Internet-Scale BGP Simulation

Maciek Wojciechowski†, Benno Overeinder†,
Guillaume Pierre*, and Maarten van Steen*

† NLnet Labs

* Vrije Universiteit Amsterdam

Outline

- BGP analysis: observation versus simulation
- Approach to large-scale/Internet-scale simulation of BGP protocol & AS network
- Protocol and network abstraction level
- Model validation and simulation verification
- Request for comments and suggestions…

Introduction

Building BGP and AS models for understanding:

- Real-world observations
  - observation ➔ hypothesis/analysis ➔ conclusions ➔ validation ➔ …
- Modeling and simulation
  - abstraction ➔ simulation/experiment ➔ hypothesis/analysis ➔ conclusions ➔ …
    - verification ➔ …
    - validation ➔ …

Pros and Cons

- Observation/analysis
  - ground truth ✓
  - analysis of now and past
- Simulation/analysis
  - abstraction ➔ away from reality ✗
  - analysis of what-if questions ✓
- Cannot proof model is correct… but one can show model is wrong (falsifiable)

What-If...

- What if topology changes?
- What if AS number/prefix number grows?
- What is the impact of different policies on protocol behavior?
- How do certain changes to its operation (e.g., route flap damping) affect convergence?
- Better understanding of protocol behavior

First Step

MODELING

Abstract From Reality

- Goal: large-scale/Internet-scale BGP simulation
  - 250k prefixes
  - 27k ASes
- Identify important and unimportant characteristics and properties
  - BGP protocol
  - AS network
- Choices are driven by experimental framework
  - which question to answer
  - which experiment to run

Modeling Choices

- Modeling of BGP protocol
  - explicit prefix tables
  - standard BGP announcement/withdrawal messages
  - FIB updates & propagation according to policy
  - freedom in individual BGP speaker behavior
- Modeling of AS network
  - no physical network modeling, only connectivity
  - 1 AS ➔ 1 node
  - simple iBGP convergence modeling

Second Step

SIMULATION

Run the Model

- Interested in dynamics, time-evolution behavior of protocol
  - time progress is explicit in simulation
- Time-driven, event-driven, or scaled-time simulation
  - time-driven: non-option for discrete event systems
  - event-driven: exact but expensive simulation methodology ➔ scalability issue
  - scaled-time: emulation… [next slide please]

Emulation

- Scaled-time
  - intuitively seeming best fitting technique to this class of problems
  - run the simulation real-time
  - apply scaling factor for real-time progress
  - less precise, but “good enough”?!?
- Requirements for emulation
  - experience same rate of progress in time
  - network not modeled ➔ network component should be constant and relative small (negligible) to BGP operation (orders of 10 sec.)

Implementation and Execution

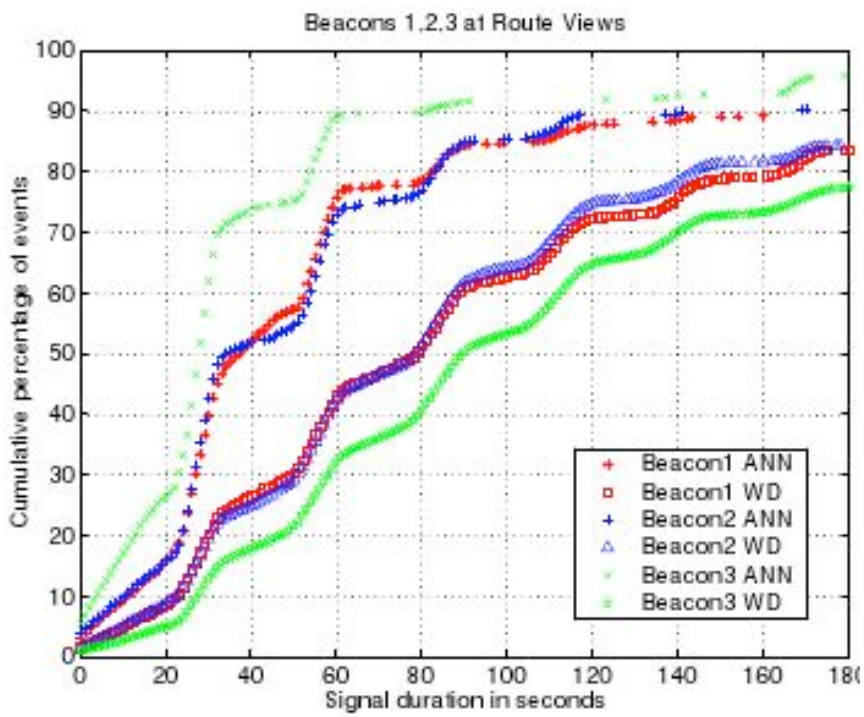
- Implemented in Java
- AS network adjacencies from CAIDA
- Runs on homogeneous cluster
  - DAS-3 VU cluster
    - 85 dual-node, dual-core cluster
    - 4 GB memory per node
    - 10 Gb/s Myrinet network
- For simulations with large memory requirements
  - prefixes stored on hard drives

Third Step

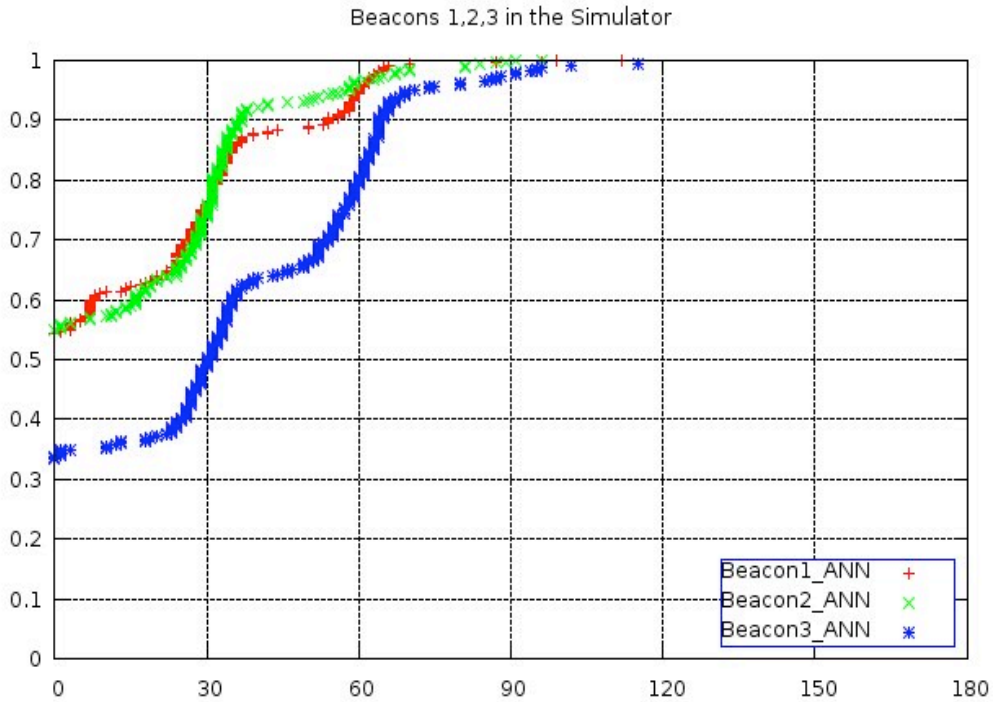
EXPERIMENTS

Validation: Is the Model Correct?

- Compare real-world observations with simulation results
- Need to identify root cause events in AS network
  - what is the observed effect
  - what is the simulated effect
- BGP beacons
  - Z. Morley Mao, Randy Bush, and others
  - attempt to reproduce results from paper (August 2002–April 2003)
  - analysis of current BGP beacon observation and simulation



(a)

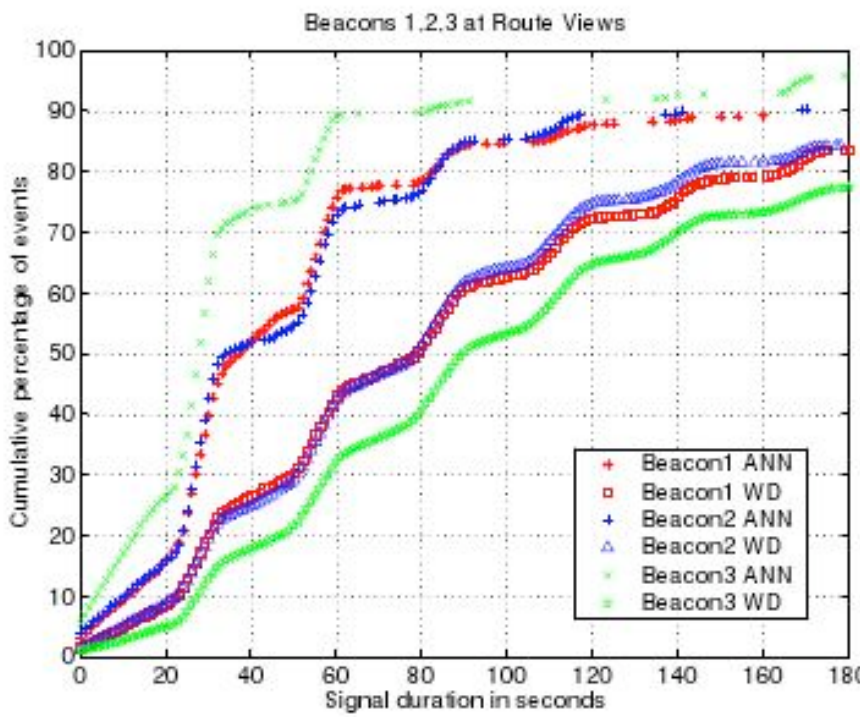


(b)

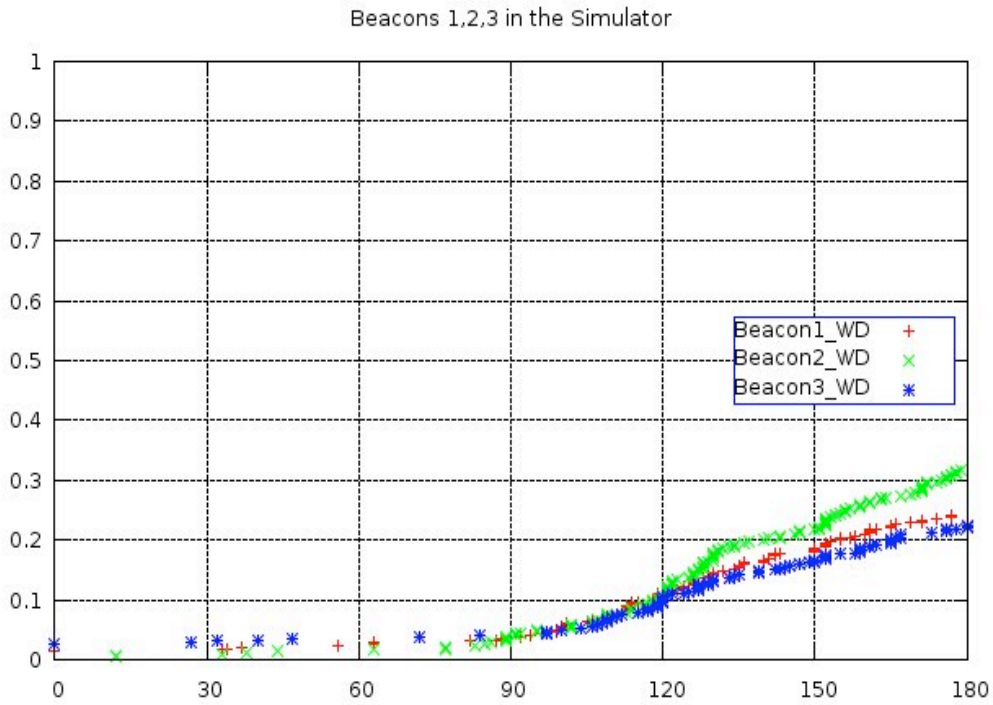
Cumulative distribution signal duration

Cumulative distribution of the signal duration for all three beacons from the BPG beacons reference study (a) and from the simulation (b) (data from January 2004).

30 seconds MRAI timers are apparent in both figures. Large difference in CDF, especially the percentage of 0 sec. signal duration.



(a)

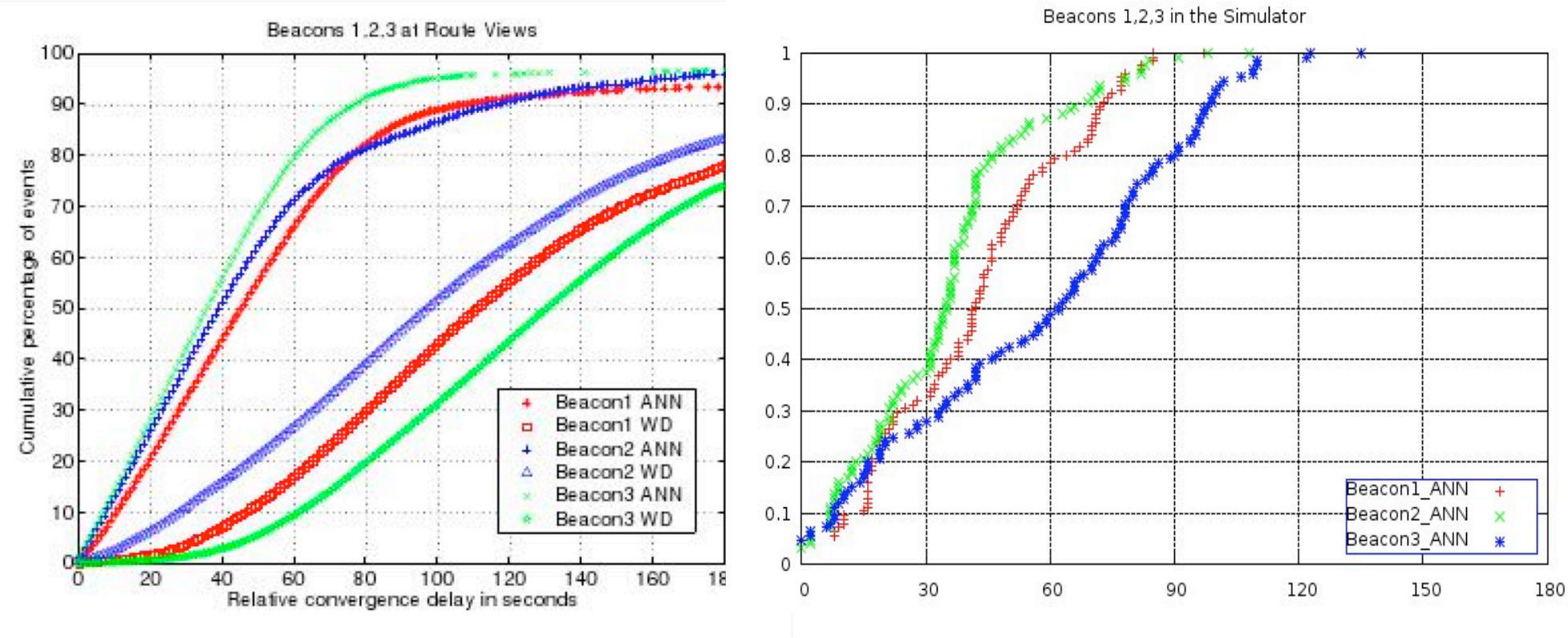


(b)

Cumulative distribution signal duration

Cumulative distribution of the signal duration for all three beacons from the BPG beacons reference study (a) and from the simulation (b) (data from January 2004).

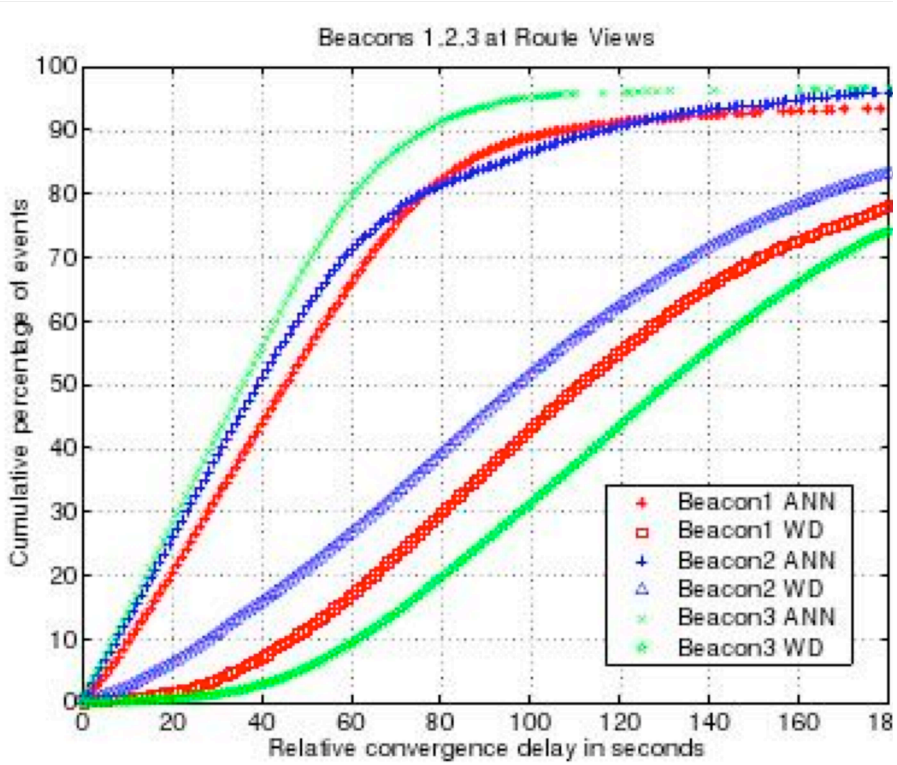
30 seconds MRAI timers are apparent in both figures. Large difference in CDF, especially the percentage of 0 sec. signal duration.



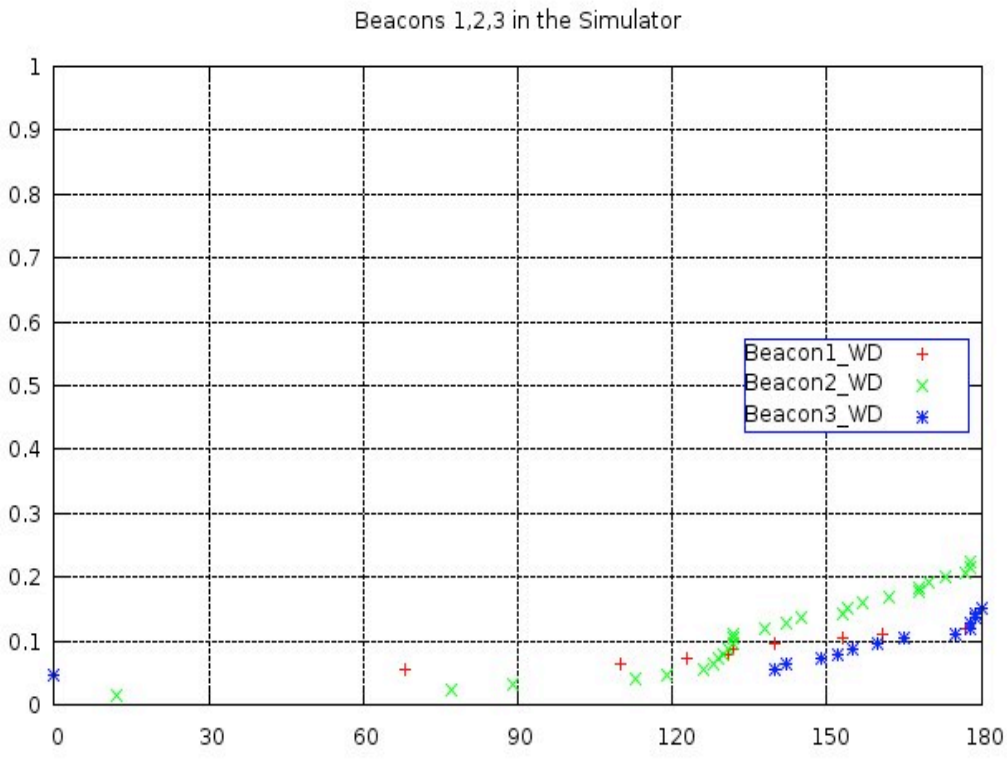
(a) (b)

Cumulative distribution of signal convergence

Cumulative distribution of relative convergence for all three beacons from the reference study (a) and from the simulation (b) (data from January 2004).



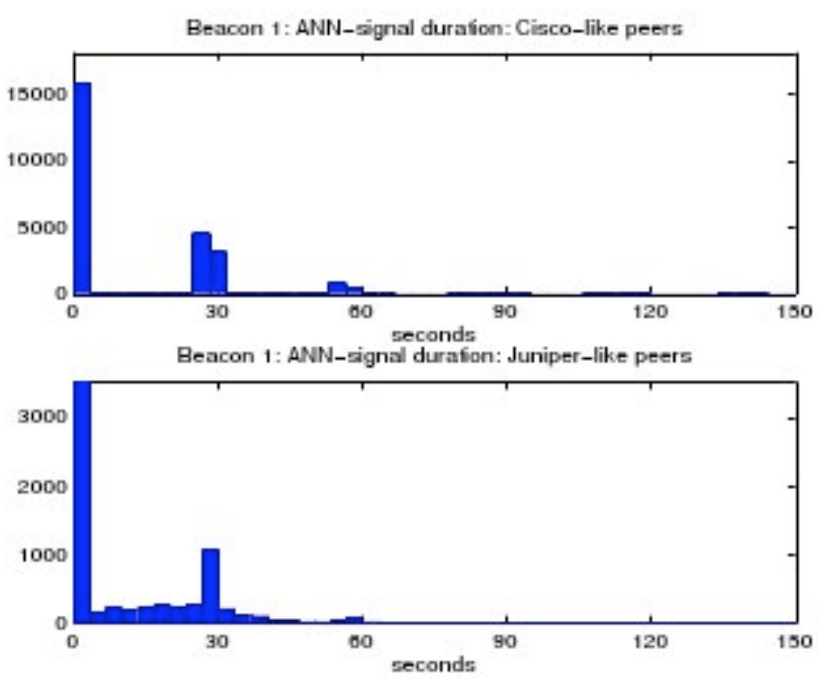
(a)



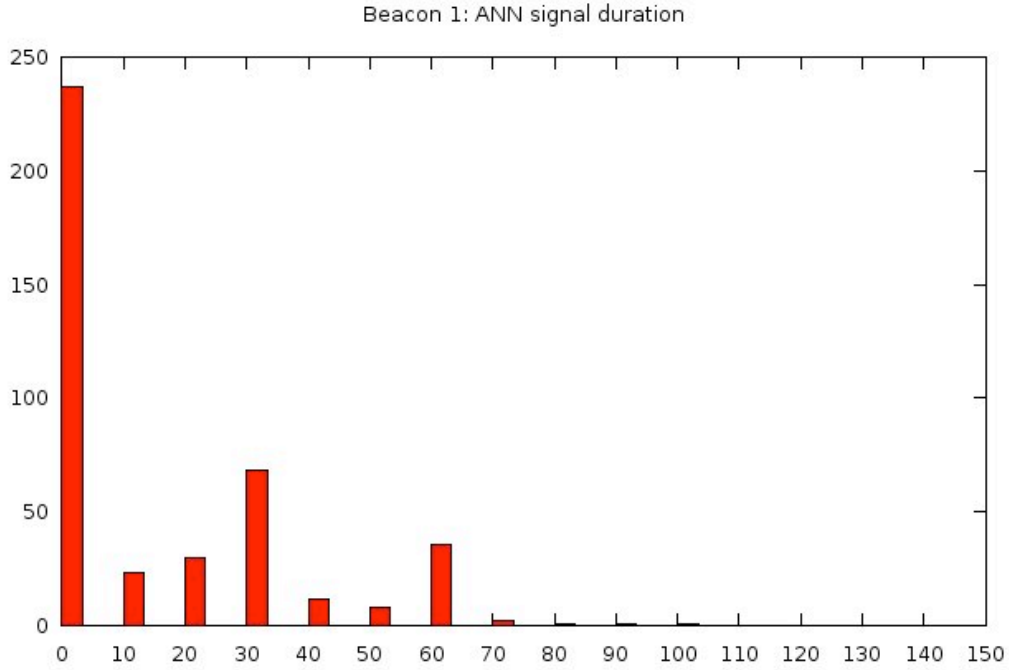
(b)

Cumulative distribution of signal convergence

Cumulative distribution of relative convergence for all three beacons from the reference study (a) and from the simulation (b) (data from January 2004).



(a)

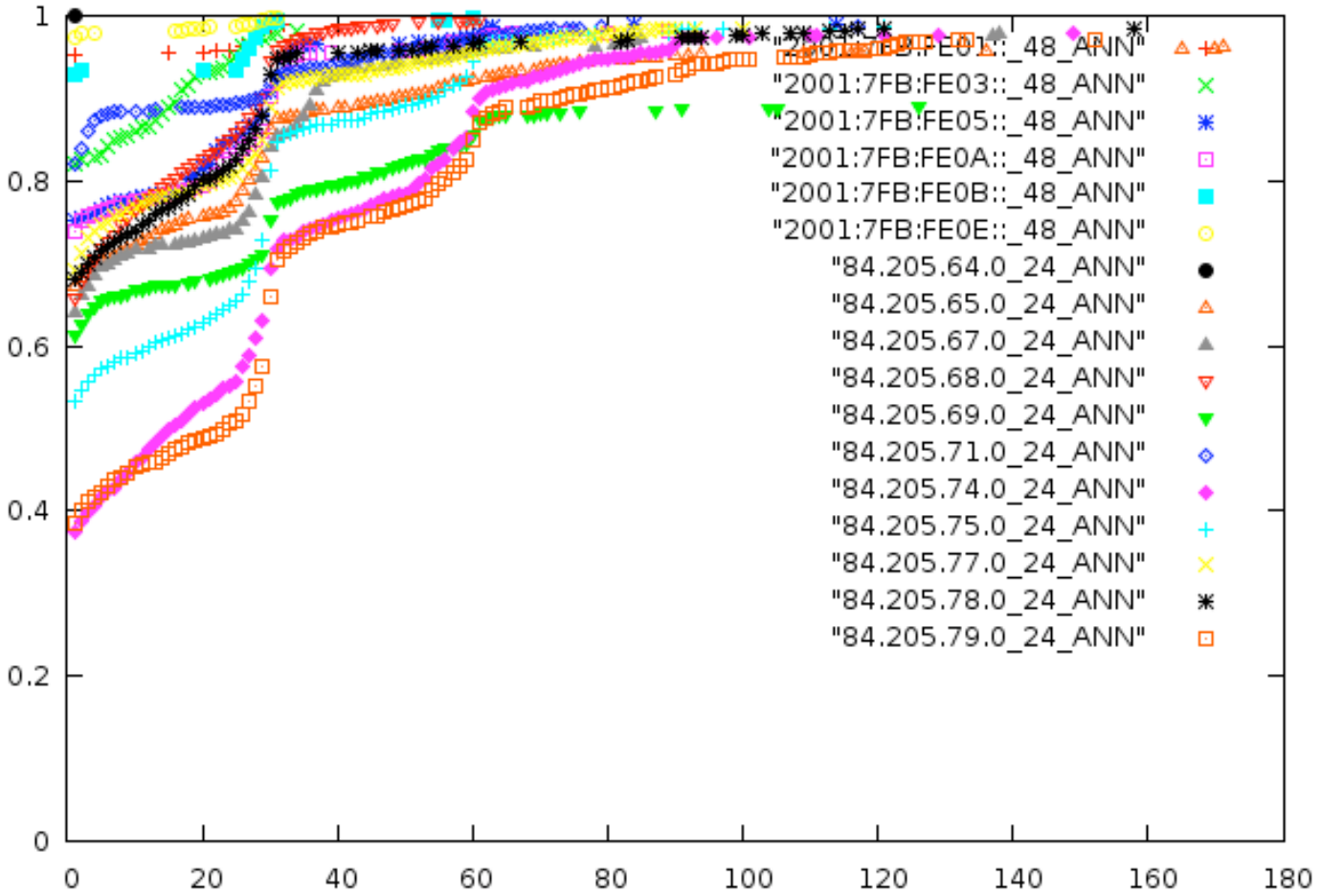


(b)

Head-to-Head Comparison

Beacon 1's announcement signal duration distribution for Cisco-like and Juniper-like peers as observed by Mao et al. (a) and as observed from the simulation (b).

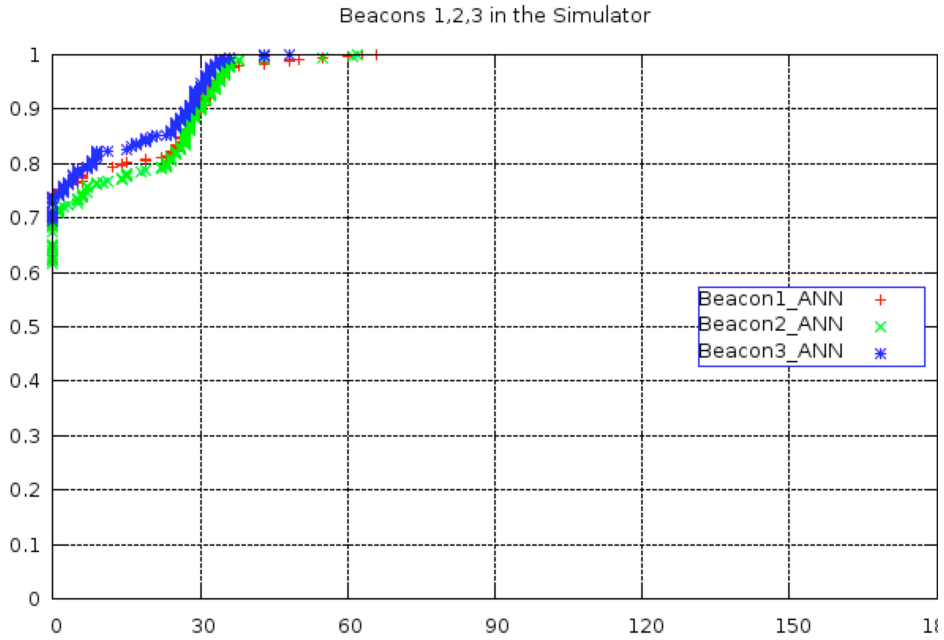
The histograms in figure (a) do not match with the CDF plot for signal duration time in figure (a) of slide 15.

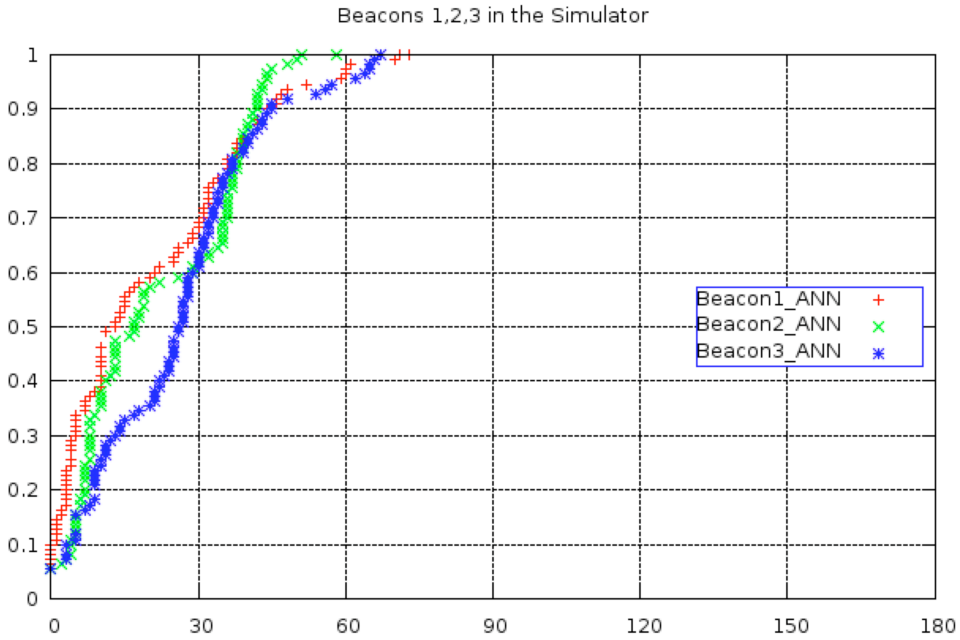


RIS beacons monitored at DECIX

Observation of RIS beacons (different prefixes) at RIS monitor at DECIX (February 2008). Other RIS monitors seem to behave similarly.

Preliminary results, but show that the simulator is in fact more accurate than initially thought.





What-if analysis: Cumulative distribution of duration and convergence time

Cumulative distribution of signal duration and relative convergence time for announcement signals for all three beacons for the network from January 2008 as observed by simulation.

Same experiments as slides 15 and 17, but for newer AS network. Monitors, beacons and so on are the same (no RIS here).

Request for Comments

- Need more and other data and experiments to validate the model and simulation
- Ideas for what-if scenarios are welcome
- Contact us: {maciek,benno}@nlnetlabs.nl!

Example Simulation Environments

- BGP++
  - BGP daemon on top of ns-2 simulator
  - (almost) 1-to-1 relation with real BGP operation
  - difficult to scale to 27k ASs
- C-BGP
  - policy evaluation
  - steady-state analysis, no dynamics
  - scalable to large number of ASs, for limited set of prefixes

Some Runtime Figures

- About 500 ms to fully propagate a new prefix
- About 2000 messages per minute
- Up to 15000 prefixes per minute
- 100 GB of compressed storage of 250k prefixes