CS 268: Lecture 11
Inter-domain Routing Protocol

Karthik Lakshminarayanan
UC Berkeley
(substituting for Ion Stoica)

(*slides from Timothy Griffin and Craig Labovitz)

Overview

- **An Introduction to BGP**
- BGP and the Stable Paths problem
- Convergence of BGP in the real world
- The End-to-End Effects of Internet Path Selection

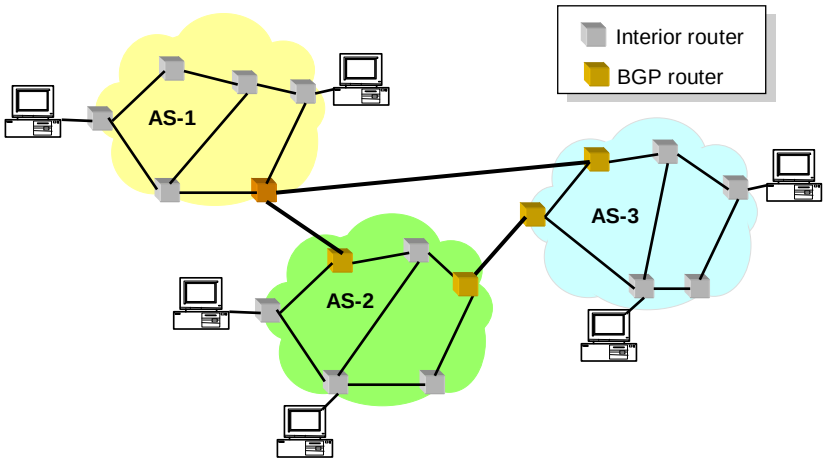
2

Internet Routing

- Internet organized as a two level hierarchy
- First level – autonomous systems (AS's)
  - AS – region of network under a single administrative domain
- AS's run an intra-domain routing protocols
  - Distance Vector, e.g., RIP
  - Link State, e.g., OSPF
- Between AS's runs inter-domain routing protocols, e.g., Border Gateway Routing (BGP)
  - De facto standard today, BGP-4

3

Example



4

Inter-domain Routing basics

- Internet is composed of over 20000 autonomous systems
- **BGP** = **B**order **G**ateway **P**rotocol
  - **Policy-Based** routing protocol
  - **De facto inter-domain routing protocol** of Internet
- Relatively simple protocol, but…
  - complex configuration: entire world can see, and be impacted by, configuration mistakes

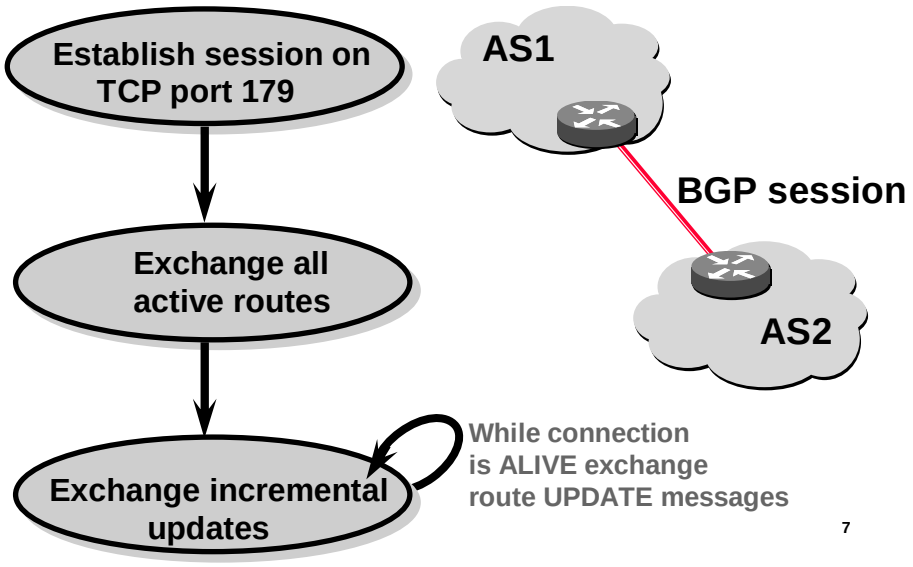
5

BGP Basics

- Use TCP
- Border Gateway Protocol (BGP), based on Bellman-Ford path vector
- AS's exchange reachability information through their BGP routers, only when routes change
- BGP routing information – a sequence of AS's indicating the path traversed by a route
- General operations of a BGP router:
  - **Learns multiple paths**
  - **Picks best path according to its AS policies**
  - **Install best pick in IP forwarding tables**

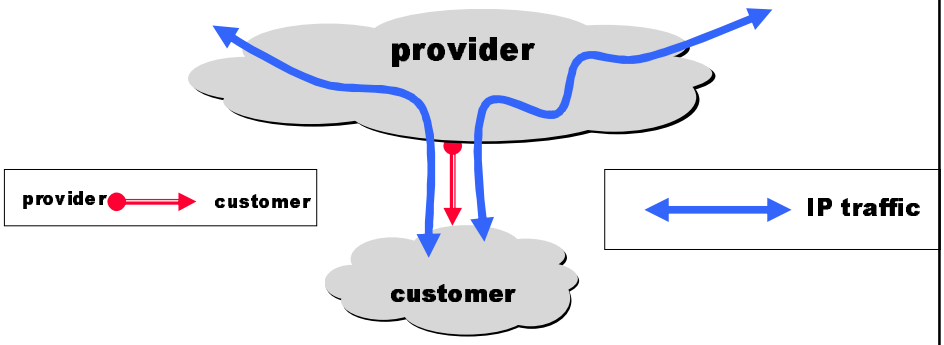
6

BGP Operations (Simplified)



7

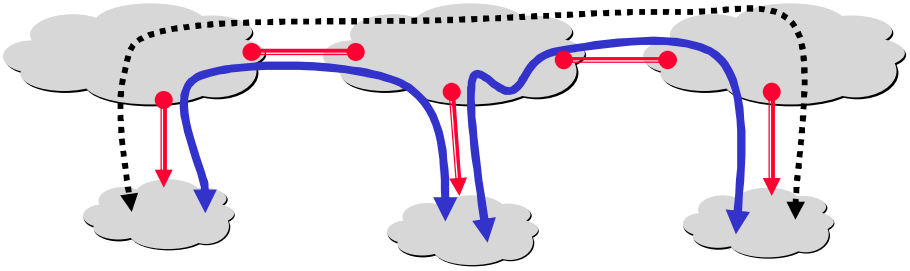
Customers and Providers



Customer pays provider for access to the Internet

8

The "Peering" Relationship



peer — peer
provider → customer

traffic allowed

traffic NOT allowed

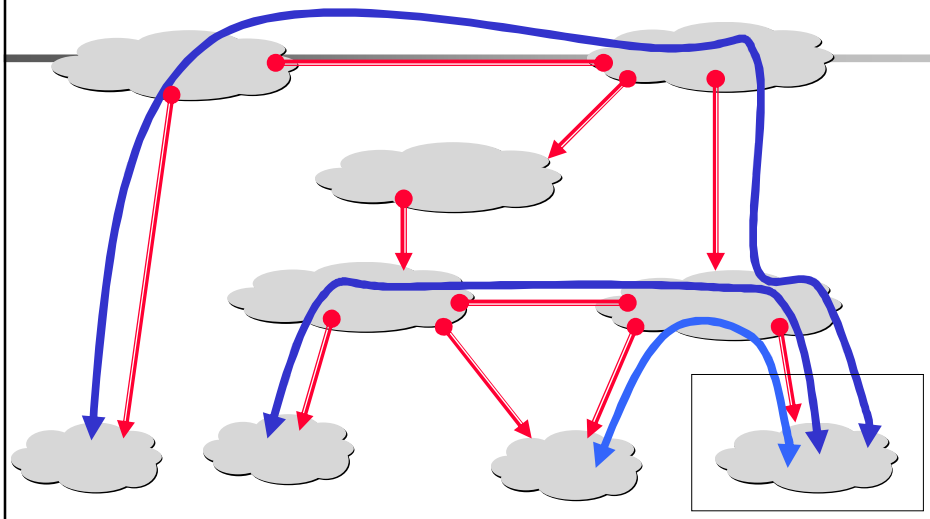
Peers provide transit between their respective customers

Peers do not provide transit between peers

Peers (often) do not exchange $$$

9

Peering Provides Shortcuts



Peering also allows connectivity between the customers of "Tier 1" providers.

peer — peer
provider → customer

10

Peering Wars

Peer

- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet ("Tier 1")

Don't Peer

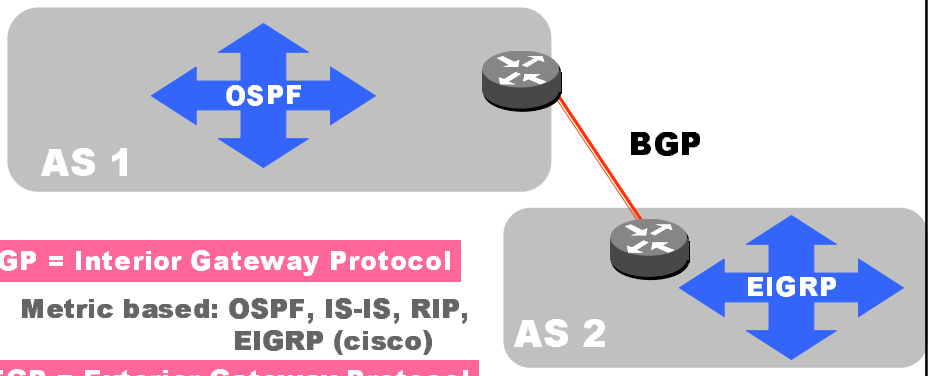
- You would rather have customers
- Peers are usually your competitors
- Peering relationships may require periodic renegotiation

Peering struggles are by far the most contentious issues in the ISP world!

Peering agreements are often confidential.

11

Architecture of Dynamic Routing



IGP = Interior Gateway Protocol

Metric based: OSPF, IS-IS, RIP, EIGRP (cisco)

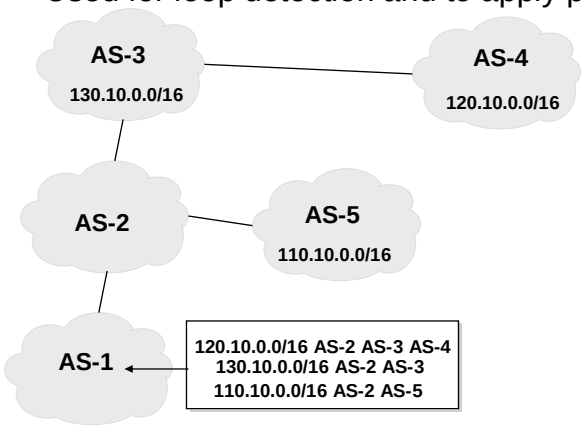
EGP = Exterior Gateway Protocol

Policy based: BGP

12

AS-Path

- Sequence of AS's a route traverses
- Used for loop detection and to apply policy



13

Four Types of BGP Messages

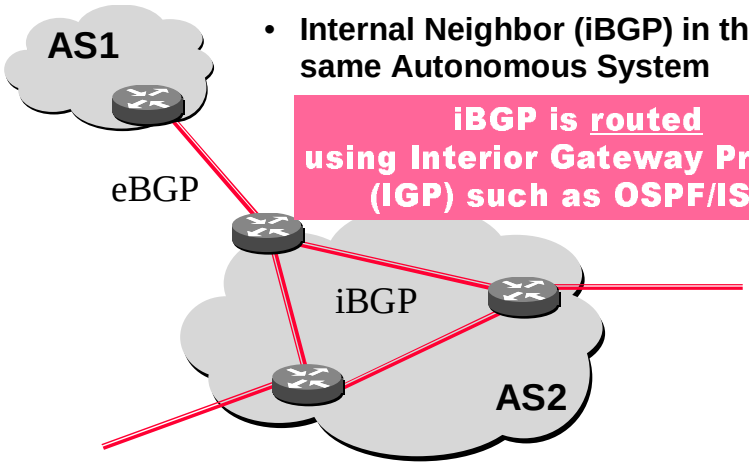
- **Open** : Establish a peering session.
- **Keep Alive** : Handshake at regular intervals.
- **Notification** : Shuts down a peering session.
- **Update** : Announcing new routes or withdrawing previously announced routes.

announcement
=
prefix + attributes values

14

Two Types of BGP Neighbor Relationships

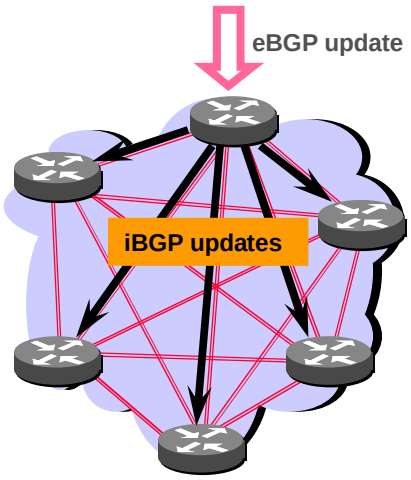
- External Neighbor (eBGP) in a different Autonomous Systems
- Internal Neighbor (iBGP) in the same Autonomous System



iBGP is routed using Interior Gateway Protocol (IGP) such as OSPF/IS-IS

15

iBGP Peers Must be Fully Meshed



iBGP neighbors do not announce routes received via iBGP to other iBGP neighbors.

- iBGP is needed to avoid routing loops within an AS
- Injecting external routes into IGP does not scale and causes BGP policy information to be lost
- BGP does not provide "shortest path" routing

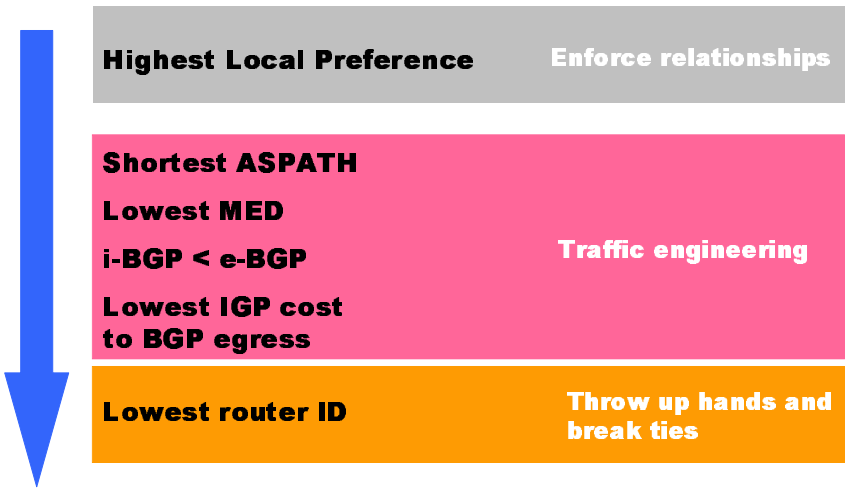
16

Important BGP attributes

- LocalPREF
  - Local preference policy to choose “most” preferred route
- Multi-exit Discriminator (MED)
  - Which peering point to choose?
- Import Rules
  - What route advertisements do I accept?
- Export Rules
  - Which routes do I forward to whom?

17

Route Selection Summary



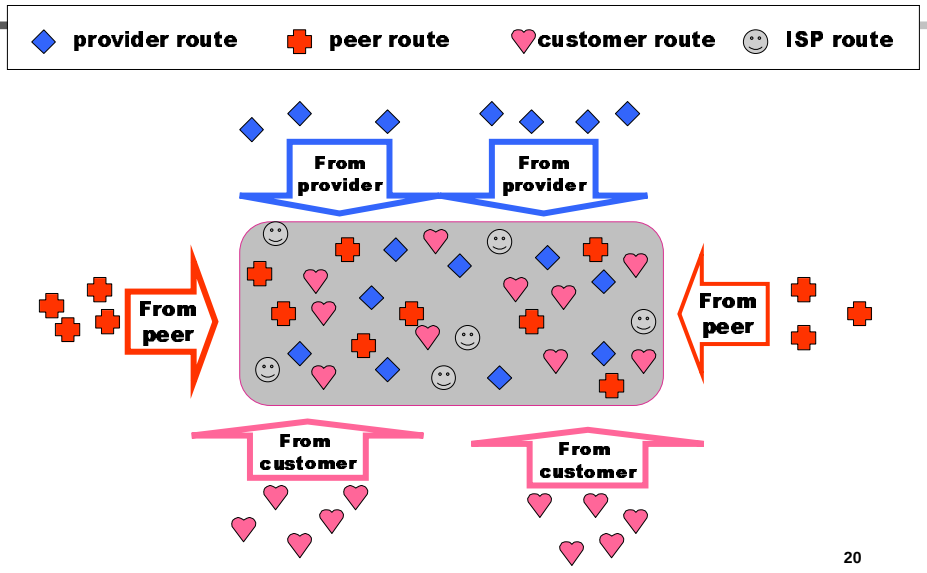
18

Implementing Customer/Provider and Peer/Peer relationships

- Enforce transit relationships
  - Outbound route filtering
- Enforce order of route preference
  - provider < peer < customer

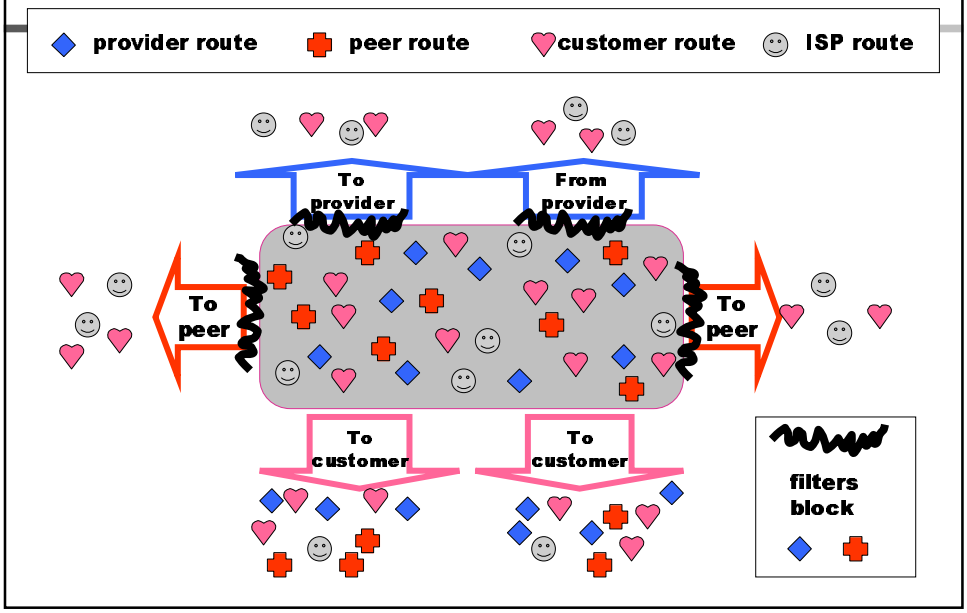
19

Import Routes



20

Export Routes



Overview

- An Introduction to BGP
- **BGP and the Stable Paths problem**
- Convergence of BGP in the real world
- The End-to-End Effects of Internet Path Selection

22

What Problem is BGP solving?

Underlying problem	Distributed means of computing a solution
Shortest Paths	RIP, OSPF, IS-IS
X?	BGP

Having an X can

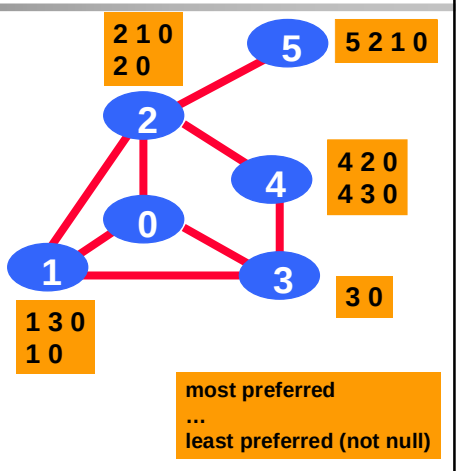
- aid in the design of policy analysis algorithms and heuristics,
- aid in the analysis and design of BGP and extensions,
- help explain some BGP routing anomalies,
- provide a fun way of thinking about the protocol

Our focus

Q : How simple can X get?
A: The *Stable Paths Problem* (SPP)

An instance of the SPP :

- graph of nodes and edges
- node 0, called *the origin*
- for each non-zero node, a set or permitted paths to the origin—this set always contains the "null path"
- ranking of permitted paths at each node—null path is always least preferred

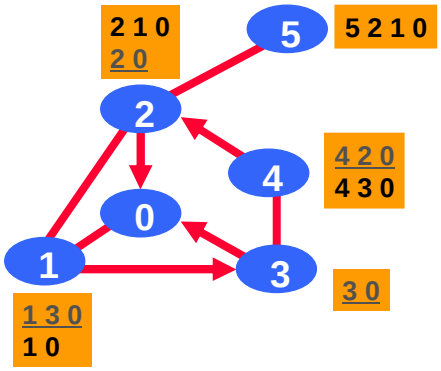


When modeling BGP : nodes represent BGP speaking border routers, and 0 represents a node originating some address block

A Solution to a Stable Paths Problem

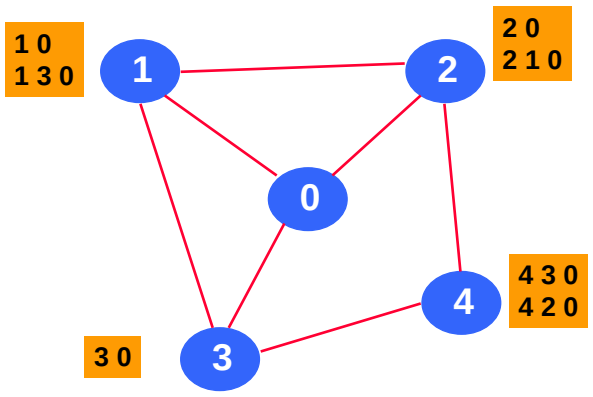
A *solution* is an assignment of permitted paths to each node such that

- node u's assigned path is either the null path or is a path uwP, where wP is assigned to node w and {u,w} is an edge in the graph,
- each node is assigned the highest ranked path among those consistent with the paths assigned to its neighbors.



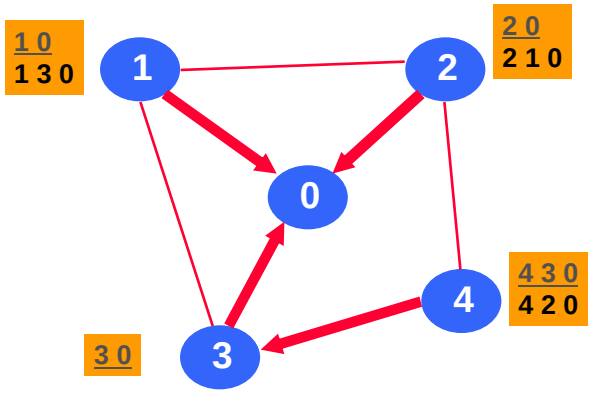
A Solution need not represent a shortest path tree, or a spanning tree.

Example: SHORTEST1



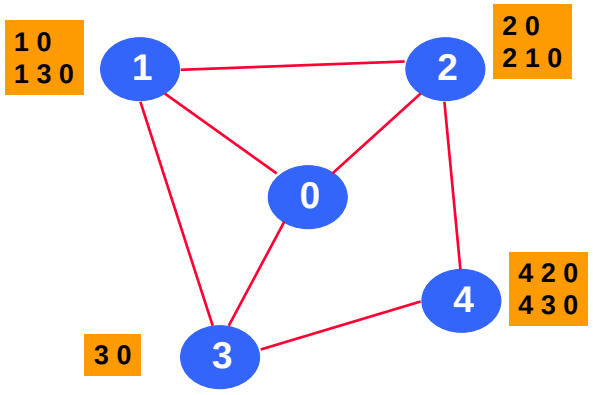
26

Example: SHORTEST1 (Solution)



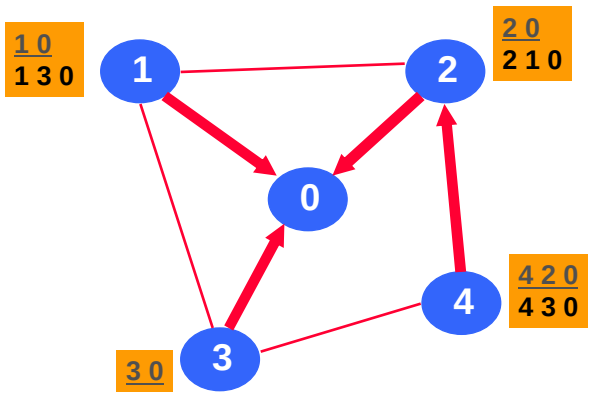
27

Example: SHORTEST2



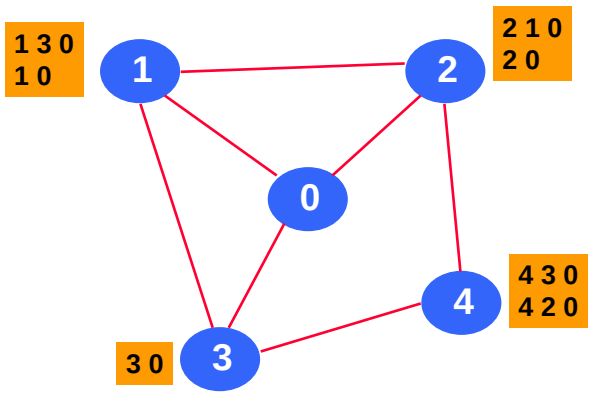
28

Example: SHORTEST2 (Solution)



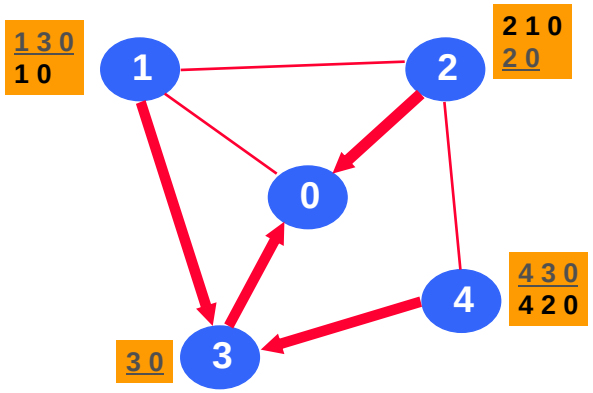
29

Example: GOOD GADGET



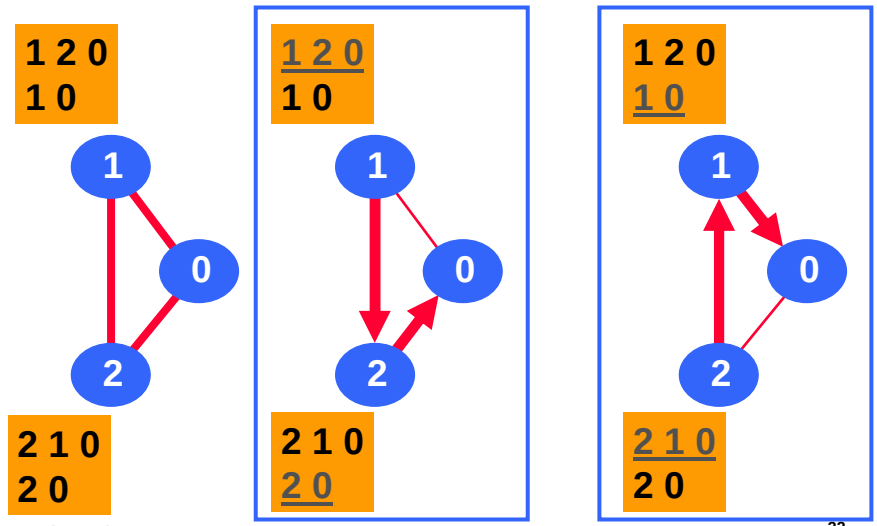
30

Example: GOOD GADGET (Solution)



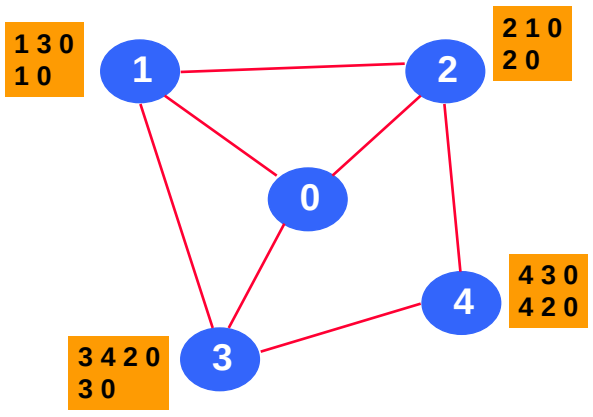
31

A Stable Paths Problem may have multiple solutions



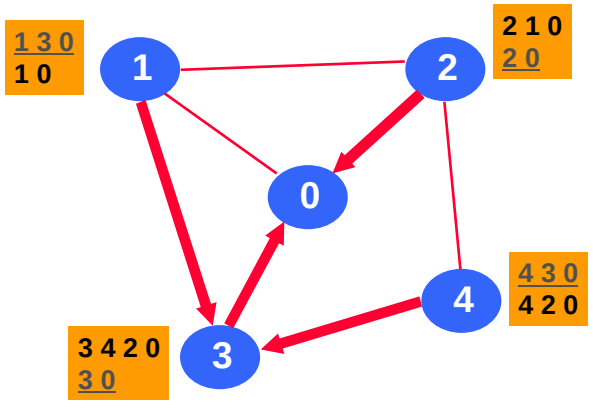
32

Example: NAUGHTY GADGET



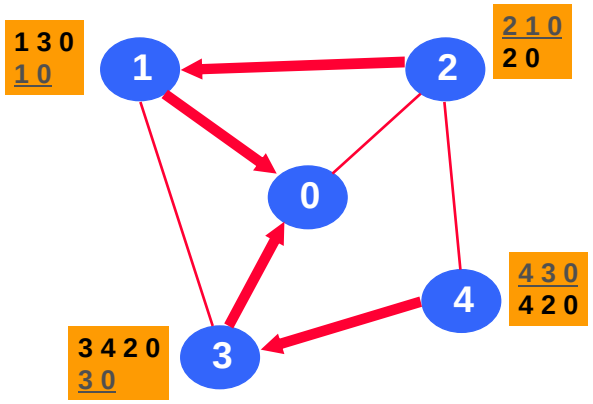
33

Example: NAUGHTY GADGET (Solution 1)



34

Example: NAUGHTY GADGET (Solution 2)

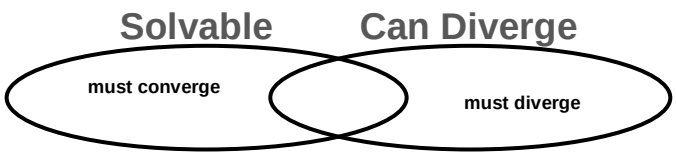


35

SPP explains possibility of BGP divergence

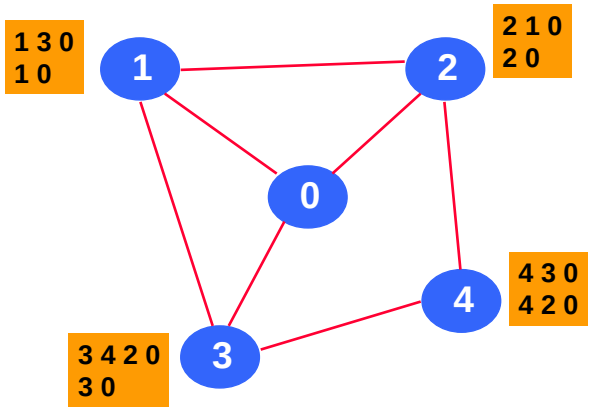
- BGP is not guaranteed to converge to a stable routing. Policy inconsistencies can lead to "livelock" protocol oscillations.
- See "Persistent Route Oscillations in Inter-domain Routing" by K. Varadhan, R. Govindan, and D. Estrin. ISI report, 1996

The SPP view :



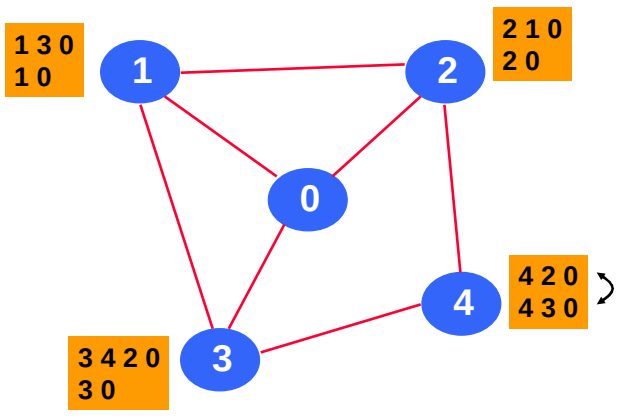
36

Example: NAUGHTY GADGET



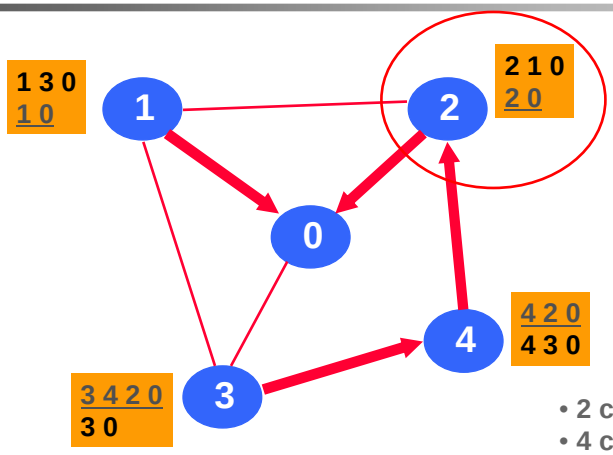
37

Example: BAD GADGET



38

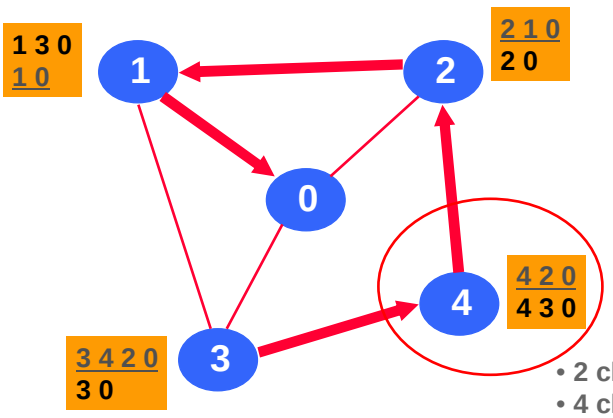
Example: BAD GADGET



- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

39

Example: BAD GADGET

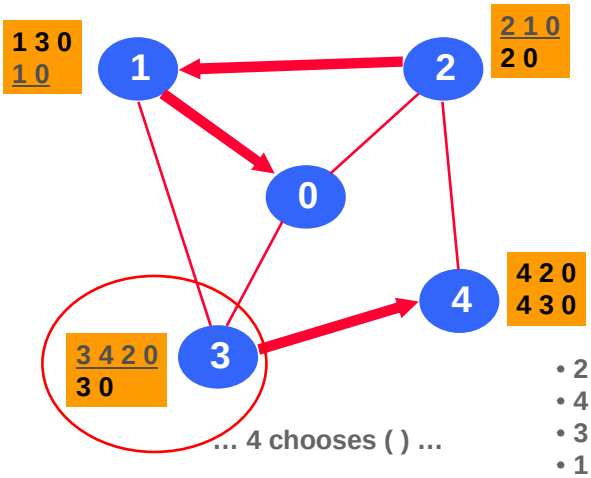


... 2 chooses (2 1 0) ...

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

40

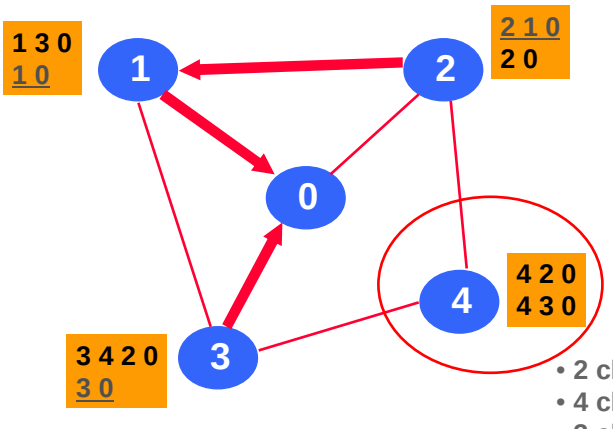
Example: BAD GADGET



- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

41

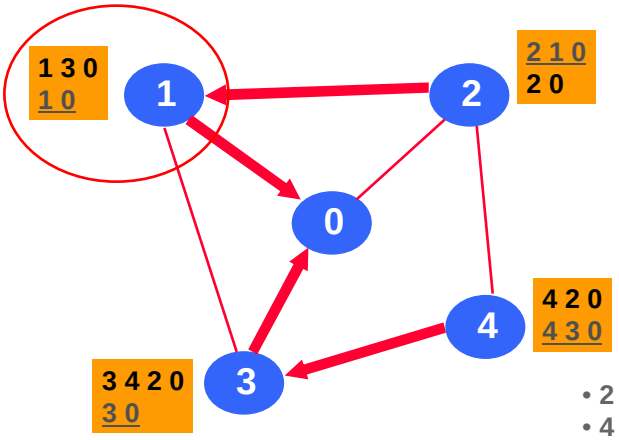
Example: BAD GADGET



- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

42

Example: BAD GADGET

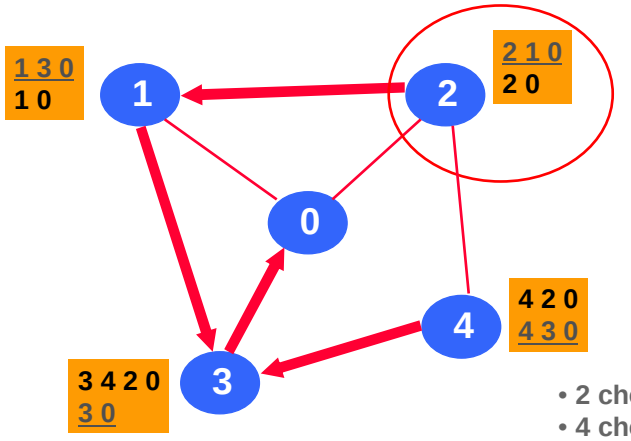


… 4 chooses (4 3 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

43

Example: BAD GADGET

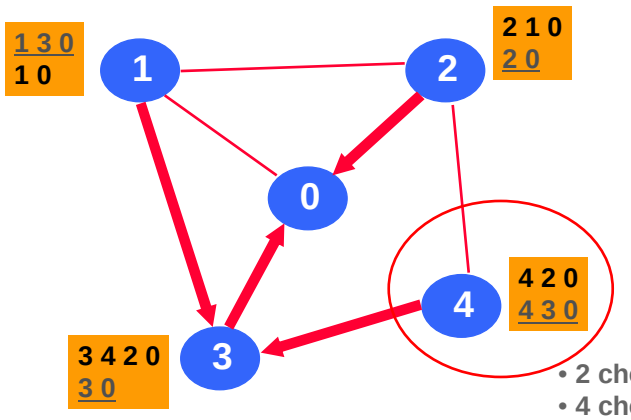


… 1 chooses (1 3 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

44

Example: BAD GADGET

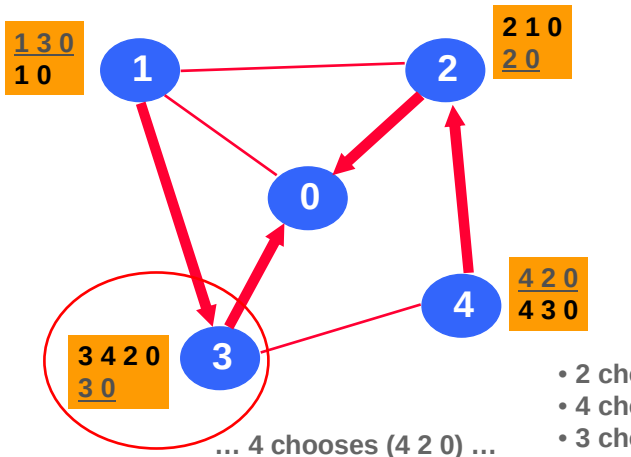


… 2 chooses (2 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

45

Example: BAD GADGET

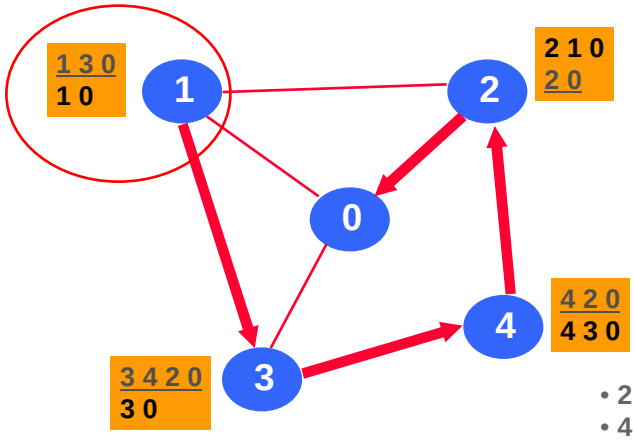


… 4 chooses (4 2 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

46

Example: BAD GADGET

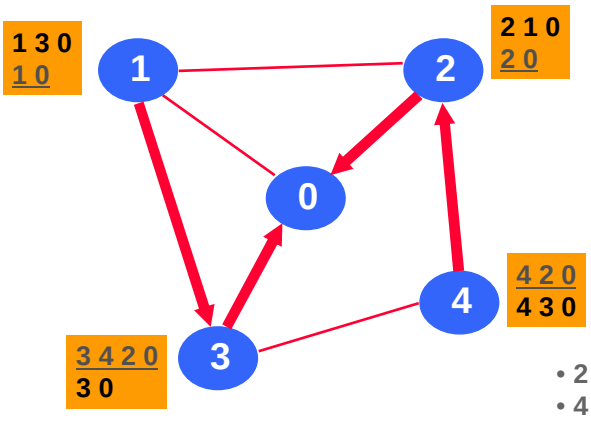


… 3 chooses (3 4 2 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

47

Example: BAD GADGET



… 1 chooses (1 0) …

- 2 chooses (2 0)
- 4 chooses (4 2 0)
- 3 chooses (3 4 2 0)
- 1 chooses (1 0)

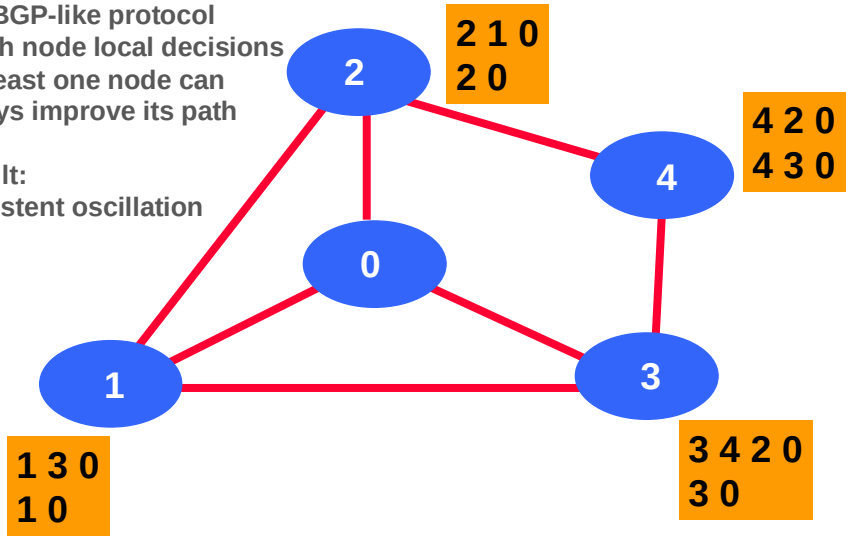
That was one round of oscillation!

48

BAD GADGET : No Solution

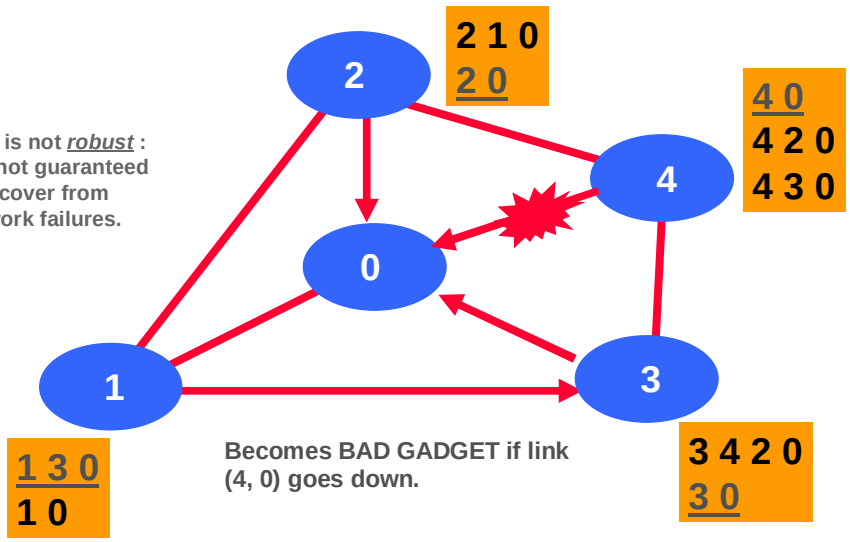
In a BGP-like protocol
- each node local decisions
- at least one node can always improve its path

Result:
persistent oscillation



SURPRISE : Beware of Backup Policies

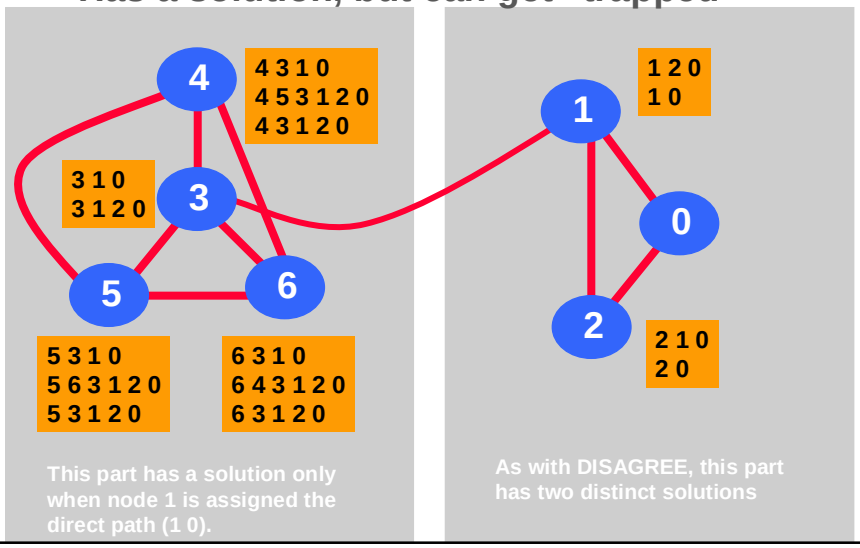
BGP is not *robust* :
it is not guaranteed to recover from network failures.



Becomes BAD GADGET if link (4, 0) goes down.

PRECARIOUS

Has a solution, but can get "trapped"

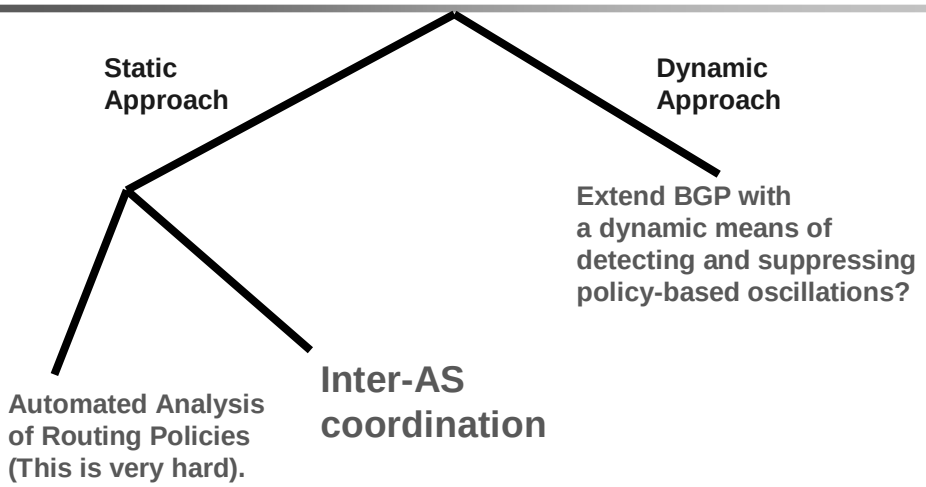


Theoretical Results

- The problem of determining whether an instance of stable paths problem is solvable is NP-complete
- Shortest path route selection is provably safe

52

What is to be done?



These approaches are <u>complementary</u>

53

Overview

- An Introduction to BGP
- BGP and the Stable Paths problem
- **Convergence of BGP in the real world**
- The End-to-End Effects of Internet Path Selection

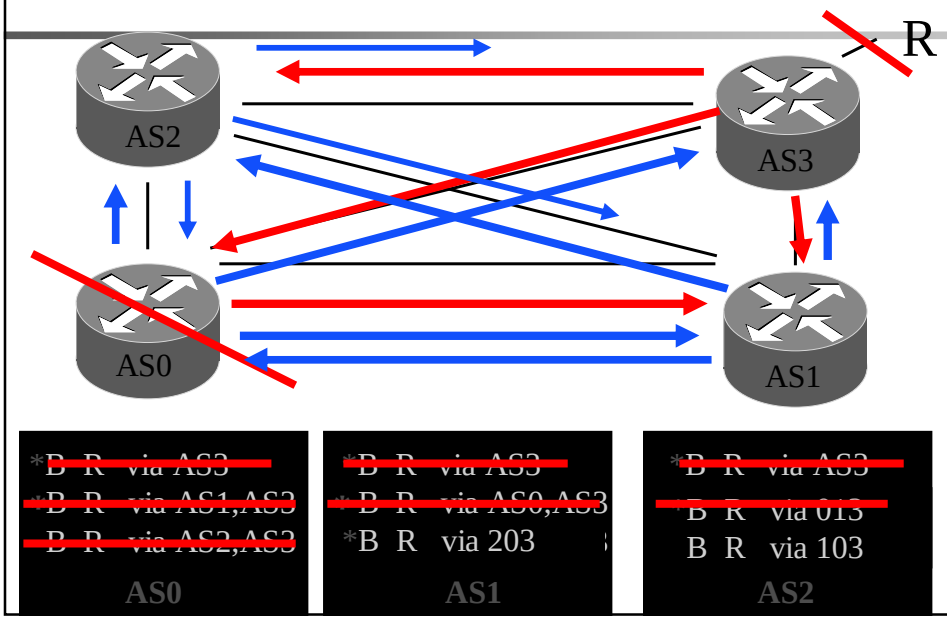
54

Convergence in the real-world?

- [Labovitz99] Experimental results from two year study which measured 150,000 BGP faults injected into peering sessions at several IXPs
- Found:
  - Internet averages 3 minutes to converge after failover
  - Some multihomed failovers (short to long ASPath) require 15 minutes

55

BGP Convergence Example



Convergence Result

- If we assume
  1. unbounded delay on BGP processing and propagation
  2. Full BGP mesh BGP peers
  3. Constrained shortest path first selection algorithm

  There exists possible ordering of messages such that BGP will explore all possible ASPaths of all possible lengths
- Convergence time of BGP is O(N!), where N number of BGP speakers

57

Overview

- An Introduction to BGP
- BGP and the Stable Paths problem
- Convergence of BGP in the real world
- **The End-to-End Effects of Internet Path Selection**

58

End-to-end effects of Path Selection

- Goal of study: Quantify and understand the impact of *path selection* on end-to-end performance
- Basic metric
  - **Let X = performance of default path**
  - **Let Y = performance of best path**
  - **Y-X = cost of using default path**
- Technical issues
  - **How to find the best path?**
  - **How to measure the best path?**

59

Approximating the best path

- Key Idea
  - **Use end-to-end measurements to extrapolate potential alternate paths**
- Rough Approach
  - **Measure paths between pairs of hosts**
  - **Generate synthetic topology – full NxN mesh**
  - **Conservative approximation of best path**
- **Question: Given a selection of N hosts, how crude is this approximation?**

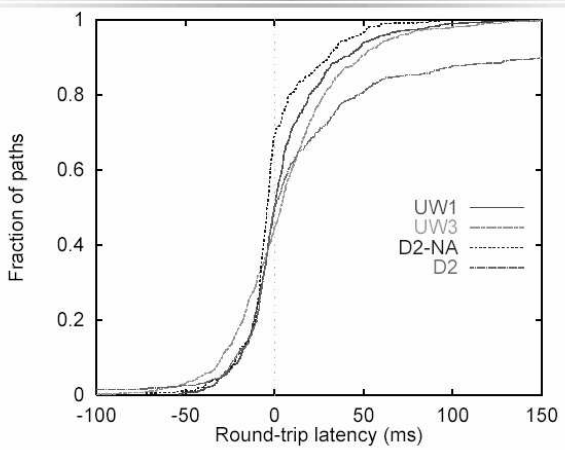
60

Methodology

- For each pair of end-hosts, calculate:
  - Average round-trip time
  - Average loss rate
  - Average bandwidth
- Generate synthetic alternate paths (based on long-term averages)
- For each pair of hosts,graph difference between default path and alternate path

61

Round-trip time

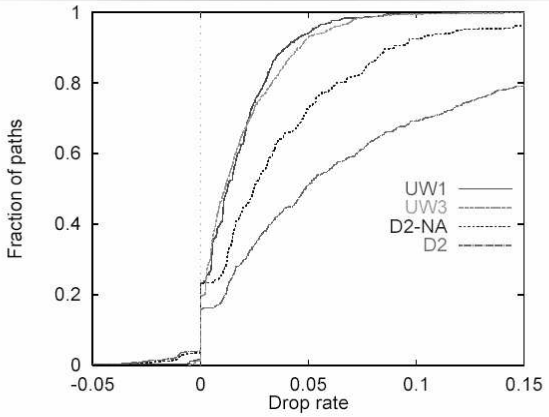


Courtesy:
Stefan Savage

30%-55% of default paths have longer round-trip times

62

Loss rate

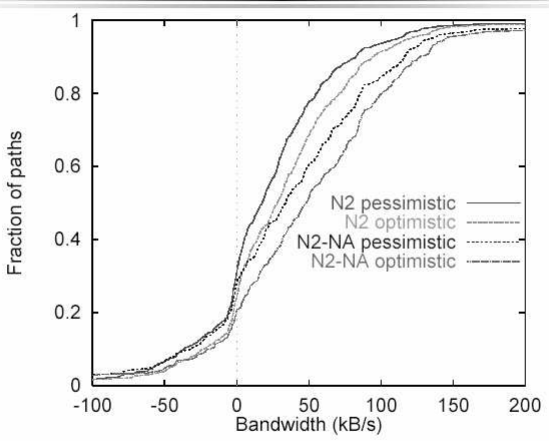


Courtesy:
Stefan Savage

75%-85% of default paths have higher loss rates

63

Bandwidth

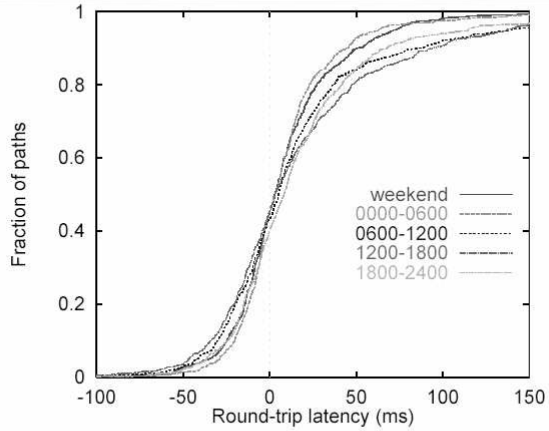


Courtesy:
Stefan Savage

70%-80% of default paths have lower bandwidth

64

Time-of-day variation (latency)



Courtesy: Stefan Savage

Effect stronger during "peak" hours

65

Why Path Selection is imperfect?

- Technical Reasons
  - Single path routing
  - Non-topological route aggregation
  - Coarse routing metrics (AS_PATH)
  - Local policy decisions
- Economic Reasons
  - Disincentive to offer transit
  - Minimal incentive to optimize transit traffic

66