An Experimental Study of

Delayed Internet Routing Convergence

ACM SIGCOMM 2000
Stockholm Aug 31, 2000

Craig Labovitz
Microsoft Research
labovit@microsoft.com

Abha Ahuja, Farnam Jahanian, Abhijit Bose
University of Michigan
{ahuja, farnam, abose}@umich.edu

Motivation

(Why we should care about convergence)

- Routing reliability/fault-tolerance on small time scales (minutes) not previously a priority
- Emerging transaction oriented and interactive applications (e.g. Internet Telephony) will require higher levels of end2end network reliability
- How well does the Internet routing infrastructure tolerate faults?

2

Conventional Routing Wisdom

(IETF, IAB, ISPs, etc)

- Internet routing is robust under faults
  - Supports path re-routing and restoral 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 always improve site fault-tolerance

3

In This Talk

We will show that most of the conventional wisdom about routing convergence is not accurate…

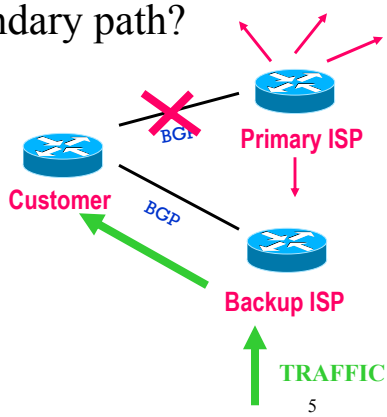
- Measurement of BGP convergence in the Internet
- Analysis/intuition behind delayed BGP routing convergence
- Modifications to BGP implementations which would improve convergence times

4

Open Question

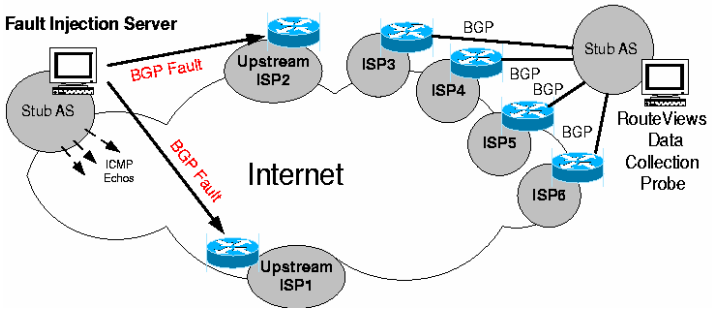
After a fault in a path to multi-homed site, how long does it take for the majority of Internet routers to fail-over to the secondary path?

- Routing table convergence (backbone routers reach steady-state) after a fault
- End-to-end paths stable ("normal" levels of loss and latency)



5

Experiments



- Inject BGP faults (announcements/withdraws) of varied prefix and ASPath lengths into topologically and geographically diverse ISP peering sessions
- Monitor impact faults through 1) recordings of default-free BGP peering sessions with 20 tier1/tier2 ISPs and 2) active ICMP measurements (512 byte/second to 100 random web sites)
- Wait two years (and 250,000 faults)

6

Fault Scenarios

- 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)

7

Major Convergence Results

- 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)

8

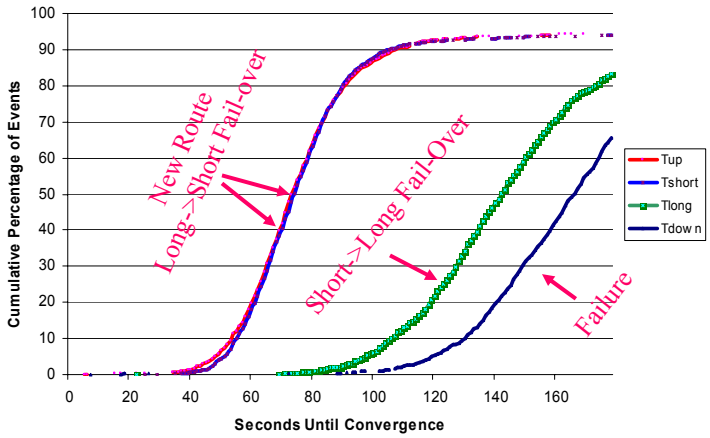
Example of BGP Convergence

TIME	BGP Message/Event
10:40:30	*Route Fails/Withdrawn by AS2129*
10:41:08	*2117 announce 5696 2129*
10:41:32	*2117 announce 1 5696 2129*
10:41:50	*2117 announce 2041 3508 3508 4540 7037 1239 5696 2129*
10:42:17	*2117 announce 1 2041 3508 3508 4540 7037 1239 5696 2129*
10:43:05	*2117announce 2041 3508 3508 4540 7037 1239 6113 5696 2129*
10:43:35	*2117 announce 1 2041 3508 3508 4540 7037 1239 6113 5696 2129*
10:43:59	*2117 sends withdraw*

- BGP log of updates from AS2117 for route via AS2129
- One BGP withdrawal triggers 6 announcements and one withdrawal from 2117
- Increasing ASPath length until final withdraw

9

CDF of BGP Routing Table Convergence Times



- Less than half of Tdown events converge within two minutes
- Tup/Tshort and Tdown/Tlong form equivalence classes
- Long tailed distribution (up to 15 minutes)

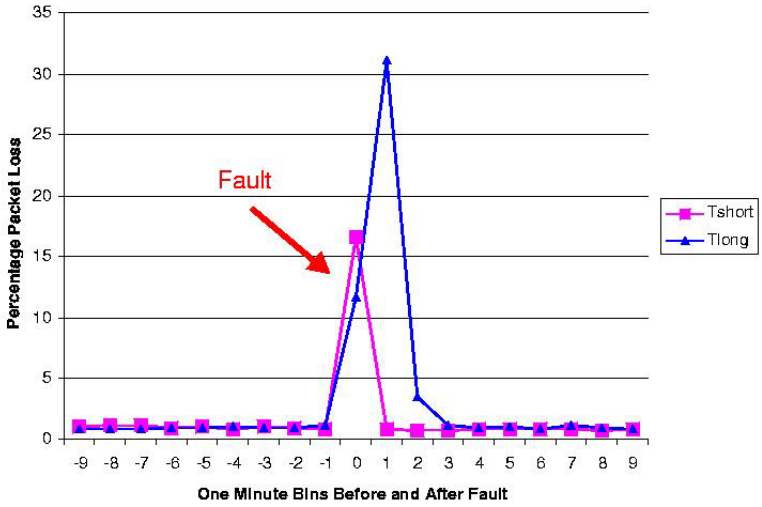
10

Impact of Delayed Convergence

- Why do we care about routing table convergence? It deleteriously impacts end-to-end Internet paths
- ICMP experiment results
  - Loss of connectivity, packet loss, latency, and packet re-ordering for an average of 3-5 minutes after a fault
  - Why? Routers drop packets for which they do not have a valid next hop. Also problems with cache flushing in some older routers.

11

End-to-End Impact Failover



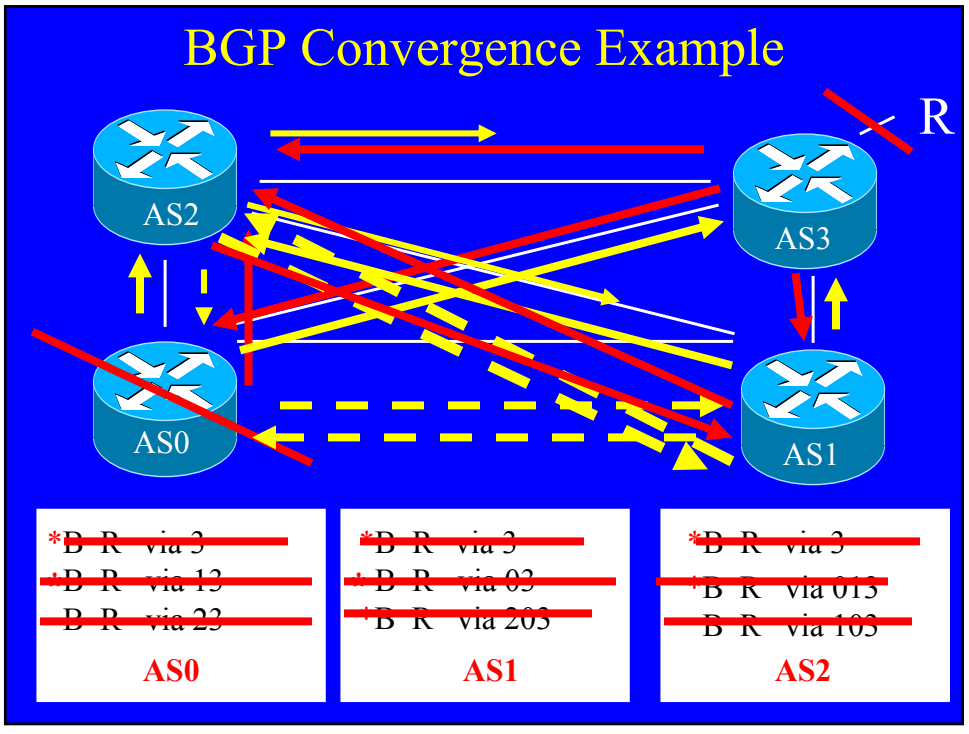
- ICMP loss to 100 randomly chosen web sites with VIF source address of our probe
- Tlong/Tshort exhibit similar relationship as before

12

Delayed Convergence Background

- Well known that distance vector protocols exhibit poor convergence behaviors
  - Counting to infinity, looping, bouncing problem
- RIP redefines infinity and adds split-horizon, poison reverse, etc.
  - Still, slow convergence and not scalable
- BGP advertises ASPaths instead of distance
  - Solves counting to infinity and RIP looping problem, but…
  - BGP can still explore "invalid" paths during convergence (i.e. the bouncing problem)

13



Intuition for Delayed BGP Convergence

- 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 in a complete graph with default policy
- Although seemingly very different protocols, BGP and RIP share very similar convergence behaviors. Major difference:
  - RIP explores metrics (1…*N*)
  - BGP ASPath provides multiple ways to represent metric (path) of length *N*, or (*N*-1)!

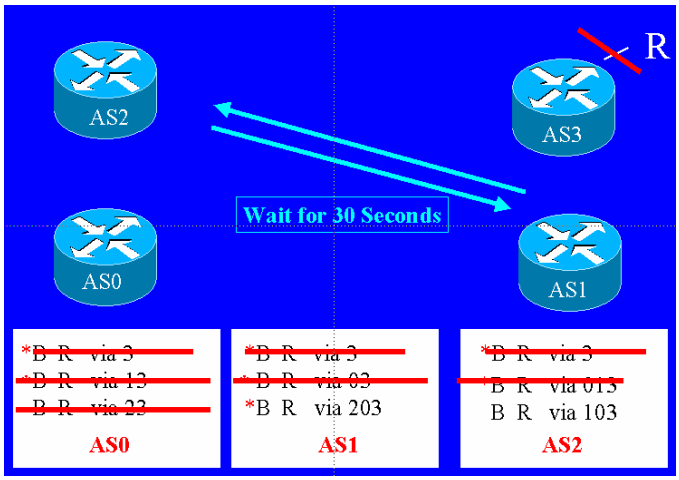
15

Lower Bound on BGP

- If assume optimal ordering of messages, what is the best we can expect from BGP?
- In practice, BGP timers (MinRouteAdver) provide synchronization and limit possible orderings of messages
  - MinRouteAdver timer specifies interval between successive updates sent to a peer for a given prefix
  - Useful for bundling updates together
  - According to RFC, MinRouteAdver applies only announcements
- But, interaction of MinRouteAdver and vendor ASPath loop detection implementation introduce "artificial" delay

16

MinRouteAdver Rounds



- Implementation of MinRouteAdver timer and receiver-side loop detection timer leads to 30 second rounds O(n-3)*30 seconds time complexity

17

Conclusion and Next Steps

- Internet does not posses effective inter-domain fail-over (15 minutes is a long time for phone call)
- Majority of BGP convergence delay due to vendor implementation decisions of MinRouteAdver and loop detection
- In practice, Internet is not a complete graph and same degree of message re-ordering unlikely. Our current work:
  - What is the impact of ISP policy and topology on BGP convergence?
  - Can we improve BGP convergence times?

18