MPLS Fast Reroute

Kireeti Kompella
kireeti@juniper.net

MPLS 2004

Agenda

- Fast reroute for IP and LDP traffic
  - We already know how to do RSVP-TE fast reroute
- Two independent but related problems
  - Fast repair
  - Safe IGP convergence (not covered here)
- For each, an analysis of
  - "coverage"
  - feasibility
  - scalability

Problem: IP/LDP Fast Reroute

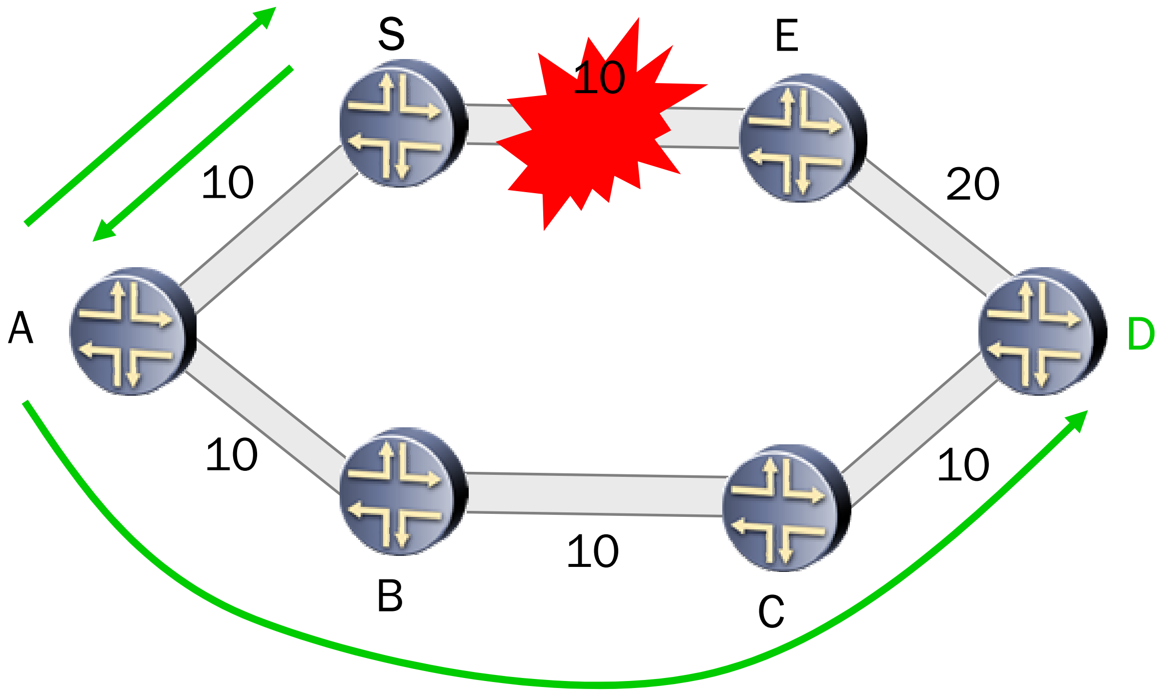
- IGP convergence within a node takes 100s of milliseconds to compute SPF and update FIB
  - Can be made faster with incremental SPF, but …
- … network convergence of a link-state IGP is a distributed problem
- Two parts
  - fast repair
  - distributed convergence

(This presentation will focus on IP fast reroute and link failures; generalization to LDP and node failures is fairly straightforward)

Fast Repair Techniques

- Loop-free alternates
- U-turn alternates
- IP tunnels
- RSVP-TE (source-routed) bypass tunnels

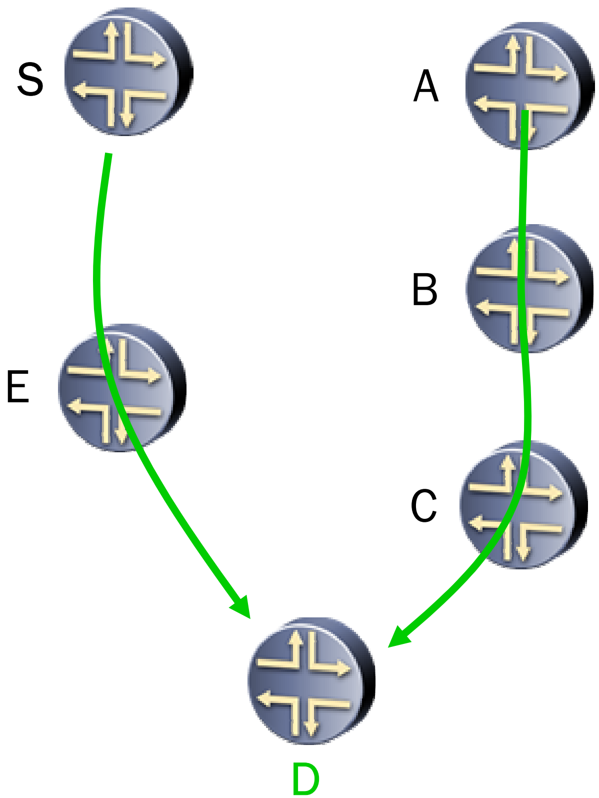
Illustrative Network Diagram



Consider traffic from **S** to **D** which flows **S→E→D**

Now, suppose that **link SE** fails

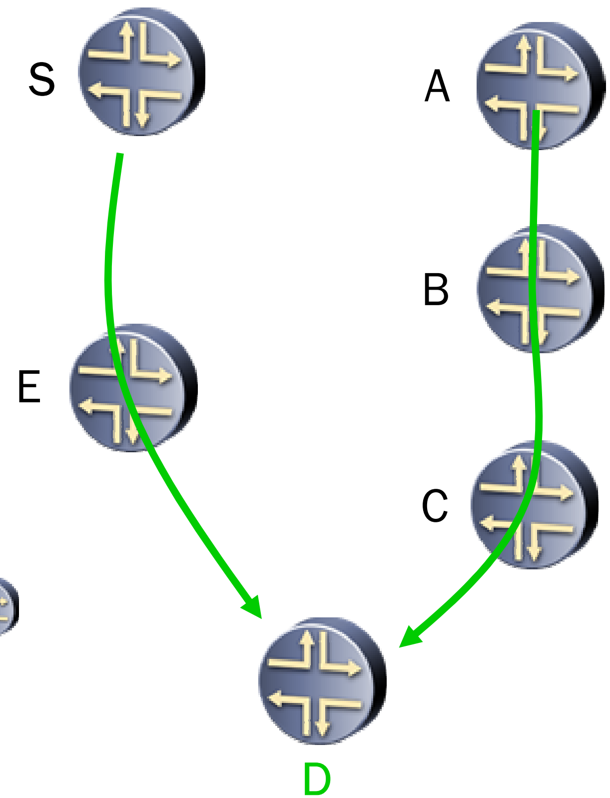
Useful Analogy



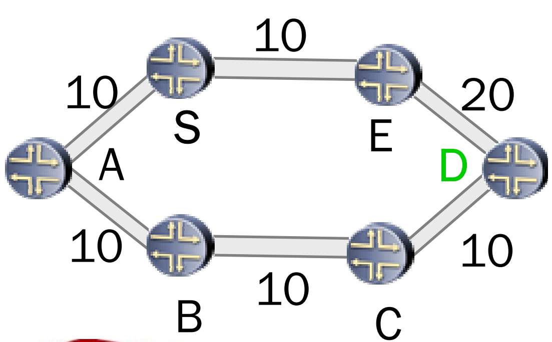
Packets flowing along the shortest path will be viewed as ***water flowing downhill***

"Downhill" depends very much on the **destination**

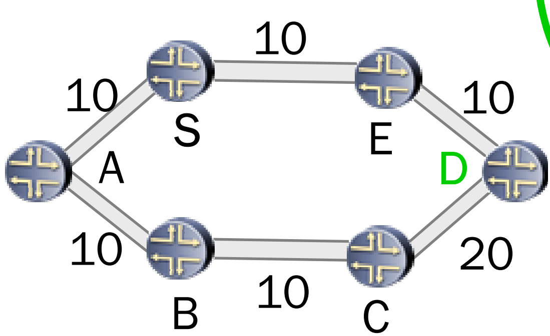
“Loop Free Alternate”

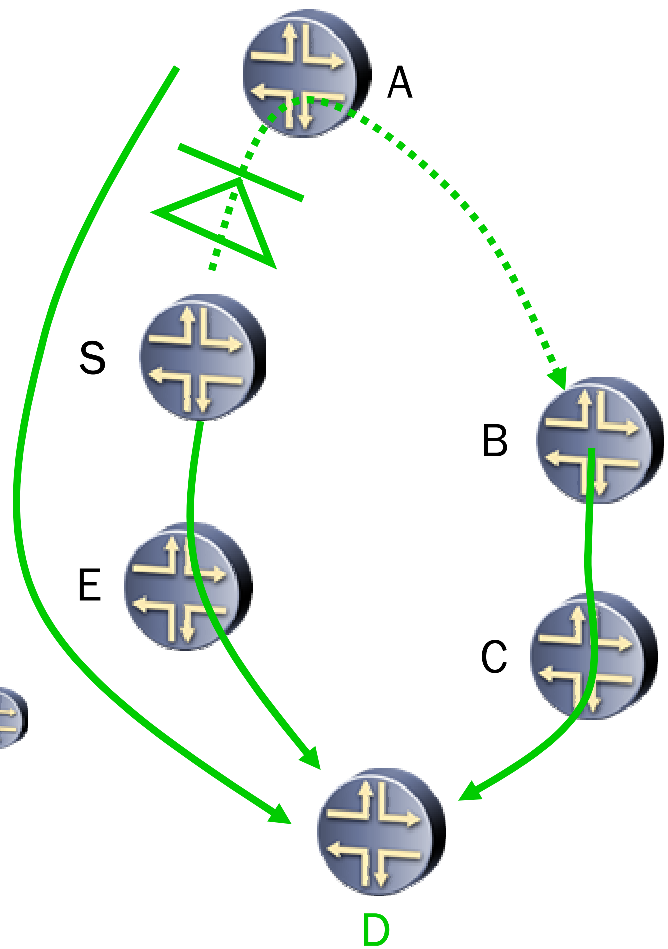


For a given destination, a neighbor of yours whose downhill path **<u>doesn’t go through you</u>** is a **“loop-free alternate”**



"U-Turn Alternate"



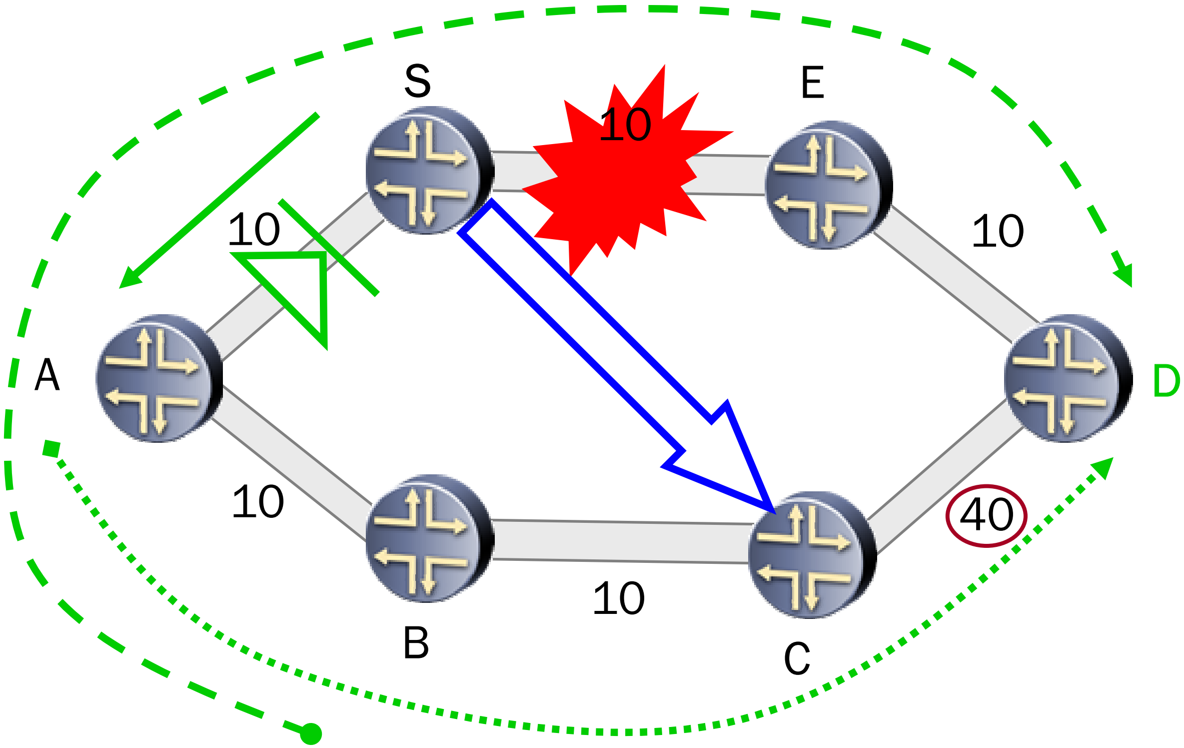


A neighbor of yours whose downhill path **<u>does</u>** go through you but is capable of diode-like behaviour

MPLS 2004

Juniper NETWORKS

IP Tunnels for Fast Reroute



Suppose in this case A is assumed to be a U-turn alternate.

On failure, S reroutes to A, who sends it to B …

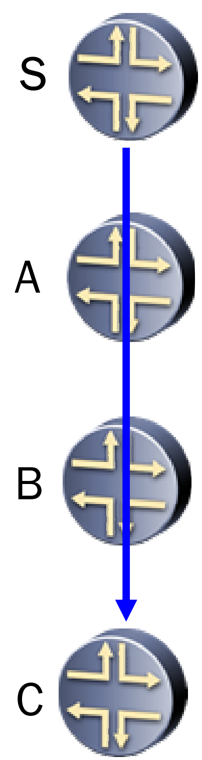
… but B sends it right back!

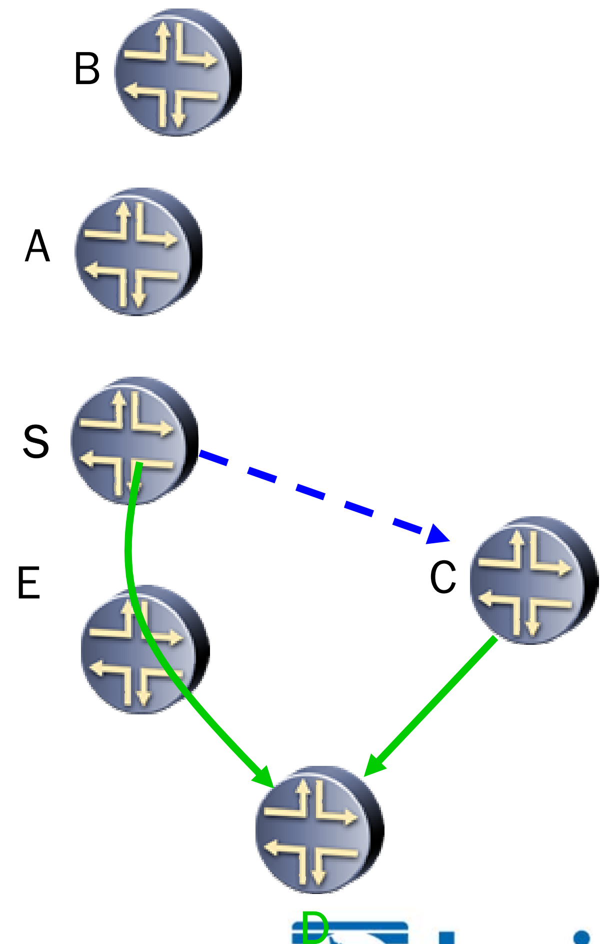
However, if S could reroute **directly to C**, it would work

IP Tunnels: How They Work

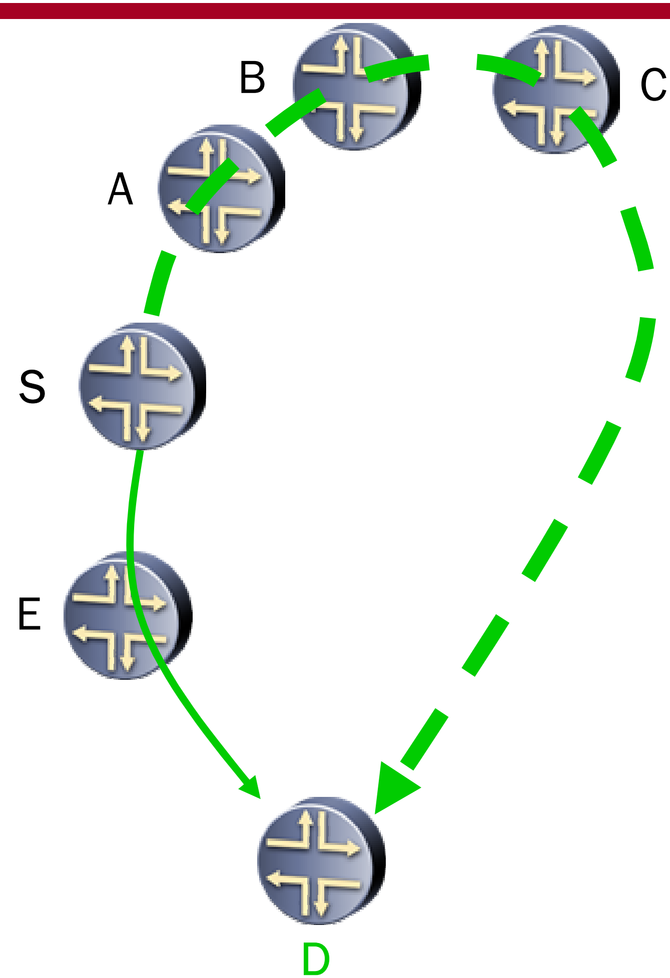
First, find a node X such that
i) X is "below" S,
ii) the destination D is below X
X is a *distant loop-free alternate*

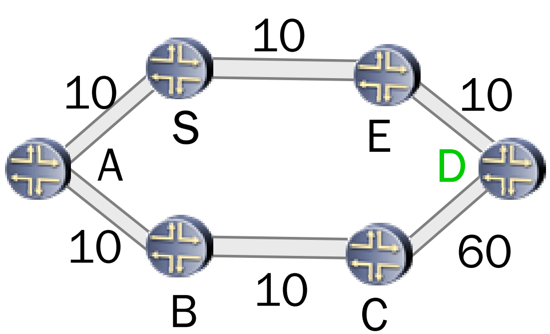
Then, (on failure) encapsulate (tunnel) packets destined to D to X





Source Routed Tunnels





Following the analogy of water flowing downhill, a source-routed tunnel is a **pump strong enough to defy gravity**

MPLS 2004

Juniper NETWORKS

Analysis: Coverage

1. Loop-free alt < U-turn alt < source routing
2. Loop-free alt < IP tunnel (*) < source routing
3. Source routing is provably 100% coverage (assuming the network topology after failure is still connected)
4. U-turn alternatives and IP tunnels are not comparable -- depends on network topology

Note: "x < y" means for **every** graph G, if technique x works, technique y works; AND for *some* graph G, y works, but x does not work

Analysis: Feasibility

- Identifying loop-free alternates requires one "normal" SPF on behalf of each neighbor; a node needs to do this for each of its destinations
- Checking if a node can be a U-turn alternative needs some funky SPF-like computations; a node has to do this for each <neighbor, destination>
- Looking for "distant" loop-free alternates involves doing one "spanning tree" computation, and one "reverse spanning tree" for each destination
- Finding source-routed paths involves one Constrained SPF (CSPF) for every adjacent link

Analysis: Scalability Parameters

- First, some concepts
  - Nexthops: immediate neighbors
  - Next-next-hops: neighbors at a distance of 2
  - Destinations: exit points in a given network
  - Routes: IP prefixes
- Typical numbers:
  - Nexthops: 3-8
  - Next-next-hops: 10-50
  - Destinations: 50-500
  - Routes: 100K-250K

Analysis: Scalability

- Control plane: computations, size of RIB
- Data (forwarding) plane: size of FIB
- Comparisons:
  - Computations done **before** a failure can be relatively long and intense
  - Computations done **in reaction to** a failure must be fast (or better, none)
  - RIB memory is large and cheap
  - FIB memory is small and expensive

The Cost of Protecting a Link L

- Loop-free alternate: for each **destination** whose shortest path traverses **L**, find a loop-free alternate (if any) in case **L** fails
- U-turn alternate:
  - For each **destination** whose shortest path traverses **L**, find a loop-free *or* U-turn alternate
  - For each **neighbor**, see if you could be a U-turn alternate for neighbor; if so, inform neighbor and continue to next step
    - for each **<destination, neighbor>**, install a "diode route" (if available) in the **FIB**

The Cost of Protecting a Link L

- IP tunnels:
  - For each **destination D** whose shortest path traverses **L**, find a *distant* loop-free alternate **X**
  - On failure, encapsulate packets to **D** with an IP destination of **X** (e.g., using IP tunnels)
- Source-routed tunnels
  - For each **nexthop N** on the other side of **L**, find a source routed path to **N** that avoids **L** (CSPF)
  - On failure of **L**, switch packets to **D** to source routed-path (e.g., by pushing an MPLS label)

Solution Comparison

	Coverage	Control Plane	FIB updates	FIB size impact	Sexiness
Loop-free Alts	Low	Low	Medium	Low	Low
U-turn Alts	High	High	Medium	High	High
IP Tunnels	High (*)	High	Medium	Medium	Medium
Source routing	100%	Low	Low	Low	Low

MPLS 2004

Juniper NETWORKS

Caveats

- The analysis done here for U-turn alternates and IP tunnels was fairly unsophisticated
  - Perhaps, on discussion with the respective authors, more sophisticated and efficient computations and/or FIB representations will emerge
  - Documenting the impact on the control plane and on the FIB is a vital part of the analysis
- There is an "allergy" factor (folks coming out in hives if MPLS or source routing is mentioned) that has been deliberately ignored here
  - Not enough value is given to the simplicity of delivering a packet where it was meant to go

Summary

- IP/LDP fast reroute is a real problem
- There are a wealth of solutions
- A well-reasoned analysis of several factors is mandatory before embarking on more detailed solutions and clearly before deployment
  - Much of this is missing in current documents
  - If all else is equal, existing and deployed mechanisms should be given an edge
- This talk is a first step: it presents some of the axes and scalability parameters for the analysis