BGP CONVERGENCE

Tawfiq Khan
TCOM 610
George Mason University

The meaning of BGP Convergence

- Time for a router from un-initialized state to fully established state
  - Mostly Up Convergence
  - Mostly for a single router reload or BGP restart
- Time for route changes viewed/accepted by remote peers or global Internet
  - Up Convergence
  - Down Convergence
  - Failover to more specific or longer path

Router BGP Convergence Tuning

- Router BGP Convergence Conditions
  - All routes are accepted, installed in routing table, InQ and OutQ are zero for all peers
- Scenarios
  - Edge routers: receive 250K paths and advertise 500 prefixes
  - Peering Routers: receive 80K paths and advertise 250K prefixes to RR
  - Route reflectors: receive 400K paths and advertise 250K prefixes per clients
- Key Factors:
  - TCP operations, Router Queues, data packaging

TCP Protocol Consideration

- MSS – Max Segment Size
  - Carries as TCP option in SYN packet
  - Cisco default: 536 bytes (RFC 791 for Packet Size < 576 bytes)
  - Safe to increase to 1460 bytes for Ethernet
  - Increasing MSS will reduce the number of packets to send for large number of prefix announcement
  - Should be set to (Path MTU – 40 bytes)
- TCP Window
  - Control the max number of packets before receiving acknowledge
  - Default: 16 KB for Cisco

Queue Optimization

- Goal: minimize packet loss due to overflow, especially for large fan-out of BGP sessions
- Packet reception process: input hold queue (with max depth), selective packet discard (SPD) headroom for high priority packet such as control traffic, system buffer: actual storage
- Hold queue size = WindowSize/(2 * MSS) * PeerCount, “hold-queue 700 in”
- “ip spd mode aggressive”, “ip spd headroom 1000” “ip spd queue min-threshold 998”
- “buffer small permanent 1000”, “buffer small min-free 250”, “buffer small max-free 1375”

Other Optimization

- Peer group: group all BGP sessions with the same outbound policy together -- same BGP messages for all peers in a group
- Dynamic peer group: automatic group identification by Cisco IOS
- Update packaging enhancement: build cache for each peer or update group so that NLRI for each attribute combination can be packed into a single update
- Transmit side loop detection: don’t send updates if the neighbor will deny due to AS_PATH loop detection. Void for MPLS-VPN (new ORF)
- How long to converge for full internet route table? Over 5 minutes, but could be tuned down to 2 minutes

Internet BGP Convergence

- Common Wisdom
  - “Internet routing is robust under faults”
    - Supports path re-routing and restoration on the order of seconds
  - “BGP has good convergence properties”
    - Does not exhibit looping/bouncing problems of RIP
  - “Internet fail-over will improve with faster routers and faster links”
  - “More redundant connections (multi-homing) to Internet will improve site fault-tolerance”
  - “Bad news travels fast, good news can go slow”
- BGP has great convergence properties
  - Modified distance vector protocol: advertise full AS_PATH
  - ASPath solved the convergence and counting to infinity problems
  - Just guarantee no looping, but no fast convergence

Internet Requirements

- Replication, round-robin DNS, etc. helps reliability of inter-domain content oriented services
- Inter-domain transaction oriented services (e.g. VoIP, EBay, database commits, etc.) still pose a challenge
- IP become the ultimate platform for all communications: VoIP, VideoOverIP, triple play, 3G/4G wireless over IP, Skyper, YouTube …
- Need to model how long it takes for the Internet to converge and fully understand Internet convergence property

Routing Protocol Convergence

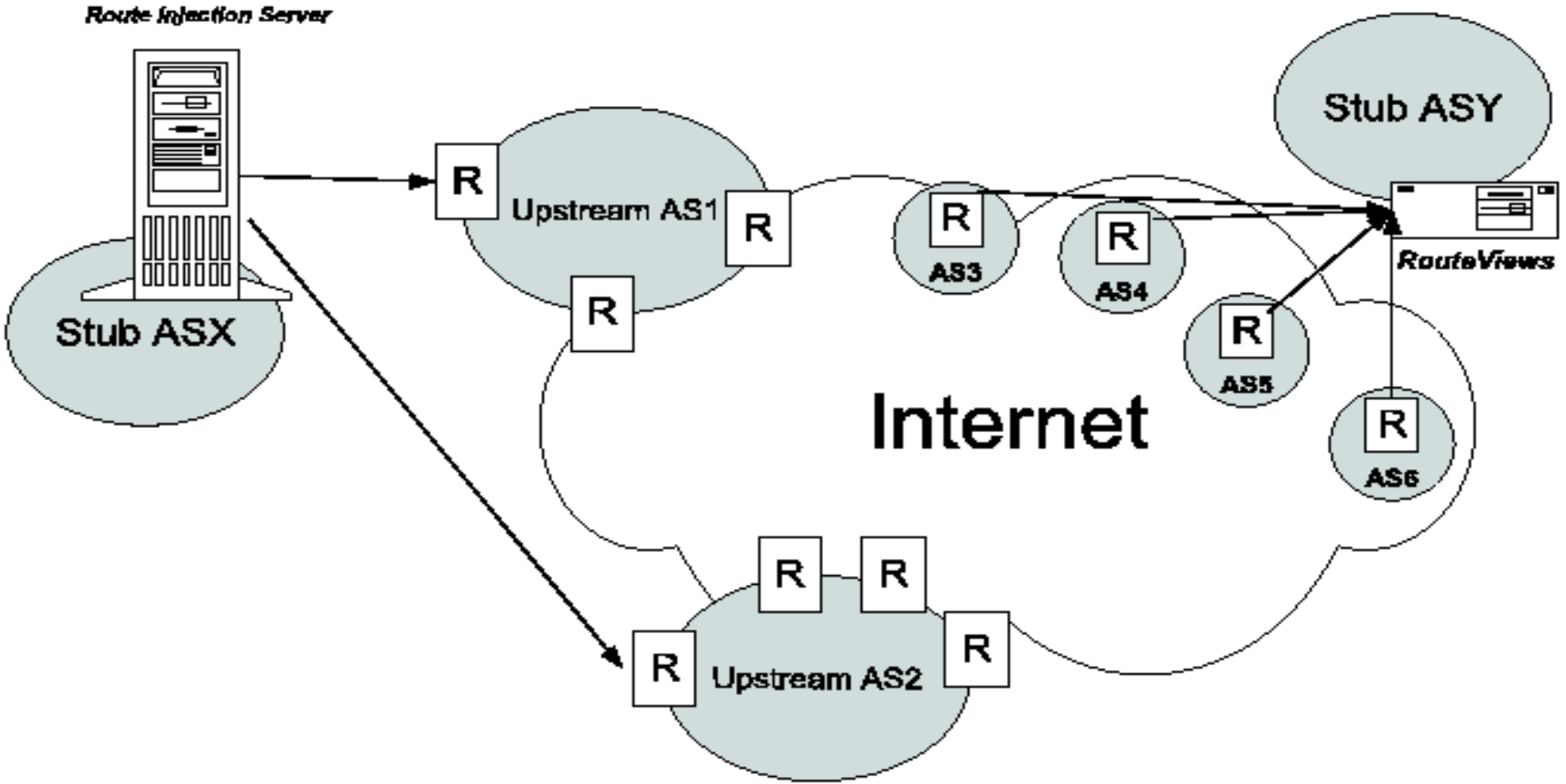
- Unlike connection oriented PSTN (~30 ms), Internet does not have fast, deterministic fail-over
- Instead, each node recalculates on a hop-per-hop basis (i.e. no flooding of changes) and make independent decision
- Distance-vector algorithms (e.g. RIP, BGP) exhibit slower convergence than link state protocols
- During convergence
  - Latency, loss, out of order
  - Micro-looping possible
  - Additional update messages (CPU processing)

Does BGP always converge?

- With unconstrained policies (Griffin99, Varadhan96)
  - Possible Divergence
  - Possible to create mutually un-satisfiable policies
  - NP-complete to identify these policies in IRR
- With constrained policies (e.g. shortest path first)
  - Transient oscillations
  - BGP usually converges
  - It might take a very long time though

BGP Convergence Analysis

- Passive: Route-view project with 30+ peers with full Internet tables, including major Tier1
  - Record all BGP events over multiple years
  - difficult to determine causal relationships
  - Mostly for BGP pathologies and failures
- Active: BGP Beacon and Merit BGP instrument
  - Inject routes into geographically and topologically diverse provider BGP peering sessions (Mae-West, Japan, Michigan, London)
  - Periodically fail and change these routes (i.e. send withdraws or new attributes) in pre-determined intervals
  - Time events using ICMP echos and NTP synchronized BGP “routeviews” monitoring machines (also http gets)
  - Correlate with active ICMP data to top 100 web sites


Route Injection Server
Stub ASX
R
Upstream AS1
R
R
Stub ASY
R
AS3
R
AS4
RouteViews
R
AS5
R
AS6
Internet
R
R
R
Upstream AS2
R

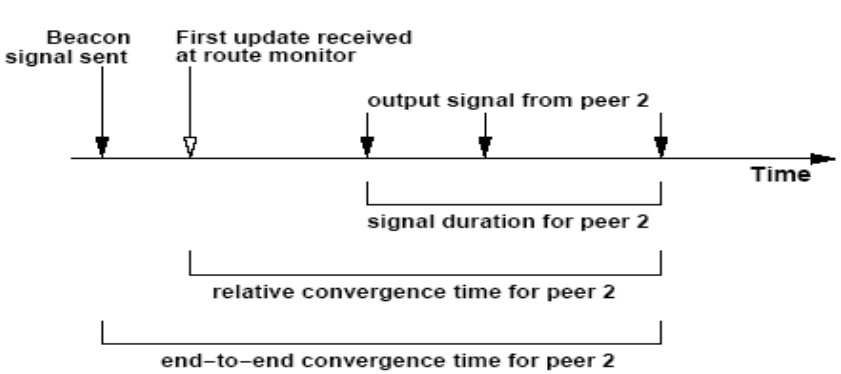
BGP Beacon

- Inject known prefix into Internet table at pre-determined intervals and record Internet response
  - 2 hours interval with periodic announce/withdraw
- Best to NTP synchronize clock from Beacon server and route-view monitors
- 4 PSG Beacons and 8 RIPE Beacons
- PSG Beacon difference:
  - Use aggregator IP address field for timestamp (seconds since the beginning of the month in 10.x.y.x and 0.x.y.z for seconds)
  - Use aggregator ASN number for sequence: 64512 to 65635 (private ASN range)
  - Anchor prefixes: statically pin-up prefixes in the host ASN to correlate network events with Beacons events

BGP Beacons

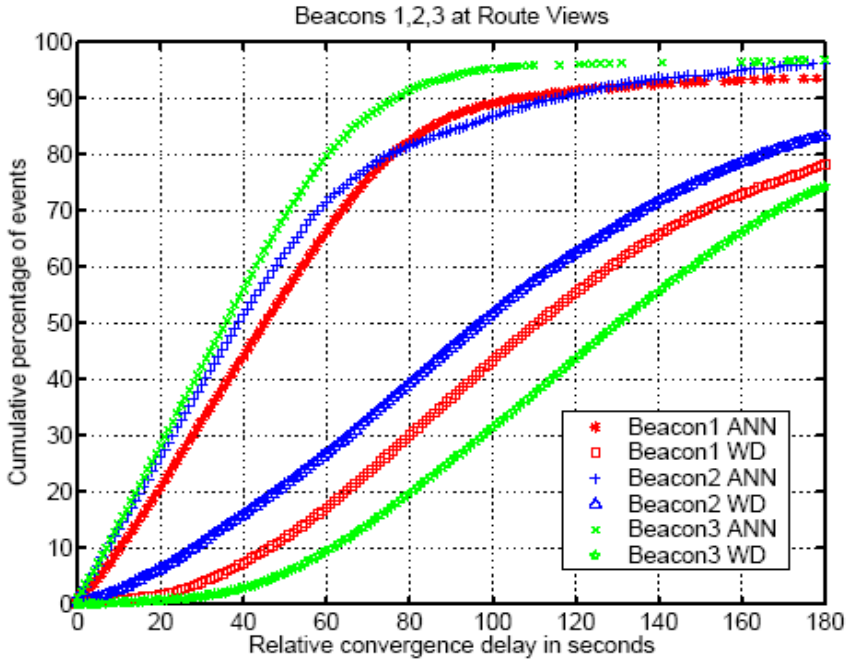
Prefix	Source AS	Upstream	Contact	Start date
198.133.206.0/24	3927	AS2914, AS1	Randy Bush	10-Aug-2002
192.135.183.0/24	5637	AS3701, AS2914	Dave Meyer	4-Sep-2002
203.10.63.0/24	1221	AS1221	Geoff Huston	25-Sep-2002
198.32.7.0/24	3944	AS2914, AS8001	Andrew Partan	24-Oct-2002
195.80.(224+n).0/24	12654	Various	ris@ripe.net	30-Sep-2002

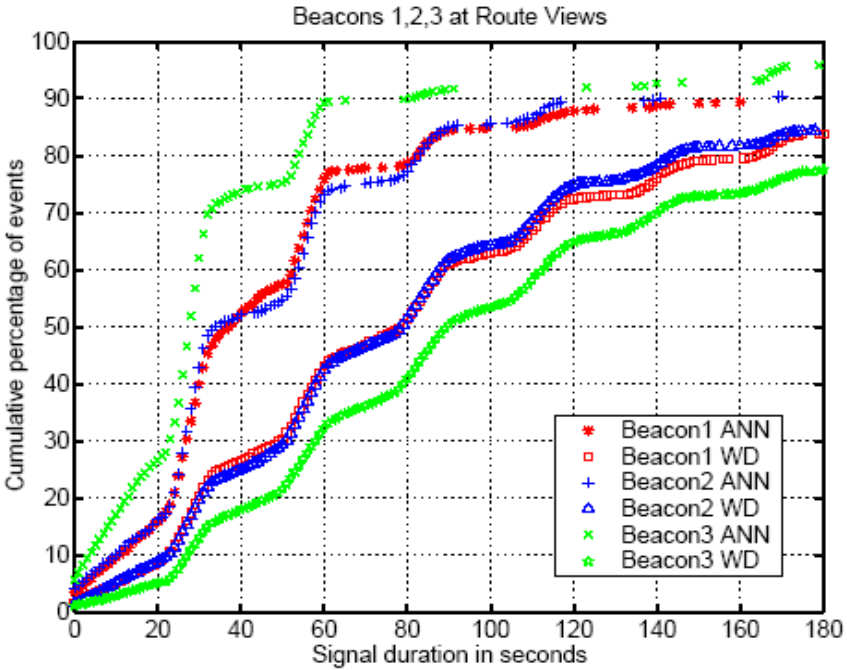
BGP Beacons



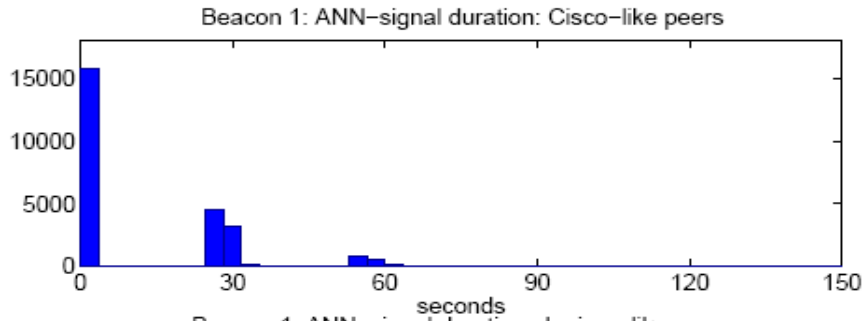
- Relative Convergence time and end-to-end convergence time
- Signal duration, signal latency, and signal length
- Correlate Beacon AS instability within W minutes (= 5 minutes) window to exclude unrelated events
- Not all updates from Beacon sources are visible through all peers

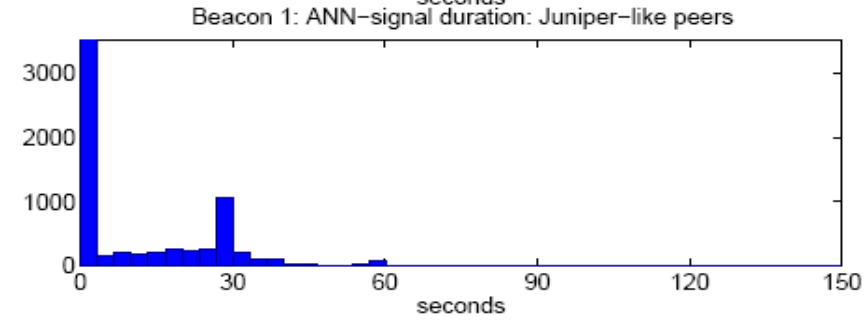
PSG Beacons Result

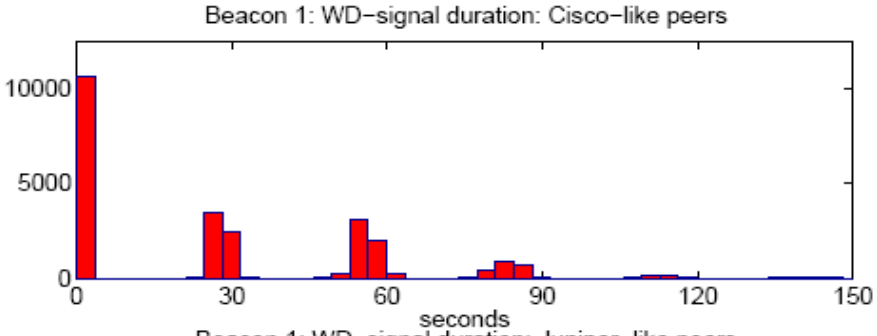


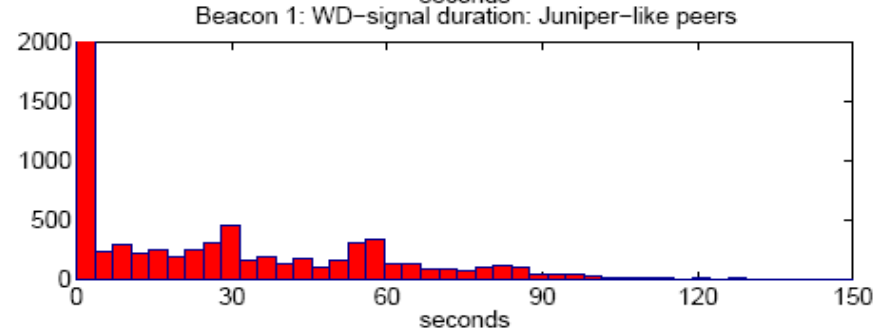


PSG Beacons

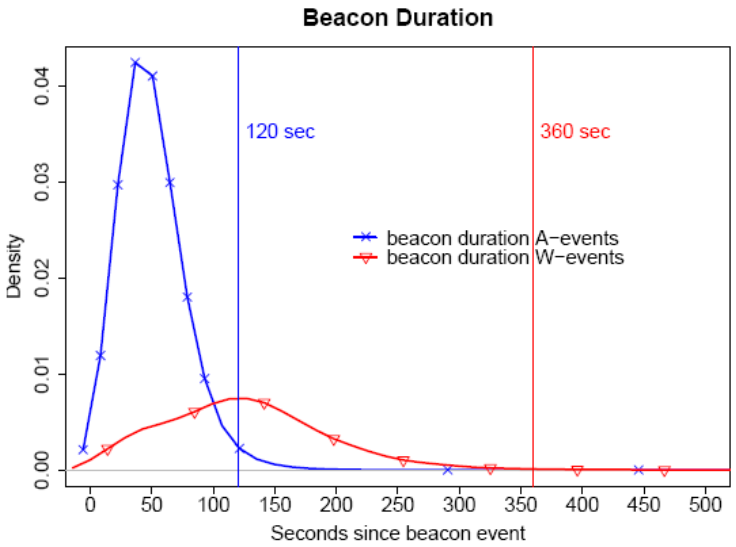

Beacon 1: ANN−signal duration: Cisco−like peers
15000
10000
5000
0
0
30
60
90
120
150
seconds

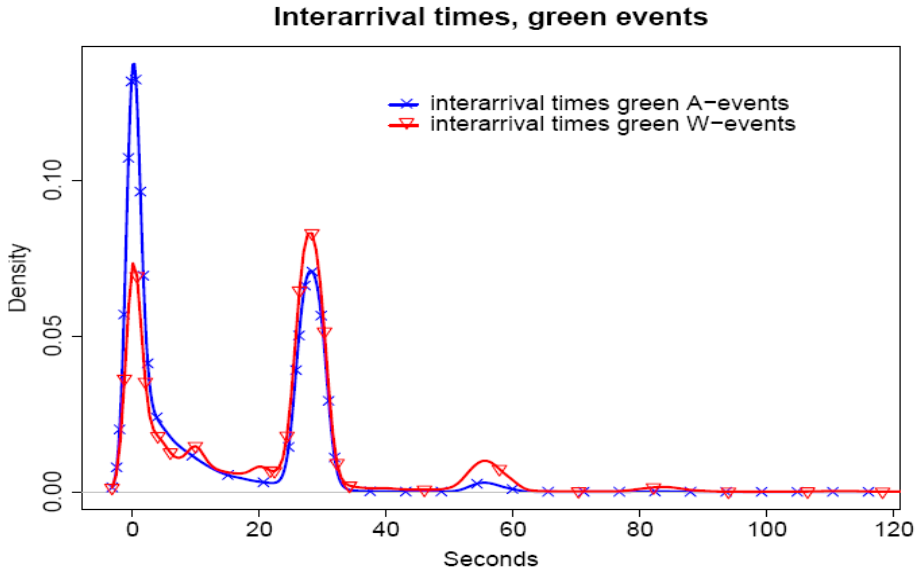

Beacon 1: ANN−signal duration: Juniper−like peers
3000
2000
1000
0
0
30
60
90
120
150
seconds


Beacon 1: WD−signal duration: Cisco−like peers
10000
5000
0
0
30
60
90
120
150
seconds


Beacon 1: WD−signal duration: Juniper−like peers
2000
1500
1000
500
0
0
30
60
90
120
150
seconds

RIP Beacons

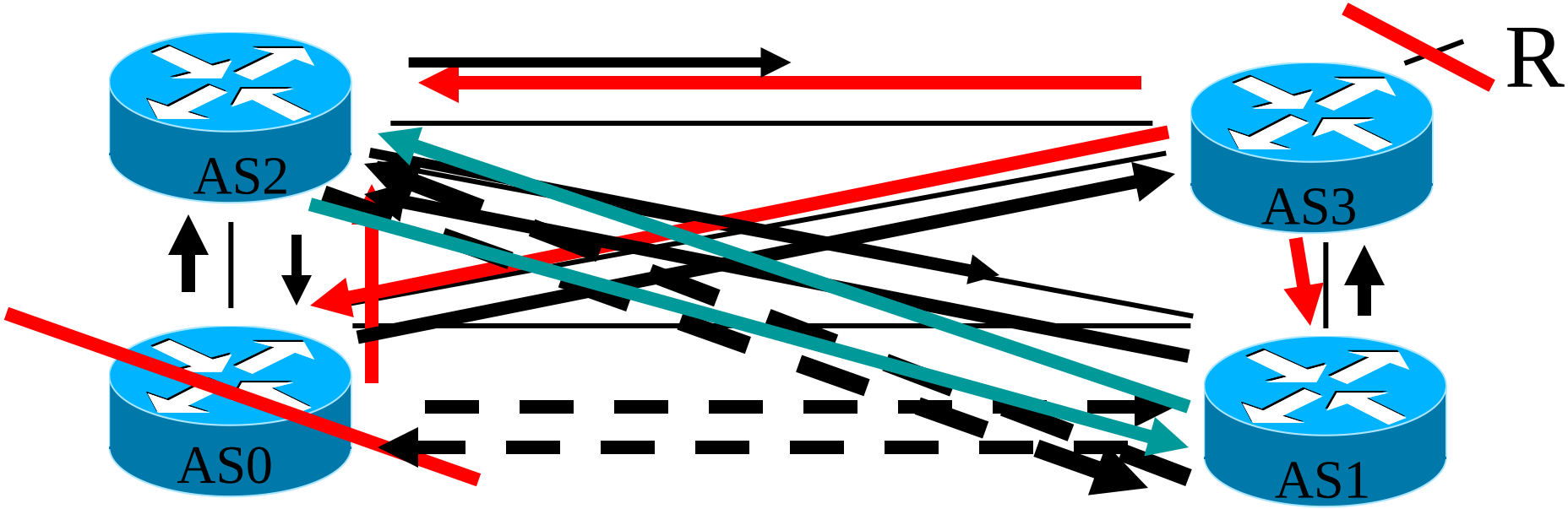

Beacon Duration
120 sec
360 sec
beacon duration A−events
beacon duration W−events
Density
0.00
0.01
0.02
0.03
0.04
0
50
100
150
200
250
300
350
400
450
500
Seconds since beacon event


Interarrival times, green events
interarrival times green A−events
interarrival times green W−events
Density
0.00
0.05
0.10
0
20
40
60
80
100
120
Seconds

RIPE Beacons Result

- Green Events
  - A: converge within 120 seconds with A (90.5%)
  - W: converge within 360 seconds with W (96.5%)
- Red Events
  - All events with long convergence (4.38%)
  - Mostly due to route-damping effect
- Orange Events
  - Converge to wrong type of events (1.8%), more A-Events
- Greg Events
  - Invisible events through certain peers, account for 40% of all events
  - Sudden appearance: during routing policy change?

BGP Convergence Update Burst

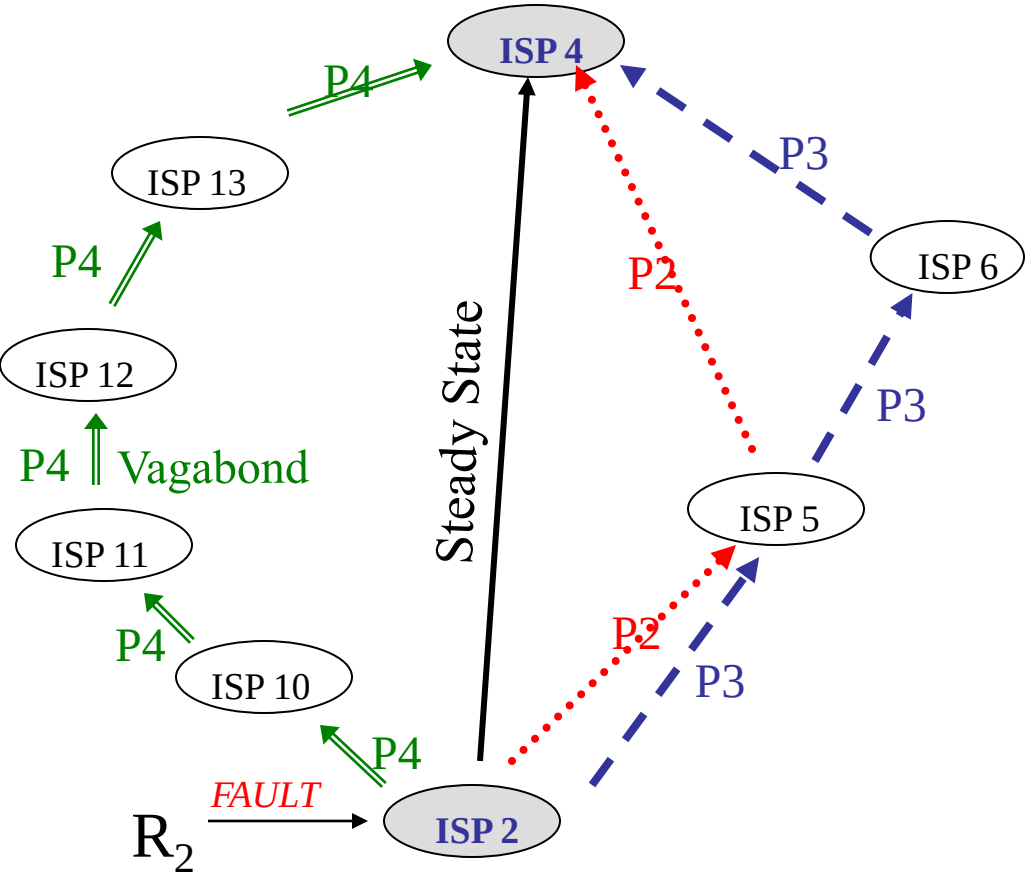


*B R via 3
*B R via 13
B R via 23
AS0

*B R via 3
*B R via 203
B R via 23
AS1

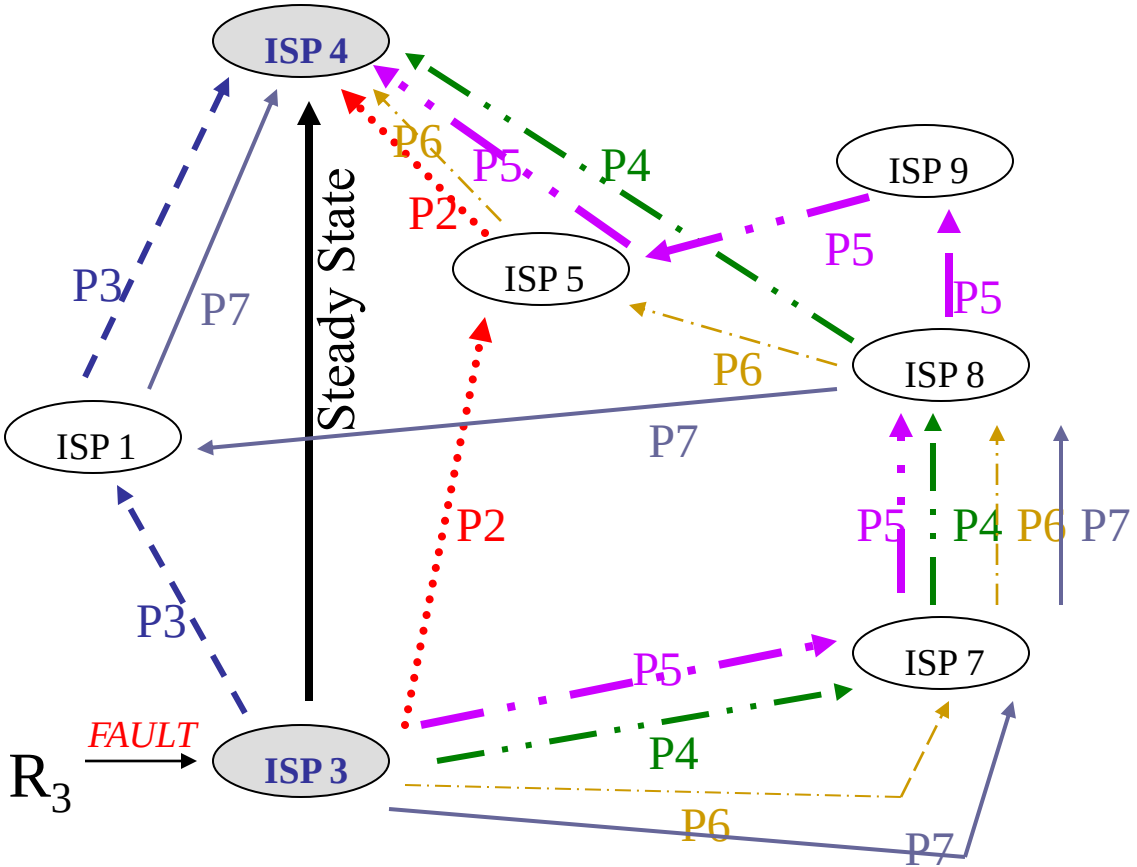
*B R via 3
*B R via 013
B R via 103
B R via 13
AS2

ISP2-ISP4 Paths During Failure



63%	Average: 79 (min/max 44/208) seconds
AS4 AS5 AS2	(35 seconds)
~~Withdraw~~	~~(79 seconds)~~
7%	Average: 88 (min/max 80/94) seconds
Announce AS4 AS5 AS2	(33 seconds)
Announce AS4 AS6 AS5 AS2	(61 seconds)
Withdraw	(88 seconds)
7%	Average: 54 (min/max 29/9) seconds
Withdraw	(54 seconds)

ISP3-ISP4 Paths During Failure



36% Average: 110 (min/max 78/135) seconds

Announce AS4 AS5 AS (52 seconds)

Withdraw (110 seconds)

35% Average: 107 (min/max 91/133) seconds

Announce AS4 AS1 AS3 (39 seconds)

Announce AS4 AS5 AS3 (68 seconds)

Withdraw (107 seconds)

2% Average:140.00 (min/max 120/142)

Announce AS4 AS5 AS8 AS7 AS3 (27)

Announce AS4 AS5AS9 AS8 AS7 AS3 (86)

Withdraw (140 seconds)

27% **Other**

Typical BGP Withdraw

7/5 19:33:25	Route **R** is **withdrawn**	
7/5 19:34:15	AS6543 **announce** **R**	6543 66665 8918 1 5696 999
7/5 19:35:00	AS6543 **announce** **R**	6543 66665 8918 67455 6461 5696 999
7/5 19:35:37	AS6543 **announce** **R**	6543 66665 4332 6461 5696 999
7/5 19:35:39	AS6543 **announce** **R**	6543 66665 5378 6660 67455 6461 5696 999
7/5 19:35:39	AS6543 **announce** **R**	6543 66665 65 6461 5696 999
7/5 19:35:52	AS6543 **announce** **R**	6543 66665 6461 5696 999
7/5 19:36:00	AS6543 **announce** **R**	6543 66665 5378 6765 6660 67455 6461 5696 999
		…
7/5 19:38:22	AS6543 **withdraw** **R**	

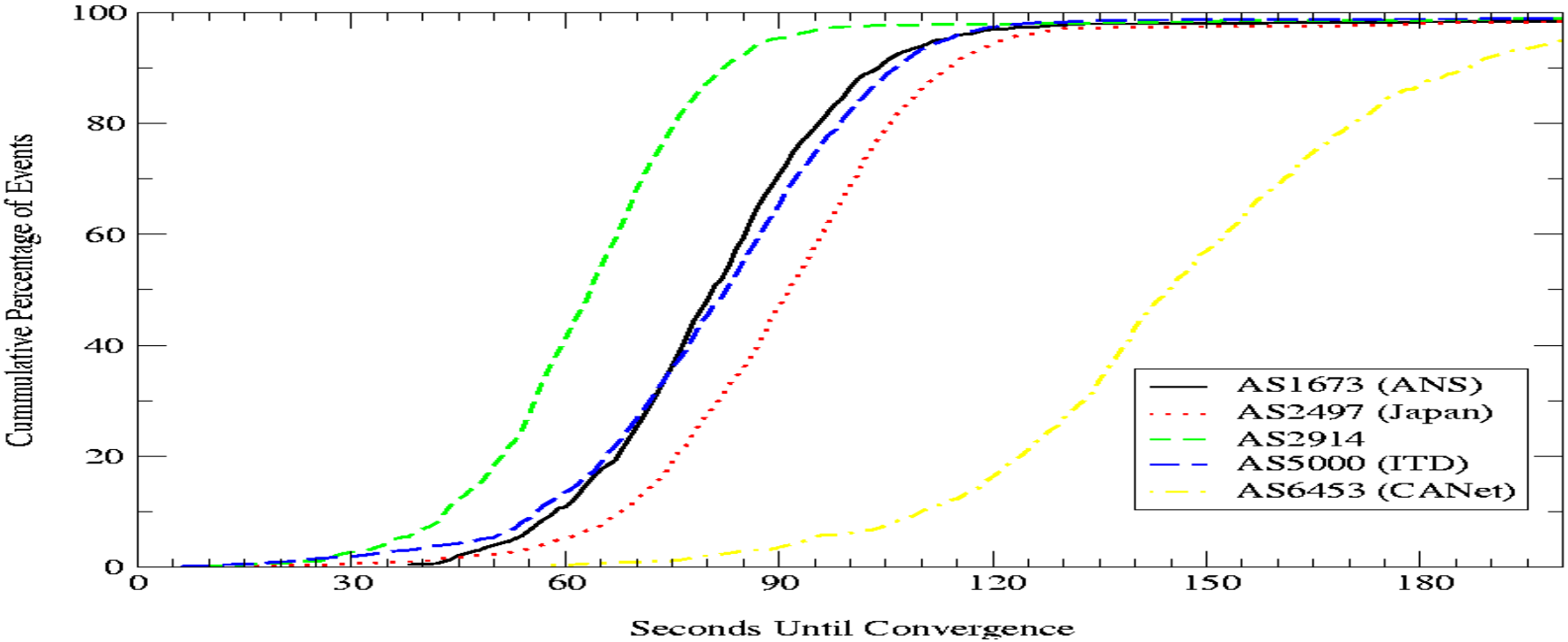
Merit -- Convergence Time

- Tup -- A new route is advertised
- Tdown -- A route is withdrawn (i.e. single-homed failure)
- Tshort -- Advertise a shorter/better ASPath (i.e. primary path repaired)
- Tlong -- Advertise a longer/worse ASPath (i.e.primary path fails)

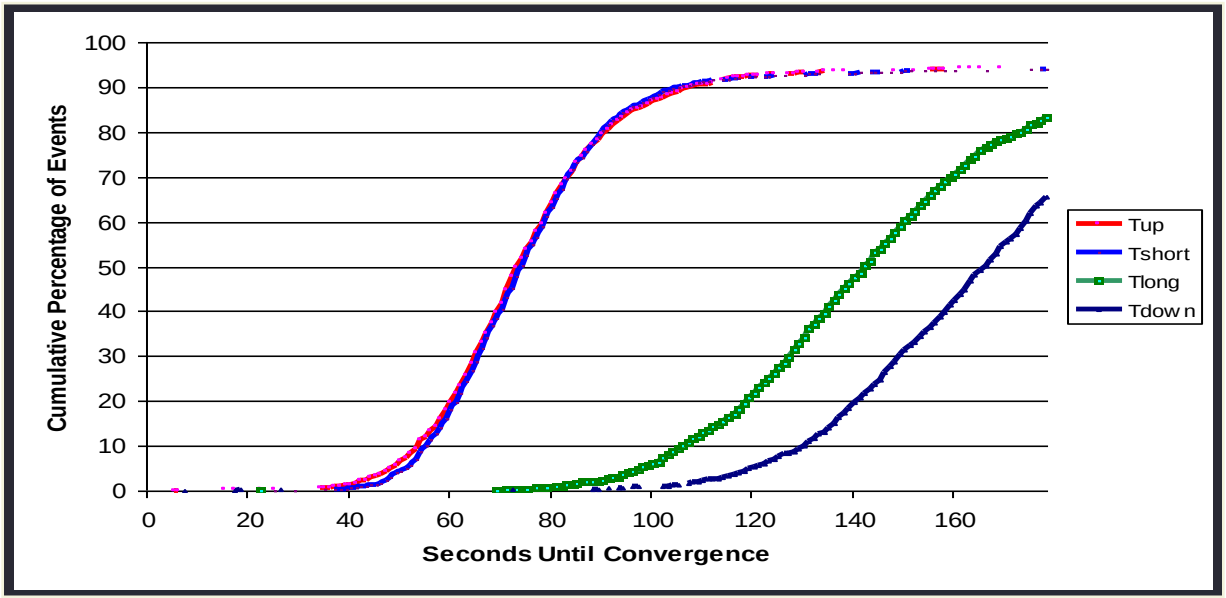
Merit Result

- Routing convergence requires an order of magnitude longer than expected (10s of minutes)
- Routes converge more quickly following Tup/Repair than Tdown/Failure events (“bad news travels more slowly”)
- Curiously, withdrawals (Tdown) generate several times the number of announcements than announcements (Tup)

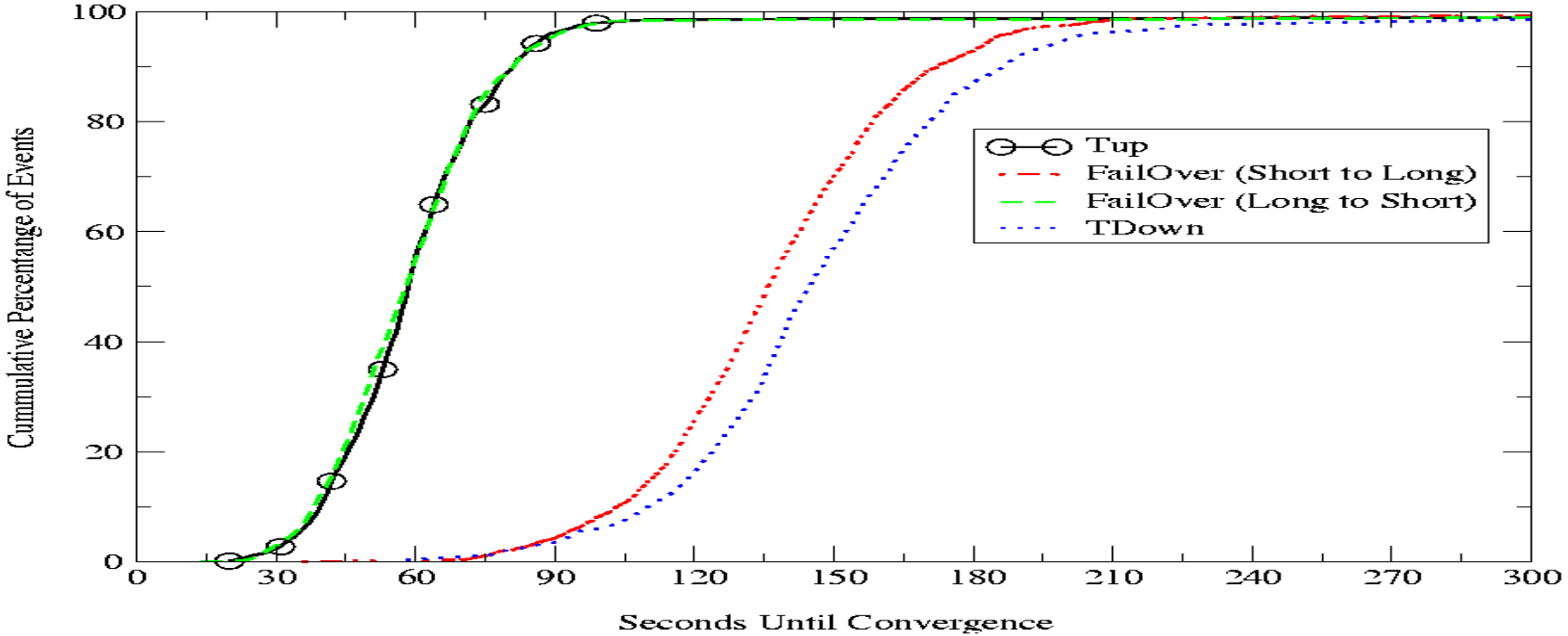
Withdraw Convergence



BGP Convergence



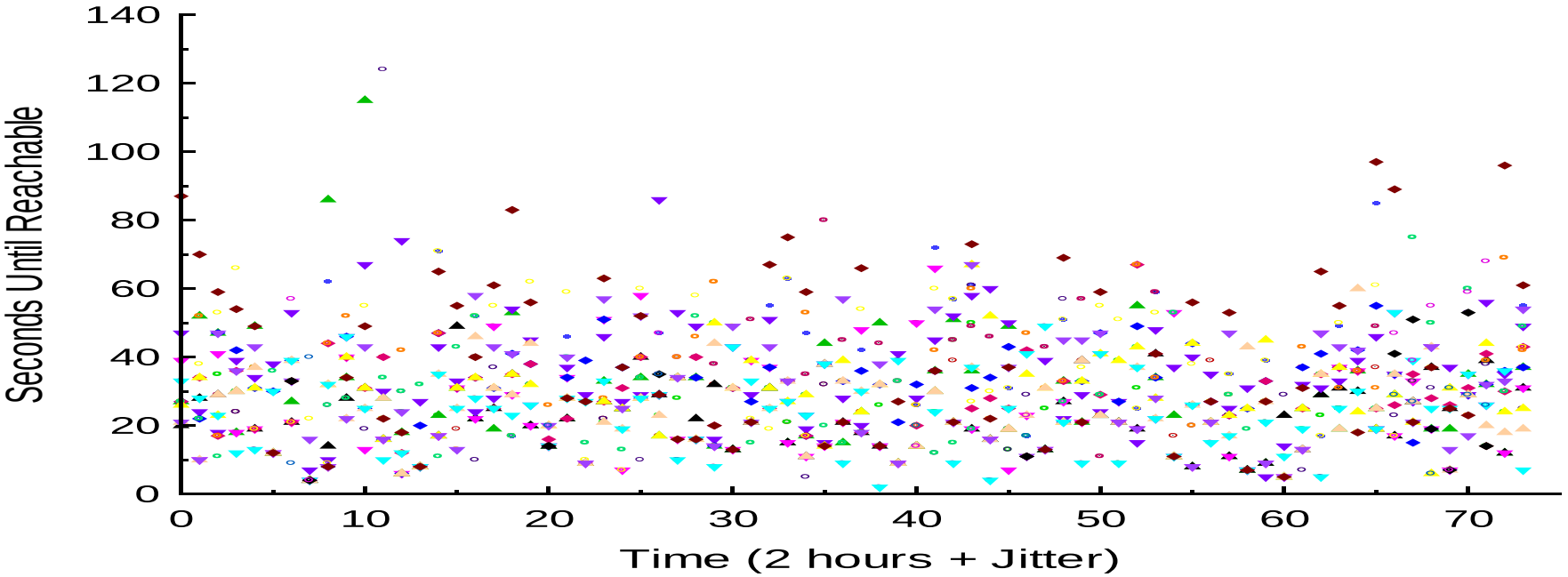
Failure, Fail-over Convergence



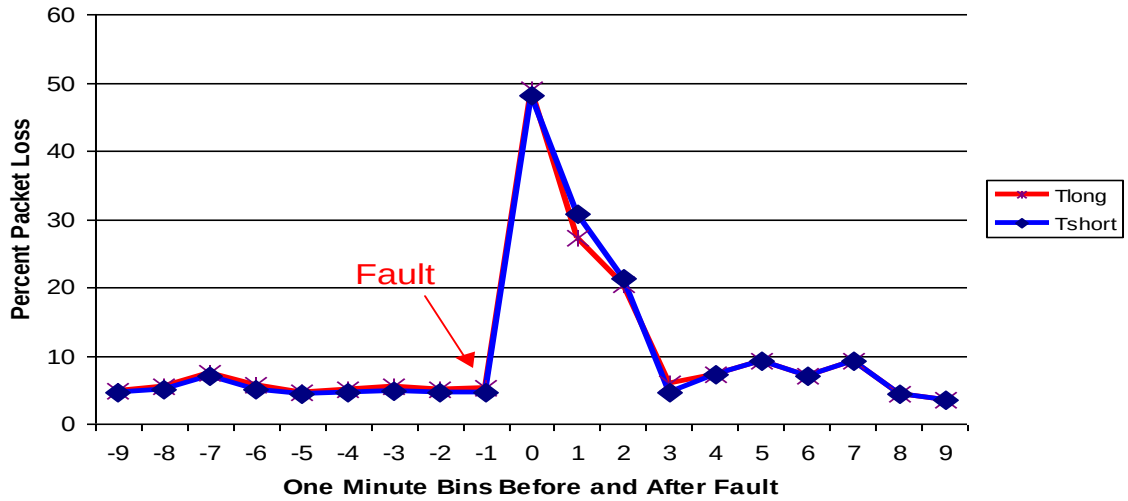
Withdraw Convergence

- 80% of withdraws from all ISPs take more than a minute
- For ISP4, 20% withdraws took more than three minutes to converge
- Failures (Tdown) and short-long fail-overs (e.g. primary to secondary path) also similar
  - Slower than Tup (e.g. a repair)
  - 60% take longer than two minutes
  - Fail-over times degrade the **greater** the degree of multi-homing!
- Internet averages 3 minutes to converge after failover
  - Some multihomed failovers (short to long ASPath) require 15 minutes

ICMP Response after Repairs



End2end Impact after Fail-over



Route damping effect

- Route damping: deal with long time scale instability
- MinAdvTimeInterval: Route short time instability, delay updates to batch consecutive updates to reduce updates
- No matter how large MinAdvTimeInterval, possible to induce damping due to single update
- Measured from route-view and use default Cisco and Juniper parameters: on average 5%, but up to 45% of updates might be suppressed!
- Route damping might be the main reason for the extended delay convergence

PSG -- Route-damping

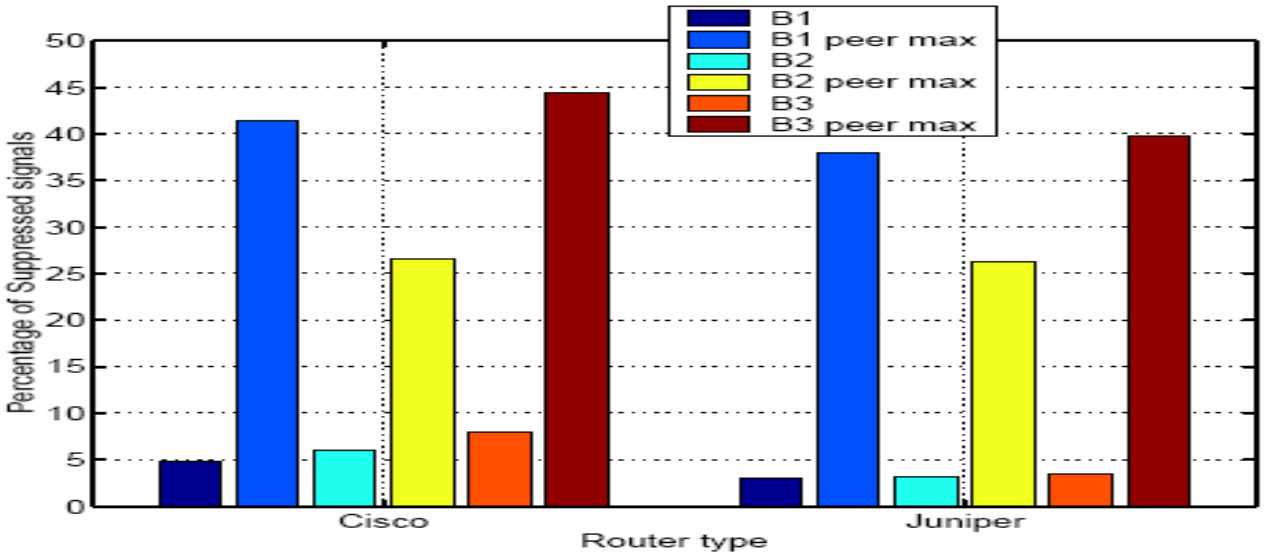


Figure 10: Overall percentage of suppressed signals due route flap damping for each Beacon and on a per peer basis for Cisco and Juniper.

BGP Model

- If complete fully-mesh ASN graph, N! upper theoretic bound and 30*(N-3) lower bound
- In practice, Internet has hierarchy and customer/provider/sibling relationships
  - Bounded by length of longest possible path
- ASPath limits “infinity” to the width of the Internet
  - Monotonically increasing
  - Upper bound?

BGP Model

- 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**
- BGP is O(N!), where N number of default-free BGP speakers

Alternative Path Enumeration

- BGP monotonically increasing. Multiple (N!) ways to represent a path metric of N.
- ***AS-PATH Enumerations***
  - ***2117 5696 2129***
  - ***2117 1 5696 2129***
  - ***2117 2041 3508 3508 4540 7037 1239 5696 2129***
  - ***2117 1 2041 3508 3508 4540 7037 1239 5696 2129***
  - ***2117 2041 3508 3508 4540 7037 1239 6113 5696 2129***
  - ***2117 1 2041 3508 3508 4540 7037 1239 6113 5696 2129***

- BGP “solved” RIP routing table loop problem by making it exponentially worse…

MRAI Timers

- MinAdvertiseInterval: timer to limit the numbers of advertisement updates per prefix. Recommend by RFC and only apply to advertisement eBGP; Usually not applied to iBGP and route withdraw
- Small timer (Juniper): more updates, short convergence time
- Longer timer (Cisco: 30 seconds): fewer updates, longer convergence time
- Implementation of MinRouteAdver timer leads to 30 second rounds
  - Time complexity is O(n-3)*30 seconds
  - State/Computational complexity O(n)
  - At its best, BGP performs as well as RIP2 (but uses exponentially more memory in the process)

MRAI Timer

- Minimum interval between successive updates sent to a peer for a given prefix
  - Allow for greater efficiency/packing of updates
  - Rate throttle
- Applied only to announcements (at least according to BGP RFC)
- Applied on (prefix destination, peer) basis, but implemented on (peer) basis
- 30*(N-3) delay due to creation mutual dependencies. Provide proof that N-3 rounds necessarily created during bounded BGP MinRouteAdver convergence
- Rounds due to
  - Ambiguity in the BGP RFC and lack receiver loop detection
  - Inclusion of BGP withdrawals with MinRouteAdver (in violation of RFC)

Findings

- Non-deterministic ordering of BGP update messages leads to
  - Transient oscillations
  - Each change in FIB adds delay (CPU, BGP bundling timer)
  - At extreme, convergence triggers BGP dampening
- Given best current routing practices, inter-domain BGP convergence times degrade **exponentially** with increase in the degree of interconnectivity for a given route and the degree of inter-connectivity (multi-homing, transit, etc) is increasing

MRAI Timer

- Cisco default: 30 seconds
- ATT BGP Convergence Simulations Results:
  - Exists optimal MRAI Mu, if above, total updates for convergence is stable
  - Exists optimal MRAI Mt where convergence time is minimized, if above, average convergence time increases linearly
  - Mt increase with average router load, and an optimal MRAI can significantly reduce convergence time (but network dependent)
- Recent Simulations Results
  - Optimal MRAI for most of network today might be between 1-5 seconds

Impact to Reality

- Great research result and provide a lot of insight into Internet BGP dynamics
- Engage talks with Cisco/Juniper to improve the behaviors and convergence
- But from practical point of view, people care more on reachability rather than absolute convergence time
- More BGP research in
  - Internet routing simulations based different timers and polices
  - Alternative routing mechanism design simulations
  - Reality check on voice/jitter/video due to route convergence and fail-over needed