



BGP with an adaptive minimal route advertisement interval

Nenad Lasković
nlaskovi@cs.sfu.ca

Communication Networks Laboratory
<http://www.ensc.sfu.ca/research/cnl>
School of Engineering Science
Simon Fraser University





Road map

- Introduction and motivation
- Border Gateway Protocol (BGP)
- **BGP with adaptive MRAI** (Minimal Route Advertisement Interval)
 - empirical model for BGP processing delay
 - reusable MRAI timers
 - the adaptive MRAI algorithm
- Performance analysis of BGP with adaptive MRAI
- Conclusions



Introduction

- The Internet consists of numerous heterogeneous networks without centralized control
- An **Autonomous System (AS)** is a group of networks controlled by a common administrative entity
- ASs communicate using **Border Gateway Protocol (BGP)** version 4, RFC 1771
- BGP is the *de facto* standard inter-domain routing protocol in today's Internet



Motivation

- One of the major problems of BGP is its **long convergence time**
 - unreachable destinations
 - packet loss
- **Solution: an algorithm that decreases BGP convergence time**
- The proposed algorithm should not change BGP messages format or BGP functioning

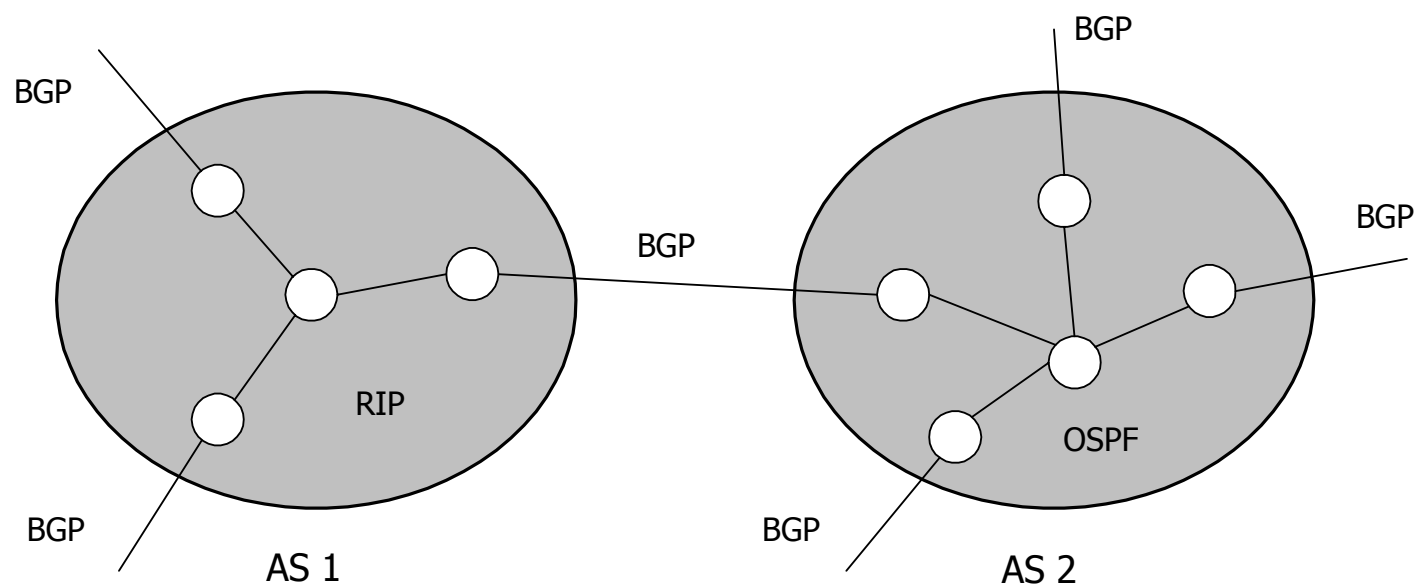


Road map

- Introduction and motivation
- **Border Gateway Protocol (BGP)**
- BGP with adaptive MRAI
 - empirical model for BGP processing delay
 - reusable MRAI timers
 - the adaptive MRAI algorithm
- Performance analysis of BGP with adaptive MRAI
- Conclusions



Routing in the Internet



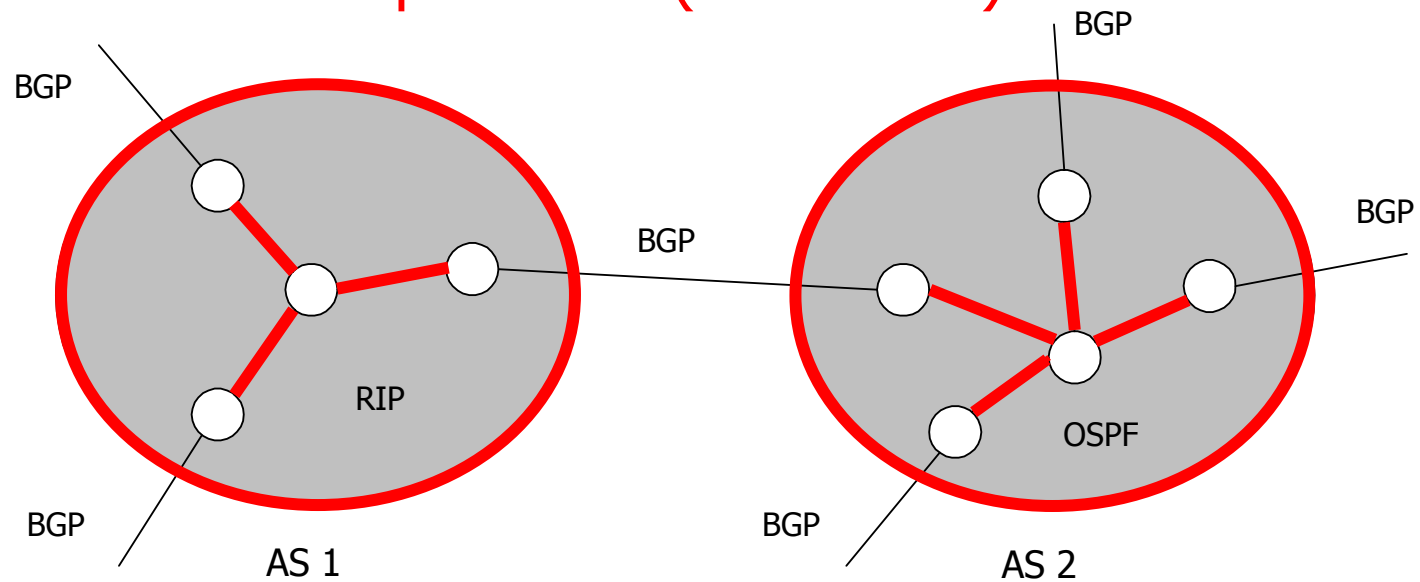
AS: Autonomous System
BGP: Border Gateway Protocol

RIP: Routing Information Protocol
OSPF: Open Shortest Path First



Routing in the Internet

intra-domain protocols (inside ASs)



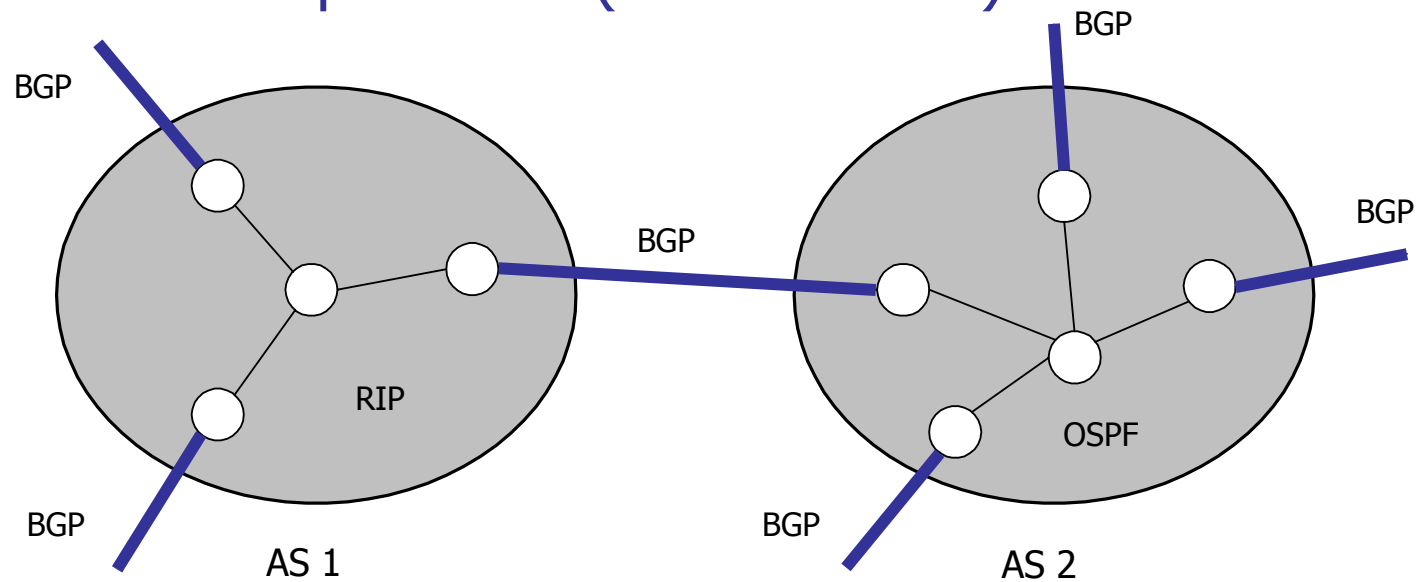
AS: Autonomous System
BGP: Border Gateway Protocol

RIP: Routing Information Protocol
OSPF: Open Shortest Path First



Routing in the Internet

inter-domain protocols (between ASs)

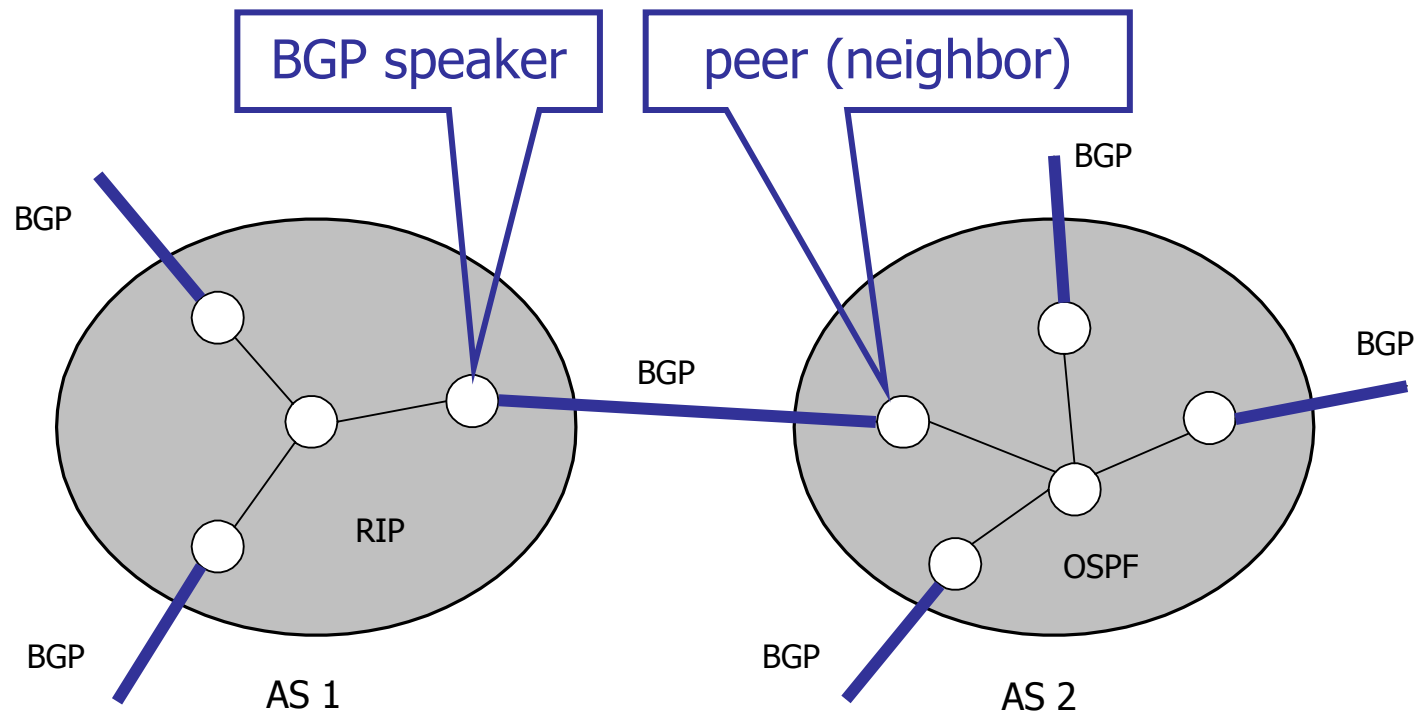


AS: Autonomous System
BGP: Border Gateway Protocol

RIP: Routing Information Protocol
OSPF: Open Shortest Path First



Routing in the Internet



AS: Autonomous System
BGP: Border Gateway Protocol

RIP: Routing Information Protocol
OSPF: Open Shortest Path First



Exchange of routing information

- Two main functions of BGP:
 - establishing routes (paths) between ASs
 - routing packets to their destinations (ASs)
- BGP distance metric: length of route in hops
- BGP speakers exchange information only when changes cause a replacement of the best routes
 - **advertisements**: new best path to a destination
 - **withdrawals**: destination unreachable



Dynamical behavior of BGP

- Changes of the Internet topology result in:
 - frequent changes of BGP routing tables
 - a large number of update messages
- **BGP convergence time**: from the time when the first update message is sent, until all update messages that are a consequence of the original update are received



Advertisement rate limiting

- Two conflicting requirements for BGP speakers:
 - minimize the number of sent update messages
 - react to changes in a timely manner
- **Minimal Route Advertisement Interval (MRAI):**
 - minimal time interval that must elapse between two consecutive advertisements of the same destination sent from one BGP speaker
 - controlled by use of **MRAI timers**
 - MRAI round is 30 s (RFC 1771)



MRAI timers

- **Per-destination** MRAI timers (RFC 1771):
 - one timer is associated with one destination
 - independent rate limiting for each destination
 - unfeasible: ~100,000 destinations per router
- **Per-peer** MRAI timers (RFC 1771):
 - one timer is associated with one peer
 - BGP speakers have less than 100 peers
 - disadvantage: all advertisements are delayed

$$\text{average delay of an advertisement} = \frac{MRAI}{2}$$



BGP convergence time

- **up (advertisement) phase**, a new destination is introduced to a network
 - convergence time T_{up} (estimation):

$$T_{up} \approx |length\ of\ the\ shortest\ path| \times \frac{MRAI}{2}$$

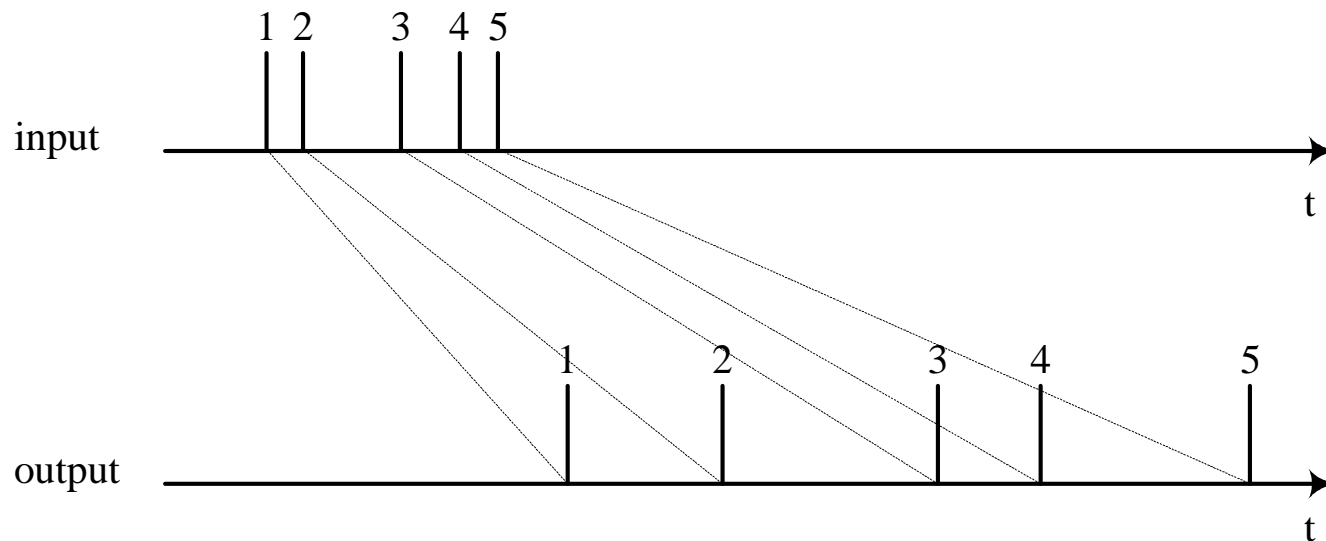
- **down (withdrawal) phase**, the only route to a destination is withdrawn from a network
 - convergence time T_{down} (upper bound):

$$T_{down} \leq |length\ of\ the\ longest\ path| \times \frac{MRAI}{2}$$



BGP processing delay: uniform model

- update messages are processed independently
- delay of each message is modeled with a **uniformly** distributed random variable



T. Griffin and B. Premore, "An experimental analysis of BGP convergence time," in *Proc. ICNP*, Riverside, CA, Nov. 2001, pp. 53–61.



BGP processing delay: measurements

- BGP speakers process groups of update messages in **fixed 200 ms processing cycles**
- 95% of messages are processed within 210 ms
 - BGP processing delay is independent of the number of received updates
- The uniform model estimates unrealistically high processing delays
 - Example: 20 messages -> ~10 s delay

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.



Previous work

- Each network has an **optimal MRAI** value that minimizes BGP convergence time
- Optimal MRAI values depend on:
 - network topology
 - traffic load
- A global MRAI value cannot be determined for the entire Internet

T. Griffin and B. Premore, "An experimental analysis of BGP convergence time," in *Proc. ICNP*, Riverside, CA, Nov. 2001, pp. 53–61.



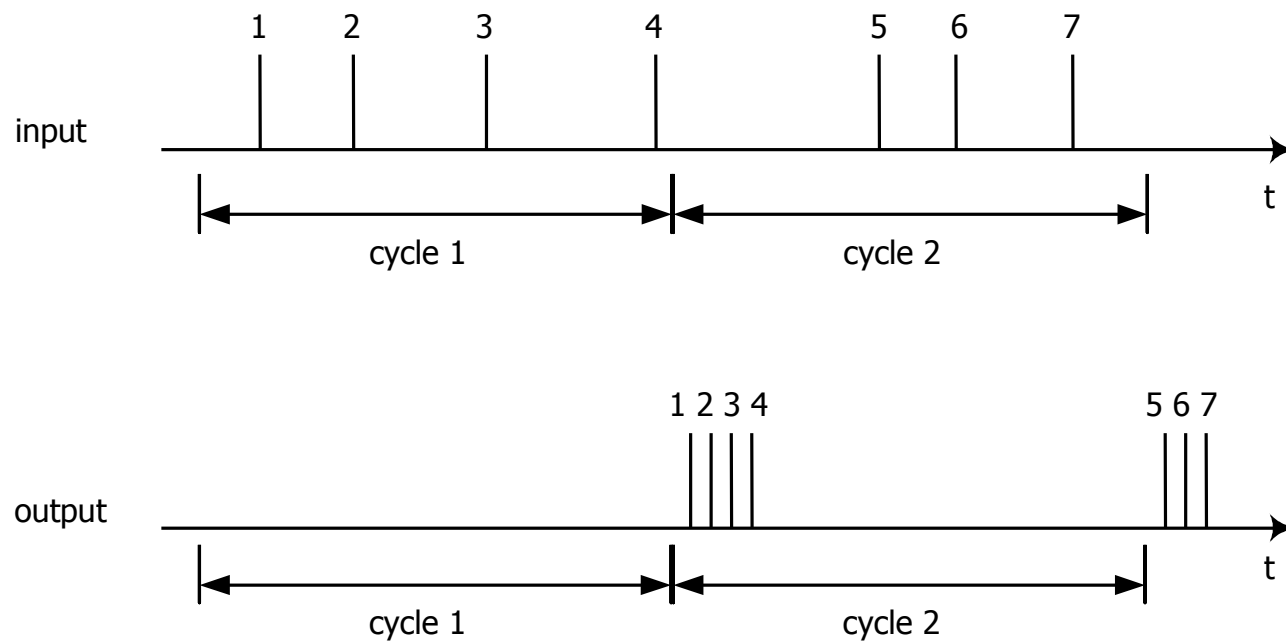
Road map

- Introduction and motivation
- Border Gateway Protocol (BGP)
- BGP with adaptive MRAI
 - empirical model for BGP processing delay
 - reusable MRAI timers
 - the adaptive MRAI algorithm
- Performance analysis of BGP with adaptive MRAI
- Conclusions

BGP processing delay: empirical model

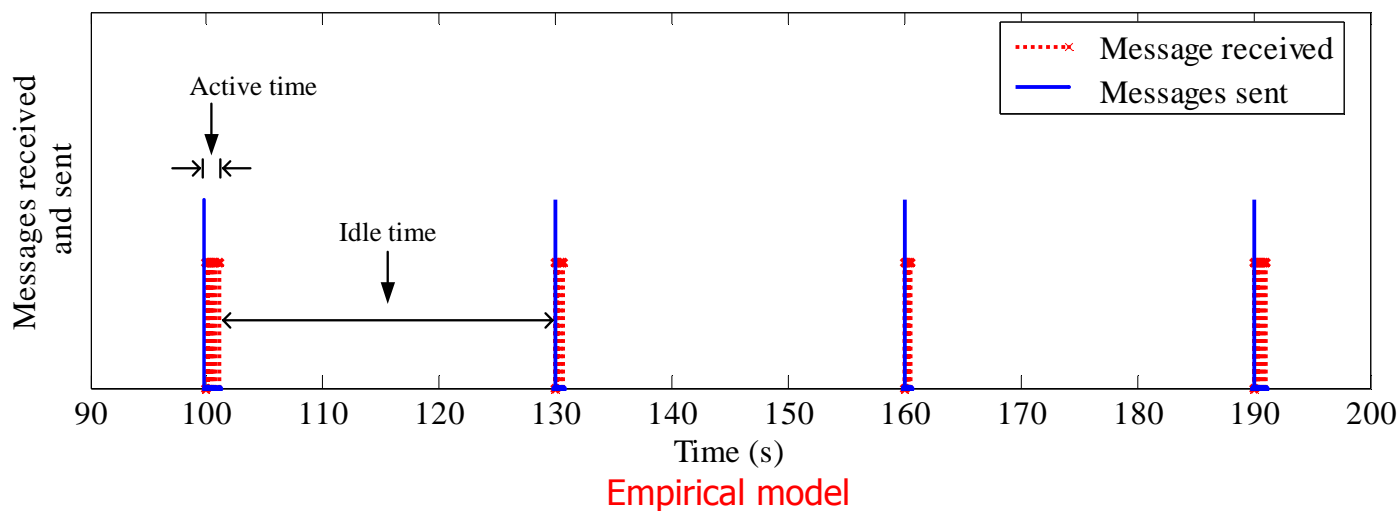
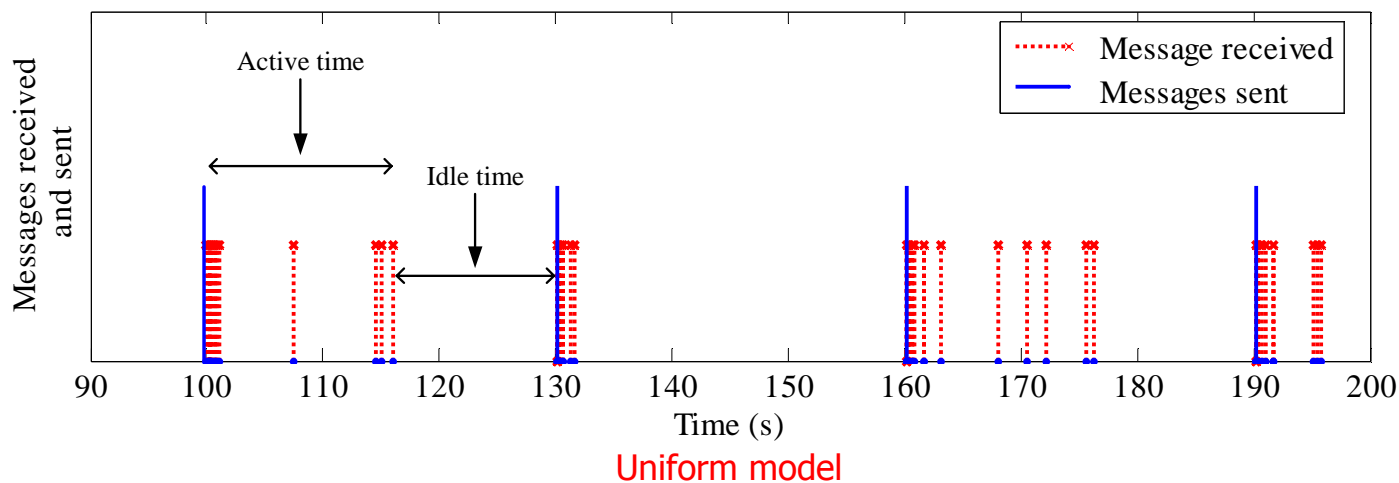


- Model based on measurements
- BGP speaker completes processing of all received updates at the end of the 200 ms processing cycle





Uniform vs. empirical model





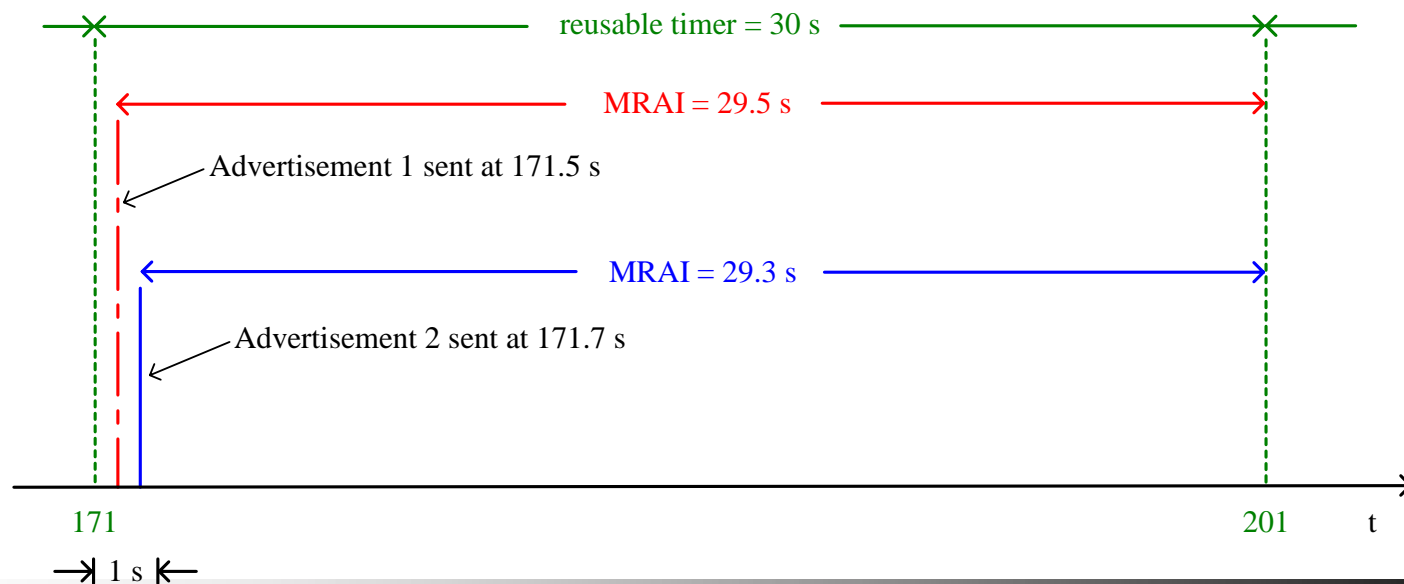
BGP with adaptive MRAI

- Goal: minimizing the idle time for each destination
 - MRAI round for each destination has to be equal to the active time
 - independent rate limiting for each destination
- BGP needs **new MRAI timers** and a **new algorithm that adaptively adjusts the duration of MRAI rounds**
- Solution: **BGP with adaptive MRAI**



Reusable MRAI timers

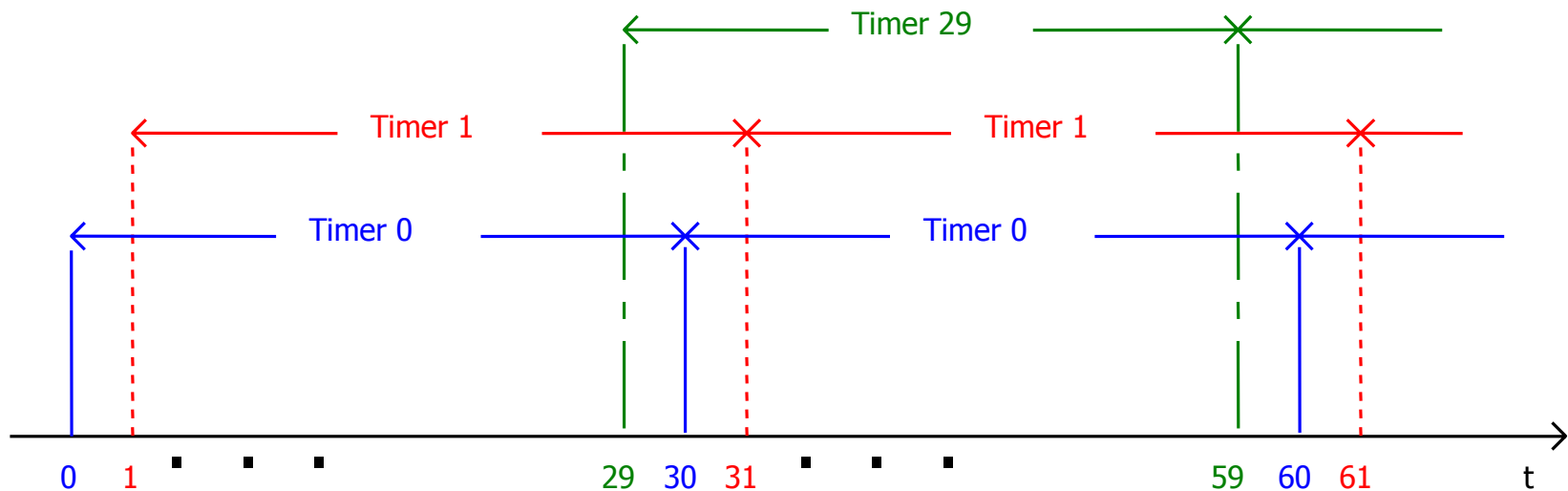
- Associating all route advertisements sent during a one second interval with a single **reusable MRAI timer**
- Instead of being exactly 30 s, an **MRAI round belongs to an interval** between 29 and 30 s





Reusable MRAI timers

- A BGP speaker needs only 30 reusable MRAI timers for all destinations and all peers
- Reusable MRAI timers achieve **independent** rate limiting





Adaptive MRAI algorithm

- Finding the optimal MRAI requires knowledge of the active time during an MRAI round
- BGP speakers may estimate the active time of the next round using information from previous rounds
- Durations of the adaptive MRAI rounds are estimated for each destination separately:

$$adaptiveMRAI_{n+1}(D) = avg_active_n(D) + 3 \times deviation_n(D)$$



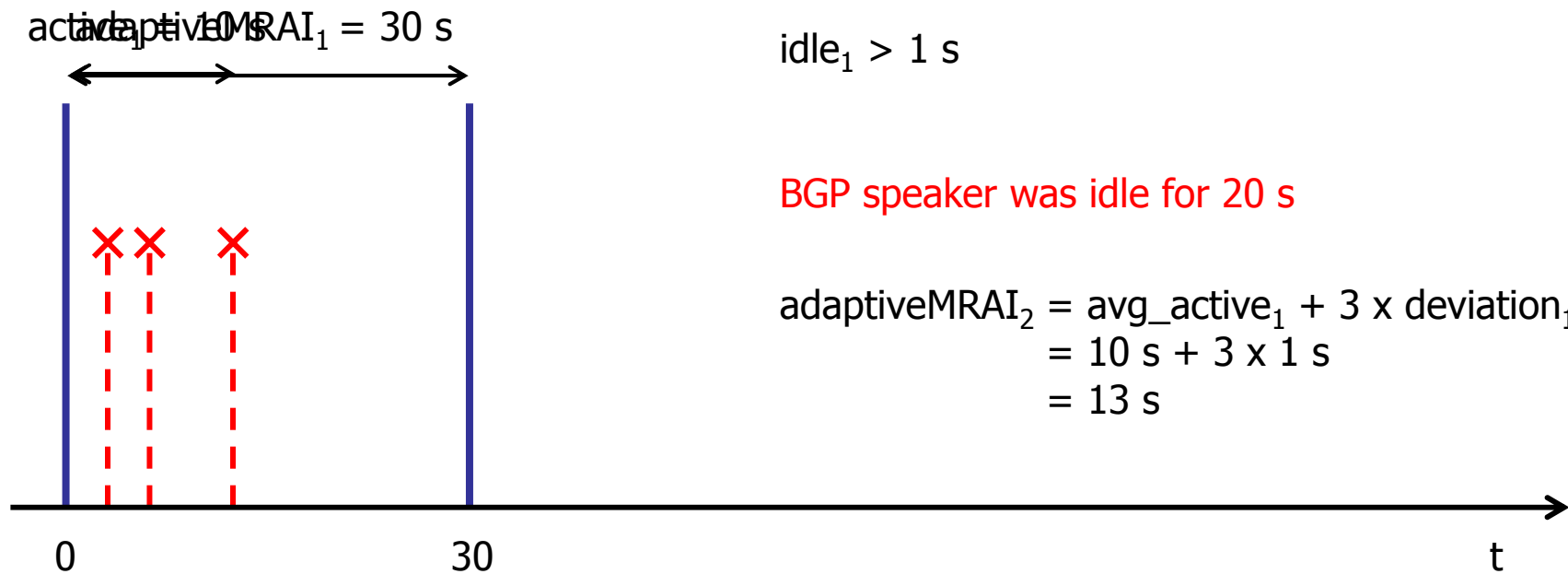
Adaptive MRAI algorithm: round 1

$$\begin{aligned} \text{idle}_1 &= \text{adaptiveMRAI}_1 - \text{active}_1 \\ &= 30 - 10 \text{ s} \\ &= 20 \text{ s} \end{aligned}$$

$$\text{idle}_1 > 1 \text{ s}$$

BGP speaker was idle for 20 s

$$\begin{aligned} \text{adaptiveMRAI}_2 &= \text{avg_active}_1 + 3 \times \text{deviation}_1 \\ &= 10 \text{ s} + 3 \times 1 \text{ s} \\ &= 13 \text{ s} \end{aligned}$$



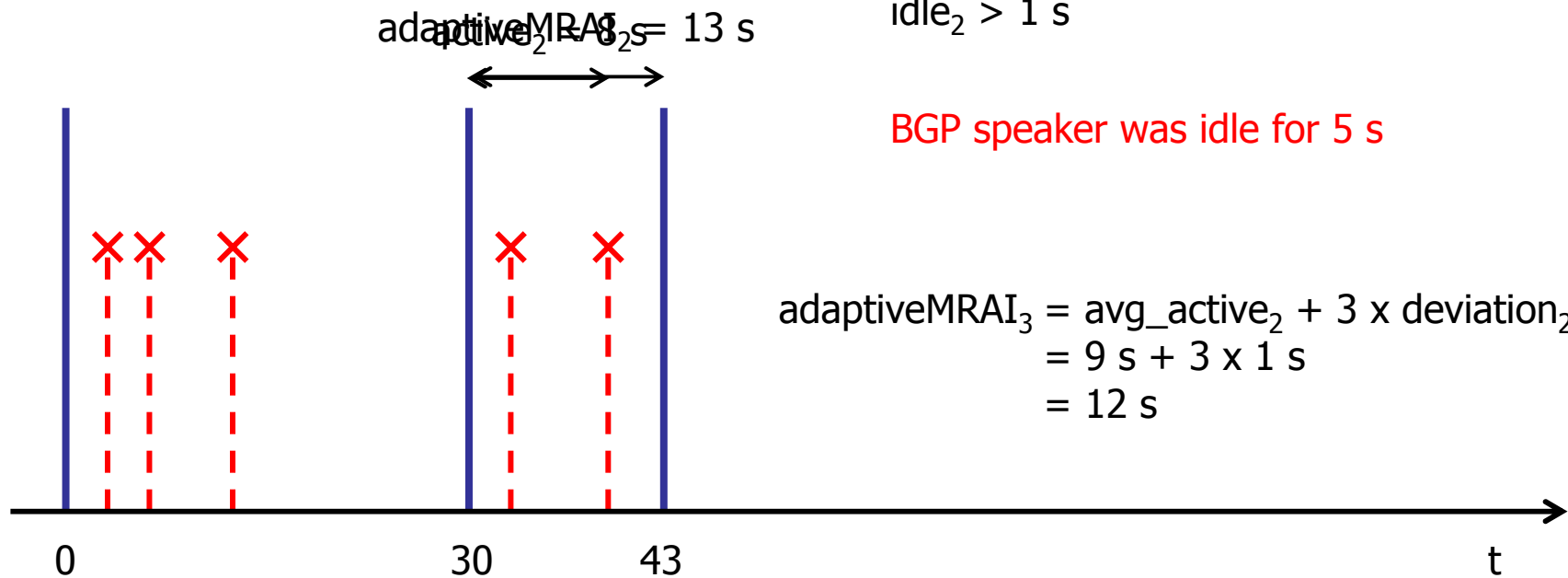


Adaptive MRAI algorithm: round 2

$$\begin{aligned} \text{idle}_2 &= \text{adaptiveMRAI}_2 - \text{active}_2 \\ &= 13 - 8 \text{ s} \\ &= 5 \text{ s} \end{aligned}$$

$$\text{idle}_2 > 1 \text{ s}$$

BGP speaker was idle for 5 s



$$\begin{aligned} \text{adaptiveMRAI}_3 &= \text{avg_active}_2 + 3 \times \text{deviation}_2 \\ &= 9 \text{ s} + 3 \times 1 \text{ s} \\ &= 12 \text{ s} \end{aligned}$$



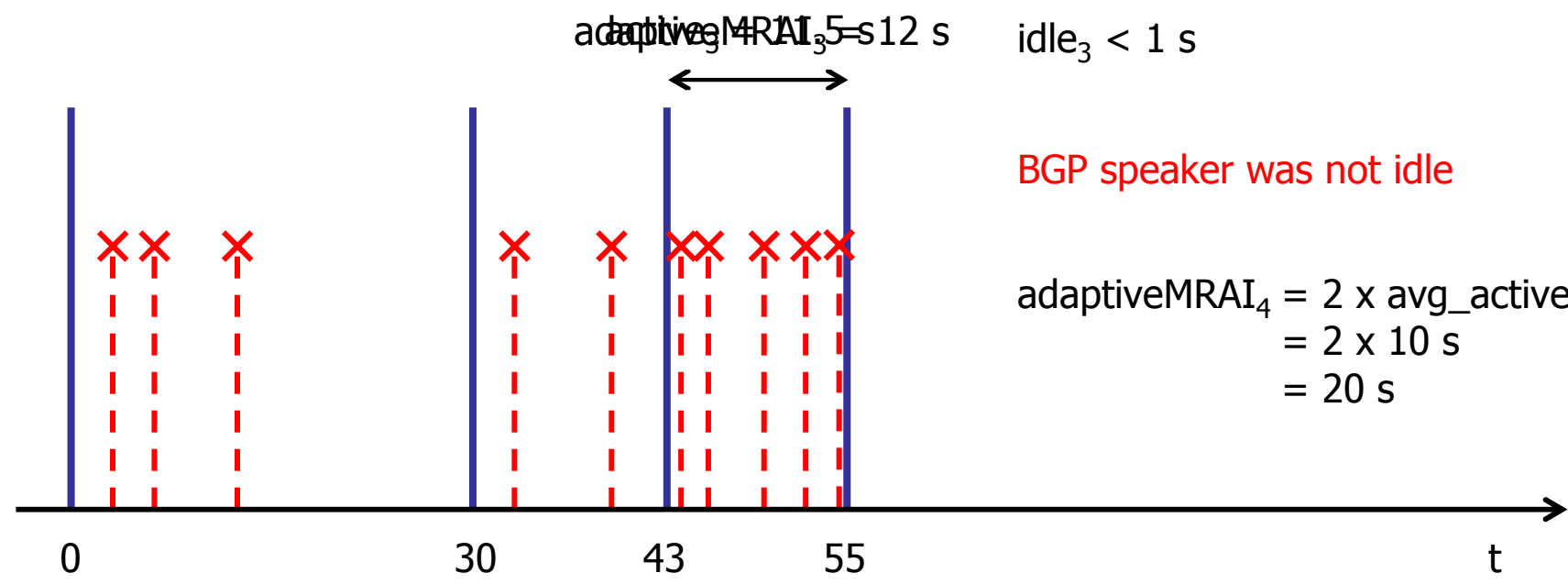
Adaptive MRAI algorithm: round 3

$$\begin{aligned} \text{idle}_3 &= \text{adaptiveMRAI}_3 - \text{active}_3 \\ &= 12 - 11.5 \text{ s} \\ &= 0.5 \text{ s} \end{aligned}$$

$$\text{adaptiveMRAI}_3 = 12 \text{ s} \quad \text{idle}_3 < 1 \text{ s}$$

BGP speaker was not idle

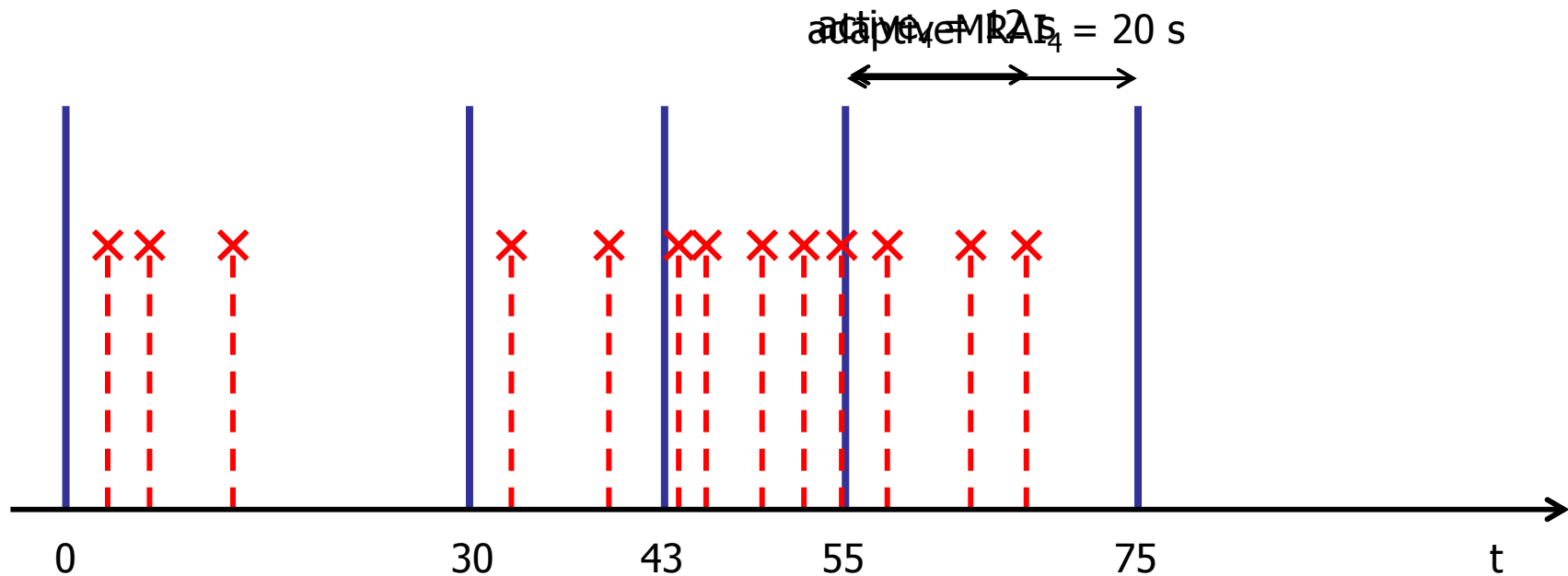
$$\begin{aligned} \text{adaptiveMRAI}_4 &= 2 \times \text{avg_active}_3 \\ &= 2 \times 10 \text{ s} \\ &= 20 \text{ s} \end{aligned}$$





Adaptive MRAI algorithm: round 4

$$\begin{aligned} \text{adaptiveMRAI}_5 &= \text{avg_active}_4 + 3 \times \text{deviation}_4 \\ &= 10.5 \text{ s} + 3 \times 2 \text{ s} \\ &= 16 \text{ s} \end{aligned}$$





The algorithm overhead

- Memory requirements:
 - reusable MRAI timers: 30 timers + pointer for each non-converged route (~ 100)
 - adaptive MRAI algorithm: 4 integers for each non-converged route
- Computational overhead of the adaptive algorithm:
 - 13 operations at the end of a MRAI round
 - the number of MRAI rounds \sim the number of non-converged routes
 - **computational complexity depends linearly on the number of non-converged routes**



Road map

- Introduction and motivation
- Border Gateway Protocol (BGP)
- BGP with adaptive MRAI
 - empirical model for BGP processing delay
 - reusable MRAI timers
 - the adaptive MRAI algorithm
- Performance analysis of BGP with adaptive MRAI
- Conclusions

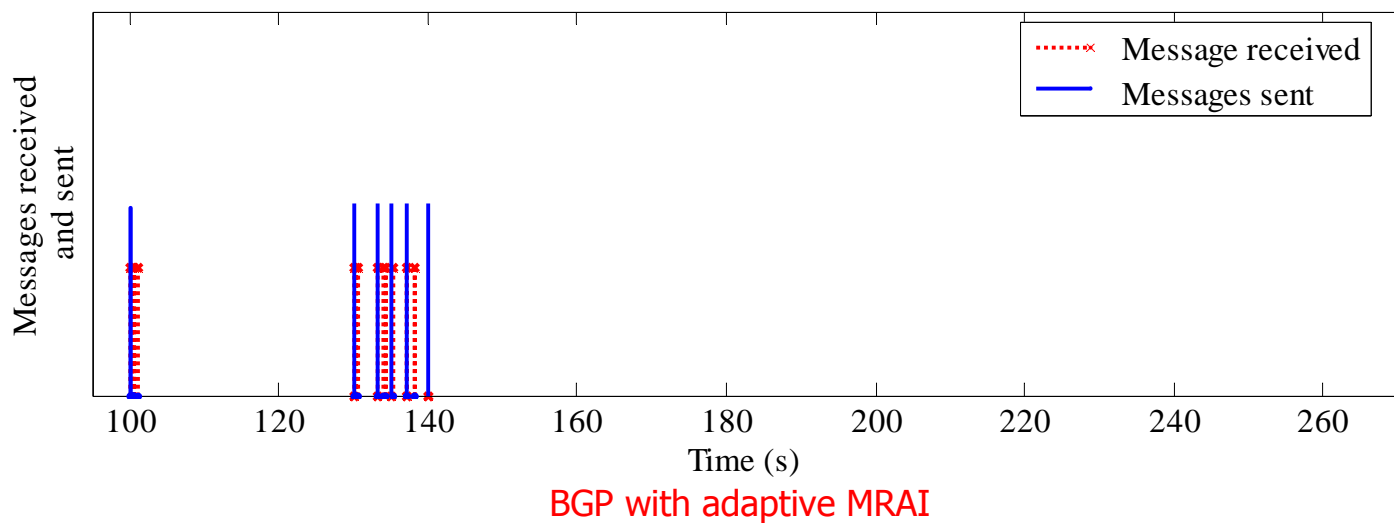
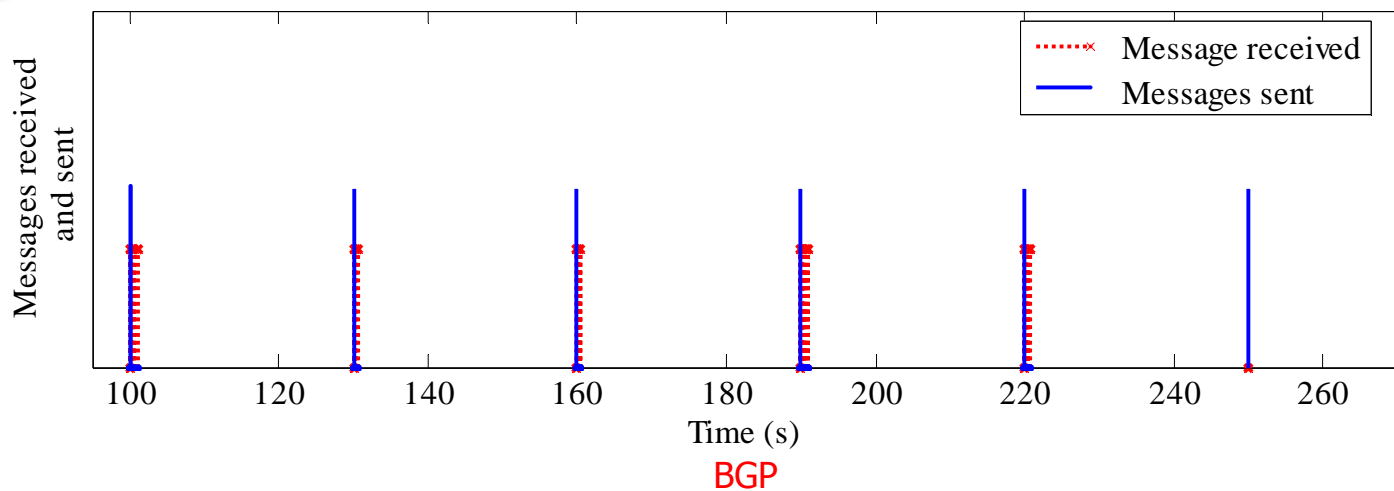


Simulation scenarios

- We implemented the algorithm using ns-2 and its BGP module (**ns-BGP 2.0**)
- Topologies:
 - completely connected graph with 15 nodes
 - two topologies (29 and 110 nodes) derived from the BGP routing tables (Route Views Project)
 - topology with 200 nodes obtained using topology generator BRITE
- Each simulation scenario is repeated **30 times** using 30 unique random number generator seeds

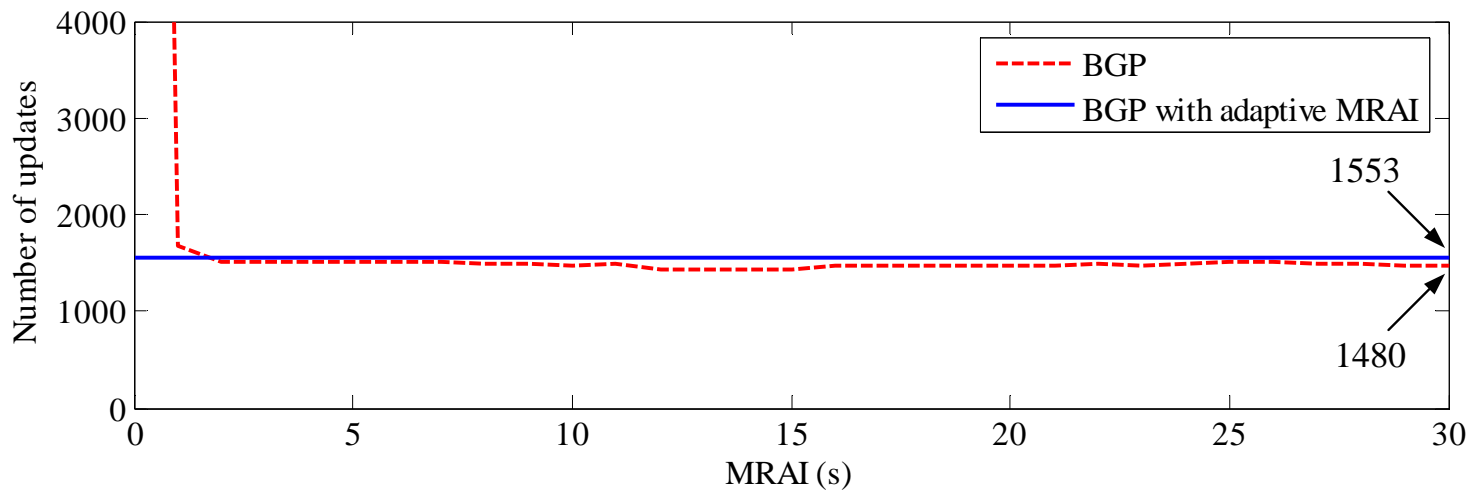
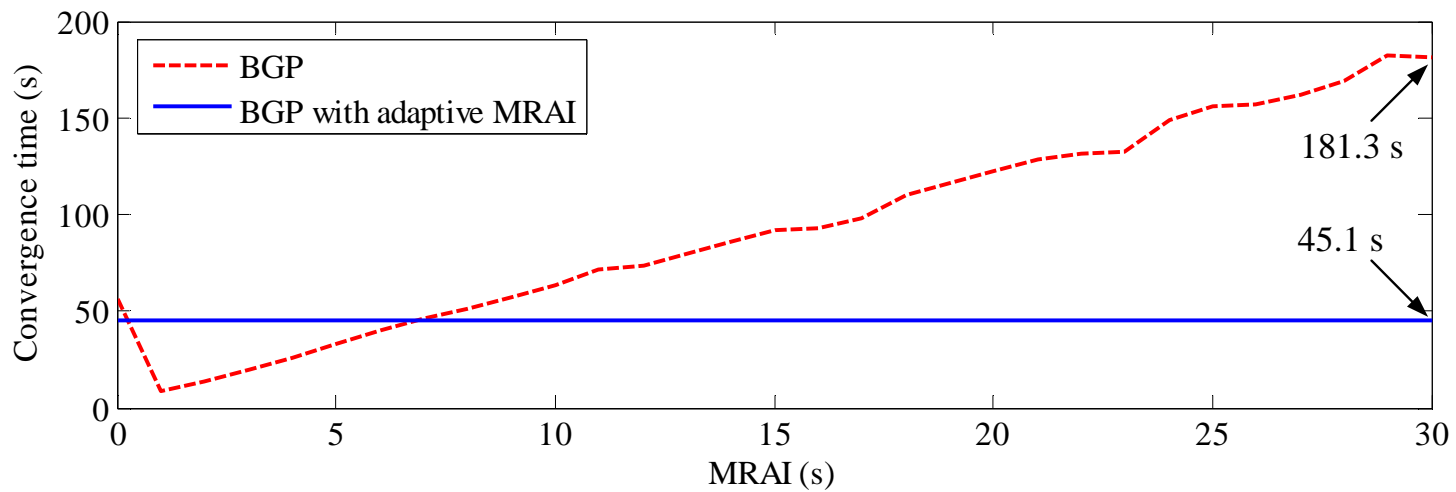


Time series of update messages



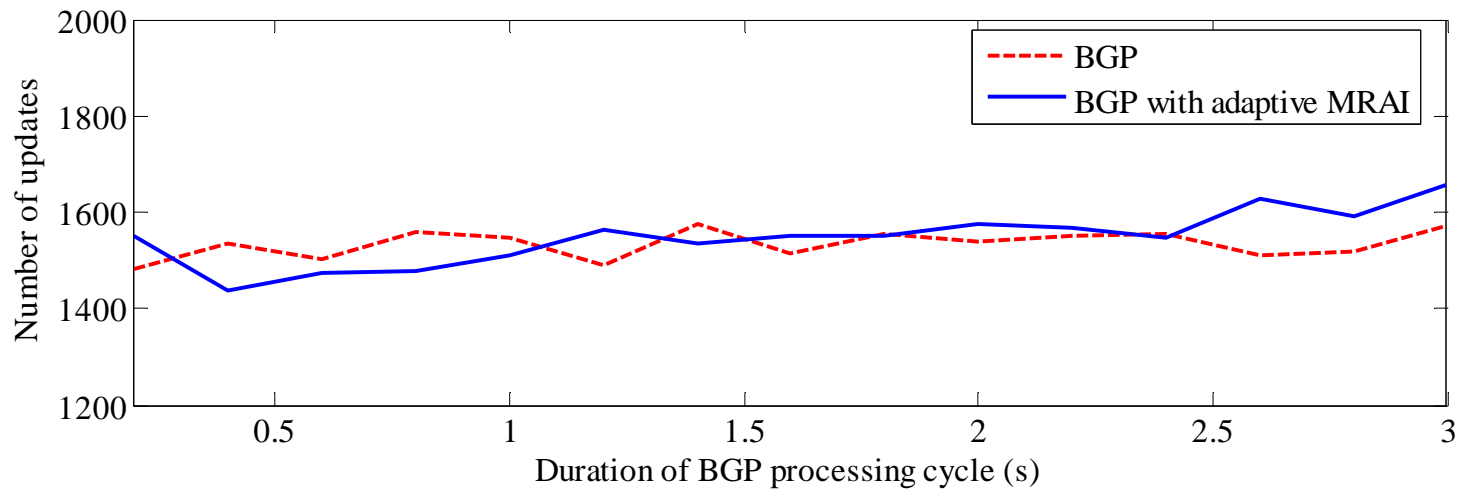
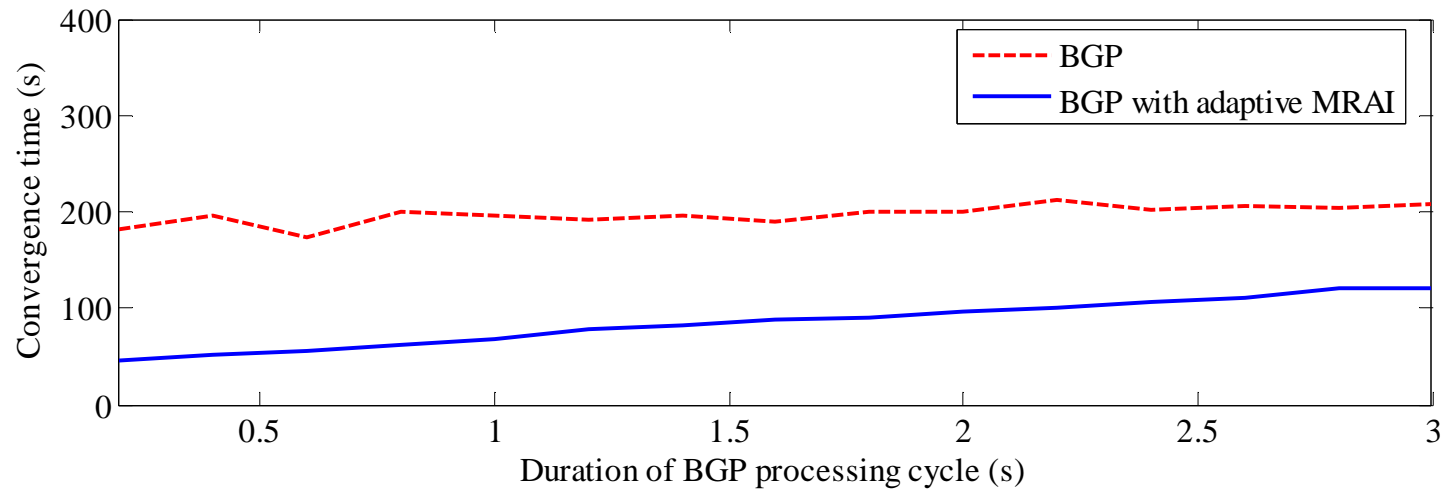


Completely connected graph



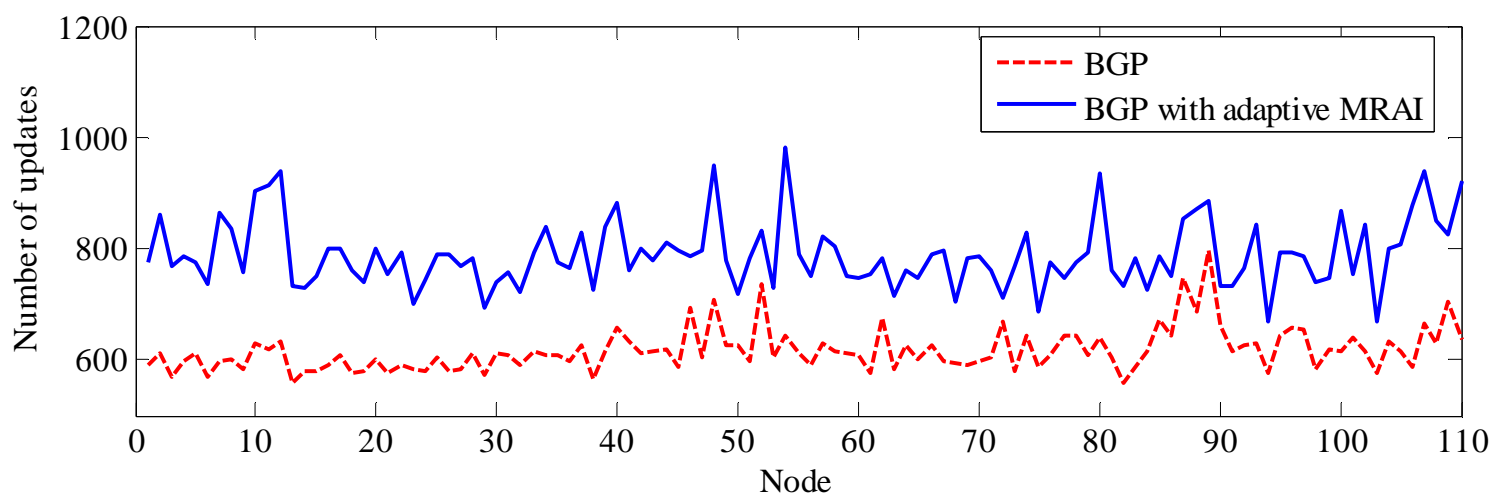
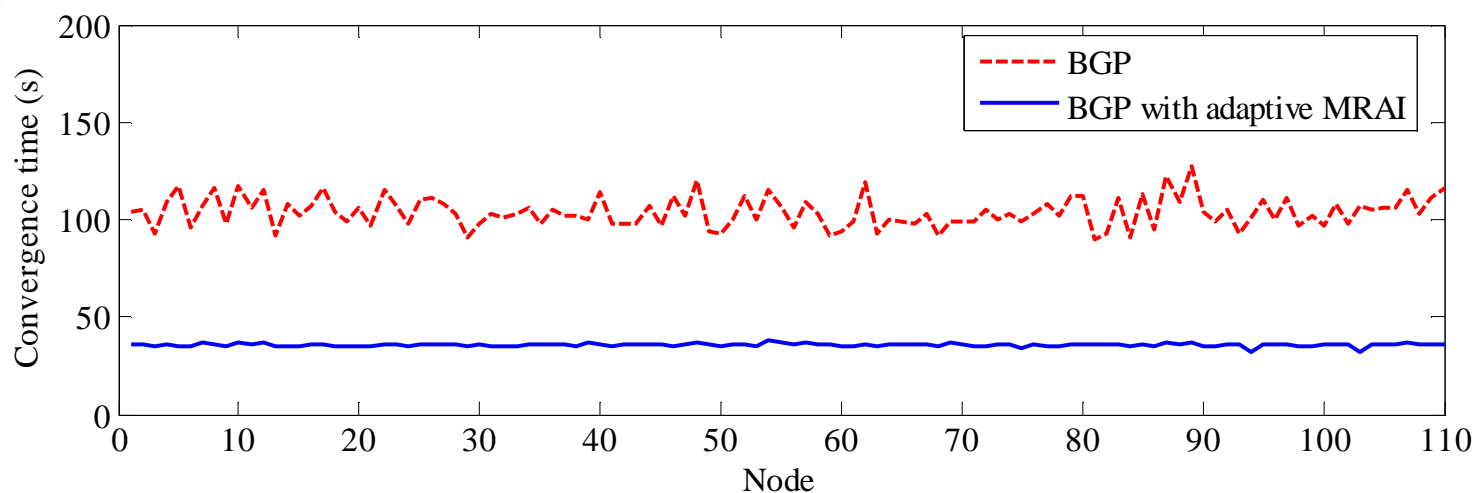


BGP processing cycles



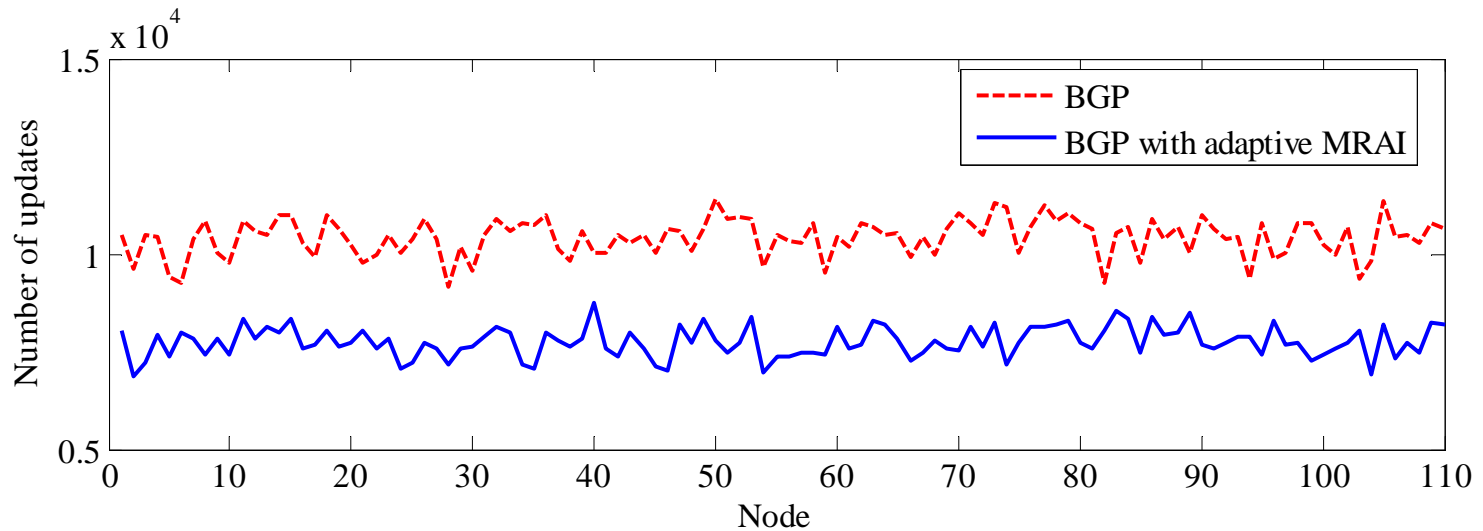
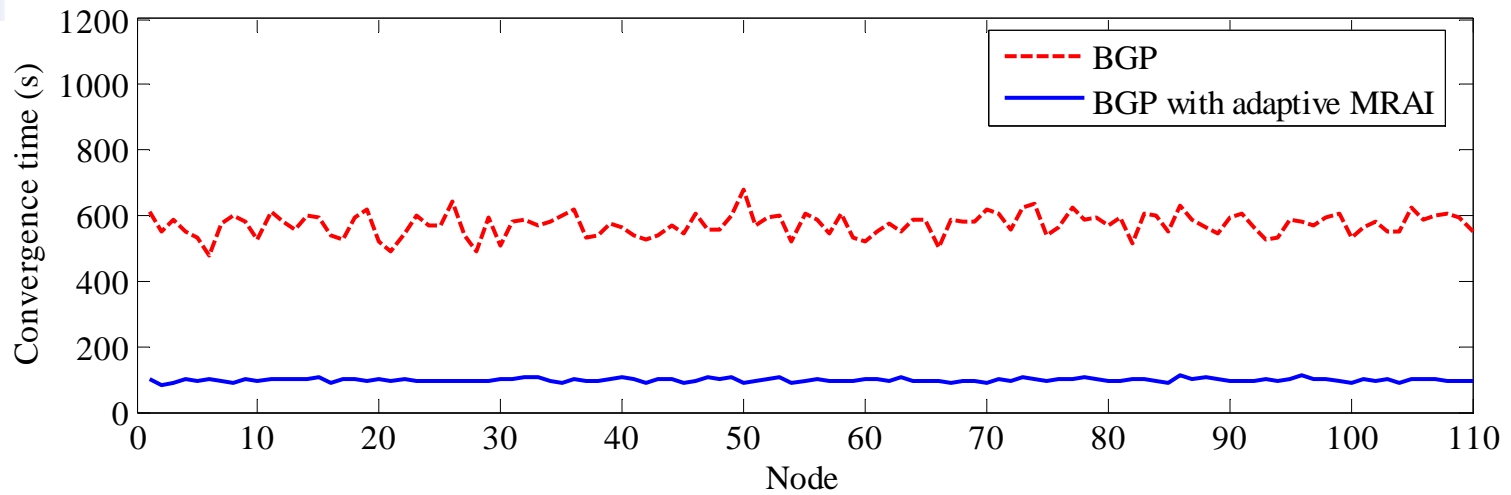


Network with 110 nodes: up phase





Network with 110 nodes: down phase





Road map

- Introduction and motivation
- Border Gateway Protocol (BGP)
- BGP with adaptive MRAI
 - empirical model for BGP processing delay
 - reusable MRAI timers
 - the adaptive MRAI algorithm
- Performance analysis of BGP with adaptive MRAI
- **Conclusions**



Conclusions

- We introduced the empirical delay model for the BGP processing delay
- We proposed **BGP with adaptive MRAI**
 - reusable MRAI timers
 - the adaptive MRAI algorithm
- BGP with adaptive MRAI results in shorter BGP convergence times for four simulated topologies
 - BGP convergence time is a linear function of the average BGP processing delay (traffic load)



Future work

- Further analysis requires additional measurement of the processing delay and the active time in various network settings
 - new value for the duration of the first MRAI round
- If fluctuations of the average active time are not too rapid the algorithm may be simplified
 - the duration of the adaptive MRAI does not have to be calculated in each round and for each destination separately



References

- Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)," *IETF RFC 1771*, Mar. 1995.
- T. Griffin and B. Premore, "An experimental analysis of BGP convergence time," in *Proc. ICNP*, Riverside, CA, 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.
- 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.
- C. Labovitz, A. Ahuja, R. Wattenhofer, and S. Venkatachary, "The impact of Internet policy and topology on delayed routing convergence," in *Proc. INFOCOM*, Anchorage, AK, Apr. 2001, pp. 537–546.
- C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," *IEEE/ACM Trans. Networking*, vol. 9, no. 3, June 2001, pp. 293–306.



References

- ns-2 (March, 2005) [Online]. Available: <http://www.isi.edu/nsnam/ns/>.
- T. D. Feng, R. Ballantyne, and Lj. Trajković, "Implementation of BGP in a network simulator," in *Proc. ATS*, Arlington, VA, Apr. 2004, pp. 149–154.
- J. Nykvist and L. Carr-Motyckova, "Simulating convergence prosperities of BGP," in *Proc. ICCCN*, Miami, FL, Oct. 2002, pp. 124–129.
- SSFNET (March, 2005) [Online]. Available: <http://www.ssfnet.org/>.
- Multi-AS topologies from BGP routing tables (March, 2005) [Online]. Available: <http://www.ssfnet.org/Exchange/gallery/asgraph/index.html>.
- The University of Oregon Route Views Project (March, 2005) [Online]. Available: <http://www.routeviews.org/>.
- Z. M. Mao, R. Bush, T. G. Griffin, and M. Roughan, "BGP beacons," in *Proc. IMC*, Miami Beach, FL, Oct. 2003, pp. 1–14.
- A. L. Barábasi and R. Albert, "Emergence of scaling in random networks," *Science*, Oct. 1999, pp. 509–512.
- D. Magoni and J.J. Pansiot, "Evaluation of Internet topology generators by power law and distance indicators," in *Proc. ICON*, Mumbai, India, Aug. 2002 pp. 401–406.



QUESTIONS?



BGP routing table (RIB)

Prefix (destination)	Peer's AS	Peer's IP	AS path
3.0.0.0/8	1755	213.174.64.80	1755 701 80
3.0.0.0/8	3130	147.28.255.2	3130 7018 701 80
3.0.0.0/8	3130	147.28.255.1	3130 7018 701 80
3.1.0.0/8	701	64.200.199.3	701 80
3.1.0.0/8	715	157.22.9.7	715 1239 80
3.1.0.0/8	3561	208.172.146.2	3561 1239 80
3.1.0.0/8	6539	216.18.31.102	6539 701 80
3.1.0.0/8	8121	199.74.221.1	8121 6461 701 80



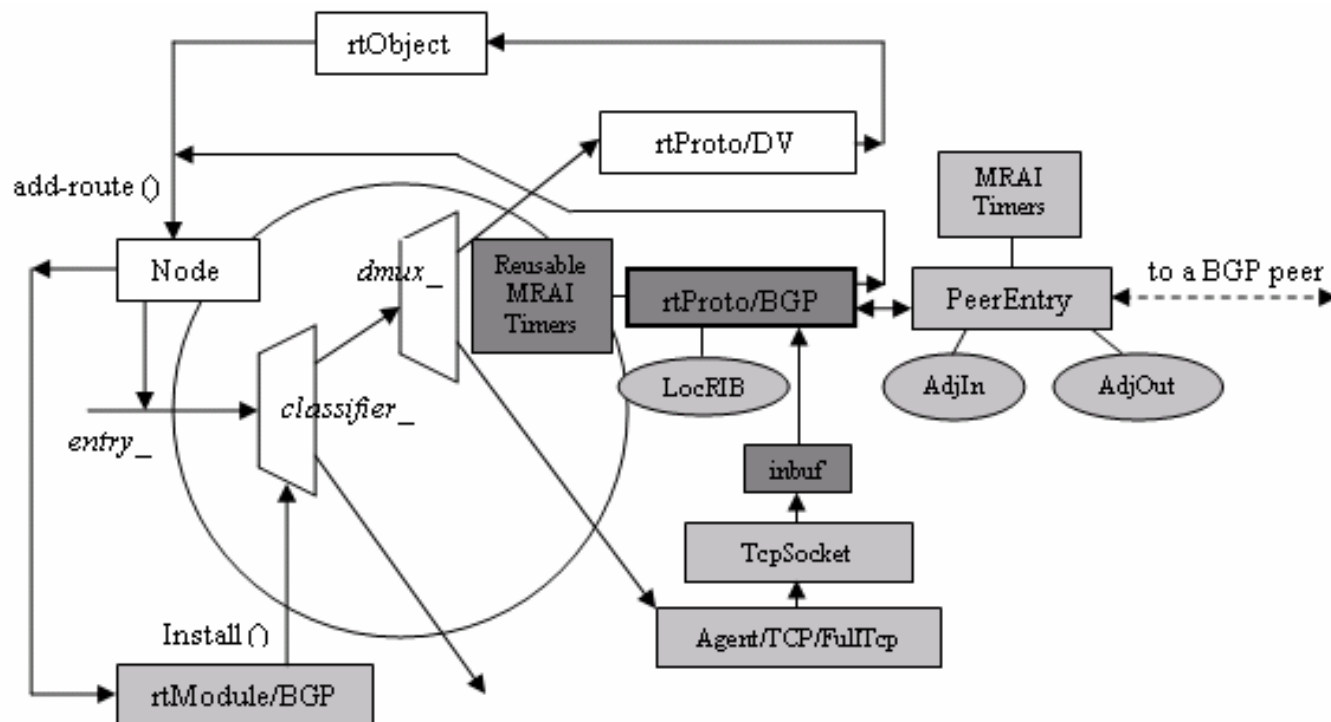
Granularity of reusable MRAI timers

Reusable timers		BGP convergence time (sec)	Number of updates
Number of timers	Granularity (sec)		
10	3	56.2	1651.7
15	2	54.9	1659.8
30	1	45.1	1552.9
60	0.5	45.8	1489.0
120	0.25	45.7	1520.0



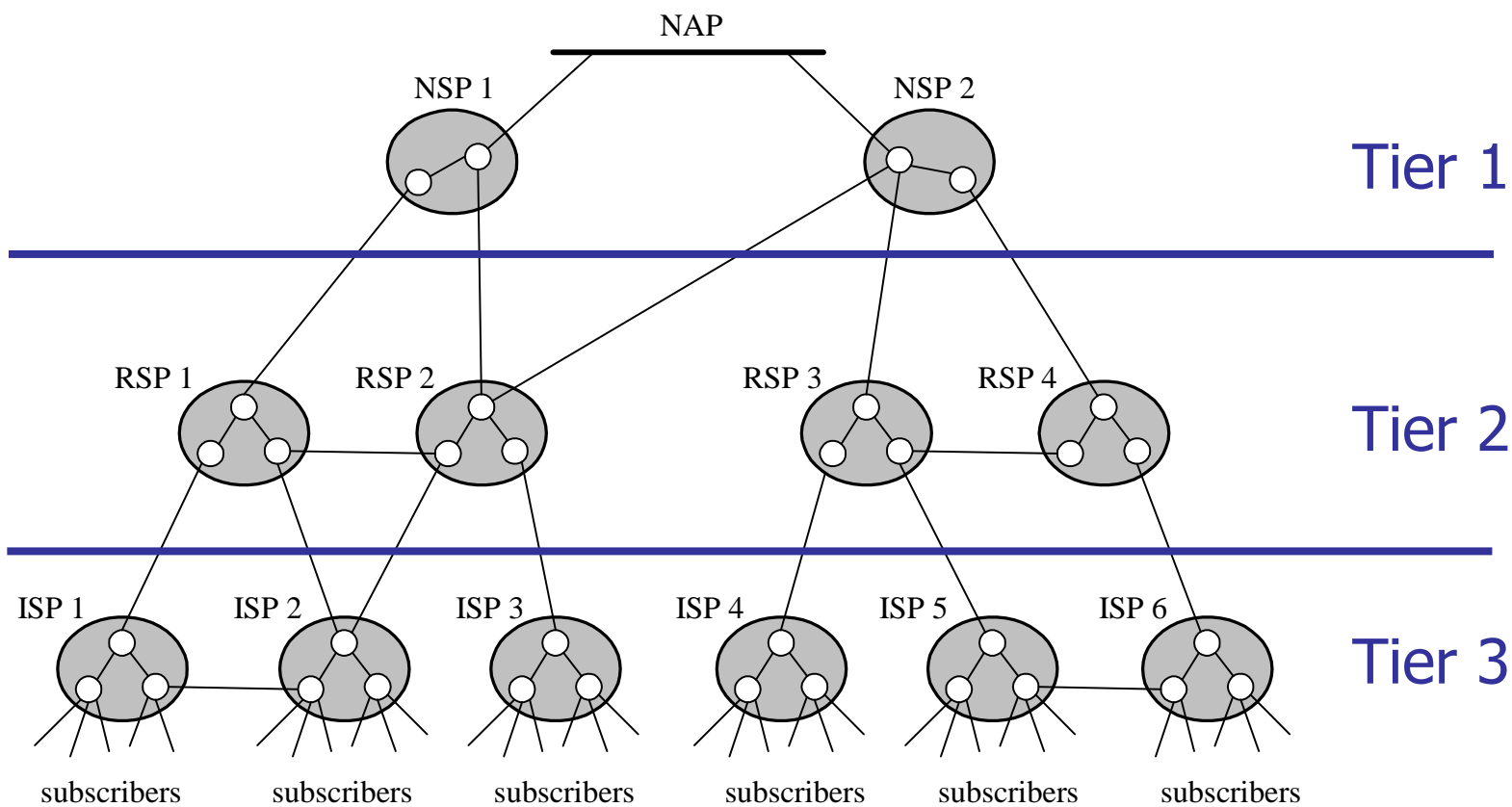
ns-2 implementation

- We used ns-2 network simulator and its BGP module **ns-BGP 2.0**





AS hierarchy

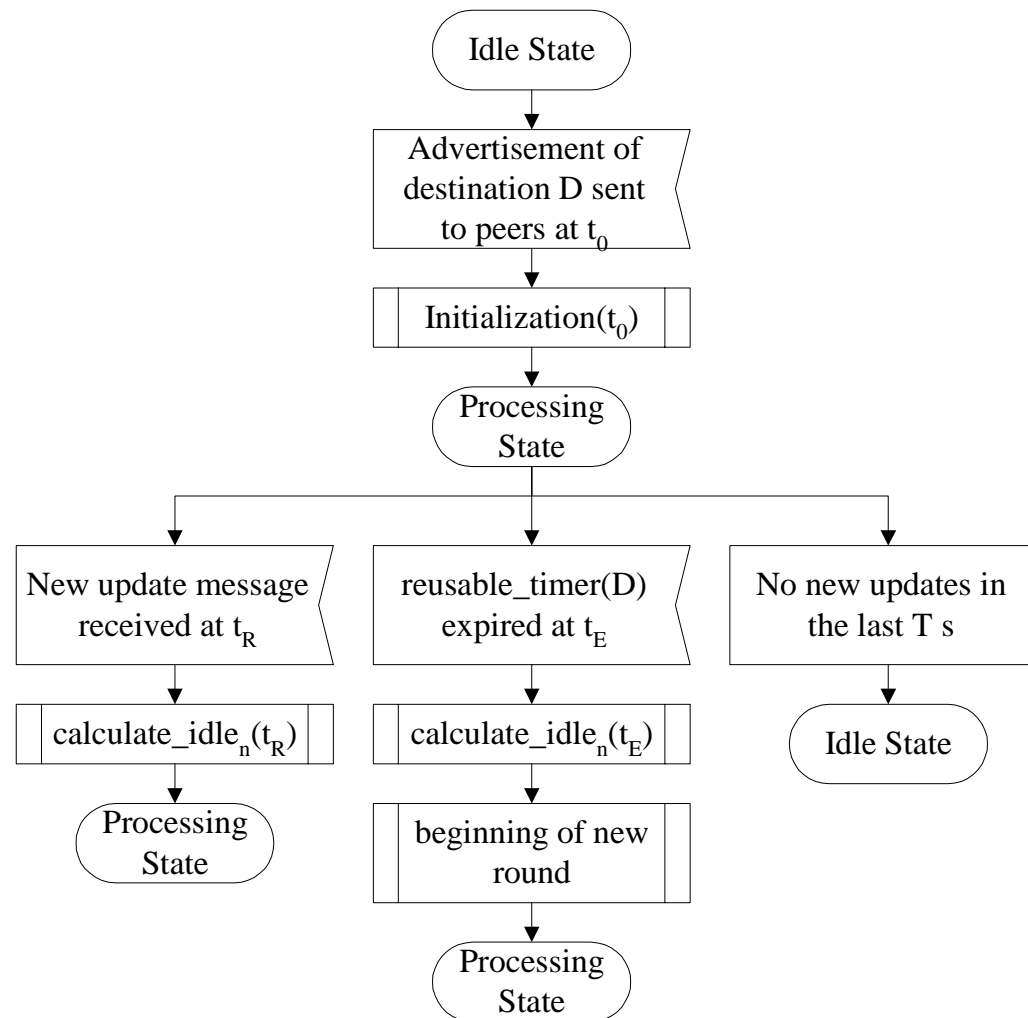


ISP: Internet Service Provider
RSP: Regional Service Provider

NSP: Network Service Provider
NAP: Network Access Point

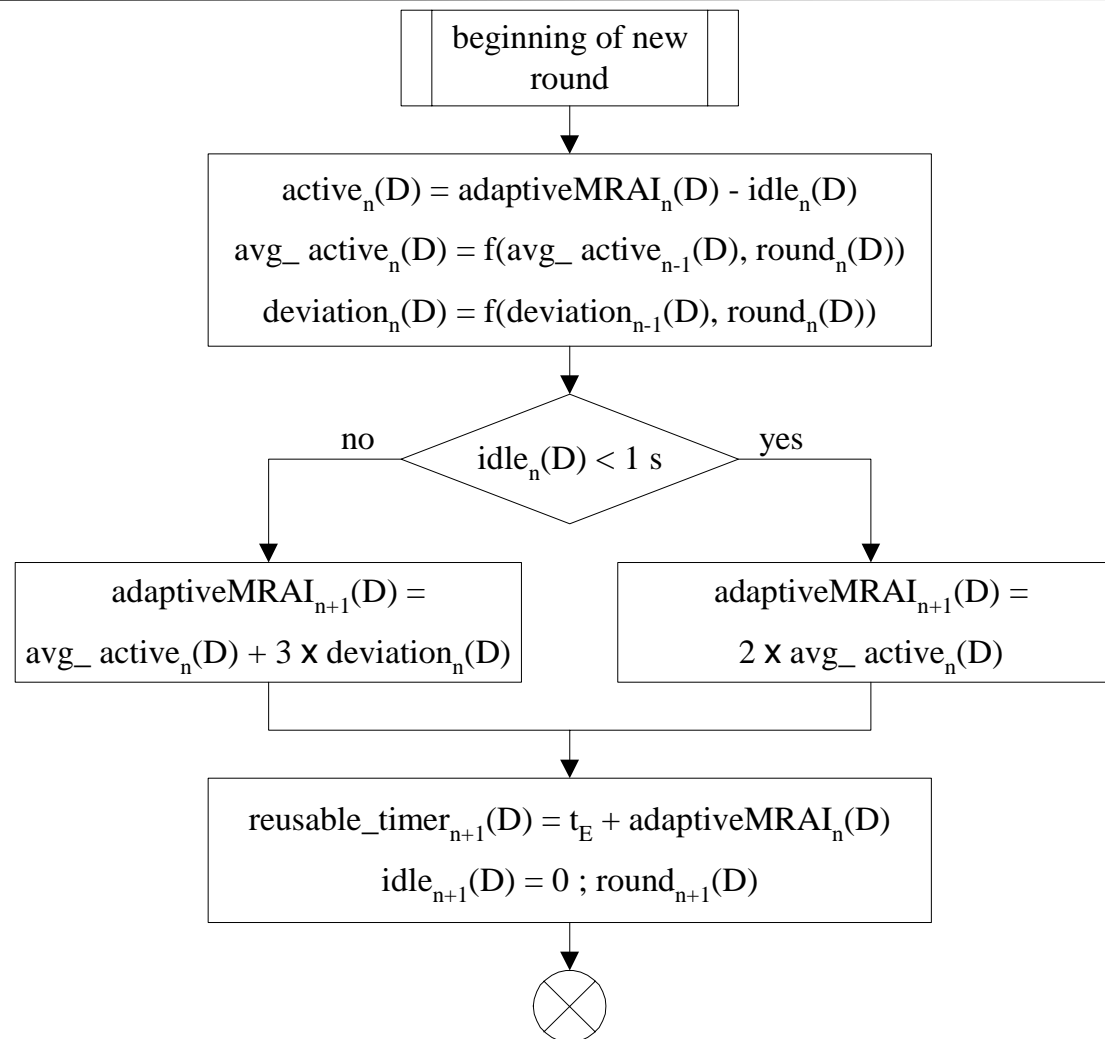


Adaptive MRAI algorithm



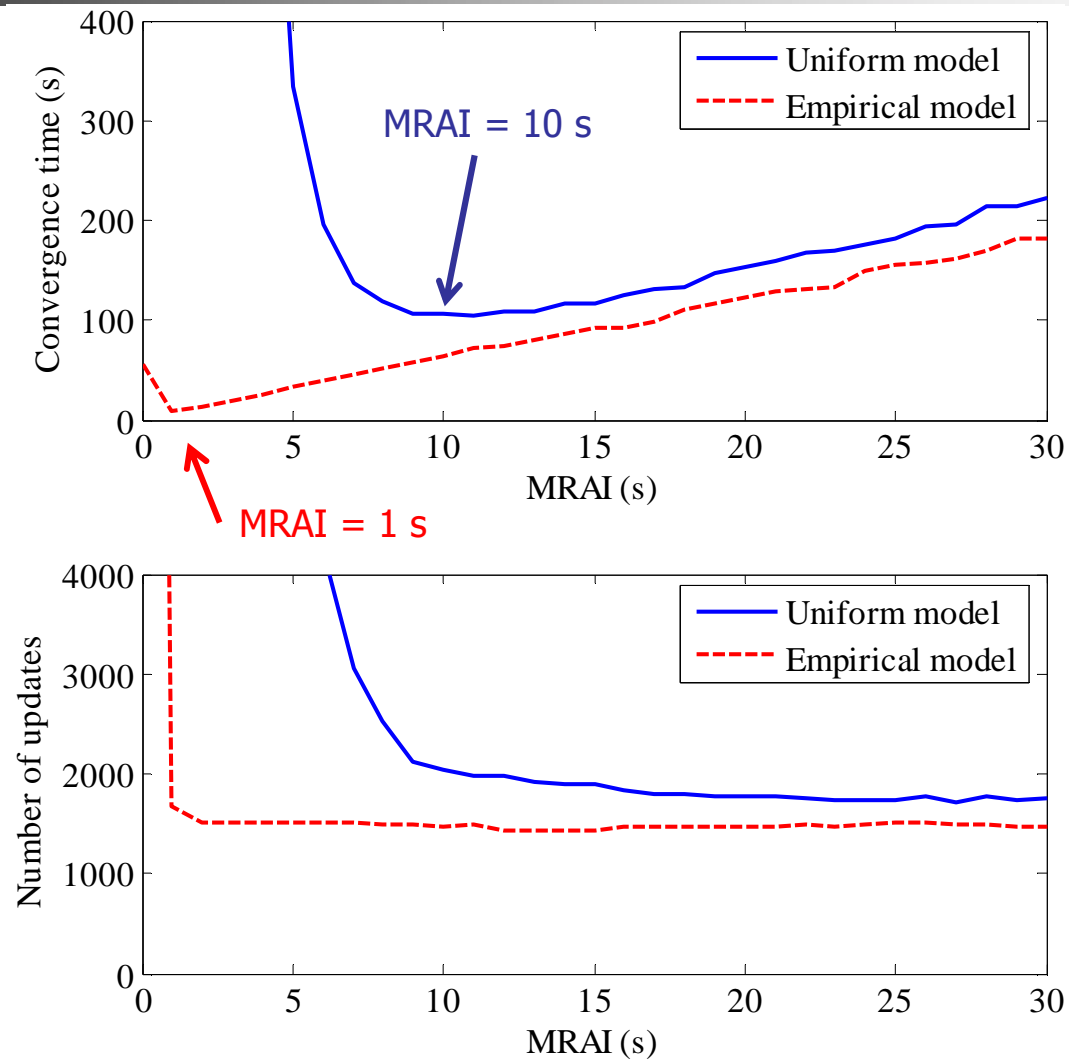


Adaptive MRAI algorithm





Empirical vs. uniform model





The average active time: calculation

$$\begin{aligned} \text{avg_active}_n(D) &= \sum_i^n \frac{\text{active}_i(D)}{n} \\ &= \frac{1}{n} \left(\sum_i^{n-1} \text{active}_i(D) + \text{active}_n(D) \right) \\ &= \frac{n-1}{n} \sum_i^{n-1} \frac{\text{active}_i(D)}{n-1} + \frac{1}{n} \text{active}_n(D) \\ &= \left(1 - \frac{1}{n}\right) \text{avg_active}_{n-1}(D) + \frac{1}{n} \text{active}_n(D) \\ &= \text{avg_active}_{n-1}(D) + \frac{(\text{active}_n(D) - \text{avg_active}_{n-1}(D))}{n} \\ &= \text{avg_active}_{n-1}(D) + \frac{\Delta_n}{n} \end{aligned}$$

The standard deviation of the active time



$$\begin{aligned} \text{deviation}_n(D) &= \sqrt{\sum_i \frac{n}{i} (\text{active}_i(D) - \text{avg_active}_n(D))^2} \\ &= \sqrt{\frac{1}{n} \left(\sum_i^{n-1} (\text{active}_i(D) - \text{avg_active}_n(D))^2 + (\text{active}_n(D) - \text{avg_active}_n(D))^2 \right)} \\ &= \sqrt{\frac{n-1}{n} \sum_i^{n-1} \frac{(\text{active}_i(D) - \text{avg_active}_n(D))^2}{n-1} + \frac{\Delta_n^2}{n}} \\ &= \sqrt{\left(1 - \frac{1}{n}\right) \text{deviation}_{n-1}^2(D) + \frac{\Delta_n^2}{n}} \\ &= \sqrt{\text{deviation}_{n-1}^2(D) + \frac{(\Delta_n^2 - \text{deviation}_{n-1}^2(D))}{n}} \end{aligned}$$