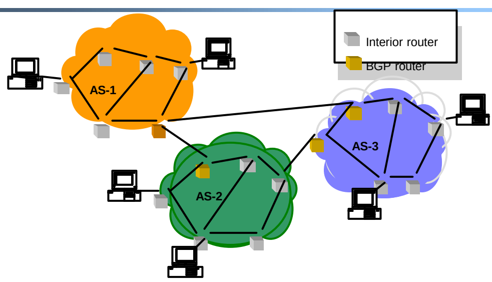
Internet Routing Dynamics

CS589 Lecture 2
Z. Morley Mao
Jan 11, 2004

Two types of Internet Routing Protocols

- Internet consists of roughly 19,000 **Autonomous Systems**
- What is an Autonomous system (AS)?
  - A network belonging to single administrative entity
  - With unified routing policies
- Intradomain routing protocol: **within** an Autonomous System
  - Distance Vector, e.g., RIP
  - Link State, e.g., OSPF, IS-IS
- Interdomain routing protocol: **between** Autonomous Systems
  - Border Gateway Protocol (BGP-4)
  - Path vector protocol

Intradomain routing vs. Interdomain routing



Intra-domain Routing Protocols Link state vs. distance vector

- Uses unreliable datagram delivery
  - Flooding at layer 2
- **Distance vector**
  - Routing Information Protocol (RIP), Bellman-Ford based
  - Each router periodically exchange reachability information with its neighbors
  - Minimal communication overhead
  - Takes long to converge, i.e., in proportion to the maximum path length
  - Has count to infinity problems
- **Link state**
  - Open Shortest Path First Protocol (OSPF), based on Dijkstra
  - Each router periodically floods immediate reachability information to other routers
  - Fast convergence
  - High communication and computation overhead

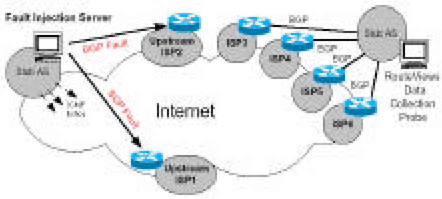
Inter-domain Routing BGP

- Use TCP for reliable transport
- Path vector protocol
- Routing messages indicate changes, no refreshes
- BGP routing information
  - AS path: a sequence of AS's indicating the path traversed by a route;
  - next hop
  - other attributes
- General operations of a BGP router:
  - Learns multiple paths
  - Picks best path according to its AS policies based on BGP decision process
  - Install best pick in IP forwarding tables

Internet Routing Instability [Labovitz et al 2000]

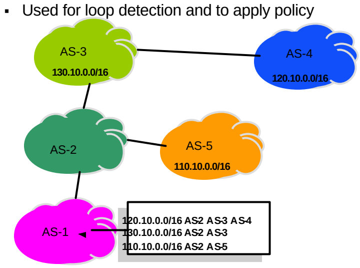
- Methodology
  - Collect routing messages from five public exchange points
- Problems caused by routing instability
  - Increased delays, packet loss and reordering, time for routes to converge (small-scale route changes)
- Relevant BGP information
  - AS-Path
  - Next hop: Next hop to reach a network
- Two routes are the same if they have the same AS-Path and Next hop
  - Other attributes (e.g., MED, communities) ignored for now

Measurement methodology



AS-Path

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

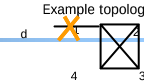


BGP Information Exchange

- **Announcements**: a router has either
  - Learned of a new route, or
  - Made a policy decision that it prefers a new route
- **Withdrawals**: a router concludes that a network is no longer reachable
  - Explicit: associated to the withdrawal message
  - Implicit: (in effect announcement) when a route is replaced as a result of an announcement message
- In steady state BGP updates should be only the result of infrequent policy changes
  - BGP is stateful requires no refreshes
  - Update rate: indication of network stability

Example of delayed convergence

Example topology:



	stage 0	1	4	9
node 2:	[1]	[41]	[431]	--
3:	[1]	[41]	[241]	--
4:	[1]	[31]	--	--

Assuming node 1 has a route to a destination, and it withdraws the route:

Stage (msg processed)	Msg queued
0:	1->{2,3,4}W
1: 1->{2,3,4}W	2->{3,4}A[241], 3->{2,4}A[341], 4->{2,3}A[431]
2: 2->{3,4}A[241]	3->{2,4}A[341], 4->{2,3}A[431]
3: 3->{2,4}A[341]	4->{2,3}A[431], 4->{2,3}W
4: 4->{2,3}A[431]	
MinRouteAdver timer expires:	4->{2,3}W, 3->{2,4}A[3241], 2->{3,4}A[2431]
... (omitted)	
9: 3->{2,4}W	

Note: In response to a withdrawal from 1, node 3 sends out 3 messages:

3->{2,4}A[341], 3->{2,4}A[3241], 3->{2,4}W

Types of Inter-domain Routing Updates

- Forwarding instability:
  - may reflect topology changes
- Policy fluctuations (Routing instability):
  - may reflect changes in routing policy information
- Pathological updates:
  - redundant updates that are neither routing nor forwarding instability
- Instability:
  - forwarding instability and policy fluctuation → change forwarding path

Routing Successive Events (Instability)

- WADiff:
  - a route is explicitly withdrawn as it becomes unreachable, and is later replaced with an alternative route (forwarding instability)
- AADiff:
  - a route is implicitly withdrawn and replaced by an alternative route as the original route becomes unavailable or a new preferred route becomes available (forwarding instability)
- WADup:
  - a route is explicitly withdrawn, and reannounced later (forwarding instability or pathological behavior)

Routing Successive Events (Pathological Instability)

- AADup:
  - A route is implicitly withdrawn and replaced with a duplicate of the original route (pathological behavior or policy fluctuation)
- WWDup:
  - The repeated transmission of BGP withdrawals for a prefix that is currently unreachable (pathological behavior)

Findings

- BGP updates more than one order of magnitude larger than expected
- Routing information dominated by pathological updates
  - Implementation problems:
    - Routers do not maintain the history of the announcements sent to neighbors
    - When a router gets topological changes they just sent these announcements to all neighbors, irrespective of whether the router sent previous announcements about that route to a neighbor or not
  - Self-synchronization – BGP routers exchange information simultaneously → may lead to periodic link/router failures
  - Unconstrained routing policies may lead to persistent route oscillations

Findings

- Instability and redundant updates exhibits strong correlation with load (30 seconds, 24 hours and seven days periods)
  - Overloaded routers fail to respond an their neighbors withdrawn them
- Instability usually exhibits high frequency
- Pathological updates exhibits both high and low frequencies
- No single AS dominates instability statistics
- No correlation between size of AS and its impact on instability statistics
- There is no small set of paths that dominate instability statistics

Conclusions

- Routing in the Internet exhibits many undesirable behaviors
  - Instability over a wide range of time scales
  - Asymmetric routes
  - Network outages
  - Problem seems to worsen
- Many problems are due to software bugs or inefficient router architectures

Lessons

- Even after decades of experience routing in the Internet is not a solved problem
- This attests the difficulty and complexity of building distributed algorithm in the Internet, i.e., in a heterogeneous environment with products from various vendors
- Simple protocols may increase the chance to be
  - Understood
  - Implemented right

Beacons [2003], Motivation: Better understanding of BGP dynamics

- Border Gateway Protocol (BGP)
  - Internet interdomain routing protocol
- Difficult to understand BGP's dynamic behavior
  - Multiple administrative domains
  - Unknown information (policies, topologies)
  - Unknown operational practices
  - Ambiguous protocol specs

Proposal: a controlled active measurement infrastructure for continuous BGP monitoring – **BGP Beacons**.

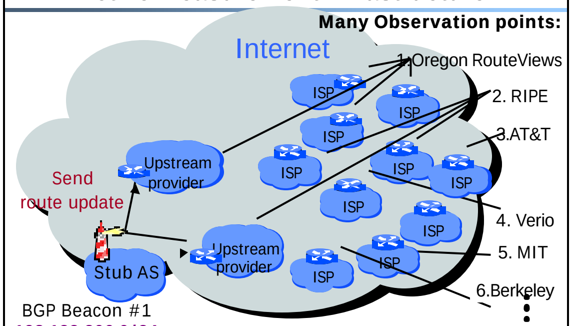
What is a BGP Beacon?

- An unused, globally visible prefix with **known** Announce/Withdrawal schedule
  - For long-term, public use
- For research purposes to study BGP dynamics
  - To calibrate and interpret BGP updates
  - To study convergence behavior
  - To analyze routing and data plane interaction
- Useful to network operators
  - Serve to debug reachability problems
  - Test effects of configuration changes:
    - e.g., flap damping setting

Related work

- Differences from Labovitz's "BGP fault-injector"
  - Long-term, publicly documented
  - Varying advertisement schedule
  - Beacon sequence number (AGG field)
  - Enabler for many research in routing dynamics
- RIPE Ris Beacons
  - Set up at 9 exchange points

Active measurement infrastructure

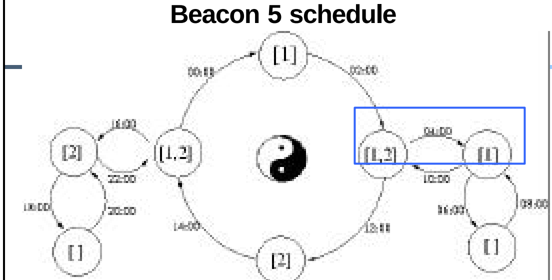


Deployed PSG Beacons

Prefix	Src AS	Start date	Upstream Provider AS	Beacon Host	Beacon Location
198.133.206.0/24	3130	8/10/02	2914, 1239	Randy Bush	WA, US
192.135.183.0/24	5637	9/4/02	3701, 2914	Dave Meyer	OR, US
203.10.63.0/24	1221	9/25/02	1221	Geoff Huston	Australia
198.32.7.0/24	3944	10/24/02	2914, 8001	Andrew Partan	MD, US
192.83.230.0/24	3130	06/12/03	2914, 1239	Randy Bush	WA, US

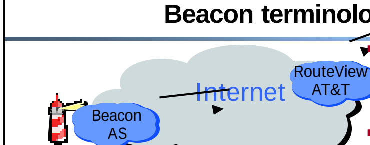
- B1, 2, 3, 5:
  - Announced and withdrawn with a fixed period
  - (2 hours) between updates
    - 1st daily ANN: 3:00AM GMT
    - 1st daily WD: 1:00AM GMT
- B4: varying period, B5: fail-over experiments
- Software available at: http://www.psg.com/~zmao

Beacon 5 schedule



- Live host behind the beacon for data analysis
- Study fail-over behavior for multi-homed customers

Beacon terminology



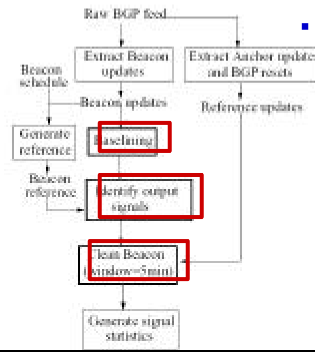
Beacon prefix: 198.133.206.0/24

- **Input signal:** Beacon-injected change
  3:00:00 GMT: Announce (A0)
  5:00:00 GMT: Withdrawal (W)
- **Output signal:**
  5:00:10 A1
  5:00:40 W
  5:01:10 A2
- **Signal length:** number of updates in output signal (3 updates)
- **Signal duration:** Time between first and last update in the signal (5:00:10 -- 5:01:10 60 seconds)
- **Inter-arrival time:** Time between consecutive updates

How to process Beacon data?

- How to identify output signals, ignore external events?
  - Data cleaning
  - Anchor prefix as reference
    - Same origin AS as beacon prefix
    - Statically nailed down
- How to minimize interference btw consecutive input signals?
  - Beacon period is set to 2 hours
- Time stamp and sequence number
  - Attach additional information in the BGP updates
    - Make use of a transitive attribute: Aggregator fields

Beacon data cleaning process

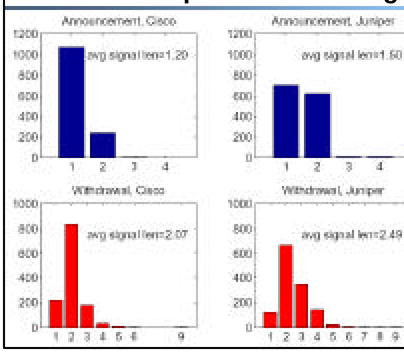


- Goal
  - Clearly identify updates associated with injected routing change
  - Discard beacon events influenced by external routing changes

Beacon example analysis

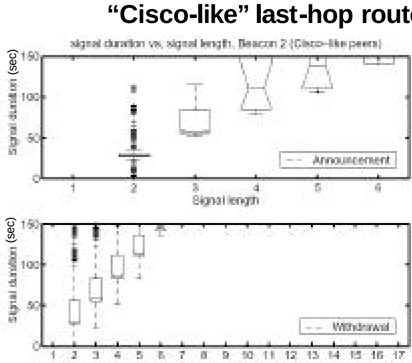
- BGP implementation impact:
  - Cisco vs. Juniper
- Route flap damping analysis
- Convergence analysis
- Inter-arrival time analysis

Cisco vs. Juniper update rate-limiting



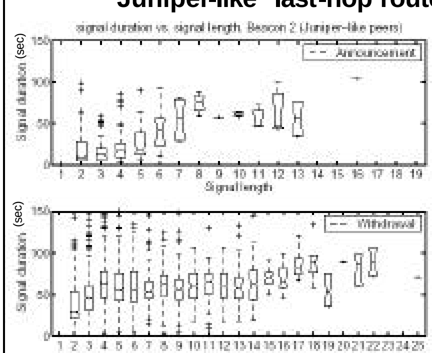
- Known **last-hop Cisco and Juniper routers** from the same AS and location
- **Average signal length:** Average number of updates observed for a **single** beacon-injected change

"Cisco-like" last-hop routers



- Linear increase in signal duration wrt signal length
- Slope=30 sec
- Due to Cisco's default rate-limiting setting

"Juniper-like" last-hop routers



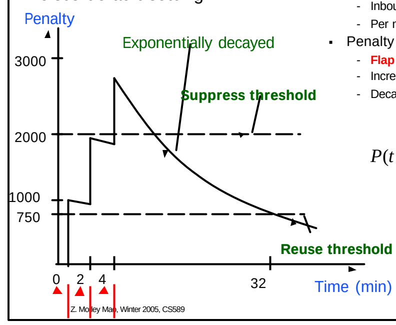
- Signal duration relatively stable wrt increase in signal length
- Shorter signal duration compared to "Cisco-like" last-hops

What is route flap damping?

- A mechanism to punish unstable routes by suppressing them
- RFC2439 [Villamizar et al. 1998]
  - Supported by all major router vendors
  - Believed to be widely deployed [AT&T, Verio]
- Goals:
  - Reduce router processing load due to instability
  - Prevent sustained routing oscillations
  - Do not sacrifice convergence times for well-behaved routes
- There is conjecture a single announcement can cause route suppression.

What is route flap damping?

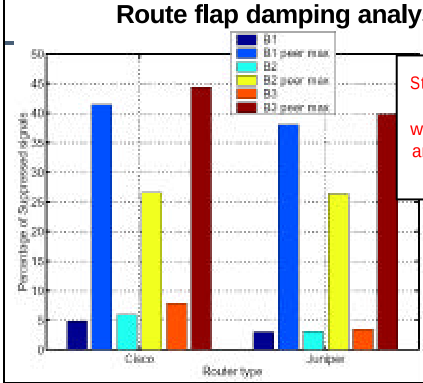
Cisco default setting:



- Scope
  - Inbound external routes
  - Per neighbor, per destination
- Penalty
  - **Flap**: route change
  - Increases for each flap
  - Decays exponentially

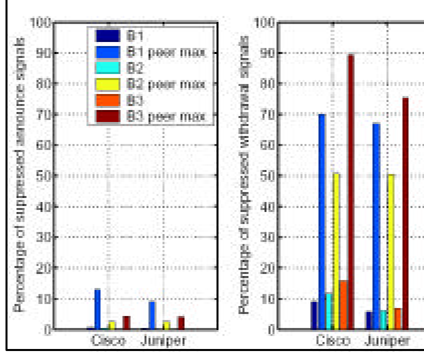
$$P(t') = P(t)e^{-l(t'-t)}$$

Route flap damping analysis



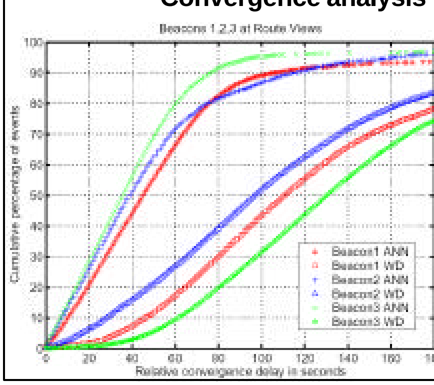
Strong evidence for withdrawal- and announcement-triggered suppression.

Distinguish between announcement and withdrawal



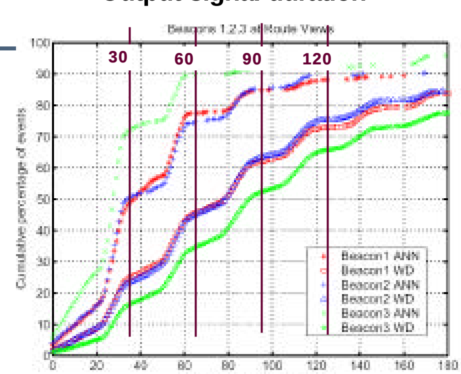
Summary:
- WD-triggered sup more likely than ANN-triggered sup
- Cisco: overall more likely trigger sup than Juniper (AAAW-pattern)
- Juniper: more aggressive for AWAW pattern

Convergence analysis

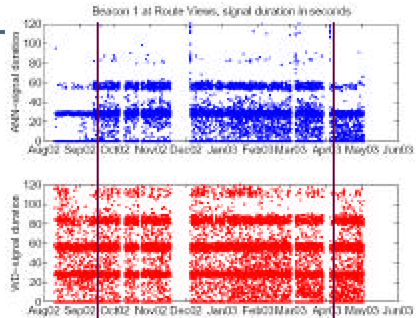


Summary:
- Withdrawals converge slower than announcements
- Most beacon events converge within 3 minutes

Output signal duration

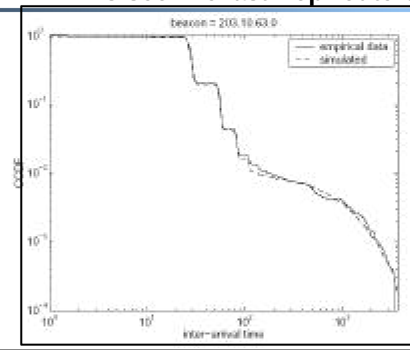


Beacon 1's upstream change



Single-homed (AS2914) **Multi-homed (AS1,2914)** **Multi-homed (AS1239, 2914)**

Inter-arrival time analysis Cisco-like last-hop routers



Complementary cumulative distribution plot

Inter-arrival time modeling

$$X = \begin{cases} 28 * (1 + \mathrm{Geom}(0.81)), & \text{with probability } 0.9524, \\ 1, & \text{with probability } 0.0381, \\ 90 + \mathrm{Exp}(970), & \text{with probability } 0.0095. \end{cases}$$

- Geometric distribution (body):
  - Update rate-limiting behavior: every 30 sec
  - Prob(missing update train) independent of how many already missed
- Mass at 1:
  - Discretization of timestamps for times<1
- Shifted exponential distribution (tail):
  - Most likely due to route flap damping

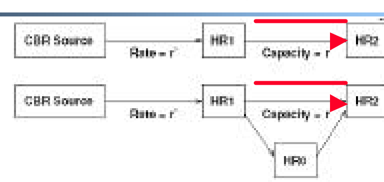
Beacon conclusion

- Beacons -- a public infrastructure for BGP analysis
- Shown examples of Beacon usage
- Future work:
  - Construction of robust and realistic model for BGP dynamics
  - Correlation with data plane
  - Analysis of RIPE Beacons

Routing stability in congested networks (Shaikh 2000)

- Investigate effects of routing control message losses on routing stability
  - Loss due to network congestion
- Previous studies reported correlation between BGP instability and network usage
- Goal: study behavior and evaluate robustness of BGP and OSPF when routing messages are dropped
- Methodology:
  - experimentation and analytical modeling

Network configuration



- Link HR1—HR2 consistently overloaded by CBR traffic
- Packet drop probability at HR1: p=(r'-r)/r'
- HR1—HR2 link overload factor: f=(r'-r)/r

Methodology

- Mean-Time-to-Flap (U2D)
- Mean-Time-to-Recover (D2U)
- Overload factors: 25-400%
- Data packet size: 64, 256, 1500 bytes
- Buffer size at HR: 4MB, 16MB

Analytical models

- Assumptions:
  - The overload factor remains constant
  - Every packet has the same probability of being dropped depending on the overload factor
  - Packet dropping probability is independent for each packet
- Markov chains to find expected values of U2D and D2U for OSPF and BGP

Conclusions

- Developed detailed analytical models
- OSPF's behavior depends only on traffic overload factor
  - Independent of buffer size, packet dropping policy
- BGP's behavior depends on overload factor and RTT
- BGP's resilience to congestion decreases as RTT increases
- There is a need to isolate routing messages from data traffic
  - Through scheduling and buffer management

Lecture summary

- Internet routing is still not well-understood
  - For example, difficult to interpret BGP update messages
  - Holy grail: root cause analysis of BGP updates, need to correlate intradomain and interdomain changes
  - Measurement is useful for understanding routing stability
- Effect of congestion on routing protocols
  - Is TCP the right transport for BGP?
  - How should router treat routing messages differently?