15-744: Computer Networking

L-3 BGP

Next Lecture: Interdomain Routing

- BGP
- Assigned Reading
  - MIT BGP Class Notes
  - [Gao00] On Inferring Autonomous System Relationships in the Internet
    - Ooops…

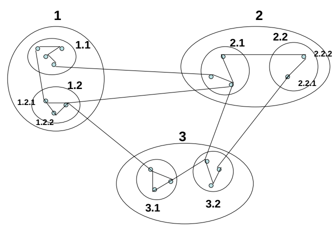
Outline

- Need for hierarchical routing
- BGP
  - ASes, Policies
  - BGP Attributes
  - BGP Path Selection
  - iBGP
  - Inferring AS relationships
- Problems with BGP
  - Convergence
  - Sub optimal routing

Routing Hierarchies

- Flat routing doesn't scale
  - Each node cannot be expected to have routes to every destination (or destination network)
- Key observation
  - Need less information with increasing distance to destination
- Two radically different approaches for routing
  - The area hierarchy
  - The landmark hierarchy

Areas

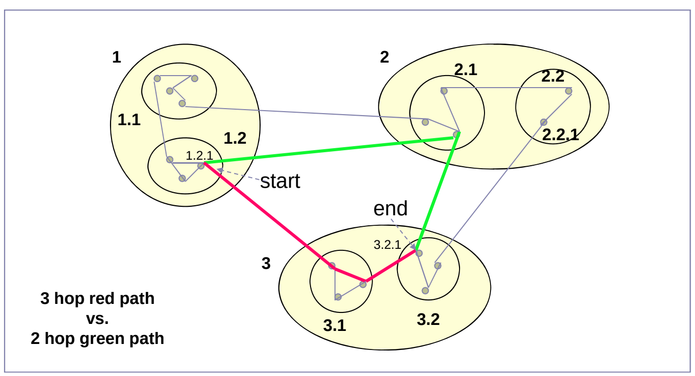


- Divide network into areas
  - Areas can have nested sub-areas
  - Constraint: no path between two sub-areas of an area can exit that area
- Hierarchically address nodes in a network
  - Sequentially number top-level areas
  - Sub-areas of area are labeled relative to that area
  - Nodes are numbered relative to the smallest containing area

Routing

- Within area
  - Each node has routes to every other node
- Outside area
  - Each node has routes for other top-level areas only
  - Inter-area packets are routed to nearest appropriate border router
- Can result in sub-optimal paths

Path Sub-optimality



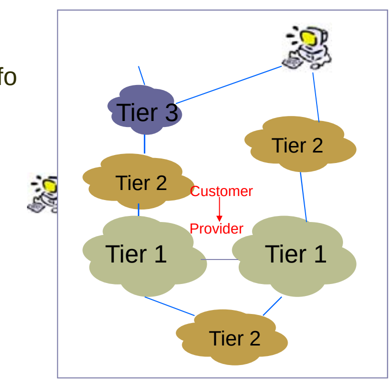
A Logical View of the Internet

- National (Tier 1 ISP)
  - "Default-free" with global reachability info

  Eg: AT & T, UUNET, Sprint
- Regional (Tier 2 ISP)
  - Regional or country-wide

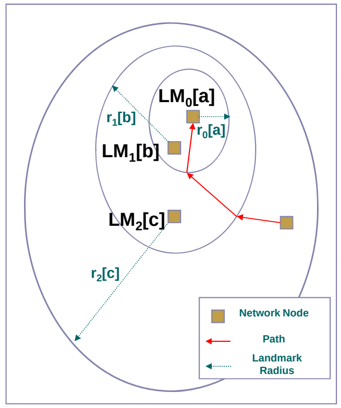
  Eg: Pacific Bell
- Local (Tier 3 ISP)

  Eg: Telerama DSL

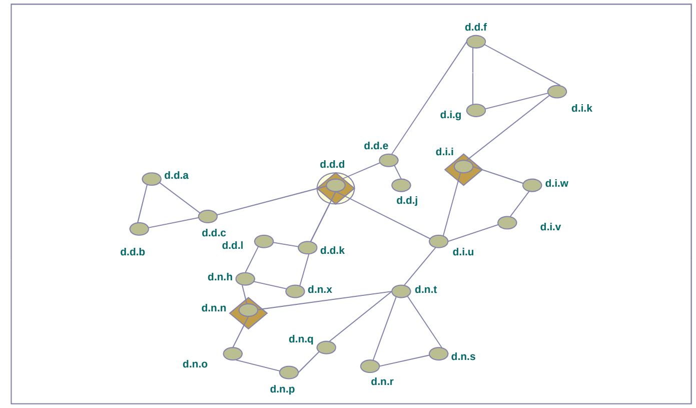


Landmark Routing: Basic Idea

- Source wants to reach $LM_0[a]$, whose address is *c.b.a:*
  - Source can see $LM_2[c]$, so sends packet towards c
  - Entering $LM_1[b]$ area, first router diverts packet to b
  - Entering $LM_0[a]$ area, packet delivered to a
- Not shortest path
- Packet may not reach landmarks



Landmark Routing: Example

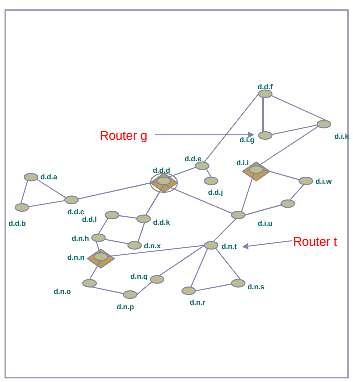


Routing Table for Router g

Landmark	Level	Next hop
$LM_2[d]$	2	f
$LM_1[i]$	1	k
$LM_0[e]$	0	f
$LM_0[k]$	0	k
$LM_0[f]$	0	f

r0 = 2, r1 = 4, r2 = 8 hops

- How to go from d.i.g to d.n.t? g-f-e-d-u-t
- How does path length compare to shortest path? g-k-I-u-t



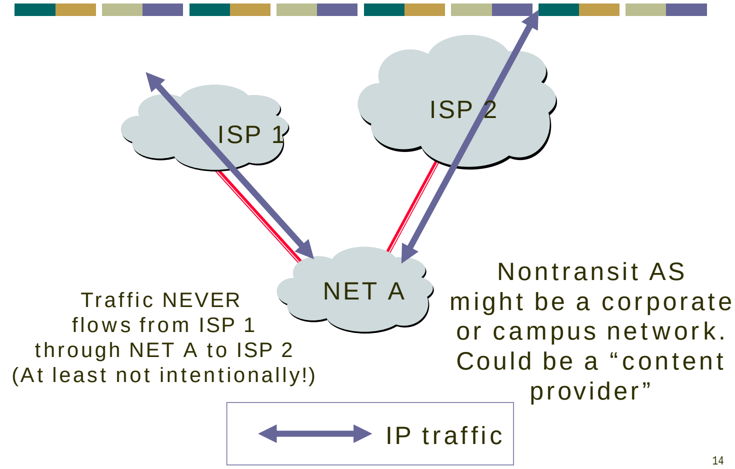
Outline

- Need for hierarchical routing
- BGP
  - ASes, Policies
  - BGP Attributes
  - BGP Path Selection
  - iBGP
  - Inferring AS relationships

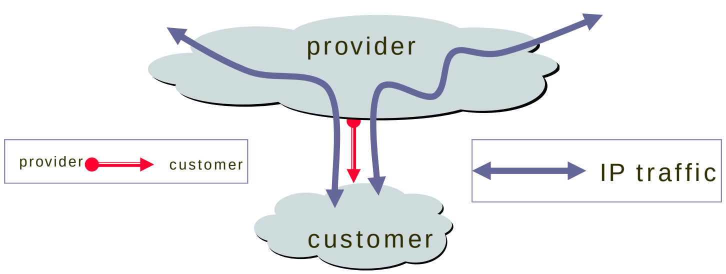
Autonomous Systems (ASes)

- Autonomous Routing Domain
  - Glued together by a common administration, policies etc
- Autonomous system – is a specific case of an ARD
  - ARD is a concept vs AS is an actual entity that participates in routing
  - Has an unique 16 bit ASN assigned to it and typically participates in inter-domain routing
- Examples:
  - MIT: 3, CMU: 9
  - AT&T: 7018, 6341, 5074, …
  - UUNET: 701, 702, 284, 12199, …
  - Sprint: 1239, 1240, 6211, 6242, …
- How do ASes interconnect to provide global connectivity
- How does routing information get exchanged

Nontransit vs. Transit ASes

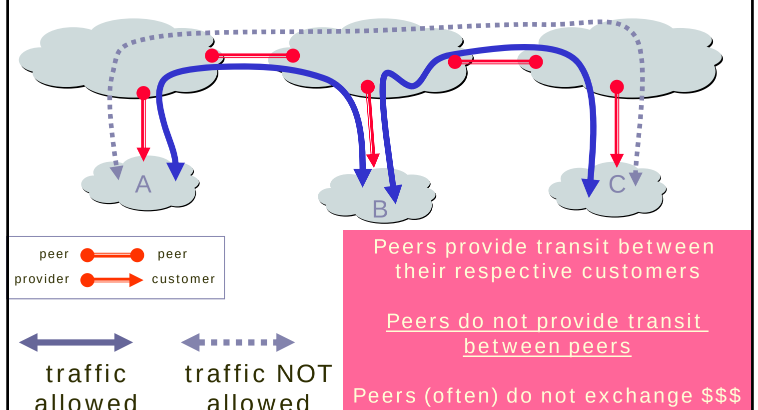


Customers and Providers



Customer pays provider for access to the Internet

The Peering Relationship



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 competition
- Peering relationships may require periodic renegotiation

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

Peering agreements are often confidential.

Routing in the Internet

- Link state or distance vector?
  - No universal metric – policy decisions
- Problems with distance-vector:
  - Bellman-Ford algorithm may not converge
- Problems with link state:
  - Metric used by routers not the same – loops
  - LS database too large – entire Internet
  - May expose policies to other AS's

Solution: Distance Vector with Path

- Each routing update carries the entire path
- Loops are detected as follows:
  - When AS gets route check if AS already in path
    - If yes, reject route
    - If no, add self and (possibly) advertise route further
- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops

BGP-4

- BGP = Border Gateway Protocol
- Is a Policy-Based routing protocol
- Is the EGP of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.

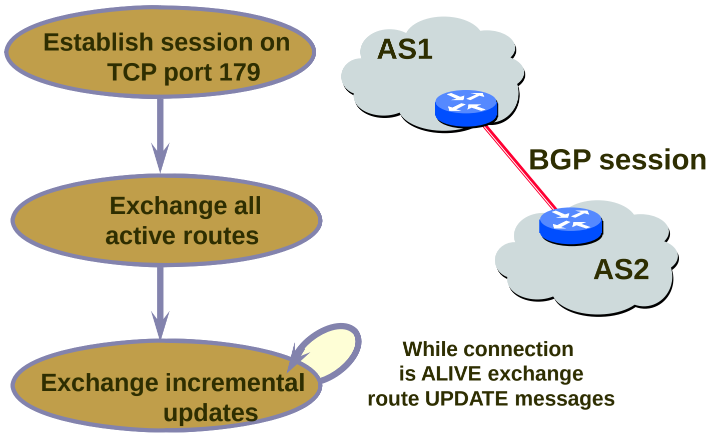
1989 : BGP-1 [RFC 1105]
- Replacement for EGP (1984, RFC 904)

1990 : BGP-2 [RFC 1163]

1991 : BGP-3 [RFC 1267]

1995 : BGP-4 [RFC 1771]
- Support for Classless Interdomain Routing (CIDR)

BGP Operations (Simplified)



21

Interconnecting BGP Peers

- BGP uses TCP to connect peers
- Advantages:
  - Simplifies BGP
  - No need for periodic refresh - routes are valid until withdrawn, or the connection is lost
  - Incremental updates
- Disadvantages
  - Congestion control on a routing protocol?
  - Inherits TCP vulnerabilities!
  - Poor interaction during high load

22

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

23

Policy with BGP

- BGP provides capability for enforcing various policies
- Policies are **not** part of BGP: they are provided to BGP as configuration information
- BGP enforces policies by choosing paths from multiple alternatives and controlling advertisement to other AS's
- Import policy
  - What to do with routes learned from neighbors?
  - Selecting best path
- Export policy
  - What routes to announce to neighbors?
  - Depends on relationship with neighbor

24

Examples of BGP Policies

- A multi-homed AS refuses to act as transit
  - Limit path advertisement
- A multi-homed AS can become transit for some AS's
  - Only advertise paths to some AS's
  - Eg: A Tier-2 provider multi-homed to Tier-1 providers
- An AS can favor or disfavor certain AS's for traffic transit from itself

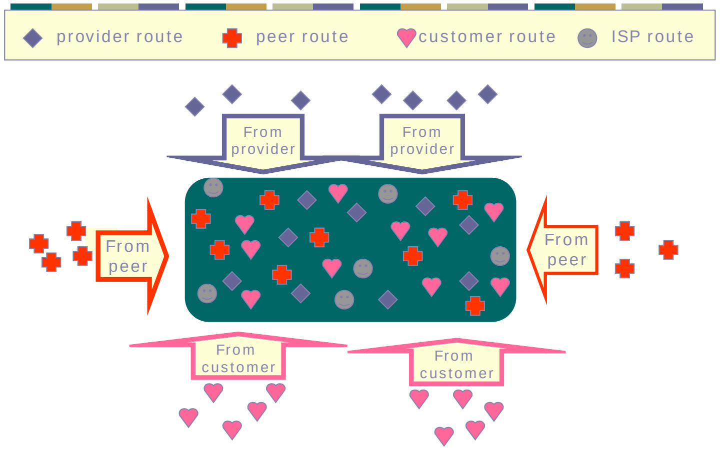
25

Export Policy

- An AS exports only best paths to its neighbors
  - Guarantees that once the route is announced the AS is willing to transit traffic on that route
- To Customers
  - Announce all routes learned from peers, providers and customers, and self-origin routes
- To Providers
  - Announce routes learned from customers and self-origin routes
- To Peers
  - Announce routes learned from customers and self-origin routes

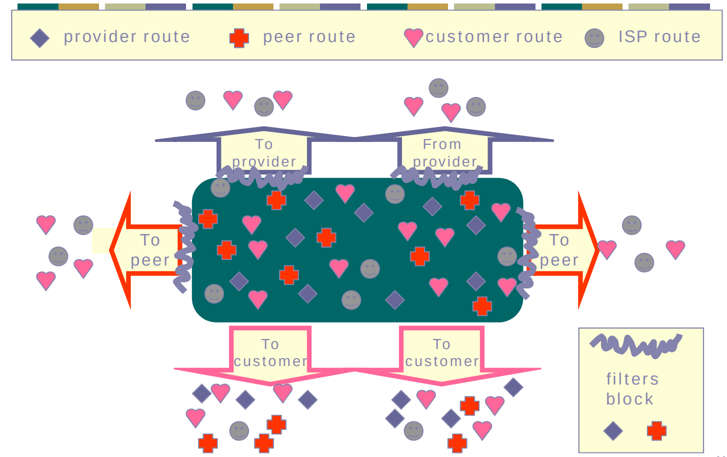
26

Import Routes



27

Export Routes



28

BGP UPDATE Message

- List of withdrawn routes
- Network layer reachability information
  - List of reachable prefixes
- Path attributes
  - Origin
  - Path
  - Metrics
- All prefixes advertised in message have same path attributes

Path Selection Criteria

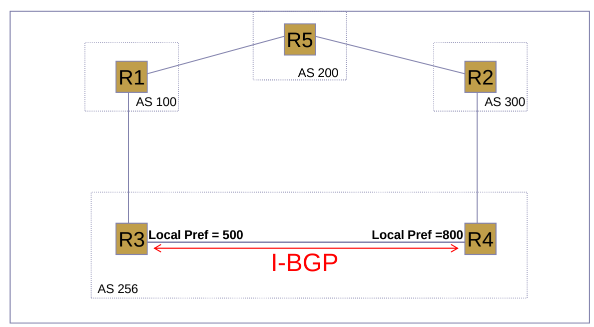
- Information based on path attributes
- Attributes + external (policy) information
- Examples:
  - Hop count
  - Policy considerations
    - Preference for AS
    - Presence or absence of certain AS
  - Path origin
  - Link dynamics

Important BGP Attributes

- Local Preference
- AS-Path
- MED
- Next hop

LOCAL PREF

- Local (within an AS) mechanism to provide relative priority among BGP routers

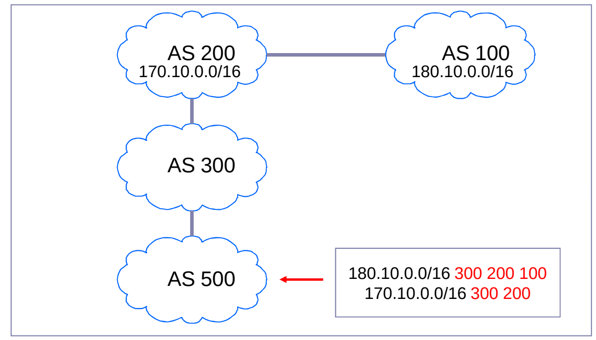


LOCAL PREF – Common Uses

- Handle routes advertised to multi-homed transit customers
  - Should use direct connection (multihoming typically has a primary/backup arrangement)
- Peering vs. transit
  - Prefer to use peering connection, why?
- In general, customer > peer > provider
  - Use LOCAL PREF to ensure this

AS_PATH

- List of traversed AS's
- Useful for loop checking and for path-based route selection (length, regexp)

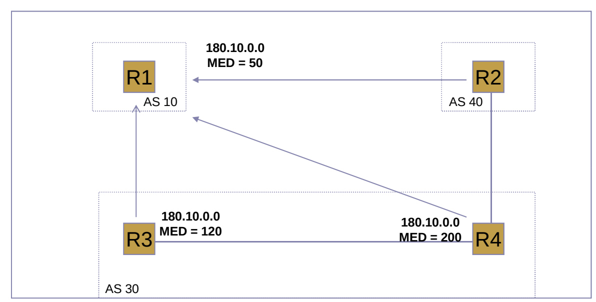


Multi-Exit Discriminator (MED)

- Hint to external neighbors about the preferred path into an AS
  - Non-transitive attribute
  - Different AS choose different scales
- Used when two AS's connect to each other in more than one place

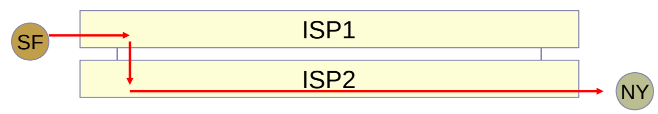
MED

- Typically used when two ASes peer at multiple locations
- Hint to R1 to use R3 over R4 link
- Cannot compare AS40's values to AS30's



MED

- MED is typically used in provider/subscriber scenarios
- It can lead to unfairness if used between ISP because it may force one ISP to carry more traffic:



- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way

37

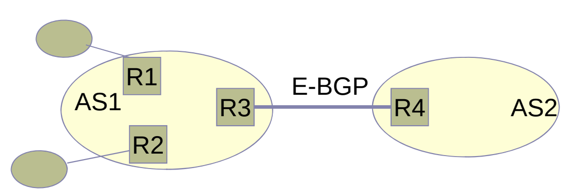
Route Selection Process

Highest Local Preference	Enforce relationships
Shortest ASPATH Lowest MED i-BGP < e-BGP Lowest IGP cost to BGP egress	Traffic engineering
Lowest router ID	Throw up hands and break ties

38

Internal vs. External BGP

- BGP can be used by R3 and R4 to learn routes
- How do R1 and R2 learn routes?
- Option 1: Inject routes in IGP
    - Only works for small routing tables
- Option 2: Use I-BGP



39

Internal BGP (I-BGP)

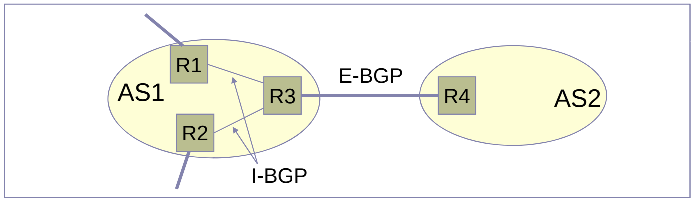
- Same messages as E-BGP
- Different rules about re-advertising prefixes:
    - Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but
    - Prefix learned from one I-BGP neighbor cannot be advertised to another I-BGP neighbor
    - Reason: no AS PATH within the same AS and thus danger of looping.

40

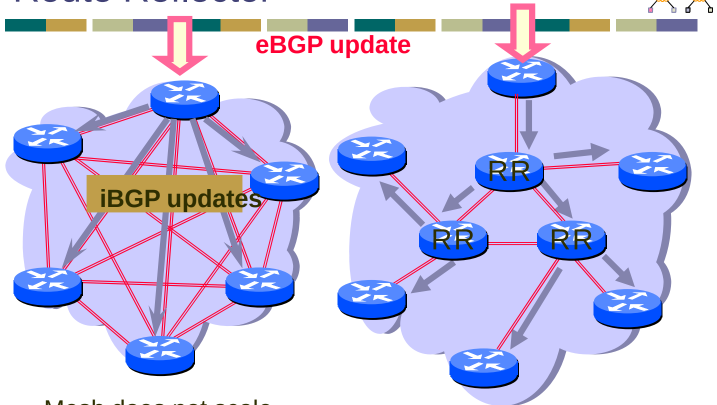
Internal BGP (I-BGP)

- R3 can tell R1 and R2 prefixes from R4
- R3 can tell R4 prefixes from R1 and R2
- R3 cannot tell R2 prefixes from R1

- R2 can only find these prefixes through a *direct connection* to R1
- Result: I-BGP routers must be fully connected (via TCP)!
  - contrast with E-BGP sessions that map to physical links



Route Reflector



Mesh does not scale

Each RR passes only best routes, no longer N^2 scaling problem

Policy Impact

- Different relationships – Transit, Peering
- Export policies → selective export
- "Valley-free" routing
  - Number links as (+1, 0, -1) for customer-to-provider, peer and provider-to-customer
  - In any path should only see sequence of +1, followed by at most one 0, followed by sequence of -1

How to infer AS relationships?

- Can we infer relationship from the AS graph
  - From routing information
  - From size of ASes /AS topology graph
  - From multiple views and route announcements
- [Gao01]
  - Three-pass heuristic
  - Data from University of Oregon RouteViews
- [SARK01]
  - Data from multiple vantage points

[Gao00] Basic Algorithm

- Phase 1: Identify the degrees of the ASes from the tables
- Phase 2: Annotate edges with "transit" relation
  - AS u transits traffic for AS v if it provides its provider/peer routes to v.
- Phase 3: Identify P2C, C2P, Sibling edges
  - P2C → If and only if u transits for v, and v does not, Sibling otherwise
  - Peering relationship ?

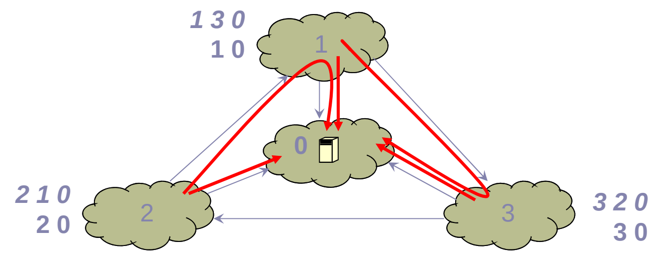
How does Phase 2 work?

- Notion of Valley free routing
  - Each AS path can be
    - Uphill
    - Downhill
    - Uphill – Downhill
    - Uphill – P2P
    - P2P -- Downhill
    - Uphill – P2P – Downhill
- How to identify Uphill/Downhill
  - Heuristic: Identify the highest degree AS to be the end of the uphill path (path starts from source)

Next Lecture: Congestion Control

- Congestion Control
- Assigned Reading
  - [Chiu & Jain] Analysis of Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks
  - [Jacobson and Karels] Congestion Avoidance and Control

Safety: No Persistent Oscillation



Varadhan, Govindan, & Estrin, "Persistent Route Oscillations in Interdomain Routing", 1996

Main Idea of Optional Paper

- Permit only two business arrangements
  - Customer-provider
  - Peering
- Constrain both filtering and ranking based on these arrangements to guarantee safety
- Surprising result: these arrangements correspond to today's (common) behavior

Gao & Rexford, "Stable Internet Routing without Global Coordination", *IEEE/ACM ToN*, 2001