Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
  Reduction of update load
  MRAI and PED convergence time compared

Conclusions and future work
Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
  Reduction of update load
  MRAI and PED convergence time compared

Conclusions and future work

About this talk

- Selected for the Applied Networking Research Prize (ANRP) based on a peer-reviewed paper:

- Paper published in the IEEE Journal on Selected Areas in Communications (JSAC), October 2010
  - A Technique for Reducing BGP Update Announcements through Path Exploration Damping.
  - Geoff Huston, Mattia Rossi, Grenville Armitage

- Project sponsored by the Cisco University Research Program Fund

1 http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5586440
What is Path Exploration Damping?

- Path Exploration Damping → PED
- Algorithm intended to replace MRAI and RFD
- Methods to reduce update churn and convergence time in BGP
- BGP is the de-facto standard for inter-domain routing
  - BGP – Border Gateway Protocol
  - MRAI – Minimum Route Advertisement Interval
  - RFD – Route Flap Damping

PED implementations

- CAIA
  - Implemented in Quagga
  - Patch available for Quagga 0.99.13
- Cisco (by Mohammed Mirza)
  - Implemented in CISCO IOS-XE Experimental Version 15.1
  - Running on Cisco ASR1002 (2RU)
  - Currently tested at APNIC Pty. Ltd., Australia
Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
  Reduction of update load
  MRAI and PED convergence time compared

Conclusions and future work

Simplified BGP speaker design

RIB: Routing Information Base – Routing Table
Adj-RIB-(in,out): Adjacency RIB (in,out)
Basic BGP dynamics

Prefix: 1.0.0.0/8

AS1
AS2 AS3 AS4
AS5
Observer

Path: 1

Prefix: 1.0.0.0/8
Basic BGP dynamics

Path: 2,1
Observer
Prefix: 1.0.0.0/8

Path: *2,1
3,2,1
Observer
A: 5,2,1
Prefix: 1.0.0.0/8
Basic BGP dynamics

Path: *2,1
3,2,1
4,3,2,1

Observer
A: 5,2,1

Path: 2,1
Path: 3,2,1

Path Exploration

Path: *2,1
3,2,1
4,3,2,1

Observer
A: 5,2,1
---
Path Exploration

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---

Path Exploration

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---
A: 5,3,2,1
Path Exploration

Path: W

Observer
A: 5,2,1
---
A: 5,3,2,1
A: 5,4,3,2,1

Prefix: 1.0.0.0/8

IETF-81
http://www.caia.swin.edu.au
mrossi@swin.edu.au
28 July, 2011
What is Path Exploration?

- An update sequence lengthening the AS-path gradually until stability is reached

MRAI and RFD

- Need to decrease BGP chattiness and Path Exploration
- Minimum Route Advertisement Interval – MRAI (RFC 1771)
  - Apply 30s (default) delay on announcements
Minimum Route Advertisement Interval – MRAI

Path: *2,1
Path: 3,2,1
Path: 4,3,2,1
Path: 2,1
Path: 3,2,1
Prefix: 1.0.0.0/8

Observer
Delay 0-90s
A: 5,2,1

Prefix: 1.0.0.0/8
Minimum Route Advertisement Interval – MRAI

Path: *3,2,1
4,3,2,1

Observer
A: 5,2,1
---

Prefix: 1.0.0.0/8

Minimum Route Advertisement Interval – MRAI

Path: *4,3,2,1

Update delayed

Observer
A: 5,2,1
---

Prefix: 1.0.0.0/8
Minimum Route Advertisement Interval – MRAI

Path: W

AS1
AS2
AS3
AS4
AS5

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---
---
---

Path: W
Path: W
Path: W
Path: W

Prefix: 1.0.0.0/8

IETF-81
http://www.caia.swin.edu.au
mrossi@swin.edu.au
28 July, 2011
Minimum Route Advertisement Interval – MRAI

Path: *4,3,2,1
Observer
A: 5,2,1
---
Prefix: 1.0.0.0/8

Path: W
Path: W
Path: W
AS3
AS4
AS2
AS5
AS1

Prefix: 1.0.0.0/8
Minimum Route Advertisement Interval – MRAI

Path: W

AS5

AS2

AS3

AS4

AS1

Path: W
Path: W
Path: W

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---
A: 5,4,3,2,1
W

MRAI and RFD

- Need to decrease BGP chattiness and Path Exploration
- Minimum Route Advertisement Interval – MRAI (RFC 1771)
  - Apply 30s (default) delay on announcements
  - MRAI on withdrawals (WRATE) allowed per RFC 4271
MRAI and RFD

- Need to decrease BGP chattiness and Path Exploration
- Minimum Route Advertisement Interval – MRAI (RFC 1771)
  - Apply 30s (default) delay on announcements
  - MRAI on withdrawals (WRATE) allowed per RFC 4271
- Route Flap Damping – RFD (RFC 2439)
  - Flapping = sequence of announcements and withdrawals
  - Suppress flapping prefixes for 1 hour (or more)
Route Flap Damping – RFD

Router 1

Flapping prefix!

Router 2

Adj-RIB-in

RIB

Adj-RIB-out

RFD

Buffer, Filter

Decision process

Router 3

Router 4

Adj-RIB-in

RIB

Adj-RIB-out

RFD

Buffer, Filter

Path: W

Path: W

AS1

Path: *3,2,1

AS2

Observer

A: 5,2,1

---

Prefix: 1.0.0.0/8

Path: W

Path: 3,2,1

AS3

AS4

AS5

Path: W
Route Flap Damping – RFD

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---
A: 5,3,2,1

Path: *4,3,2,1

Observer
A: 5,2,1
---
A: 5,3,2,1
W
Route Flap Damping – RFD

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---
A: 5,3,2,1
W
A: 5,4,3,2,1

Prefix: 1.0.0.0/8

Observer
A: 5,2,1
---
A: 5,3,2,1
W
A: 5,4,3,2,1

Prefix: 1.0.0.0/8
What are the problems?

- MRAI exhibits unpredictable behavior
- WRATE creates black holes
- RFD penalizes prefix owners even if the misbehavior happened further upstream
- We want something better!
Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
  Reduction of update load
  MRAI and PED convergence time compared

Conclusions and future work

PED algorithm

- Delay BGP announcements if the announced AS path is longer than the previously known AS path
PED algorithm illustrated

PEDI: Path Exploration Damping Interval

PED algorithm illustrated
PED algorithm illustrated

1. Router 1
   - Adj-RIB-in: Buffer, Filter
   - Announce new best path, no delay
   - A: 1.0.0.0/8 first announcement / previously withdrawn

2. Router 2
   - Decision process
   - Buffer, Filter

3. Router 3
   - Adj-RIB-out: Buffer, Filter
   - Update

4. Router 4
   - Adj-RIB-out: Buffer, Filter
   - Longer path!
   - Queue update, start PEDI timer
PED algorithm illustrated

Router 1 -> Adj-RIB-in
Buffer, Filter -> Decision process

Router 2 -> Adj-RIB-in
Buffer, Filter

Router 3

Router 4

buffer, filter

shorter path!
send update,
delete PED!
timer

Buffer, Filter

Adj-RIB-out

withdrawal!
send update,
delete PED!
timer

Buffer, Filter

Adj-RIB-out

Adjacent-RIB-in
RIB
Adjacent-RIB-out

update
Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
- Reduction of update load
- MRAI and PED convergence time compared

Conclusions and future work

PED – Data analysis

- Experiments using 24 hours of real BGP updates
- Two datasets:
  1. APNIC Pty. Ltd., Australia (2 peers)
PED – Data analysis

- Replayed using the *Quagga-Accelerator*

Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
  - Reduction of update load
  - MRAI and PED convergence time compared

Conclusions and future work
Reduction of update load – APNIC

Reduction of update load – Routeviews
Convergence time approximation

- Control Plane convergence (Optimality)
- Data plane convergence – Forwarding path (Reachability)
- Convergence measured from a single point of view (Optimality)
  - Router 2
- Convergence defined as a route being stable for 5 minutes

<table>
<thead>
<tr>
<th>240s</th>
<th>30s</th>
<th>30s</th>
<th>30s</th>
<th>120s</th>
<th>5min</th>
</tr>
</thead>
</table>

Time to converge
Detailed convergence time analysis

- Investigate Reachability vs. Optimality
- Analysis of a whole BGP system
- Impact on convergence of 4 events causing instability:
  - Link failure along the path – alternative path exists
  - Link recovery
  - Prefix withdrawal
  - Prefix announcement

Detailed convergence time analysis

- Experimental analysis over 20 testruns
- Simple example topology
  - ASes are single BGP speakers
- Example: 30s MRAI on all ASes and 35s PEDI on all ASes

Prefix Origin: 1.0.0.0/8

AS_1 -- AS_2 -- AS_3 -- AS_4 -- AS_5 -- AS_6

AS_10 -- AS_11 -- AS_12 -- AS_13

AS_20 -- AS_21 -- AS_22
Optimality – Announcement of initial route

Stable System:

Prefix Origin: 1.0.0.0/8

* 1
* 2,1
* 3,2,1
* 4,3,2,1
6,13,12,11,10,1
* 13,12,11,10,1
5,4,3,2,1
6,13,12,11,10,1
* 12,11,10,1
* 11,10,1
22,21,20,10,1
* 10,1
* 1
* 10,1
* 20,10,1
* 21,20,10,1
12,11,10,1

Announcement of initial route at AS₆:

- PED: 0 seconds
- MRAI: 60-120 seconds

Reachability – Link failure

Link failure between AS₁₀ and AS₁₁:

Prefix Origin: 1.0.0.0/8

* 1
* 2,1
* 3,2,1
* 4,3,2,1
6,13,12,11,10,1
* 13,12,11,10,1
5,4,3,2,1
12,11,10,1

Reachability achieved (AS₁₁)

- PED: 0 seconds
- MRAI: 0-4 or 29-30 seconds
Optimality – Link failure

Link failure between AS_{10} and AS_{11}:

Optimality achieved:
- PED: 66 seconds (+-jitter)
- MRAI: 2-58 seconds

Optimality – Link recovery

Link recovery between AS_{10} and AS_{11}:

Optimality achieved:
- PED: 0 seconds
- MRAI: 31-33 and 55-60 seconds
Optimality/Reachability – Prefix withdrawn

Prefix withdrawal at $A_S_1$

Prefix Origin: 1.0.0.0/8

- Optimality achieved (route withdrawn on every AS):
  - PED: 0 seconds
  - MRAI: 0 seconds

Optimality/Reachability – Prefix (re)-announced

Prefix announcement at $A_S_1$

Prefix Origin: 1.0.0.0/8

- Optimality achieved (same as initial announcement):
  - PED: 0 seconds
  - MRAI: 32-34, 58-60, 76-90 seconds
Conclusions

In this talk:

- PED decreases update load
- PED converges to Reachability as fast or faster than MRAI
- PED converges to Optimality slower than MRAI in one case

In the paper:

- PED interacts well with MRAI
- PED can be deployed incrementally
- A single PED speaker is beneficial to the BGP system
- 35s PEDI is a safe default value in the MRAI dominated Internet

In the future:

- Dynamic PEDI per prefix
- More heuristics