Data Center Traffic and Measurements: Available Bandwidth Estimation

Hakim Weatherspoon

Assistant Professor, Dept of Computer Science

CS 5413: High Performance Systems and Networking

November 14, 2014

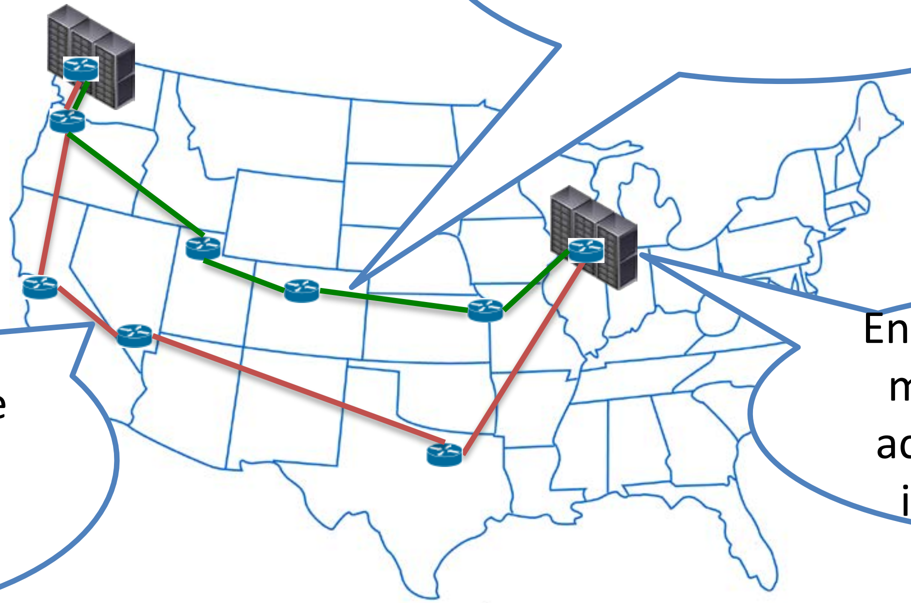
Slides from ACM SIGCOMM conference on Internet measurement (IMC), 2014, presentation of “MinProbe: Accurate, Minimum Overhead, Available Bandwidth Estimation in High Speed Wired Networks”

Goals for Today

- MinProbe: Accurate, Minimum Overhead, Available Bandwidth Estimation in High Speed Wired Networks
  - Ki Suh Lee, Erluo Li, ChiunLin Lim, Kevin Tang and Hakim Weatherspoon. In Proceedings of the 14th ACM SIGCOMM conference on Internet measurement (IMC), November 2014.

Available Bandwidth Estimation

- Basic building block
  - Network Protocol
  - Networked Systems
  - Distributed Systems



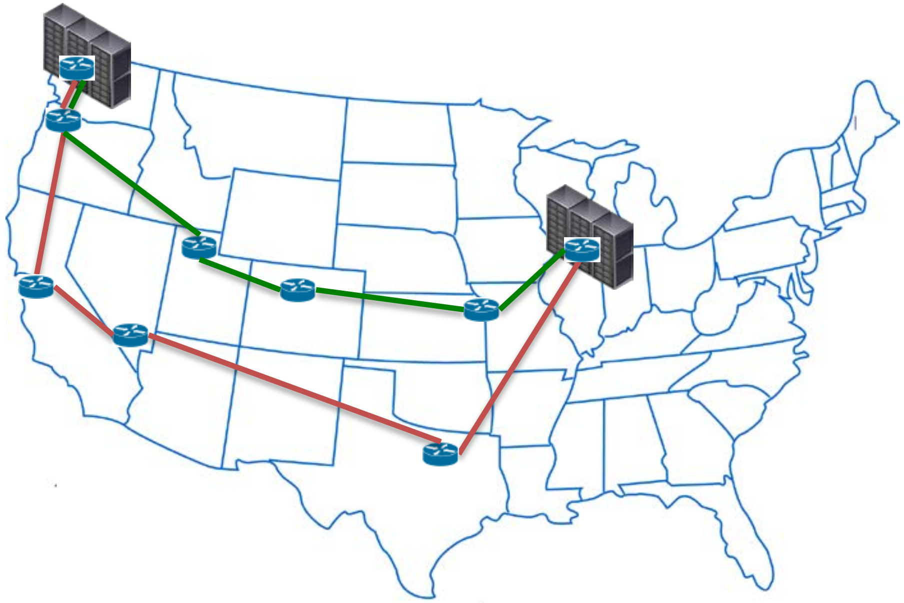
Which of the two paths has more available bandwidth?

How do I measure with minimum overhead?

End-to-end: How to measure without access to anything in the network?

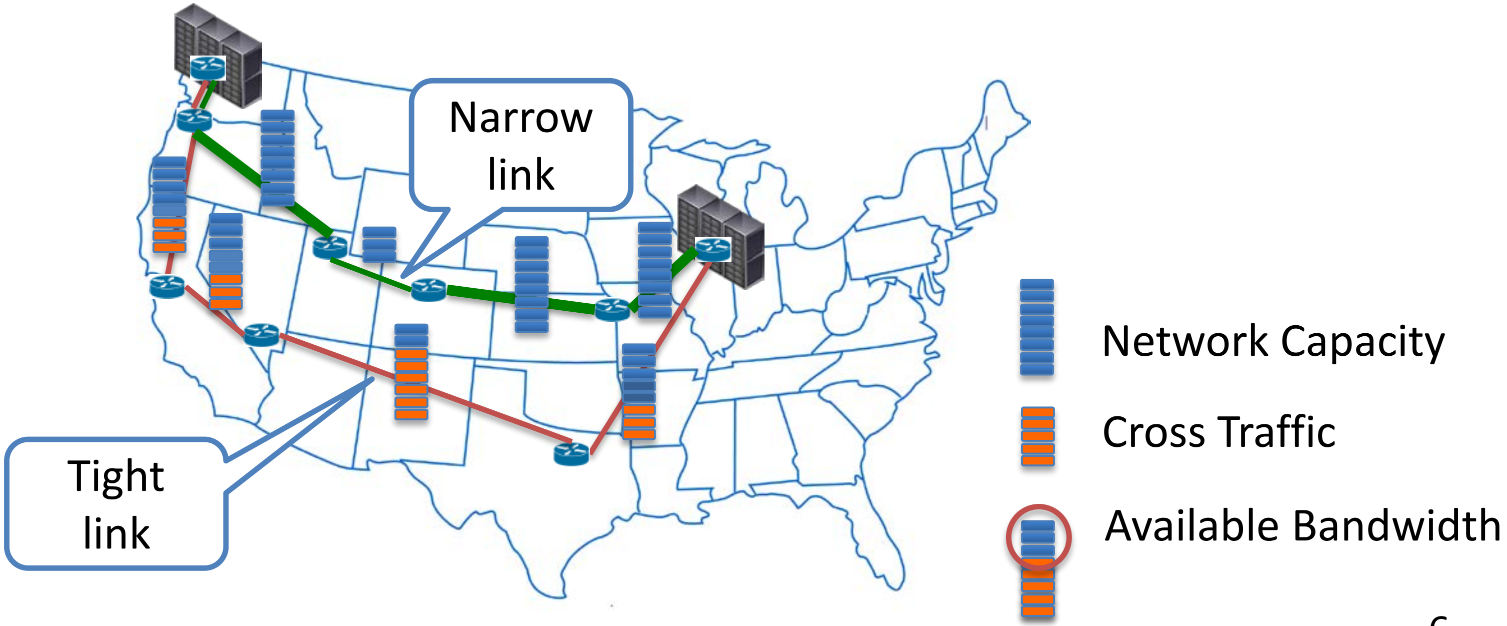
Available Bandwidth Estimation

- Passive Measurement
  - Polling counters: Port Stats or Flow Stats
- Active Measurement
  - *Probe Packets:* Packet Pair or Packet Train



Active Measurement

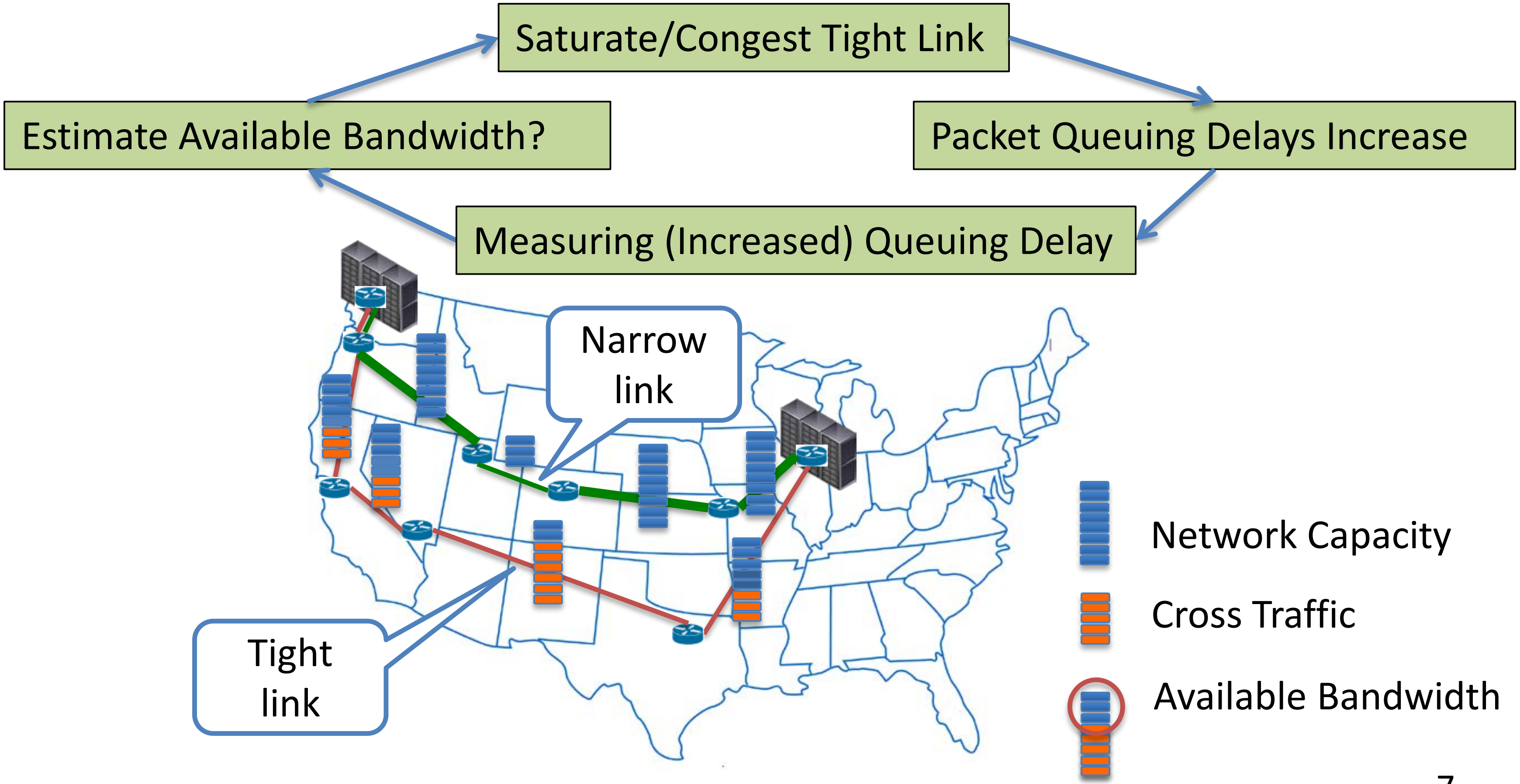
- Narrow link: least capacity
- Tight link: least available bandwidth



Narrow link
Tight link
Network Capacity
Cross Traffic
Available Bandwidth

Active Measurement

- Estimate available bandwidth by saturating the tight link



Saturate/Congest Tight Link

Packet Queuing Delays Increase

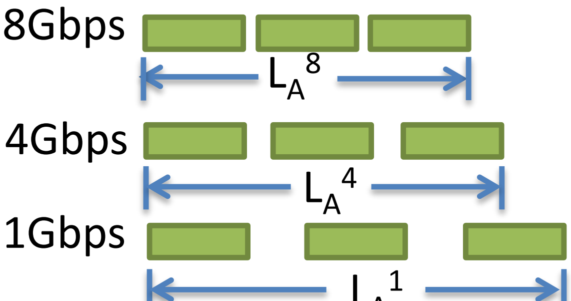
Measuring (Increased) Queuing Delay

Estimate Available Bandwidth?

Narrow link
Tight link
Network Capacity
Cross Traffic
Available Bandwidth

Active Measurement

- Estimate available bandwidth by saturating the tight link

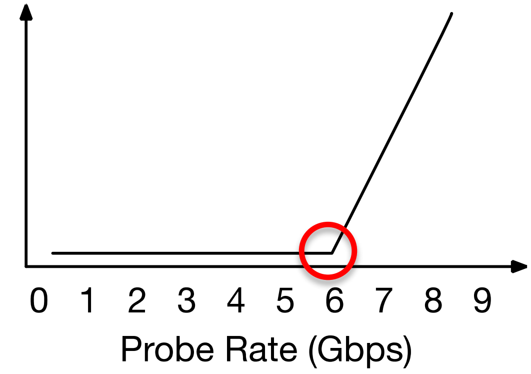

8Gbps
L_A^8
4Gbps
L_A^4
1Gbps
L_A^1

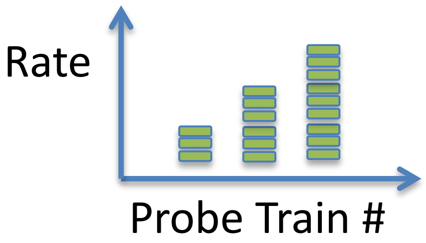
$L_A^8 < L_B^8$, congestion!

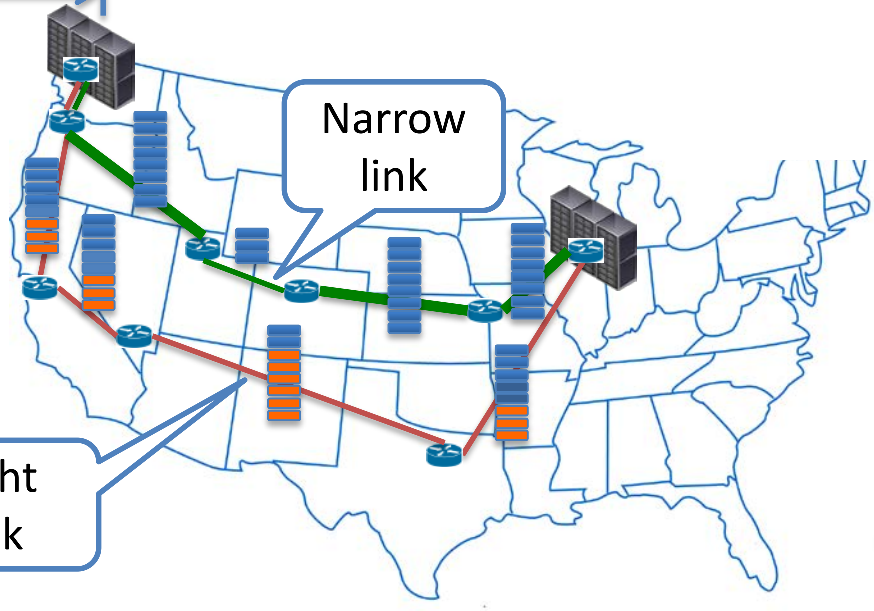
$L_A^4 == L_B^4$, no congestion

$L_A^1 == L_B^1$, no congestion

Increase in Queuing Delay Variance






Narrow link
Tight link

🟦 Network Capacity

🟧 Cross Traffic

○ Available Bandwidth

By measuring the *increase in packet train length**, we can compute the *queuing delay* experienced, hence estimate the available bandwidth.

* Increase in packet train length == increase in sum of inter-packet gap

Limitations of Available Bandwidth Estimation

- Intrusive/Expensive
  - 100s of probe packets per packet train
- Inaccurate
  - Especially in high speed network
- Does not work well for bursty traffic

Goals of Available Bandwidth Estimation

- Cheap
  - Use as little probe packets as possible
- Accurate
  - How close is the estimation to the actual value
- Works for all traffic patterns

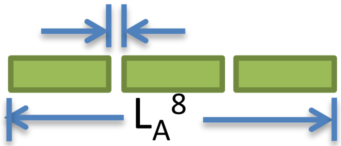
MinProbe: Bandwidth Estimation in PHY

- Cheap
  - Use as little probe packets as possible
  - ~~100s of probe packets per train~~ → 20 packets per train
  - Using application traffic as probe
- Accurate
  - How close is the estimation to the actual value
  - Error < 0.4Gbps, with a resolution of 0.1Gbps
- Works for all traffic patterns
  - Works with bursty cross traffic

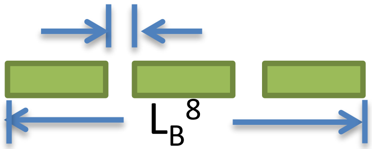
Outline

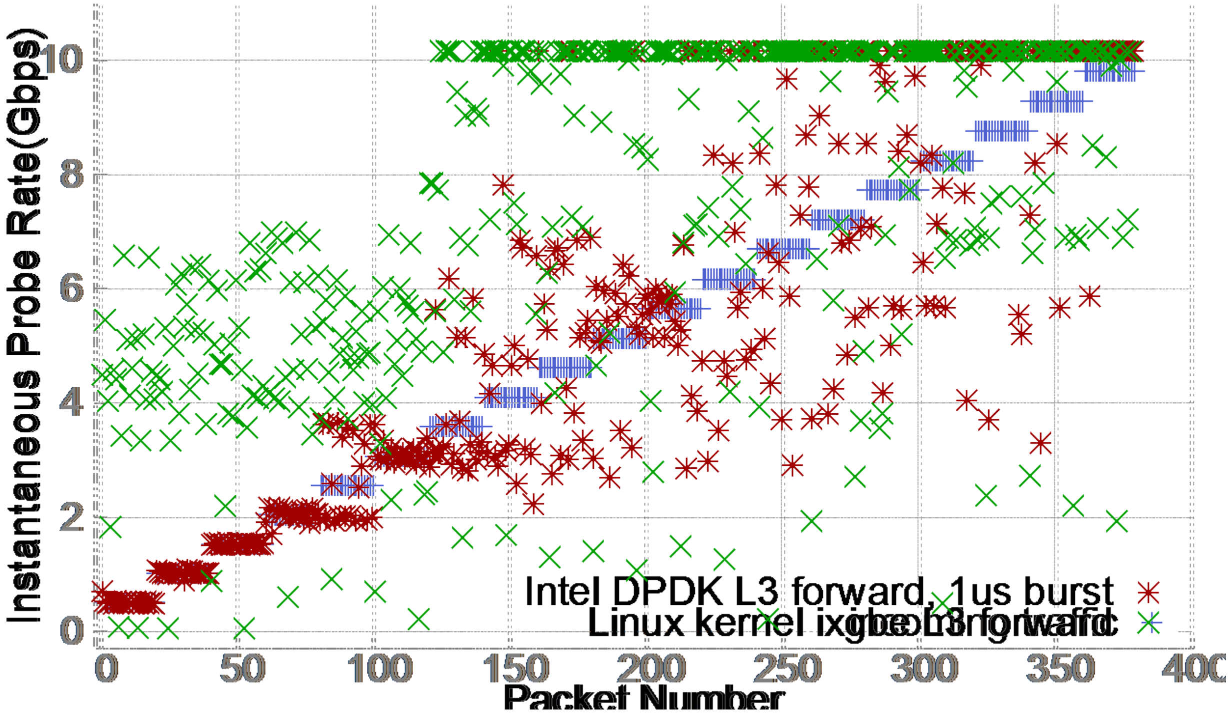
Challenges

Cannot Control at 100ps



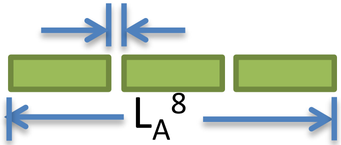
Cannot Measure at 100ps



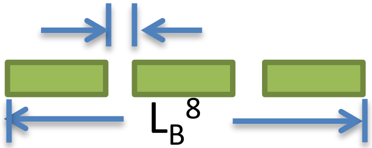


Challenges

Cannot Control at 100ps



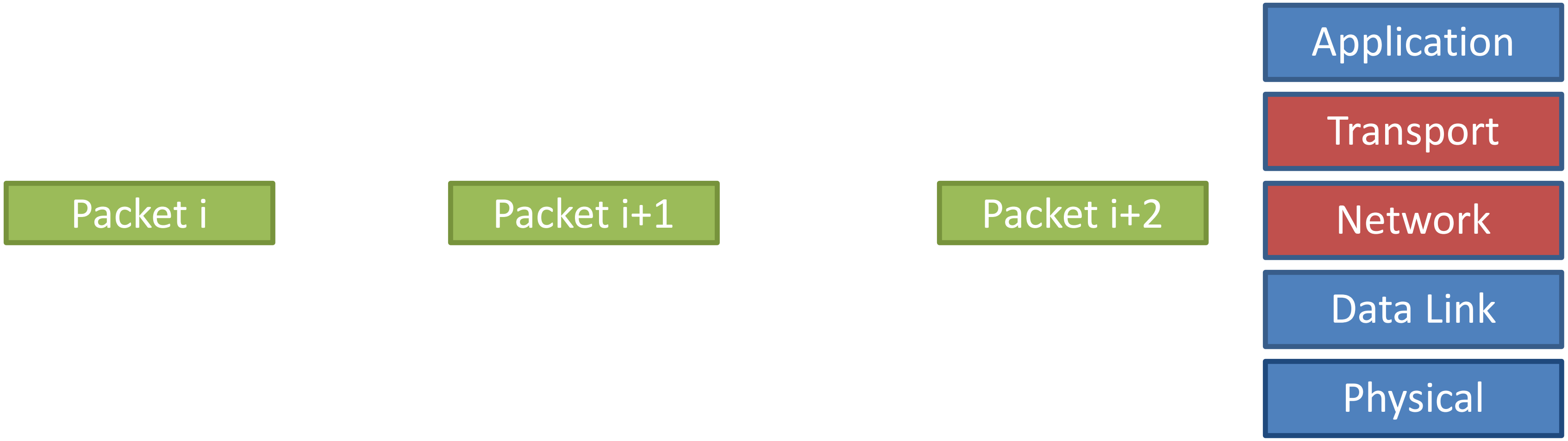
Cannot Measure at 100ps



State-of-art (software) tools do ***not*** work at high speed because they cannot control and capture inter-packet spacing with required precision.

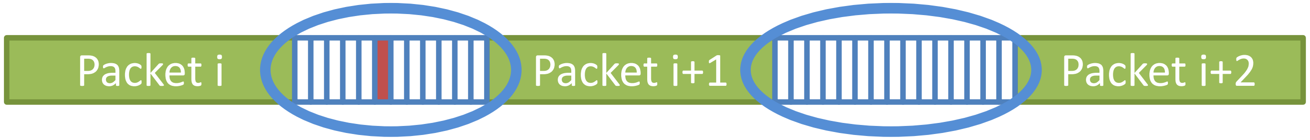
Outline

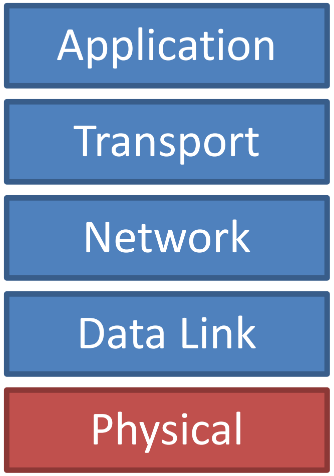
MinProbe: Better Accuracy


Packet i
Packet i+1
Packet i+2
Application
Transport
Network
Data Link
Physical

MinProbe: Better Accuracy

- Idle Characters (/I/)

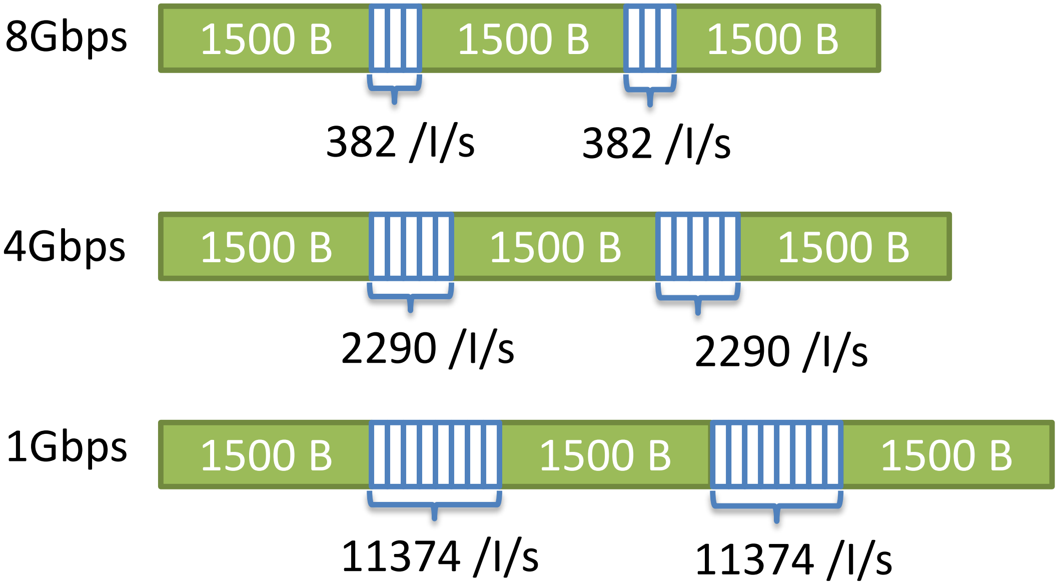

Packet i
Packet i+1
Packet i+2

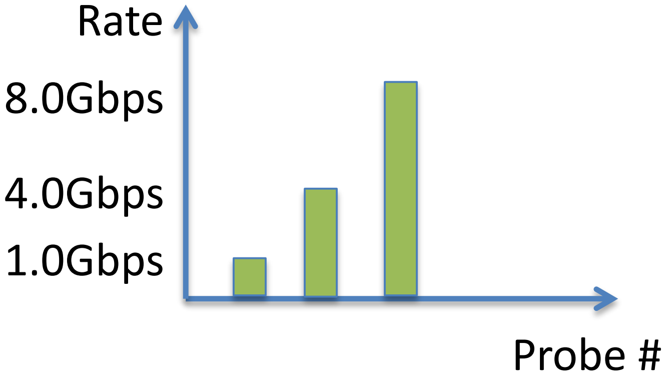

Application
Transport
Network
Data Link
Physical

  - Each bit ~100 picoseconds
  - 7~8 bit special character in the physical layer
  - 700~800 picoseconds to transmit
  - Only in PHY

MinProbe: Better Accuracy

- Probe Generation

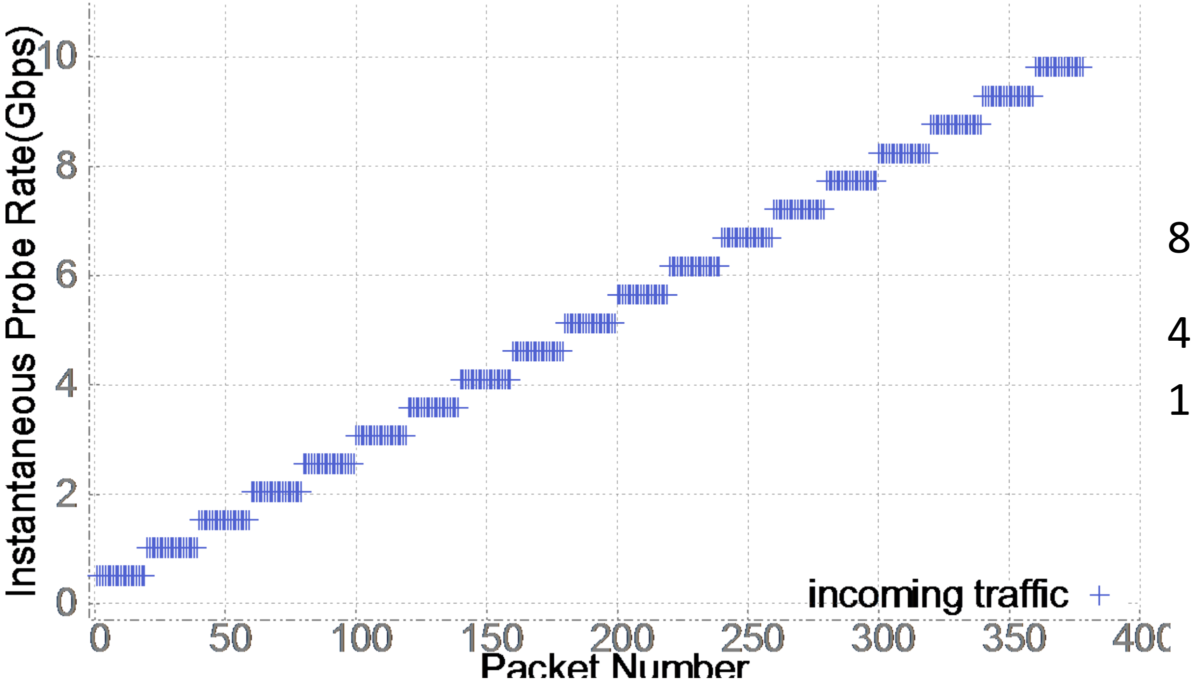

8Gbps
1500 B
1500 B
1500 B
382 /I/s
382 /I/s
4Gbps
1500 B
1500 B
1500 B
2290 /I/s
2290 /I/s
1Gbps
1500 B
1500 B
1500 B
11374 /I/s
11374 /I/s

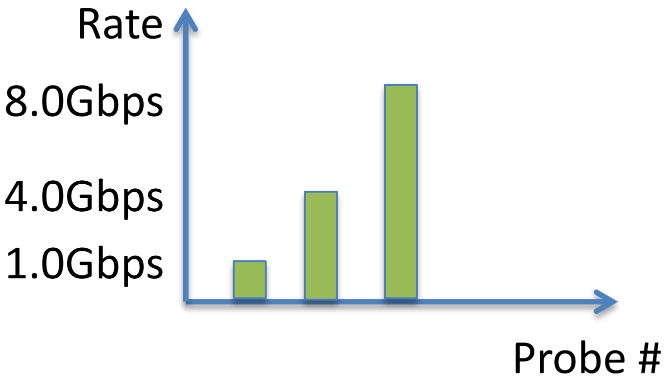


By modulating Inter-packet gap at PHY layer, we can generate accurate probe rate.

MinProbe: Better Accuracy

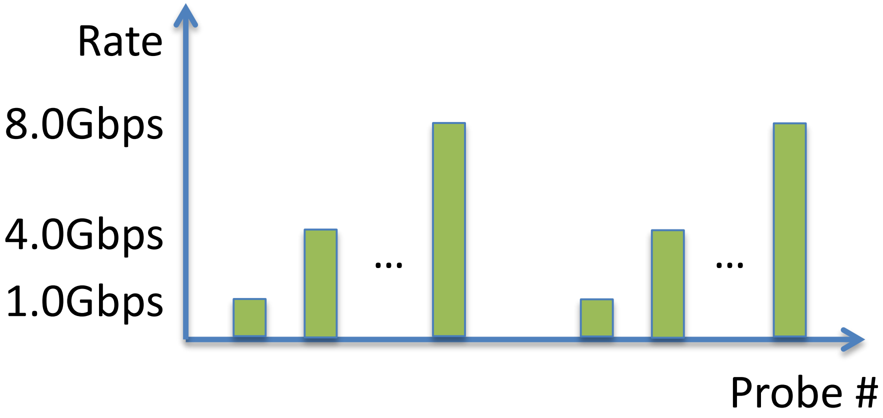
- Probe Generation




Rate
8.0Gbps
4.0Gbps
1.0Gbps
Probe #

MinProbe: Better Accuracy

- With shared N and (Rmin, Rstep, Rmax)



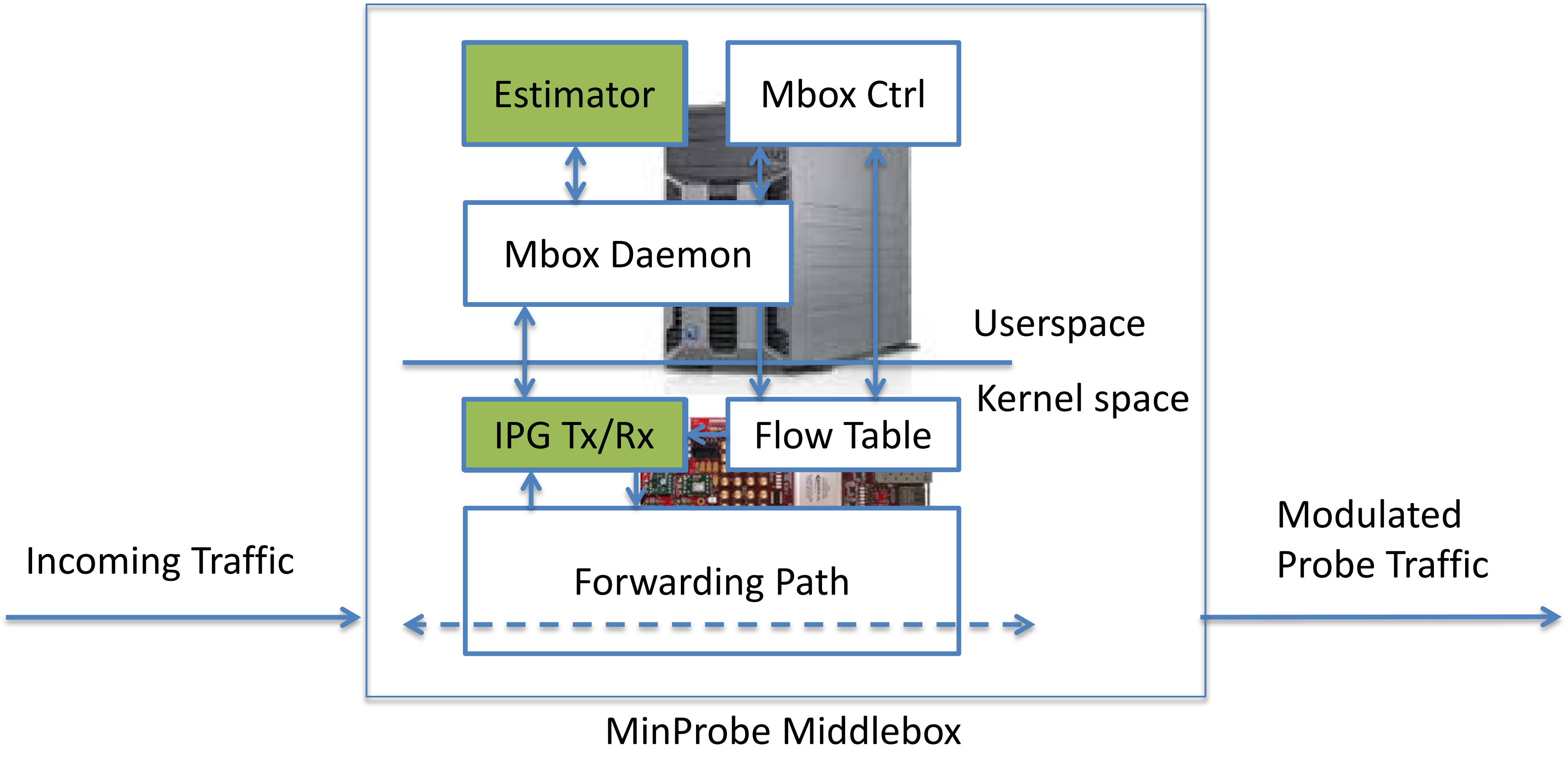
- Emulate existing bandwidth estimation algorithm
  - Pathload, Pathchirp, SLoPS …

MinProbe: Low Overhead

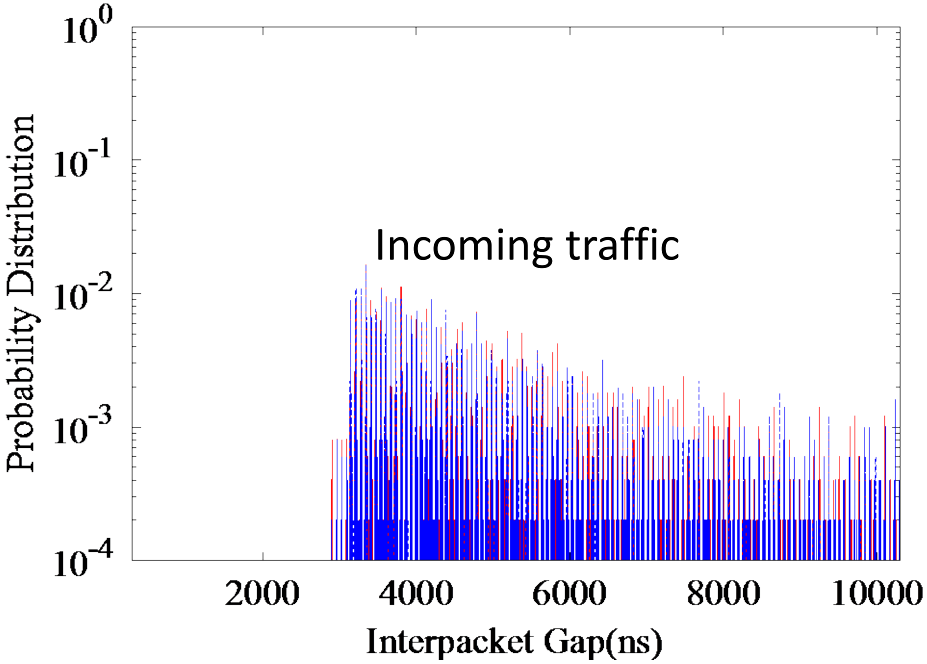
- Reduce the # of probe packets required

# of probe packet per probe train	**Error (%)**
5	+15%
20	-1%
40	-2%
60	-3%
80	-3%
100 (Pathload)	-4%

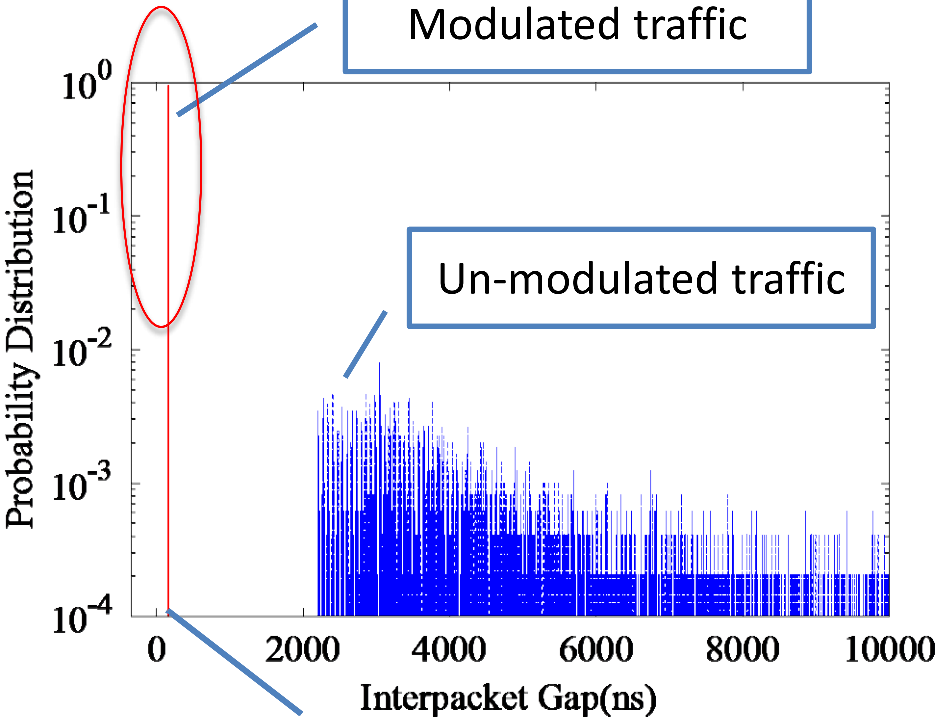
MinProbe: Very Low Overhead


Estimator
Mbox Ctrl
Mbox Daemon
Userspace
Kernel space
IPG Tx/Rx
Flow Table
Forwarding Path
Incoming Traffic
Modulated
Probe Traffic
MinProbe Middlebox

MinProbe: Bandwidth Estimation in PHY



Incoming traffic



Modulated traffic

Un-modulated traffic

200 /I/s @ 8Gbps

- Packet Size 792 Bytes

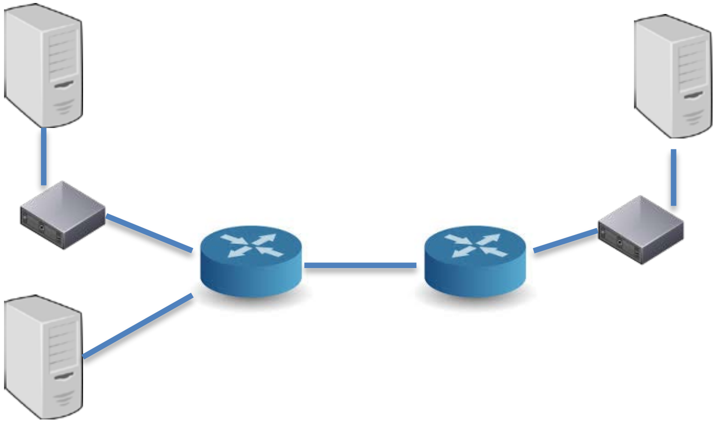
Outline

- Introduction
- Challenges
- Design: MinProbe
- Evaluation
- Conclusion

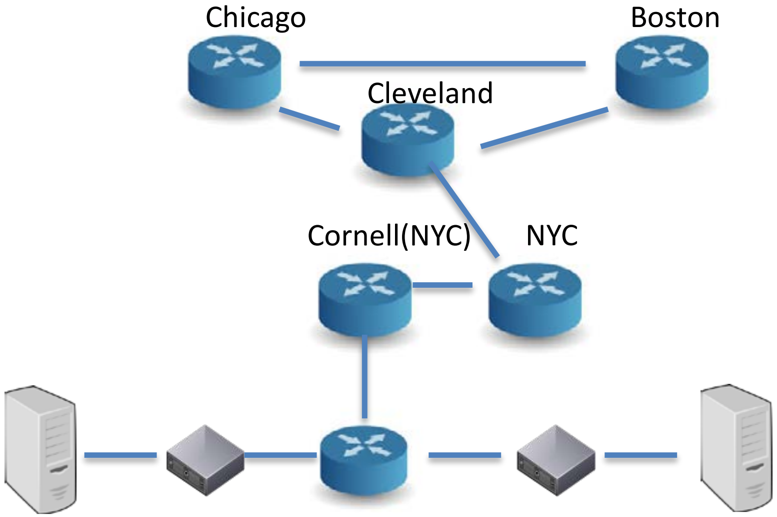
Questions:

- Can MinProbe accurately estimate avail-bw at 10Gbps?
- Can existing estimation algorithms work with MinProbe?
- How do the following parameters affect accuracy?
  - Packet train length
  - probe packet size distribution
  - cross packet size distribution
  - cross packet burstiness
- Does MinProbe work in the wild, Internet?
- Does MinProbe work in rate limiting environments?

Experiment Setup



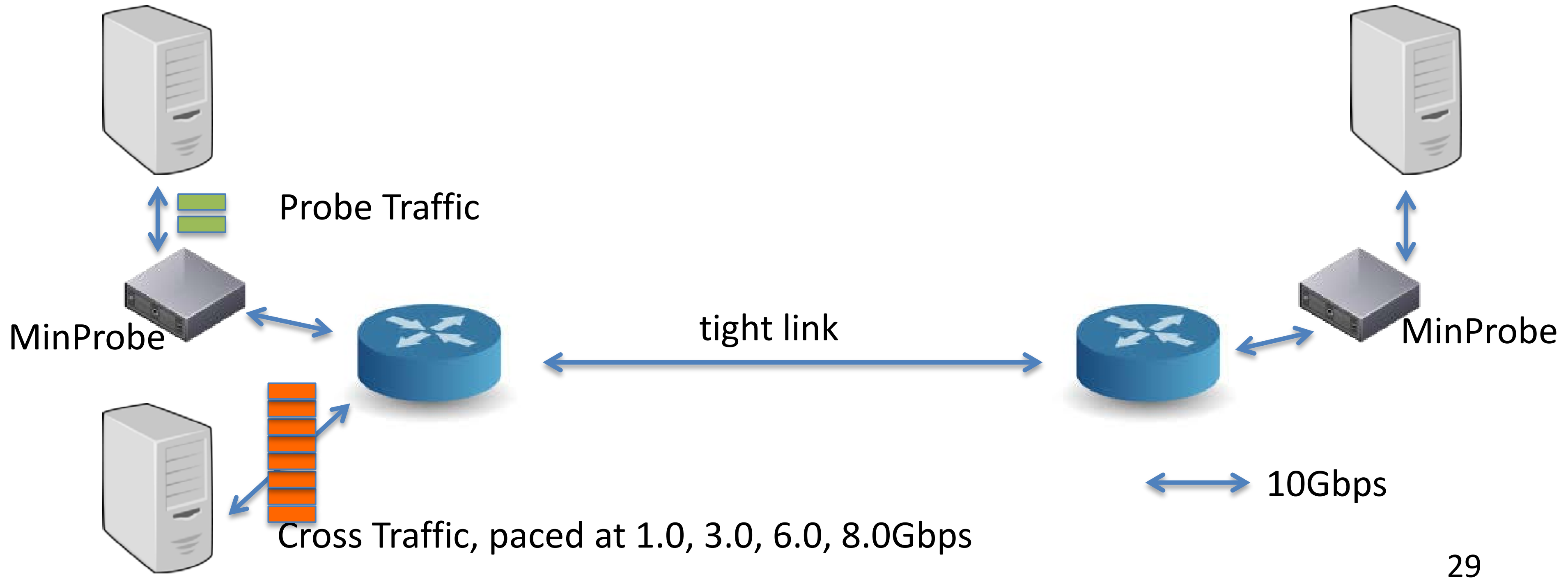
Controlled Environment


Chicago
Boston
Cleveland
Cornell(NYC)
NYC

National Lambda Rail

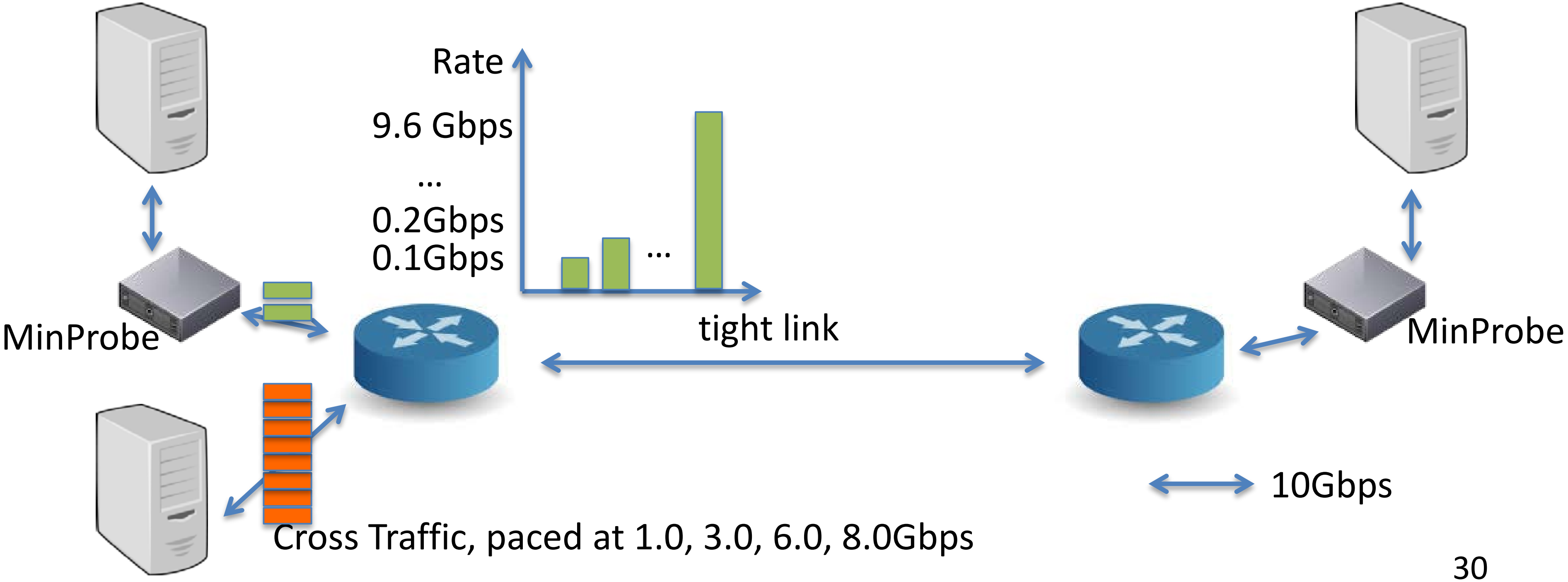
MinProbe: Can we measure at 10Gbps?

- Cross Traffic
- Probe Traffic


Probe Traffic
MinProbe
tight link
MinProbe
10Gbps
Cross Traffic, paced at 1.0, 3.0, 6.0, 8.0Gbps

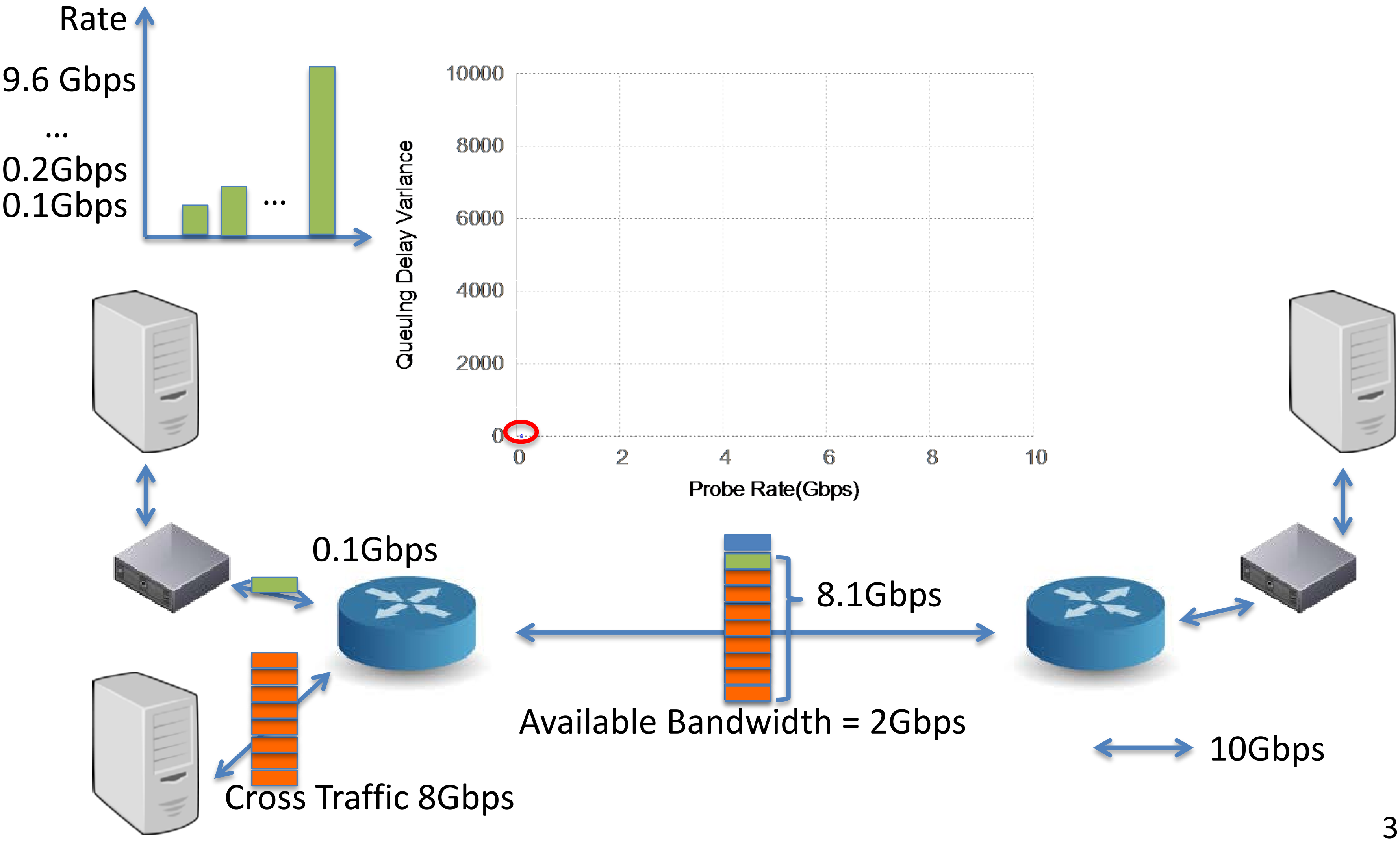
MinProbe: Can we measure at 10Gbps?

- Cross Traffic
- Probe Traffic


Rate
9.6 Gbps
…
0.2Gbps
0.1Gbps
…
MinProbe
tight link
MinProbe
10Gbps

Cross Traffic, paced at 1.0, 3.0, 6.0, 8.0Gbps

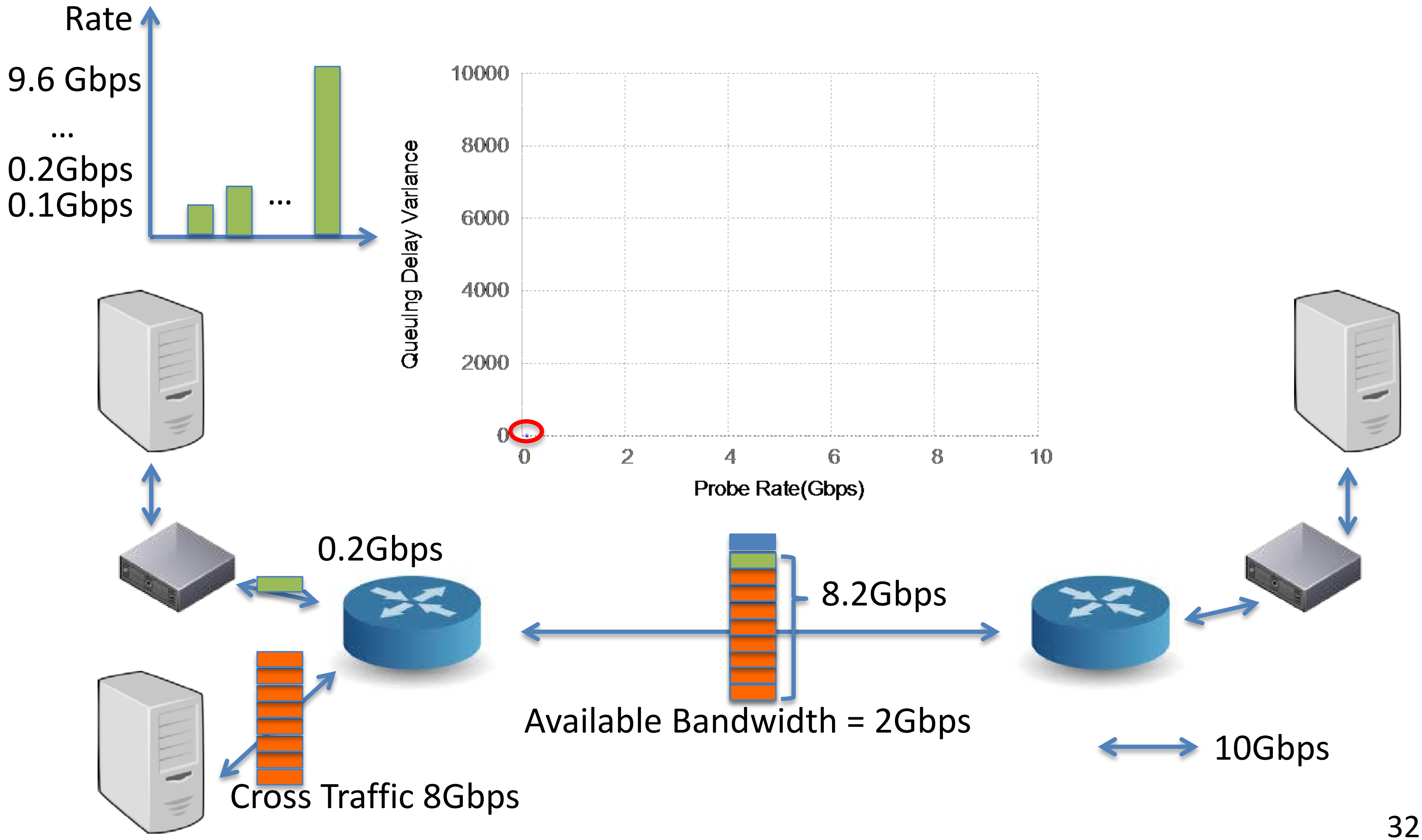
MinProbe: Can we measure at 10Gbps?


Rate
9.6 Gbps
…
0.2Gbps
0.1Gbps
…
Queuing Delay Variance
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
Probe Rate(Gbps)
0.1Gbps
8.1Gbps
10Gbps

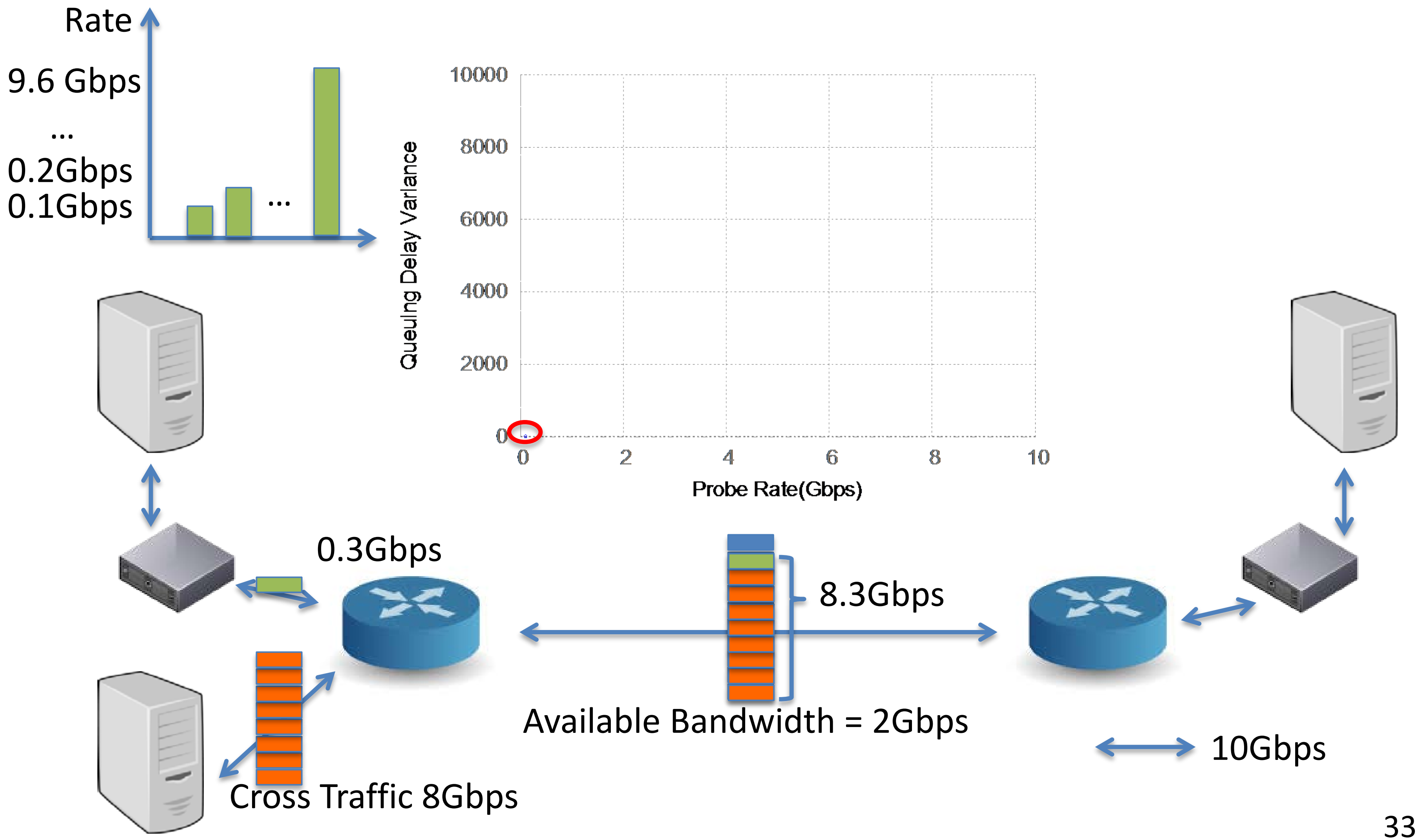
Available Bandwidth = 2Gbps

Cross Traffic 8Gbps

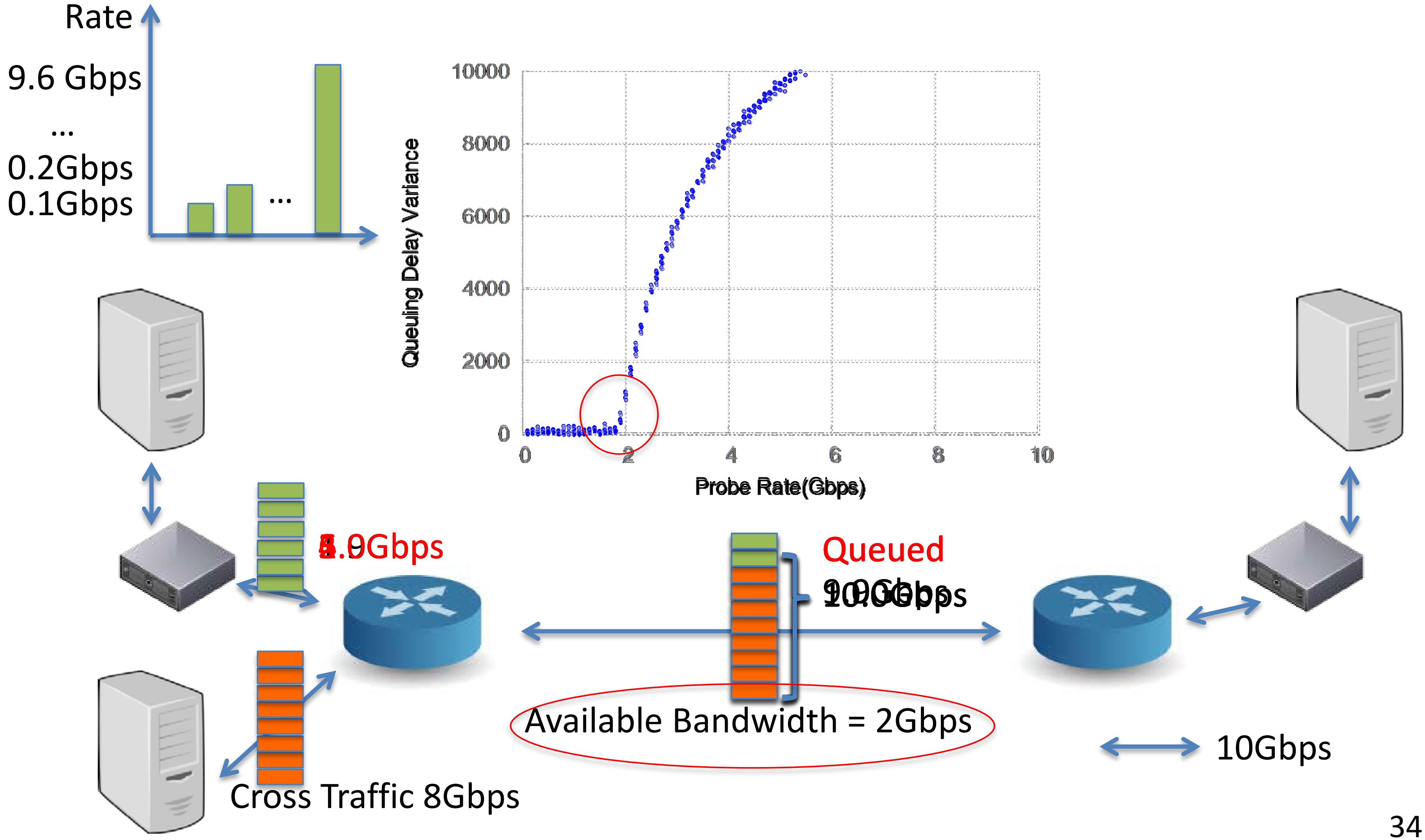
MinProbe: Can we measure at 10Gbps?


Rate
9.6 Gbps
…
0.2Gbps
0.1Gbps
…
Queuing Delay Variance
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
Probe Rate(Gbps)
0.2Gbps
8.2Gbps
Available Bandwidth = 2Gbps
10Gbps
Cross Traffic 8Gbps

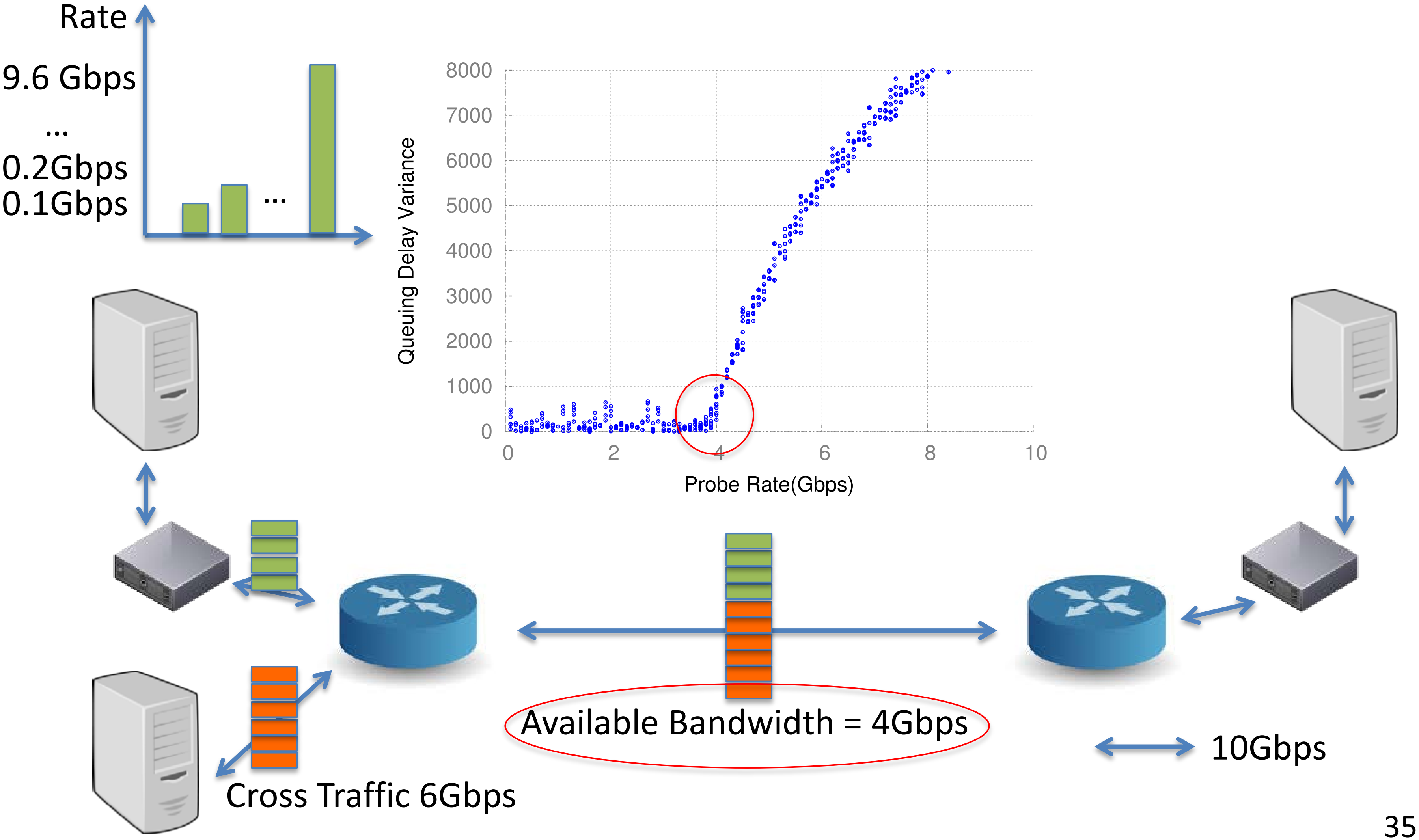
MinProbe: Can we measure at 10Gbps?


Rate
9.6 Gbps
…
0.2Gbps
0.1Gbps
…
Queuing Delay Variance
10000
8000
6000
4000
2000
0
0
2
4
6
8
10
Probe Rate(Gbps)
0.3Gbps
8.3Gbps
Available Bandwidth = 2Gbps
10Gbps
Cross Traffic 8Gbps

MinProbe: Can we measure at 10Gbps?



MinProbe: Can we measure at 10Gbps?

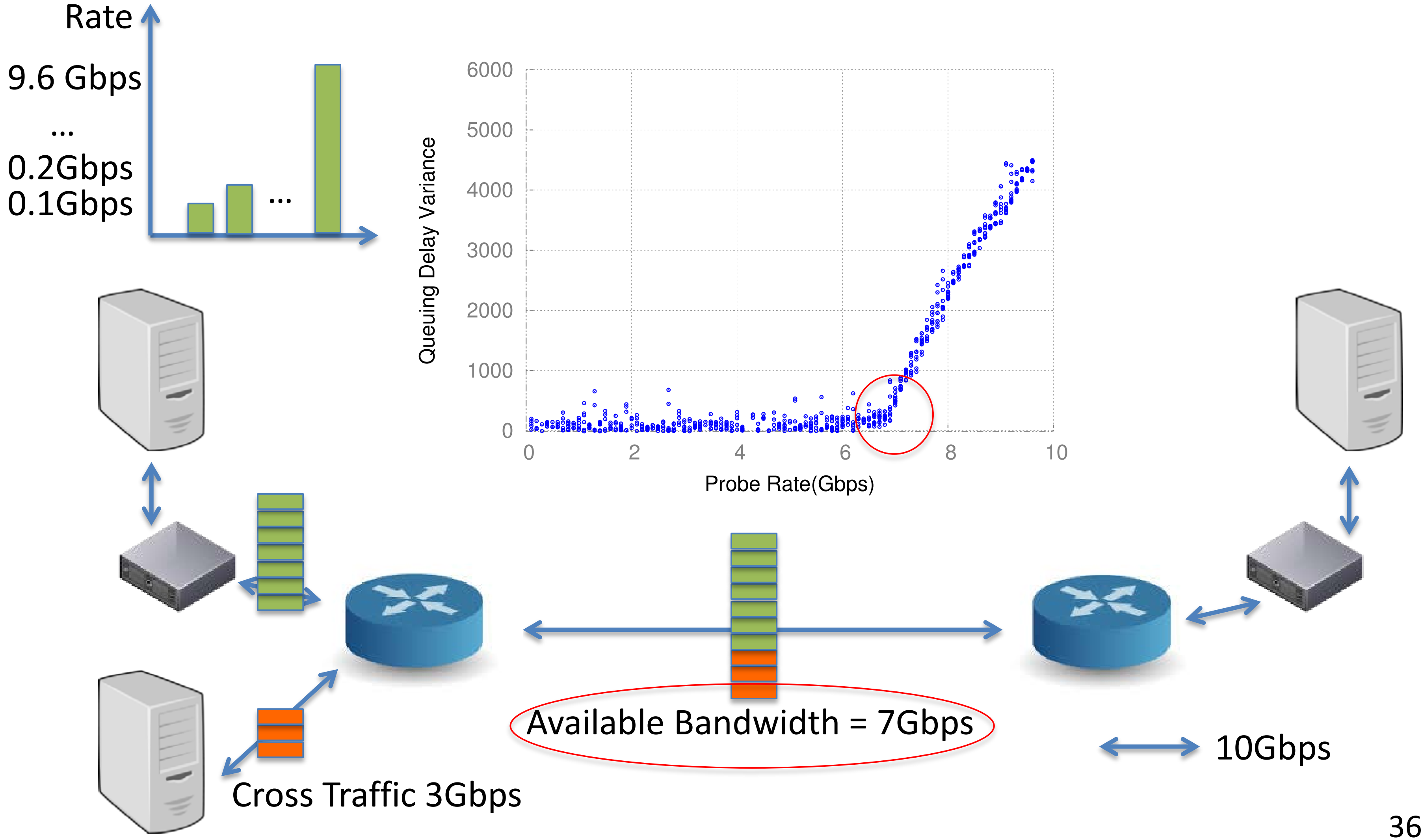

Rate
9.6 Gbps
…
0.2Gbps
0.1Gbps
…
Queuing Delay Variance
0
1000
2000
3000
4000
5000
6000
7000
8000
0
2
4
6
8
10
Probe Rate(Gbps)

Available Bandwidth = 4Gbps

Cross Traffic 6Gbps

10Gbps

MinProbe: Can we measure at 10Gbps?

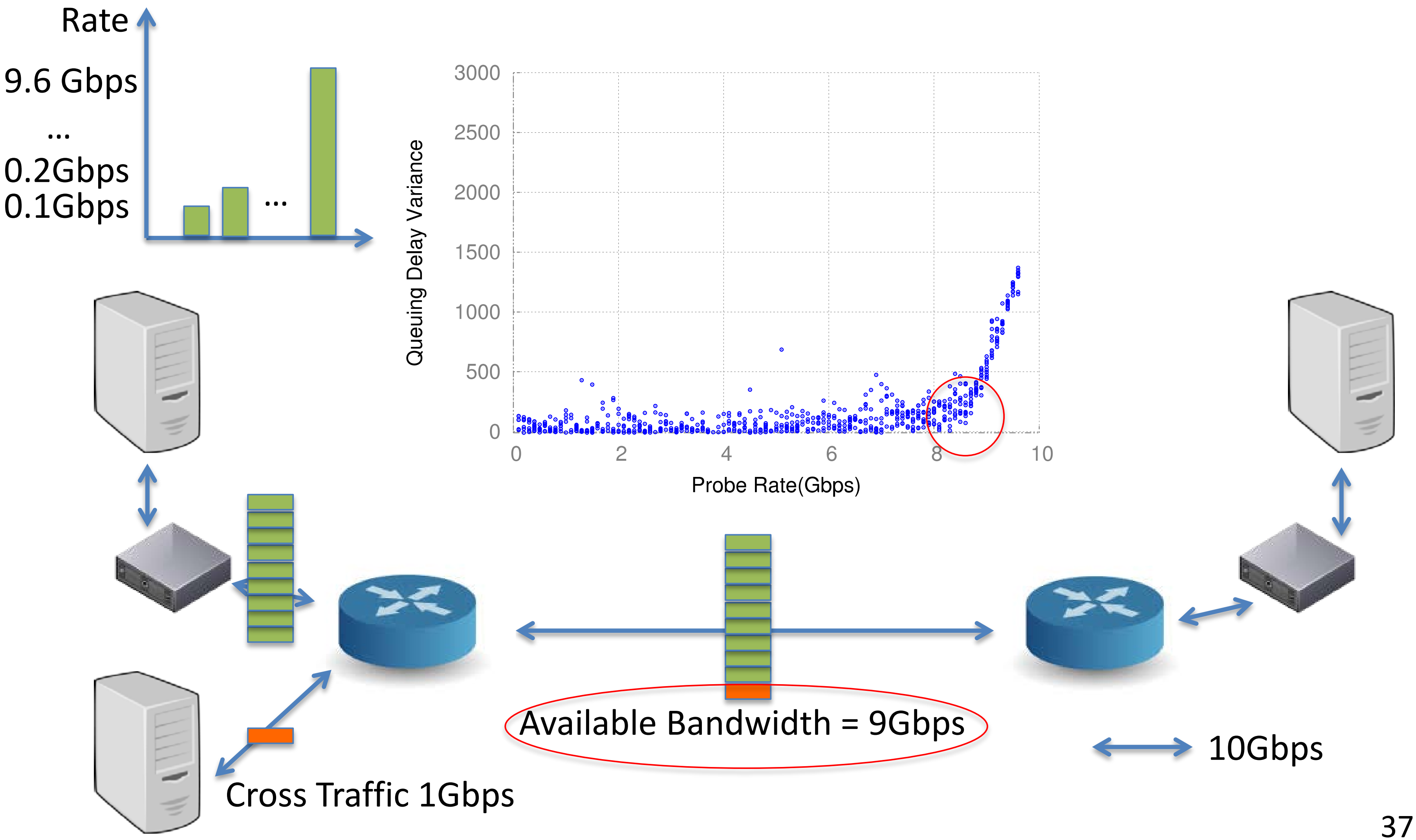


Available Bandwidth = 7Gbps

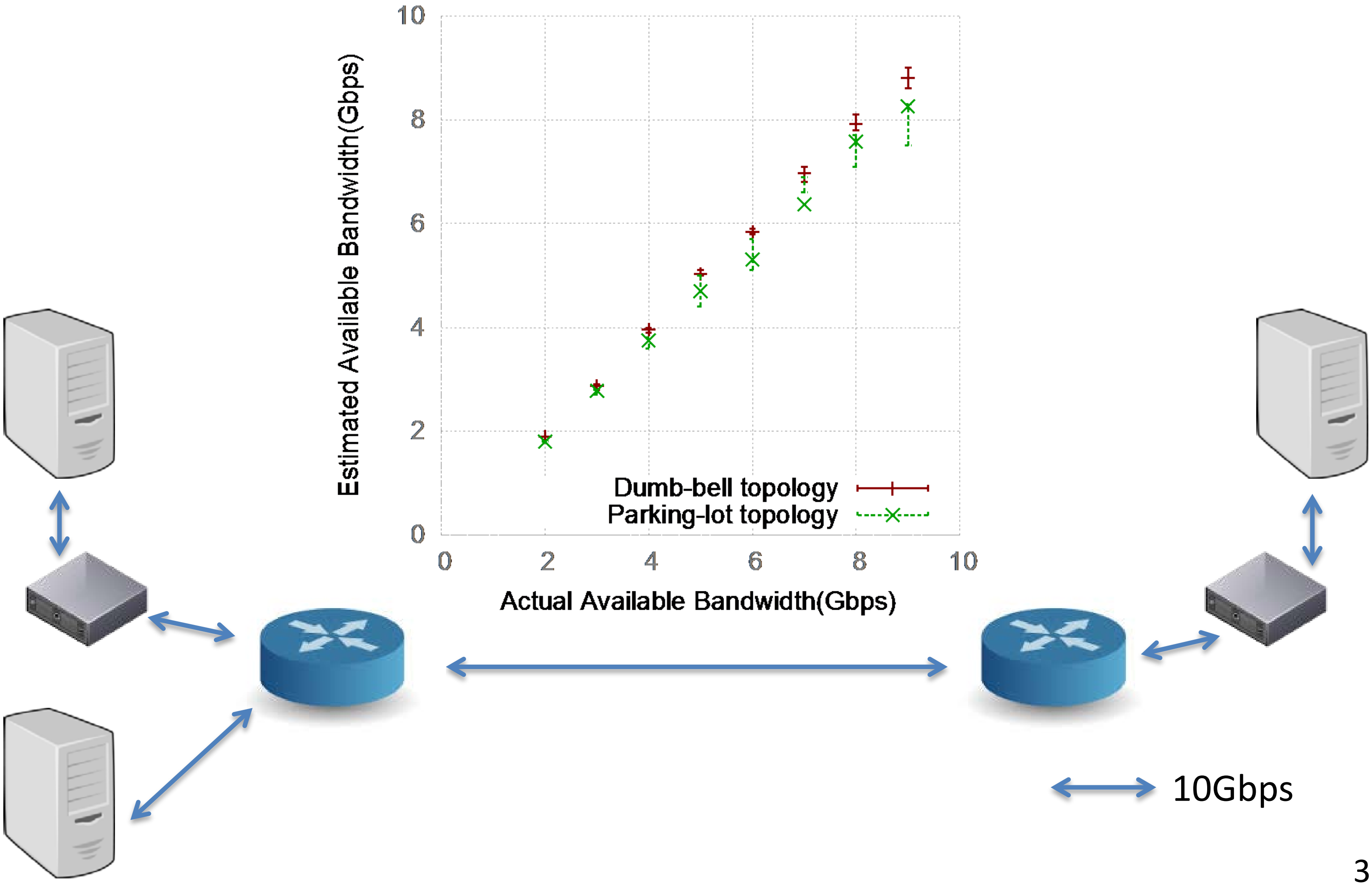
Cross Traffic 3Gbps

⟷ 10Gbps

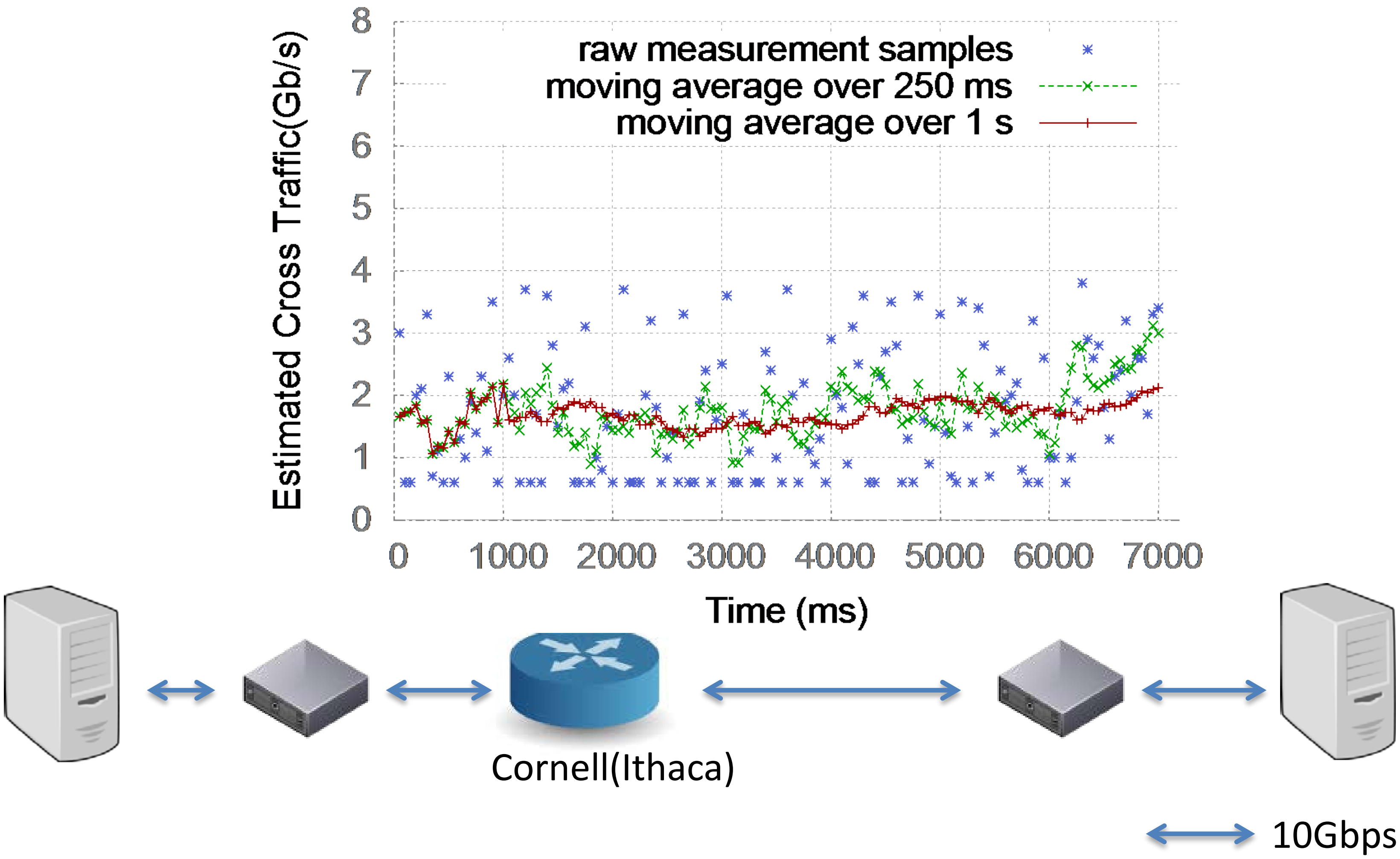
MinProbe: Can we measure at 10Gbps?


Rate
9.6 Gbps
...
0.2Gbps
0.1Gbps
...
Queuing Delay Variance
3000
2500
2000
1500
1000
500
0
0
2
4
6
8
10
Probe Rate(Gbps)
Available Bandwidth = 9Gbps
Cross Traffic 1Gbps
10Gbps

MinProbe: Can we measure at 10Gbps?



MinProbe: National Lambda Rail

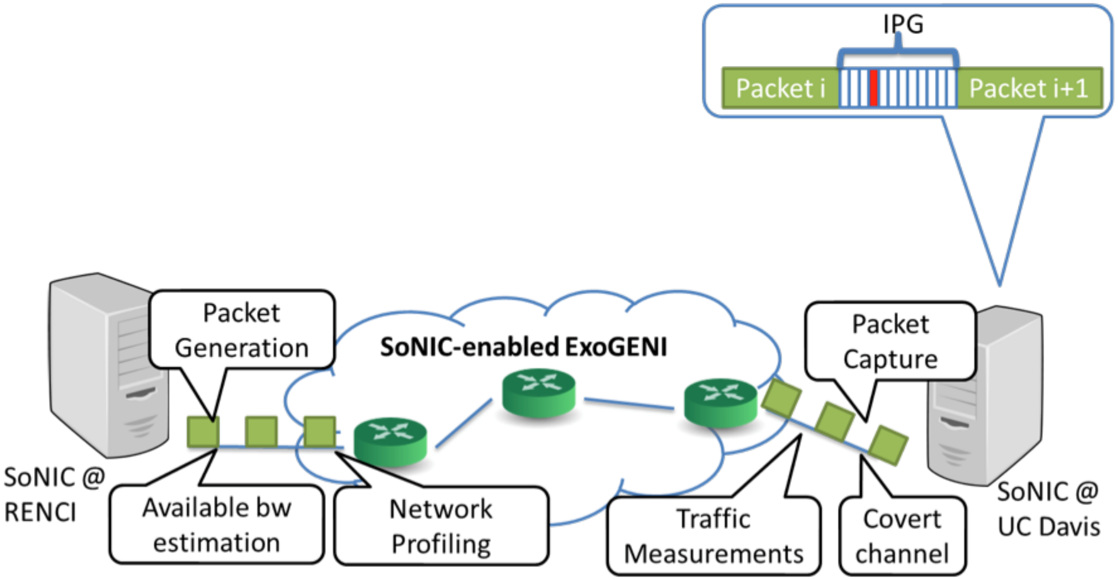


Questions:

- Can MinProbe accurately estimate avail-bw at 10Gbps?
- Can existing estimation algorithms work with MinProbe?
- How do the following parameters affect accuracy?
  - Packet train length
  - probe packet size distribution
  - cross packet size distribution
  - cross packet burstiness
- Does MinProbe work in the wild, Internet?
- Does MinProbe work in rate limiting environments?

MinProbe: In rate limiting environment?

- Rate Limiters may be an issue? No


IPG
Packet i
Packet i+1
Packet Generation
SoNIC-enabled ExoGENI
Packet Capture
SoNIC @ RENCI
Available bw estimation
Network Profiling
Traffic Measurements
Covert channel
SoNIC @ UC Davis

  - Provisioned 1Gbps virtual network
  - If the probe train is less than 1200 packets, no packet loss

Outline

Conclusion

- Modulation of probe packets in PHY
- Accurate control & measure of packet timing
- Enabled available bandwidth estimation in 10Gbps

Accurate, Minimum Overhead, Available Bandwidth Estimation in High Speed Wired Networks

Before Next time

- Project Interim report
  - **Due Monday, November 24.**
  - And meet with groups, TA, and professor
- Fractus Upgrade: Should be back online

- ***Required review and reading for Monday, November 17***
  - The Xen-Blanket: Virtualize Once, Run Everywhere, D. Williams, H. Jamjoom, and H. Weatherspoon. ACM European Conference on Computer Systems (EuroSys), April 2012, pages 113-126.
  - http://dl.acm.org/citation.cfm?id=2168849
  - http://fireless.cs.cornell.edu/publications/xen-blanket.pdf

- Check piazza: http://piazza.com/cornell/fall2014/cs5413
- Check website for updated schedule