

MPLS for Dummies

Richard A Steenbergen <ras@nlayer.net> nLayer Communications, Inc.

Purpose of This Tutorial

- There are lot of IP people out there who still don't like MPLS.
 - Many of the concepts are completely foreign to pure IP networks.
 - Many parts of MPLS smell like ATM, a technology which did a lot of things wrong as it was applied to the IP world.
 - Many aspects of MPLS could be called overly complicated, or at least have been presented in an overly complicated way in the past.
 - Even networks who claim to run MPLS networks often have only the most basic features turned on, and may not fully utilize it.
- But, MPLS can be a powerful tool for any network.
 - It's not just for the buzzword compliant or the crazy telco-heads.
- With any luck, this tutorial should:
 - Introduce the concepts of MPLS for people who are new to it.
 - Show you how MPLS can help you run your network better.

Target Audience



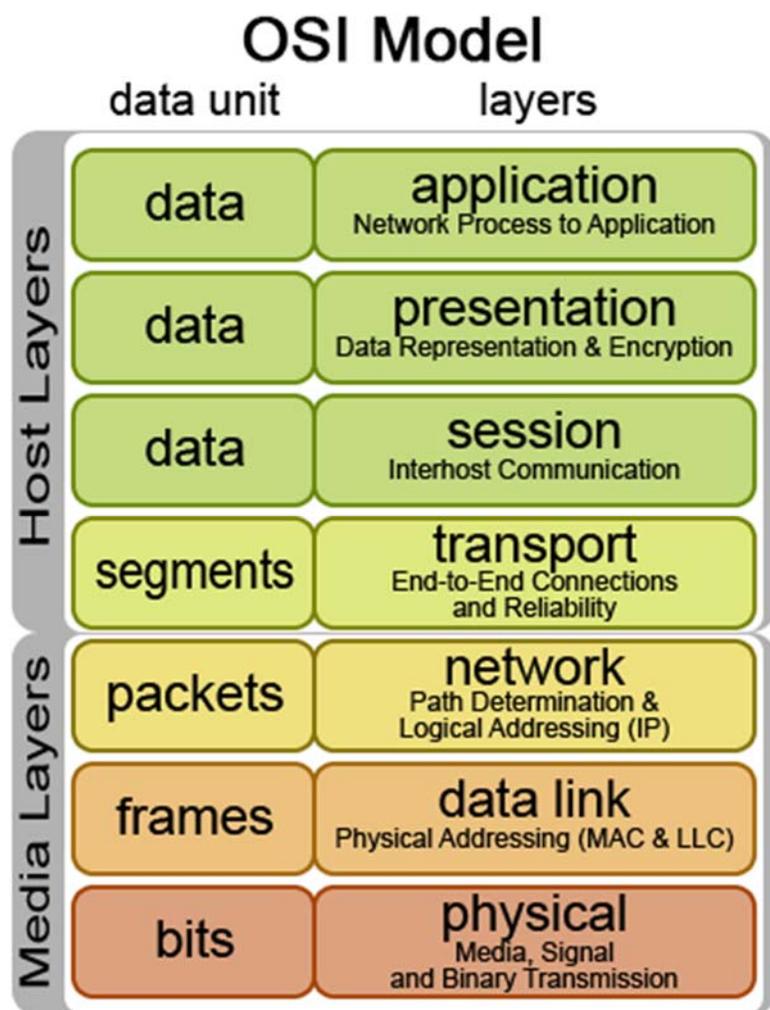
MPLS Isn't ATM 2.0, I Promise



The Basics

What is MPLS?

- MPLS stands for “Multi-Protocol Label Switching”.



MPLS is best summarized as a “Layer 2.5 networking protocol”.

In the traditional OSI model:

- Layer 2 covers protocols like Ethernet and SONET, which can carry IP packets, but only over simple LANs or point-to-point WANs.
- Layer 3 covers Internet-wide addressing and routing using IP protocols.
- MPLS sits between these traditional layers, providing additional features for the transport of data across the network.

What is Label Switching?

- In a traditional IP network:
 - Each router performs an IP lookup (“routing”), determines a next-hop based on its routing table, and forwards the packet to that next-hop.
 - Rinse and repeat for every router, each making its own independent routing decisions, until the final destination is reached.
- MPLS does “label switching” instead:
 - The first device does a routing lookup, just like before:
 - But instead of finding a next-hop, it finds the final destination router.
 - And it finds a pre-determined path from “here” to that final router.
 - The router applies a “label” (or “shim”) based on this information.
 - Future routers use the label to route the traffic
 - Without needing to perform any additional IP lookups.
 - At the final destination router the label is removed.
 - And the packet is delivered via normal IP routing.

What is the Advantage of Label Switching?

- Originally, it was intended to reduce IP routing lookups.
 - When CIDR was introduced, it had unintended consequences.
 - CIDR introduced the concept of “longest prefix matching” for IP routing.
 - Longest prefix match lookups have historically been very difficult to do.
 - The classic software algorithm for routing lookups was called a PATRICIA trie, which required many memory accesses just to route a single packet.
 - Exact matches were comparatively much easier to implement in hardware.
 - Most early hardware routing “cheated” by doing the first lookup in software, then did hardware-based exact matching for future packets in the “flow”.
 - Label switching (or “tag switching”) lookups use exact matching.
 - The idea was to have only the first router do an IP lookup, then all future routes in the network could do exact match “switching” based on a label.
 - This would reduce load on the core routers, where high-performance was the most difficult to achieve, and distribute the routing lookups across lower speed edge routers.

What is the Advantage of Label Switching?

- Modern ASICs have eliminated this issue... Mostly.
 - Today, commodity ASICs can do many tens of millions of IP routing lookups per second, relatively cheaply and easily.
 - However, they still make up a significant portion of the cost of a router.
 - Exact matching is still much cheaper and easier to implement.
 - A layer 2 only Ethernet switch (which does exact matching) may be 1/4th the cost and 4x the capacity of a similar device with layer 3 capabilities.
- So why do people still care about MPLS? Three reasons:
 - Implementing Traffic-Engineering
 - The ability to control where and how traffic is routed on your network, to manage capacity, prioritize different services, and prevent congestion.
 - Implementing Multi-Service Networks
 - The ability to deliver data transport services, as well as IP routing services, across the same packet-switched network infrastructure.
 - Improving network resiliency with MPLS Fast Reroute.

How MPLS Works

How MPLS Works – Basic Concepts

- **MPLS Label Switched Path (“LSP”)**
 - One of the most important concepts for the actual use of MPLS.
 - Essentially a unidirectional tunnel between a pair of routers, routed across an MPLS network.
 - An LSP is required for any MPLS forwarding to occur.
- **MPLS Router Roles/Positions**
 - **Label Edge Router (“LER”) or “ingress node”**.
 - The router which first encapsulates a packet inside an MPLS LSP.
 - Also the router which makes the initial path selection.
 - **Label Switching Router (“LSR”) or “transit node”**
 - A router which only does MPLS switching in the middle of an LSP.
 - **Egress Node**
 - The final router at the end of an LSP, which removes the label.

How MPLS Works – Basic Concepts

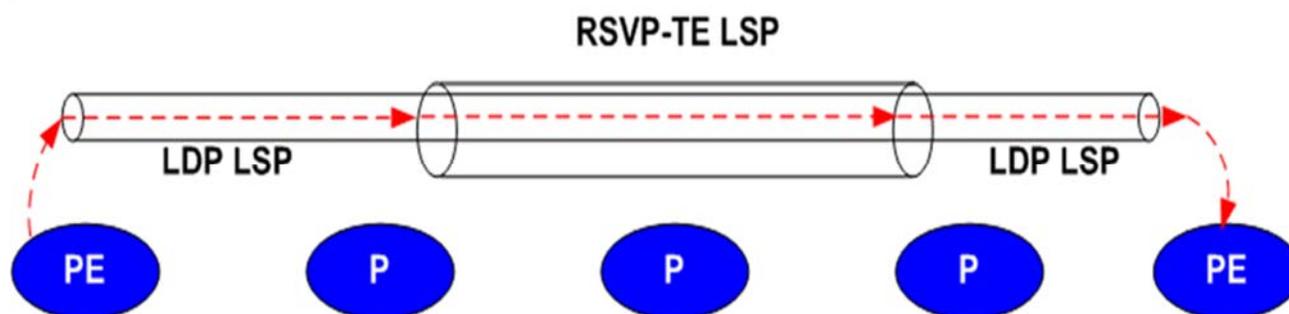
- MPLS router roles may also be expressed as “P” or “PE”:
 - Terms which come from the description of VPN services.
 - P – Provider Router
 - A core/backbone router which is doing label switching only.
 - A pure P router can operate without any customer/Internet routes at all.
 - This is common in large service provider networks.
 - PE – Provider Edge Router
 - A customer facing router which does label popping and imposition.
 - Typically has various edge features for terminating multiple services:
 - Internet
 - L3VPN
 - L2VPN / Pseudowires
 - VPLS
 - CE is the “Customer Edge”, the customer device a PE router talks to.

MPLS Signaling Protocols

- To use an LSP, it must be signaled across your routers.
 - An LSP is a network-wide tunnel, but a label is only a link-local value.
 - An MPLS signaling protocol maps LSPs to specific label values.
 - There are two main MPLS routing protocols in use today:
 - Label Distribution Protocol (“LDP”)
 - A simple non-constrained (doesn’t support traffic engineering) protocol.
 - Resource Reservation Protocol with Traffic Engineering (“RSVP-TE”)
 - A more complex protocol, with more overhead, but which also includes support for traffic-engineering via network resource reservations.
 - Most complex networks will actually need to use both protocols.
 - LDP is typically used by MPLS VPN (data transport) services.
 - But RSVP-TE is necessary for traffic engineering features.
 - Most networks will configure LDP to tunnel inside RSVP.

MPLS Label Stacking

- MPLS labels can also be stacked multiple times.
 - The top label is used to control the delivery of the packet.
 - When destination is reached, the top label is removed (or “popped”), and the second label takes over to direct the packet further.
- Some common stacking applications are:
 - VPN/Transport services, which use an inner label to map traffic to specific interfaces, and an outer label to route through the network.
 - “Bypass” LSPs, which can protect a bundle of other LSPs to redirect traffic quickly without having to completely re-signal every LSP, in the event of a router failure.



Penultimate Hop Popping

- There are two ways to terminate an LSP:
 - Implicit Null
 - Also called “Penultimate Hop Popping” (PHP).
 - Just a long way of saying “remove the label on the next-to-last hop”.
 - Explicit Null
 - Preserve the label all the way to the very last router.
- What’s the difference?
 - Implicit null is an optimization technique.
 - Since the label is already removed on the next-to-last router, the last router can more easily begin to route the packet after it exits the LSP.
 - Otherwise, the packet has to make “two trips” through the last router.
 - One pass through the forwarding path to pop the label.
 - Another pass to route the packet based on the underlying information.

Vendor Terminology Warning

- Cisco and Juniper both use somewhat confusing terms to describe the same thing.
- Example:
 - Cisco Affinities Juniper Admin-Groups
 - Cisco Autoroute Announce Juniper TE Shortcuts
 - Cisco Forwarding Adjacency Juniper LSP-Advertise
 - Cisco Tunnel Juniper LSP
 - Cisco Make-Before-Break Juniper Adaptive
 - Cisco Application-Window Juniper Adjust-Interval
 - Cisco Shared Risk Link Groups Juniper Fate-Sharing

MPLS Traffic Engineering

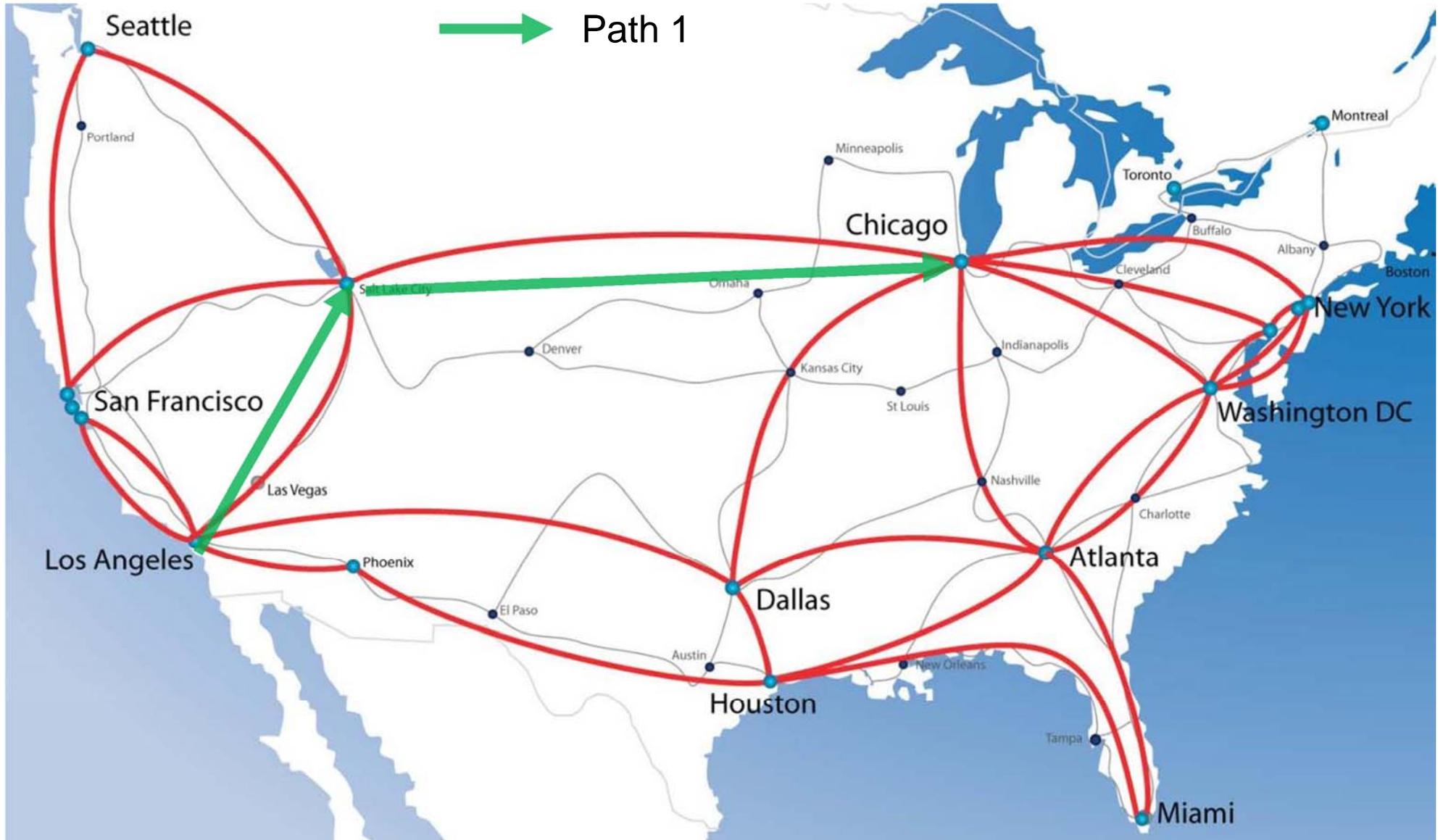
What is Traffic Engineering

- What is Traffic Engineering?
 - Classic IGPs use non-TE routing, i.e. a metric (cost) per link, and a shortest path first (SPF) algorithm to find the shortest path.
 - Traffic Engineering takes this, and adds additional constraints.
 - For example, find the shortest path that also has available bandwidth.
 - This is also called constrained routing, using a CSPF algorithm.
 - The principal is simple: It is better to take an uncongested path even though the latency may be higher, than to congest the shortest path on one link while leaving available bandwidth unused on another link.
- Why can't I just do this manually with my IGP costs?
 - You can, but this only scales up to a certain point.
 - As networks become more complex, this gets harder to manage.
 - Changing an IGP cost by 1 can easily affect routing dozens of hops away.

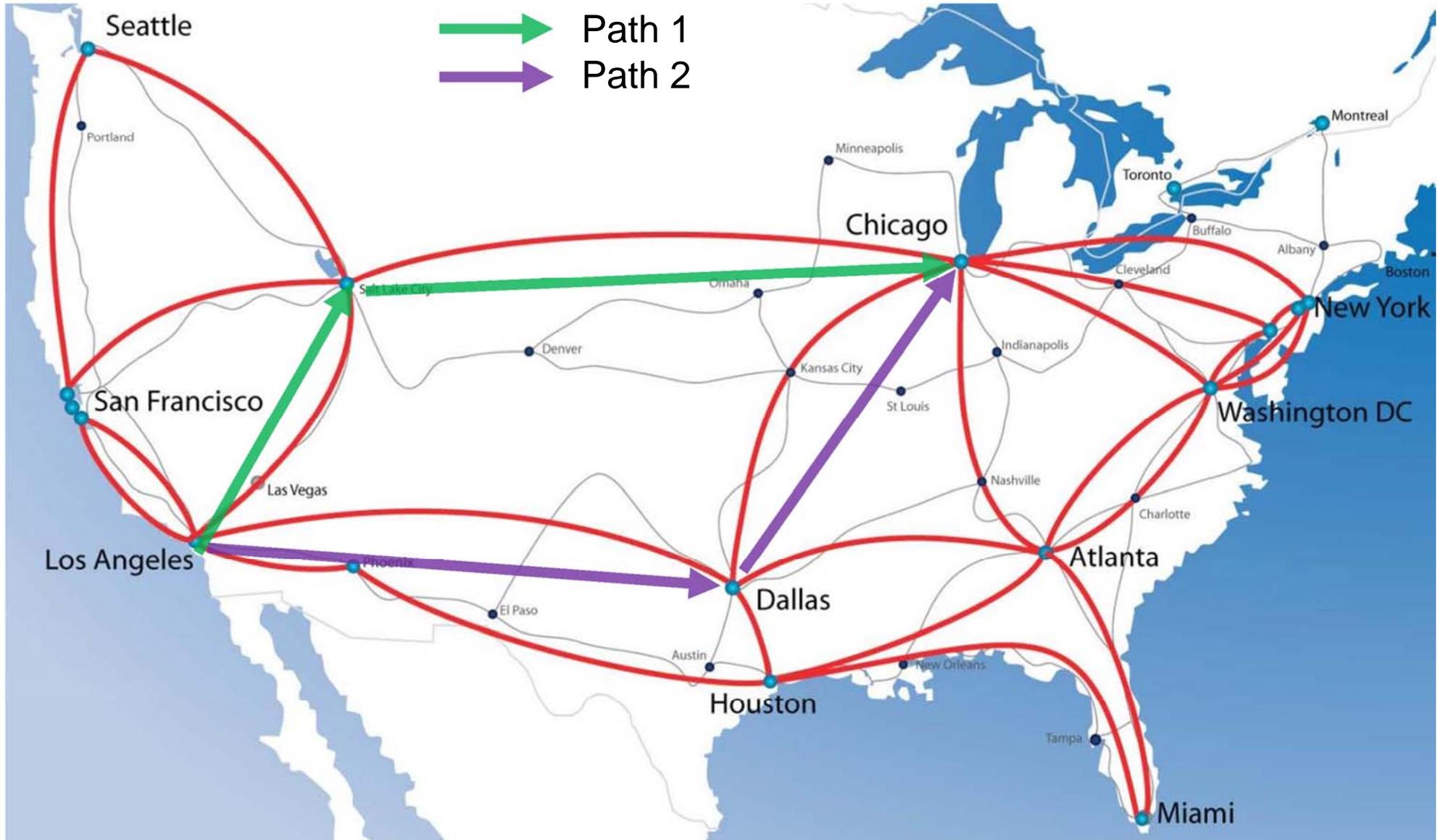
How to Route from Los Angeles to Chicago



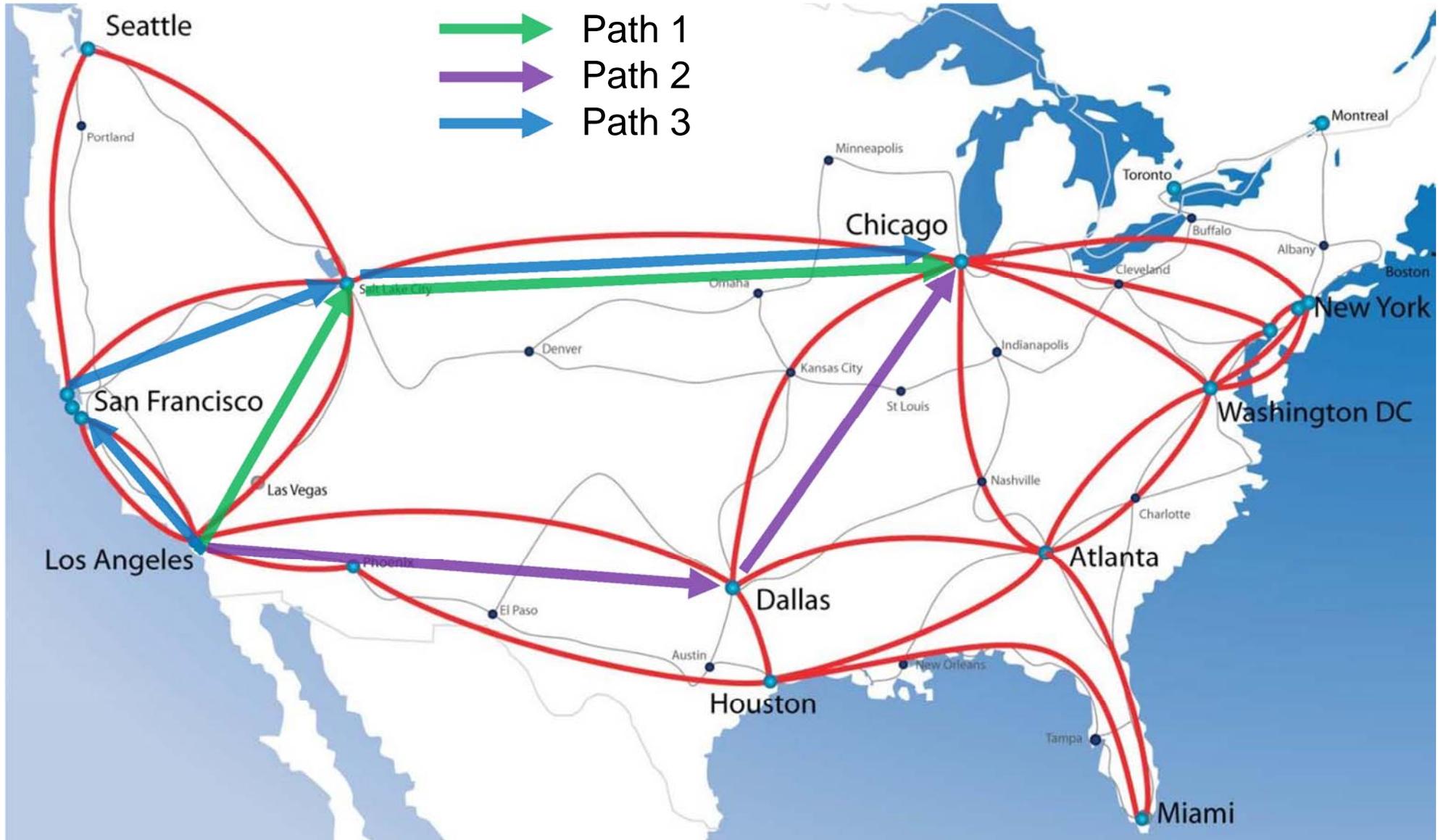
How to Route from Los Angeles to Chicago



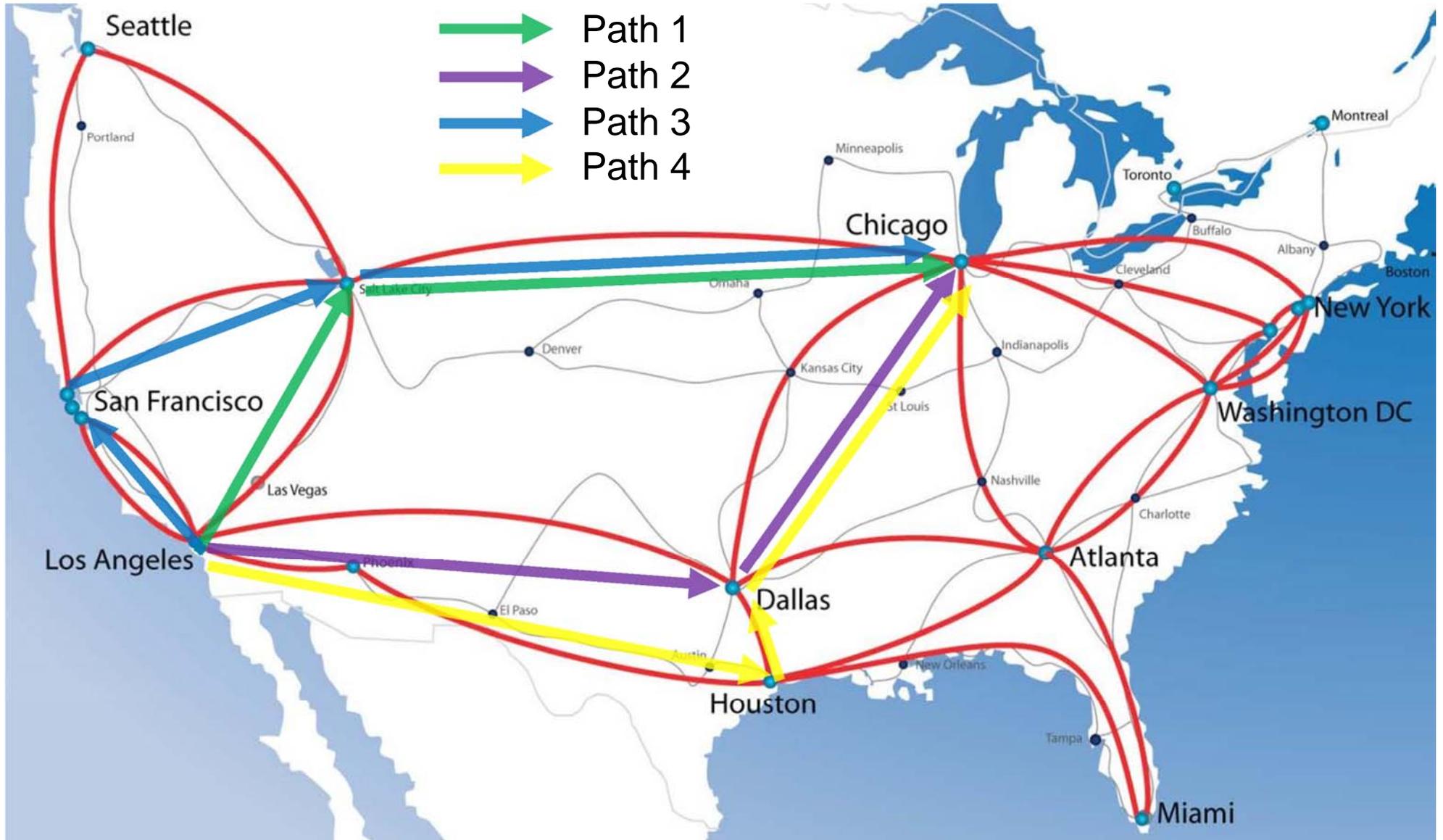
How to Route from Los Angeles to Chicago



How to Route from Los Angeles to Chicago



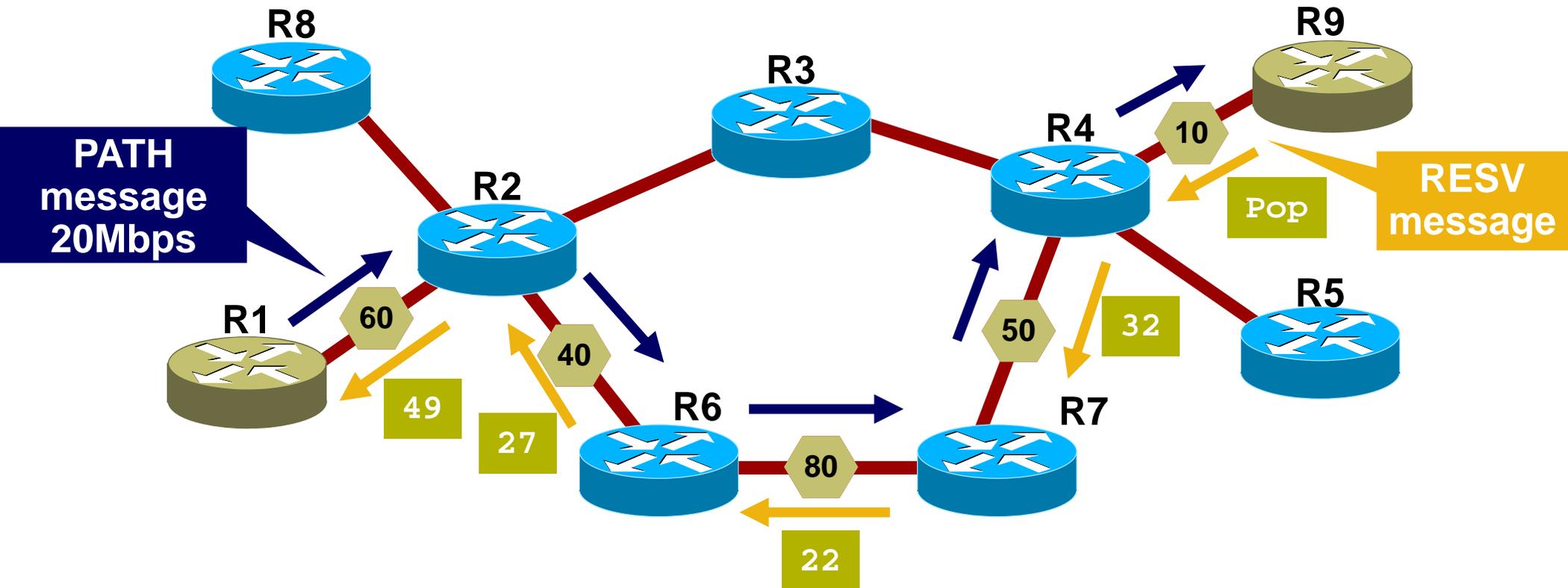
How to Route from Los Angeles to Chicago



How Does MPLS Traffic Engineering Work?

- Using RSVP-TE to reserve bandwidth across the network.
 - Remember, an LSP is a “tunnel” between two points in the network.
 - Under RSVP, each LSP has a bandwidth value associated with it.
 - Using constrained routing, RSVP-TE looks for the shortest path with enough available bandwidth to carry a particular LSP.
 - If bandwidth is available, the LSP is signaled across a set of links.
 - The LSP bandwidth is removed from the “available bandwidth pool”.
 - Future LSPs may be denied if there is insufficient bandwidth.
 - They’ll ideally be routed via some other path, even if the latency is higher.
 - Existing LSPs may be “preempted” for new higher priority LSPs.
 - You can create higher and lower priority LSPs, and map certain customers or certain traffic onto each one.
 - This isn’t traditional QoS, no packets are being dropped when bandwidth isn’t available, you’re simply giving certain traffic access to shorter paths.

How RSVP-TE Reserves Bandwidth



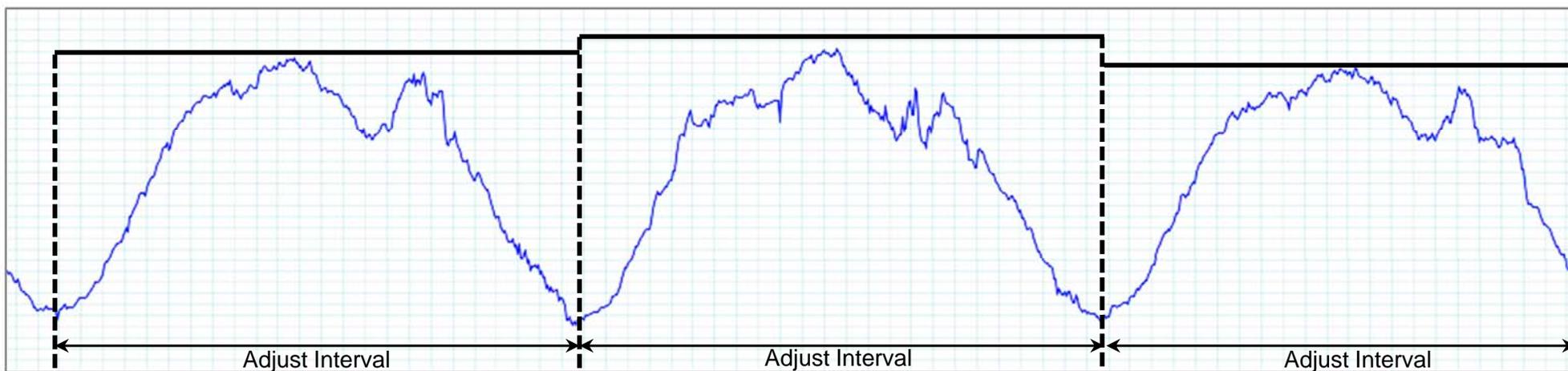
-  RSVP PATH: R1 → R2 → R6 → R7 → R4 → R9
-  RSVP RESV: Returns labels and reserves bandwidth
-  Bandwidth available on each link
-  Label value returned via RESV message

How Do You Determine an LSP Bandwidth?

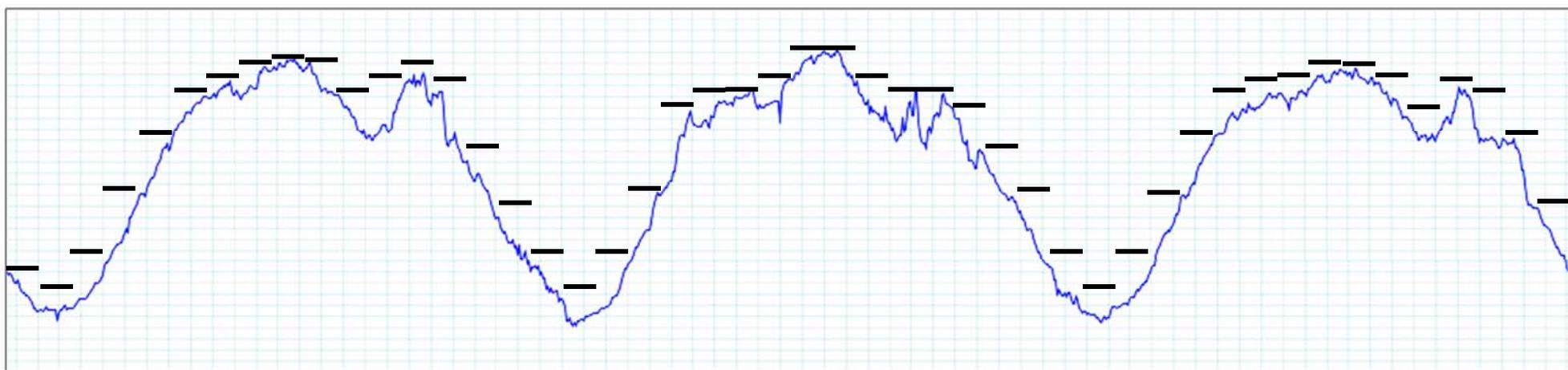
- How do you determine the bandwidth of particular LSP?
 - After all, IP networks are dynamic and packet switched.
 - Bandwidth use can change in an instant, and be unpredictable.
- There are basically two main ways to do it:
 - Offline Calculation
 - Calculation which occurs outside of the router, typically based on some bandwidth “modeling”, and often using a third party script or tool.
 - This is how MPLS was first implemented, and is still commonly used today by most large networks and early MPLS adopters.
 - Auto-Bandwidth
 - The bandwidth value is calculated on the router, by periodically measuring how much traffic is actually forwarding over the LSP.
 - The RSVP reservation is then periodically updated with the new number.

LSP Bandwidth – More Often is Better

24 Hour Adjust Interval



1.5 Hour Adjust Interval – More Efficient Bandwidth Use



Offline Calculation vs. Auto-Bandwidth

- Offline Calculation
 - You can implement any algorithm you'd like.
 - Some extremely complex LSP modeling software is available from 3rd party vendors, allowing you to do detailed LSP planning.
 - But you either have to write the software yourself, or buy it.
- Auto-Bandwidth
 - Because it runs directly on the router, it can respond to changing traffic conditions much more rapidly, with less overhead.
 - Most offline calculations are based expectations of stable traffic patterns.
 - Unusual traffic spikes can cause congestion or inefficient bandwidth use.
 - Easier to implement (just turn the knob on your router, it's free).

MPLS Data Transport Services

MPLS Pseudowires

- Layer 2 Pseudowire or VLL (Virtual Leased Line)
 - An emulated layer-2 point-to-point circuit, delivered over MPLS.
 - Currently standardized by the “PWE3” IETF Working Group.
 - Can be used to interconnect two different types of media:
 - For example, Ethernet to Frame Relay.
 - Useful for migrating legacy transport (e.g. ATM) to an MPLS network.
 - Can be difficult to load balance, due to lack of visibility into the packet.
 - The payload is unknown, so you can’t hash on the IP header inside, etc.
 - Historically two competing methods for signaling:
 - LDP-signaled / Draft Martini
 - The simpler of the two methods, and more commonly implemented.
 - BGP-signaled / Draft Kompella / L2VPN
 - More complex, but with auto-discovery support for multi-point.

MPLS L3VPNs

- L3VPN
 - An IP based VPN.
 - Networks build virtual routing domains (VRFs) on their edge routers.
 - Customers are placed within a VRF, and exchange routes with the provider router in a protected routing-instance, usually BGP or IGP.
 - Can support complex topologies and interconnect many sites.
 - Usually load-balancing hash friendly (has exposed IP headers).
 - But can add a significant load to the service provider infrastructure.
 - Since the PE device must absorb the customer's routing table, consuming RIB and FIB capacity.
 - Typically seen in more enterprise environments.
 - Signaled via BGP within the provider network.

MPLS VPLS

- VPLS (Virtual Private LAN Service)
 - Creates an Ethernet multipoint switching service over MPLS.
 - Used to link a large number of customer endpoints in a common broadcast domain.
 - Avoids the need to provision a full mesh of L2 circuits.
 - Emulates the basic functions of a layer 2 switch:
 - Unknown unicast flooding
 - Mac learning
 - Broadcasts
 - Typically load-balancing friendly since the L2 Ethernet headers are examined and used, unlike L2 pseudowires where they are passed transparently.

MPLS Fast Reroute

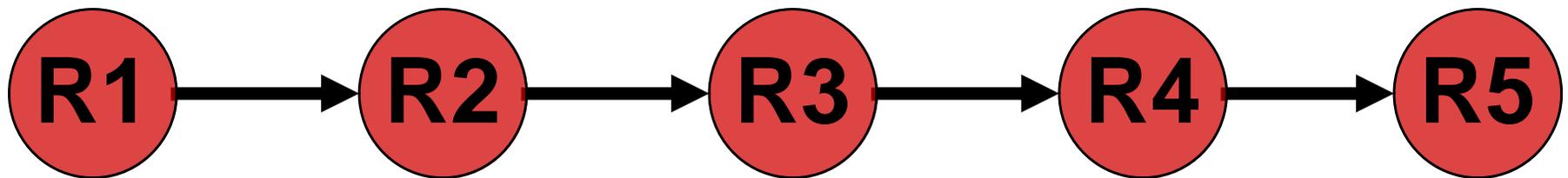
What Does Fast Reroute Do?

- MPLS Fast Reroute improves convergence during a failure.
 - By pre-calculating backup paths for potential link or node failures.
 - In a normal IP network
 - The best path calculation happens on-demand when a failure is detected.
 - It can take several seconds to recalculate best paths and push those changes to the router hardware, particularly on a busy router.
 - A transient routing loop may also occur, as every router in the networks learns about the topology change.
 - With MPLS Fast Reroute
 - The next best path calculation happens before the failure actually occurs.
 - The backup paths are pre-programmed into the router FIB awaiting activation, which can happen in milliseconds following failure detection.
 - Because the entire path is set within the LSP, routing loops cannot occur during convergence, even if the path is briefly suboptimal.

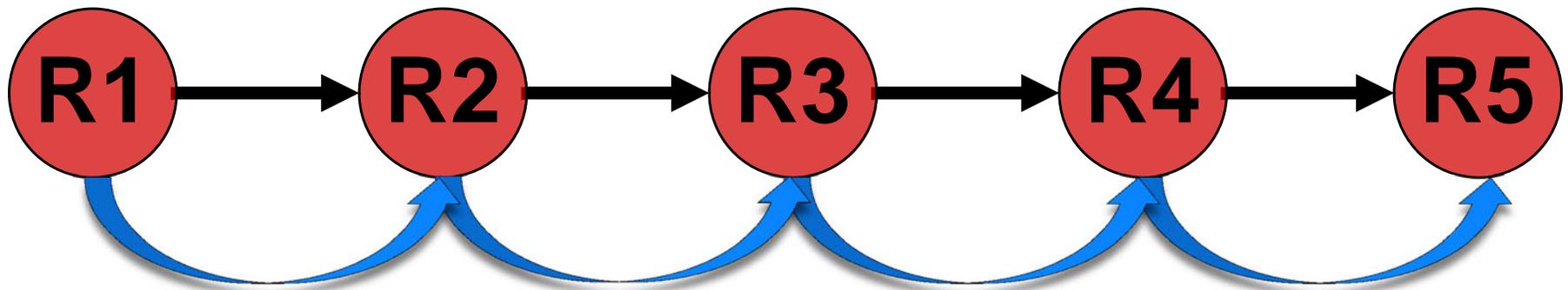
MPLS Protection Schemes

- There are two different ways to provide LSP protection:
 - One-to-One Protection / Detour
 - An individual backup path is fully signaled through RSVP for every LSP, at every point where protection is provided (i.e. every node).
 - The label depth remains at 1, but this can involve a huge number of reservations, and can cause significant overhead.
 - Many-to-One Protection / Facility Backup
 - A single bypass LSP is created between two nodes to be protected.
 - During a failure, multiple LSPs are rerouted over the bypass LSP.
- Also different types of failures that can be protected against:
 - Link Protection / Next-Hop Backup
 - A bypass LSP is created for every possible link failure.
 - Node Protection / Next-Next-Hop Backup
 - A bypass LSP is created for every possible node (router) failure

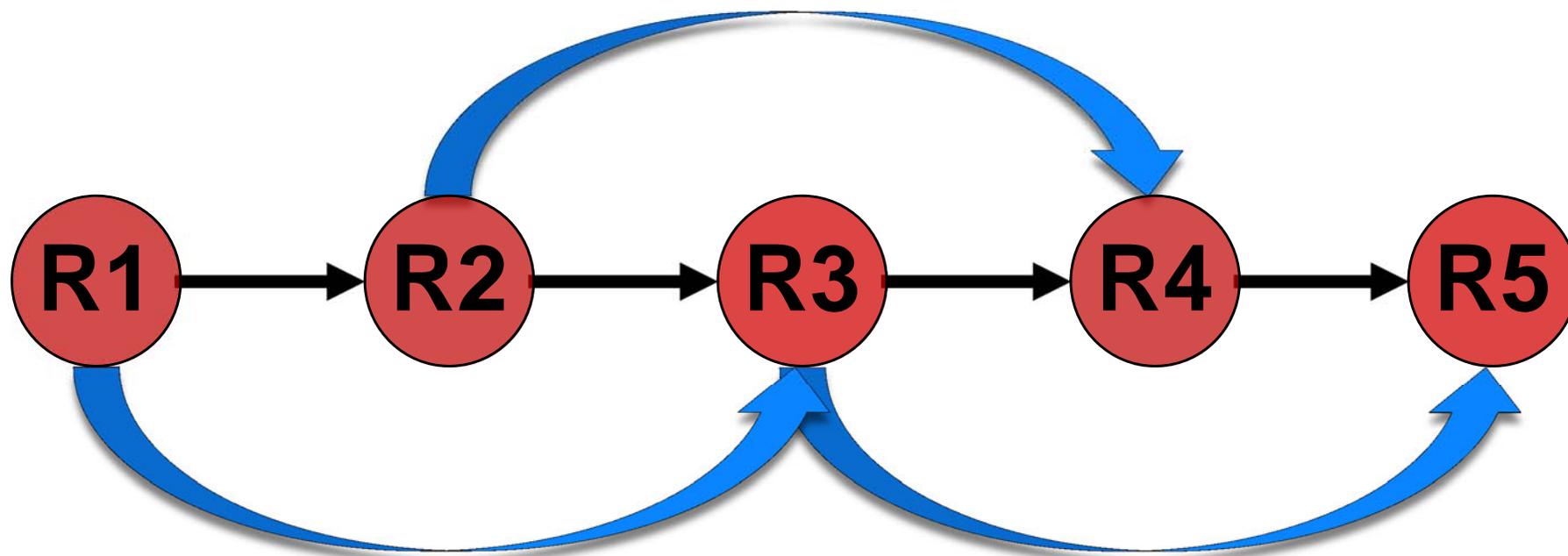
MPLS With No Protection



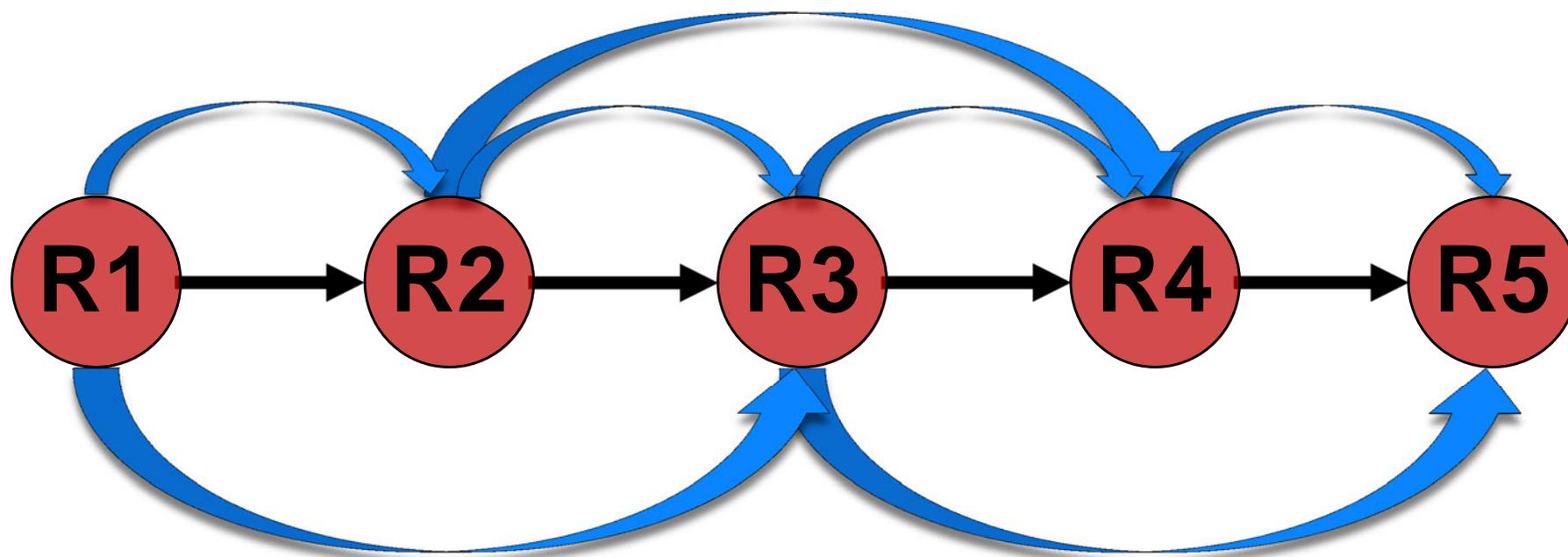
MPLS Link Protection



MPLS Node Protection



MPLS Link and Node Protection

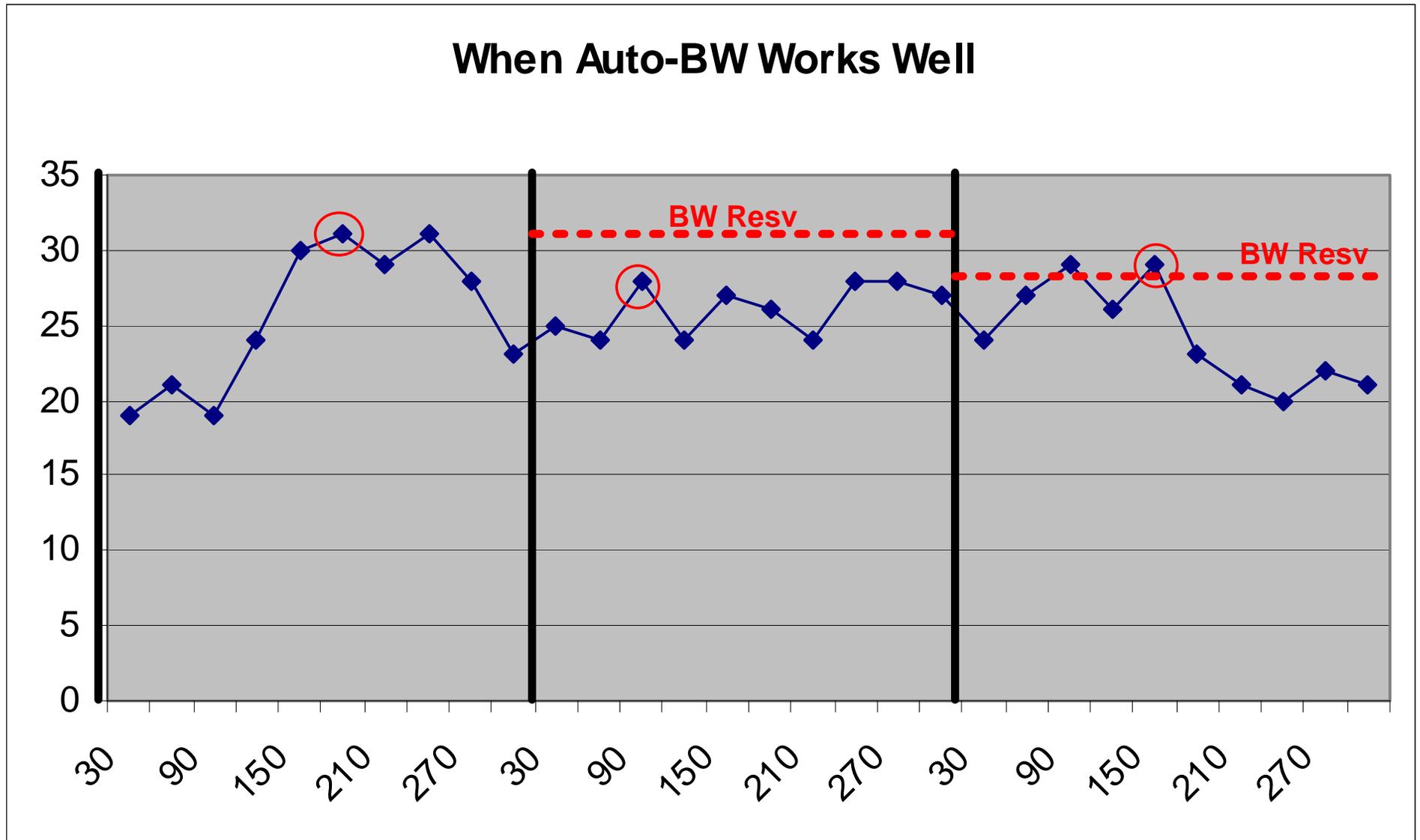


MPLS Auto-Bandwidth

How Does Auto-Bandwidth Work?

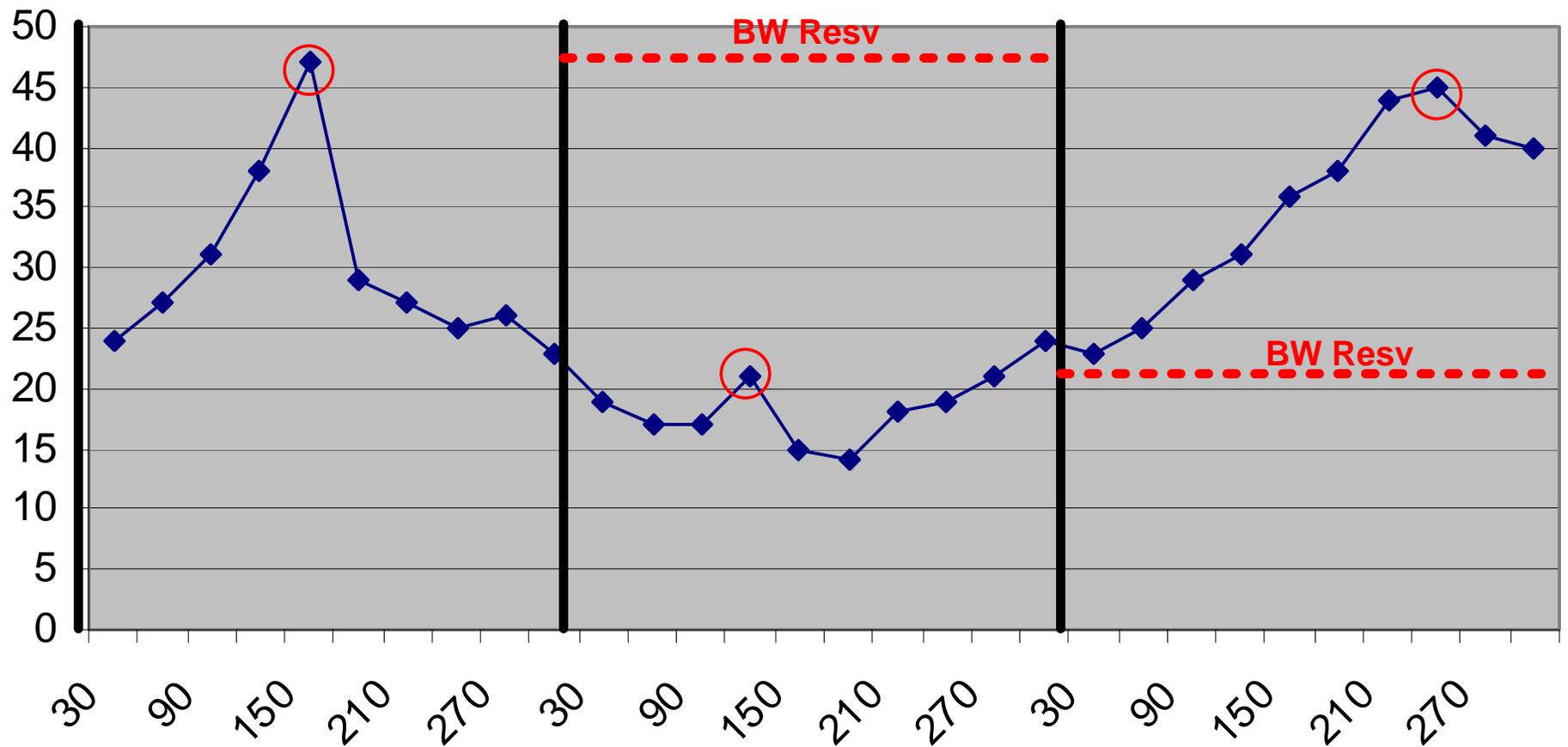
- Technically each algorithm is router/vendor independent.
 - But both Cisco and Juniper implement it in much the same way.
- Auto-Bandwidth performs the following basic steps.
 1. Every Statistics Interval, bandwidth over an LSP is measured.
 - For example, you might configure this to every 60 seconds.
 2. Every Adjust Interval, the largest sample from the process above is used to calculate the new LSP bandwidth.
 - For example, you might use 5 samples and adjust every 300 seconds.
 3. If the change is larger than a user configured minimum threshold, the new bandwidth value is re-signaled across RSVP.
 - Ideally using a make-before-break configuration.

When Auto-Bandwidth Works Well



When Auto-Bandwidth Doesn't Work Well

When Auto-BW May Not Work So Well



Overflow and Underflow

- Another common optimization is Overflow and Underflow.
 - This logic says: If a certain number of Statistics Samples exceed a certain % difference, kick off an Adjust event before the normal time.
 - Instead of having a low Adjust Interval, you can have a higher interval, and allow on Overflow/Underflow to detect significant changes to the % of bandwidth used.
 - This can help catch major bandwidth changes such as during failure events or sudden traffic spikes.
 - Warning: Not every vendor supports Underflow
 - If traffic shifts from one LSP to another, Overflow may catch the increase of traffic on the new LSP, but without Underflow you won't catch the decrease in traffic on the old LSP until the next adjust interval.
 - This can lead to suboptimal routing, or even failure to signal the LSP at all, if the network cannot find bandwidth for both LSPs simultaneously.

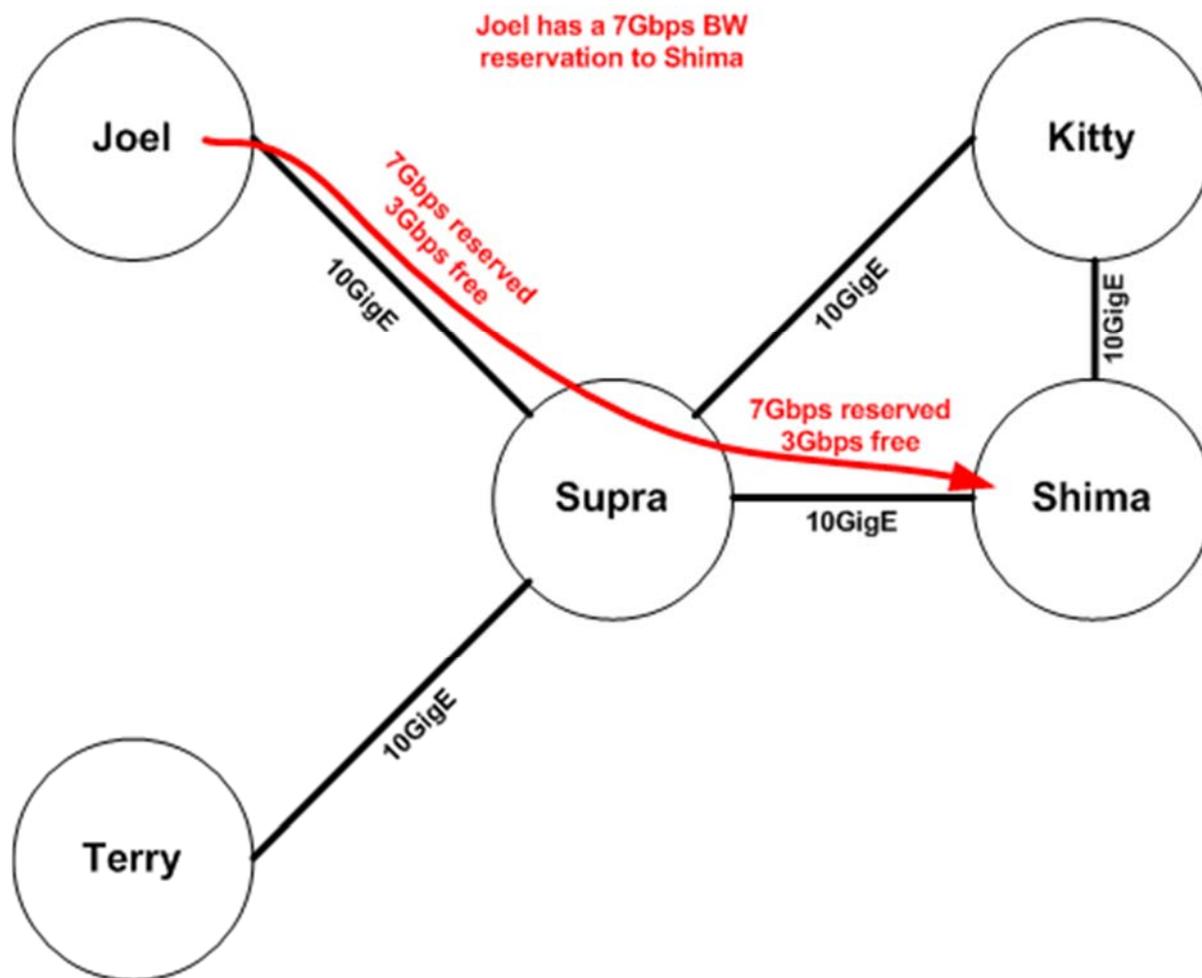
More MPLS Details

RSVP-TE LSP Priorities

- LSPs have the ability to preempt each other to acquire available bandwidth based upon a defined priority value.
 - Each LSP has a SETUP and a HOLD priority.
 - SETUP is the priority value defined when the tunnel is establishing.
 - HOLD is the priority value defined when a tunnel has already been setup.
 - 8 priority values are available (0-7), 0 having the highest priority.
- Two types of preemption on routers:
 - Hard
 - LSP is torn down in a disruptive manner.
 - Soft
 - Gives the LSP to be preempted time (usually tens of seconds – configurable) to find a new path and tear itself down.
 - Mostly non-disruptive.

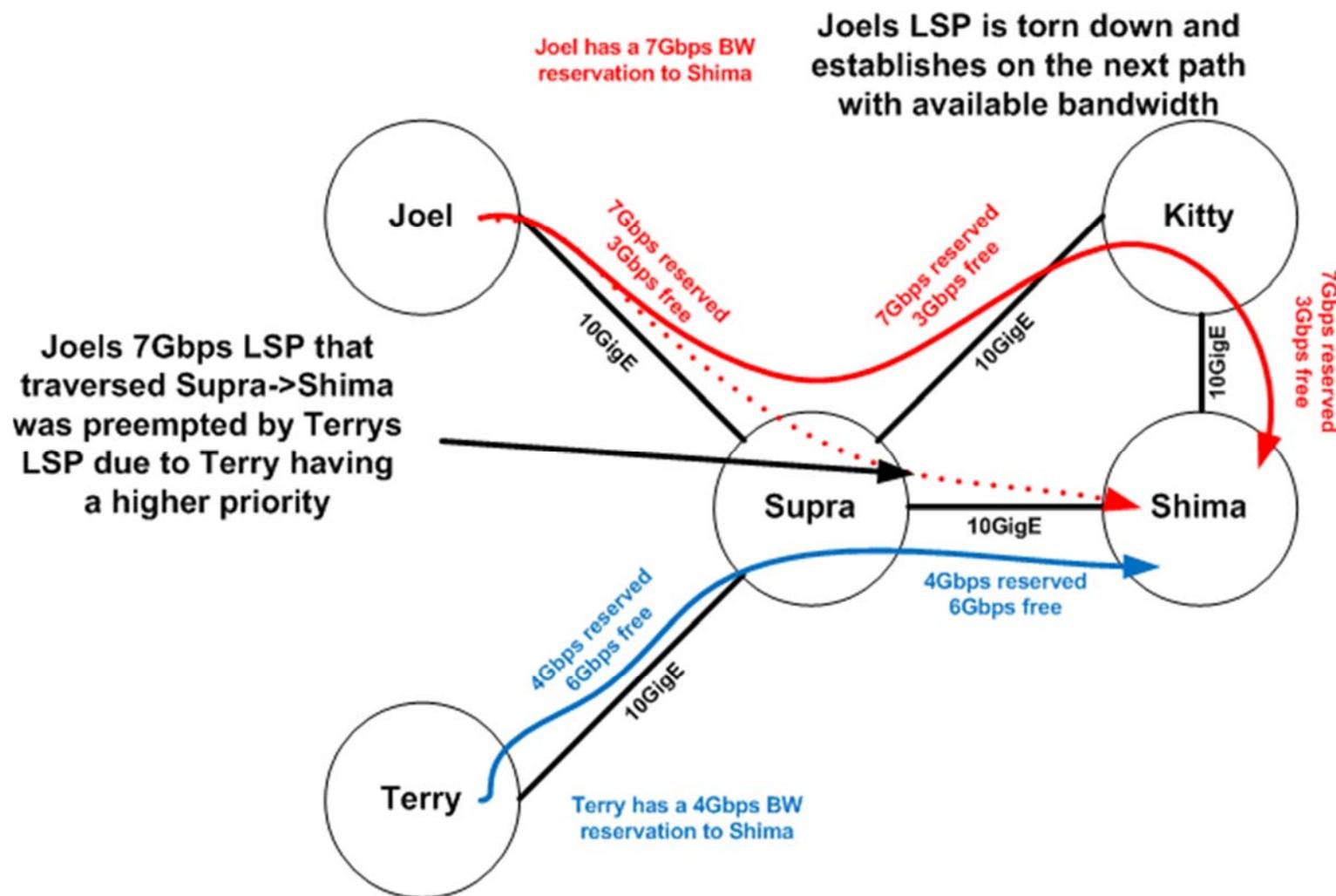
MPLS Preemption Example - Before

- LSP from Joel to Shima with a HOLD priority of 5 and 7Gbps



MPLS Preemption Example - After

- LSP from Terry to Shima with a SETUP priority of 3 and 4Gbps



LSP Optimization

- Over time, network topologies can change.
 - IGP cost changes, new links, failed links, etc.
- The optimization process will re-compute the LSP paths.
 - Normally reoptimization is a periodic process that operates in the background, looking for a better path for one or all LSPs.
 - It can also be triggered manually if necessary.
 - If a better path is found, the router will attempt to rebuild the LSP on the new path.
- Many routers can do “smart” optimization.
 - For example, a link coming back up can trigger an optimization event before the normal timer.
 - But only once, to prevent a flapping link from going nuts.

Make Before Break / Adaptive LSPs

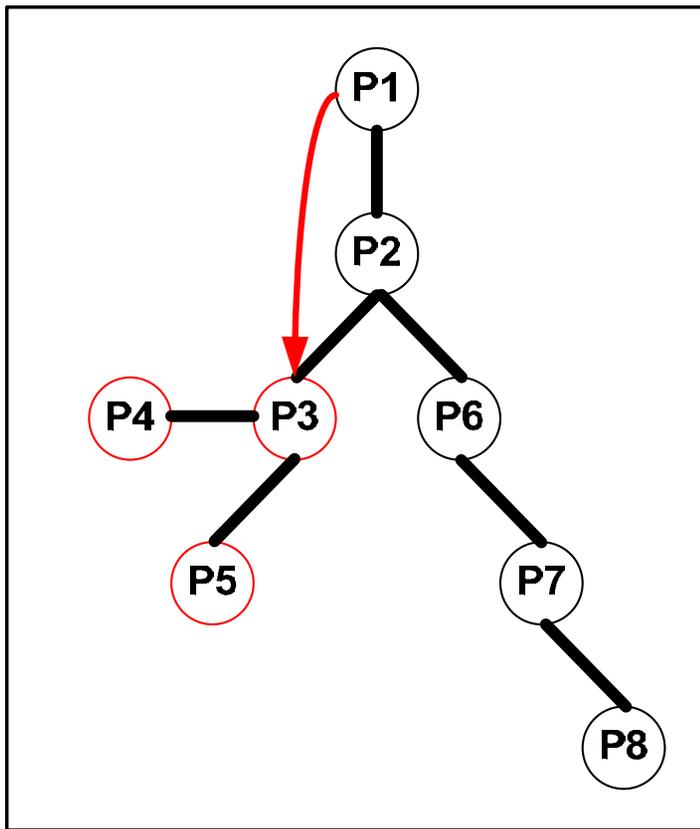
- When an LSP is re-signaled for any reason, the old LSP is completely torn down, and a new one is built in its place.
 - Reoptimization, bandwidth reservation updating, etc.
- To avoid traffic disruption, a make-before-break option fully signals the new LSP before tearing down the old one.
- But this may cause transient double-counting of bandwidth.
 - When the old and new LSPs share the same path, double counting can be avoided.
 - But if the paths are different, the LSP bandwidth may be reserved twice.

Using an LSP for your IP Traffic

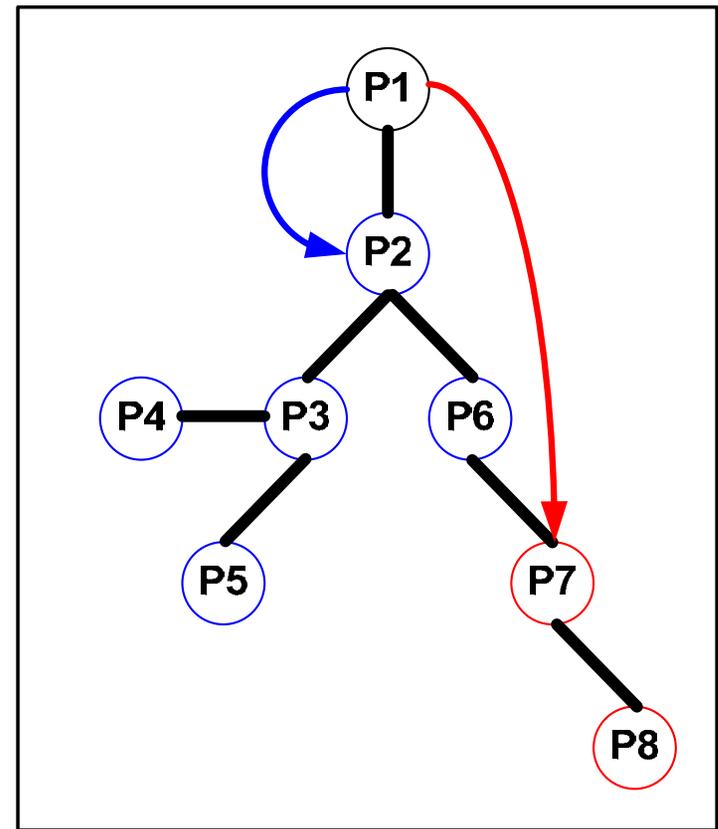
- Now that you have these LSPs, how do you use them for IP?
 - Static Routes
 - Not very practical, though these can be useful in some limited scenarios.
 - Handy way to do some quick and dirty traffic engineering for a prefix.
 - Juniper TE Shortcuts / Cisco Autoroute Announce
 - Let the local router use MPLS LSPs as next-hops for BGP/IGP routes.
 - Cisco implements this transparently by modifying the SPF algorithm, Juniper adds the LSPs to the inet.3 table, but the result is the same.
 - Since this is a local router feature, it can be enabled or disabled on specific routers, and is not advertised to other routers.
 - Even if the destination endpoint doesn't speak MPLS, the LSP that goes to the last MPLS speaking router along the path will be used.

IGP-Shortcut / Autoroute Announce

P1 -> P3 LSP



P1 -> P2 and P1 -> P7



MPLS and Traceroute

- MPLS can also let you hide traceroute hops.
 - Since you aren't actually doing IP forwarding, there is no need to decrement the IP TTL field as you MPLS forward the packet.
 - And if you don't, the LSP shows up as a single hop in traceroute.
 - Some networks prefer this behavior, as it hides the internals of their network, and makes for shorter / prettier traceroutes.
 - Some networks also run MPLS-only cores, which carry no IP routes.
 - This presents a problem, since if they did want to show the hops in traceroute, the router can't do IP routing to return the ICMP TTL Exceed.
 - To solve this problem, an "icmp tunneling" feature was implemented.
 - If an ICMP message is generated inside an LSP, the ICMP message is carried all the way to the end of the LSP before being routed back.
 - This can make traceroute look really weird, since you see all the hops along the LSP, but they all appear to have the same latency as the final hop. This causes much end-user confusion.

Link Coloring (Affinities / Admin-Groups)

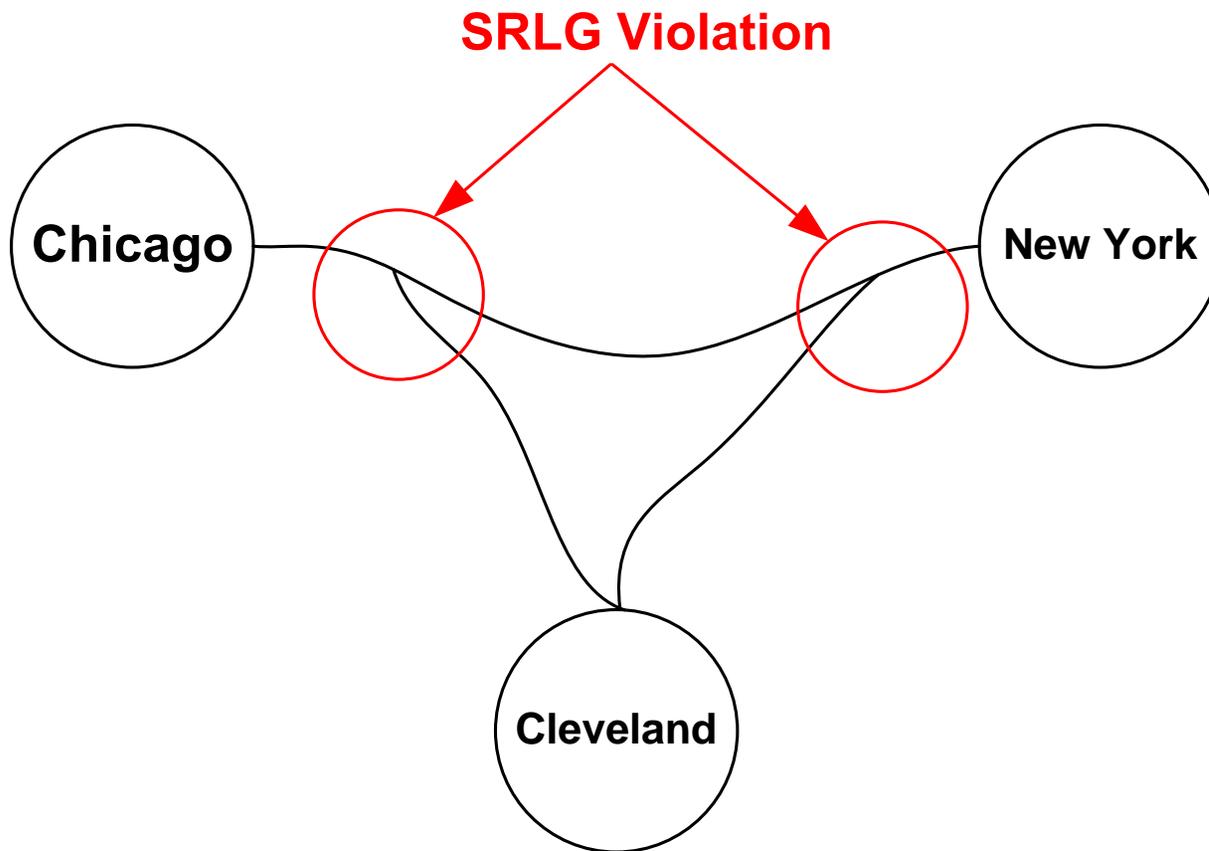
- An additional constraint in the CSPF algorithm.
 - Allows for 32 unique “color” markings that can be placed on a link.
 - Multiple color markings can be applied on a link.
 - Link colors are advertised as a link attribute.
 - Periodically flooded out just like Bandwidth information.
 - The operator can use these markings in any way they wish.
- This is created for specifying:
 - Geographic / Political boundaries
 - Prefer to keep traffic routing within a specific country/region/continent
 - Cost-Out/Maintenance Activities
 - Can instruct all LSPs to immediately move off a path
 - Prevent traffic from traversing specific links/paths
 - Don't have “core-to-core” LSPs traverse edge routers or metro networks

SRLG - Shared Risk Link Groups

- By default, backup paths may not provide full redundancy.
 - For example, the “next best path” that goes into fast reroute may ride on the same transport equipment, physical path, conduit, etc.
 - If both paths fail simultaneously, you don’t get a fast reroute.
- SRLGs let you define links that share common risks.
 - This can then be used to force backup paths to use different SRLG links, even if the backup path is less optimal by IGP cost.

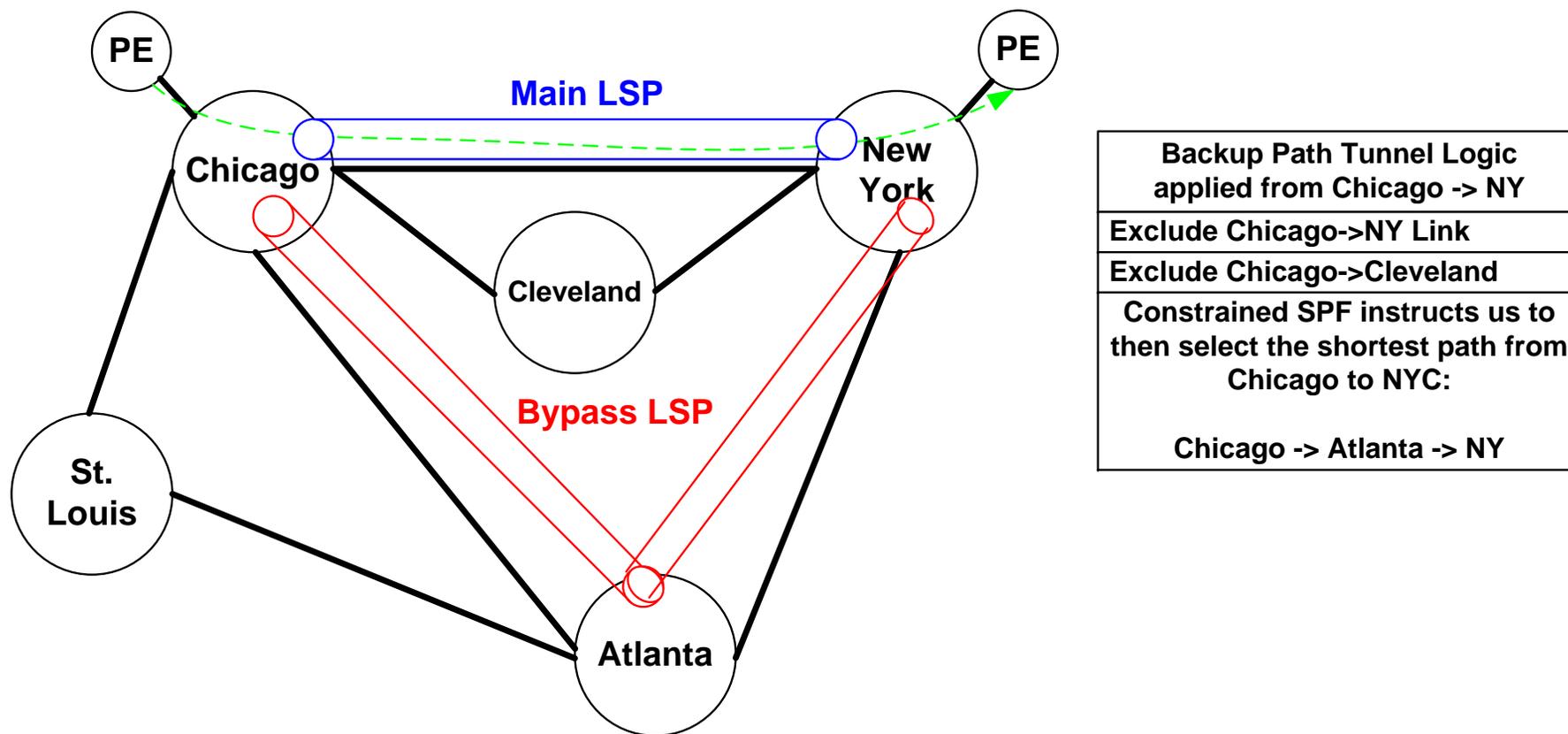
SRLG - Shared Risk Link Groups

- Layer 1 Example



SRLG - Shared Risk Link Groups

- Solution – Build a backup path that avoids the SRLG

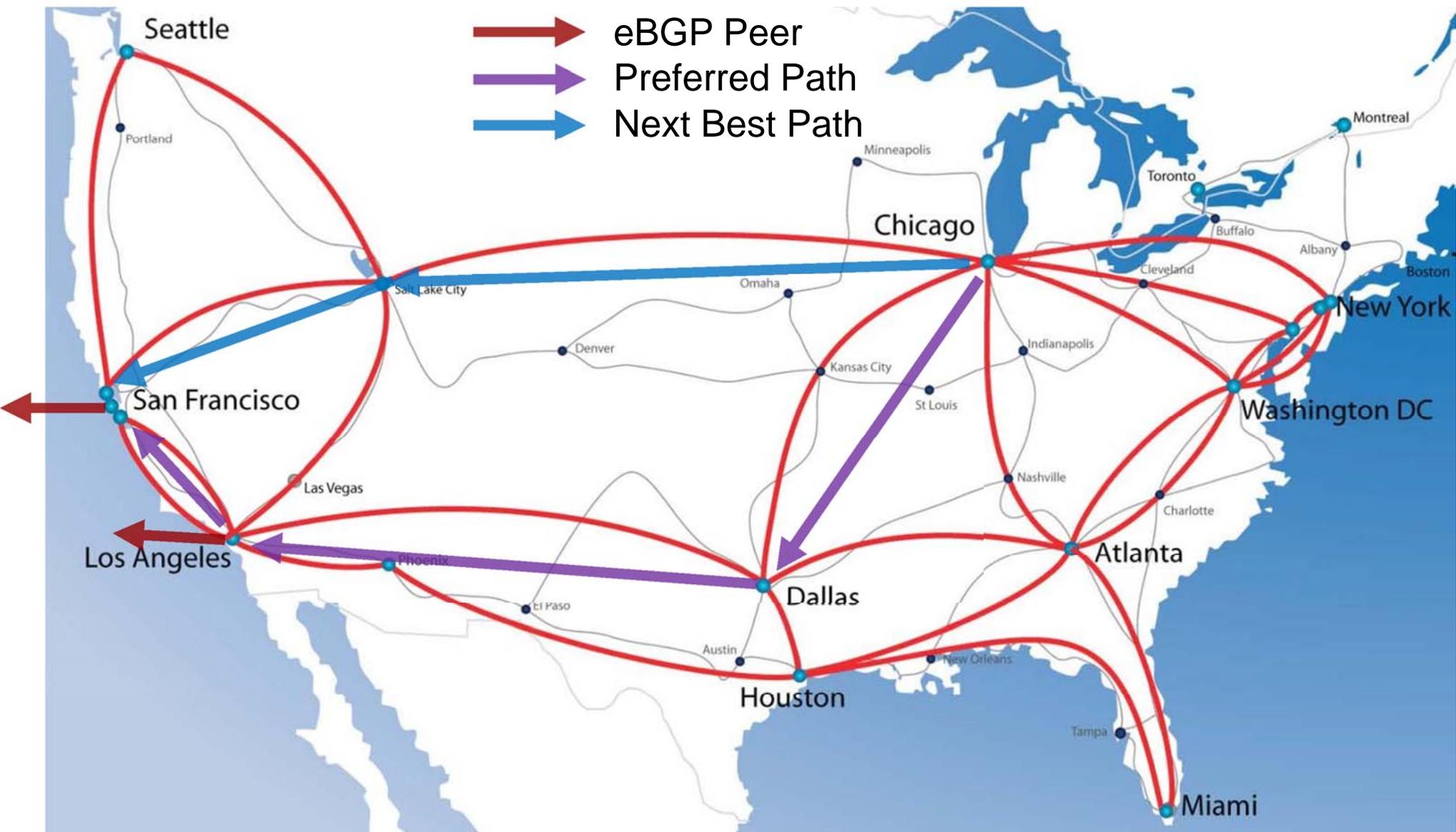


The Downsides of MPLS

The Downsides of MPLS

- No protocol is perfect, MPLS least of all.
 - One major drawback is that it hides suboptimal topologies from BGP, where multiple exits may exist for the same route.
- For example:
 - Say you peer with a major network in San Jose and Los Angeles.
 - Traffic coming from Chicago would normally go directly to San Jose.
 - But because of a capacity issue, the LSP is forced to go via Los Angeles first.
 - In an IP network, the packet would probably be diverted to the local Los Angeles peer as it passes through Los Angeles.
 - But MPLS will hide the suboptimal topology, the packet will continue to San Jose because that's what Chicago saw as the best exit.
 - This can be a good or a bad thing depending on your goals.

MPLS Blocks Use of a Second Exit



MPLS LSPs Don't Create Themselves

- Unlike other protocols, MPLS isn't entirely auto-magic.
 - There are no protocols to auto-discover MPLS speaking nodes.
 - The MPLS “protocols” just exchange label values for an LSP.
 - They have no involvement with the creation of the LSPs.
 - Building the full mesh of LSP tunnels is left up to the operator.
 - Essentially this means operator supplied scripts are a necessity.
 - Or else an operator purchased commercial software solution.
 - Examples include WANDL, Cariden, etc.
- Some vendors offer some very basic Auto-Mesh capabilities.
 - For example, Cisco can auto-create a mesh of LSPs from a template, using a list of router IPs supplied in an access-list.
 - But this leaves you no way to control a specific LSP configuration.
 - Oh and if you want to remove a node from the mesh you have to remove the entire ACL, bringing down every dynamic auto-mesh LSP on the box.

Large LSPs Can't Fit Down Small Pipes

- An LSP can only be moved as an atomic unit.
 - So if you have relatively large LSPs relative to the size of the circuits they're traversing, you may not be able to efficiently pack them.
 - For example, say you have (3) 6 Gbps LSPs and 2x10G circuits.
 - You'll only be able to fit 2 of the 3 LSPs above.
 - The other LSP will have to find another longer path, if one exists at all.
 - And your 2 circuits will be left with 4 Gbps of unfilled capacity.
 - Another example, say you have mixed OC192 and OC48 circuits.
 - A 3 Gbps LSP will never be able to fit down an OC48 circuit.
- One workaround is to create multiple parallel LSPs
 - Instead of having (3) 6 Gbps LSPs you could have (9) 2 Gbps LSPs.
 - But so far no router vendor auto-mesh systems support parallel LSPs.
 - Ideally you would want auto-bandwidth to "fork" an LSP doing $> \#$ BW
 - But no vendor implementation can do this either.

The Gotchas of Auto-Bandwidth

- Auto-Bandwidth isn't perfect either.
 - We've already seen some examples of incorrect sizing.
- Auto-Bandwidth + Oversubscribed Links = Bad Things
 - Auto-Bandwidth doesn't know anything about congestion on links.
 - Say you oversubscribe a link, RSVP fills it, and you get packet drops.
 - Drops cause TCP to throttle back, and the IP traffic rate goes down.
 - Auto-Bandwidth adapts to this new rate, and thinks everything is fine.
 - This leads to sustained congestion requiring manual intervention.
- Also, be careful if your router doesn't "see" L2 overhead.
 - A 28 byte UDP flood consumes 84 bytes over the wire on Ethernet.
 - A DoS attack of small packets can result in in congestion that is completely invisible to auto-bandwidth.

Send questions, comments, complaints to:

Richard A Steenbergen ras@nlayer.net