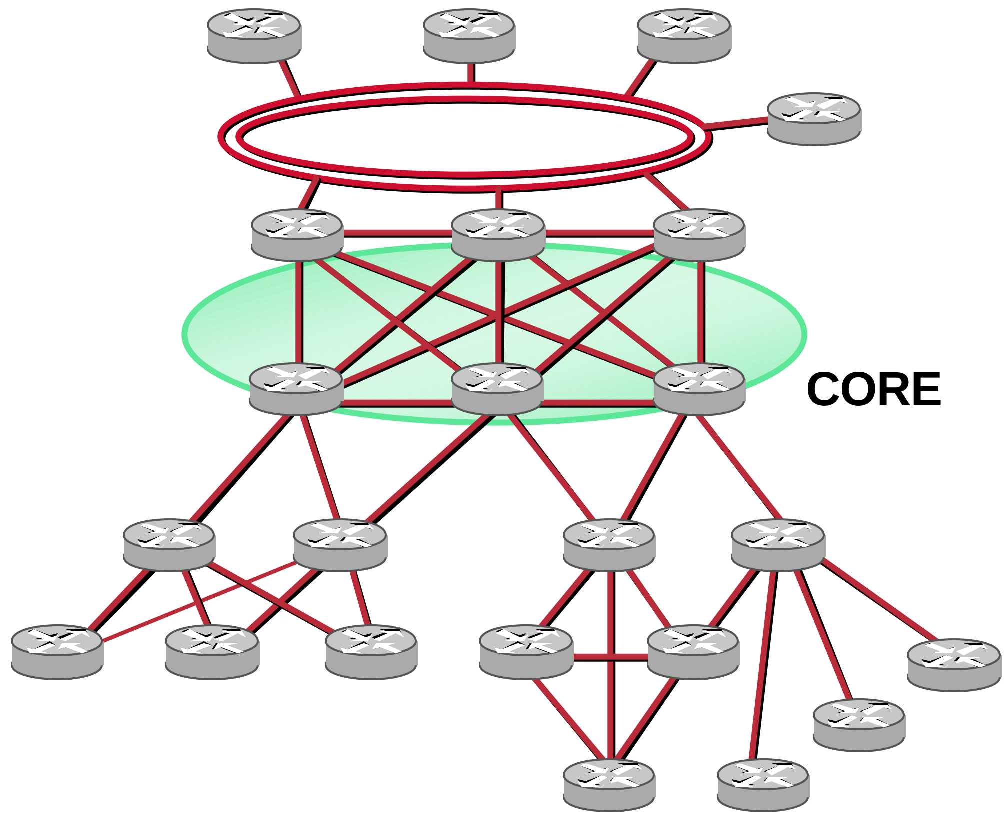
Hierarchy

- **Hierarchy refers to the simplification of the functions in a network by clearly defining levels of responsibility.**

    Backbone/core

    Distribution

    Access



Hierarchy

- **Each level of the hierarchy is responsible for specific functions. For example:**

  Core/Backbone: unified connectivity, packet switching

  Distribution: route summarization, traffic aggregation, services

  Access: "customer" entrance to the network, packet filtering and classification

- **Facilitates the scalability of the network by providing a clear indication of where the growth is needed.**

- **Enhances the stability of the network by isolating functions, traffic, routes, etc.**

Redundancy

- **Redundancy provides alternative (or duplicate) capacity to the network.**

  Usually in the form of additional links that provide connectivity between the nodes.

- **Allows for the provisioning of back-up and/or equal-cost paths.**

  Higher aggregate bandwidth contributes to the ability of the network to handle an increasing load.

  Contributes to the stability by supplying existing alternate paths.

Redundancy

- **BUT…too much redundancy may be harmful ☹**

  More protocol adjacencies must be maintained.

  The number of possible paths in the network increases, as does the complexity of the algorithms used to find them.

  Routing information must be propagated (flooding) through a larger number of paths, which will result in an increase in the duplicate data received.

  Convergence may be delayed and resources may be exhausted.

Addressing and Summarization

- **Address assignment must be made with summarization in mind.**
- **Proper addressing and summarization results in the reduction of routing information.**

    Allows for faster convergence (less routes).

    Increased stability (the components are not propagated).

    Contributes to scalability by allowing the routers to store more information.

Addressing and Summarization

- **Improper summarization may lead to loss of specific information, resulting in non-optimal routing.**

  Each protocol has its own way to summarize and allow the leaking of specific components.

Agenda

- Scope of the Presentation
- Scalability Building Blocks
  - Hierarchy
  - Redundancy
  - Addressing and Summarization
- **Link State Scalability**
  - **ISIS Scalability**
  - **OSPF Scalability**
- BGP Scalability

Link State Scalability – Issues

- **Link-State Packet (LSP) Flooding**

  number of neighbours, redundant paths, buffers, speed of links and size of the network, detection speed

- **Shortest Path First (SPF) Computation**

  forwarding continues during SPF

Agenda – Link State Scalability

- **Hierarchy**
  - **Area types and flow of routing information**
  - **Use and limitations of Hierarchical Networks**
  - **LSA Filtering/Route Leaking**
- **Detection and propagation of changes**
  - **Fast Hellos**
  - **LSA/LSP Generation**
  - **SPF Runs**
  - **Exponential Backoff**
- **Other tips...**

Agenda – Link State Scalability

- **Hierarchy**
  - **Use and limitations of Hierarchical Networks**
  - **Area types and flow of routing information**
  - **LSA Filtering/Route Leaking**
- Detection and propagation of changes
  - Fast Hellos
  - LSA/LSP Generation
  - SPF Runs
  - Exponential Backoff
- Other tips...

Hierarchical Networks

- **Benefits**

  By creating areas you hide instability in one part of the network from the other parts

  Only partial SPF needs to be run when network flaps in another area

- **The most expensive part of route computation (actual Dijkstra) is run over intra-area topology only**

  The SPT is built from nodes in the area

  Dividing networks into areas means less CPU cycles spent on Dijkstra

Hierarchical Networks

Cisco.com

Area Routing

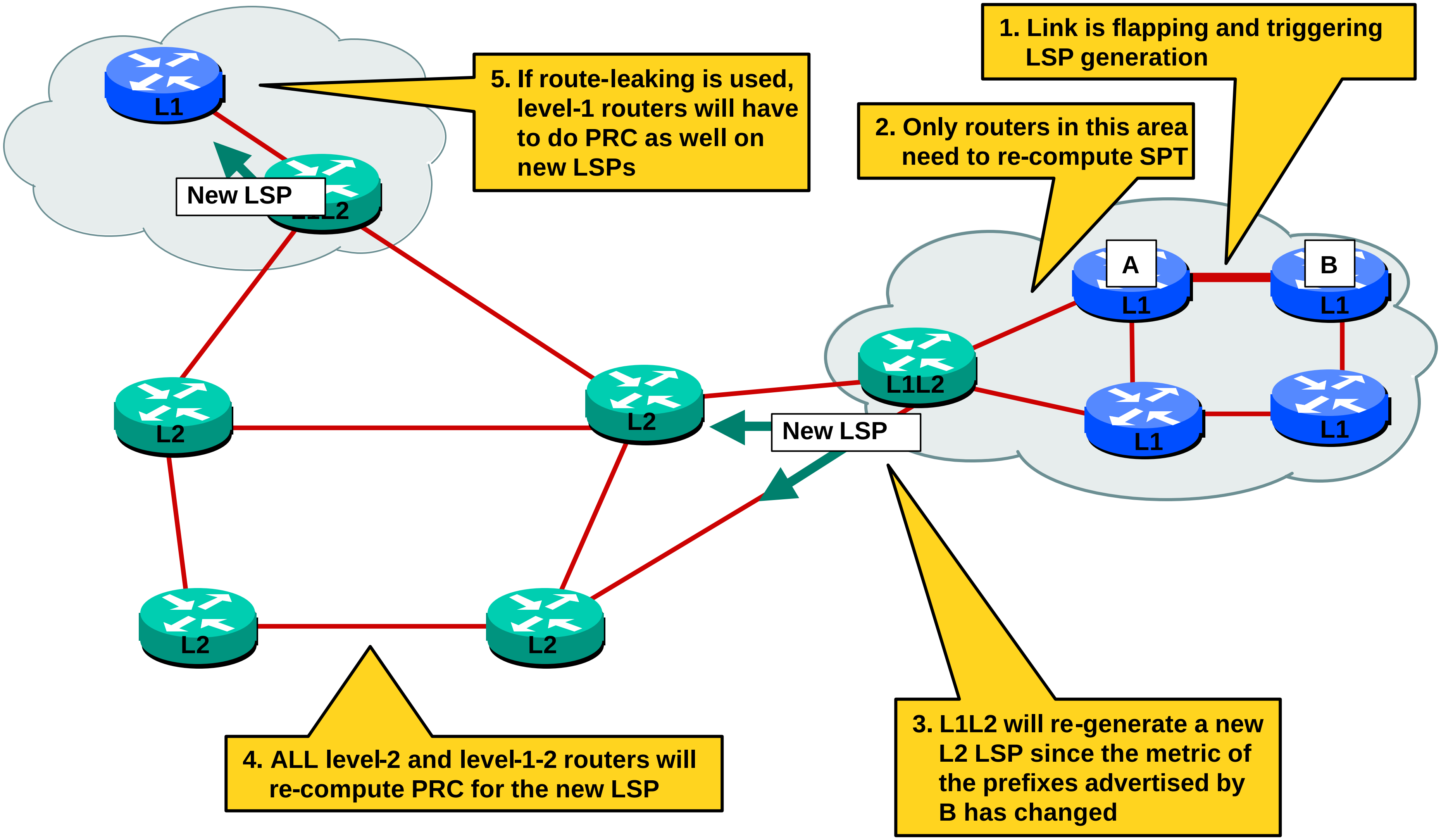
- **Area routing has the benefit to reduce the size of the SPT each router will have to compute**

  SPF will take less CPU (if needed….)

- **However, even with area routing, changes in one area may have an impact on other areas.**

Hierarchical Networks

Cisco.com



Hierarchical Networks

Address Summarization

- **Reduces the number of prefixes and adds stability.**
- **Summarization in ISIS**

  From L1 areas into the L2 backbone

  From L2 leaking down into L1 areas

  When redistributing into L2 or L1

- **Summarization in OSPF**

  At Area Border Routers (ABR)

  At AS Border Routers (ASBR)

  When translating type-7 LSAs into type-5

OSPF Summarization

- **Instead of advertising many specific routes, advertise only one summary route**

  Summarize at ABR

  area <areaID> range <network> <mask>

  Summarize at ASBR

  summary-address <network> <mask>

- **Reduces LSA database and routing table size. Drawback is possible sub-optimal routing.**

Hierarchical Networks

Sub-Optimal Routing

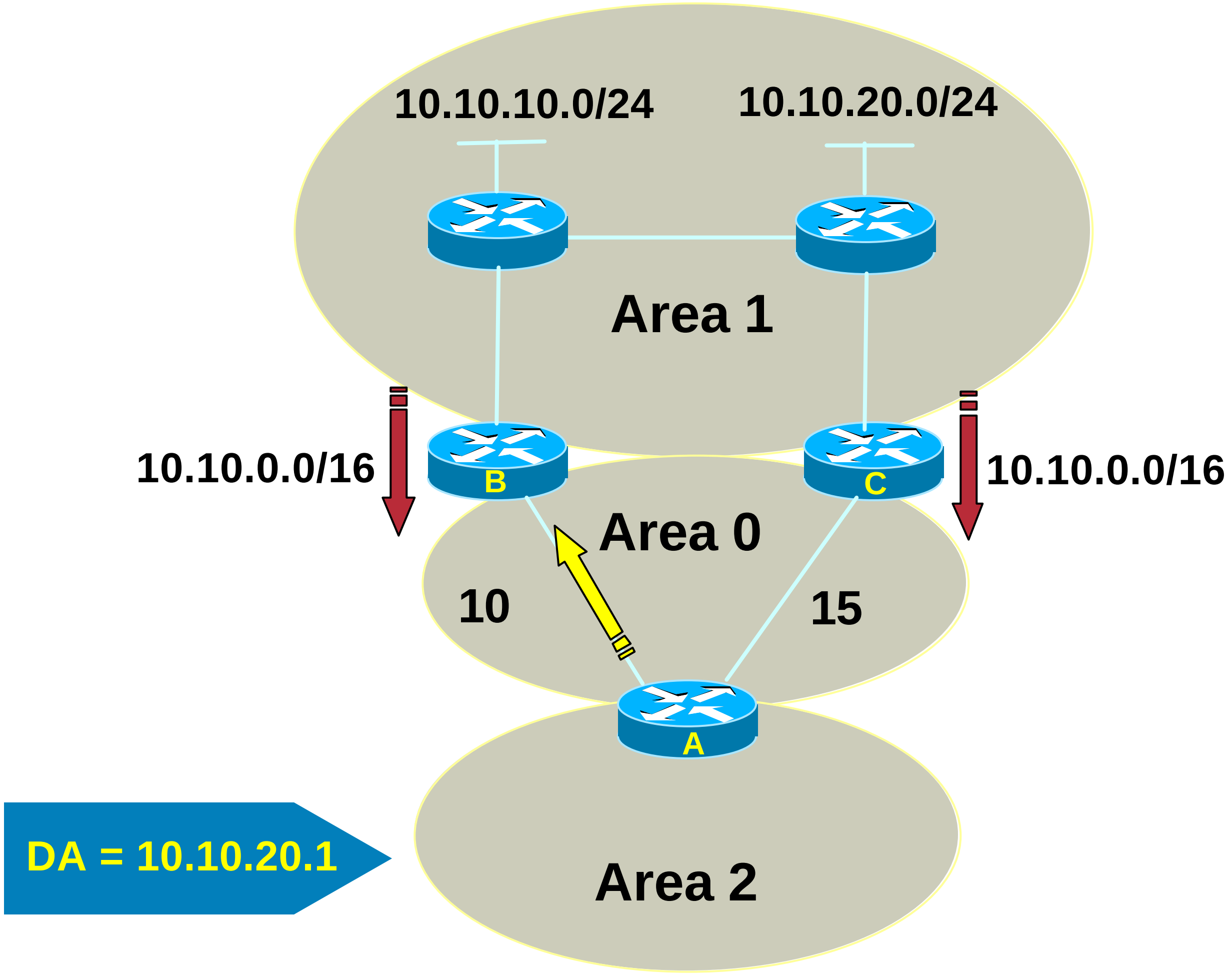
- **If the end-to-end information is not advertised to all the areas, then a sub-optimal path may be used (if more than one exit exists from the area).**

  ISIS: before *route leaking*, L1 areas won't learn any information from other areas.

  OSPF: stub areas (stub, NSSA, etc.) don't accept (in varying degrees) information from other areas.

Summarization Example

Sub-optimal Routing



Hierarchical Networks – Nuts and Bolts

- **Two Layer Hierarchy**

  Contiguous backbone and areas connected to it.

  ISIS: the backbone is level-2, areas are called level-1. A node can be part of both levels, in the same area (L1L2 router).

  OSPF: area 0 is the backbone. A router can be part of multiple areas (ABR).

- **Same algorithms apply to both layers**

  Separate SPF, PRC and/or partial SPF for each level or area.

Hierarchical Networks – Types of Routers

- **Internal Routers**

  Neighbors only in the same area and may only have information about own area

  OSPF: inside an area (not area 0)

  ISIS: L1-only routers (look at the attached bit in L1 LSPs to find the closest L1L2 router)

- **Backbone Routers**

  Have information about the backbone topology. Know which destinations are reachable outside the backbone and how to reach them through the backbone topology.

  OSPF: inside area 0 (has Summary LSAs describing other areas)

  ISIS: L2-only routers

Hierarchical Networks – Types of Routers

Cisco.com

- **Border Routers**

  Connect two (or more) levels (or areas). May have neighbors in any area and has one LSDB for each level (or area) it belongs to.

  OSPF: Area Border Routers (ABR)

  ISIS: L1L2 routers

- **AS Border Routers**

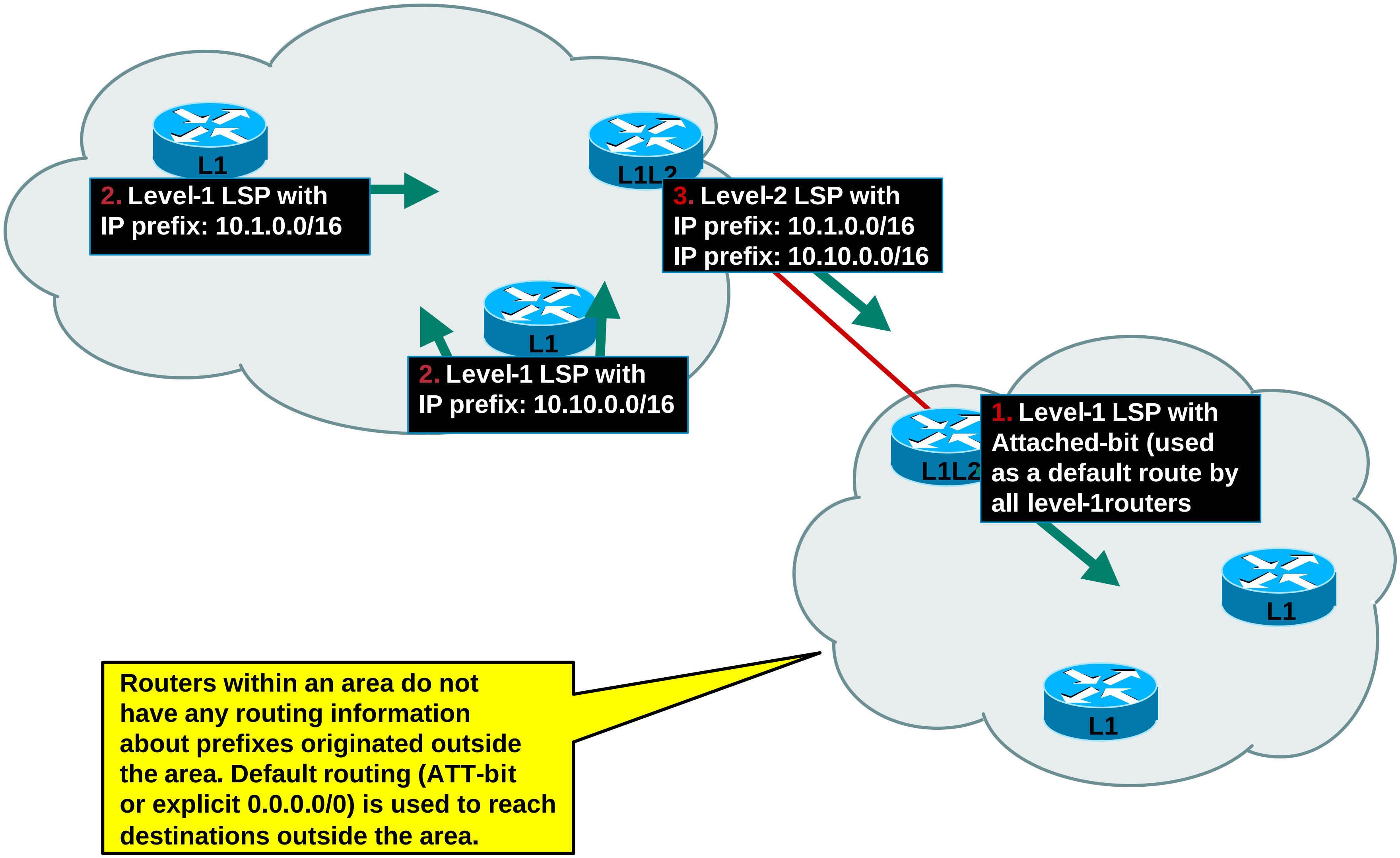
  Accept external information into the local domain.

  OSPF: ASBRs may not be placed in Stub areas (only NSSA and “normal” areas).

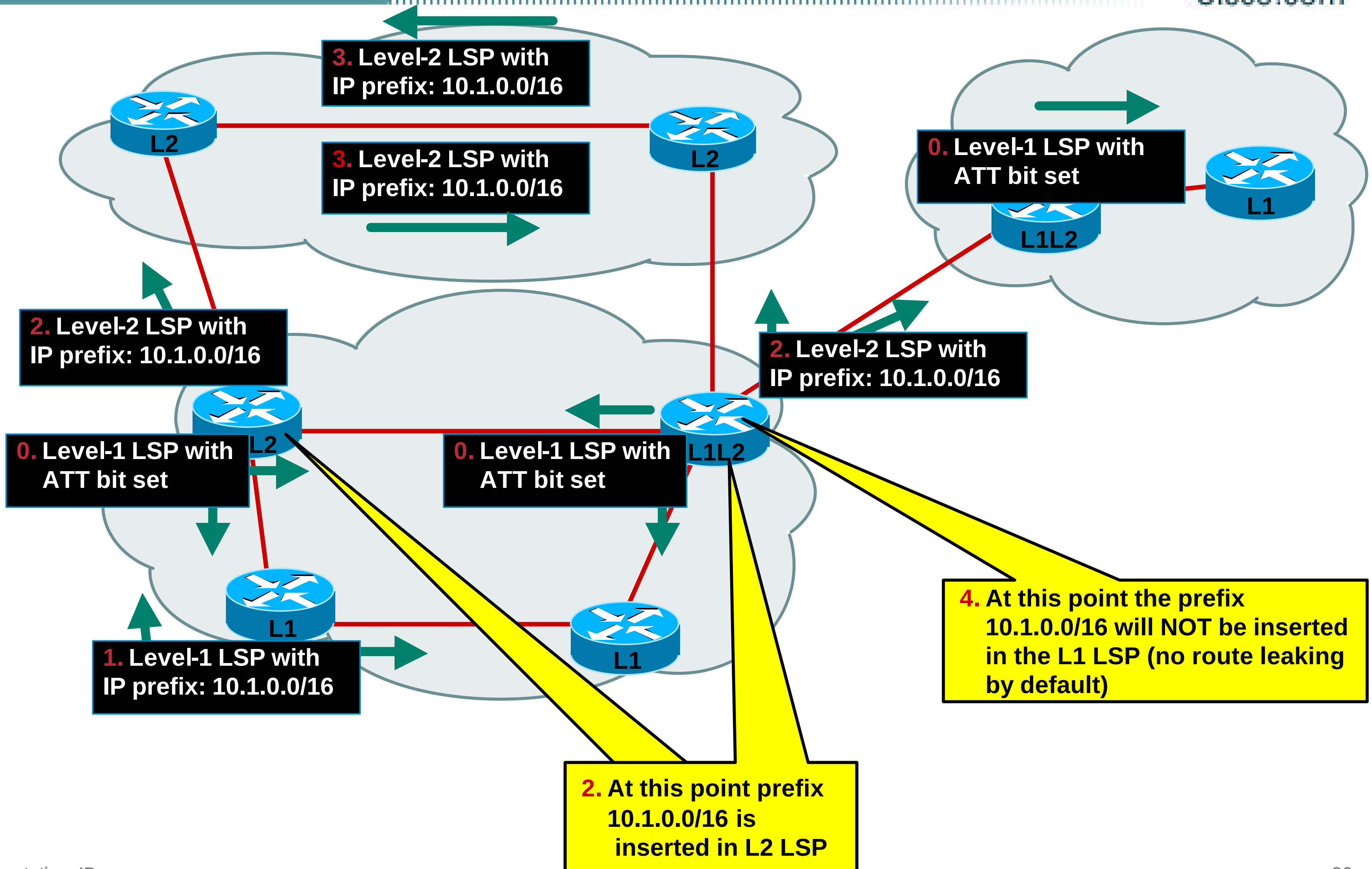
  ISIS: any node can perform this function..

Hierarchy in IS-IS: Routing Levels

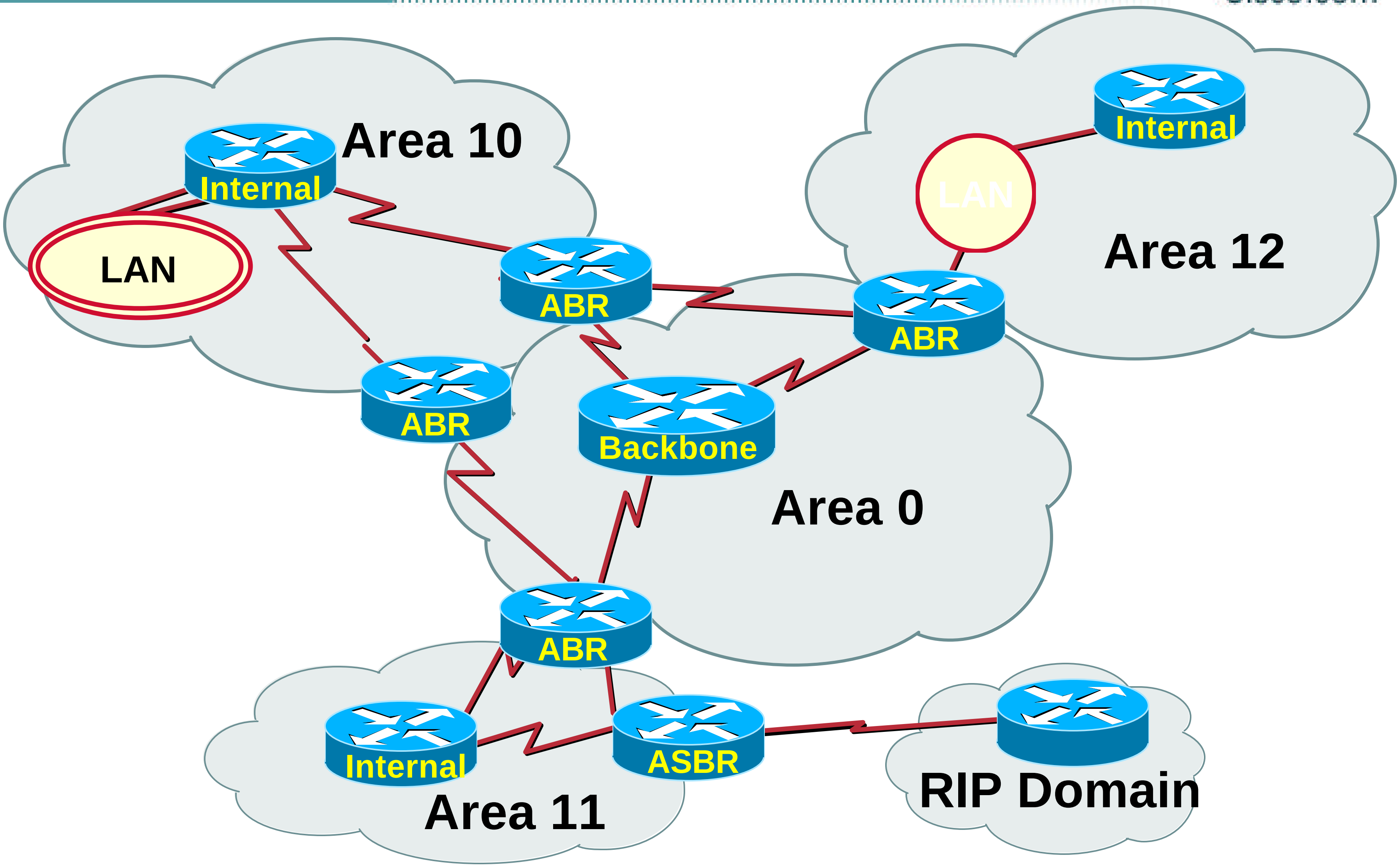
Cisco.com



Hierarchy in IS-IS: Routing Levels



Hierarchical OSPF Network - Example



L1 OR L2-only Networks

- **A router can't tell whether it is a transit node**

  cisco default is L1L2

  this will make the backbone larger than necessary

  always configure L1-only or L2-only when possible

- **Start with Level 2 !**

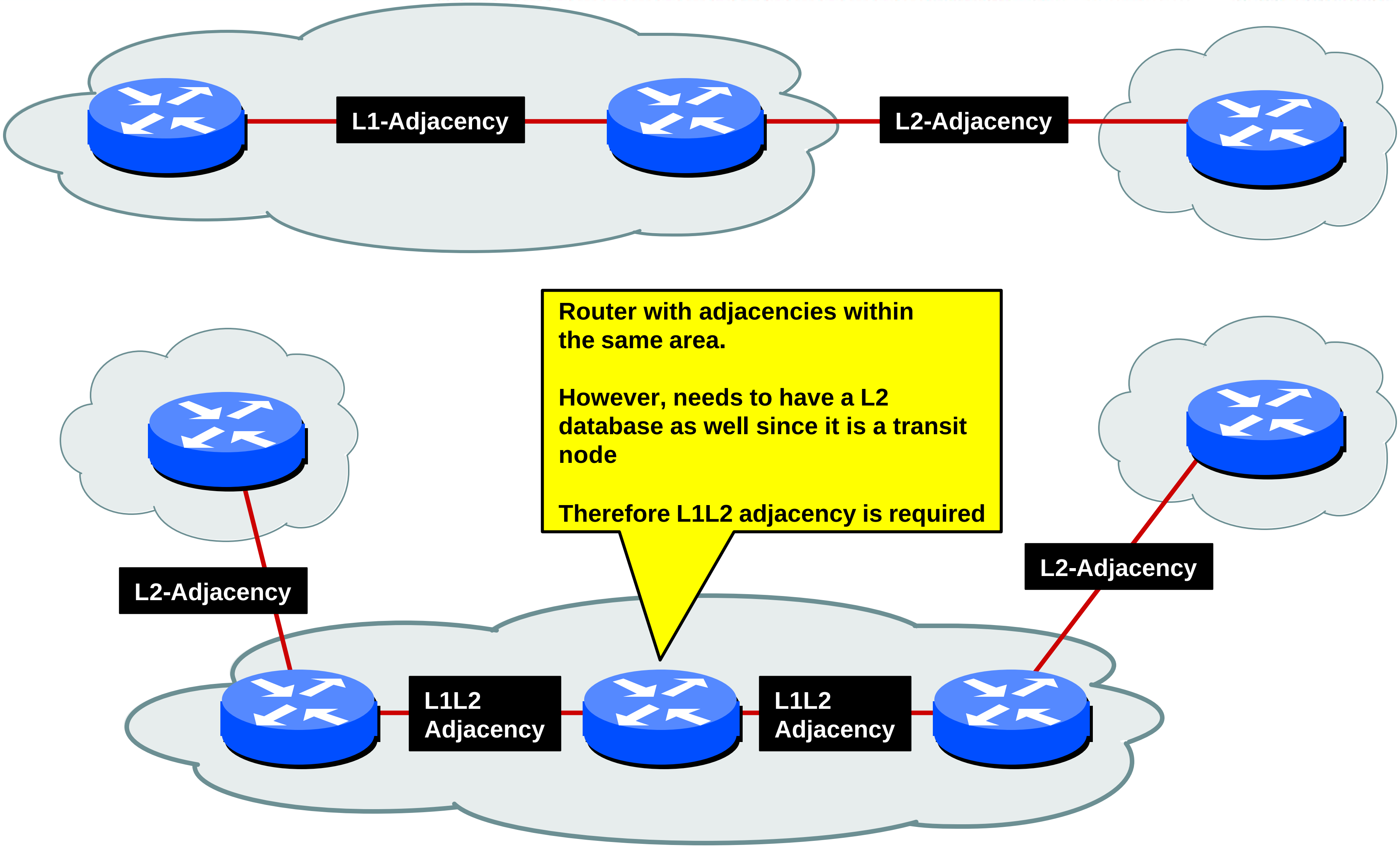
  Why ?

  easier to migrate to hierarchical topology

  level 2 is the backbone !

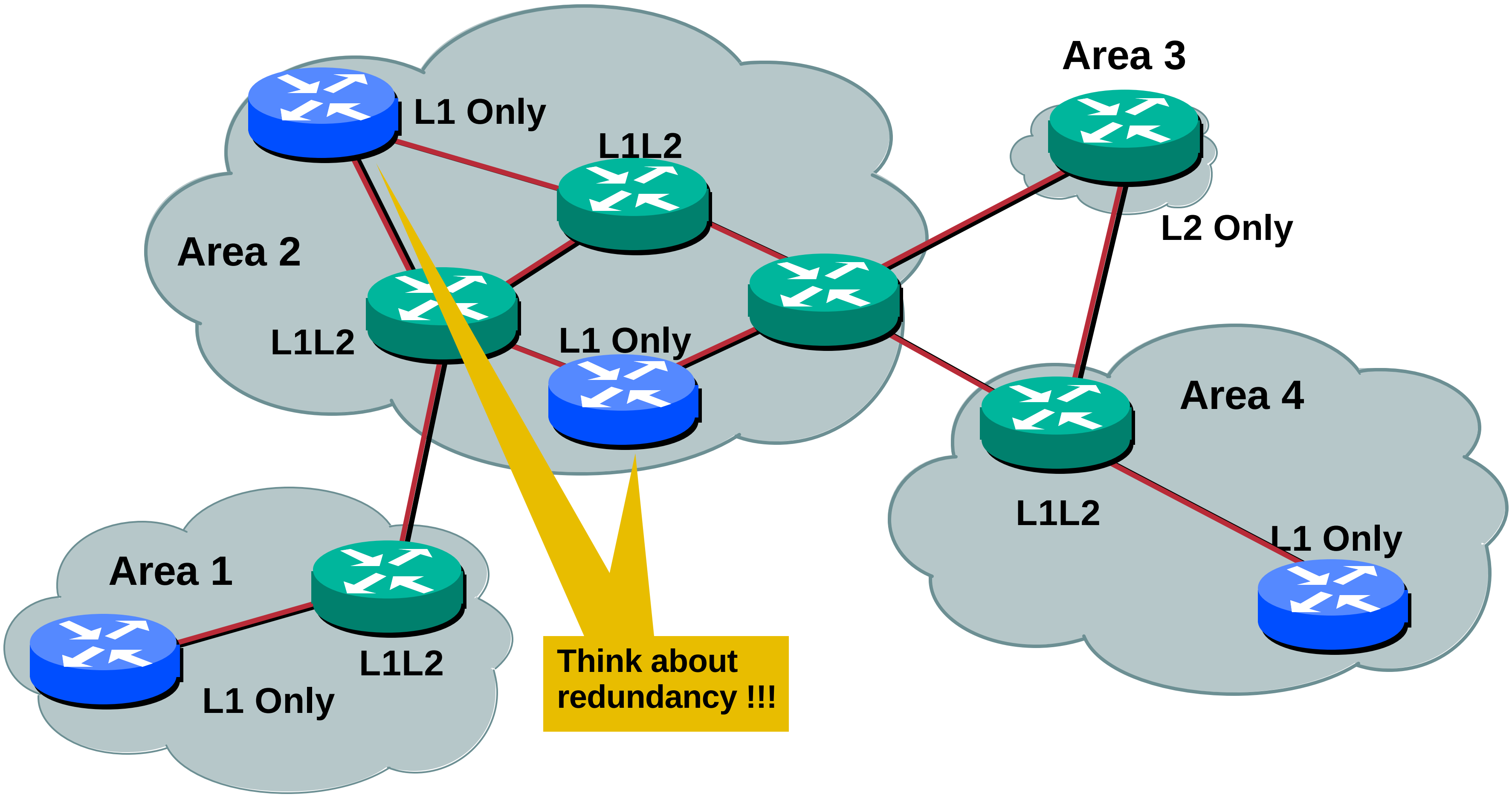
  can add level 1 areas as and when required

L1 or L2-only Networks: Adjacency Levels

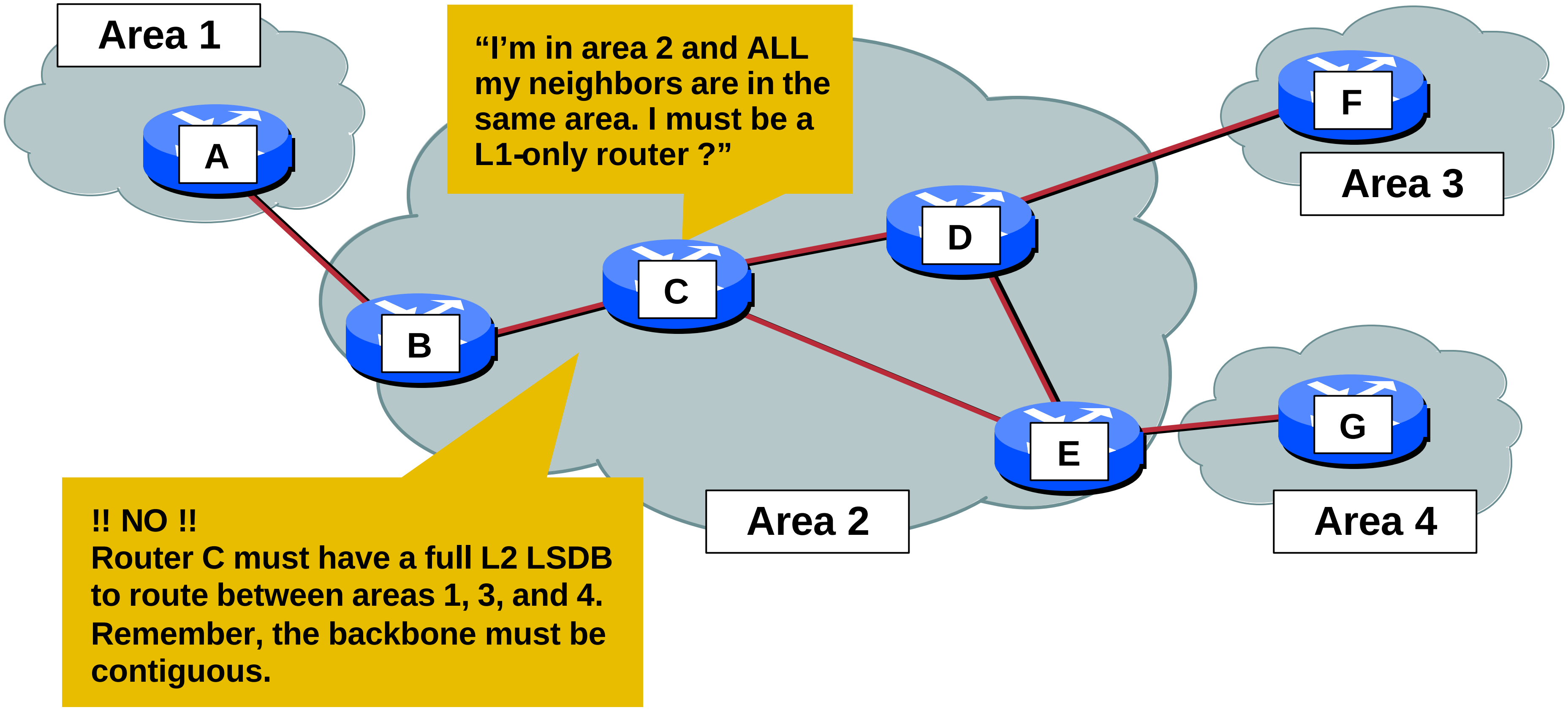


L1 or L2-only Networks: L2 Backbone

- Backbone **must be** L2 contiguous



L1 or L2-only Networks: L2 Backbone



Remember, the *Backbone Must Be Contiguous:*

An IS-IS router cannot determine if it is required to be L1, L2 or L1L2 - so all routers are configured as L1L2 by default

Hierarchical Networks – OSPF Areas

Cisco.com

- **Stub Areas**

  External routes are not permitted. Default route is injected into the Stub area as an inter-area route

  area <areaID> stub

- **Totally Stubby Areas**

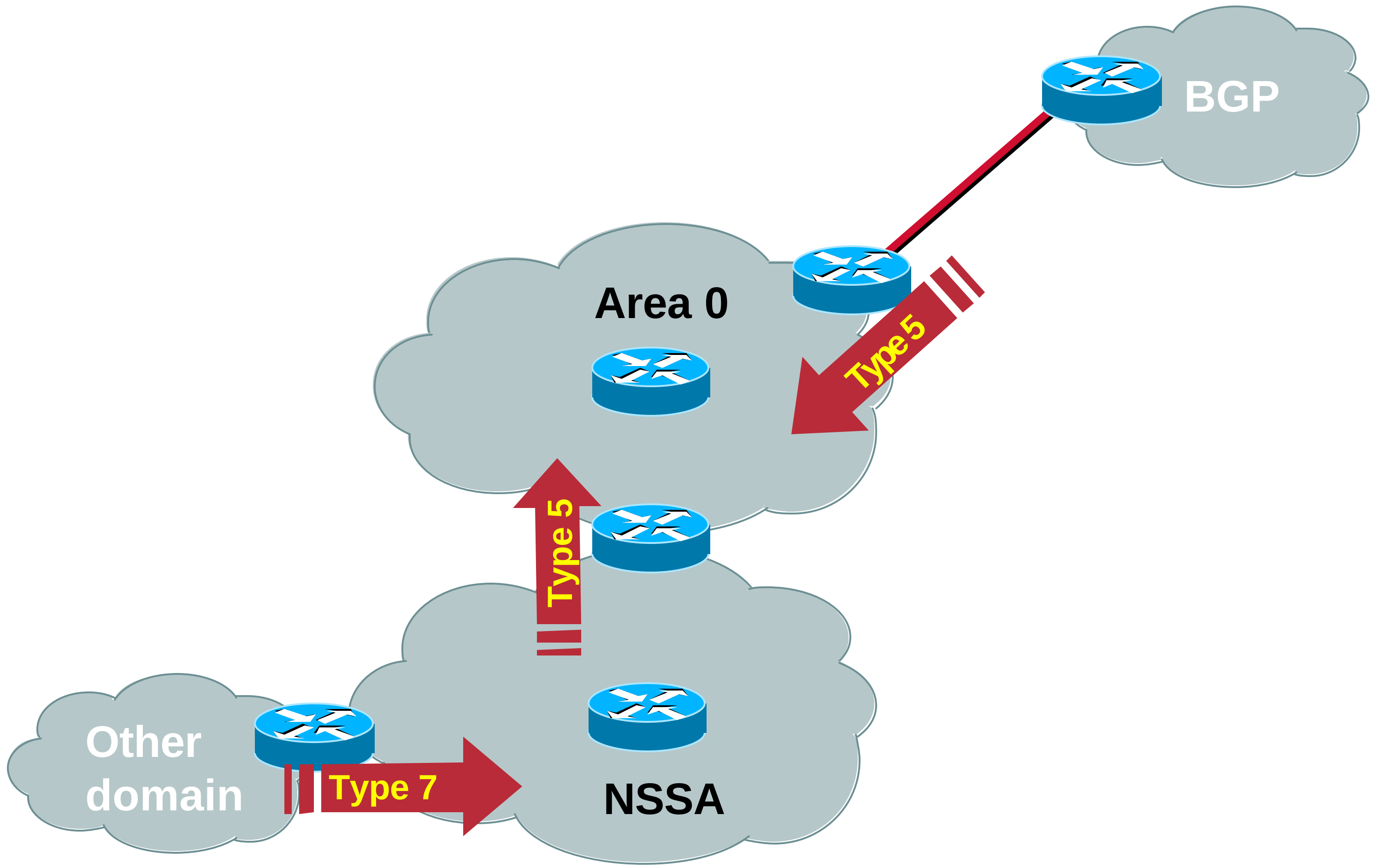
  Allows intra-area routes only. Default route injected into area as inter-area.

  area <areaID> stub no-summary

- **Not-So-Stubby Area**

  Same benefits of stub area, but an ASBR is allowed. Type 7 LSA flooded within the NSSA area. Converted into Type 5 LSA by the ABR when flooded into Area 0

NSSA Area - Example



Filtering LSAs in OSPF

- “Normal” OSPF areas receive all inter-area routes.

  The total number of routes may be reduced using summarization.

  To control the number of summary LSAs even further, filter in/out of an area specific Type 3 LSAs at ABR.

  area <area-id> filter-list prefix-list <list> <in|out>

Route Leaking - ISIS

Cisco.com

- **Allows L1L2 routers to insert in their L1 LSP IP prefixes learned from L2 database if also present in the routing table**
- **Works with new-style (wide) TLVs: draft-ietf-isis-traffic-04**

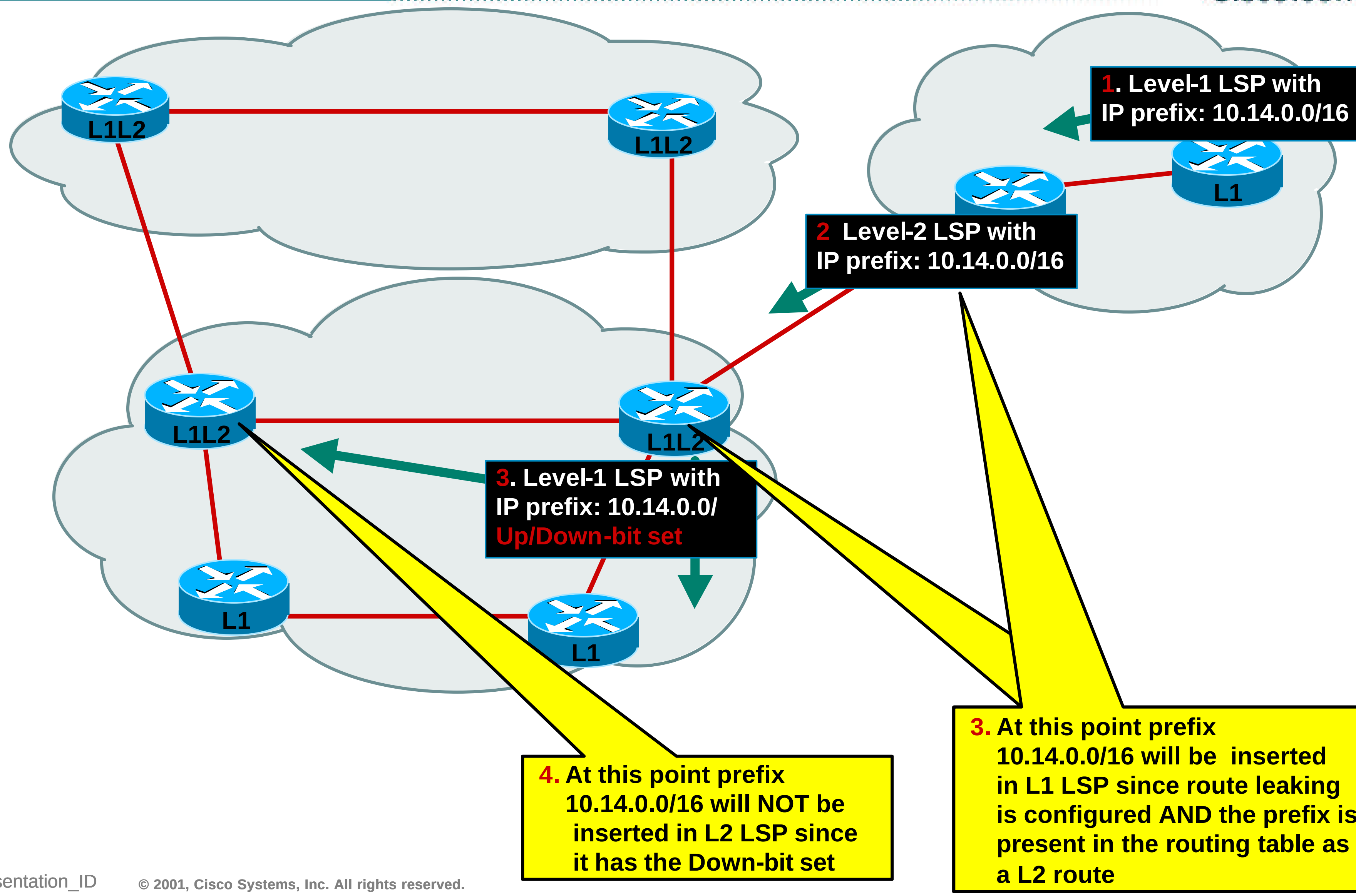
  Extended IP Reachability TLV (135)

- **Works with old-style (narrow) TLVs: RFC 2966**

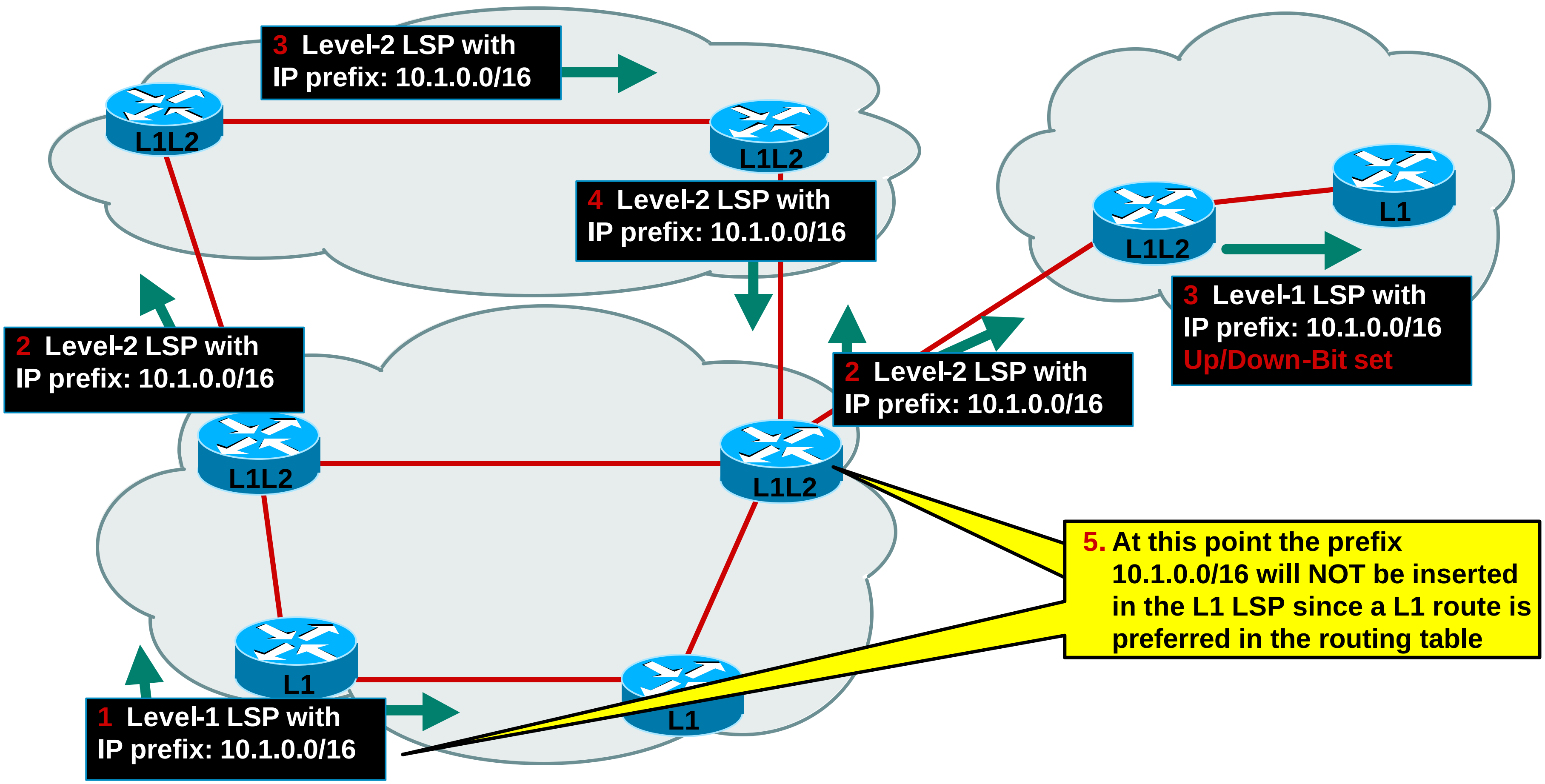
  IP Internal Reachability Information (TLV 128)

  IP External Reachability Information (TLV 130)

Route Leaking - ISIS



Route Leaking - ISIS



Route Leaking– ISIS

- **UP/Down bit used to prevent leaked routes being re-injected into the backbone**

  Extended IP Reachability TLV (135) contains Up/Down bit

  described in draft-ietf-isis-traffic-04

- **UP/Down bit is set each time a prefix is leaked into a lower level**

- **Prefixes with Up/Down bit set are NEVER propagated to an upper level**

Route Leaking – ISIS

- **TLVs 128 and 130 have a metric field that consists of 4 TOS metrics**

  The first metric, the so-called "default metric", has the high-order bit reserved (bit 8). Routers must set this bit to zero on transmission, and ignore it on receipt

- **The high-order bit in the default metric field in TLVs 128 and 130 becomes the Up/Down bit**

- **Recommendation:**
  Use Wide metrics (TLV 135)

  metric-style wide

Hierarchical Design Summary

- Use areas where necessary
- Summarize where ever possible
  - Match topology with addressing hierarchy
- Define routers to be in backbone first

Need for a Hierarchical Design?

- **Size of the network**

  Network has become too large. CPU, memory and link utilization requirements have increased

- **Stability of the network**

  Instability can be masked by incorporating areas. Problematic routers/links can be contained within an area so as not to disrupt the rest of the network.

Agenda – Link State Scalability

- Hierarchy
  - Use and limitations of Hierarchical Networks
  - Area types and flow of routing information
  - LSA Filtering/Route Leaking
- Detection and propagation of changes
  - Fast Hellos
  - LSA/LSP Generation
  - SPF Runs
  - Exponential Backoff
- Other tips...

Detection and Propagation: Convergence

Main Dependencies:

- Link failure detection
- Change propagation
- Initial wait for SPF computation
- Time to run SPF computation

Detection and Propagation: Failure Detection

- Router keeps track of the state of its interfaces

  looks at the physical state

- Layer 2 keepalives

  like HDLC or PPP or keep track of end-to-end state (ATM OAM)

- LAN's cannot detect all failure modes

  no indication of full LAN connectivity status

  use routing protocol itself – HELLO/IIH

Detection and Propagation: Failure Detection

- **ISIS**
  - **Each IIH carries a hold-time field**
  - **indicates time before removing adjacency**
  - **minimum holdtime is 1 second**
  - **By default routers limit IIHs to 1 per second**
- **OSPF**
  - **Lowest value for hello-interval is 1 second**
  - **Lowest value recommended for dead-interval is 3 seconds.**

Detection and Propagation: Failure Detection

- **When a link/interface goes down**
  - **layer 2 keepalives are lost**
  - **interface goes into shut or cable removed**
- **Don't have to wait for *hold-time* to *expire***
  - **interface is immediately removed from layer 2 adjacency table**
- **Adjacency is *immediately torn down***

Fast HELLOs

- **Advantages**

  reduced link failure detection time !

- **Disadvantages**

  increased BW/buffer/CPU usage can cause missed hellos

  potential increased adjacency flapping can cause instability

Fast HELLOs

- **ISIS**

  Minimum holdtime has now been reduced to 1 second (configurable)

    isis hello-interval minimal

  Advertised hold-time will now be 1 second; hello-interval will be 1 second divided by configured hello multiplier.

- **OSPF**

  Lowest value for hello-interval is 1 second

  Lowest value recommended for dead-interval is 3 seconds.

    ip ospf hello-interval

    ip ospf dead-interval

ISIS: HELLO Padding

- **Hellos are padded to full MTU size to aid in detecting MTU mismatch.**
  - **Inefficient use of bandwidth**
  - **May use significant number of buffers**
  - **Processing overhead when using authentication.**
- **Can be suppressed selectively:**
  - **no hello padding {multi-point|point-to-point}**
  - **no isis hello padding**

LSP Generation: What Triggers A New LSP ?

- When something changes …
  - Adjacency came up or went down
  - Interface up/down (connected IP prefix)
  - Redistributed IP routes change
  - Inter-area IP routes change
  - An interface is assigned a new metric
  - Most other configuration changes
  - Periodic refresh

LSP Generation: New LSP

- Create new LSP, install in your own LSPDB and mark it for flooding
- Send the new LSP to all neighbors
- Neigbors flood the LSP further

LSP Generation: Frequency Control

- LSP generation

  ISIS: lsp-gen-interval

  Controls the "frequency" of LSP generation

  Prevents from flapping links causing a lot of LSPs to be flooded throughout the network

  Default = 5 sec

  OSPF: LSA generation can't be throttled

SPF Runs: Algorithm and Complexity

- **Dijkstra's algorithm**
  - The goal is to find the topology in the form of a shortest path tree (SPT)
  - From the SPT we build routing tables
- **Depends on many factors**
  - in theory on # of routers, # of links
  - SPF complexity is O(n log n), where n is the number of routers.

SPF Runs: Theory verses Reality

- **CPU usage depends on other stuff**

  # links is important, also because of bi-directional check, # IP routes, stability of adjacencies, frequency of SPF, L1/L1L2, number of areas per ABR, etc.

- **Route installation is expensive**
- **Flooding is just as important**

  SPF is event driven and periodic

  flooding happens all the time

Exponential Backoff

- **Throttling of events may slow down convergence, while not throttling may cause melt downs.**
- **Exponential Backoff is a compromise:**

  The scope is to react fast to the first events, but under constant churn slow down to avoid a collapse.

Exponential Backoff

- **Backoff algorithm uses 3 timers**
  - **Maximum interval**

    Maximum amount of time the router will wait between consecutives executions
  - **Initial delay**

    Time the router will wait before starting execution
  - **Incremental interval**

    Time the router will wait between consecutive executions

    This timer is variable and will increase until it reaches *maximum-interval*

Exponential Backoff

- **Maximum-interval default values:**
  - **SPF -> ISIS: 10 sec; OSPF: 10 sec**
  - **PRC -> ISIS: 5 sec; OSPF: N/A**
  - **LSP-Generation -> ISIS: 5 sec; OSPF: N/A**
- **Initial-wait default values:**
  - **SPF -> ISIS: 5.5 sec; OSPF: 5 sec**
  - **PRC -> ISIS: 2 sec; OSPF: N/A**
  - **LSP-Generation -> ISIS: 50 ms; OSPF: 500 ms**

Exponential Backoff

- **Incremental-interval default values:**

  SPF -> ISIS: 5.5 sec; OSPF: 5 sec

  PRC -> ISIS: 5 sec; OSPF: N/A

  LSP-Generation -> ISIS: 5 sec; OSPF: N/A

Exponential Backoff

- **Extended syntax**

  ISIS: spf-interval <a> [<b> <c>]

  OSPF: timers throttle spf <b> <c> <a>

  <a> max time between SPF runs (seconds)

  <b> milliseconds between first trigger and SPF

  <c> milliseconds between first and second SPF

Exponential Backoff

- Example: spf-interval 10 100 1000

  (a) (b) (c)

- We decide to run SPF

  Wait 100 msecs, then run SPF (b=100 milliseconds)

  Wait at least 1 second before running a second SPF if needed (c = 1000 milliseconds)

  If we need to run a 3rd SPF, right after, wait at least 2 seconds (c = 2c)

  Wait at least 4 sec before next SPF, then 8 sec, then 10 sec, 10 sec, … (c= MIN(2c, a))

Exponential Backoff

- **When the network calms down, and there were no triggers for 2 times the minimum interval (20 sec in this example), go back to fast behaviour (100 ms initial wait)**

SPF Back-off Algorithm Behavior

- **Example: timer throttle spf 10 100 1000**

  X Y Z

- **Decide to run the first SPF:**

  SPF scheduled to run X msecs after first event

  Next SPF at Y msecs after first event

  if we need to run a third SPF, we wait (2 * value Y)

  – **Third run: (2*Y) = 200msec**

  – **Fourth run: (2*Y) = 400msec**

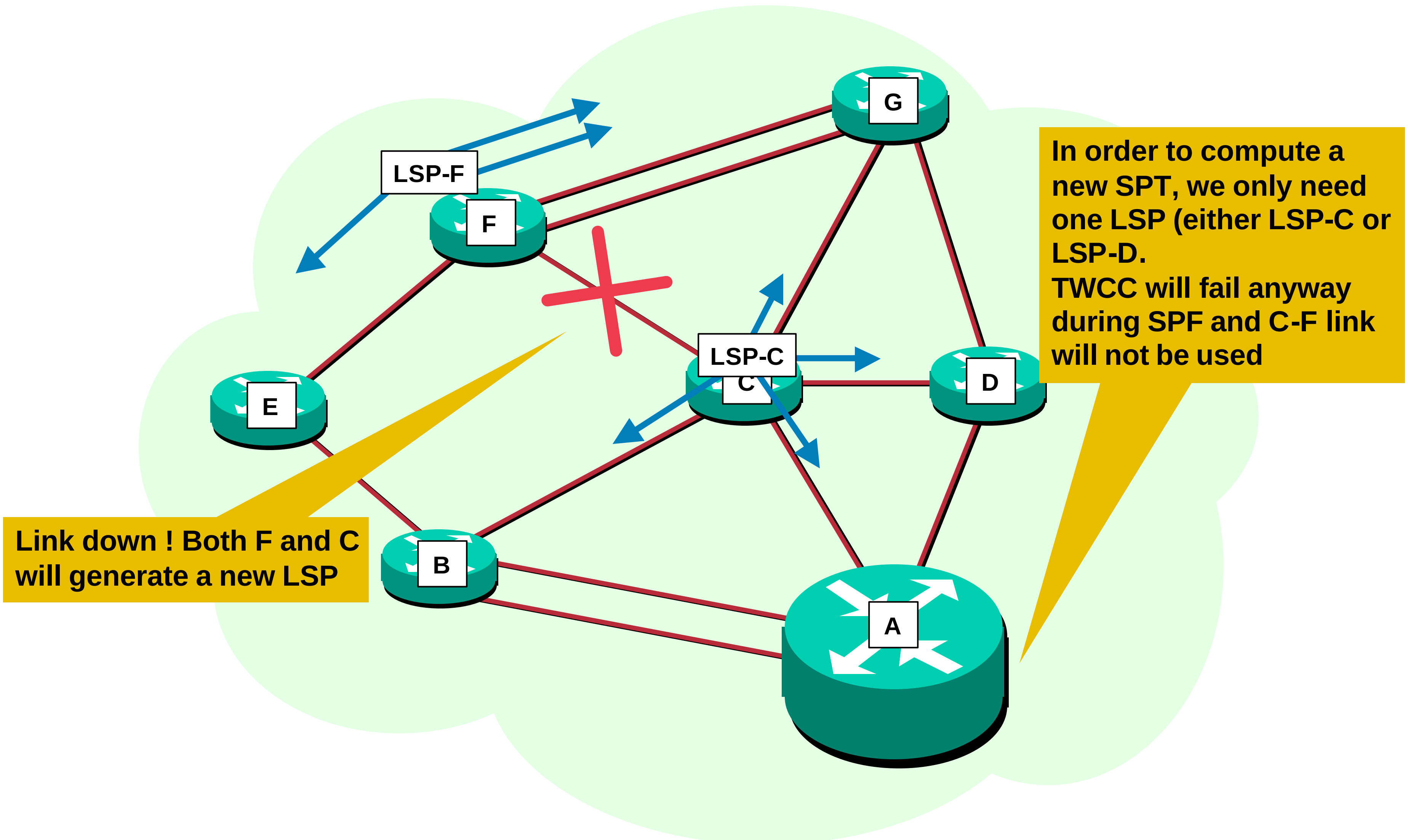
  – **Fifth run: (2*Y) = 800msec**

  Interval never to exceed Z msecs (maximum)

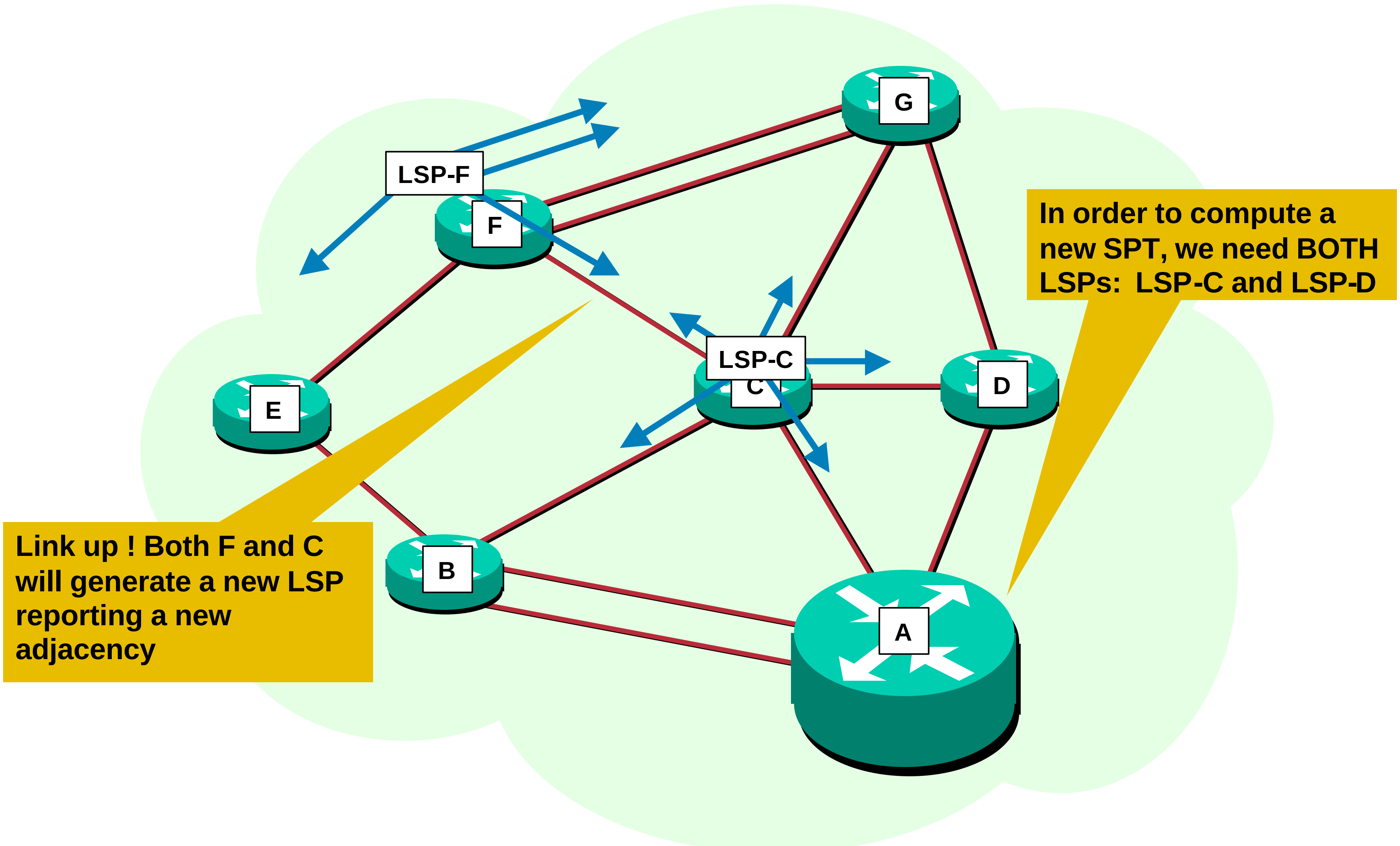
SPF Back-off Algorithm Behavior

- **When the network calms down, and there are no triggers for 2 times the maximum interval - value Z msec, we go back to fast behaviour (X msec initial wait)**
- **Old syntax still applies but can now configure the initial wait and a maximum wait interval**
- **Experience with the timers will show how the defaults can be tuned to more appropriate values. Each network is different!**
- **Use with *care* !**

Exponential Backoff: Configuration



Exponential Backoff: Configuration



Exponential Backoff: Configuration

- **Back-off Algorithm for SPF**

  timers throttle spf <spf-start><spf-hold><spf-max-wait>

  <spf-start> Delay between receiving a change to SPF calculation in milliseconds. Range 1-600000 milliseconds.

  <spf-hold> Delay between first and second SPF calculation in milliseconds. Range 1-600000 milliseconds.

  <spf-max-wait> Maximum wait time in milliseconds for SPF calculations. Range 1-600000 milliseconds.

Exponential Backoff: Configuration

- **Back-off Algorithm for LSA Gen**

timers lsa throttle <delay> <hold> <max-wait>

<delay> Delay in milliseconds between generating the first intra-area LSA. Range 1-5000 milliseconds.

<hold> Minimum delay in milliseconds while generating intra-area LSAs. Range 1-10000 milliseconds.

<max-wait> Maximum wait time in milliseconds while generating intra-area LSAs. Range 1-100000

Exponential Backoff: Configuration

- **spf-interval <a> <b> <c>**
- **<a> maximum SPF interval *(seconds)***

  This value is in seconds (backward compatibility). Use 1 second as long as your SPF doesn't take more than 1000 msecs (very large networks)

- **<b> initial wait *(milliseconds)***

  Give the router a chance to flood the LSP which triggered SPF before starting computation. Initial-wait can be any short value as long as the SPF computation doesn't take more than ~40 milliseconds

- **<c> incremental wait *(milliseconds)***

  This value will be doubled at each run anyway so you can start with small values. Start with number of milliseconds taken to complete the SPF computation

Exponential Backoff: Configuration

- prc-interval <a> <b> <c>
- <a> maximum PRC interval *(seconds)*

  This value is in seconds (backward compatibility). Use 1 second as the time to complete PRC will be short

- <b> initial wait *(milliseconds)*

  Allow the router time to flood the LSP which triggered PRC. The PRC will take just a few milliseconds and in most of the cases the delay will be insignificant (~20 IP prefixes processed per millisecond)

- <c> incremental wait *(milliseconds)*

  This value will be doubled at each run so start with small values. Start with number of milliseconds taken to complete the PRC computation

Exponential Backoff: Configuration

- **lsp-gen-interval <a> <b> <c>**
- **<a> maximum LSP generation interval *(seconds)***

  This value is in seconds (backward compatibility).
  Use 1 second as LSP generation will NEVER take this long
- **<b> initial wait**

  As soon as an event triggers a new LSP generation, you don't want to wait. So initial-wait has to be set to 1 (msec)
- **<c> incremental wait**

  Use a value that can rapidly be increased to a real wait value (i.e.: if you use 1, it will take 5 LSP generation before seeing an interval of 16 msecs)

Exponential Backoff: LSP Pacing

Cisco.com

- **Exponential Backoff should protect against constant LSP generations**
- **Therefore LSP pacing can be reduced in order to speed up end to end flooding**
- **Also, *Bad News* requires fewer number of LSPs in order to be processed. Therefore pacing has less impact on bad news**
- **Default 33msecs between successive LSPs**

  Reduce the pacing gap by using the <*lsp-interval*> interface configuration command (in msecs):

  lsp-interval 10

Controlling Background Flooding

- **Increase LSA refresh interval. Sets DNA bit on LSAs but does not suppress hellos. Receiving router does-not-age the received LSAs.**

  ip ospf flood-reduction

- **Adjust LSA group pacing**

  timers lsa-group-pacing *seconds*

  Created to control the synchronization of LSA check-summing, aging and refreshing processes.

  New format in 12.2 IOS

  timers pacing lsa-group

Throttling LSAs

- **LSA Flood Pacing**

  timers pacing flood

  Allows pacing of LSAs queued for flooding. Default is 33 milliseconds. Range is 5 to 100 milliseconds. Available in 12.2 IOS.

- **LSA Retransmission Pacing**

  timers pacing retransmission

  Allows pacing of LSAs queued for re-transmission. Default is 66 milliseconds. Range is 5 to 200 milliseconds. Available in 12.2 IOS.

Throttling LSAs – (cont.)

- **LSA Retransmission Timer**

  Ip ospf retransmit-interval

  Delay, in seconds, between retransmission of unacknowledged LSAs. Available since 10.0 IOS.

Agenda – Link State Scalability

Cisco.com

- Hierarchy
  - Use and limitations of Hierarchical Networks
  - Area types and flow of routing information
  - LSA Filtering/Route Leaking
- Detection and propagation of changes
  - Fast Hellos
  - LSA/LSP Generation
  - SPF Runs
  - Exponential Backoff
- **Other tips...**

Overload Bit

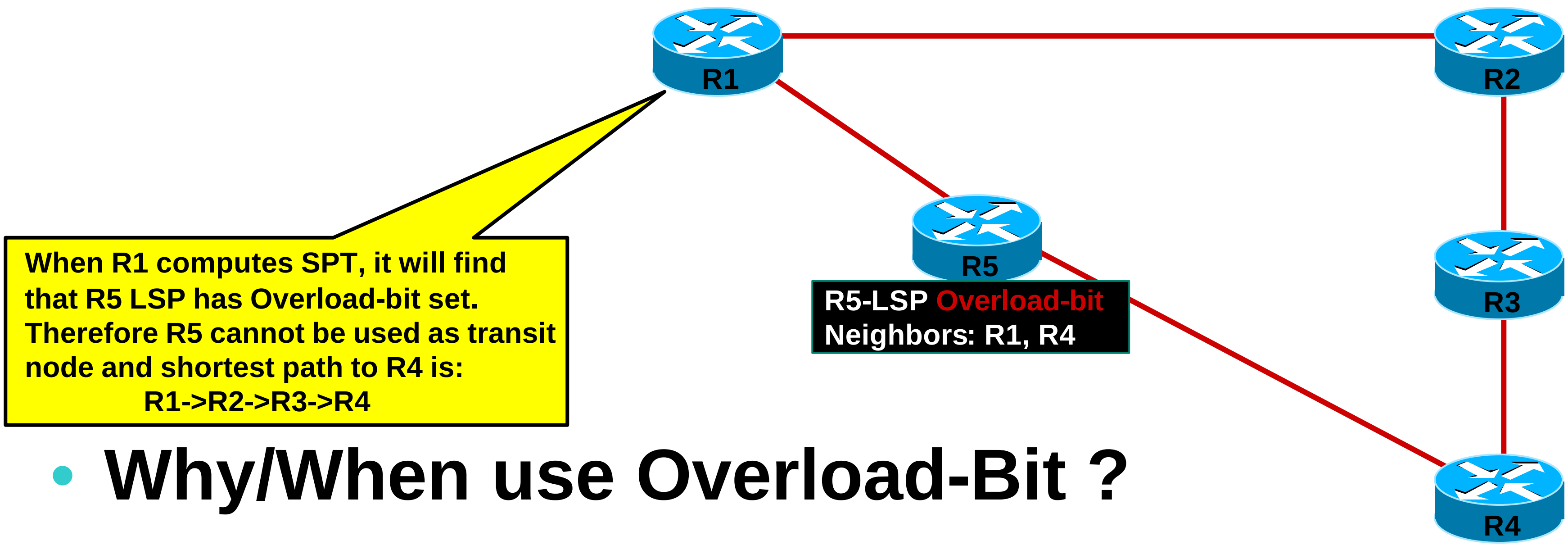
Cisco.com

- **ISO 10589 defines for each LSP a special bit called the LSPDB Overload (OL) Bit**
- **The Overload Bit may be set when a router experiences problems (such as a corrupt database)**
- **Once set, it will not be used for transit by other routers**
- **Connected IP prefixes still reachable**

Overload Bit

- IS-IS allows the manual setting of the Overload Bit
- This router will therefore never be used for transit, but it is still reachable
- Use for routers in the lab, routers aggregating management PVCs, etc

Overload Bit



- **Why/When use Overload-Bit ?**

  When the router is not ready to forward traffic for ALL destinations

  Typically when ISIS is up but BGP (or even MPLS) not yet

  When the router has other functions (Network Management)

Overload Bit

- **BGP will typically converge much slower than the IGP**
- **During this time, other routers in the AS will use this new router for transit**
- **But if the new router does not have all BGP routes yet, it will drop traffic**
- **New router should first converge BGP before carrying traffic**

Overload Bit

- **IS-IS can set the OL bit after each reboot, and allow BGP to converge before it advertises itself as transit by unsetting the OL bit**
- **Network admin needs to specify how long IS-IS should wait for BGP to converge**

  typically 2 to 5 minutes

Overload Bit

- **BGP can tell IS-IS to unset the Overload-bit immediately**
- **Default BGP update delay is 2 min**
- **When BGP never informs ISIS, the Overload-bit will be cleared after 10 minutes**

Overload Bit

Manually Setting Overload Bit

router isis

set-overload-bit

set-overload-bit on-startup <sec>

set-overload-bit on-startup wait-for-bgp

router bgp 100

bgp update-delay <sec>

Overload Bit

- **Overload-bit on-startup recommended in MPLS networks**
- **During boot-up a router may have all IGP routes but not all labels**
- **During this time it's better not to use the router as a transit point**

```
router isis
set-overload-bit on-startup 120
```

Fast Convergence at Adjacency Set-Up

- **Packets forwarded to a reloading router could be lost !**

    For instance, on a BGP border router, the IGP (OSPF) may converge faster than BGP. Traffic may be forwarded to the reloading router with no where to go.

Fast Convergence at Adjacency Set-Up

- **Recent code enables the reloading router to immediately flood its router-LSA**
- **All router link metrics within the router-LSA are set to infinity (0xffff) so it will NOT be used for transit**
- **LSA with "max-metric" set can be advertised for a specific amount of time or wait for BGP to signal it has converged.**

    max-metric router-lsa <on-startup {wait-for-bgp | <announce-time>}>

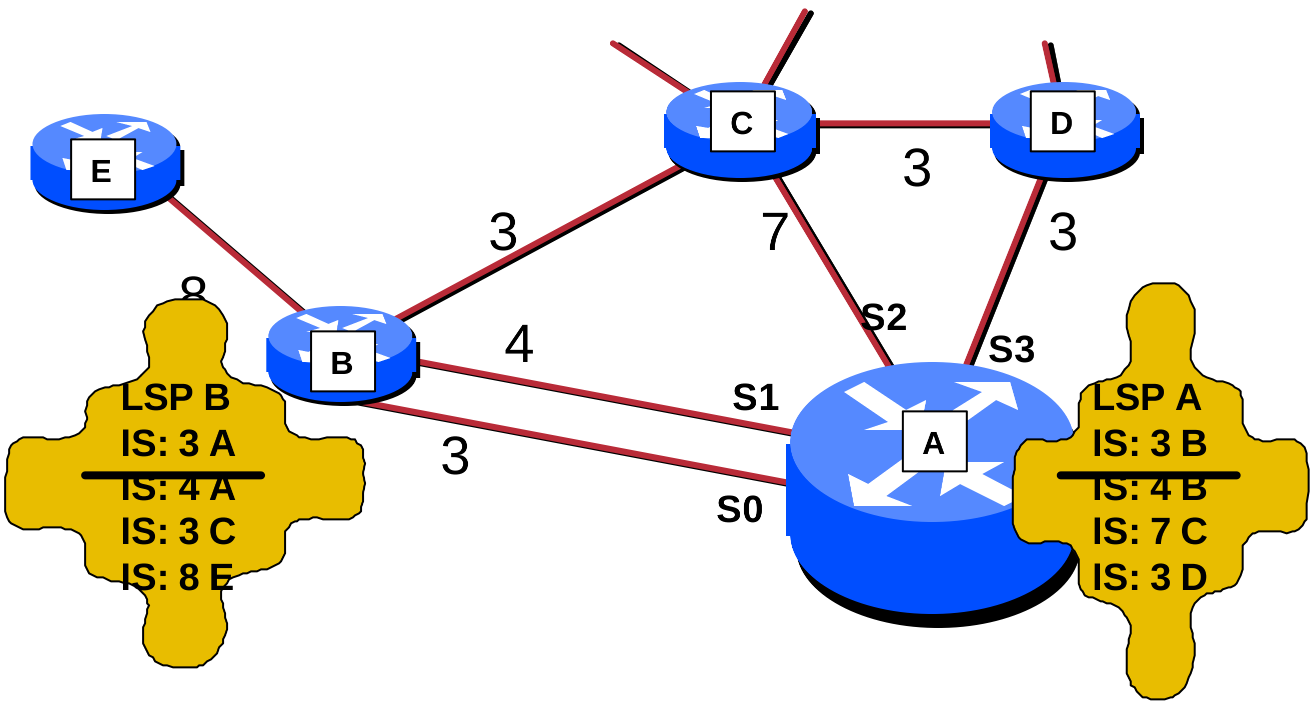
Other Tips: Parallel P2P Adjacencies

- When building an IS-IS LSP all adjacencies are inserted from the adjacency database
- Parallel adjacencies may therefore be included and advertised in the LSP
- Only need to advertise parallel point-to-point adjacencies once
- SPF uses only the best cost adjacency between two routers anyway

Other Tips: Parallel P2P Adjacencies

- **Number of advantages for not advertising parallel adjacencies**
  - **LSP's will be smaller and use less bandwidth when flooded**
  - **LSP's have lower chances of being fragmented**
  - **SPF calculations will be more efficient**
  - **Flapping of one of a set of parallel links will be invisible to the rest of the network**

Other Tips: Parallel P2P Adjacencies



- **Only the best parallel adjacency is reported**

Other Tips: Parallel P2P Adjacencies

- **With Traffic Engineering this is a problem:**
  - **All adjacencies need to be advertised with their own bandwidth characteristics**
  - **P2P optimisation is automatically turned off if TE is configured**

Other Tips: P2P Adjacencies Over a LAN

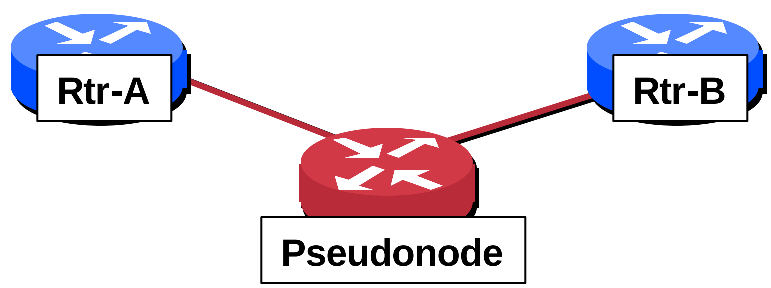
- When LAN interfaces (fast-ethernet, giga-ethernet, …) are used between two routers, tell ISIS to behave as p2p:
  - Avoid DIS election
  - Avoid CSNP transmissions
  - Reduce number of nodes in SPT (no pseudonode)
- New interface configuration command:

```
interface fastethernet1/0
 isis network point-to-point
```

Other Tips: P2P Adjacencies Over a LAN

LAN topology

SPT topology

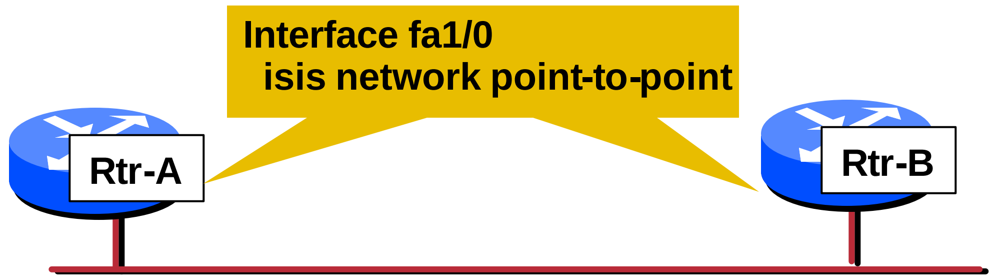


- SPF doesn't know anything about LANs
- All links are p2p
- Achieved by using Pseudonodes (same as OSPF type-2)

Other Tips: P2P Adjacencies Over a LAN

Cisco.com

LAN topology



SPT topology

- One step less in SPF computation
- No DIS election
- No CSNP flooding

Agenda

- Scope of the Presentation
- Scalability Building Blocks
  - Hierarchy
  - Redundancy
  - Addressing and Summarization
- Link State Scalability
  - ISIS Scalability
  - OSPF Scalability
- **BGP Scalability**

Agenda – BGP Scalability

Cisco.com

- iBGP Full Mesh: Route Propagation Requirements
- Peer-Groups: Configuration Grouping and UPDATE Generation
- Route Reflectors
  - Deployment (Hierarchy)
- Confederations
  - Deployment
  - Interaction with IGPs
- Detection and Propagation of Changes
  - minRouteAdvertisementInterval
  - NEXT_HOP Reachability
  - Route Dampening

Agenda – BGP Scalability

iBGP Full Mesh: Route Propagation Requirements

Peer-Groups: Configuration Grouping and UPDATE Generation

Route Reflectors

- Deployment (Hierarchy)

Confederations

- Deployment
- Interaction with IGPs

Detection and Propagation of Changes

- minRouteAdvertisementInterval
- NEXT_HOP Reachability
- Route Dampening

iBGP Full Mesh

"When a BGP speaker receives an UPDATE message from an internal peer, the receiving BGP speaker shall not re-distribute the routing information contained in that UPDATE message to other internal peers..."

draft-ietf-idr-bgp4-13

Section 9.2.1

iBGP Full Mesh

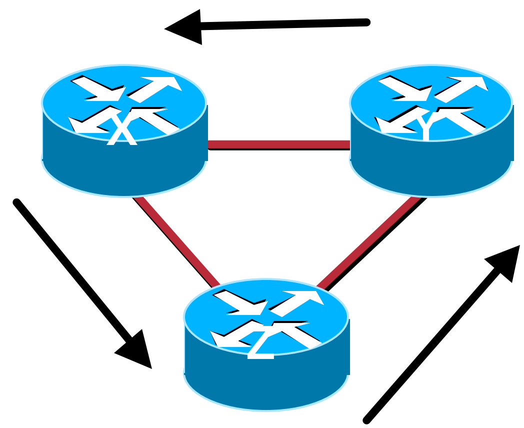
- **Why have a restriction?**

  No mechanism to detect an UPDATE loop exists in iBGP.

- **What may be the consequences of not having a full iBGP mesh?**

  Black holes and routing loops.

  UPDATE loops.



- **HINT:** ***Only the border routers (or the originators of routing information) MUST maintain a session with all the other routers in the AS.***

iBGP Full Mesh

Scalability Concerns

- **Administration**

  Configuration Management on increasingly large number of routers.

- **Number of TCP Sessions**

  Total number of sessions = n(n-1)/2

  Maintaining extreme numbers of TCP sessions creates extra overhead.

- **BGP Table Size**

  A higher number of neighbors generally translates to a higher number of paths for each route.

  Memory consumption.

Agenda – BGP Scalability

Cisco.com

iBGP Full Mesh: Route Propagation Requirements

Peer-Groups: Configuration Grouping and UPDATE Generation

Route Reflectors

- Deployment (Hierarchy)

Confederations

- Deployment
- Interaction with IGPs

Detection and Propagation of Changes

- minRouteAdvertisementInterval
- NEXT_HOP Reachability
- Route Dampening

Peer-groups

- **Peer-groups address two scalability issues**
  - **Configuration size**
  - **UPDATE replication/advertisement**
- **A "peer-group" is a configuration tool that is used to apply the same commands to multiple peers without explicitly configuring those commands for each peer.**
- **Members of a peer-group will receive the same BGP UPDATEs. As a result, all members of a peer-group must have the same outbound policy.**

Peer-groups

Configuration Example

```
neighbor 1.1.1.1 remote-as 100
neighbor 1.1.1.1 update-source
   Loopback 0
neighbor 1.1.1.1 send-community
neighbor 1.1.1.1 version 4
neighbor 1.1.1.2 remote-as 100
neighbor 1.1.1.2 update-source
   Loopback 0
neighbor 1.1.1.2 send-community
neighbor 1.1.1.2 version 4
```

```
! Define the peer-group
neighbor iBGP peer-group
neighbor iBGP remote-as 100
neighbor iBGP update-source
   Loopback 0
neighbor iBGP send-community
neighbor iBGP version 4
! Assign peers to the peer-group
neighbor 1.1.1.1 peer-group iBGP
neighbor 1.1.1.2 peer-group iBGP
```

BEFORE → AFTER

Peer-groups

Application Rules

- **All members MUST share a common outbound policy.**

  The same UPDATE message is sent to all the peers.

- **Examples:**

  RR-clients, but not a mixture of clients and iBGP peers

  iBGP OR eBGP peers, but not both in the same peer-group

  NEXT_HOP is an exception to the rule. Peers A and B can be in a peer-group and receive a different NEXT_HOP for an UPDATE. Accomplished by doing the *NEXT_HOP re-write* at the last minute

Peer-groups

Cisco.com

- Three common eBGP peer-groups
  - Advertise default-route only
  - Advertise customer routes
  - Advertise full routes
- All should filter bogus inbound information
  - Address space that you use in your IGP!!
  - RFC 1918 address space
  - Class D and E addresses
  - Prefixes that are too specific (Class A /32s for example)
  - Un-assigned Class A, B, and C address space (optional)
  - "max-prefix" can be used for additional protection

Peer-groups


```
neighbor eBGP-default peer-group
neighbor eBGP-default route-map bogus_filter in
neighbor eBGP-default route-map default_only out
neighbor eBGP-default version 4
!
neighbor eBGP-customer peer-group
neighbor eBGP-customer route-map bogus_filter in
neighbor eBGP-customer route-map customer_routes out
neighbor eBGP-customer version 4
!
neighbor eBGP-full peer-group
neighbor eBGP-full route-map bogus_filter in
neighbor eBGP-full route-map full_routes out
neighbor eBGP-full version 4
```

Peer-groups

- **Problem: Advertise 100,000+ routes to hundreds of peers. BGP will need to send a few hundred megs of data in order to converge all peers.**
- **Solution: Use peer-groups!**

  UPDATE generation is done once per peer-group.

  The UPDATEs are then replicated for all peer-group member.

- ***Scalability is enhanced because more peers can be supported!***

Peer-groups

- **UPDATE generation without peer-groups**

  The BGP table is walked once, prefixes are filtered through outbound policies, UPDATEs are generated and sent…per peer!!

- **UPDATE generation with peer-groups**

  A peer-group *leader* is elected for each peer-group. The BGP table is walked once (for the leader only), prefixes are filtered through outbound policies, UPDATEs are generated and sent to the peer-group leader and replicated for peer-group members that are *synchronized* with the leader.

  Replicating an UPDATE is much easier/faster than formatting an UPDATE. Formatting requires a table walk and policy evaluation, replication does not.

Peer-groups

Synchronization

- **A peer-group member is *synchronized* with the leader if all UPDATEs sent to the leader have also been sent to the peer-group member**

  The more peer-group members stay in sync the more UPDATEs BGP can replicate.

- **A peer-group member can fall out of sync for several reasons**

  Slow TCP throughput

  Rush of TCP Acks fill input queues resulting in drops

  Peer is busy doing other tasks

  Peer has a slower CPU than the peer-group leader

Peer-groups

Synchronization

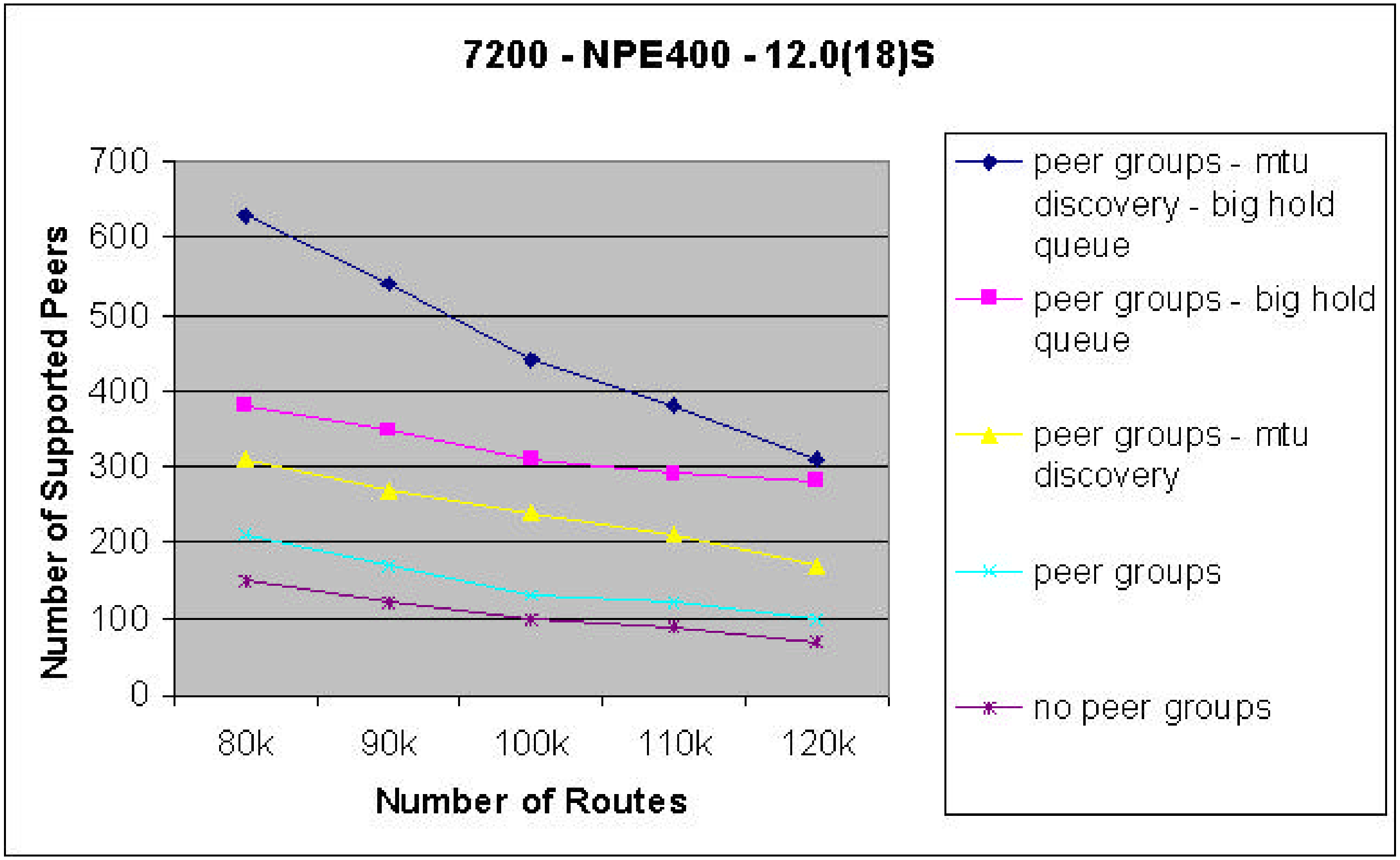
- **TCP throughput can be increased by reducing TCP overhead.**

  ip tcp path-mtu-discovery **allows TCP to use an optimal Max Segment Size (MSS – default = 536 bytes). The MSS will be based on the smallest MTU of the links between the two peers**

- **Advertising UPDATEs to many peers in a short period of time can induce a rush of TCP acknowledgements.**

  These Acks are destined for the router and can fill process level input queues. Increasing these queue depths (*hold-queue 1000 in*) can reduce the number of dropped TCP Acks

Peer-groups



- ***Using peer-groups, "ip tcp path-mtu-discovery" and larger input queues together will improve BGP scalability by reducing convergence times!***

Peer-groups – Summary

- **Peer-group scalability benefits:**

  UPDATE Generation and Replication

  Configuration Grouping

Agenda – BGP Scalability

- iBGP Full Mesh: Route Propagation Requirements
- Peer-Groups: Configuration Grouping and UPDATE Generation
- **Route Reflectors**
  - **Deployment (Hierarchy)**
- Confederations
  - Deployment
  - Interaction with IGPs
- Detection and Propagation of Changes
  - minRouteAdvertisementInterval
  - NEXT_HOP Reachability
  - Route Dampening

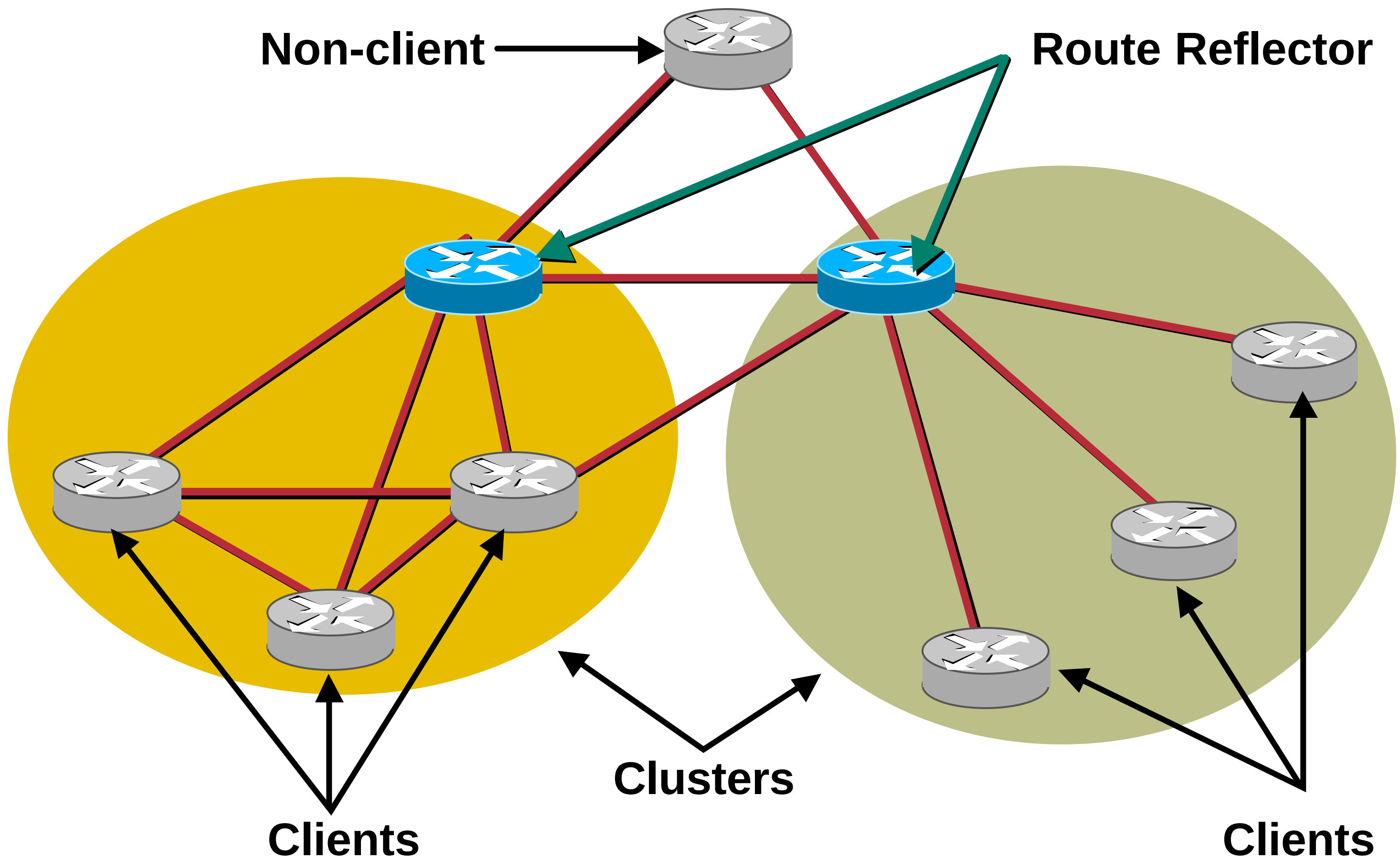
Route Reflectors

- **Defined in rfc2796.**
- **Allows a router (route reflector – RR) to advertise routes received from an iBGP peer to other iBGP peers.**

    Between clients and from clients to non-clients, and vice versa.

- **The ORIGINATOR_ID and CLUSTER_LIST attributes are used to perform loop detection.**
- **Provides a scalable alternative to an iBGP full mesh.**

Route Reflectors - Terminology



Lines Represent Both Physical Links and BGP Logical Connections

Route Reflectors

Reflection Decisions

- **Only the best path is propagated.**
  - From an eBGP peer, send the path to everyone
  - From a RRC, reflect the path to clients and non-clients, send the path to eBGP peers
  - From a regular iBGP peer (non-client), reflect the path to RRCs and send the path to eBGP peers
- **When a route is reflected the RR appends its BGP_ID (or configured *bgp cluster-id*)to the CLUSTER_LIST.**

Route Reflectors - Redundancy

- **A RRC may peer with more than one reflector, in different clusters.**

  A RRC that peers to only one RR has a single point of failure

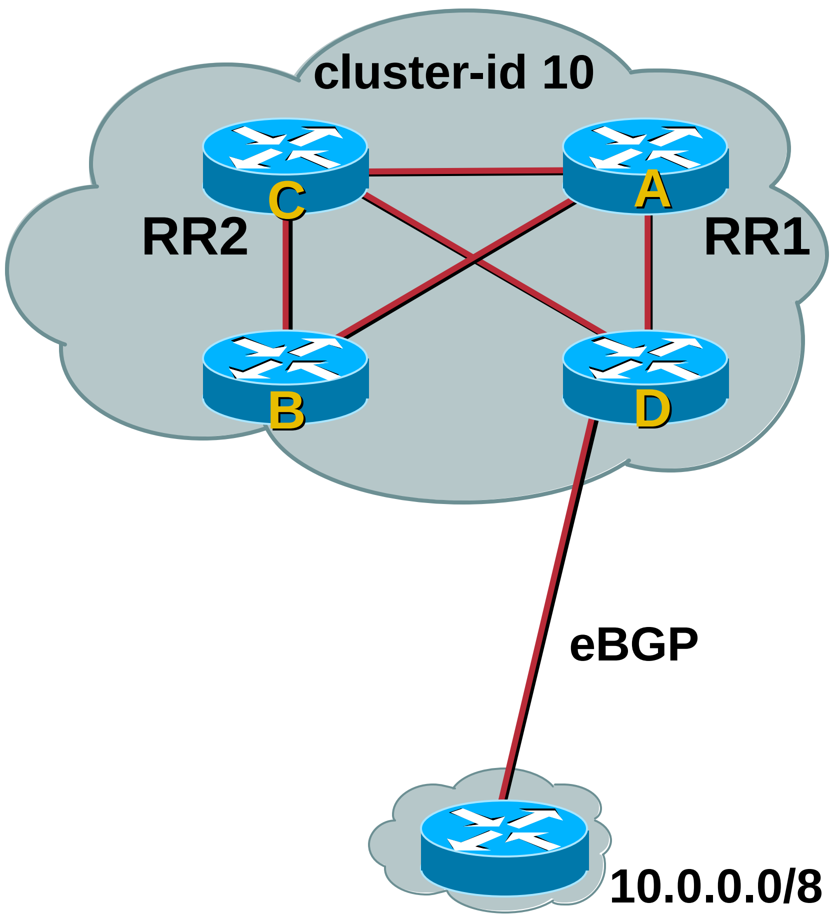
  RRC should peer to at least two RRs to provide redundancy

- **The million dollar question**

  Should redundant RRs be in the same cluster or should they be in separate clusters?

Route Reflectors - Redundancy

- RRs A and C have the same Cluster-ID
- C will deny routes reflected from A due to cluster-list loop detection
- If session from C to D fails, C will not be able to reach 10.0.0.0/8
- If session from B to A fails, B will not be able to reach 10.0.0.0/8
- D has some redundancy, but not 100%

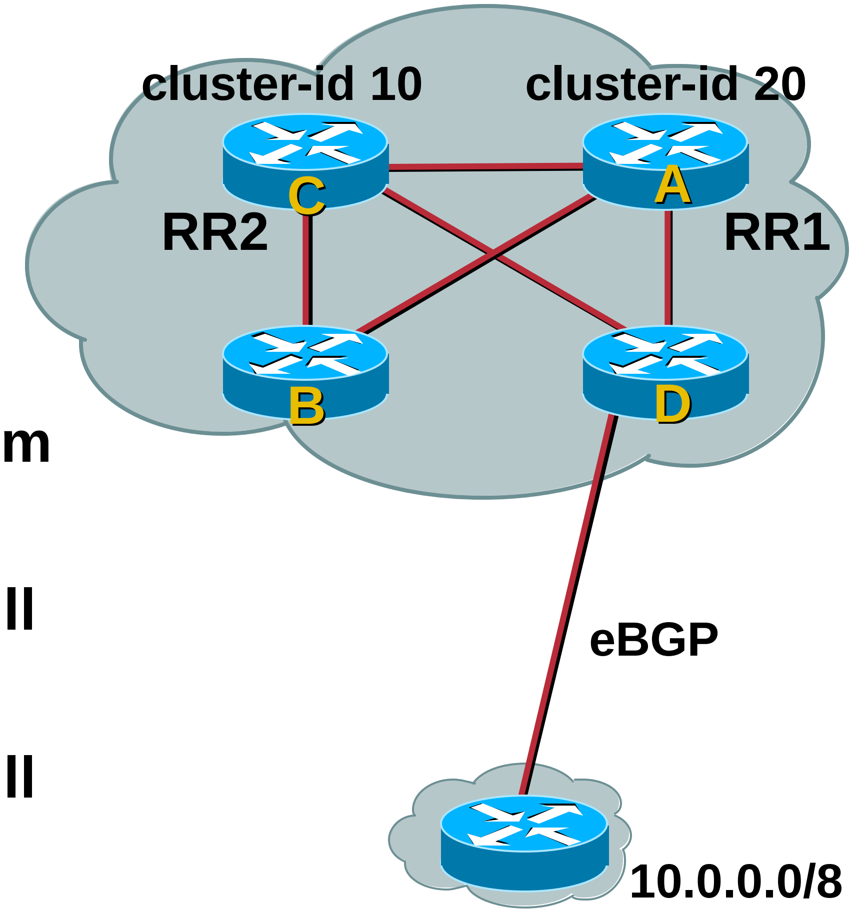


Lines Represent Both Physical Links and BGP Logical Connections

Route Reflectors - Redundancy

Cisco.com

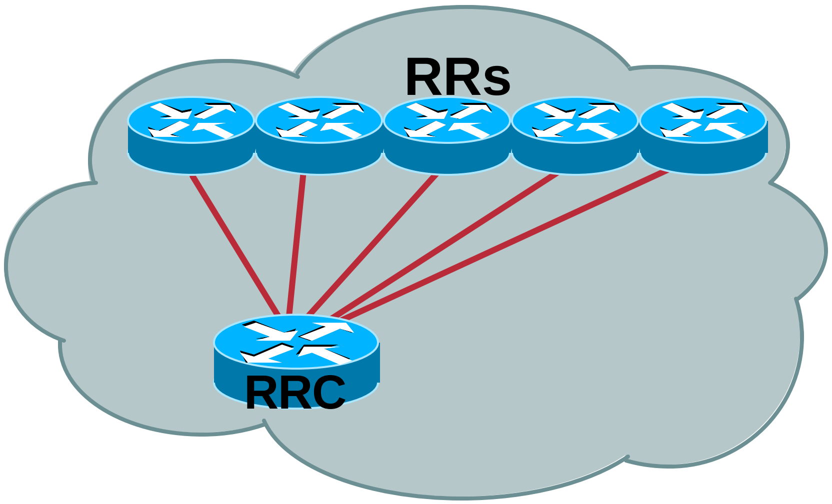
- RRs A and C have different Cluster-IDs
- C will not deny routes reflected from A
- C will know about 10.0.0.0/8 from A and D
- If C to D session fails, C can still reach 10.0.0.0/8 via A
- If B to A session fails, B can still reach 10.0.0.0/8 via C
- D has true redundancy



Lines Represent Both Physical Links and BGP Logical Connections

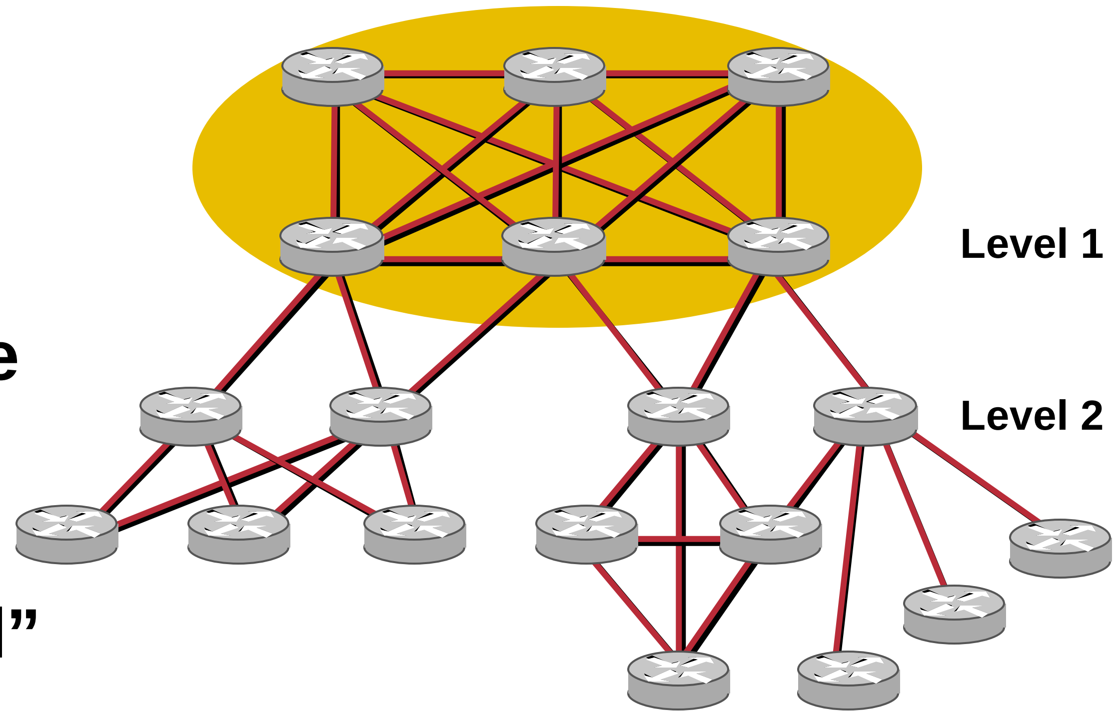
Route Reflectors - Redundancy

- **Can a RRC have too much redundancy?**
- **RRC will receive an additional view for each extra RR it peers with, which will consume extra memory.**
- ***Redundancy is a good thing, but too much redundancy can cost memory without adding significant benefit.***



Route Reflectors - Hierarchy

- **Clusters may be configured hierarchically**
- **RRs in a cluster are clients of RRs in a higher level**
- **Provides a "natural" method to limit routing information sent to lower levels**



Route Reflector - Deployment

1. **Divide the network into multiple clusters**
2. **Each cluster contains at least one RR.**

   Clients can peer with RRs in other clusters for redundancy.
3. **Top Level RRs are fully meshed via iBGP.**
4. **Still use single IGP — NEXT_HOP unmodified by RR unless via explicit route-map.**

Follow the physical topology!!

Route Reflectors – Summary

- **Hierarchical Deployment**

  Follow Physical Topology!!

- **Per-client Redundancy**

  Clients may peer to RRs in different clusters.

  Each additional RR supplies an extra route view to the client.

- **RRs provide a “natural” way to aggregate the amount of routing information sent to the clients.**

  Only the best path is propagated.

Agenda – BGP Scalability

- iBGP Full Mesh: Route Propagation Requirements
- Peer-Groups: Configuration Grouping and UPDATE Generation
- Route Reflectors
  - Deployment (Hierarchy)
- **Confederations**
  - **Deployment**
  - **Interaction with IGPs**
- Detection and Propagation of Changes
  - minRouteAdvertisementInterval
  - NEXT_HOP Reachability
  - Route Dampening

Confederations

- **Described in rfc3065.**
- **An AS is split into multiple Sub-Ases; still looks like a single AS to eBGP peers.**

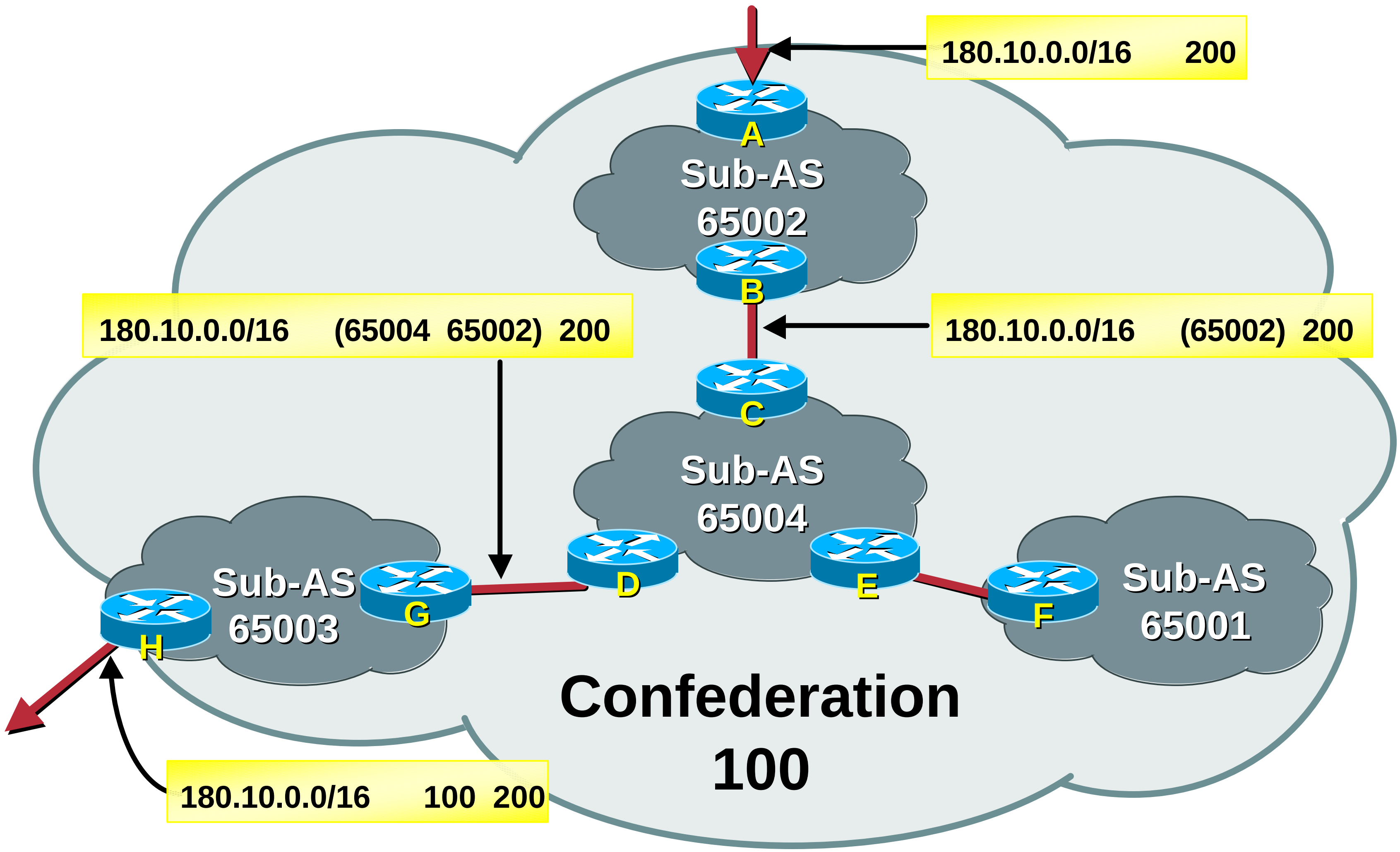
  Sub-AS numbers should come from private AS range

- **BGP sessions between each Sub-AS is similar to eBGP**

  Preserve NEXT_HOP, LOCAL_PREF and MED.

  AS_CONFED_SEQUENCE is used to perform loop detection.

Confederations – AS_CONFED_SEQ



Confederations – Deployment

- **No graceful way to migrate an existing network from a full mesh to confederations.**
- **Easy to define policies per Sub-AS.**

  “independent sub-AS administration”

- **NEXT_HOP can be reset when advertising routes from one Sub-AS to another**

  Makes it possible to run a separate IGP per Sub-AS!!

- **Provides quick and dirty method of integrating a network into an existing one.**

Confederations – Summary

- *Simplify* the network topology.

  Allow contained hierarchy per sub-AS.

- Policy may be defined per sub-AS

  Ease of network integration.

- Migration to/from confederations is not straight forward.

RRs or Confederations

	External Connectivity	Multi-Level Hierarchy	Policy Control	Scalability	Migration Complexity
Confederations	Anywhere In the Network	Yes	Yes	Medium	Medium To High
Route Reflectors	Anywhere In the Network	Yes	Yes	Very High	Very Low

Agenda – BGP Scalability

- iBGP Full Mesh: Route Propagation Requirements
- Peer-Groups: Configuration Grouping and UPDATE Generation
- Route Reflectors
  - Deployment (Hierarchy)
- Confederations
  - Deployment
  - Interaction with IGPs
- **Detection and Propagation of Changes**
  - **minRouteAdvertisementInterval**
  - **NEXT_HOP Reachability**
  - **Route Dampening**

minRouteAdvertisementInterval

"MinRouteAdvertisementInterval determines the minimum amount of time that must elapse between advertisement of routes to a particular destination from a single BGP speaker."

draft-ietf-idr-bgp4-13

Section 9.2.3.1

minRouteAdvertisementInterval

- **Studies* have been made to study the effects of the minRouteAdvertisementInterval on BGP convergence**
- **In a nutshell**

  Keeping the timer per peer instead of per prefix has some negative effects

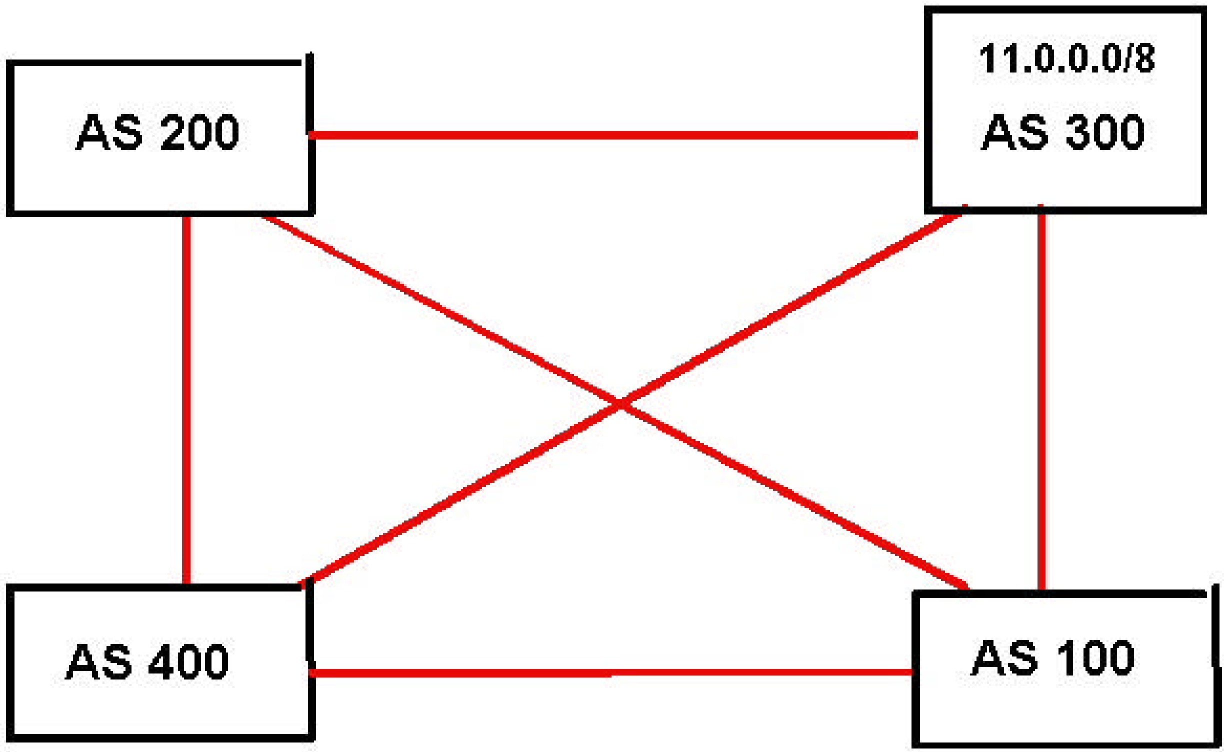
  The default MinAdvInterval of 30 seconds may be too long

  TX Loop Detection should be implemented

  using an outbound filter to prevent advertising routes to a peer that will deny them due to AS_PATH loop detection

"An Experimental Study of Internet Routing Convergence"
- Labovitz, Ahuja, Bose, Jahanian

minRouteAdvertisementInterval



- **Topology used to perform internal testing to study the effects when flapping the 11.0.0.0/8 prefix.**

minRouteAdvertisementInterval -- Test Matrix

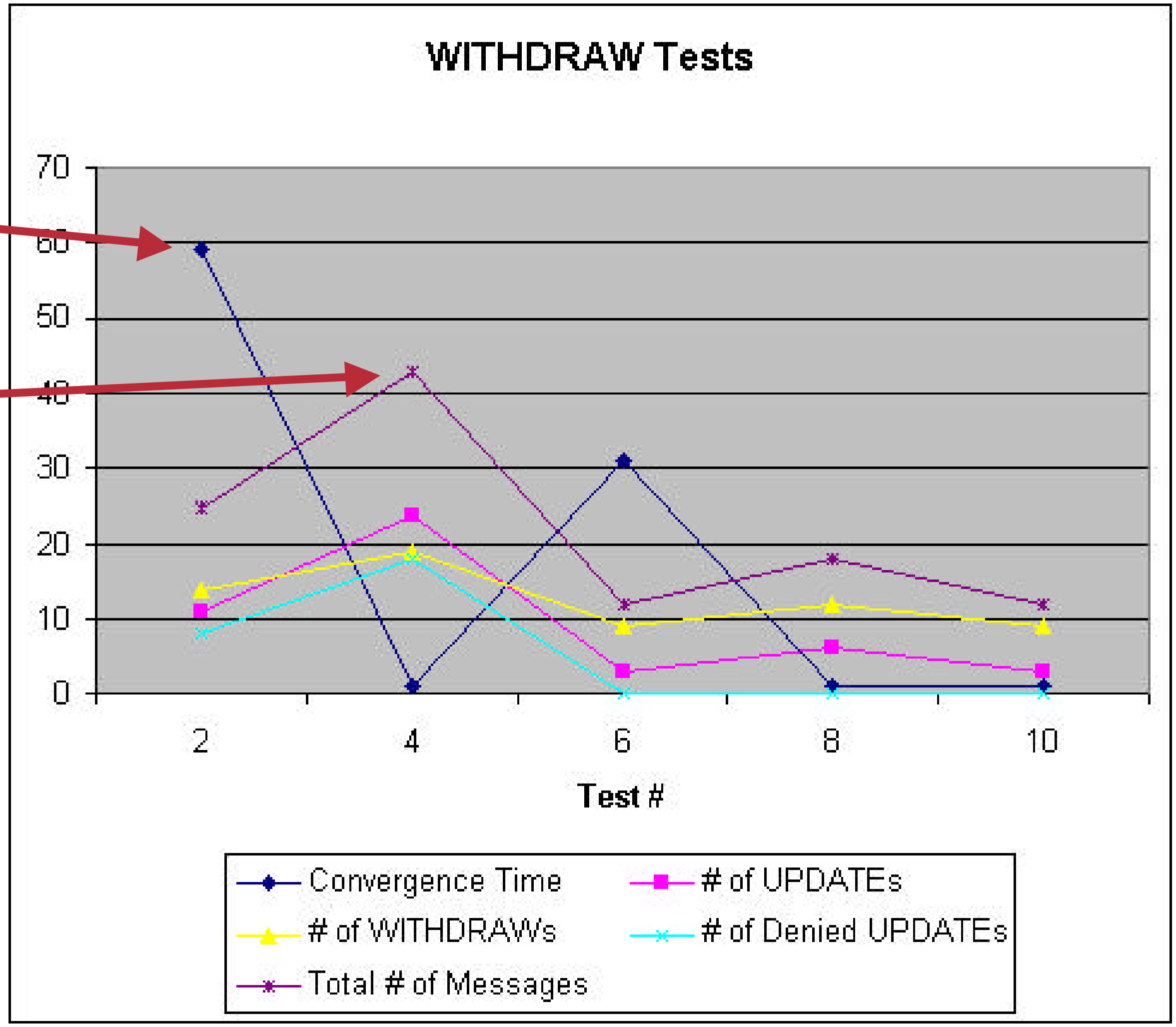
Cisco.com

	Message Type.	Timer (sec)	TX Loop Detection	# Msgs Total	Denied UPDATES	Conv. (sec)
Test 1	UPDATE	30		9		< 1
Test 2	WITHDRAW	30		25	8	59
Test 3	UPDATE	0		9		< 1
Test 4	WITHDRAW	0		43	18	< 1
Test 5	UPDATE	30	X	9		< 1
Test 6	WITHDRAW	30	X	12		31
Test 7	UPDATE	0	X	9		< 1
Test 8	WITHDRAW	0	X	18		< 1
Test 9	UPDATE	1	X	9		< 1
Test 10	WITHDRAW	1	X	12		< 1

minRouteAdvertisementInterval - Conclusions

Cisco.com

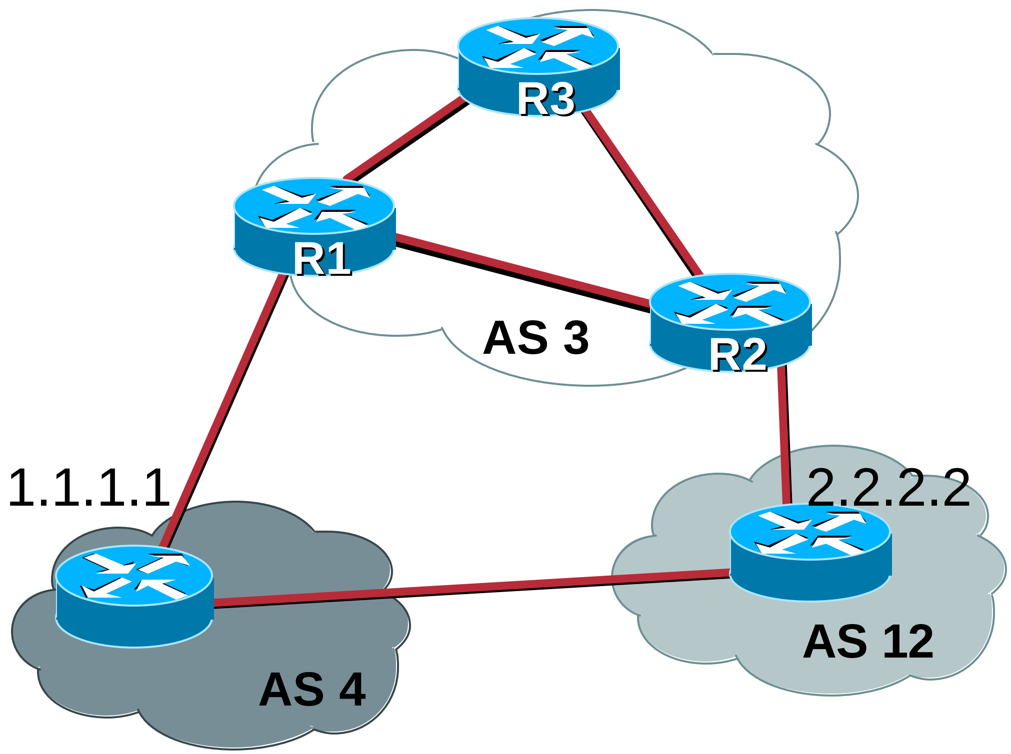
- **Default behavior takes almost 1 minute to converge**
- **Using a MinAdvInterval of 0 results results in a flurry of messages (43) for a single route-flap (see Test 4)**
- **Using TX Loop Detection reduces the number of messages sent (see Tests 6, 8, and 10)**
- **Best results are in test 10 which uses TX Loop Detection with Min Adv Interval of 1 second**



NEXT_HOP Reachability

- **The NEXT_HOP MUST be reachable for the BGP path to be valid.**
  - **Reachability should be provided by the IGP.**
- **Other route characteristics also important for best path selection**
  - **IGP metric to NEXT_HOP**
- **Change in the reachability characteristics of the NEXT_HOP (availability, cost) may impair the ability to forward traffic and/or cause black holes or routing loops.**
  - **BGP depends on the underlying IGP to provide fast and consistent notification of any change**

NEXT_HOP Reachability



- R1 and R2 advertise routes to R3 with NEXT_HOPs of 1.1.1.1 and 2.2.2.2
- R3 must have a route to these two addresses
- Black Holes and severe route flapping can occur if R3 does not have a proper route to both NEXT_HOPs

NEXT_HOP Reachability

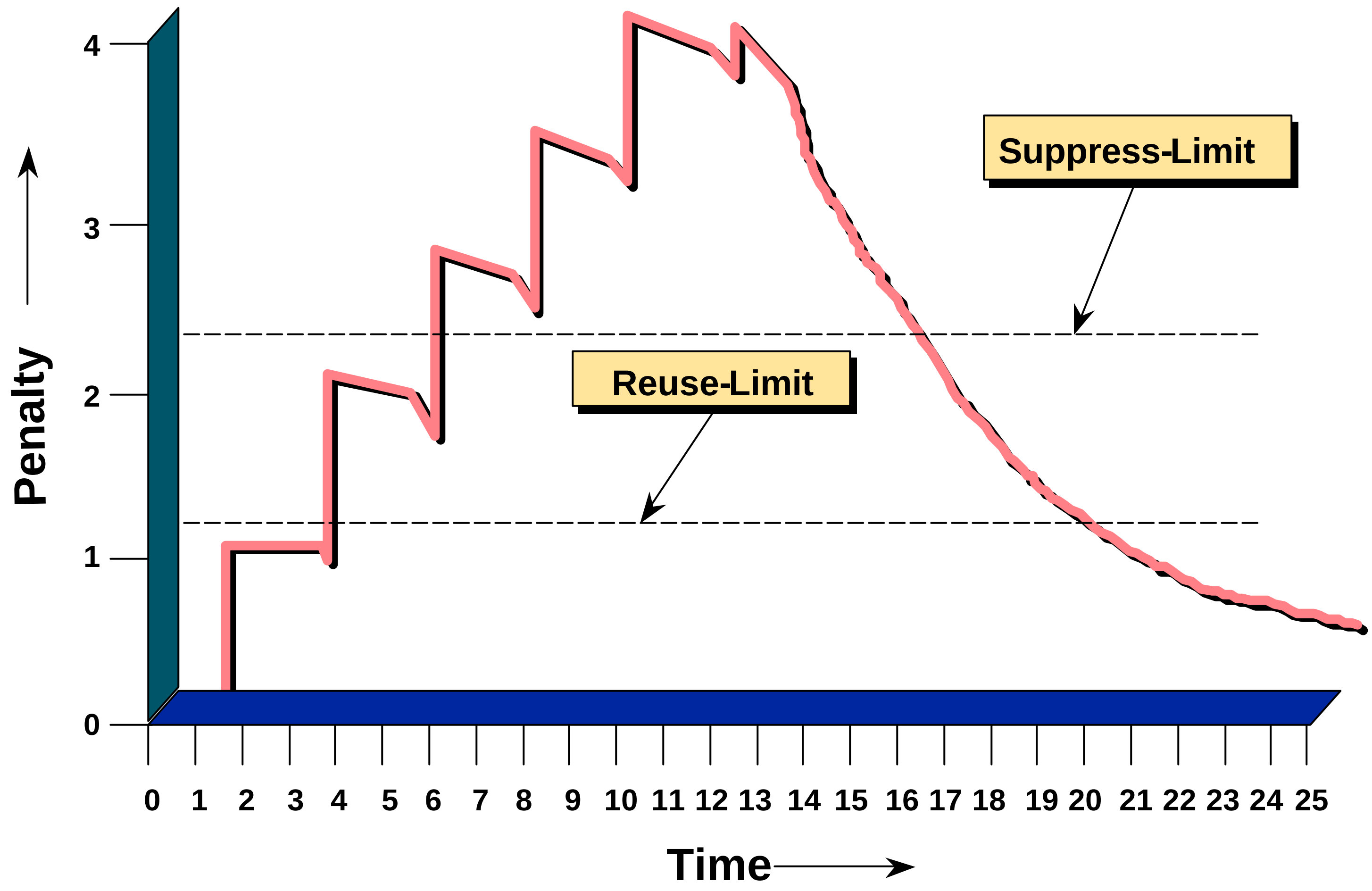
- **Common methods to provide routing information about the NEXT_HOP**

1. **Enable the IGP in the external links (use *passive-interface*).**
2. **Use *redistribute connected*. The information becomes external to the IGP.**
3. **Use *next-hop-self* to make the information internal without adding extra information to the IGP.**

Dampening

- **Defined in rfc2439.**
- **Route flap: The bouncing up and down of a path or a change in its characteristics.**
  - **A flap ripples through the entire Internet**
  - **Consumes CPU cycles, causes instability**
- **Solution: Reduce scope of route flap propagation**
  - **History predicts future behavior**
  - **Suppress oscillating routes**
  - **Advertise stable suppressed routes**
  - **Only external routes are dampened.**

Dampening


Penalty
4
3
2
1
0
Suppress-Limit
Reuse-Limit
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Time

Dampening

- **A route can only be suppressed when receiving an advertisement.**
  - **Not when receiving a WITHDRAW.**
  - **Attribute changes count as a flap (1/2).**
- **In order for a route to be suppressed the following must be true:**
  - **The penalty must be greater than the suppress-limit**
  - **An advertisement for the route must be received while the penalty is greater than the suppress-limit**
  - ***A route will not automatically be suppressed if the suppress-limit is 1000 and the penalty reaches 1200. The route will only be suppressed if an advertisement is received while the penalty is decaying from 1200 down to 1000.***

Dampening – Deployment

- **Configurable parameters:**

  half-life – The number of minutes it takes for the penalty to decay by 1/2

  reuse-limit – If a route is suppressed the penalty must decay to this value to be unsuppressed

  suppress-limit – The penalty must be greater than this threshold when an advertisement is received for a route to be suppressed

  max-suppress-time – The maximum number of minutes a route may be suppressed

Dampening – Deployment

- **Calculated parameters:**

  max-penalty – The maximum penalty a route may have that will allow the penalty to decay to reuse-limit within max-suppress-time

  max-penalty = reuse-limit * 2^(max-suppress-time/half-life)

  If half-life is 30, reuse-limit is 800, and max-suppress-time is 60 then the max-penalty would be 3200. If we allowed the penalty to reach 3201 it would be impossible for the penalty to decay to 800 within 60 minutes.

IOS will generate a warning message if the max-penalty is above 20,000 or less than the suppress-limit.

Dampening – Example

- **Small suppress window:**

  Half-life of 30 minutes, reuse-limit of 800, suppress-limit of 3000, and max-suppress-time of 60

  max-penalty is 3200

- **Advertisement must be received while penalty is decaying from 3200 down to 3000 for the route to be suppressed**

  A 3 min 45 second (rough numbers) window exist for an advertisement to be received while decaying from 3200 to 3000.

Dampening – Example II

- **No window:**

  Half-life of 30 minutes, reuse-limit of 750, suppress-limit of 3000, and max-suppress-time of 60

  max-penalty = 750 * 2^(60/30) = 3000

  Here the max-penalty is equal to the suppress-limit

- **The penalty can only go as high as 3000.**

  The decay begins immediately, so the penalty will be lower than 3000 by the time an advertisement is received.

  A route could consistently flap several times a minute and never be suppressed

Detection and Propagation of Changes – Summary

Cisco.com

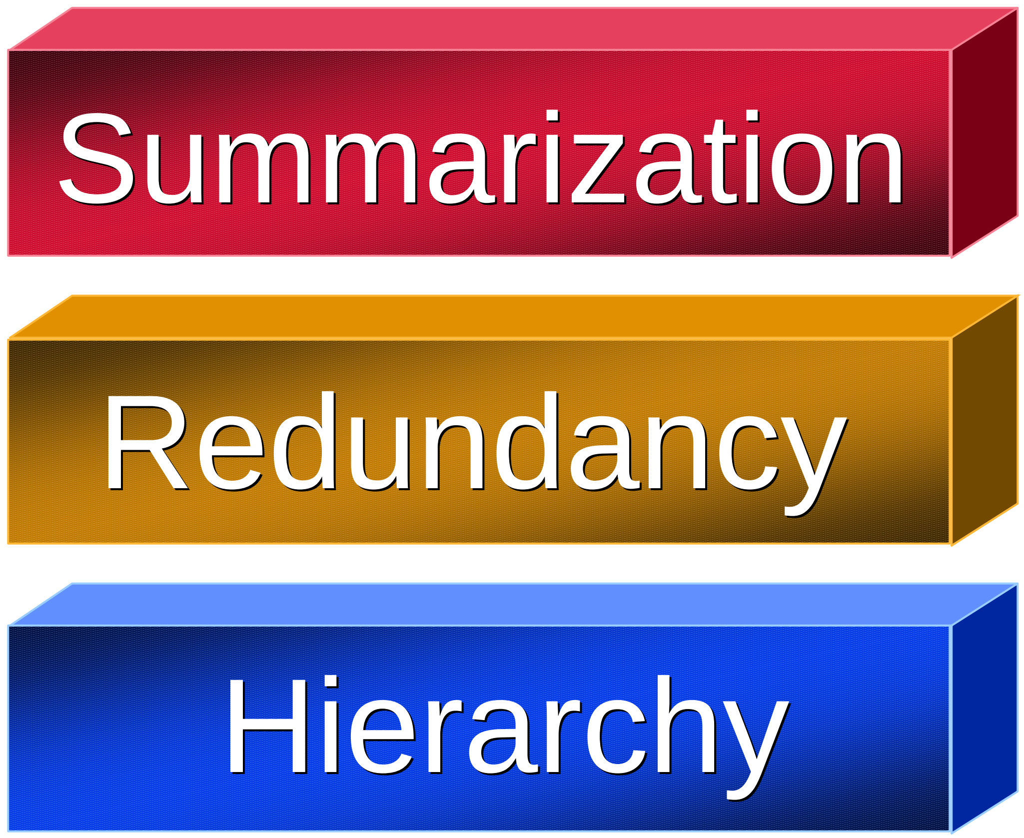
- Use the minRouteAdvertisementInterval and TX Loop Detection to reduce the number of messages generated and the convergence time.
- Choose the appropriate IGP to meet your convergence requirements.
- Implement dampening at all the borders to reduce the impact of external instability in your network.

BGP Scalability – Summary

- Use peer-groups!!
- Eliminate iBGP full mesh with route-reflectors, confederations, or both
- Enable “ip tcp path-mtu-discovery” to improve TCP efficiency
- Increase interface input queues to reduce drops during rush of TCP Acks
- Redundancy is good but too much redundancy only chews up memory without adding much benefit
- Choose dampening parameters with care.

Scalability Building Blocks

Cisco.com



Network Design Goals

- Fast Convergence
- Stable
- Scalable

IP Routing Protocol Scalability

Alvaro Retana (aretana@cisco.com)

IP Routing Deployment and Scalability

Additional Slides

Cisco.com

ISIS Flooding

Detection and Propagation: Flooding

- **All routers generate an LSP**
- **All LSPs need to be flooded to all routers in the network**

  if LSPDB is not synchronised, routing loops or black holes might occur

- **IS-IS' two components are the SPF computation and reliable flooding**

Detection and Propagation: Flooding

- When receiving an LSP, compare with old version of LSP in LSPDB
- If newer:

  install it in the LSPDB

  Acknowledge the LSP with a PSNP

  Flood to all other neighbors

  Check if need to run SPF

Detection and Propagation: Flooding

Cisco.com

Basic Flooding Rules

- **If same age:**

  acknowledge the LSP with a PSNP

- **If older:**

  acknowledge the LSP with a PSNP

  send our version of the same LSP

  wait for PSNP

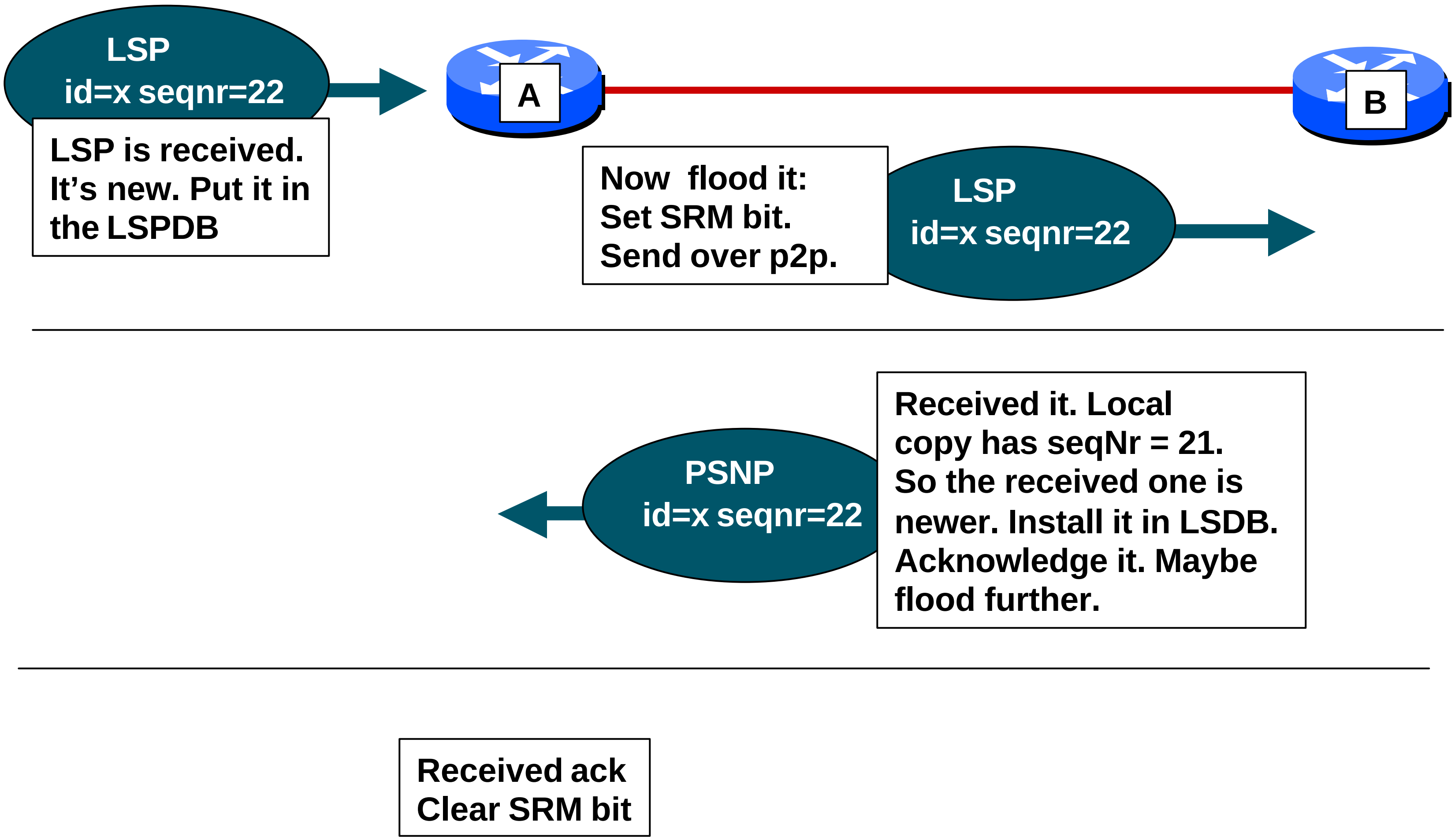
Detection and Propagation: Flooding

Sequence Number

- Each LSP (and LSP fragment) has its own Sequence Number
- When the router boots, set the SeqNr to one
- When there is a change, the SeqNr is incremented, a new version of the LSP is generated with the new SeqNr
- Higher Sequence Number means newer LSP

Detection and Propagation: Flooding on P2P

Cisco.com



Detection and Propagation: DIS

- **DIS is the *Designtated Intermediate System***
- **DIS is only on LANs, not on p2p**
- **DIS has two tasks**

  create/update pseudonode LSP

  conduct flooding over the LAN
- **DIS sends persiodic CSNPs**

  LSP-ID, SeqNr, Checksum, Lifetime of all LSPs present in the LSPDB

Detection and Propagation: DIS

- **No Backup DIS in ISIS**

  not necessary since the DIS doesn't "do" the flooding (a la OSPF)

  flooding is performed directly by all routers on the LAN

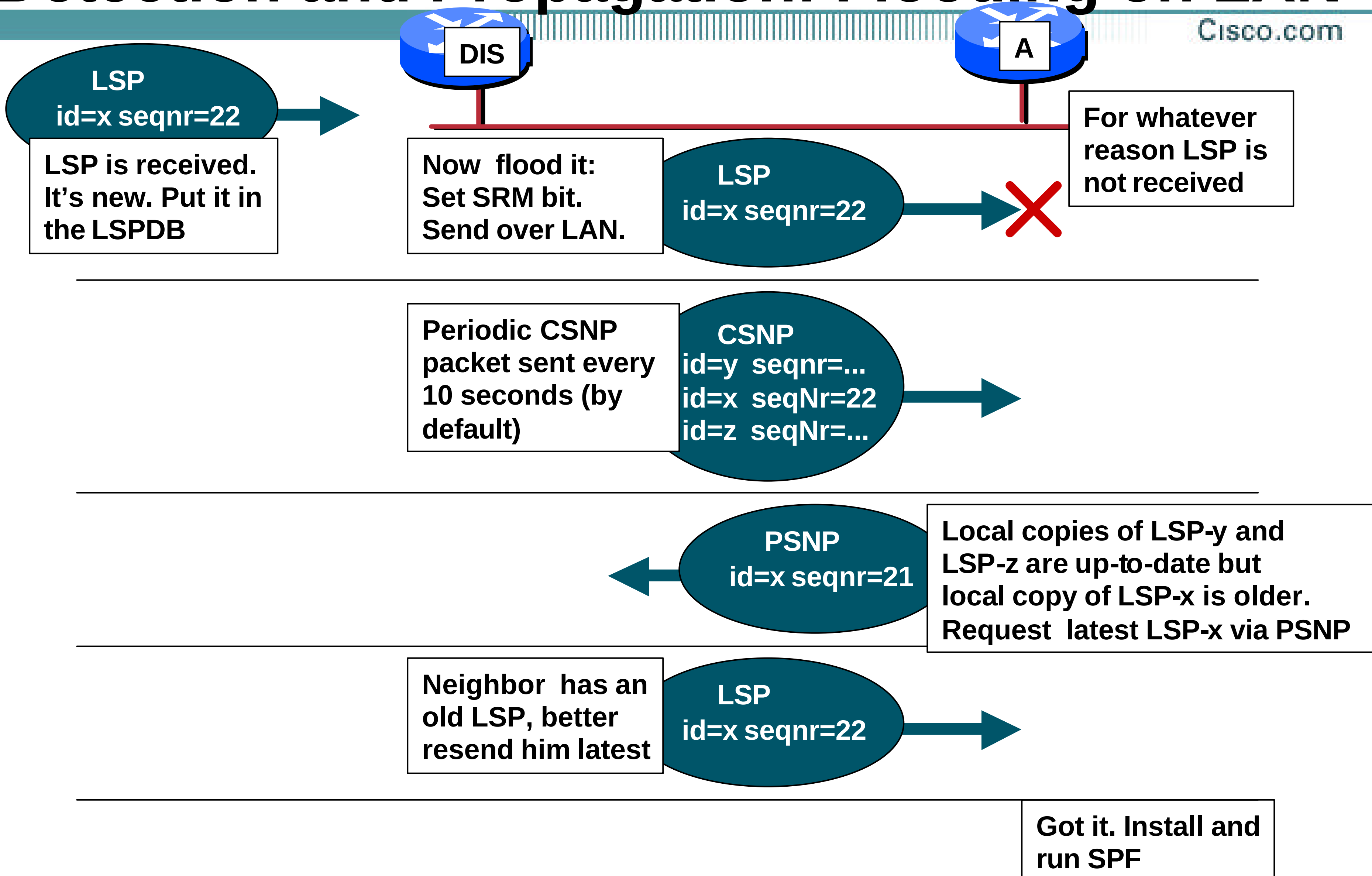
- **DIS is elected by priority and MAC**

  it is "self elected"

- **LAN circuitID shows who is DIS**

  use <show clns interface>

Detection and Propagation: Flooding on LAN



Detection and Propagation: Flooding on LAN

- ISO 10589 states LSP flooding on a LAN should be limited to 30 LSP's per second
- IOS throttles over both LAN and point-to-point interfaces
- Default time between consecutive LSP's is a minimum of 33 milliseconds
- Pacing timer is configurable

Detection and Propagation: Flooding on LAN

- LAN flooding usually doesn't encounter any problems
- No retransmission over LANs
- No ACKs on LANs
- DIS only sends periodic CSNPs
- CSNP timer is configurable

Detection and Propagation: Bad & Good News

- Intuitively we would like to run SPF/PRC/LSP-gen without any initial delay knowing that exponential backoff will protect us from damage ….
- However, link state changes can be categorized as:
  - Bad News
  - Good News

Detection and Propagation: Bad News

- **Bad News:**

  concerns neighbor/adjacency loss or the same neighbor with a worse metric

- **Bad News needs to be processed AS FAST AS POSSIBLE**

  in order to possibly converge to another path

  lose as little packets as possible

Detection and Propagation: Good News

- **Good News:**
  - **concerns new neighbor/adjacency or the same neighbor with a better metric**
- **Good News may wait a little before being processed**
  - **we have been worse off up to now anyway, so we can go on like this a little longer ….**
  - **no packet loss when converging to a better path**

Detection and Propagation: TWCC

Two Way Connectivity Check

- **Remember ?**

  In order for a node to be moved into TENT, it has to report an adjacency to its parent

- **When an adjacency goes down, both ends will generate and flood a new LSP**

- **In order for all other routers to process this change (*Bad News*) only one LSP is needed**

  TWCC will fail during SPF anyway

- ***Bad News* requires one LSP**

Detection and Propagation: TWCC

- **When a new adjacency is advertised, the calculating router must have BOTH LSPs (LSPs from both ends of the adjacency) in order for the adjacency to be considered during SPF**

    in order for a node to be moved into TENT, it has to report an adjacency to its parent

- ***Good News* requires more LSPs**

    another reason to wait a bit longer for good news

OSPF Flooding

Flooding Concepts

The Need for Flooding

- **All routers generate LSAs**
- **All routers must have a consistent view of the network (area).**
- **All LSAs need to be forwarded to all routers in the network (area)**

  if LSDB is not synchronised, routing loops might occur

Propagation of LSAs

Factors influencing propagation

- Speed of light
- Network diameter

Network Diameter

- Each hop takes time to propagate LSA
- If LSA is not received (acked) first time …

  there will be re-transmissions (5 seconds later)
- If LSAs are dropped

  they may be received via different paths

Types of Flooding

- Flooding on point-to-point links
- Flooding on LANs
- General background flooding

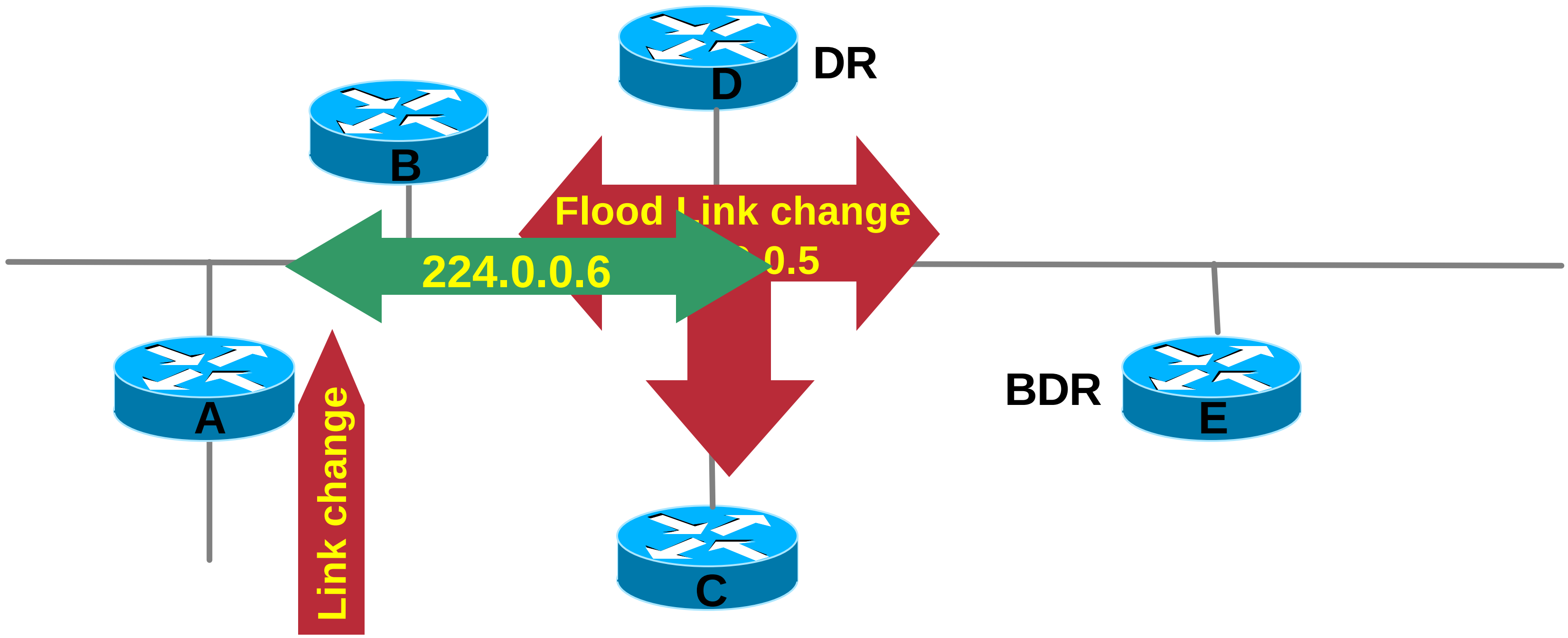
Flooding on point-to-point links

- No concept of Designated Router (DR) or Backup DR. Still reach Full adjacency status.
- Hellos and Updates carried on 224.0.0.5

Flooding on a LAN

- All routers on a LAN synchronize LSDB with DR/BDR.
- DR/BDR listen for updates on 224.0.0.6, then flood updates to other routers on 224.0.0.5
- DR is responsible for originating type-2 LSA (network) for the LAN segment.

Flooding on LAN



What Triggers a New LSA ?

When something changes ...

- Adjacency up / down
- Interface up / down
- Redistributed routes change
- Inter-area routes change
- An interface is assigned a new metric

Basic Flooding Rules

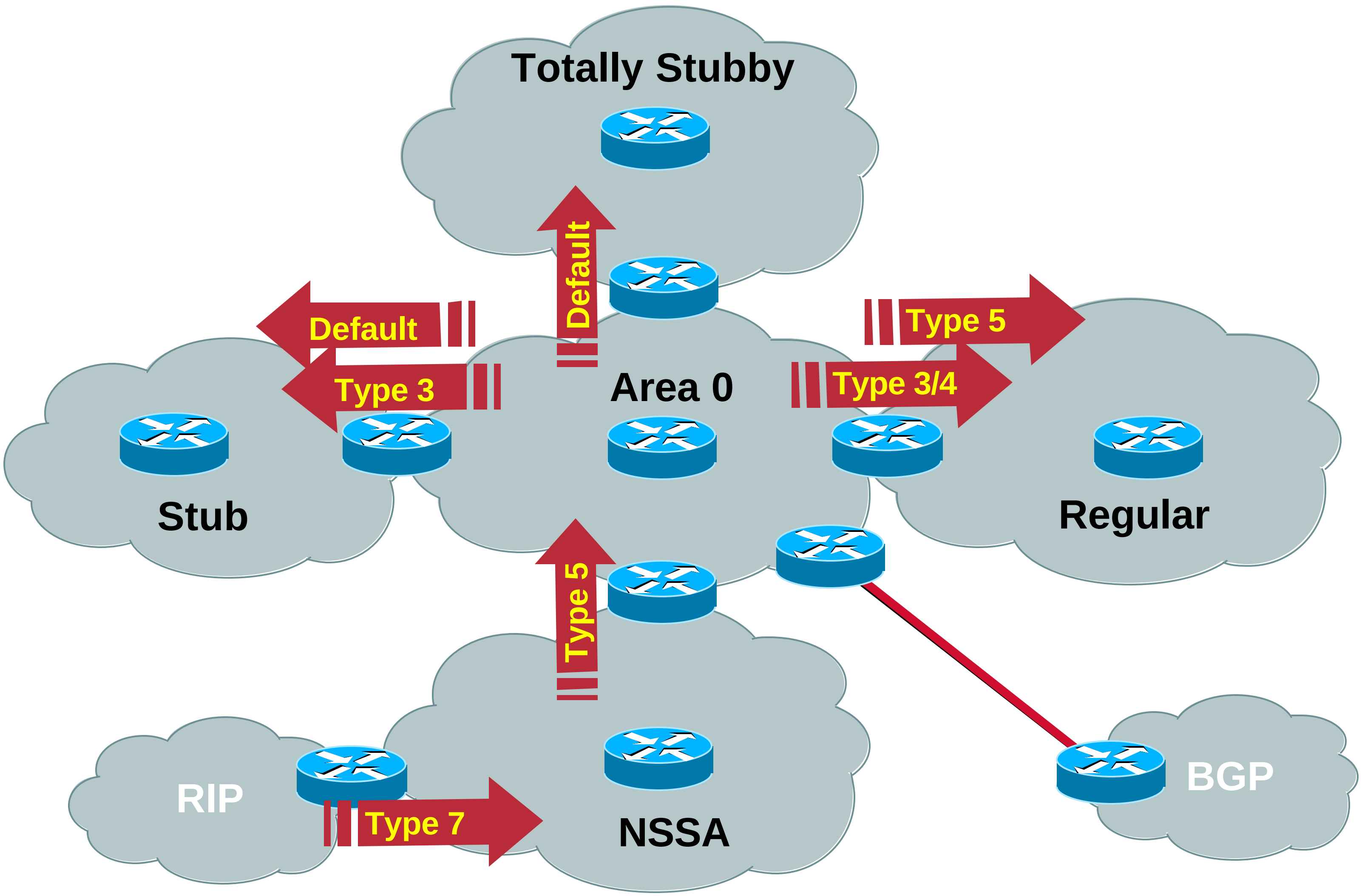
- **When receiving a new LSA, compare with old version of LSA in LSDB.**

  if newer, install it in the LSDB, flood to all other neighbours, (except the one you got it from, send that one an ACK). Check if you need run SPF/Partial-SPF

  if same age, only send ACK, don't flood

  if older, send latest LSP from our LSDB

Areas & LSA Flooding Example



Synchronisation of the LSDB

Cisco.com

For both LANs and Point-to-Point

- Synchronization process begins after bidirectional connectivity established
- Routers exchange Database Descriptors (DBD)
- Link State Requests sent
- Link State Updates sent
- Become Fully Adjacent

Background Flooding

Background Flooding

Cisco.com

- LSAs are still flooded even in stable networks
- Every 1800 seconds (30 minutes) routers re-fresh LSAs they originate
- Only the originating router can re-create and re-flood its own LSAs
- Can cause unnecessary overhead and limit scalability

LSA Age - MaxAge

- **Used to age out and purge old LSAs**

  when the LSA originator has ceased to be active

- **Periodic refresh required**
- **OSPF starts at 0 and counts to 3600 seconds (1 hour)**
- **Received LSAs with MaxAge set (3600) is removed from the LSDB. Setting LSAs to MaxAge can also be used to pre-maturely purge LSAs.**

Controlling Flooding

Controlling Flooding

Scalability means *CONTROLLING* flooding !!

- **Can be resource intensive !**
  - **CPU**
  - **Memory**
  - **Buffers**
- **Bandwidth utilization**
  - **rate-limit flooding on low bandwidth links**

Controlling Flooding – (cont.)

What can be done?

- **Apply good design techniques**
  - use area hierarchy where required
  - summarize
- **Use throttle timers**
  - background flooding timers
  - network specific timers (pt2pt, LAN)

Controlling Background Flooding

- **Increase LSA refresh interval. Sets DNA bit on LSAs but does not suppress hellos. Receiving router does-not-age the received LSAs.**

  ip ospf flood-reduction

- **Adjust LSA group pacing**

  timers lsa-group-pacing *seconds*

  Created to control the synchronization of LSA check-summing, aging and refreshing processes.

  New format in 12.2 IOS

  timers pacing lsa-group

Throttling LSAs

- **LSA Flood Pacing**

  timers pacing flood

  Allows pacing of LSAs queued for flooding. Default is 33 milliseconds. Range is 5 to 100 milliseconds. Available in 12.2 IOS.

- **LSA Retransmission Pacing**

  timers pacing retransmission

  Allows pacing of LSAs queued for re-transmission. Default is 66 milliseconds. Range is 5 to 200 milliseconds. Available in 12.2 IOS.

Throttling LSAs – (cont.)

- **LSA Retransmission Timer**

  Ip ospf retransmit-interval

  Delay, in seconds, between retransmission of unacknowledged LSAs. Available since 10.0 IOS.

Useful LSA Throttling commands

Cisco.com

- **Useful router configuration commands**

  timers pacing flood

  timers pacing lsa-group

  timers pacing retransmission

- **Useful interface configuration commands**

  ip ospf database-filter

  Ip ospf retransmit-interval

  ip ospf flood-reduction

SPF Algorithm

SPF Runs: Shortest Path First Algorithm

- **We maintain three lists**

  UNKNOWN list: all nodes start on this list

  TENTative list: all nodes we are currently examining. Also called the *candidate list*

  PATHS list: all nodes to which we have calculated final paths. Also called the *known list*

SPF Runs: Shortest Path First Algorithm

- **We execute *N* steps:**

  typically N is the number of nodes in the network. During each step we find the path(s) to one node

- **We initialise the computation by moving ourselves onto the TENT list**

SPF Runs: Shortest Path First Algorithm

- **At each step:**
  - **Find the node amongst all nodes on TENT that has the lowest cost, and move it from TENT into PATHS**
  - **Find all prefixes advertised by this node and install them in the RIB**
  - **Find all neighbours reachable from that node and move them into TENT**

SPF Runs: Shortest Path First Algorithm

- **Two Way Connectivity Check (TWCC)**

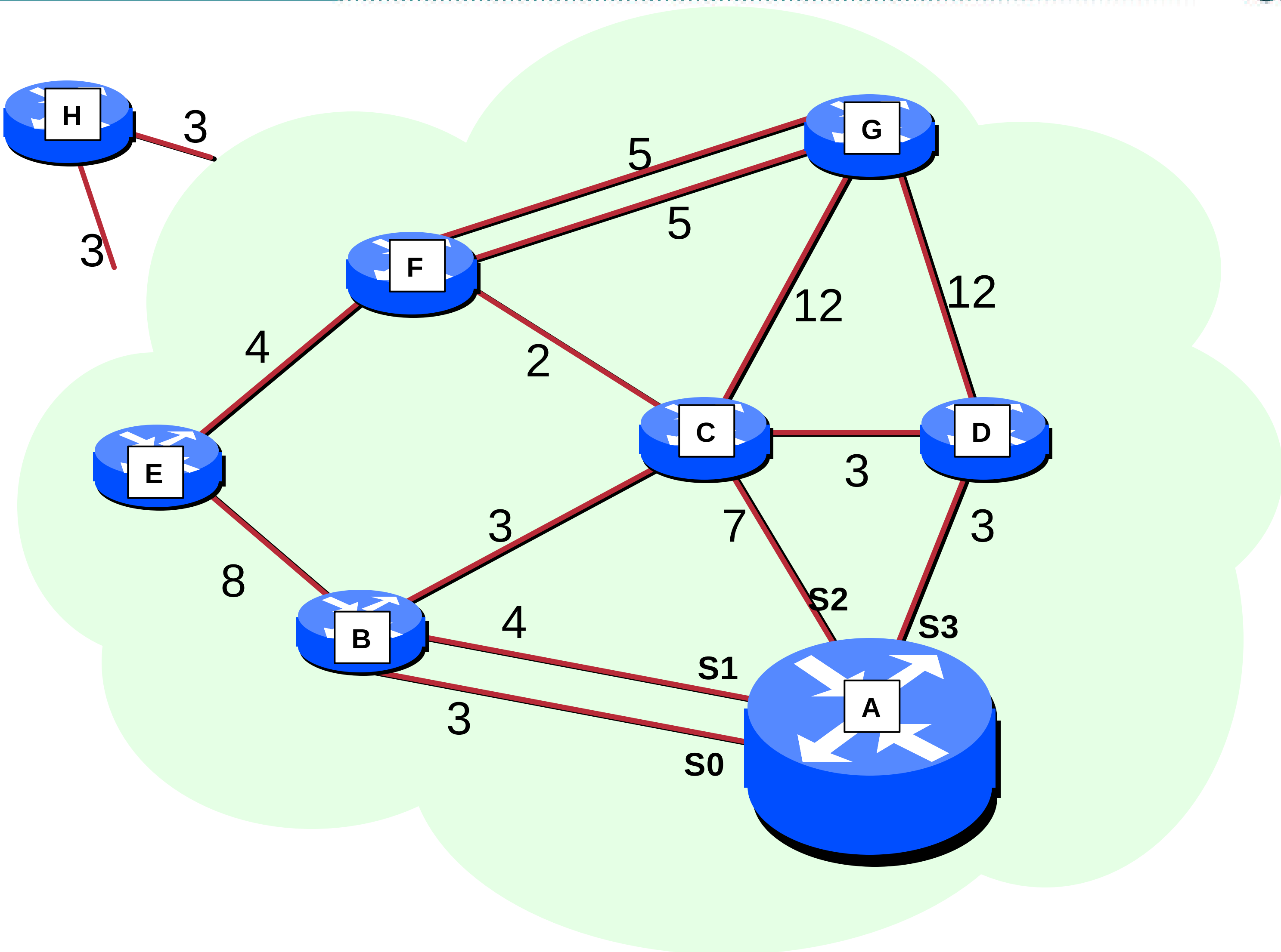
  Before moving a node into TENT, we want to be sure the parent has the same visibility as its child

  The node we want to move to TENT has to report the adjacency to its parent

SPF Runs: Shortest Path First Algorithm

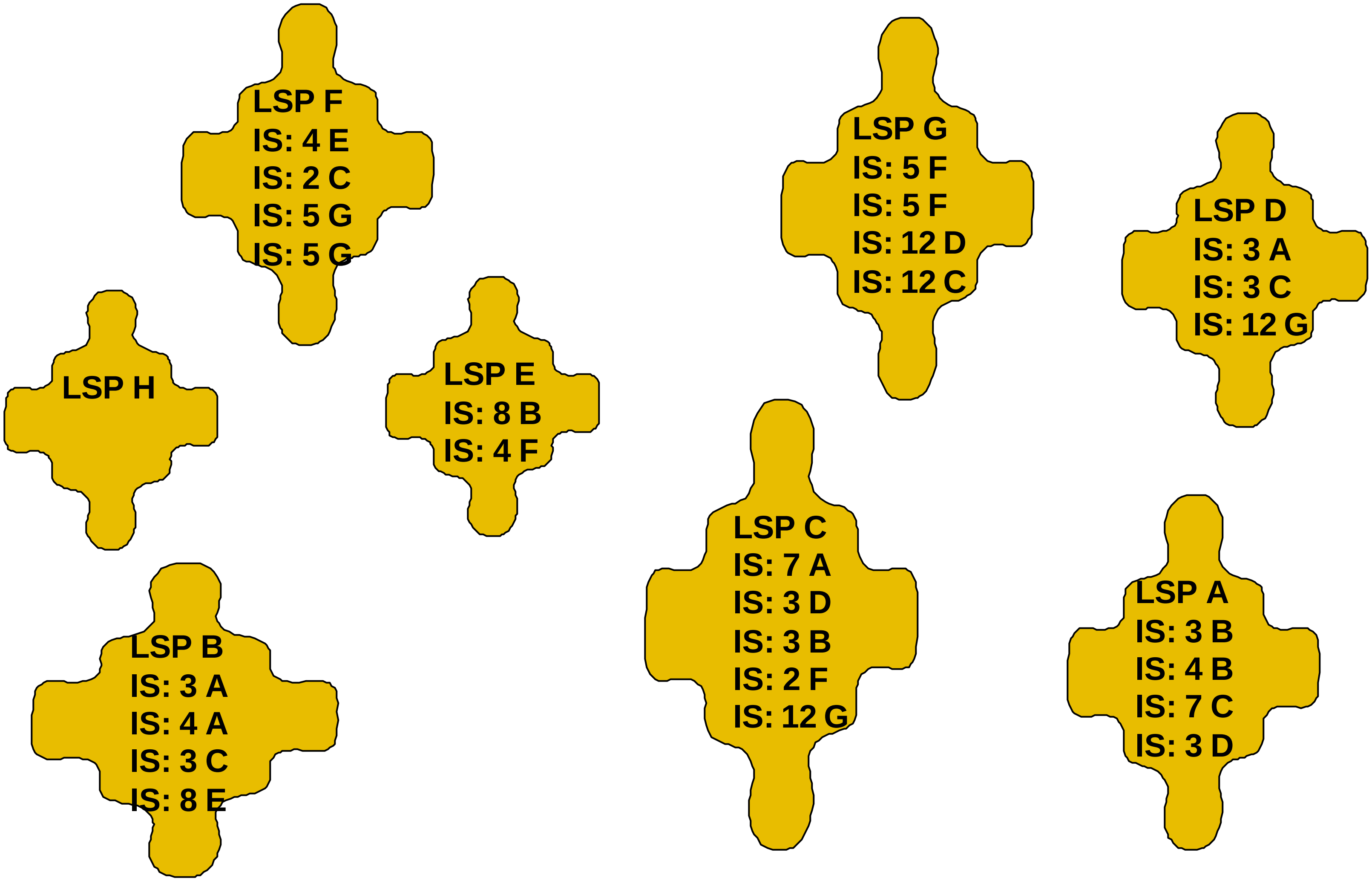
- **Special actions**
  - **If a node is directly connected to us, search the first-hop info in the adjacency database**
  - **If a node is not directly connected to us, copy the first-hop info from the parent(s)**
  - **For each node on TENT, maintain the cost to get there from the root, and the first-hop info**

SPF Runs: Shortest Path First Algorithm



SPF Runs: Shortest Path First Algorithm

Cisco.com



SPF Runs: Shortest Path First Algorithm

Neighbor	Interface	Cost
B	serial0	3
B	serial1	4
C	serial2	7
D	serial3	3

SPF Runs: Partial Route Calculation

- **During SPF, when a node is moved into PATHS, all IP prefixes advertised by that node are inserted into the routing table**
- **In ISIS, IP prefixes are *leaf nodes* of the Shortest Path Tree**
  - **We don’t use IP prefixes to build the Shortest Path Tree**
  - **Routers are identified through CLNS System-IDs**

SPF Runs: Partial Route Calculation

Cisco.com

- **When new LSPs are received, each router will check what has changed in the LSP**
- **If only leaf routes have changed, the SPT need not to be re-built**
  - **As long as the new LSP still advertises the same neighbors with the same metric**
- **In this case we only re-install IP prefixes of the newly received LSP into the routing table**
- **PRC takes the new LSP and re-installs its prefixes into the routing table**

Cisco.com

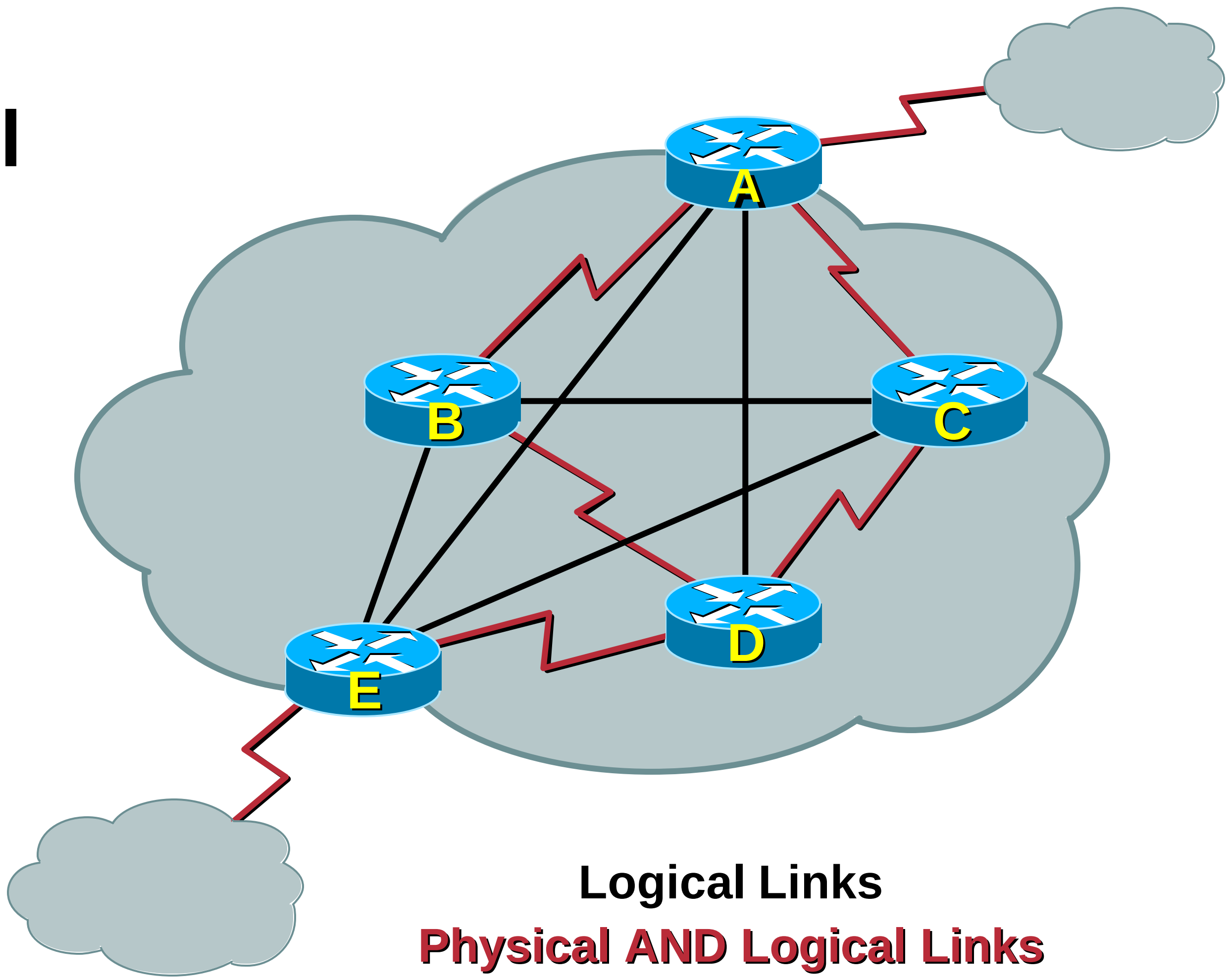
RR Migration

Route Reflectors - Migration

- **Migration is easy**
  - **Configure one RR at a time**
  - **Eliminate redundant iBGP sessions**
  - **Place one RR per cluster**
- **Repeat as needed…**

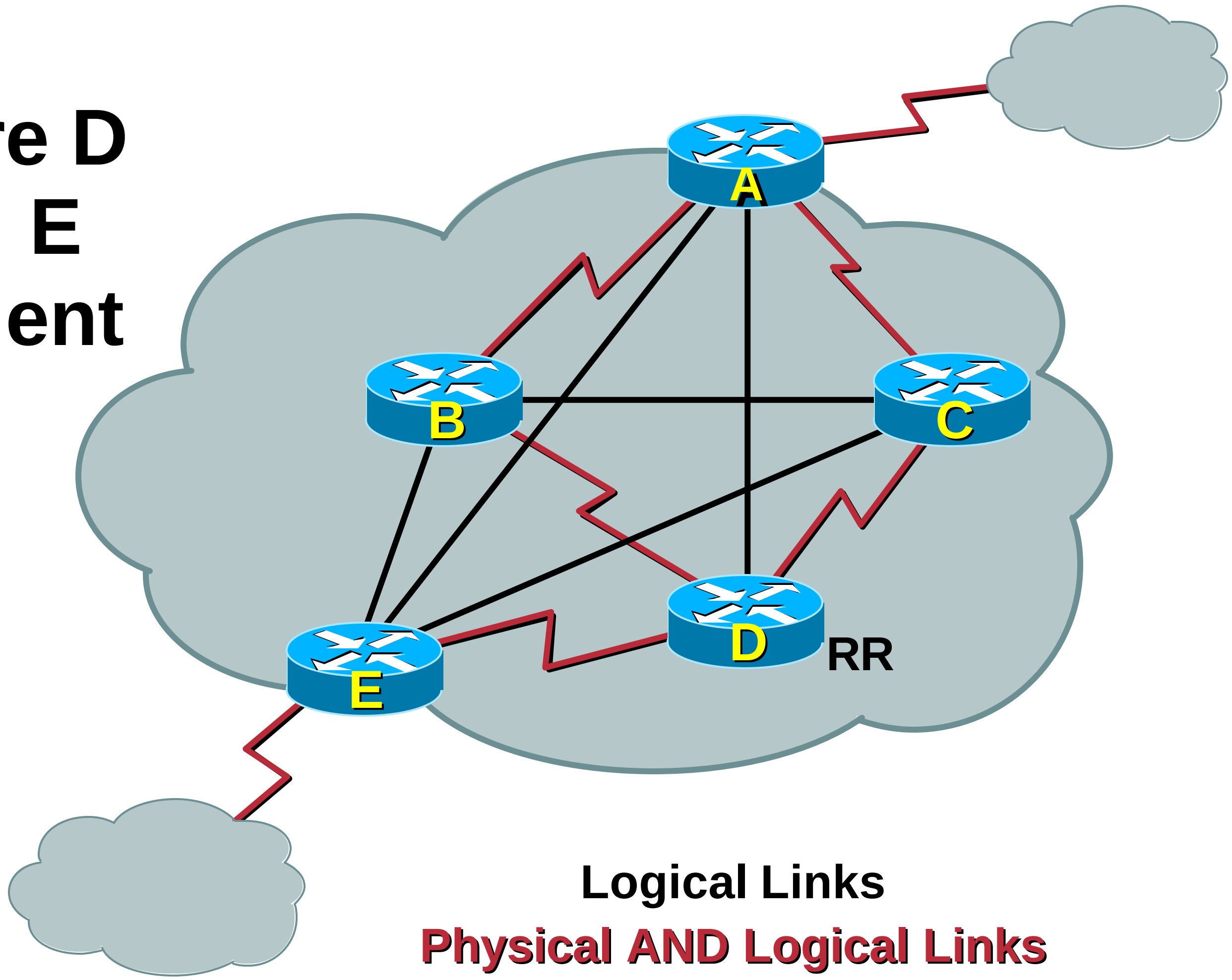
Route Reflectors - Migration

- **Step 0: iBGP full mesh**



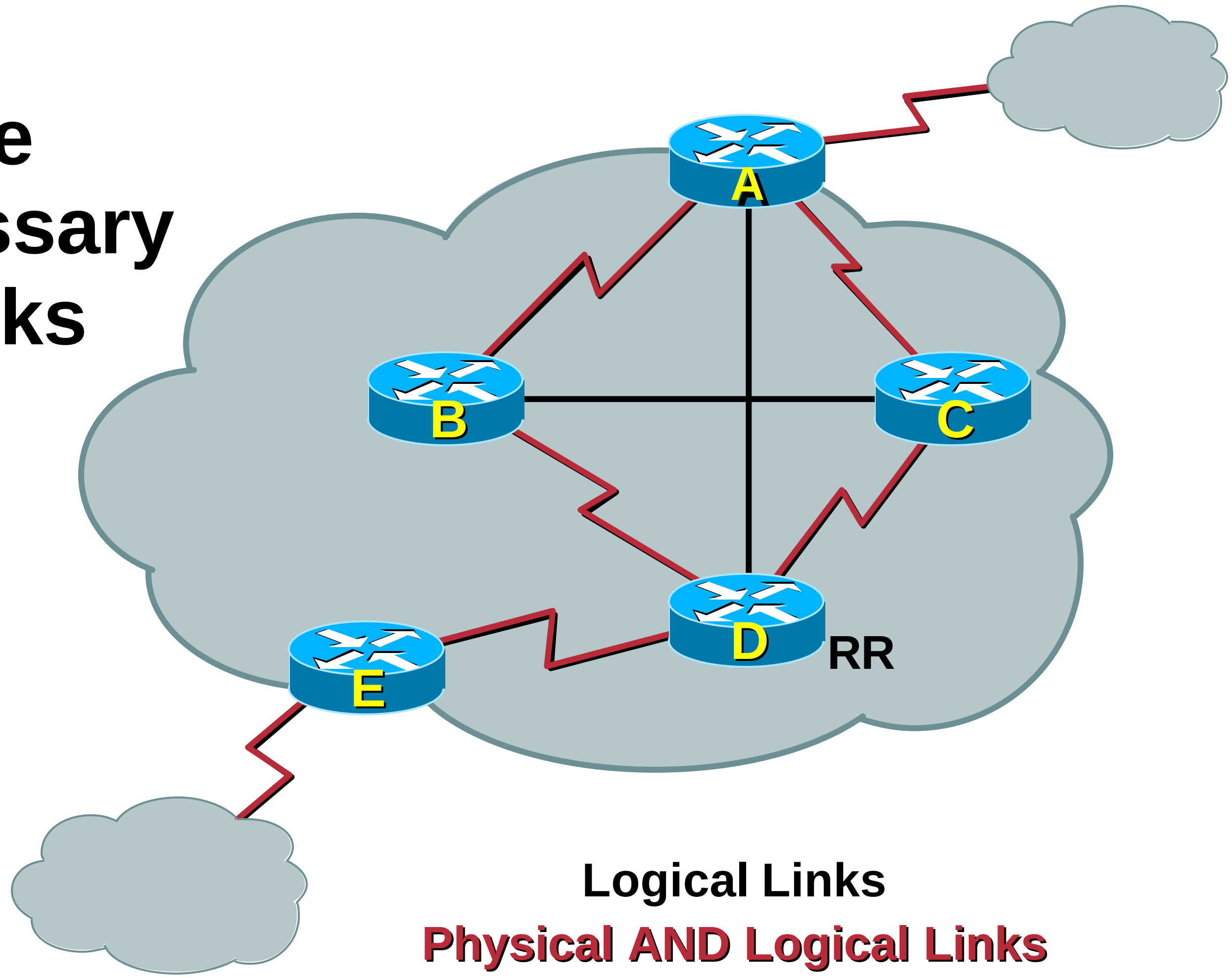
Route Reflectors - Migration

- **Step 1: configure D as a RR; E is the client**



Route Reflectors - Migration

- **Step 2: eliminate unnecessary iBGP links**



Route Reflectors - Migration

- **Step 3: repeat for other clusters and iBGP links.**
- **Finished!!**