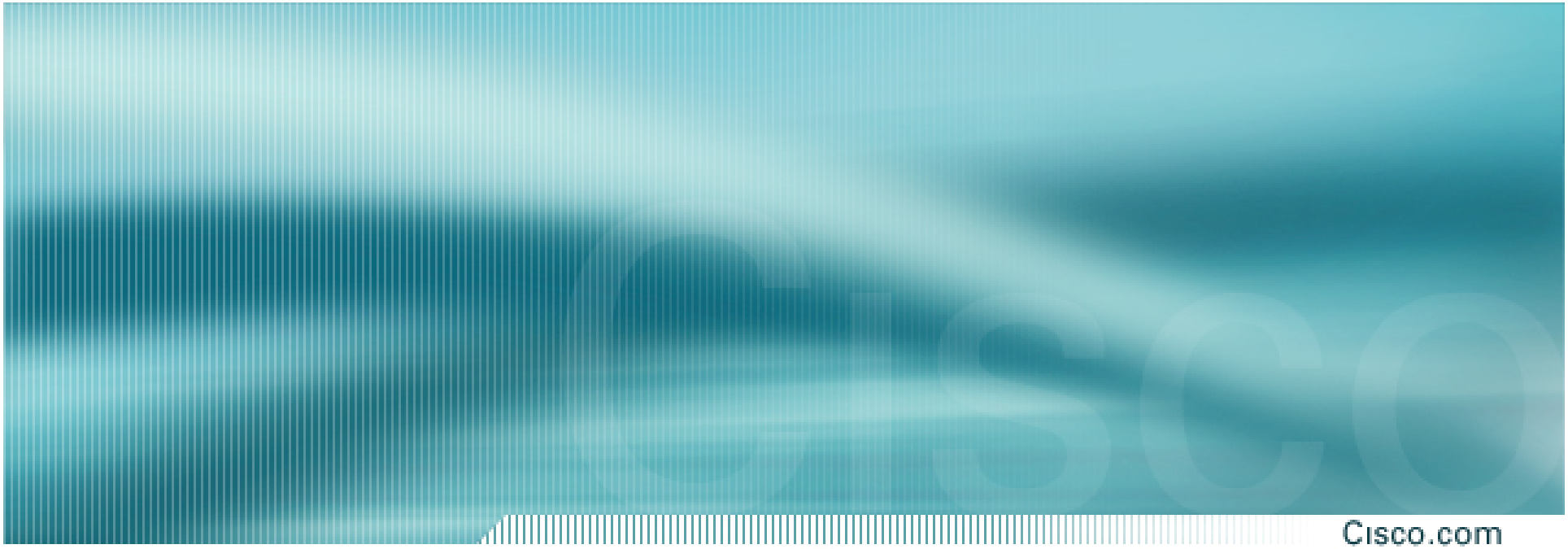


Optimizations for Routing Protocol Stability and Convergence

Brian Daugherty, CSE, CCIE #5879
Higher Education and Research Network Initiatives
bdaugher@cisco.com

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence?

Cisco.com

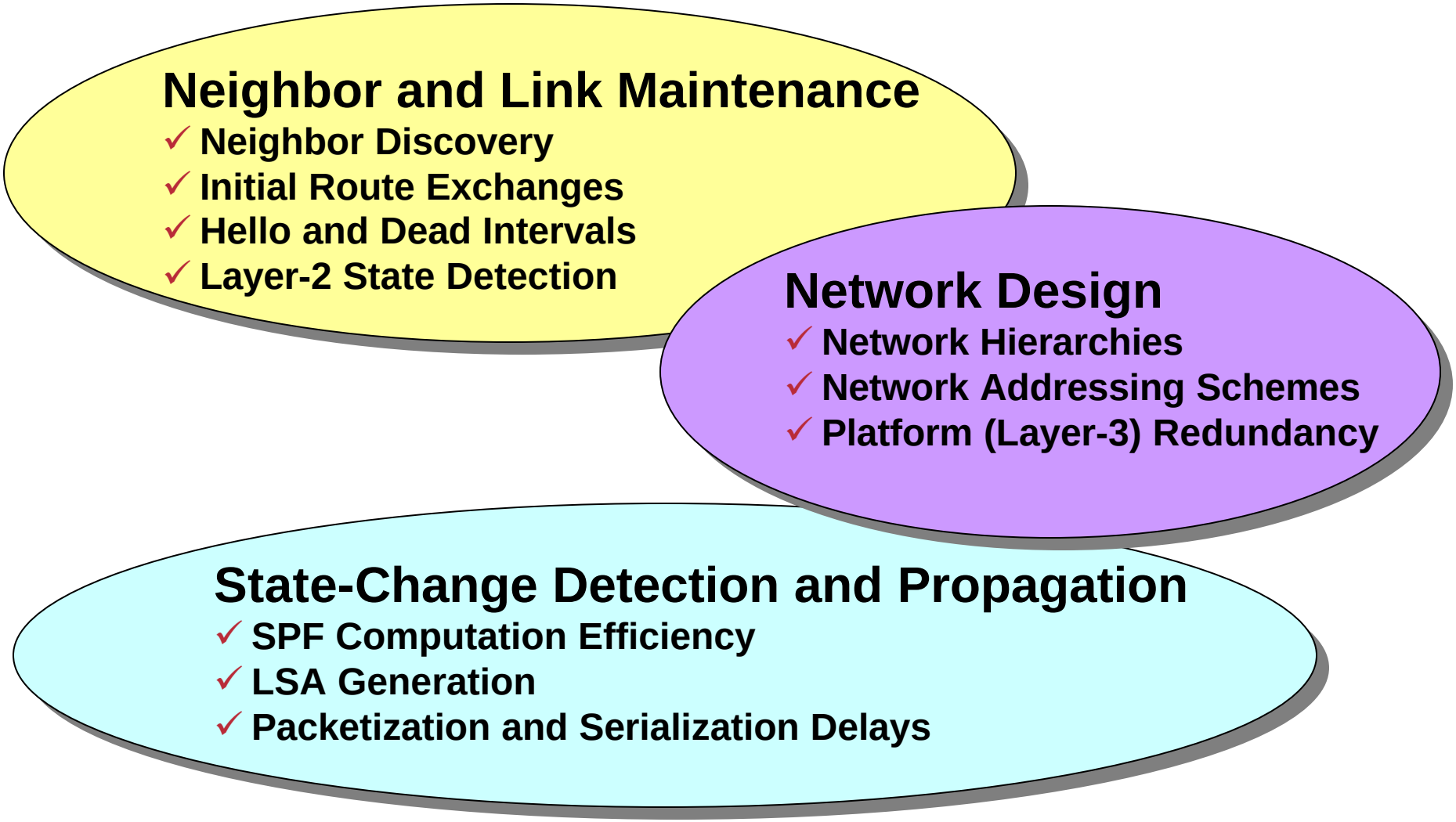
con·ver·gence (k?n-vûr´j?ns) *n.*

1. The act, condition, quality, or fact of converging.
2. Mathematics. The property or manner of approaching a limit, such as a point, line, function, or value.
3. The point of converging; a meeting place: a town at the convergence of two rivers.
4. Physiology. The coordinated turning of the eyes inward to focus on an object at close range.
5. Biology. The adaptive evolution of superficially similar structures, such as the wings of birds and insects, in unrelated species subjected to similar environments.
6. *Networking. The period beginning when a topology change occurs and ending the moment all routers have a consistent view of the network.*

1-5: The American Heritage® Dictionary of the English Language, Fourth Edition
Copyright © 2000 by Houghton Mifflin Company.
Published by Houghton Mifflin Company. All rights reserved

© 2002, Cisco Systems, Inc. All rights reserved.

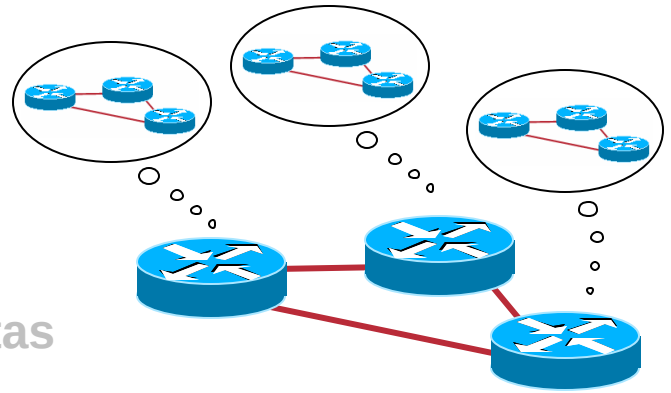
Elements Affecting Convergence



© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda

- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**



© 2002, Cisco Systems, Inc. All rights reserved.

Fast Hello Handling
Motivation

- **Interface (Layer-2) state is not always a reliable indicator of peer health**
  - **Not suitable for multi-hop or LAN connections**
  - **Cannot immediately detect peer software failures**
- **Lower values for Hello, Dead, and Holding Timers can decrease peer failure detection times**
- **OSPF and ISIS can set Dead and Hold timers to a minimum of 1 second and send Hellos at sub-second intervals (12.2S/12.0S)**

© 2002, Cisco Systems, Inc. All rights reserved.

Fast Hello Handling
Advantages and Disadvantages

- **Advantages**
  - **Reduced link failure detection time**
- **Disadvantages**
  - **Increased line bandwidth, buffer, and CPU utilization can cause missed Hellos**
  - **Increased potential for improper adjacency flapping can cause routing instability**
- ***Not a recommended solution!***

© 2002, Cisco Systems, Inc. All rights reserved.

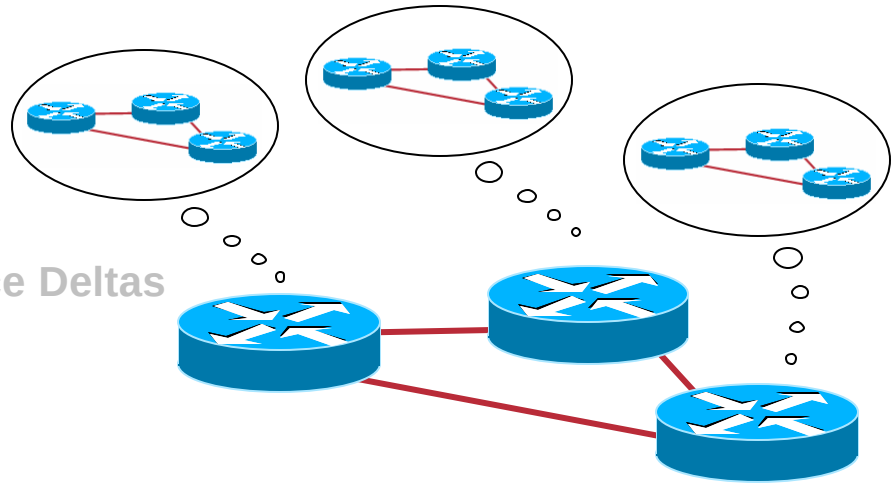
Fast Hello Handling
Scalability

- **Scaling is A Major Issue**
  - **300 interfaces x 10 neighbors/interface = 3000 neighbors**
  - **3 Hello packets per second on each interface**
  - ***Router has to generate 900 hellos per second***
  - **3000 neighbors each send 3 Hello packets per second to this router**
  - **Router has to accept and process 9000 Hello packets per second**
  - ***Router has to deal with 9900 Hello packets per second – one every 1/10000 of a second***

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda

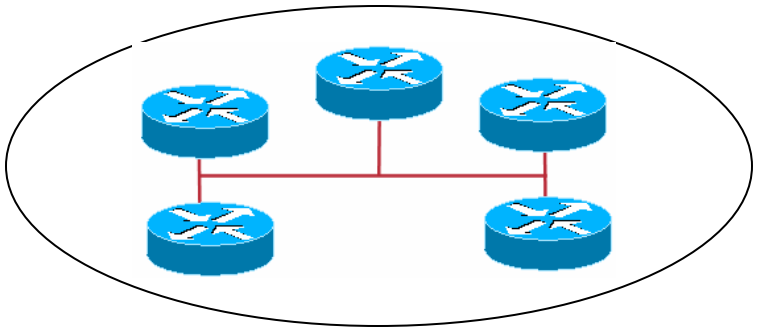
- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**



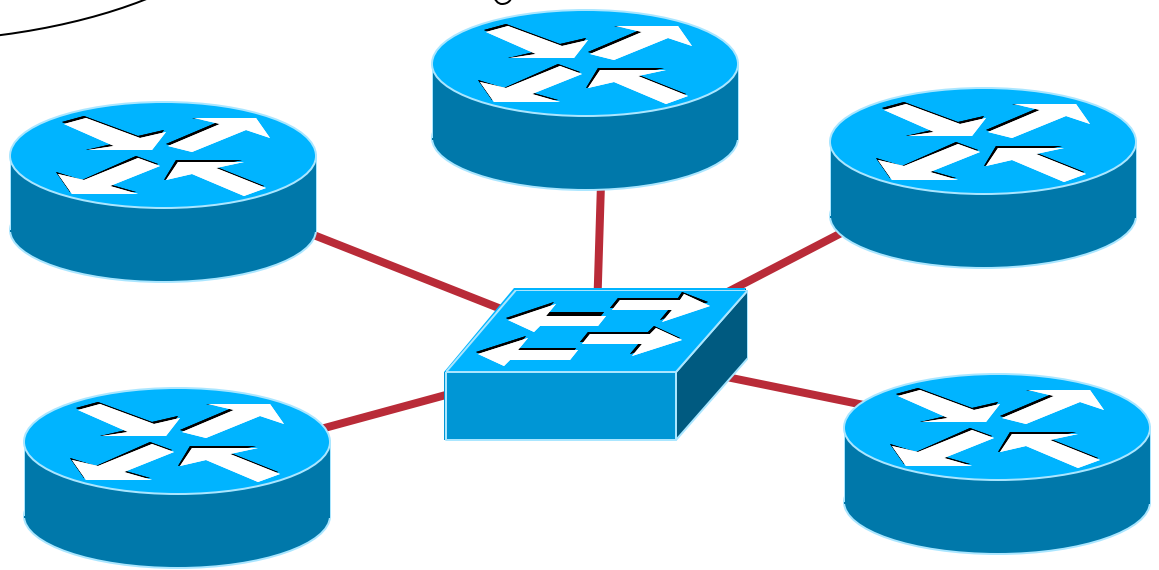
© 2002, Cisco Systems, Inc. All rights reserved.

Multi-Access Reachability Protocol (MARP)

To the router this is a broadcast network....

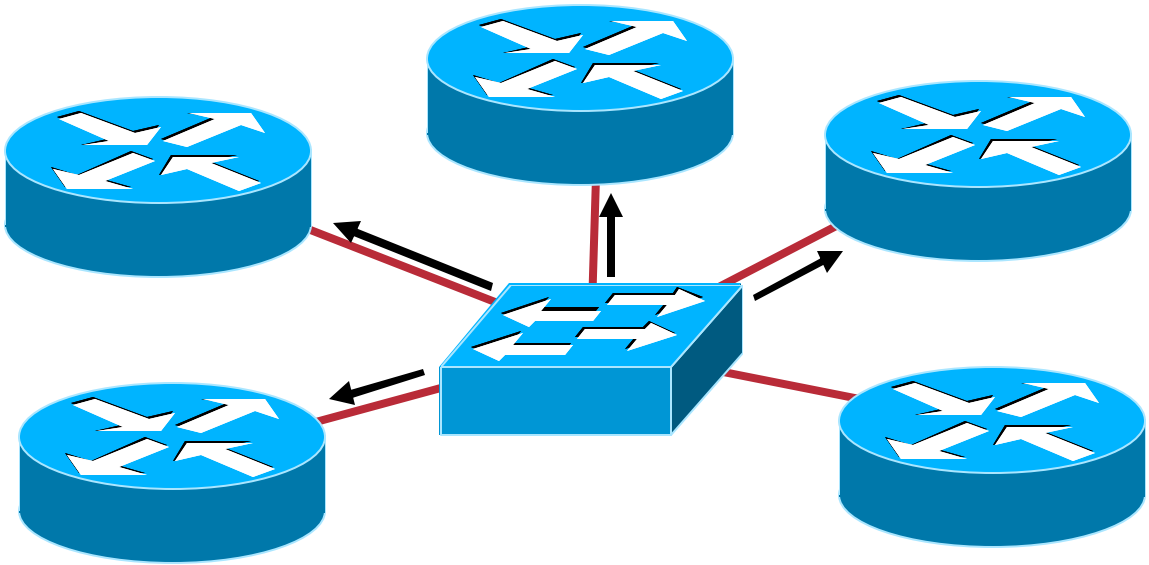


To the switch this is a point-to-point network!



© 2002, Cisco Systems, Inc. All rights reserved.

Multi-Access Reachability Protocol
Concept



With MARP* the switch can notify connected routers of link failures immediately – no need to wait for the Dead- and Hold-Timers to expire

*** Protocol in development…EFT potentially at the end of CY02.**

© 2002, Cisco Systems, Inc. All rights reserved.

Multi-Access Reachability Protocol
Operation

Cisco.com

- **Routers use MARP to advertise a set of Layer-2 addresses for important neighbors**
- **Switches listen for these advertisements and monitor the ports on which these addresses appear**
- **If a monitored port fails, MARP link-failure messages are sent via multicast to everyone on the broadcast domain**

Simple packet format uses TLVs for future address types and extended flexibility

© 2002, Cisco Systems, Inc. All rights reserved.

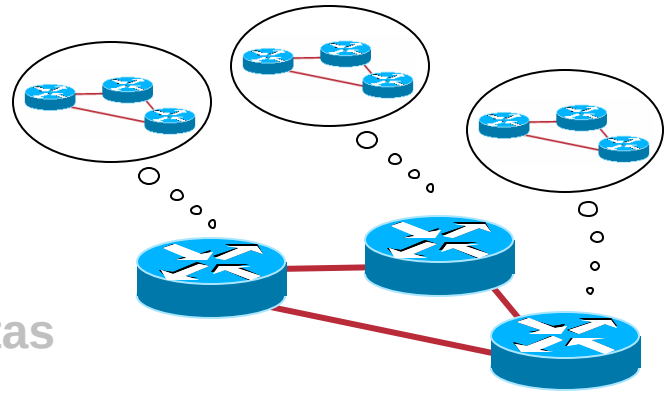
Multi-Access Reachability Protocol
Summary

- **An alternative to Fast Hellos**
  - **Allows for sub-second convergence times**
  - **Imposes no overhead on the router**
- **Currently under development**
  - **EFT possible by end of CY02**
  - **Initially supported in Cat 6500 and 4000 switches and routing platforms**
  - **Routing protocols are the first step – other protocols (such as HSRP) will follow**
- **Cisco-Only solution today, but…**
  - **MARP will be submitted for standardization at IETF**

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda

- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**



© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Introduction

Cisco.com

- **BGP startup involves advertising 130K+ routes to potentially hundreds of peers**
- **BGP implementations play a major roll in how fast a router can converge after startup**
- **A series of enhancements and fixes have recently been introduced into IOS**
  - ***NOTE: All the following graphs show the number of peers that can be converge (y-axis) within 10 minutes given the number of routes (x-axis) to be advertised***

© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Using BGP Peer-Groups

Cisco.com

- **Problem: Advertise 130K+ routes to hundreds of peers – BGP will need to send gigabytes 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***

© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Operation

- **UPDATE generation without peer-groups**
  1. **The BGP route table is walked once**
  2. **Prefixes are filtered through outbound policies**
  3. **UPDATEs are generated and sent per peer**
- **UPDATE generation with peer-groups**
  1. **A peer-group *leader* is elected for each peer-group**
  2. **The BGP table is walked once (for the leader only)**
  3. **Prefixes are filtered through outbound policies**
  4. **UPDATEs are generated and sent to the peer-group leader, then 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…

© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Synchronization

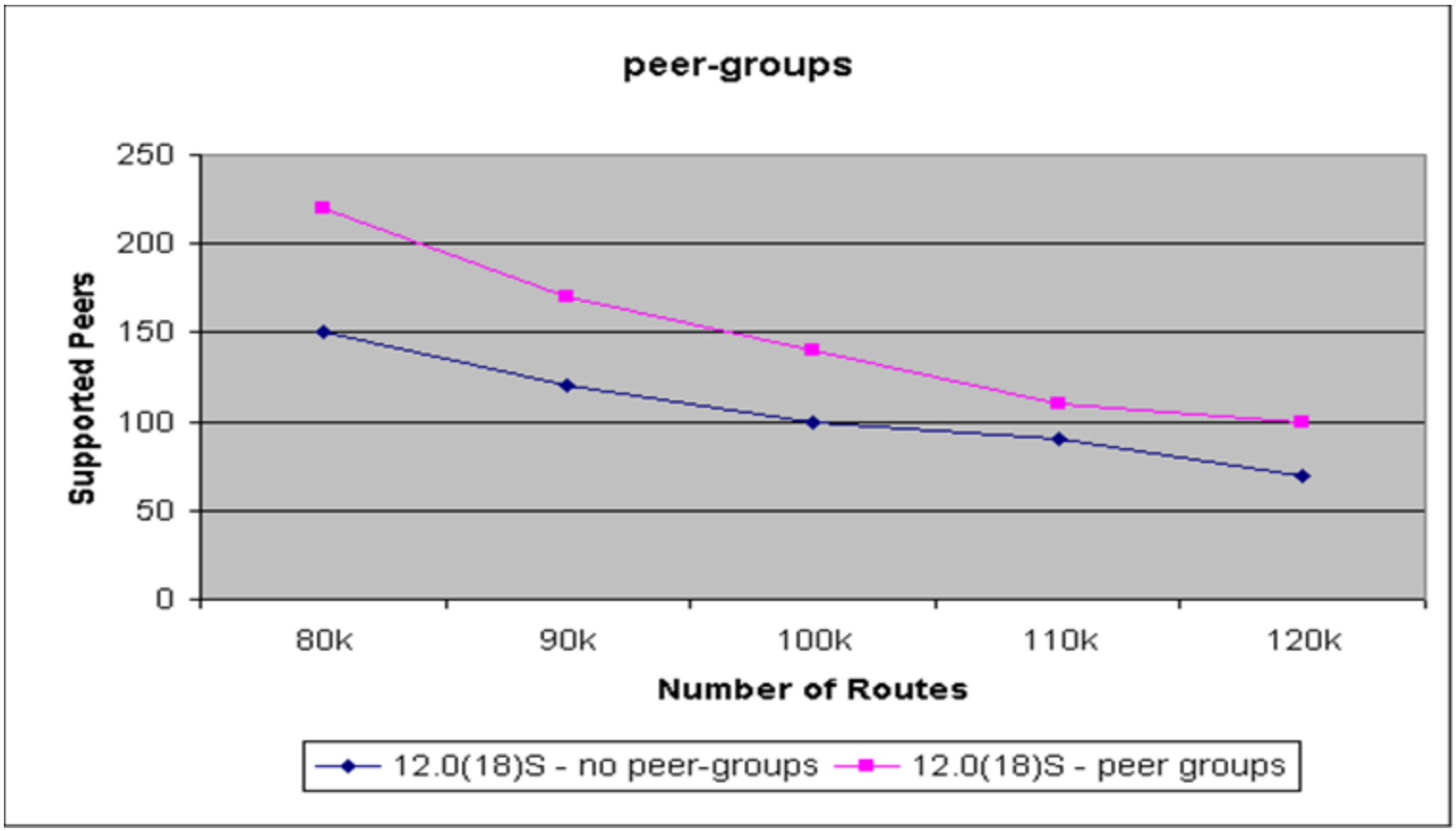
Cisco.com

- **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**

© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Peer-Groups – Effectiveness

- **Using BGP Peer-Groups increases scalability between 35% - 50%**



© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Using Larger Input Queues

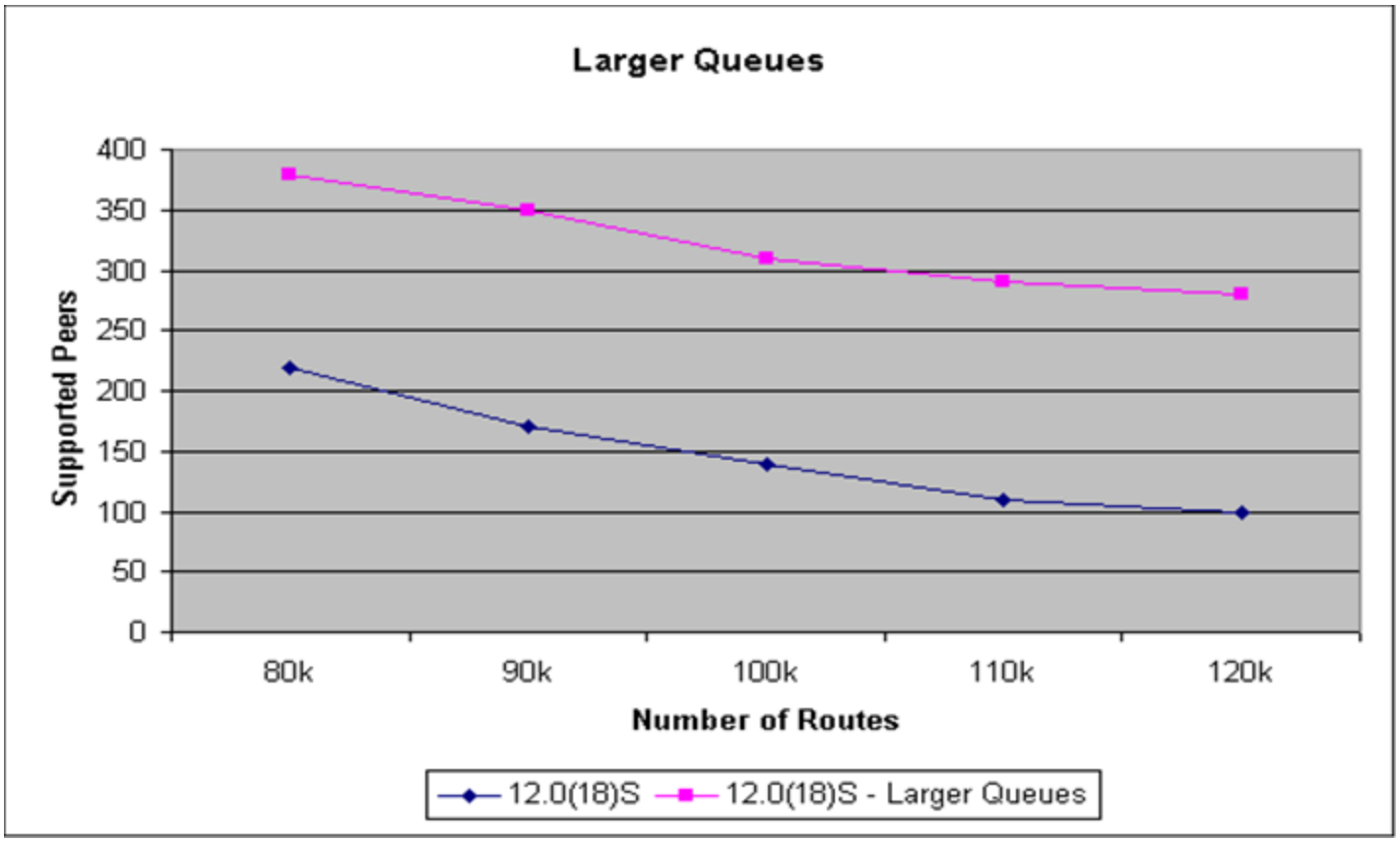
Cisco.com

- **Input Queue Saturation**
  - **Convergence times can degrade due to the enormous numbers of dropped TCP ACKs (130K+) on peer-facing input queues**
  - **A typical ISP gets ~½ million drops in 15 minutes on an average route reflector**
- **Increasing the size of the input queues**
  - **Reduces the number of dropped TCP ACKs – therefore reducing retransmissions**
  - **Improves BGP convergence times**
  - **Improves BGP scalability**

© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Larger Input Queues – Effectiveness

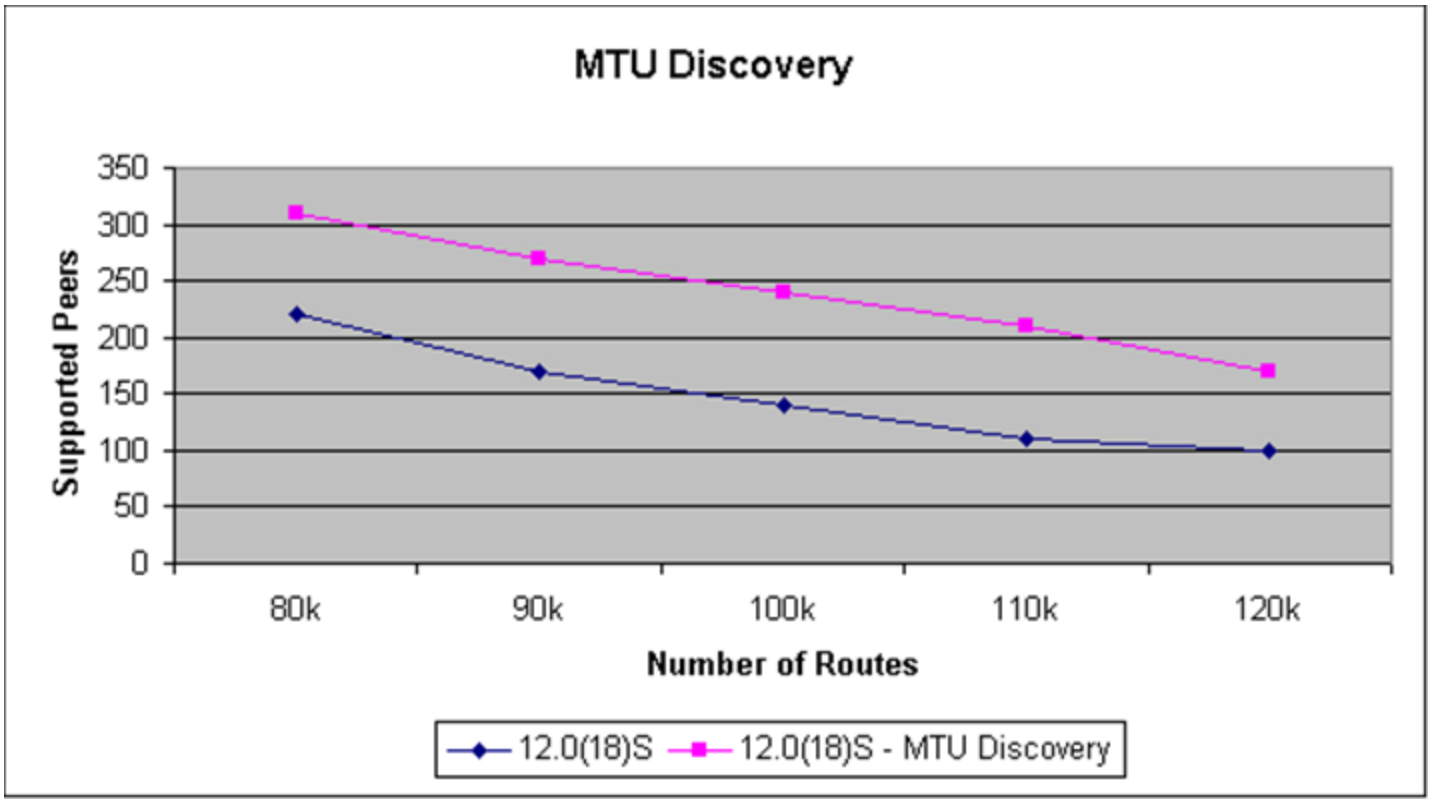
- **Rush of TCP Acks from peers can quickly fill the process-level input queue at default depth (75)**
- **Increasing queue depths (4096) improves BGP scalability**



© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Using MTU Discovery

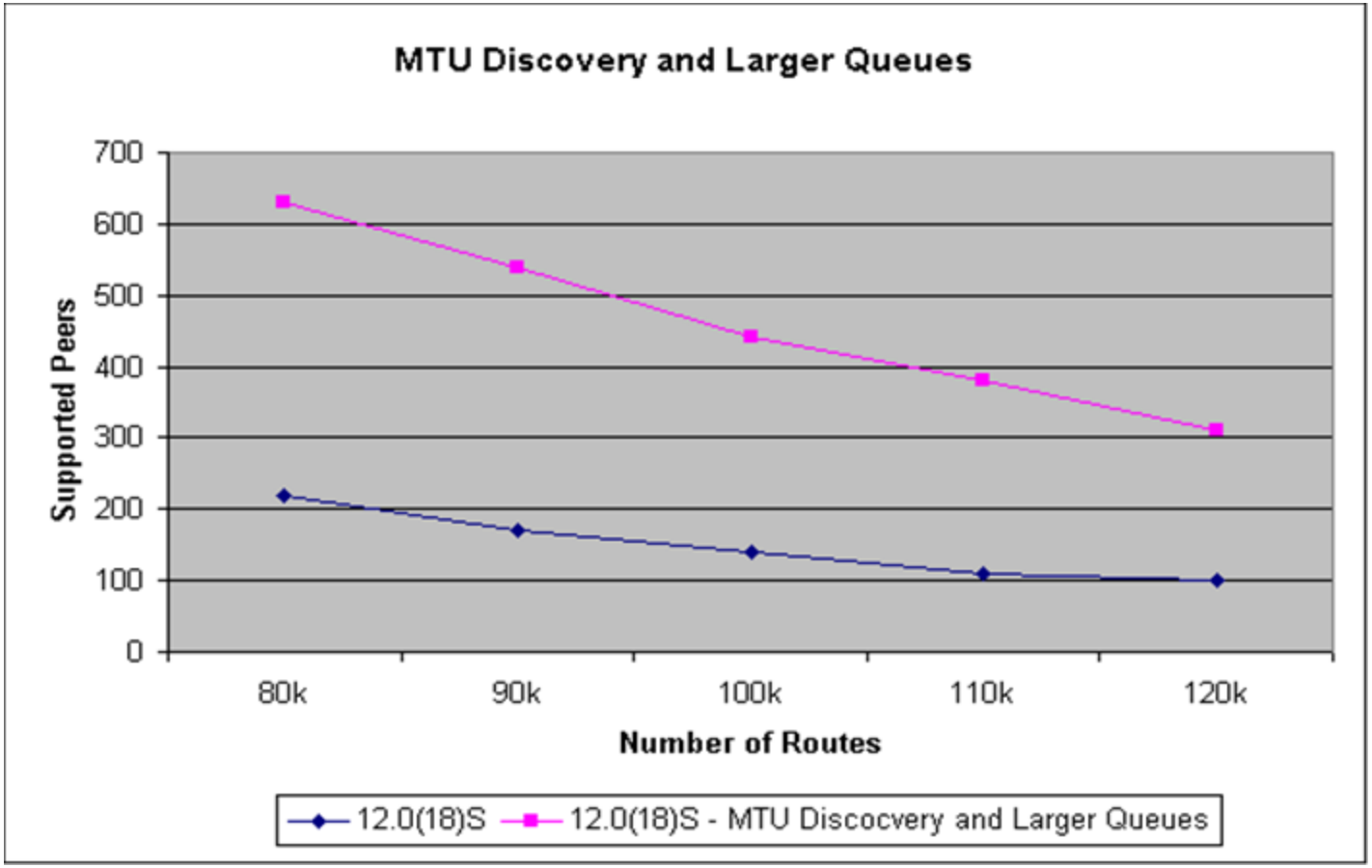
- **Default MSS (Max Segment Size) is 536 bytes**
- **Default MSS is inefficient on current POS and Ethernet networks**
- **ip tcp path-mtu-discovery improves convergence**



© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
MTU Discovery and Larger Input Queues – Effectiveness

- **Simple config changes can give 3x improvement**



© 2002, Cisco Systems, Inc. All rights reserved.

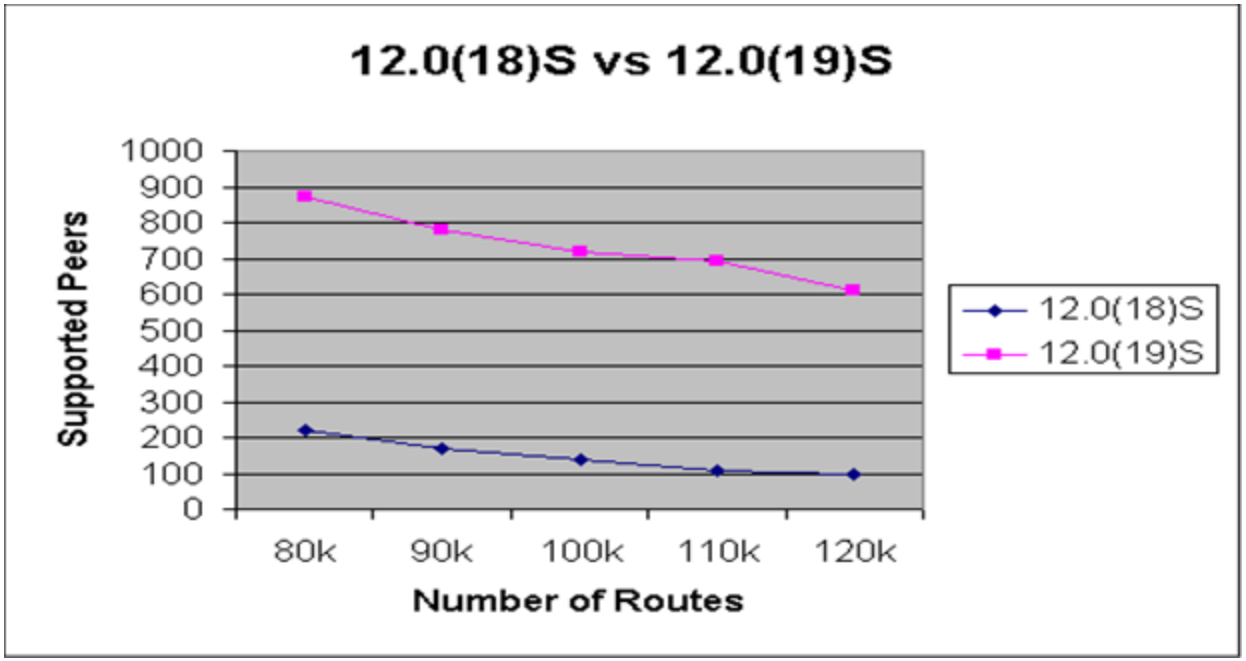
BGP Route Convergence Optimizations
BGP UPDATE Packing

- **A BGP UPDATE contains a group of attributes that characterize one (or more) prefixes**
  - **Ideally – all the prefixes with common attributes should be advertised in the same UPDATE message**
  - **For example:**
    - ✓ **A BGP table containing 100K routes and 15K attribute combinations can be advertised in 15K updates – 100% UPDATE packing is achieved**
    - ✓ **If it takes you 100K updates then 0% UPDATE packing is achieved**
- **Convergence times vary greatly with**
  - **The number of attribute combinations**
  - **BGP UPDATE packing efficiency**

© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
BGP UPDATE Packing – Effectiveness

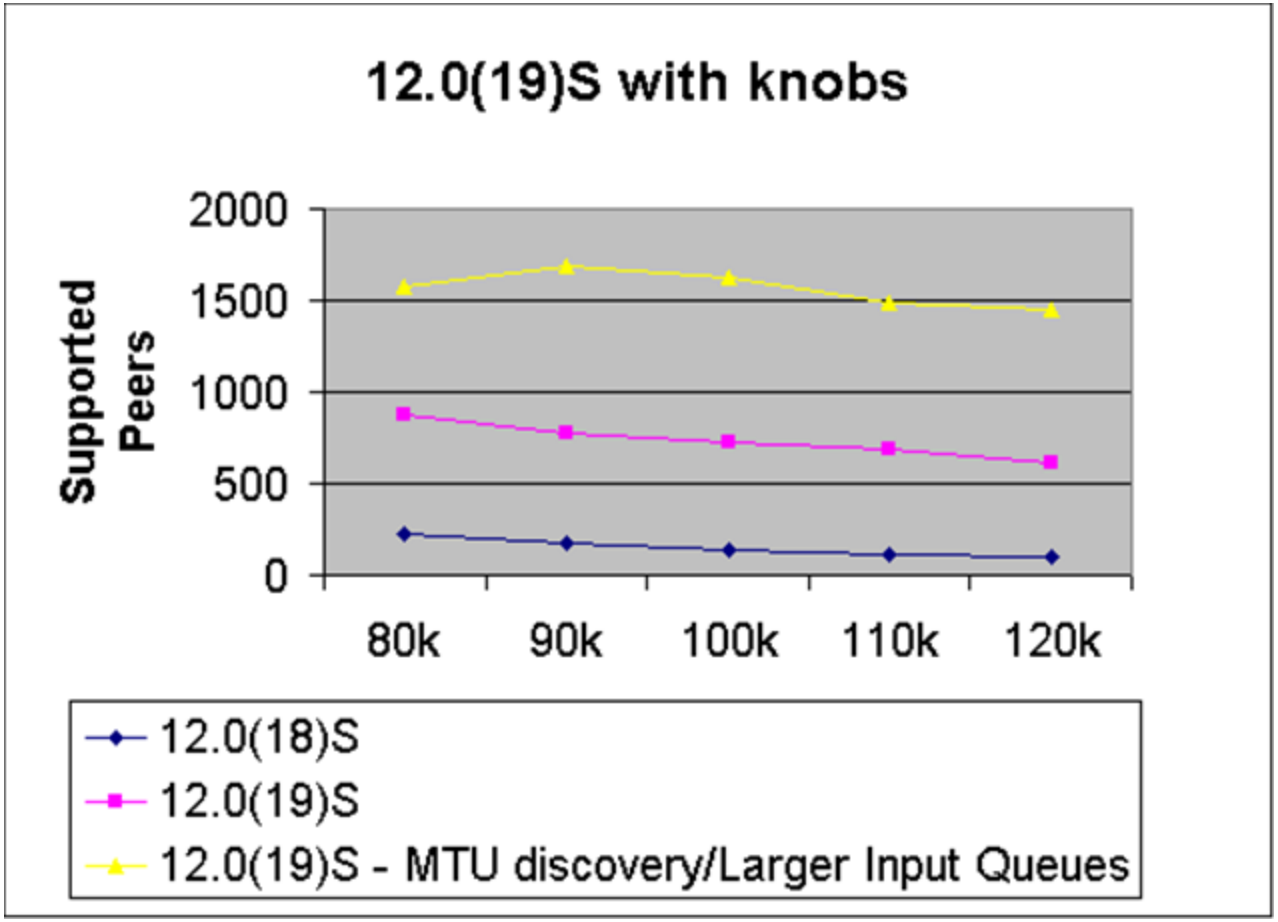
- **Improved update generation algorithm**
  - **100% update packing – attribute distribution no longer makes a significant impact**
  - **100% peer-group replication – no longer have to worry about peers staying synchronized**
  - **4x to 6x Improvement in convergence times**



© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Putting it all together

- **UPDATE packing with MTU discovery and Larger Input Queues provides a 14x improvement**



© 2002, Cisco Systems, Inc. All rights reserved.

BGP Route Convergence Optimizations
Summary

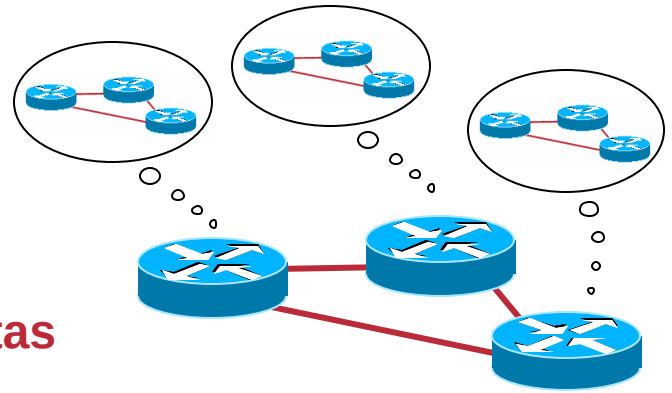
Cisco.com

- **Significant improvements can be gained just by using configuration knobs**
  - **Use peer-groups**
  - **Adjust Input Queues**
  - **Use Path MTU Discovery**
- **100% Update Packing and 100% Peer-Group Replication improves convergence times immensely**
- **Network upgrade requirements can be mitigated – enhancements are software-only**
- **No interoperability issues**

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda

- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas
Problem Description

- **Packets forwarded to a reloading router could be lost on a reloading BGP border router:**
  - **The IGP (OSPF or ISIS) may converge faster than BGP**
  - **Traffic may be forwarded to the reloading router when no next-hop is available**
- **Solution: Converge BGP before advertising transit routes via IGP**
  - **Router should advertise itself as reachable, but unavailable for transit until BGP converges**
  - **Advertise availability for transit traffic after BGP converges**
    - ✓ **Default IGP update delay is 2 minutes**
    - ✓ **IGP advertises transit routes after 10 minutes as a failsafe**

© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas
IS-IS – The Overload-Bit

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

© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

Preventing Black-Holes using IS-IS Overload-Bit

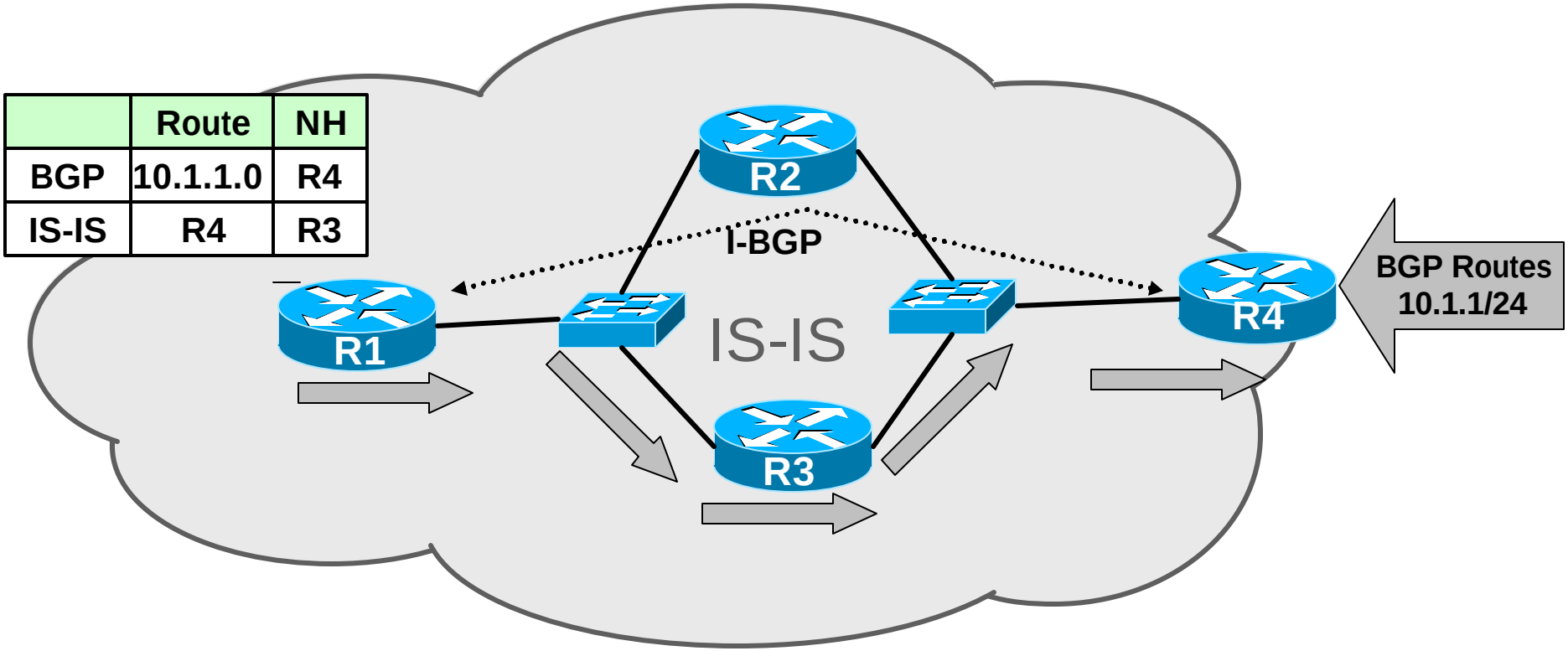
Cisco.com

- **IS-IS can set the Overload-Bit after each reboot**
  - **Wait for <time>, or for BGP to converge**
  - **Advertise as transit by unsetting the Overload-Bit**
- **Network administrator needs to specify how long IS-IS should wait for BGP to converge**
  - **set-overload-bit [on-startup {<time>|wait-for-bgp}]**
  - **Typically 2 to 5 minutes…**

© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

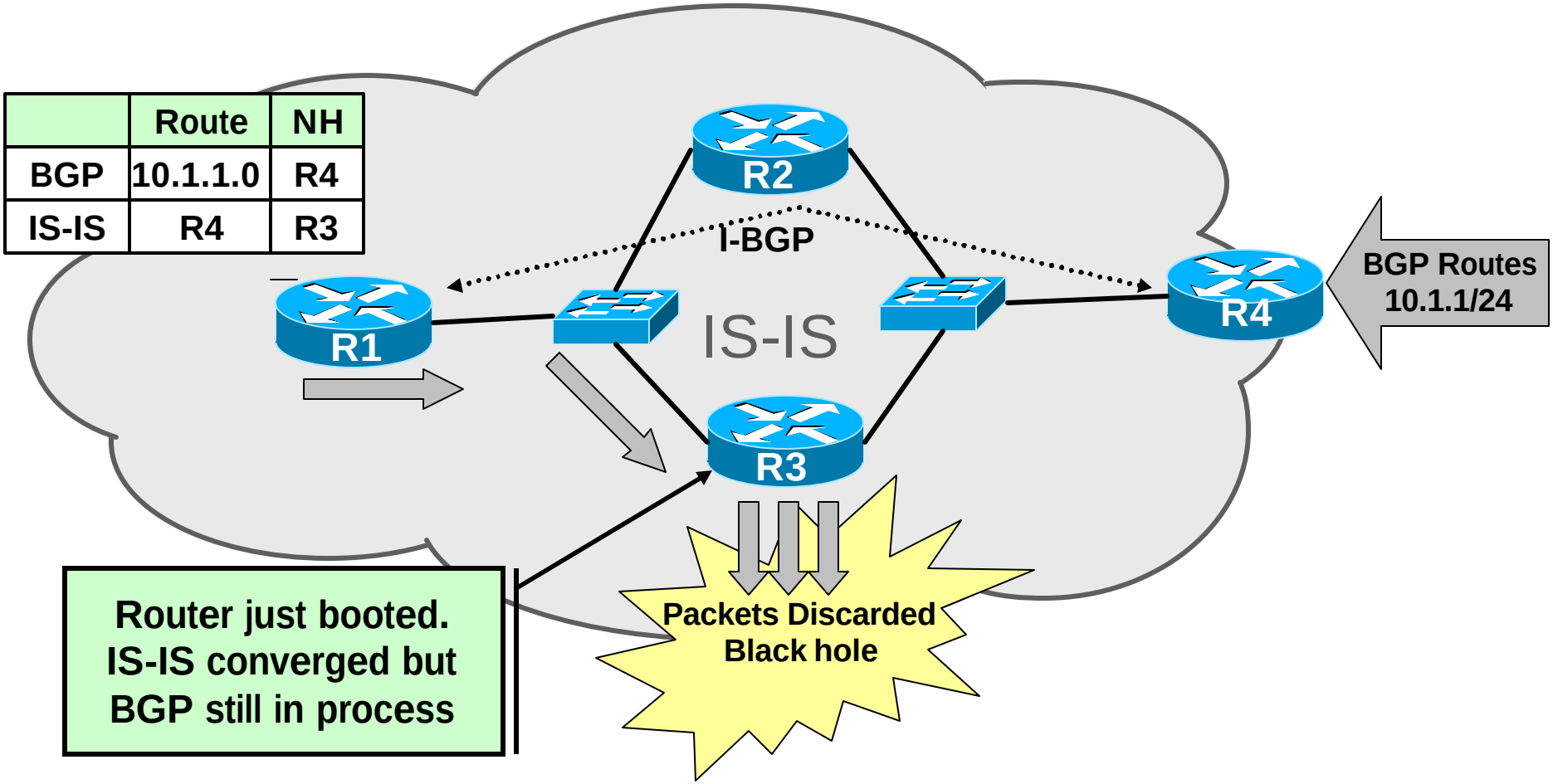
IS-IS Overload-Bit – BGP Example



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

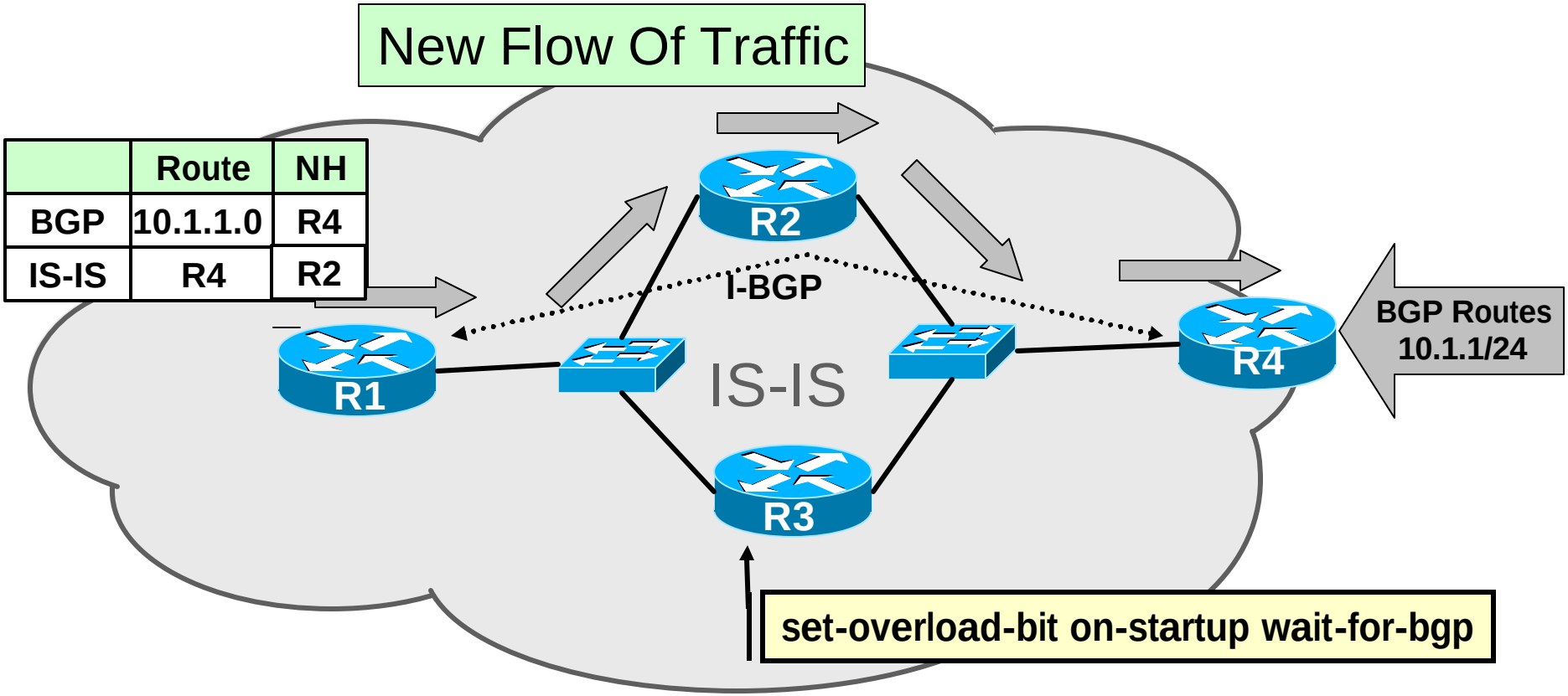
IS-IS Overload-Bit – BGP Example



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

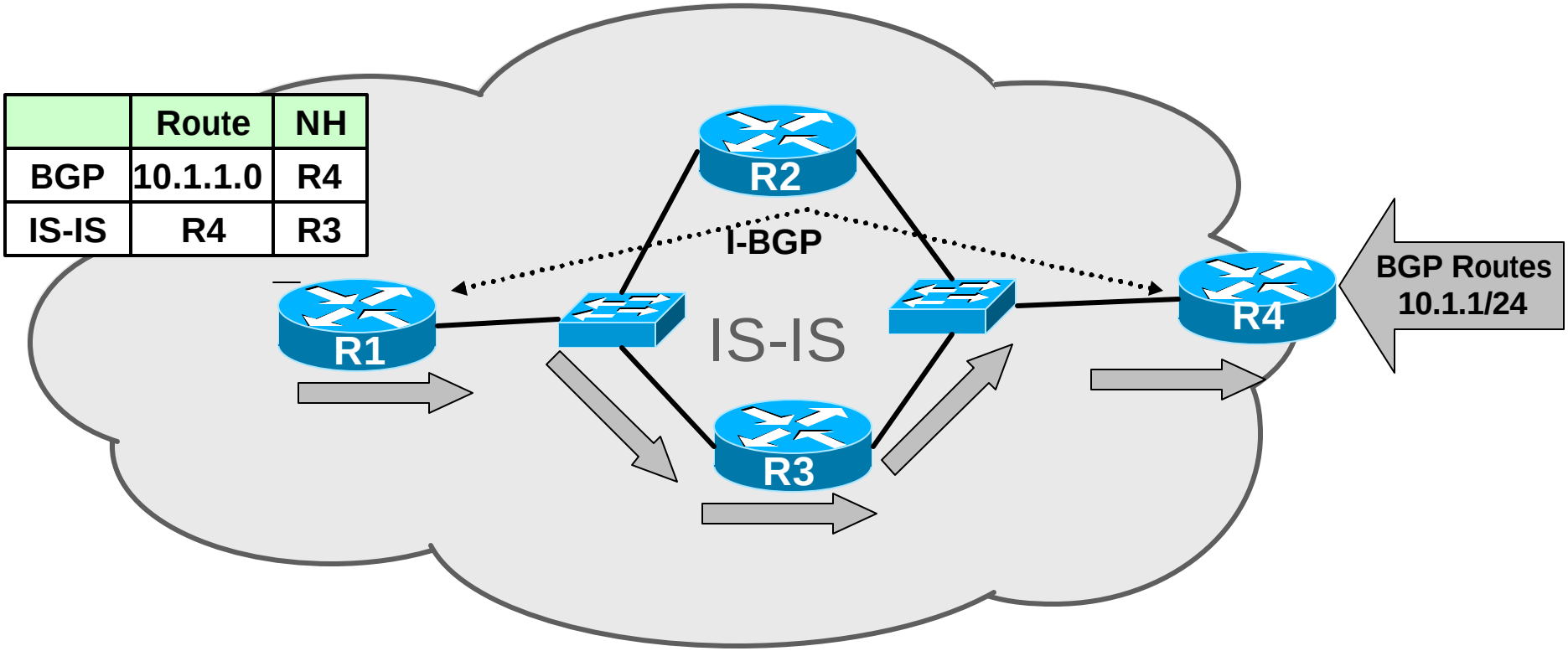
IS-IS Overload-Bit – BGP Example



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

IS-IS Overload-Bit – BGP Example



Traffic resumes normal path once BGP converges

© 2002, Cisco Systems, Inc. All rights reserved.

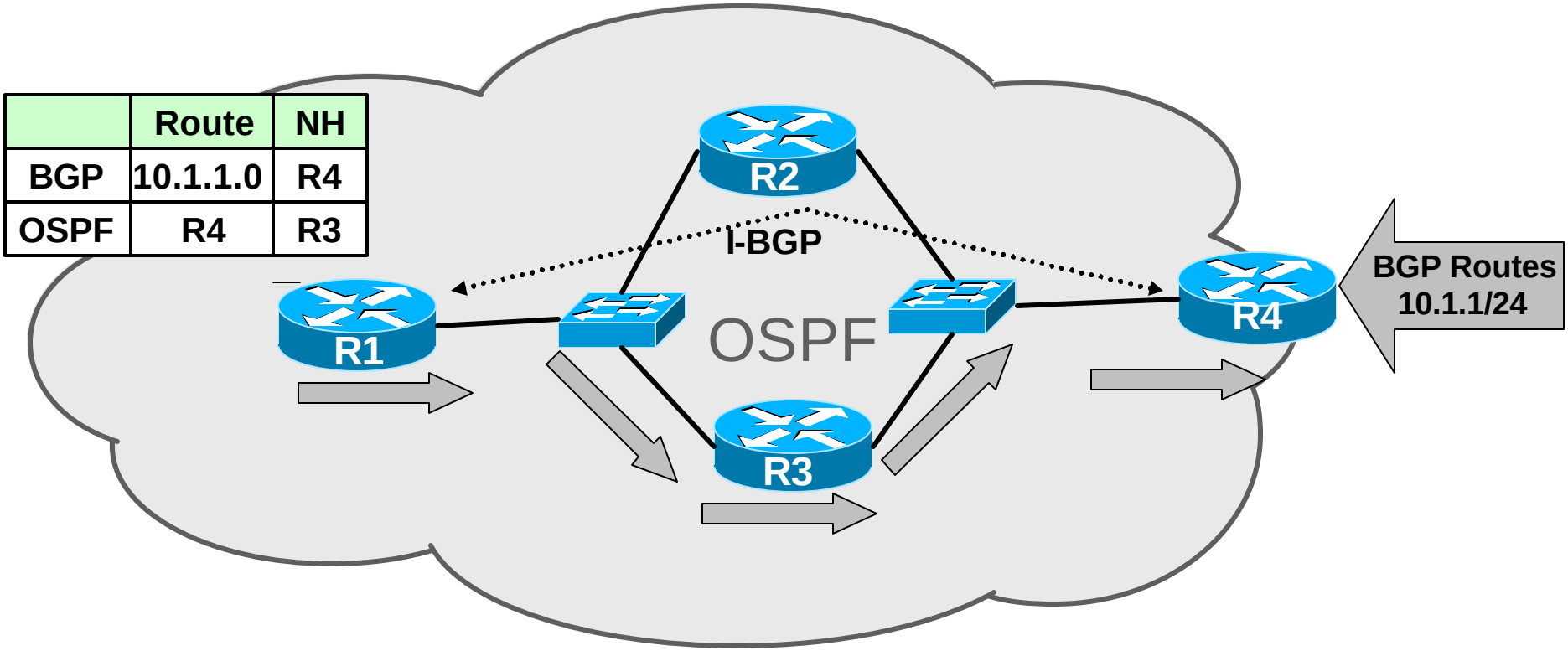
Controlling IGP/BGP Convergence Deltas
OSPF Router-LSAs at Infinity

- **RFC3137 describes a backward-compatible technique that an OSPF router can use to:**
  - **Advertise its unavailability to forward transit traffic to a set of destinations**
  - **Lower the preference level for the transit paths advertised as reachable through the router**
  - **All router link metrics within the router-LSA are set to infinity (0xffff) so it will NOT be used for transit**
- **LSAs 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 | <time>}]**

© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

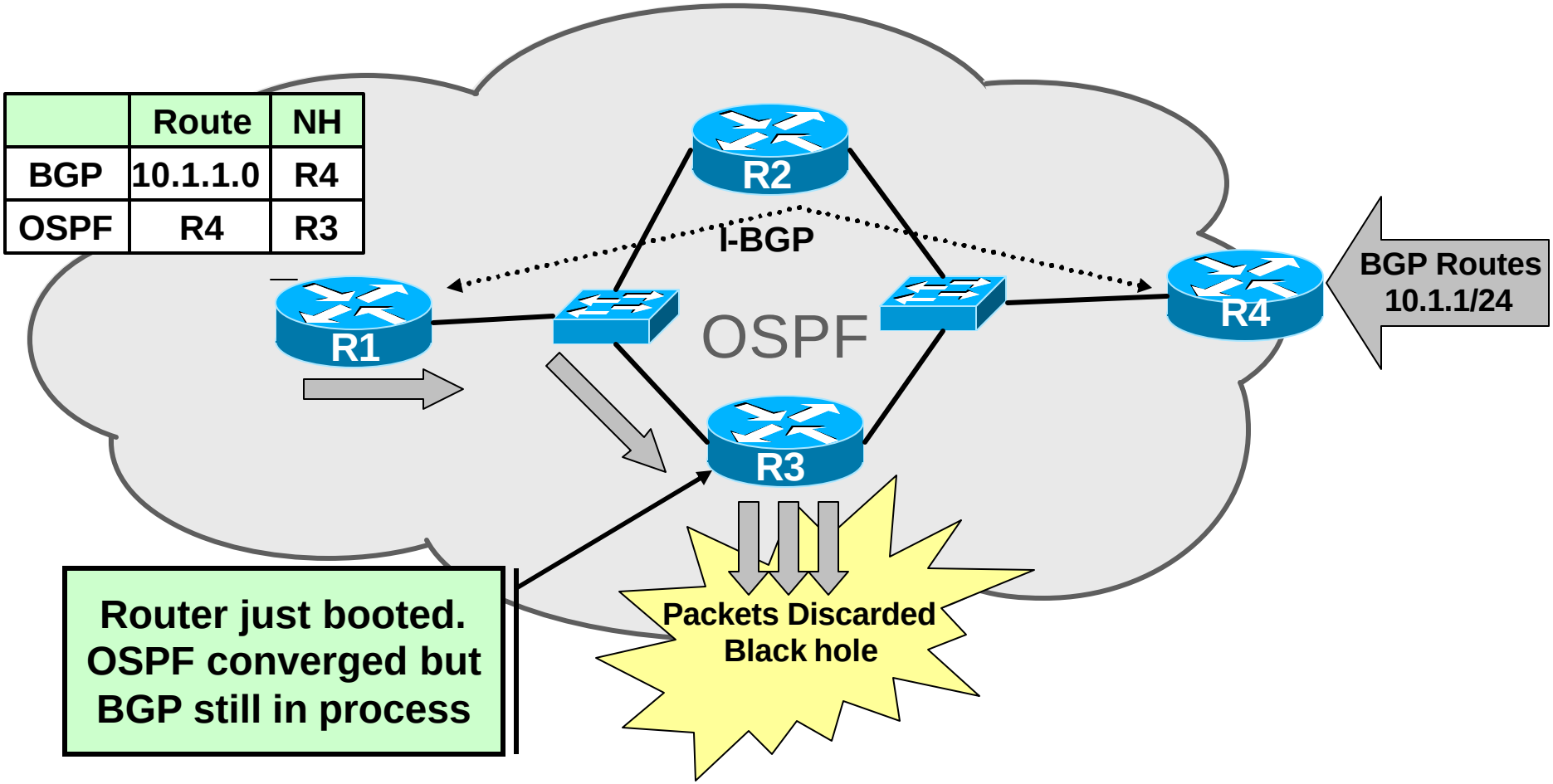
OSPF Router-LSAs at Infinity – BGP Example



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

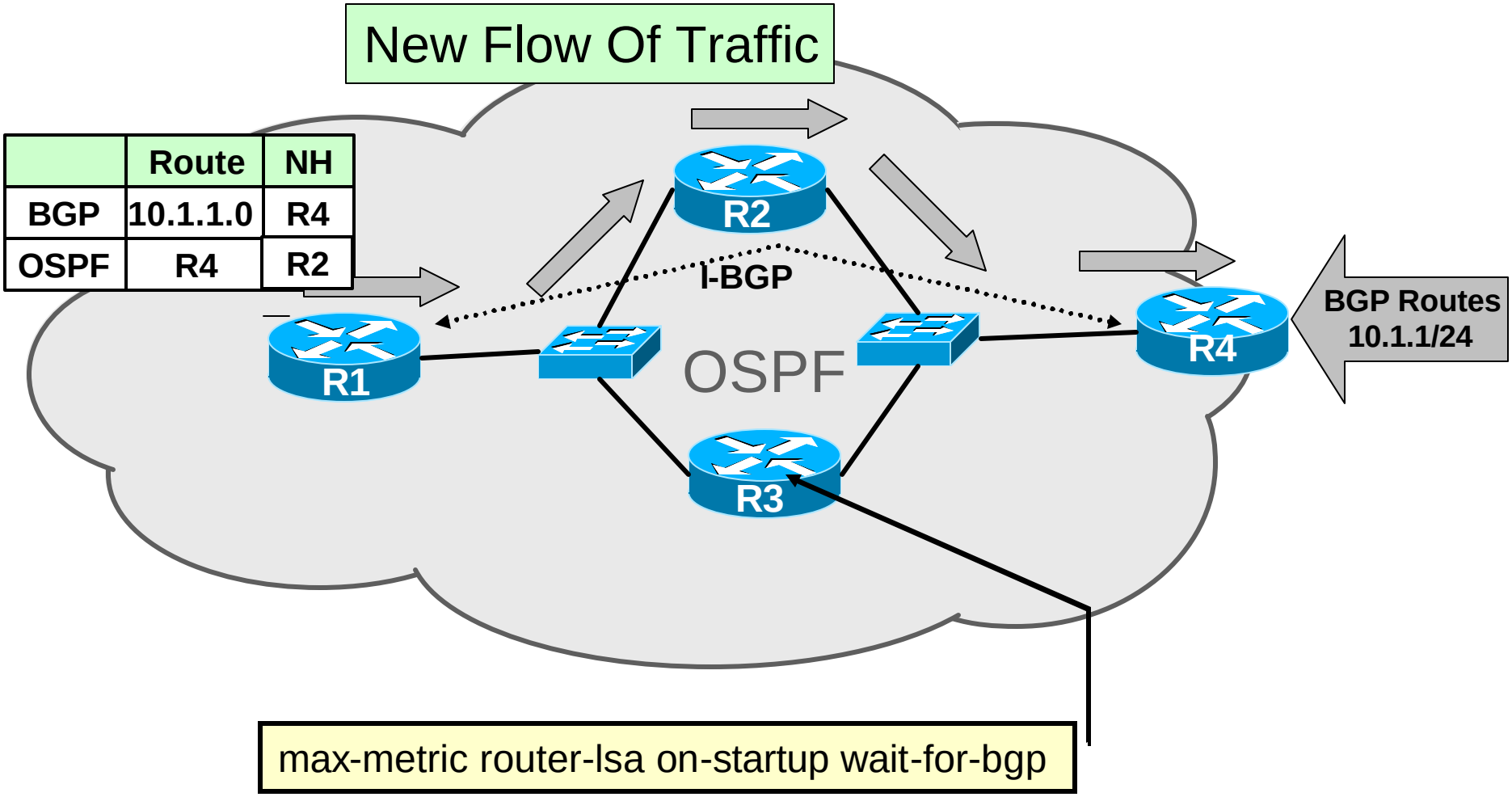
OSPF Router-LSAs at Infinity – BGP Example



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

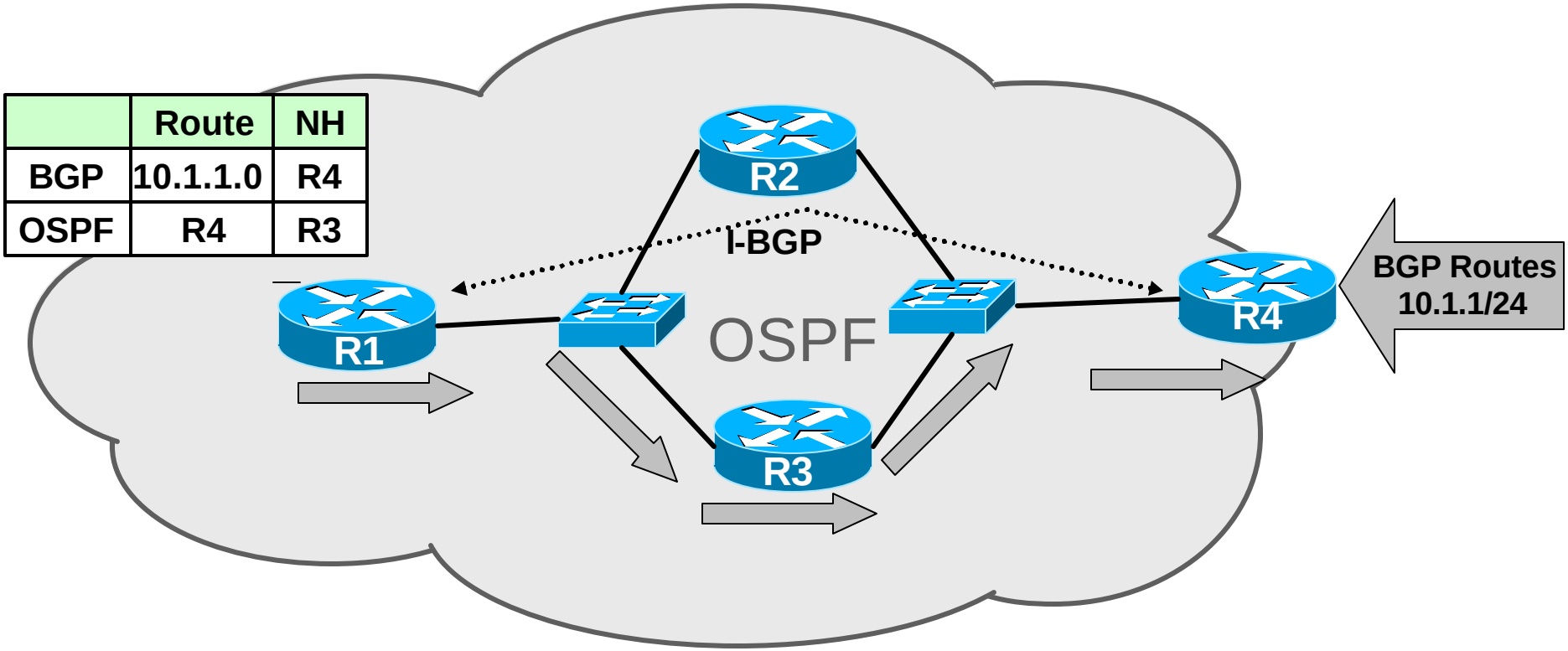
OSPF Router-LSAs at Infinity – BGP Example



© 2002, Cisco Systems, Inc. All rights reserved.

Controlling IGP/BGP Convergence Deltas

OSPF Router-LSAs at Infinity – BGP Example



Traffic resumes normal path once BGP converges

© 2002, Cisco Systems, Inc. All rights reserved.

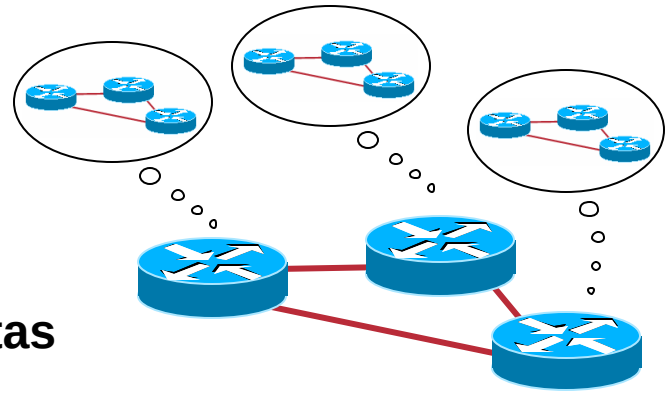
Controlling IGP/BGP Convergence Deltas
Summary

Cisco.com

- **Two standards-based ways to circumvent Black-Holes caused by differences in IGP and BGP convergence times**
  - **ISIS – set and clear the Overload-bit**
  - **OSPF – manipulate the Router-LSA metrics**
- **Available in IOS…**
  - **ISIS: 12.0(7)S, 12.1(9)E and 12.2(2)S, and above…**
  - **OSPF: 12.0(15)S, 12.1(8)E, and above…**

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda

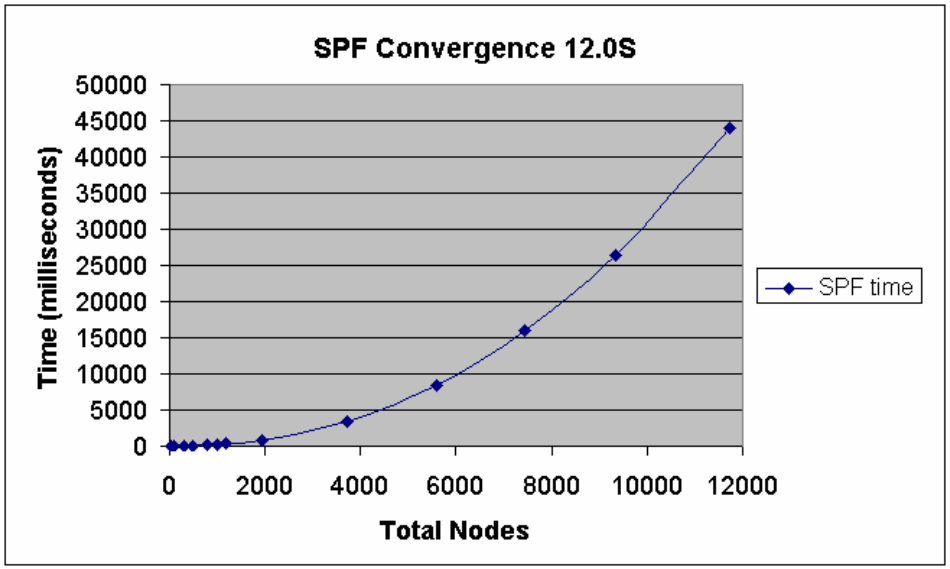


- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**

© 2002, Cisco Systems, Inc. All rights reserved.

Link-State IGP Performance Tuning
Performance Factors

- **Dijkstra's SPF Algorithm**
  - **Defines the topology in the form of a Shortest-Path Tree (SPT)**
  - **From the SPT we build the routing and forwarding tables**
  - **SPF runs are computationally expensive – control is required**
- **SPF performance depends on many factors**
  - **Number of routers – SPF computational cost is $F(n^{(\log n)})$ where n is the number of routers**
  - **Number of Links**
  - **Stability of adjacencies**
  - **Number of areas per ABR**
  - **Frequency of SPF**



© 2002, Cisco Systems, Inc. All rights reserved.

Link-State IGP Performance Tuning
LSA Generation and Flooding

- **LSA generation and flooding is important too!**
  - **Adjacency state changes – peer goes up/down**
  - **Interface state changes for connected IP subnets**
  - **Redistributed IP routes change**
  - **Inter-Area IP routes change**
  - **An interface is assigned a new metric**
  - **Periodic refreshes**
- **LSA Generation and Flooding Process**
  - **Generate the new LSA**
  - **Install it in your own LSPDB**
  - **Mark it for flooding**
  - **Flooding process sends LSA to all adjacencies**

© 2002, Cisco Systems, Inc. All rights reserved.

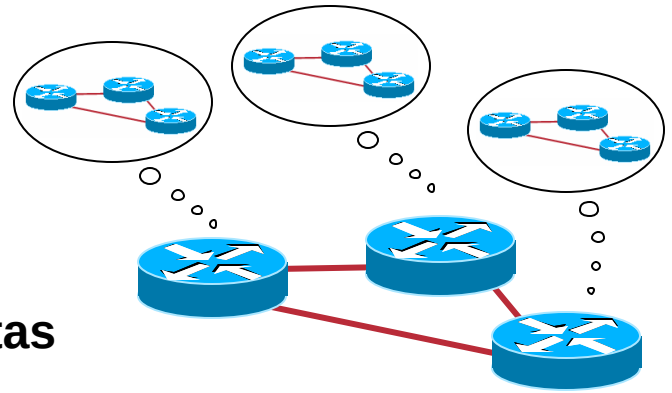
Link-State IGP Performance Tuning
Control SPF Runs and LSA Flooding

- **Controlling SPF Runs and LSA Flooding can improve stability and decrease resource requirements**
  - **Use SPF delay and hold timers to avoid excessive SPF calculations (or use Exponential Back-Off – next slides...)**
    - ✓ **OSPF: timers spf *spf-delay spf-holdtime***
  - **Use LSA Group-Pacing for refresh, checksum, and aging functions – avoids synchronization**
    - ✓ **OSPF: timers lsa-group-pacing**
  - **Use Do-Not-Age markers in stable topologies**
    - ✓ **OSPF: ip ospf flood-reduction**

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda

- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**



© 2002, Cisco Systems, Inc. All rights reserved.

IS-IS/OSPF Exponential Back-Off
Overview

Cisco.com

- **Exponential Back-Off may slow convergence but prevents IGP melt-downs by throttling SPF runs**
- **Exponential back-off is a compromise:**
  - **React fast to the first events**
  - **Under constant churn slow down to avoid a collapse**
- **Back-Off algorithm uses three timers**
  - **Maximum Interval – maximum amount of time the router will wait between consecutive SPF calculations**
  - **Initial Delay – time the router will wait before starting SPF calculations**
  - **Increment Interval – time to wait between consecutive SPF calculations. This timer is variable and increases to maximum-interval**

© 2002, Cisco Systems, Inc. All rights reserved.

IS-IS/OSPF Exponential Back-Off

Extended Syntax for IS-IS and OSPF

Cisco.com

- **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**
- **IS-IS Example: spf-interval 10 100 1000**
  - **After SPF is called, wait 100 msec, then run**
  - **Wait 1000 msec before running SPF a second time**
  - **Wait 2*1000 msec if before running SPF a third time**
  - **Wait 4*1000 msec before running SPF a fourth time**
  - **Wait a maximum of 10 seconds between SPF runs**
  - **Return to fast behaviour after no SPF run for 2*10 seconds**

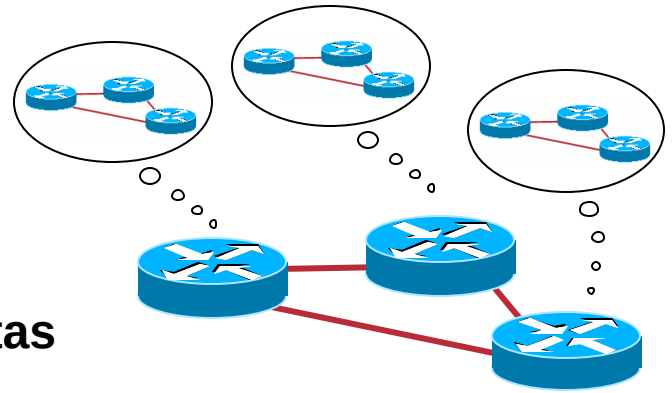
© 2002, Cisco Systems, Inc. All rights reserved.

IS-IS/OSPF Exponential Back-Off
Summary

- **Aggressive response to changes, then slow down if instability persists.**
  - **Bad news propagated faster than good news.**
  - **Helps maintain network stability!**
- **Cisco only solution today, but…**
  - **No interoperability issues because calculations done independently.**
  - **ISIS = 12.0(21)S and above**
  - **OSPF = planned for 12.0(25)S**
- **Know current network operating parameters before using – use nerd-knobs with care**

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda



- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **Non-Stop Forwarding**

© 2002, Cisco Systems, Inc. All rights reserved.

OSPF/IS-IS Incremental SPF
Background

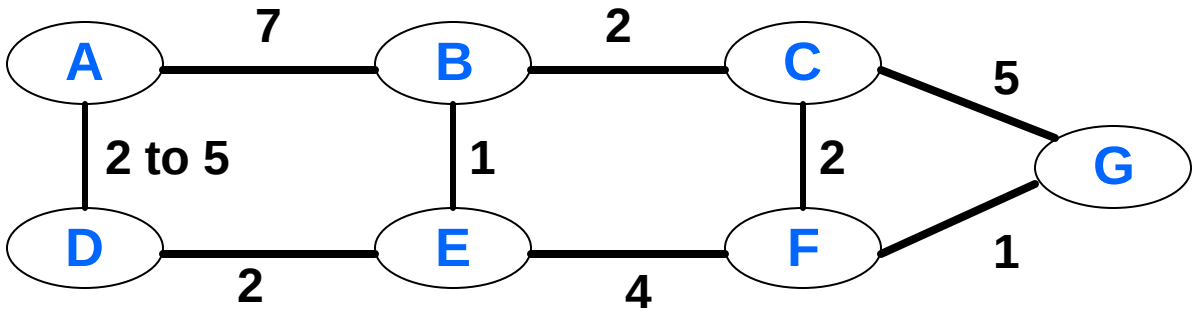
- SPF computation is triggered by the receipt of new LSAs/LSPs – including
  - Routers and Links Up/Down
  - Link cost changes
  - Adding a stub network
- The computation involves all routers in the same routing area or domain, however…
  - Some changes effect only specific branches of the SPT
  - Some changes do not effect the SPT at all
- Thus – the receipt of an LSA/LSP may not require running SPF on the whole SPT – or even at all!

© 2002, Cisco Systems, Inc. All rights reserved.

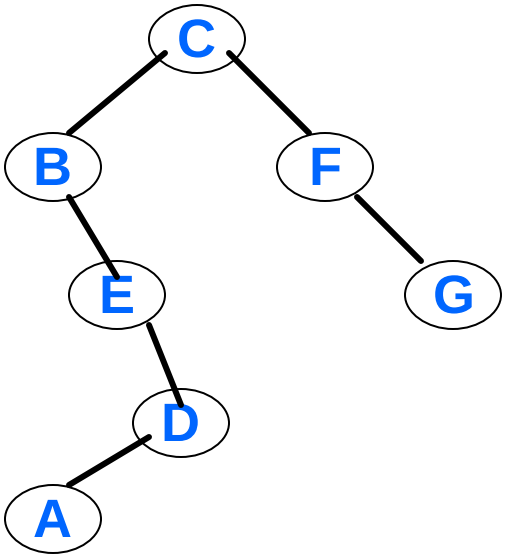
OSPF/IS-IS Incremental SPF

Example – Changes to a Stub Link

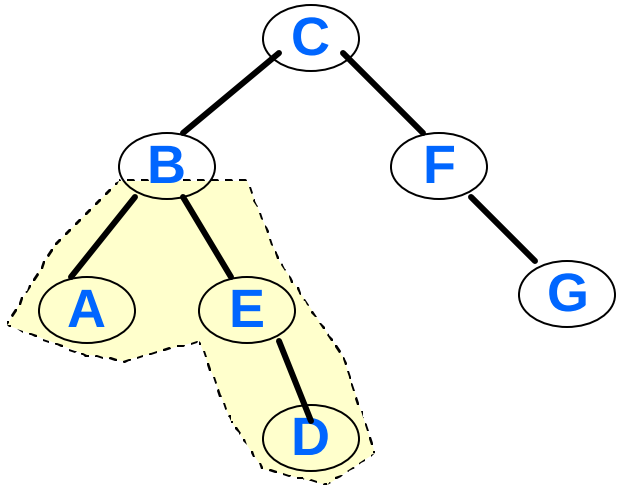
Routing Topology



Node C SPT before change



Node C SPT after Change



Changes to a stub link will not impact the whole SPT – but SPF will run anyway!

© 2002, Cisco Systems, Inc. All rights reserved.

OSPF/IS-IS Incremental SPF

Basis of Operation

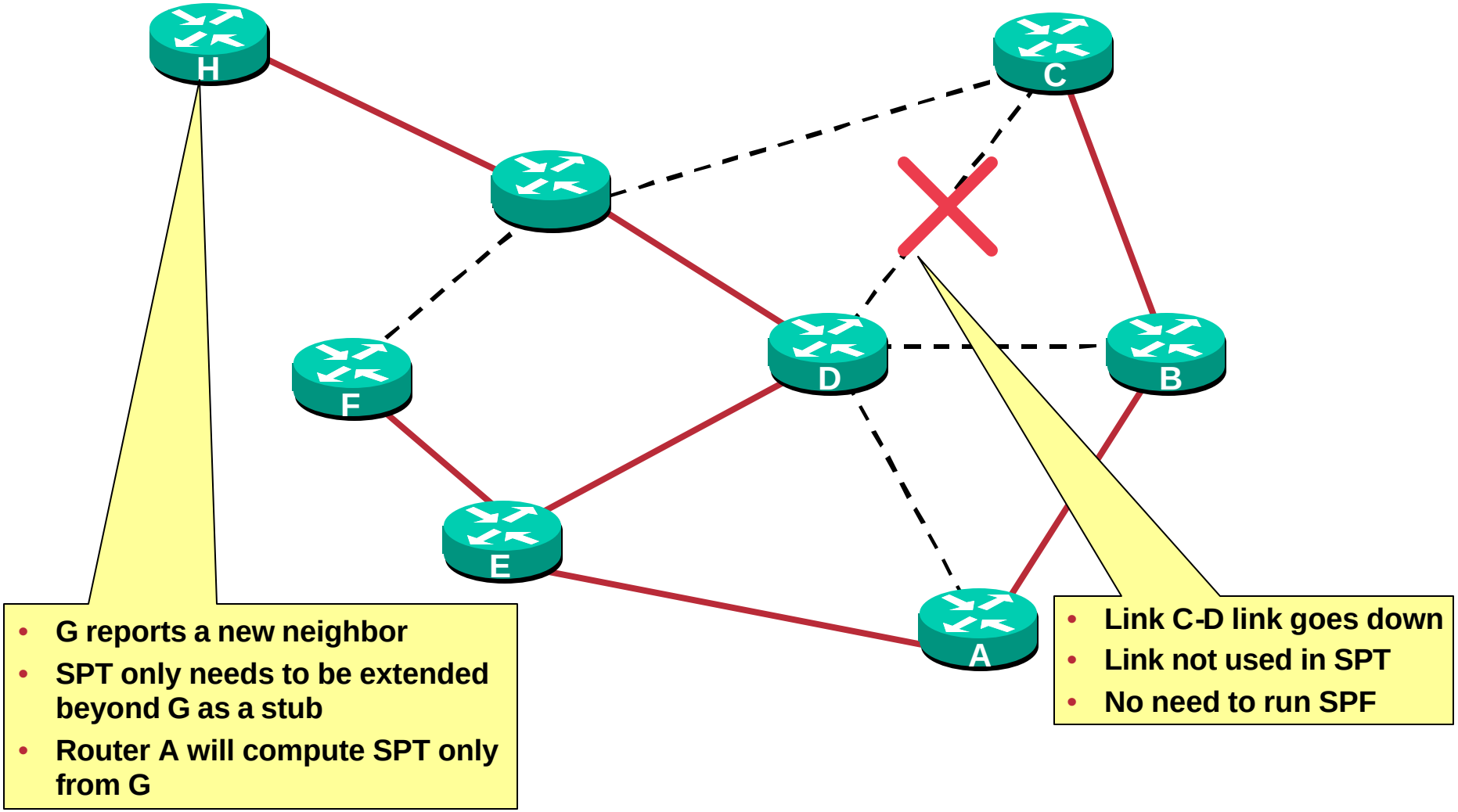
Cisco.com

- **Incremental SPF (iSPF) allows routers to**
  - **Intelligently determine the impact of a change to the SPT after receiving an LSA/LSP - based on parents and neighbors**
  - **Re-compute only the affected areas – modify instead of recalculate the SPT**
  - **Example – no need run SPF if a link that wasn't in the SPT is reported down**
- **Convergence times are reduced due to the decrease in SPF processing time**
- **Reductions in CPU usage and convergence times depend on**
  - **How much of the SPT can remain stable**
  - **In general – savings are proportional to the distance from the change**

© 2002, Cisco Systems, Inc. All rights reserved.

OSPF/IS-IS Incremental SPF

Examples – Unused Link Loss and Stub Changes



© 2002, Cisco Systems, Inc. All rights reserved.

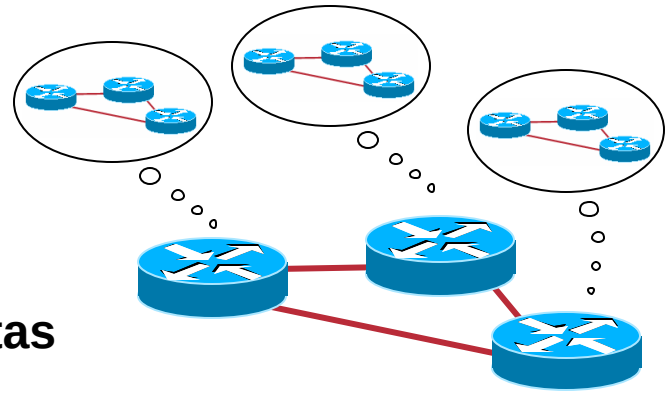
OSPF/IS-IS Incremental SPF
Summary

Cisco.com

- SPF is run only for affected areas of the SPT
  - Smaller scope of SPF runs result in faster convergence and less resource demand
  - Significant benefits for large topologies with many nodes or stubs
- Configuration
  - OSPF: incremental-spf
  - IS-IS: incremental-spf [level-1|level-2|level-1-2] [<1-100>]
    - ✓ [<1-100>] is the number of full Dijkstra runs before incremental SPF kick in

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda



- **Neighbor Maintenance**
  - **Fast HELLOs**
  - **MARP**
  - **BGP Initial Route Exchange**
  - **Controlling IGP/BGP Convergence Deltas**
- **Link-State IGP Tuning**
  - **SPF Runs and LSA Flooding**
  - **OSPF/ISIS Exponential Back-Off**
  - **OSPF/ISIS Incremental SPF**
  - **IP Event Dampening**
- **Network Design**
  - **High Availability**
  - **Non-Stop Forwarding**

© 2002, Cisco Systems, Inc. All rights reserved.

IP Event Dampening
Concept

Cisco.com

- **Apply the concept of BGP Route-Dampening to an interface so all routing protocols can benefit**
  - **Tracks interfaces and imposes a penalty to flapping interfaces**
  - **Puts an interface in a ‘down’ state – from routing protocol perspective – if the penalty is over a threshold**
  - **Uses an exponential decay algorithm to bring an interface back to an ‘up’ state after the penalty period**

© 2002, Cisco Systems, Inc. All rights reserved.

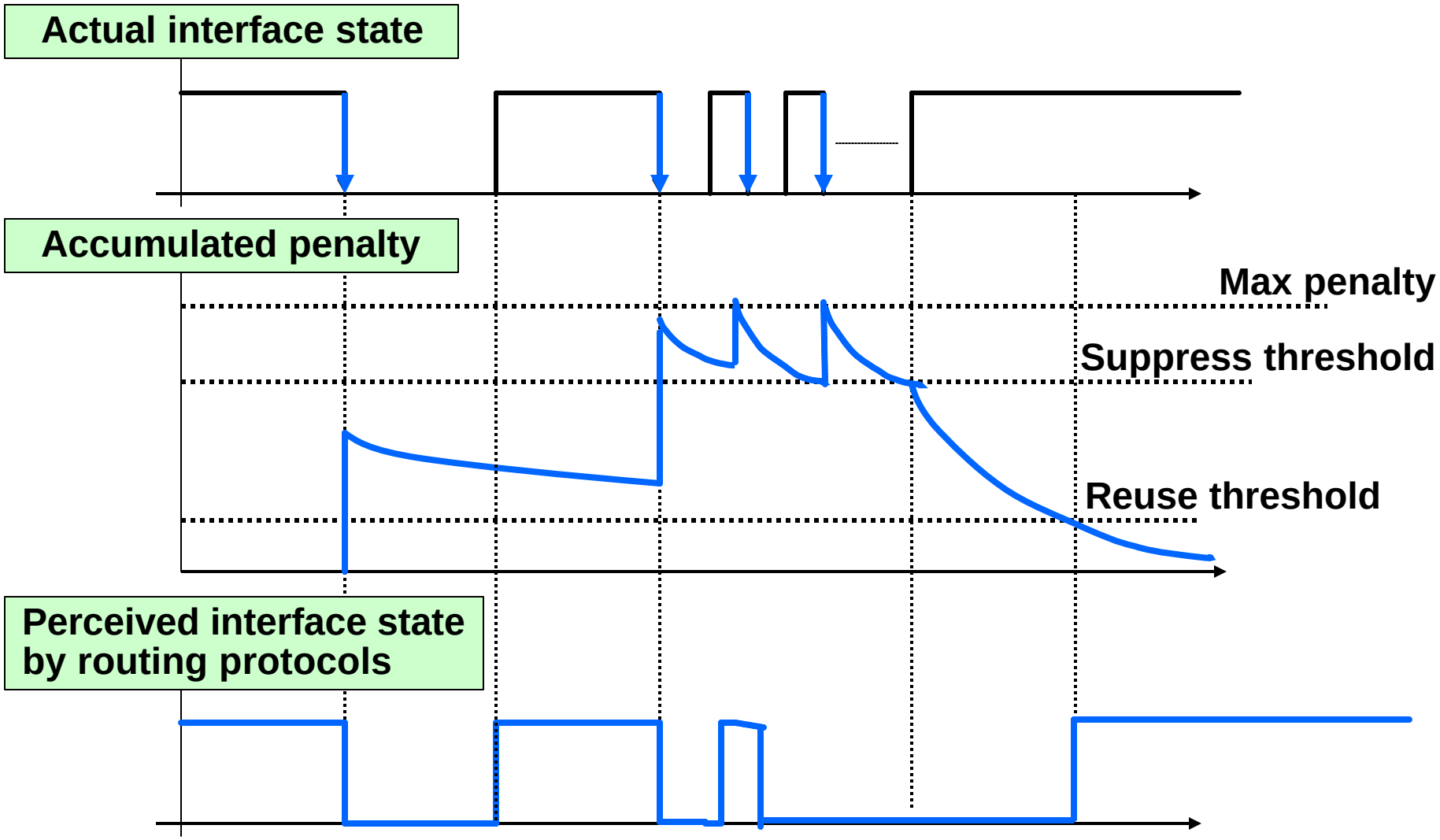
IP Event Dampening
Algorithm and Configuration

Interface: dampening <penalty> [half-life] [reuse suppress max-suppress] [restart <restart-penalty>]

- **penalty: value applied to the interface each time it flaps**
- **half-life: amount of time that must elapse without a flap to reduce penalty by half**
- **suppress: penalty threshold above which the interface is suppressed from a routing protocol perspective**
- **reuse: reactivate the interface when the penalty falls below this threshold**
- **max-suppress: maximum amount of time an interface can be suppressed**
- **restart-penalty: determines the initial penalty (if any) to be applied to an interface when system boots**

© 2002, Cisco Systems, Inc. All rights reserved.

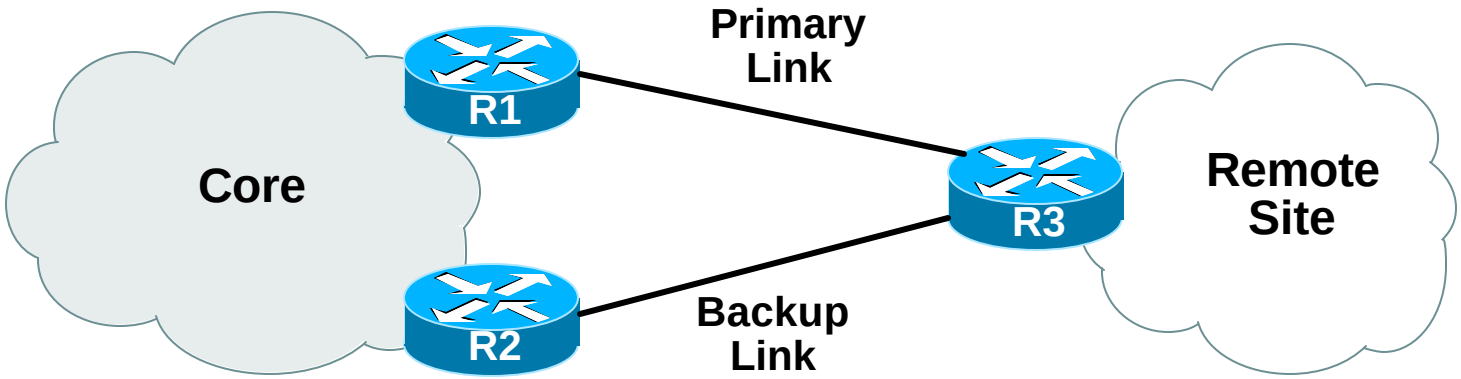
IP Event Dampening
Algorithm Behavior



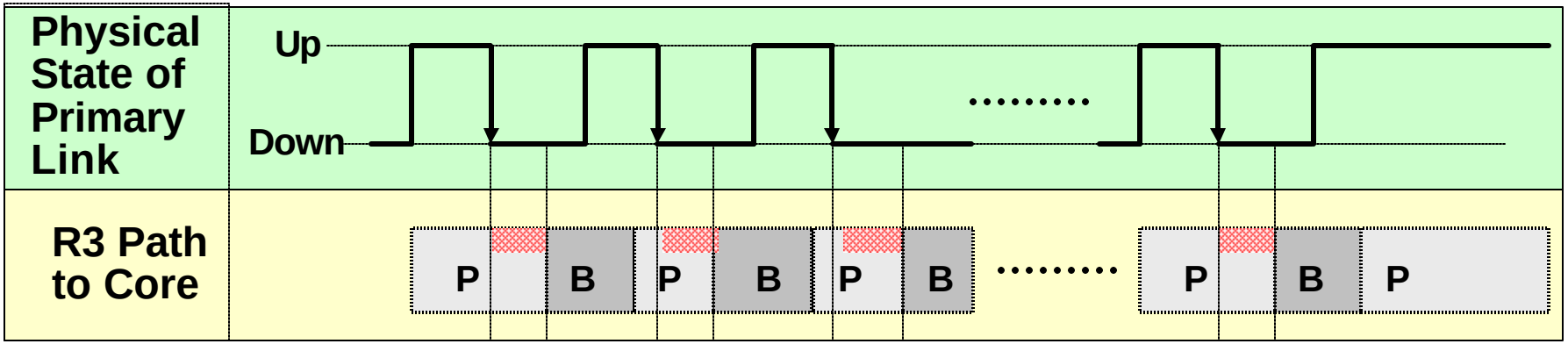
© 2002, Cisco Systems, Inc. All rights reserved.

IP Event Dampening

Deployment Example – Without Dampening



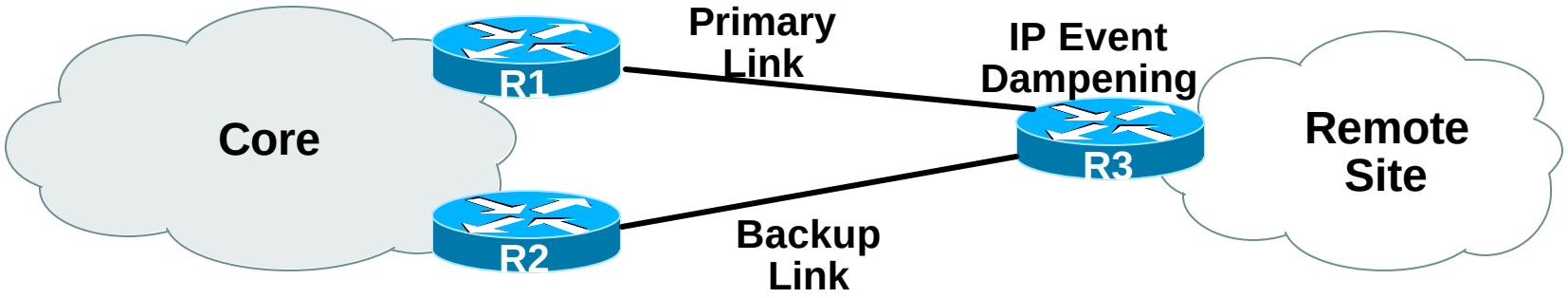
Link flapping causes routing re-convergence and packet loss



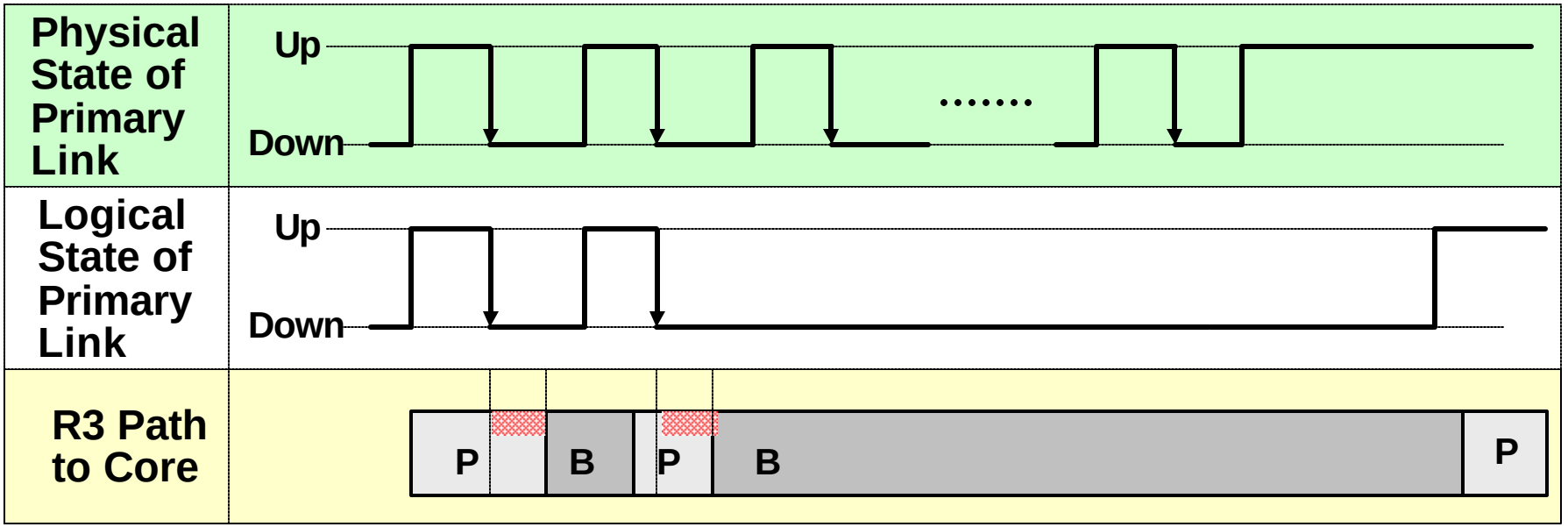
© 2002, Cisco Systems, Inc. All rights reserved.

IP Event Dampening

Deployment Example – With Dampening



IP Event Dampening absorbs link flapping effects on routing protocols



© 2002, Cisco Systems, Inc. All rights reserved.

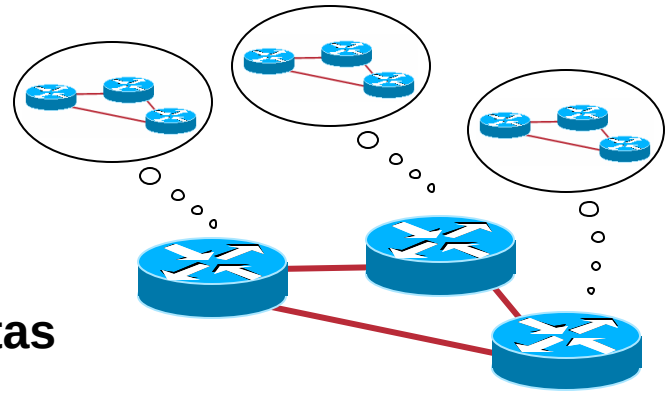
IP Event Dampening
Summary

Cisco.com

- **Prevents routing protocol churn caused by constant interface state changes**
- **Supports all IP routing protocols**
  - **Static Routing, RIP, EIGRP, OSPF, IS-IS, BGP**
  - **In addition, it supports HSRP and CLNS routing**
  - **Applies on physical interfaces and can't be applied on sub-interfaces individually**
- **Available in 12.0(22)S**

© 2002, Cisco Systems, Inc. All rights reserved.

Convergence Elements - Agenda



- **Neighbor Maintenance**
  - Fast HELLOs
  - MARP
  - BGP Initial Route Exchange
  - Controlling IGP/BGP Convergence Deltas
- **Link-State IGP Tuning**
  - SPF Runs and LSA Flooding
  - OSPF/ISIS Exponential Back-Off
  - OSPF/ISIS Incremental SPF
  - IP Event Dampening
- **Network Design**
  - Non-Stop Forwarding

© 2002, Cisco Systems, Inc. All rights reserved.

Non-Stop Forwarding
Description

- **A High-availability feature that allows for uninterrupted data forwarding during a switchover from an active to a standby processor**
  - **Active route processor synchronizes information with Standby route processor**
    - ✓ **Routing Tables**
    - ✓ **Forwarding Tables**
    - ✓ **Adjacency Information**
  - **Standby processor immediately takes control when the Active processor is compromised**
  - **Standby processor becomes and remains the Active processor**

© 2002, Cisco Systems, Inc. All rights reserved.

Non-Stop Forwarding
Impact on Routing Protocols

- **Switchovers must not impact traffic flow**
  - **Layer-2 adjacency information must be maintained**
  - **Layer-3 traffic must continue to be forwarded**
- **Switchovers must complete before Dead and Hold timers expire**
- **The FIB must remain unchanged during switchover**
  - **Routes are marked as stale during the switchover**
  - **Routes are re-marked as active once the flow routing information resumes**
  - **Transient routing loops or black-holes may be introduced if**
    - ✓ **The routing topology changes during the switchover**
    - ✓ **The FIB is updated before routes are re-marked as active**
- **Adjacencies must not be reset by either side when switchover is complete**

© 2002, Cisco Systems, Inc. All rights reserved.

Non-Stop Forwarding
Restart Processing

Cisco.com

- **During restart or switchover, the router indicates it is restarting to its peers**
- The peers then implement the following strategy
  - **Containment: The peers will maintain adjacencies and will not inform any other peers of the restart event**
  - **Perseverance: All prefixes that depended on the restarting router are marked as stale but are still *used for forwarding***
  - **Sharing: After the restart the peers re-exchange all their routing information**
  - **Concluding: The re-synchronization process must be finite**
    - ✓ **Both routers should know when the route-exchange process is finished so routes can be re-marked as active**

© 2002, Cisco Systems, Inc. All rights reserved.

Non-Stop Forwarding
What If?

Cisco.com

- **The restarting router is removed from the network or doesn't restart in time?**
  - **The Dead-Timers on the peers expire**
  - **Normal re-convergence occurs**
- **The re-convergence process takes too long?**
  - **The Stale-Path timers expire**
  - **Stale prefixes are cleared**
  - **"Re-convergence occurs**
- **Something else changes in the network while reconverging?**
  - **Worse case – a transient black-hole or routing loop may exist until the exchange of routing information can complete**
  - **Typically – a period of sub-optimal routing will occur - but not black-holes or loops**

© 2002, Cisco Systems, Inc. All rights reserved.

Non-Stop Forwarding
Summary

Cisco.com

	BGP	EIGRP	ISIS	OSPF
Restart Capability Indication	X		X	
Restart Indication	X Implicit in TCP	X	X	X
Forwarding State	X	Implicit	Implicit	Implicit
Peer Requirements	All iBGP	None	None	All
Delayed Updates	X	X	X	X

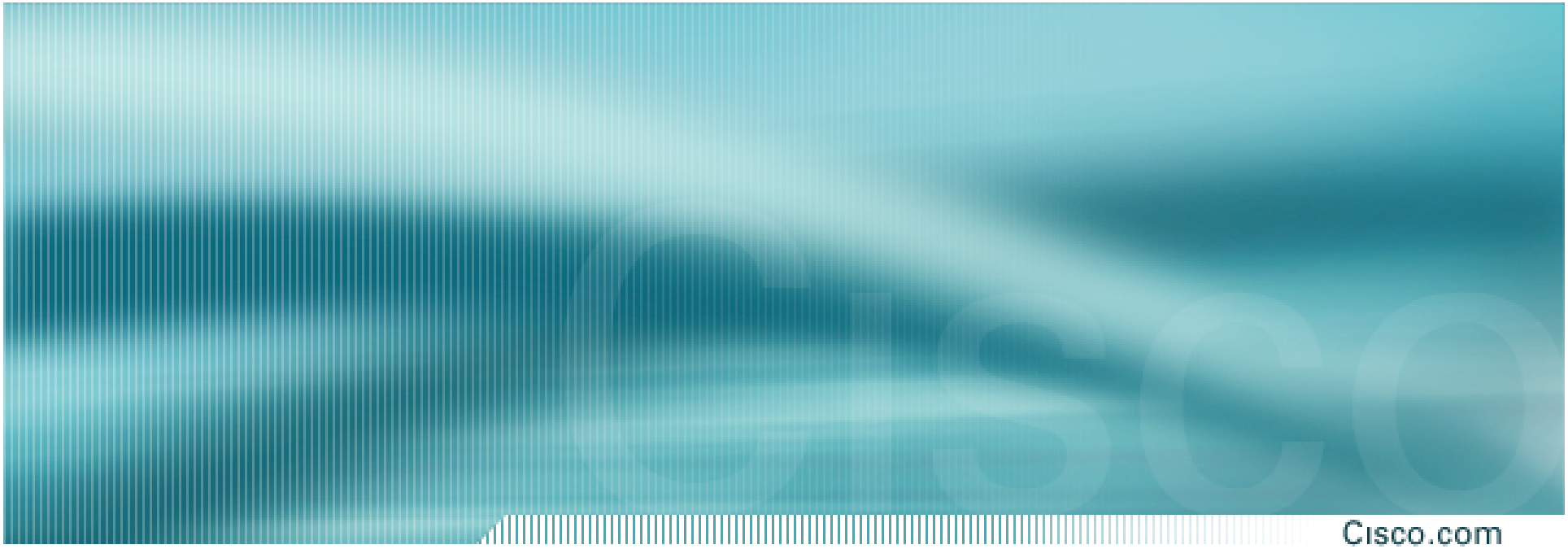
© 2002, Cisco Systems, Inc. All rights reserved.

Summary

Cisco.com

- The design of the network plays a VERY important role in the performance of the routing protocols
- Ability to change the protocol is tied to the specification, implementation and the deployed base of routers and software
- Chose the IGP that better meets your overall needs – convergence is just one part of it
- Use knobs with care

© 2002, Cisco Systems, Inc. All rights reserved.



Thanks!

© 2002, Cisco Systems, Inc. All rights reserved.