Effect of MRAI Timers and Routing Policies on BGP Convergence Times

Rajvir Gill, Ravinder Paul, and Ljiljana Trajković
Simon Fraser University
Vancouver, British Columbia, Canada

E-mail: {rajvirg, rpa28, ljilja}@sfu.ca
http://www.ensc.sfu.ca/~ljilja/cnl

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

Introduction

- BGP is de-facto inter AS routing protocol
- Operates successfully in a network of the Internet's size
- Supports CIDR
- Peer routers exchange four types of messages:
  - open
  - update
  - notification
  - keepalive
- BGP utilizes a path vector algorithm called the best path selection algorithm to select the best path

BGP : Border Gateway Protocol
AS : Autonomous System
CIDR: Classless Inter-Domain Routing

Minimal Route Advertisement Interval: MRAI

- MRAI is the interval limitation that affects BGP convergence
- Default value: 30 s
- MRAI timers control the MRAI value:
  - per-destination
  - per-peer
- Optimal MRAI value depends on:
  - network size
  - topology
  - traffic volume
  - network conditions

MRAI timers

- Per-destination:
  - associated with each network destination
  - may not be used because of the Internet size
- Per-peer:
  - associated with each peer in the network
  - starts ticking when the source router sends a route advertisement to peers
  - adversely affect advertisements to each destination

Processing delay

- Total time of an update waiting in the queue and the time required for a BGP router to process
- Uniform processing delay:
  - BGP router processes update messages sequentially
  - the delay in processing updates affects the processing time of update messages that follow
- Measurements:
  - update messages are processed within 200 ms
  - average processing time is 101 ms with the upper bound of 400 ms

- T. G. Griffin and B. J. Premore, “An experimental analysis of BGP convergence time,” in *Proc. ICNP*, Riverside, CA, USA, Nov. 2001, pp. 53–61.
- A. Feldmann, H. Kong, O. Maennel, and A. Tudor, “Measuring BGP pass-through times,” in *Proc. PAM*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 267–277.

Routing policies

- BGP allows a user to configure peer-to-peer policies
- Cause persistent route oscillations that affect the BGP stability and convergence time
- RFD prevents BGP route oscillations caused by network instabilities
- Routing policies affect:
  - BGP convergence time
  - number of updates
  - number of flaps

RFD: Route Flap Damping

Proposed algorithm and routing policies

- MRAI with Flexible Load Dispersing (FLD-MRAI) algorithm:
  - propose modifications to reusable timers
  - change MRAI durations based on BGP advertisement events
  - heterogeneous and large networks
  - perform well in networks with unspecified traffic load
- Two BGP routing policies are implemented in the routing policy module (BGP-RP) in ns-2.34:
  - AS path list policy
  - community list policy

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

BGP convergence

- BGP convergence time: the time interval between first update message sent, until all update messages that are a consequence of the original update are received
- Adaptive MRAI timers decrease the BGP convergence time and guarantees network stability
- PED algorithm proposed timer: 35 s
- Processing efficiency of router's CPU affects the BGP convergence time
    - N. Laskovic and Lj. Trajkovic, "BGP with an adaptive minimal route advertisement interval," in *Proc. IPCCC*, Phoenix, AZ, USA, Apr. 2006, pp. 142–151.
    - G. Huston, M. Rossi, and G. Armitage, "A technique for reducing BGP update announcements through path exploration damping," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 8, pp. 1271–1286, Oct. 2010.
    - T. G. Griffin and B. J. Premore, "An experimental analysis of BGP convergence time," in *Proc. ICNP*, Riverside, CA, USA, Nov. 2001, pp. 53–61.

PED: Path Exploration Damping

BGP convergence

- SSLD and optimal values for MRAI reduce the BGP convergence time
- Delay due to router or link failure increases BGP convergence time
- Shortest path to destination decreases BGP convergence time
- A router's CPU load depends on the number of BGP messages
- High CPU utilization of a BGP router causes delay
- Mismatch in policy configurations between two ASes may also cause network instabilities
  - S. Aggarwal and M. Aggarwal, "Dynamic load balancing based on CPU utilization and data locality in distributed database using priority policy," in *Proc. ICSTE*, Phuket, Thailand, Oct. 2010, vol. 2, pp. 388–391.
  - S. Agarwal, C. Chuah, S. Bhattacharyya, and C. Diot, "Impact of BGP dynamics on router CPU utilization," in *Proc. PAM*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 278–288.

SSLD: Sender Side Loop Detection
CPU : Central Processing Unit

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

FLD-MRAI algorithm: CPU utilization and processing delay

- Processing of updates:
  - normal load: DoP prefers the shortest path
  - high load: DoP prefers a longer path in the presence of the shortest path
  - empirical value: 200 ms
  - MRAI consists of two states: idle and processing

DoP: Degree of Preference

FLD-MRAI algorithm: CPU utilization and processing delay

- Calculation of available CPU:
  - CPU utilization is high, then the router responds slowly to subsequent requests in the queue
  - based on the priority of updates:

$$CPU_{available} = 100 - CPU_{active}$$
$$CPU_{active} = 100 * (CPU_{current}/CPU_{max})$$

$CPU_{available}$: percentage of available CPU of the neighboring router
CPU_{active} : percentage of active CPU utilization of the neighboring router
$CPU_{current}$: current CPU utilization
CPU_{max} : maximum CPU utilization

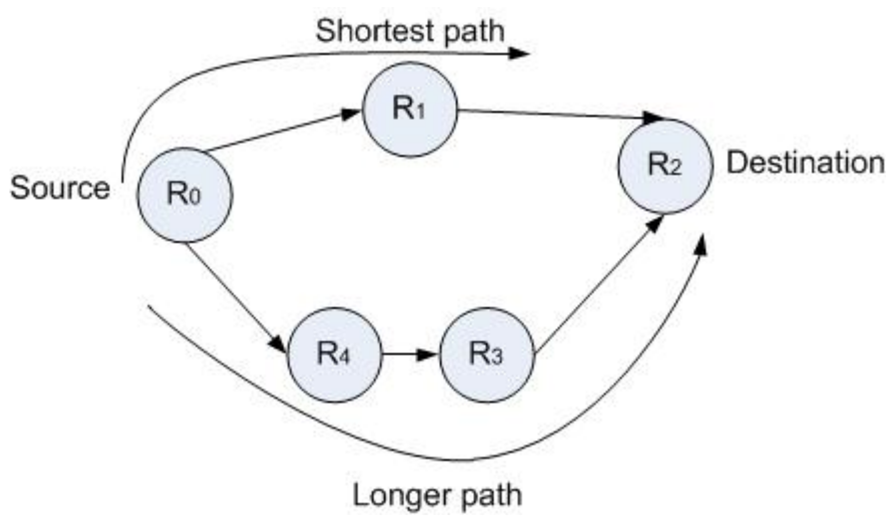
  - calculated every time a router receives the updates of a new or withdrawn route

FLD-MRAI algorithm: CPU utilization and processing delay

- Modified DoP (DoP_{mod}):
  - function of $Route_{info}$ and $CPU_{available}$, where $Route_{info}$ is the route having the shortest path
  - new, replaced, and/or withdrawn route
  - path with the highest value of DoP_{mod} is given the highest priority

FLD-MRAI algorithm: example with five routers

- R_0 is the source router and it advertises to destination R_2
- Two possible paths: R_0-R_1-R_2 and R_0-R_4-R_3-R_2
- Original BGP router: R_0-R_1-R_2
- FLD-MRAI: R_0-R_1-R_2 (normal load) and R_0-R_4-R_3-R_2 (high load)



Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

FLD-MRAI algorithm: modified reusable timers

- MRAI timer depends on network conditions and advertisement events:
  - Tdown: link failure
  - Tup : link failure recovery
  - Tlong : router failure
  - Tshort: router failure recovery
- Occurrence of advertisement events:

Events	Number of events occurring during BGP convergence period
Tdown	43.4%
Tup	39.9%
Tlong	7.3%
Tshort	7.4%
Unidentified	2.0%

- D. Pei and J. V. Merwe, “BGP convergence in virtual private networks,” in *Proc. IMC*, Rio de Janeiro, Brazil, Oct. 2006, pp. 283–288.

FLD-MRAI algorithm: modified reusable timers

- After the processing time, three events may occur:
  - new update message is received
  - no new update is received
  - MRAI reusable timer has expired
- Calculation of idle time:

$$T_{idle}\ (D) = MRAI_{total} - M_{last}$$

T_{idle} (D) : is thidle time of the destination

$MRAI_{total}$: total MRAI interval

M_{last} : time instance of the last message received

FLD-MRAI algorithm: modified reusable timers

- One reusable timer is required for all paths advertised during a short time interval
- Number of rounds per reusable MRAI timer controls the duration of MRAI round:

$$MRAI_{duration} = R_n * (t_n * g)$$

$MRAI_{duration}$: duration of MRAI round
R_n : number of rounds per reusable MRAI timer
t_n : number of reusable MRAI timers
g : granularity

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

FLD-MRAI algorithm: duration of MRAI

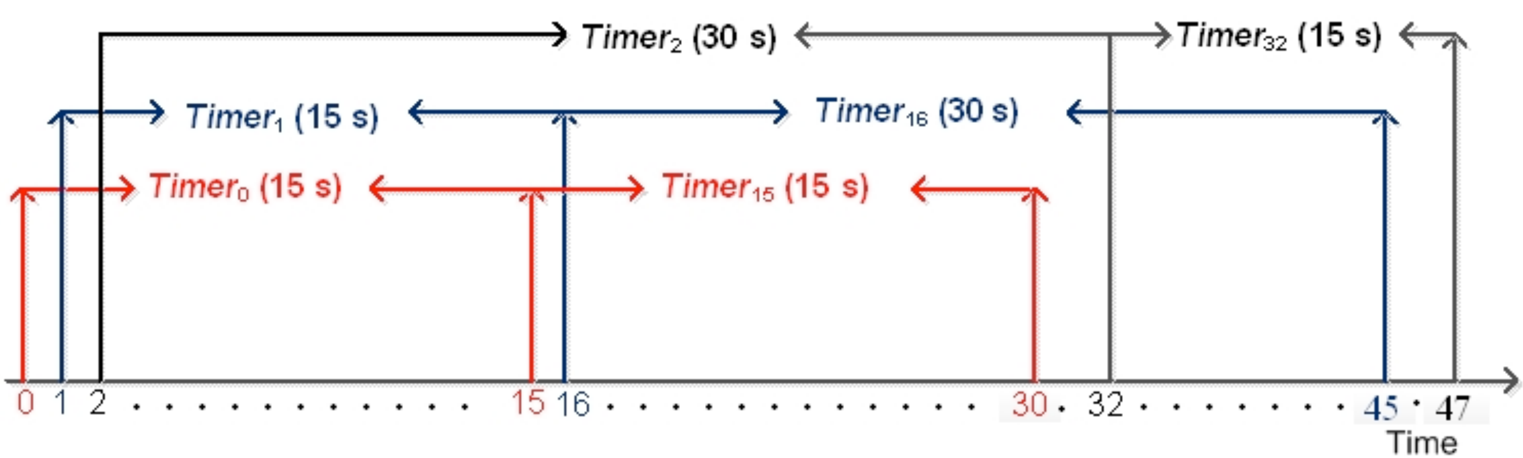
- Duration of MRAI timers:
  - longer duration: Tdown and Tlong
  - smaller duration: Tup and Tshort
- MRAI value:
  - same MRAI value: Tdown and Tlong
  - same MRAI value: Tup and Tshort
- Minimum duration: 15 s

FLD-MRAI algorithm: duration of MRAI

- Idle time longer than 1 s:
  - Tshort or Tup
  - shortest path becomes available
  - duration of MRAI round: 15 s
  - use one MRAI round: 15 s
- Idle time smaller than 1 s:
  - Tlong or Tdown
  - router or link failure
  - duration of MRAI round: 30 s
  - use two MRAI rounds: 15 s

FLD-MRAI algorithm: example of reusable timers

- 15 reusable MRAI timers with granularity 1 s
- $Timer_0$ lasts one round of 15 s (Tshort or Tup)
- $Timer_2$ lasts one round of 30 s (Tdown or Tlong)
- After expiration, the duration of timers depends on the idle time



FLD-MRAI algorithm: pseudocode

```
when sending advertisement of destination D to peers at t_o
        set (S_i)    // priority numbers on received updates according to the shortest path
        if (C1< C2) // calculate and compare the available CPU utilization of 1st and 2nd priority
                neighboring router
        if W(t) < T(t) // calculate and compare waiting time and transmission time
        else (wait in queue of the 1st priority path)
        if dop2 < dop1 // calculate and compare the degree of preference
        choose 2nd priority path
        MRAI = 30 s
        goto processing state
        else (wait in queue of the 1st priority path)
        else if (C1> C2)
        wait in queue of the 1st priority path // duration of MRAI is based on the idle time
        goto processing state

when initiation of new round
        if (Idle(D) > 1 s) // Tshort or Tup may occur
        set modified_reusable timer = 15 s
        else if (e ∈ network failure) // event changes due to the network failure
        choose 2nd priority path  // after expiration of the timer
        set modified_reusable timer = 30 s
        goto processing state
        else if (e ∉ network failure)
        goto processing state
        else if (Idle(D) < 1 s)  // Tlong or Tdown may occur
        set modified_reusable timer = 30 s
        else if (Pt∈ Ps) // if the shortest path becomes available
        choose shortest path // after expiration of the timer
        set modified_reusable timer = 15 s
        goto processing state
        else (Pt ∉ Ps) // if the shortest path is not available
        goto processing state
```

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

FLD-MRAI algorithm: space and time complexity

- Space complexity:
  - depends on the number of non-converged routes (n)
  - router keeps three variables of each non-converged route: $CPU_{current}$, CPU_{max}, and M_{last}
  - variables are integer counters that a router may easily store
  - space complexity is $O(n)$

FLD-MRAI algorithm: space and time complexity

- Time complexity:
  - division, multiplication, and subtraction operations : $CPU_{available}$, T_{idle}, and $MRAI_{duration}$
  - division and multiplication depends on n
  - subtraction is constant
  - approximate these variables with constants equal to their maximum values
  - calculation of variables do not depend on n
  - FLD-MRAI requires two subtractions, three multiplications, and one division

FLD-MRAI algorithm: space and time complexity

- Time complexity:
  - BGP router may send only one advertisement and one withdrawal during a single MRAI round
  - number of neighbors and non-converged routes during one MRAI round affects the maximum number of update messages
  - time complexity of the computation of variables and idle time is O(n) if the number of neighbors is constant

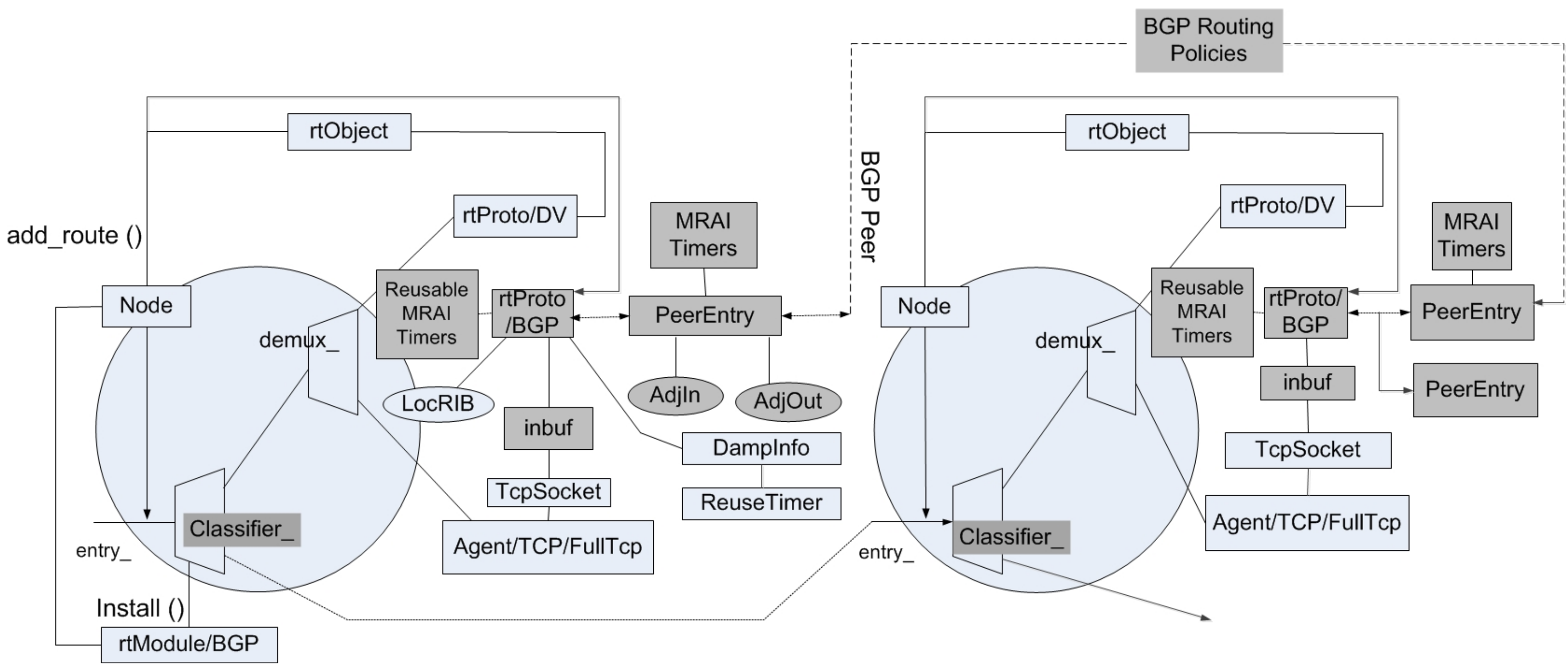
Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

Implementation: FLD-MRAI and BGP-RP

- ns-2.34 network simulator
- ns-BGP 2.0 developed module (ported from the SSFNET simulator)
- Routing structure of a modified ns-2 node:
  - forwarding plane: categorizes the received packets whether to be processed or forwarded to neighboring nodes
  - control plane: controls computation, maintenance, and implementation of routes in routing tables

Implementation: FLD-MRAI and BGP-RP



Implementation: BGP-RP

- BGP-RP configures AS path list and community list routing policies by using regular expressions
- Uses tre-0.8.0 library for regular expressions
- AS path list:
  - filters the BGP AS path attributes that define the entire set of AS numbers
  - uses a regular expression string to identify the attribute pattern to deny or permit the list

Implementation: BGP-RP

- Community list:
  - numbered or named
  - identifies and filter the routes according to common attributes between two networks

when receiving a route r with prefix *d* from neighbor
if (rouPolicy = 1)
        // routing policy of source matches with neighbor
update RIB
        // update the Routing Information Base
else if (rouPolicy = 0)
        // routing policy of source does not match with neighbor
update RIB
        // update the Routing Information Base

Simulation scenarios: performance comparisons

- FLD-MRAI algorithm:
  - FLD-MRAI-30
  - FLD-MRAI-15
  - default-MRAI-30
  - default-MRAI-15
  - adaptive MRAI
- BGP-RP:
  - BGP without policies
  - BGP with AS path list policy
  - BGP with community list policy

Simulated topologies

Topology	Number of nodes	Topology generator
Topology 1	67	Manually from BCNET BGP traffic
Topology 2	100	GT-ITM
Topology 3	200	GT-ITM
Topology 4	300	BRITE
Topology 5	500	BRITE

GT-ITM: Georgia Tech Internetwork Topology Models
BRITE : Boston university Representative Internet Topology gEnerator

Assumptions

- Do not consider RFD when evaluating the performance of FLD-MRAI
- Each AS consists of a single BGP router
- BGP convergence procedure is complete if BGP router receives no update message from other BGP routers within 60 s

RFD: Route Flap Damping

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

Performance evaluation: network Topology 1

- FLD-MRAI algorithm: average convergence time (s) for network Topology 1 for various BGP options

Scenarios	default-MRAI-30	default-MRAI-15	adaptive MRAI	FLD-MRAI-15	FLD-MRAI-30	FLD-MRAI
Tshort	126.66	145.49	131.90	68.73	69.40	66.93
Tlong	138.47	143.81	142.62	78.45	79.60	77.07
Tup	126.39	145.50	132.02	66.03	67.80	65.33
Tdown	138.52	145.62	141.40	74.73	75.86	74.73
High load	1,192.07	1,196.21	1,047.42	767.62	782.62	764.63

Performance evaluation: network Topology 1

- FLD-MRAI algorithm: overall number of update messages for network Topology 1 for various BGP options

Scenarios	default-MRAI-30	default-MRAI-15	adaptive MRAI	FLD-MRAI-15	FLD-MRAI-30	FLD-MRAI
Tshort	726	1304	870	375	373	374
Tlong	608	1073	1142	452	456	445
Tup	681	1262	763	399	420	391
Tdown	673	1251	751	394	415	386
High load	14,911	27,566	12,362	10,094	10,549	4,526

Performance evaluation: network Topology 1

- BGP-RP: comparison of BGP module with and without policies: convergence time, number of updates, and number of flaps

Scenario	Convergence time (s)	Number of updates	Number of flaps
BGP without policies	129.29	745	302
BGP-RP with AS path list	130.65	1,102	437
BGP-RP with community path list	147.13	970	430

Performance evaluation

- The average percentage of improvement of FLD-MRAI over default MRAI (30 s) based on different network topologies

Events	Convergence time (s)	Overall number of updates
Tshort	24%	47%
Tlong	26%	51%
Tup	27%	47%
Tdown	25%	46%

Roadmap

- Introduction
- BGP convergence
- FLD-MRAI algorithm:
  - CPU utilization and processing delay
  - modified reusable timers
  - duration of MRAI
  - space and time complexity
- Implementation of FLD-MRAI and BGP-RP
- Performance evaluation
- Conclusions

Conclusions

- Proposed FLD-MRAI algorithm employs:
  - BGP modifications to reduce the convergence time and number of update messages exchanged during normal and high traffic loads
  - modified DoP that depends on the calculation of available CPU
  - separate durations of MRAI intervals for different events that occur during BGP advertisements
  - modified reusable MRAI timers
- Simulation results show that FLD-MRAI performs better than other BGP options at the cost of computing available CPU of neighboring routers

Conclusions

- The CPU processing capability and duration of MRAI timers greatly affect the BGP convergence time
- Simulation results also show that AS path list policy performs similarly to BGP without any policy setup and performs better than the community list policy
- FLD-MRAI and BGP-RP with AS path list policy exhibit the best performance in networks with large diameter and may help improve performance of today's Internet

References: http://www.sfu.ca/~ljilja/cnl

- T. Farah, S. Lally, R. Gill, N. Al-Rousan, R. Paul, D. Xu, and Lj. Trajković, “Collection of BCNET BGP traffic,” in *Proc. ITC*, San Francisco, CA, USA, Sept. 2011, pp. 322–323.
- S. Lally, T. Farah, R. Gill, R. Paul, N. Al-Rousan, and Lj. Trajković, “Collection and characterization of BCNET BGP traffic,” in *Proc. IEEE PACRIM*, Victoria, BC, Canada, Aug. 2011, pp. 830–835.
- N. Laskovic and Lj. Trajković, “BGP with an adaptive minimal route advertisement interval,” in *Proc. IPCCC*, Phoenix, AZ, USA, Apr. 2006, pp. 142–151.
- W. Shen and Lj. Trajković, "BGP route flap damping algorithms," in *Proc. SPECTS* 2005, Philadelphia, PA, July 2005, pp. 488-495.
- T. D. Feng, R. Ballantyne, and Lj. Trajković, “Implementation of BGP in a network simulator,” in *Proc. ATS*, Arlington, VA, USA, Apr. 2004, pp. 149–154.

References: BGP

- S. Aggarwal and M. Aggarwal, “Dynamic load balancing based on CPU utilization and data locality in distributed database using priority policy,” in *Proc. ICSTE*, Phuket, Thailand, Oct. 2010, vol. 2, pp. 388–391.
- D. Pei and J. V. Merwe, “BGP convergence in virtual private networks,” in *Proc. IMC*, Rio de Janeiro, Brazil, Oct. 2006, pp. 283–288.
- R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, “Dynamics of hot-potato routing in IP networks,” in *Proc. ACM SIGMETRICS*, New York, NY, USA, June 2004, pp. 307–319.
- S. Agarwal, C. Chuah, S. Bhattacharyya, and C. Diot, “Impact of BGP dynamics on router CPU utilization,” in *Proc. PAM*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 278–288.
- G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos, “Power-laws and the AS-level Internet topology,” *IEEE/ACM Trans. Networking*, vol. 11, no. 4, pp. 514–524, Aug. 2003.
- C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, “Delayed Internet routing convergence,” in *Proc. ACM* SIGCOMM, Stockholm, Sweden, Aug. 2000, pp. 175–187.
- M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the Internet topology,” in *Proc. SIGCOMM*, Cambridge, MA, USA, Sept. 1999, pp. 251–262.
- T. Griffin and G. Wilfong, “An analysis of BGP convergence properties,” in *Proc. SIGCOMM*, Cambridge, MA, USA, Aug. 1999, pp. 277– 288.
- C. Labovitz, G. Malan, and F. Jahanian “Origins of Internet routing instability,” in *Proc. INFOCOMM*, New York, NY, USA, Mar. 1999, pp. 218–226.
- Y. Rekhter and T. Li, “A border gateway protocol 4 (BGP-4),” *IETF* RFC 1771, Mar. 1995.
- Operational Experience with the BGP-4 protocol [Online]. Available: http://tools.ietf.org/html/draft-ietf-idr-bgp4-op-experience-01.

References: MRAI

- B. Wang, “The research of BGP convergence time,” in *Proc. ITAIC*, Chongqing, China, Aug. 2011, vol. 2, pp. 354–357.
- A. Fabrikant, U. Syed, and J. Rexford, “There's something about MRAI: timing diversity can exponentially worsen BGP convergence,” in *Proc. INFOCOMM*, Shanghai, China, Apr. 2011, pp. 2975–2983.
- G. Huston, M. Rossi, and G. Armitage, “A technique for reducing BGP update announcements through path exploration damping,” *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 8, pp. 1271–1286, Oct. 2010.
- S. Deshpande and B. Sikdar, “On the impact of route processing and MRAI timers on BGP convergence times,” in *Proc. GLOBECOM*, Dallas, TX, USA, Nov. 2004, vol. 2, pp. 1147–1151.
- A. Feldmann, H. Kong, O. Maennel, and A. Tudor, “Measuring BGP pass-through times,” in *Proc. PAM*, Antibes Juan-les-Pins, France, Apr. 2004, pp. 267–277.
- B. Premore, An analysis of convergence properties of the border gateway protocol using discrete event simulation, Ph. D. Thesis, Dartmouth College, 2003.
- T. G. Griffin and B. J. Premore, “An experimental analysis of BGP convergence time,” in *Proc. ICNP*, Riverside, CA, USA, Nov. 2001, pp. 53–61.

References: routing policies and RFD

- W. Sun, Z. M. Mao, and K. G. Shin, “Differentiated BGP update processing for improved routing convergence,” in *Proc. ICNP*, Santa Barbara, CA, USA, Nov. 2006, pp. 280–289.
- Z. Mao, R. Govindan, G. Varghese, and R. Katz, “Route flap damping exacerbates Internet routing convergence,” *Computer Communication Review Journal*, vol. 32, no. 4, pp. 221–233, Aug. 2002.
- L. Gao and J. Rexford, “Stable Internet routing without global coordination,” in *Proc. ACM SIGCOMM*, San Diego, CA, USA, Dec. 2001, pp. 681–692.
- C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja, “The impact of Internet policy and topology on delayed routing convergence,” in *Proc. INFOCOMM*, Anchorage, AL, USA, Apr. 2001, pp. 537–546.
- K. Varadhan, R. Govindan, and D. Estrin, “Persistent route oscillations in inter-domain routing,” *Computer Networks Journal*, vol. 32, no. 1, pp. 1–16, Mar. 2000.
- C. Villamizar, R. Chandra, and R. Govindan, “BGP route flap damping,” *IETF* RFC 2439, Nov. 1998.
- Route flap damping considered useable [Online]. Available: http://www.iepg.org/2010-11-ietf79/101107.iepg-rfd.pdf.
- BGP++ [Online]. Available: http://www.ece.gatech.edu/research/labs/MANIACS/BGP++/.
- TRE [Online]. Available: http://laurikari.net/tre/.

References: topology generator

- BCNET [Online]. Available: https://wiki.bc.net/atl-conf/display/Content/Home.
- GT-ITM [Online]. Available: http://www.cc.gatech.edu/projects/gtitm/.
- BRITE [Online]. Available: http://www.cs.bu.edu/brite.

Thank you