Improved BGP Convergence via Ghost Flushing
Bremler-Barr, Afek, Schwarz, 2003

BGP-RCN: Improving Convergence Through Root Cause Notification
Pei, Azuma, Massey, Zhang, 2005

Spring 2013
Obsolet paths (C B A) and (E D B A) explored before converging on valid path (I H G F A)
Potential to Explore N! Paths

Paths Explored by A

A,C,S
Link C,S fails
A,B,C,S
A,D,C,S
A,B,D,C,S
A,D,B,C,S
....
No route
Theoretically can explore 
N! paths before no route
Some Routing Terminology

• Tup = route to previously unreachable prefix is announced.
• Tdown = route to current reachable prefix is withdrawn and no replacement exists
• Tshort = route to current reachable prefix switches to shorter path
• Tlong = route to current reachable prefix switches to a longer path
• Other terminology
  – Tdown = fail-down
  – Tlong = fail-over
BGP MRAI Time and Convergence

- Minimum Route Advertisement Interval (MRAI) timer:
  - Within $M=30$ seconds, at most one announcement from A to B
  - not for the first announcement, not for the withdrawal

- Impact:

  a. suppress transient changes
  b. delay convergence

A’s path changes: $P_1 \xrightarrow{w} P_2 \xrightarrow{} P_3 \xrightarrow{} P_4 \xrightarrow{} P_5$

Msgs from A to B: $P_1 \xrightarrow{w} P_4 \xrightarrow{} P_5$
[BAS03] Improving BGP Convergence

- **Objective:**
  - Improve convergence time after a legitimate route change.

- **Approach:**
  - Flush out ghost information that is blocked by the MRAI timer
    - P4 in previous slide is ghost information

- **Contributions:**
  - Simple, easily deployed, and clever approach to improve convergence
  - Theoretical understanding of convergence behavior
  - Improves on 2002 result from Pei et al.
Basic Model

• Each AS is treated as one node
  – Though not strictly required in ghost flushing

• Routers use shortest path routing policy
  – Helps with analysis, but not strictly required

• SPVP (simple path vector protocol) approximates BGP

• MRAI timer between updates
  – Minimum Route Advertisement Interval
  – Two consecutive updates must be at least MRAI time apart.
Ghost Information

- Obsolete path information stored at node
  - Could be preferred route or backup route stored at a node.

- MRAI timer can block removal of ghost information
  - Router cannot announce its current choice of paths because it recently announced a different path.
  - Typical MRAI value is 30 seconds
  - Can lead to increased convergence time and increased chance of selecting ghost paths.
Ghost Flushing

• Very Simple Rule for BGP Routers

When route to P is updated to a worse path and MRAI timer is delaying path announcement send withdraw(P) (no route to P)
Path Length and Time

• Assume Tdown Event
• Let $H = \text{message passing time}$
• **Claim at time $K*H$, every message or node has ASPath length > K**
  – By induction. True at time $H$ since neighboring routers received withdraw
  – Assume true at time $KH$, all paths longer than $KH$.
  – Suppose $K$ or less path exists at time $(K+1)H$
    • Must have come from some peer $P$ with path length $KH$.
    • Path must have been removed prior to time $KH$
      – Withdraw or longer path announced prior to time $KH$
      – Must be received prior to time $(K+1)H$ (contradiction)
Implications of Time/Length

• Shown that at the K*H, every message or node has ASPath length > K

• Implications:
  – Longest possible path has length N
  – At time N*H, all paths are longer than longest possible path
  – By time N*H, all routers know that path is withdrawn

• Convergence time is (N*H)
  – Reduced from N*MRAI
Message Complexity

• Claim at most 2 messages sent during each MRAI timer interval

• Resulting complexity
  – Number of MRAI rounds is $NH/(MRAI)$
  – Updates per round is $2E$

Complexity is $O(2ENH/MRAI)$

(BGP complexity is $EN$)
Tlong (fail-over) Complexity

• Expect good results, but no theoretical results presented here
  – Simulations show solid improvement
  – Other simulations (not shown here) show some surprises…
• Theoretical results later determined by Pei et al.
  – Covered next week….
[PA+05] Improving BGP Convergence

- **Objective:**
  - Improve convergence time after a legitimate route change.

- **Approach:**
  - Signal the cause of the path failure

- **Contributions:**
  - Dramatic reduction in convergence time plus ability to improve other parts of BGP
  - Theoretical understanding of convergence behavior
Observation: if Z knows [B A] failed, it could’ve avoided the obsolete paths
Root Cause Notification

- The node who detects the failure attaches *root cause* to msg
- Other nodes copy the root cause to outgoing messages

the first msg is enough for Z to remove all the obsolete paths

Z’s Candidate paths:

- ()
- (C B A)
- (E D B A)
- (I H G F A)

Diagram:

- Nodes Z, B, A, F, C, E, D, I, H, G
- Arrows indicating message paths and failures
- Node B with a cross indicating failure
- Node Z receiving a msg with root cause [B A] failure

The first msg is enough for Z to remove all the obsolete paths.
Overlapping Events

- Another topology change happens before the previous change’s convergence finishes.

- Propagation along lower path is slower than upper path.
Overlapping Events

Path: (B A)

[B A] failure

[B A] recovery

[B A] recovery
Overlapping Events

- Observation: need to \textit{order} the relative timing of the root causes

Wrong!
Path: (B A)

\begin{itemize}
  \item [B A] failure
  \item [B A] recovery
\end{itemize}
Solution: adding sequence number

• Node B maintains a sequence number for link [B A]
• Incremented each time the link status changes
Solution: adding sequence number

(B A), [B A] recovery, seqnum=2

Path: (C B A), seqnum of [B A]=2

[B A] failure, seqnum=1

dest.
Solution: adding sequence number

- Sequence number *orders* the relative timing of the root causes

Path: (B A), seqnum of [B A]=2

[B A] failure, seqnum=1

dest.
Evaluation: analysis and simulation

- Two types of topology changes:
  - Fail-over: nodes switch to worse paths
  - Fail-down: destination becomes unreachable

![Diagram of network topology changes]
Fail-down convergence delay (worst case) bound

Withdrawals are not delayed by MRAI!

Along shortest path: it takes at most $d \times h$ seconds

$d$: network diameter

$d \ll N-1$ and $h \ll M$

<table>
<thead>
<tr>
<th>RCN</th>
<th>$d \times h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGP</td>
<td>$(N-1) \times (h+M)$</td>
</tr>
</tbody>
</table>

Length of the longest possible path $(N)$

MRAI value
Fail-down simulation results

- 2-3 orders of magnitudes reduction

![Convergence Time Graph](image)

- Number of nodes: 14, 28, 56, 112
- Seconds: 1, 10, 100, 1000
- BGP
- RCN

**Convergence Time**

Seconds

Number of nodes
Border nodes in fail-over convergence

Border node Z:
- connected to an unaffected node H
- its eventual path is through H

Z’s eventual path has always been available

unaffected nodes

Affected nodes

D  Z  H  I  J  B  C  A

dest.
RCN’s fail-over delay bound

Node D’s convergence:
Phase 1: Z receives the root cause
Phase 2: Z’s path is propagated to D
(MRAI delay applies in this phase)

RCN \((M + 2h)d_{\text{affected}}\)
BGP’s fail-over delay bound

Node D’s convergence:

Phase 1: Z explores paths shorter than Z’s eventual path

Phase 2: the same as in RCN

| BGP | $(M+h) \times \min\{d' - J, |V_{affected}| + d_{affected} - 1\} $ |
Fail-over simulation results

- BGP does fine: \((M+h) \times d'\)
- \(d'\): 2~6

\(d'\): length of the longest path from any affected node to the destination

---

Constructed topologies with large \(d'\): RCN has much more pronounced improvement
RCN Overhead

- Transmission & storage of a path: doubled path:seqnum (Z C B A):(3 2 2 1)
- Storage overhead in the routing table:
  - doubled
- Transmission overhead reduced
  - 1~2 orders of magnitudes reduction in msg counts
Related Work

- Reducing negative impact of MRAI:
  - [Griffin:ICNP01], Ghost-Flushing [Bremler-Barr:Infocom03]
  - don’t deal with path exploration

- Reducing path exploration
  - Consistency Assertion [Pei:Infocom02]
  - path exploration still exists

- Explicitly signaling failure
  - RCO [Luo:Globecom02], BGP-CT [Wattenhofer:talkslides03]: may result in wrong routing decision
  - EPIC [Chandrasheka:Infocom05]: encoding difference