Measurement-based Inter-domain Traffic Engineering

W. Shao, J.L. Rougier, L. Iannone
Telecom ParisTech (not an operator)
member of University of Paris-Saclay

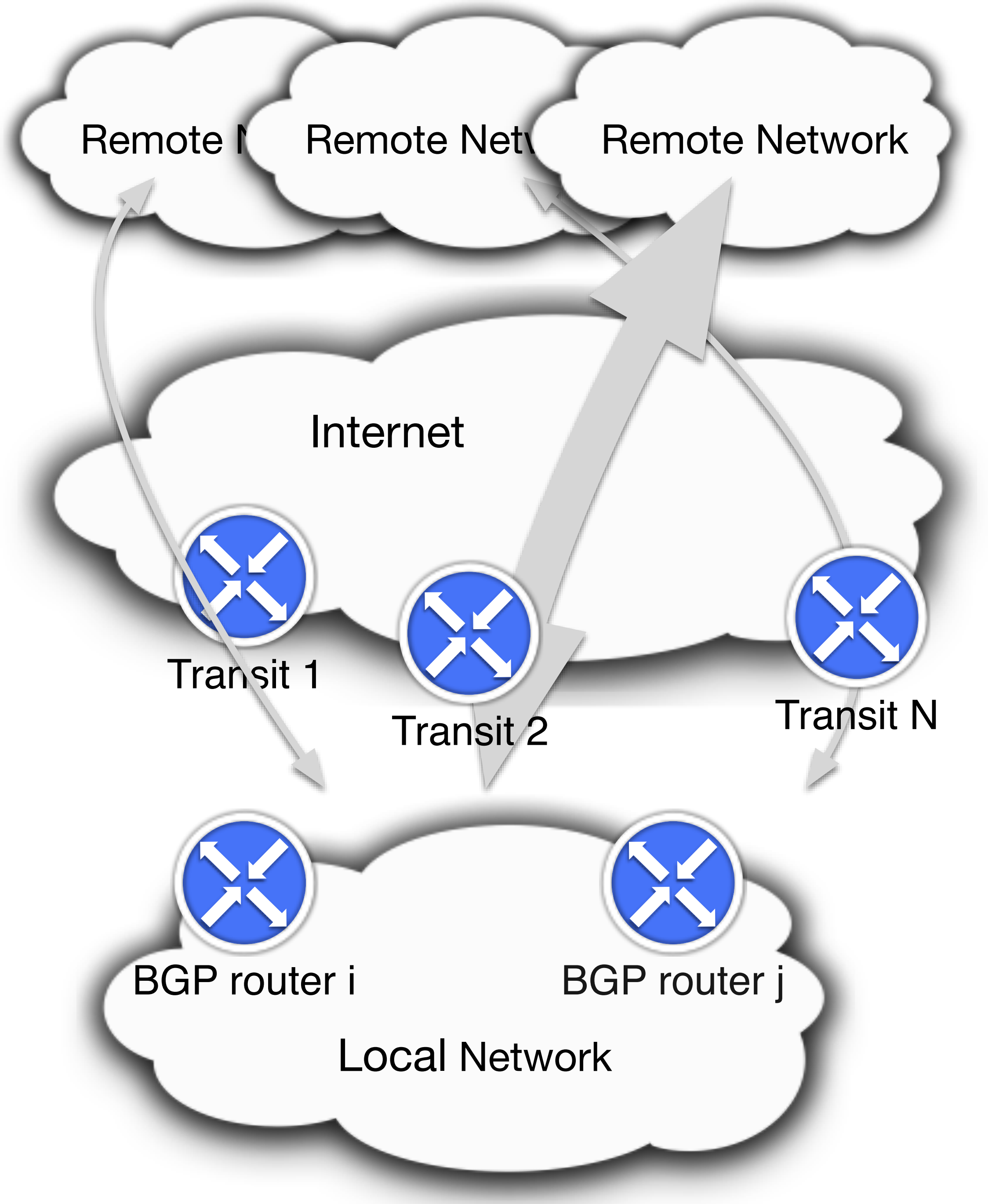
F. Devienne, M. Viste
Border 6

NANOG 67
June 13-15, 2016

Outline

- **Measurement-based** inter-domain traffic engineering
- A problem of scalability

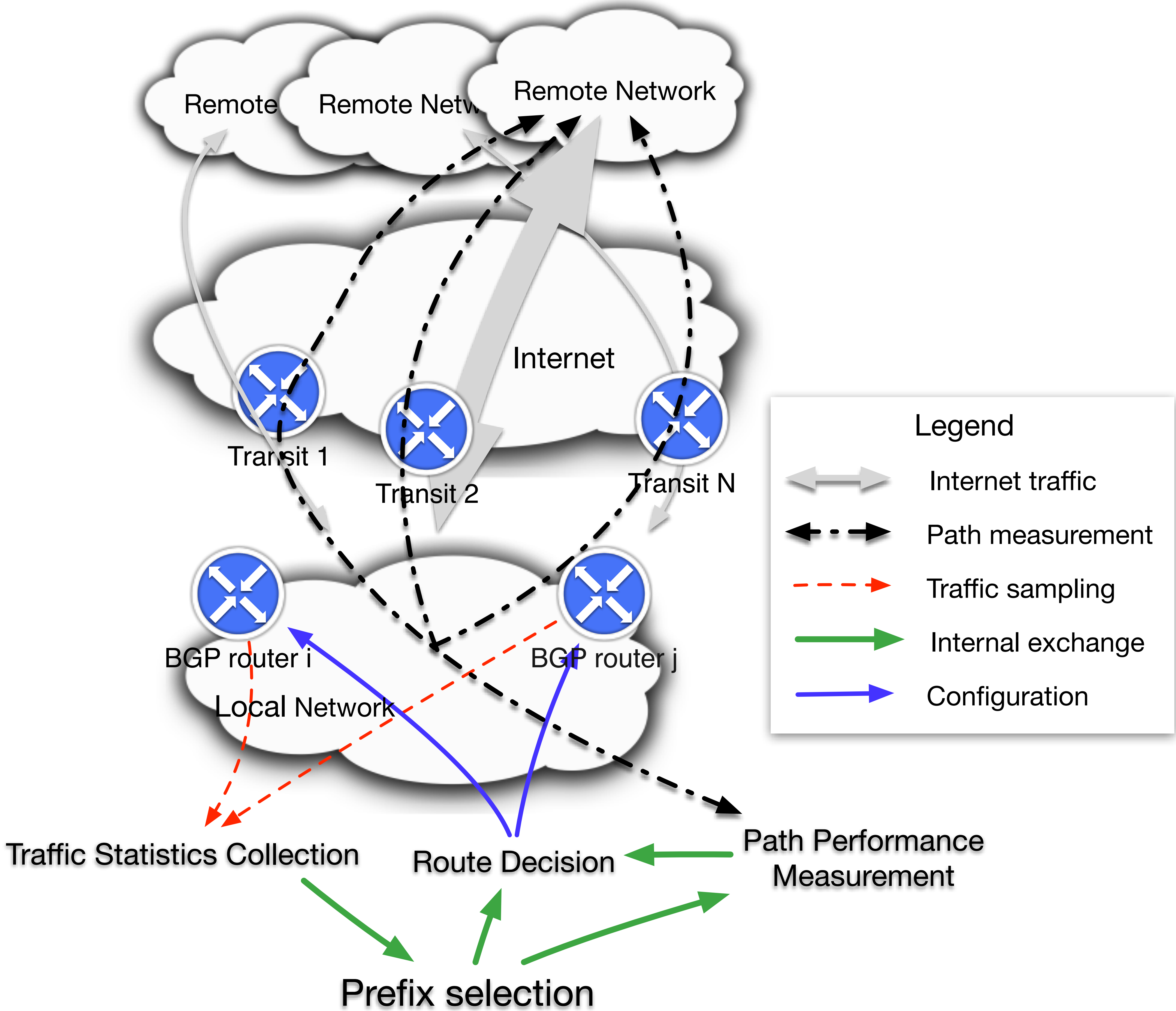
A scenario: Out-bound TE for multi-homed stub AS



1. Local_Pref
2. AS_Path
3. MED
4. eBGP > iBGP
5. tie-breaking
    1. IGP costs
    2. oldest path
    3. etc

An old scenario

- N. Feamster, D. G. Andersen, H. Balakrishnan, and M. F. Kaashoek, "**Measuring the effects of internet path faults on reactive routing**," ACM SIGMETRICS, vol. 31, no. 1, p. 126, **2003**.
- A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "**A measurement-based analysis of multihoming**," SIGCOMM, **2003**.
- D. K. Goldenberg, L. Qiu, H. Xie, Y. R. Yang, and Y. Zhang, "**Optimizing cost and performance for multihoming**," CCR, vol. 34, no. 4, p. 79, Oct. **2004**.
- A. Akella, B. Maggs, S. Seshan, and A. Shaikh, "**On the Performance Benefits of Multihoming Route Control**," IEEE/ACM Trans. Netw., vol. 16, no. 1, pp. 91–104, Feb. **2008**.


Remote
Remote Netw
Remote Network
Internet
Transit 1
Transit 2
Transit N
BGP router i
BGP router j
Local Network
Traffic Statistics Collection
Route Decision
Path Performance
Measurement
Prefix selection
Legend
Internet traffic
Path measurement
Traffic sampling
Internal exchange
Configuration

Traffic statistics collection

Purpose: to select a set of ‘managed prefix’, i.e. destinations of importance;

Collector: netflow/sflow collector, PMACCT;

Storage: RAM, PostgreSQL.

Active measurements

Target: probes discovered in ‘managed’ prefixes

Method: ping, TCP SYN -> RTT and loss;

Path: via **all available** BGP next-hops;

Steering: source-based routing, SDN, etc;

Storage: RAM, PostgreSQL.

Route decision

Objective: performance, availability & transit cost;

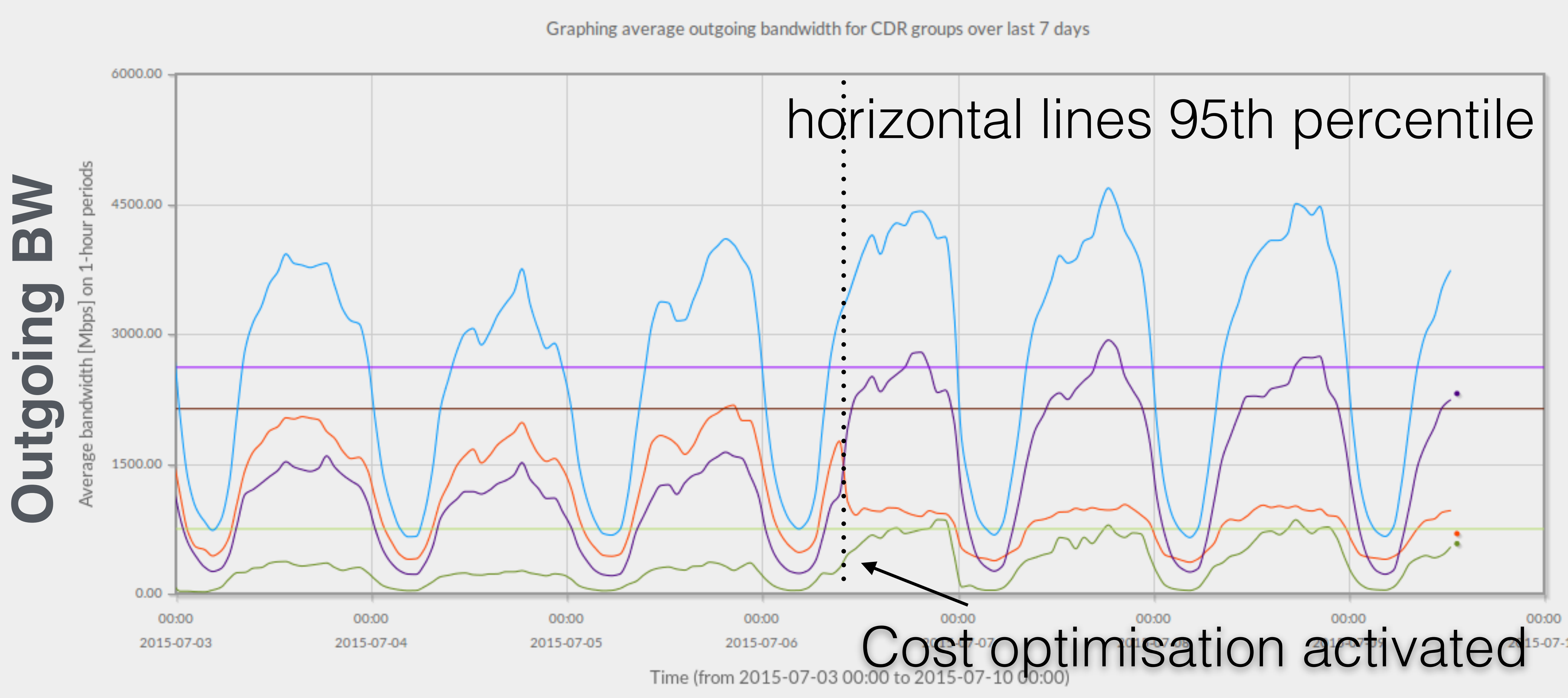
Metrics: RTT, loss and BW w.r.t. CDR;

Algorithm: depends on client needs;

Steering: BGP as SDN southbound interface.

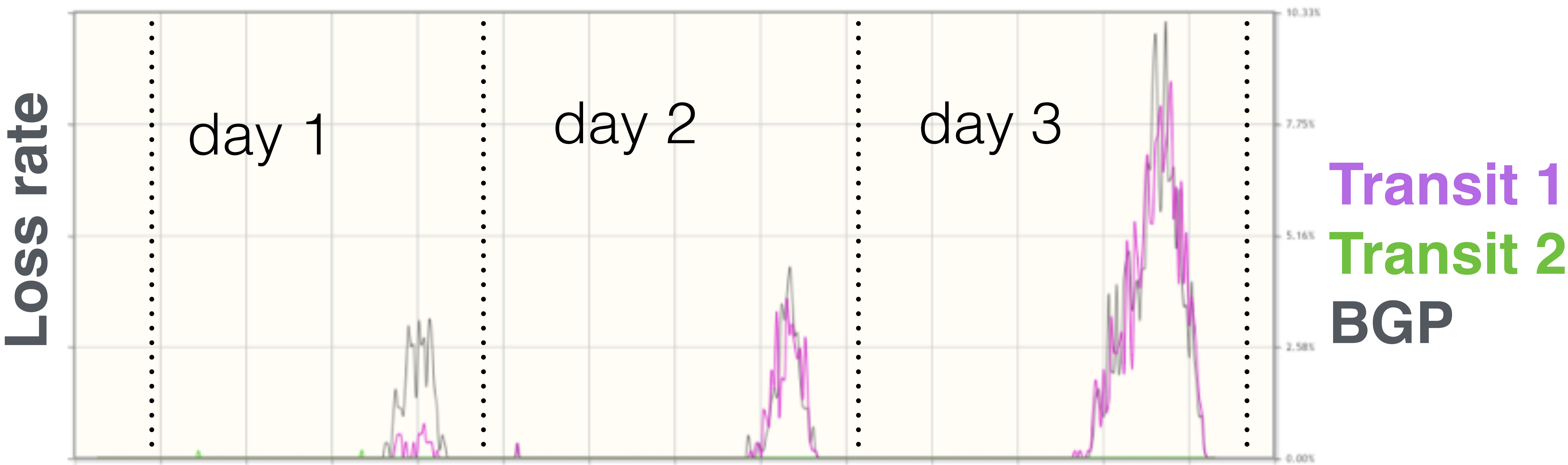
Algo example—Cost

Total Transit1 Transit2 Transit3



Transit1 has the lowest BW price.

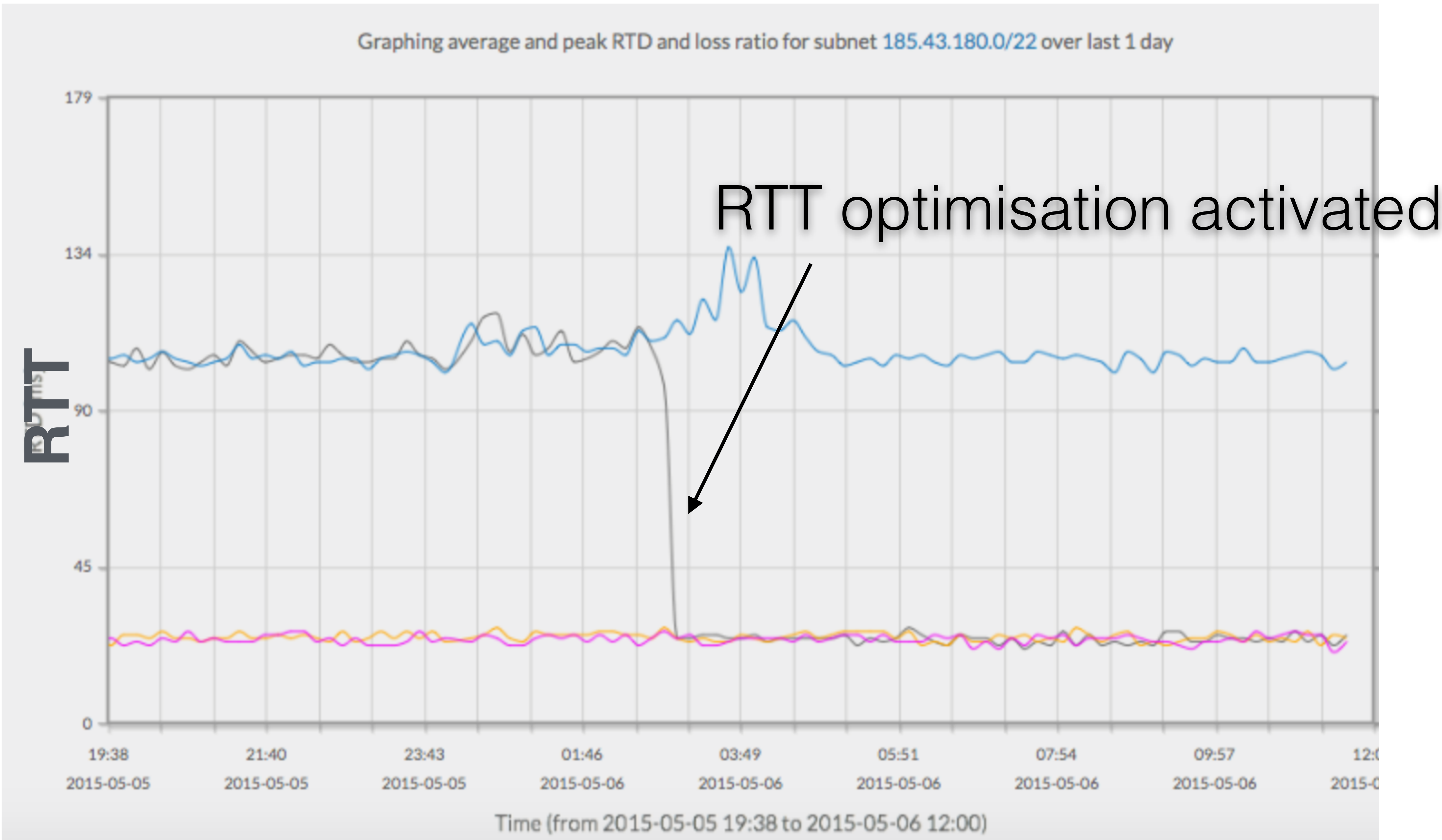
Algo example—Availability



Packet loss due to **consistent congestion**
can be avoided
by simply change a BGP next-hop.

Algo example—smaller RTT

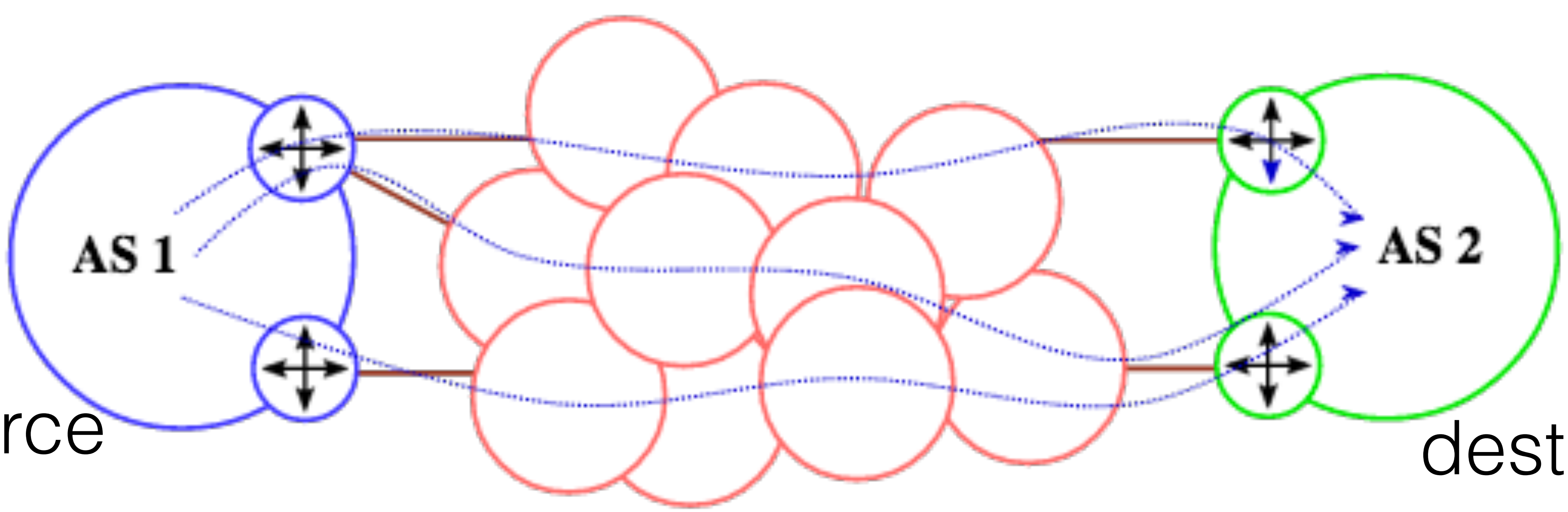
Transit 1 **Transit 2** **BGP**



Other functions

Route Preference Protocol (RPP) for **Inbound** TE

1. Tell traffic sourcing AS what is you favourite ingress point;
2. Traffic sourcing AS then does its best.



source

destination

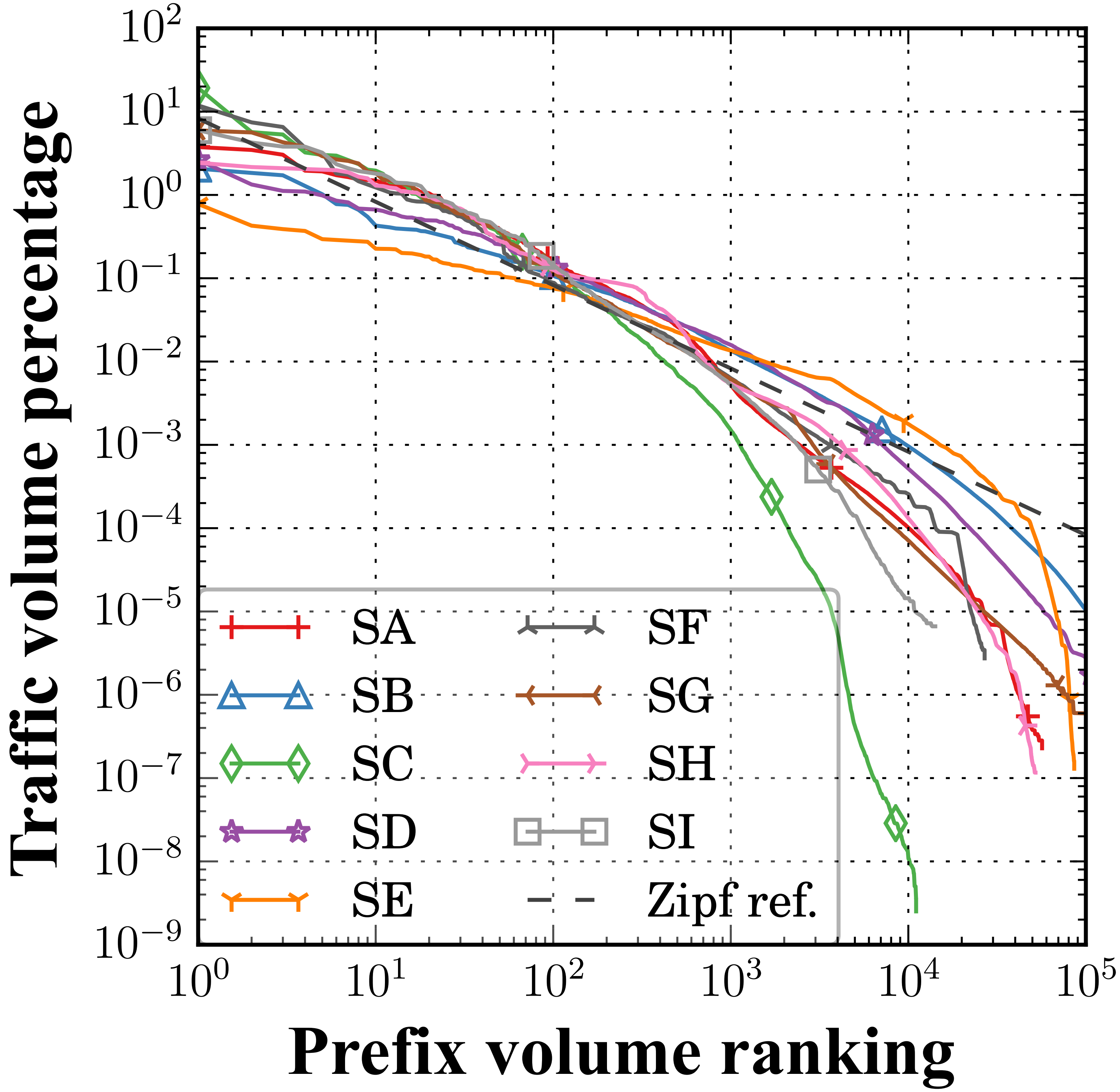
http://rpp.border6.com/
https://github.com/mateuszviste/rpp

the scalability problem
>500K BGP prefixes

How to **select prefix** of importance?

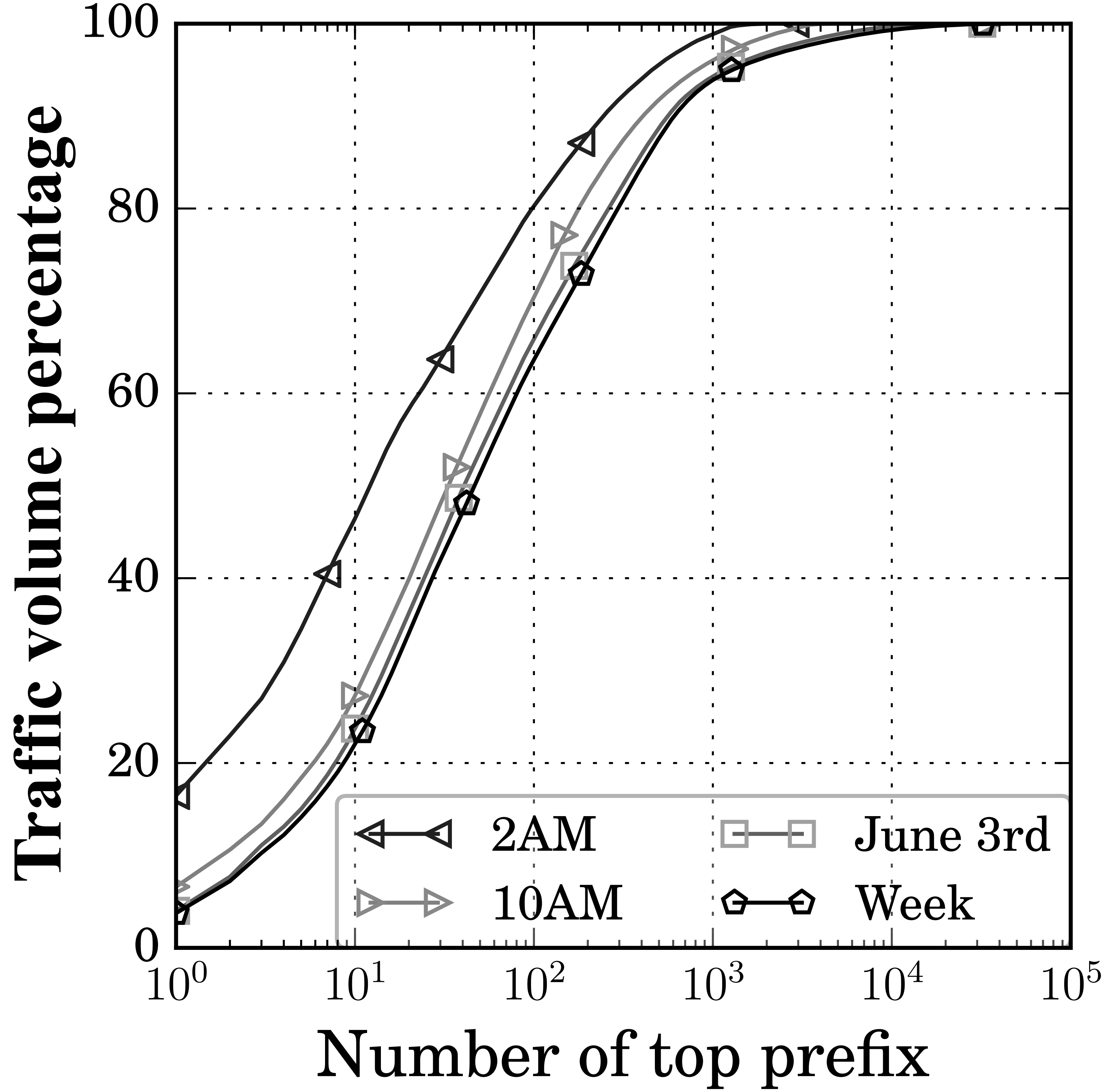
W. Shao, L. Ianonne, J.L. Rougier, F. Devienne, and M. Viste, "**Scalable BGP Prefix Selection for Effective Inter-domain Traffic Engineering**," IEEE/IFIP NOMS, 2016.

A large portion traffic is concentrated on a small number of prefixes (log-log graph not CDF)



However…

traffic value associated to each prefix evolves over time



However…

we have **some ten thousands** of time-series/prefixes to predict.

Predictively select BGP destination prefixes that stand for **a large portion** of traffic, with **simple and efficient** method.

FIB caching

- W. Zhang, J. Bi, J. Wu, and B. Zhang, "**Catching popular prefixes at AS border routers with a prediction based method**," Comput. Networks, vol. 56, no. 4, pp. 1486–1502, Mar. 2012.
  - **GM(1,1)** outperforms LFU, LRU

David Barroso, Spotify
Building an extensible ***SDN*** *Internet Router with commodity hardware*
https://youtu.be/o1njanXhQqM

Difference from FIB caching

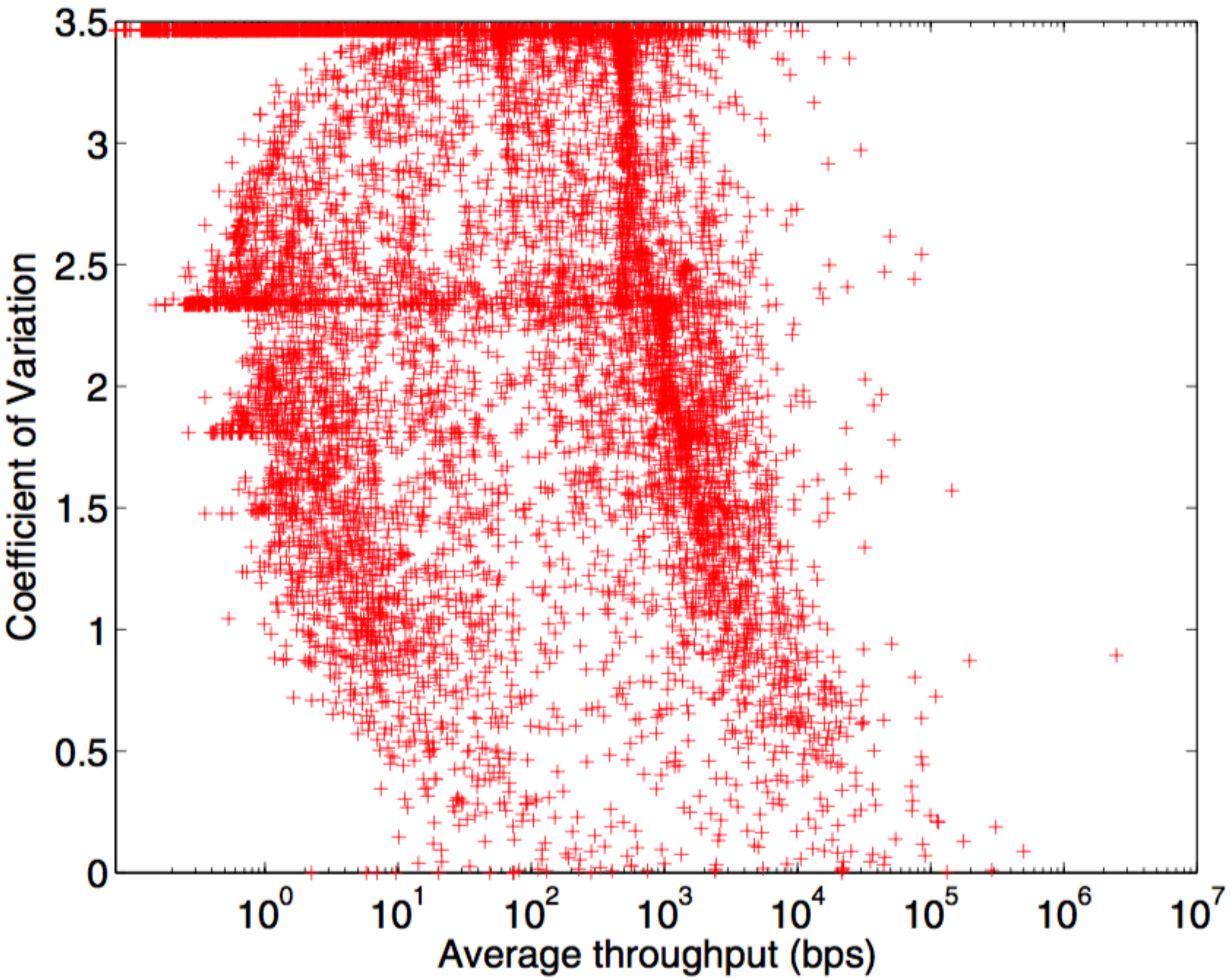
Basically, a matter of time scales

FIB caching, 5min interval or less, memory of hours;
Prefix selection, 1 hour interval, memory of days till weeks.

Why?
Not only data-plane is involved.
Probes discovery, probe selection, long term pattern and bursty-traffic, prefix churn, etc.

And more than that….

Traffic dynamism



K. Papagiannaki, N. Taft, and C. Diot, "**Impact of Flow Dynamics on Traffic Engineering Design Principles**," INFOCOM, 2004.

"… shows that there is no clear correlation between the mean and the coefficient of variation of the bandwidth of a network prefix flow."

What is case for traffic aggregated by BGP prefix over longer interval?

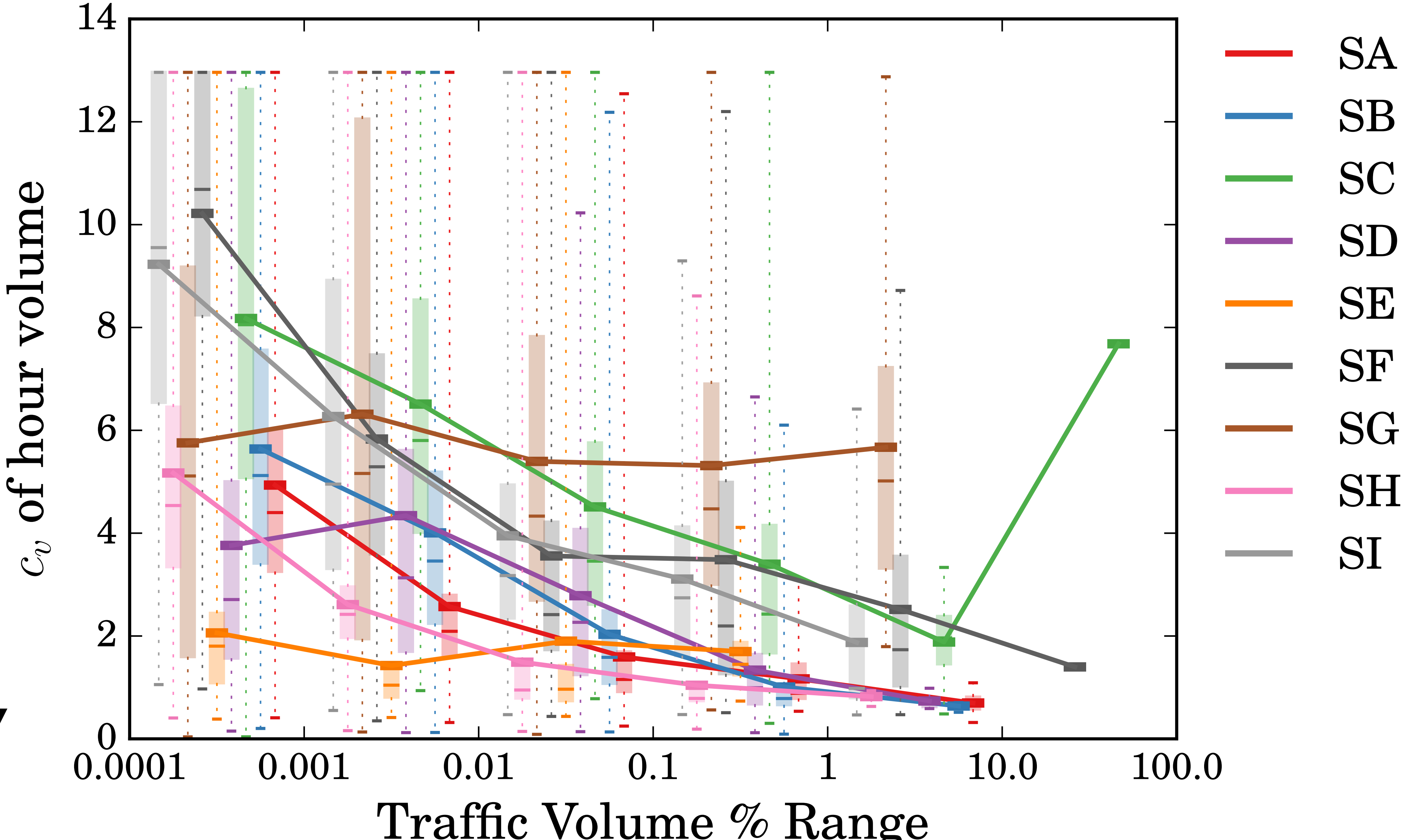
Volume importance vs predictability

Prefix volume share — Cv

Cv = std/mean

unstable

predictable



prefix of **small** volume share

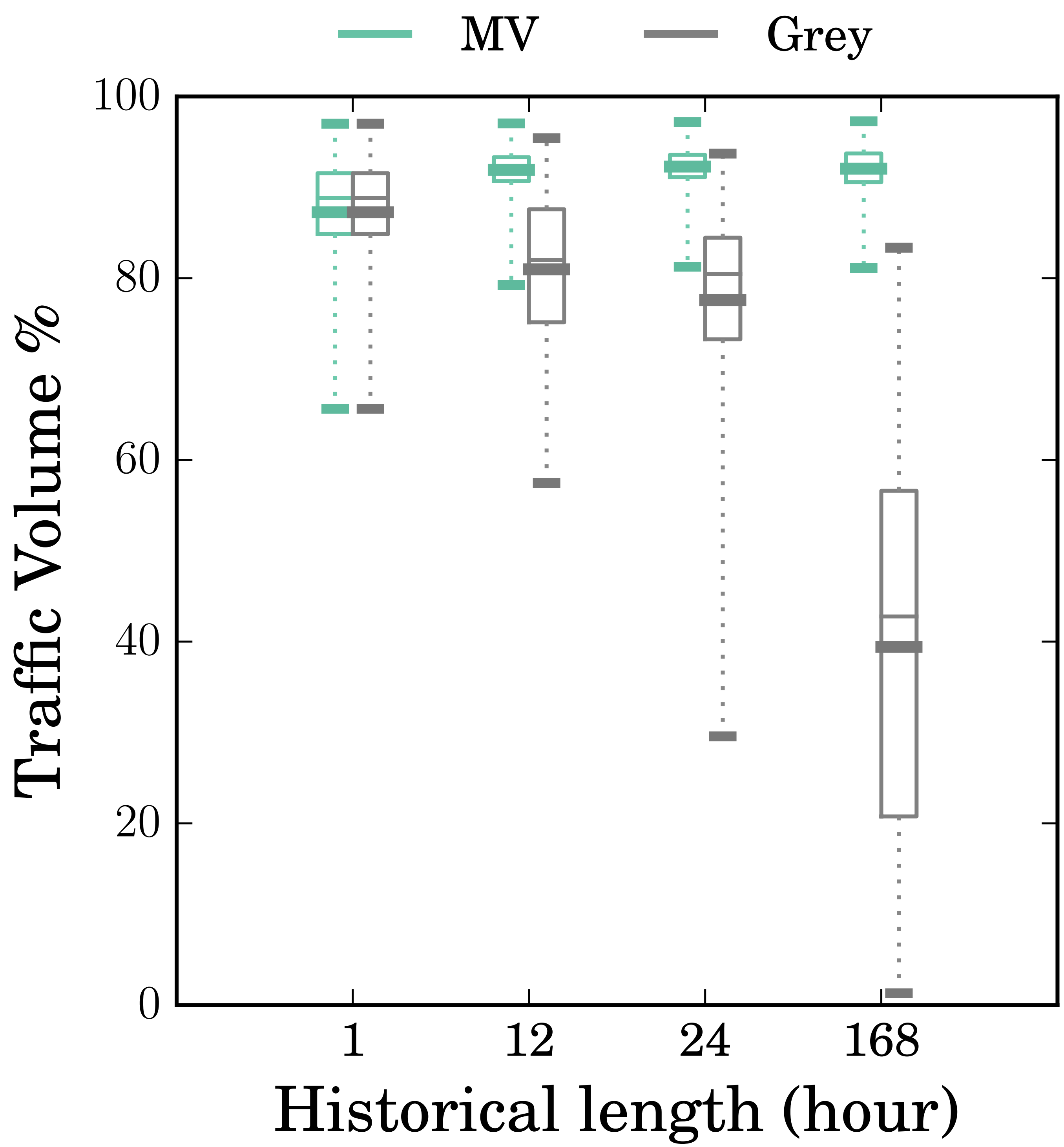
prefix of **big** volume share

A solution

as simple as
Moving Average

Volume coverage of MV compared to Grey Model GM(1,1)

good coverage



Poor coverage

short memory

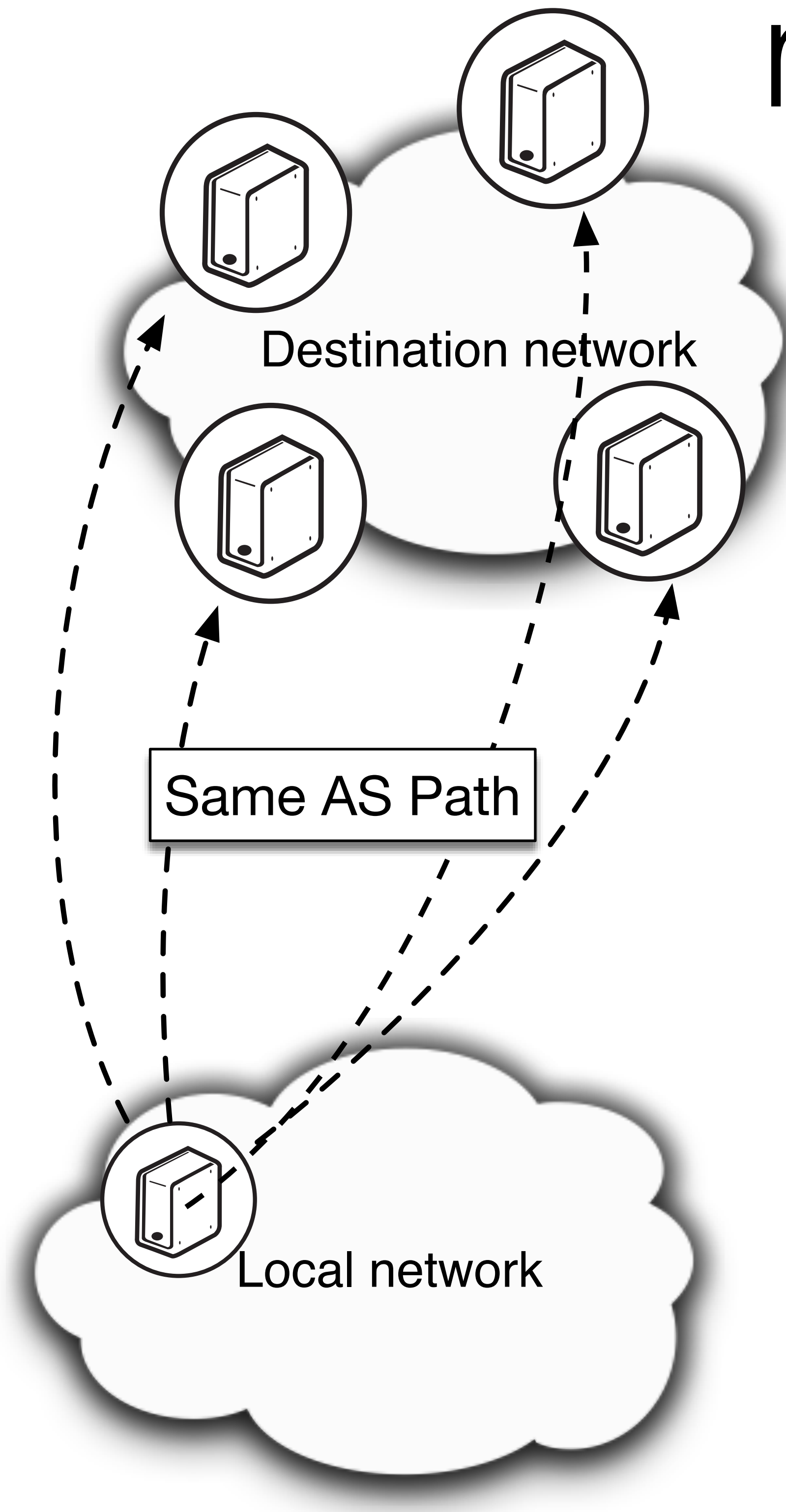
long memory

Summary

- An old scenario with many remaining challenges;
- A possible approach realising it;
- Measurement is indispensable.

Many other challenges…

Measure a same AS path with multiple probes



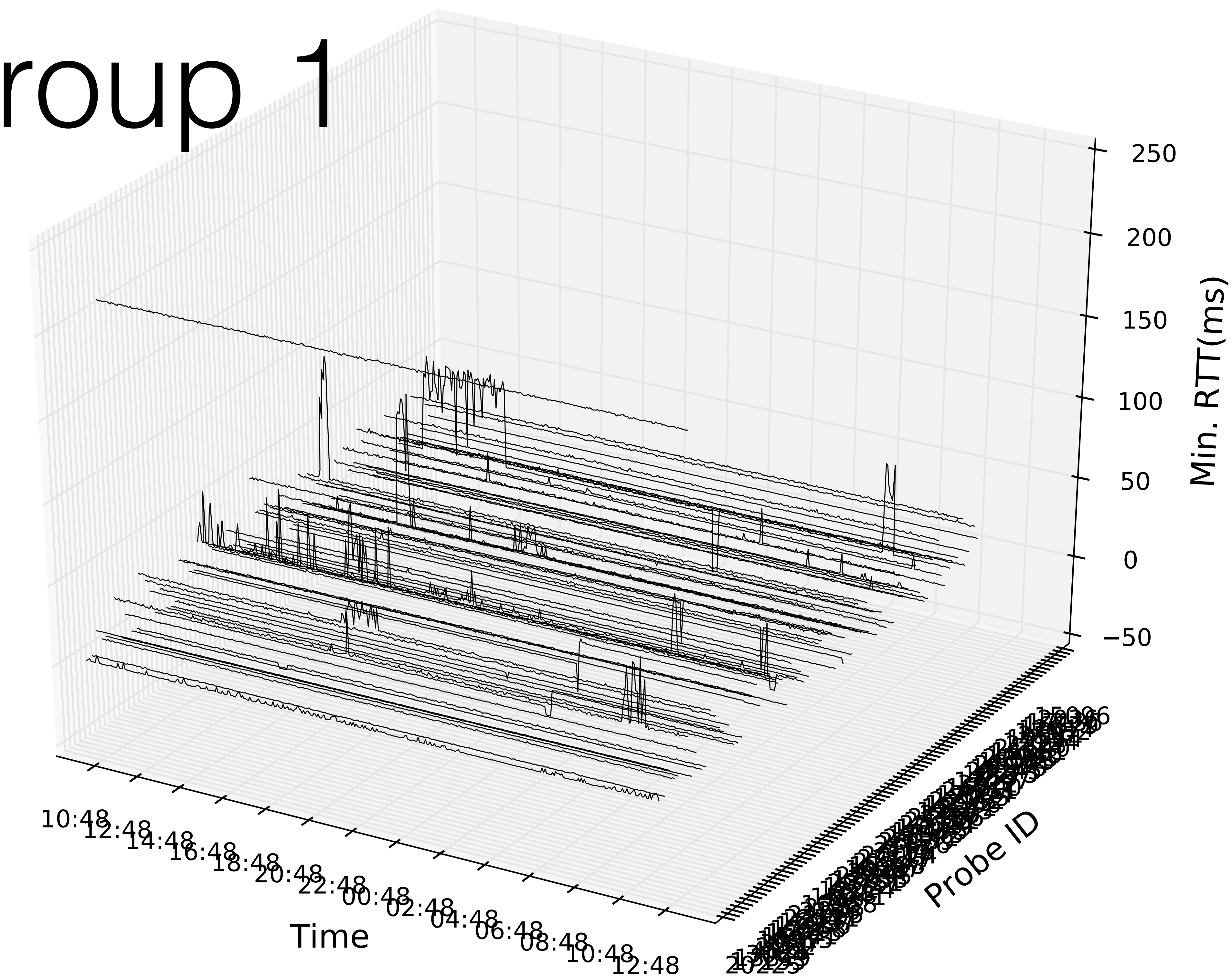
Are these measurements different? How?

Where does the difference take place?

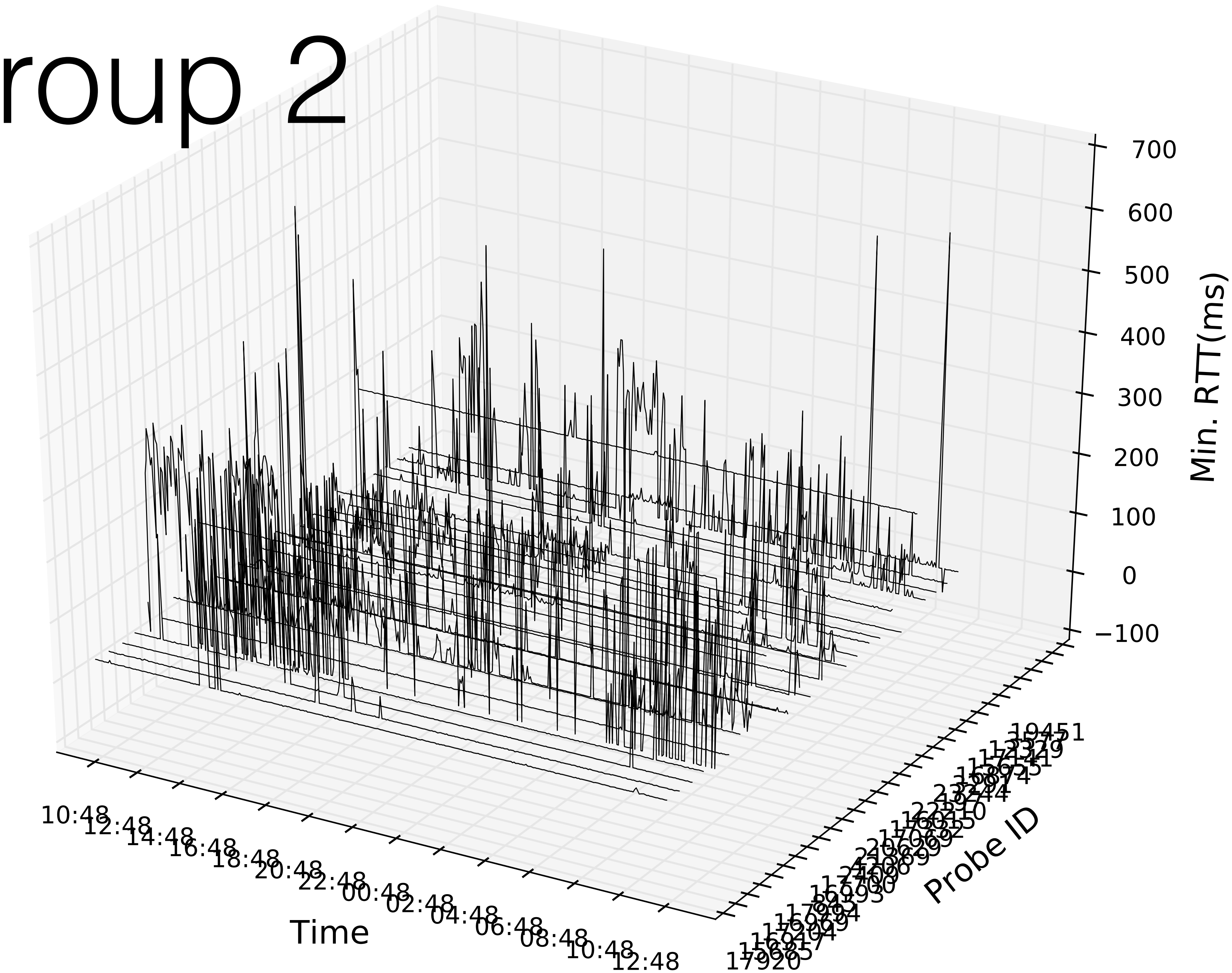
Which probe should we use?

W. Shao, J.L. Rougier, F. Devienne, and M. Viste, “**Improve RTT Measurement Quality via Clustering in Inter-Domain TE** ,” IEEE/IFIP AnNet, 2016.

end2end RTT
Group 1



end2end RTT
Group 2



Questions for you

- Does the scenario make sense?
- Your experiences on measurements and TE.

Appendix

Not all references are given.
The listed ones could be a good starting point.

https://github.com/WenqinSHAO/NANOG67_prez.git

Appendix-I

Measurement-based Inter-domain TE

- N. Feamster, D. G. Andersen, H. Balakrishnan, and M. F. Kaashoek, "Measuring the effects of internet path faults on reactive routing," ACM SIGMETRICS Perform. Eval. Rev., vol. 31, no. 1, p. 126, 2003.
- A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," SIGCOMM, 2003.
- D. K. Goldenberg, L. Qiu, H. Xie, Y. R. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," CCR, vol. 34, no. 4, p. 79, Oct. 2004.
- A. Akella, B. Maggs, S. Seshan, and A. Shaikh, "On the Performance Benefits of Multihoming Route Control," IEEE/ACM Trans. Netw., vol. 16, no. 1, pp. 91–104, Feb. 2008.
- W. Shao, L. Ianonne, J.L. Rougier, F. Devienne, and M. Viste, "Scalable BGP Prefix Selection for Effective Inter-domain Traffic Engineering," IEEE/IFIP NOMS, 2016.

Appendix-II

Traffic volume forecasting

- ARMA family, e.g. ARIMA, SARIMA, FARIMA, O(L^2), pre-process needed for each individual trace, L being historical record length.
- Artificial Neural Network family, e.g. TLFN, O(L*M), M for number of hidden nodes, usually bigger than L.
- Wavelet,O(L), pre-process needed, less accurate than FARIMA and ANN. H. Feng and Y. Shu, “Study on network traffic prediction techniques,” Proceedings. Int. Conf. Wirel. Commun. Netw. Mob. Comput., vol. 2, no. 3, pp. 995-998, 2005.
- Grey model GM(1,1) predicts the accumulated value of a time series, O(L). D. Julong, “Introduction to Grey System Theory,'' J. Grey Syst., vol. 1, pp. 1-24, 1989.

Appendix-III

Traffic dynamism

- K. Papagiannaki, N. Taft, and C. Diot, “Impact of Flow Dynamics on Traffic Engineering Design Principles,” INFOCOM, 2004.
  - no clear correlation between throughput and its stability.
- J. J. Wallerich and A. Feldmann, “Capturing the variability of internet flows across time,” INFOCOM, 2006.
  - throughput ranking of flows can change drastically over time.
- W. Zhang, J. Bi, J. Wu, and B. Zhang, “Catching popular prefixes at AS border routers with a prediction based method,” Comput. Networks, vol. 56, no. 4, pp. 1486–1502, Mar. 2012.
  - assumed positive correlation between popularity and stability.

Appendix-IV

FIB caching

- L. Iannone and O. Bonaventure, “On the cost of caching locator/ID mappings,” CoNEXT, 2007.
- C. Kim, M. Caesar, A. Gerber, and J. Rexford, “Revisiting route caching: The world should be flat,” PAM, 2009.
- H. Ballani, P. Francis, T. Cao, and J. Wang, “Making routers last longer with ViAggre,” NSDI, pp. 453–466, 2009.
- N. Sarrar, S. Uhlig, A. Feldmann, R. Sherwood, and X. Huang, “Leveraging Zipf’s law for traffic offloading,” CCR, vol. 42, no. 1, p. 16, Jan. 2012.
- W. Zhang, J. Bi, J. Wu, and B. Zhang, “Catching popular prefixes at AS border routers with a prediction based method,” Comput. Networks, vol. 56, no. 4, pp. 1486–1502, Mar. 2012.
    - GM(1,1) outperforms LFU, LRU