Riptide: Jump Starting Back-Office Connections in Cloud Systems

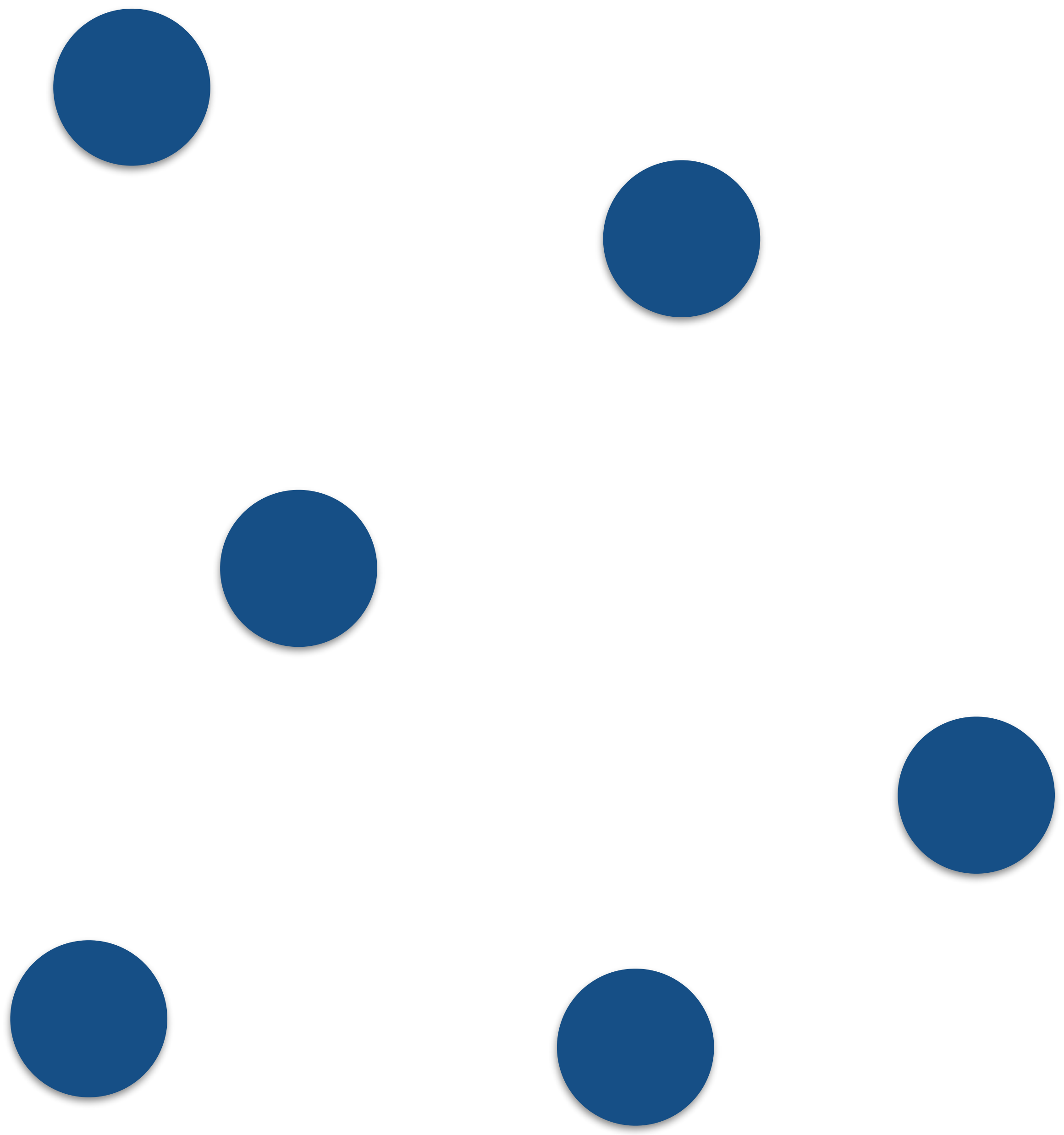
Marcel Flores - Northwestern University
Amir R. Khakpour - Verizon Digital Media Services
Harkeerat Bedi - Verizon Digital Media Services

Cloud systems

- Large scale global services:
  - CDNs, web services.
- *Back-office* traffic between Points of Presence (PoPs).
  - Control messages, small transfers.

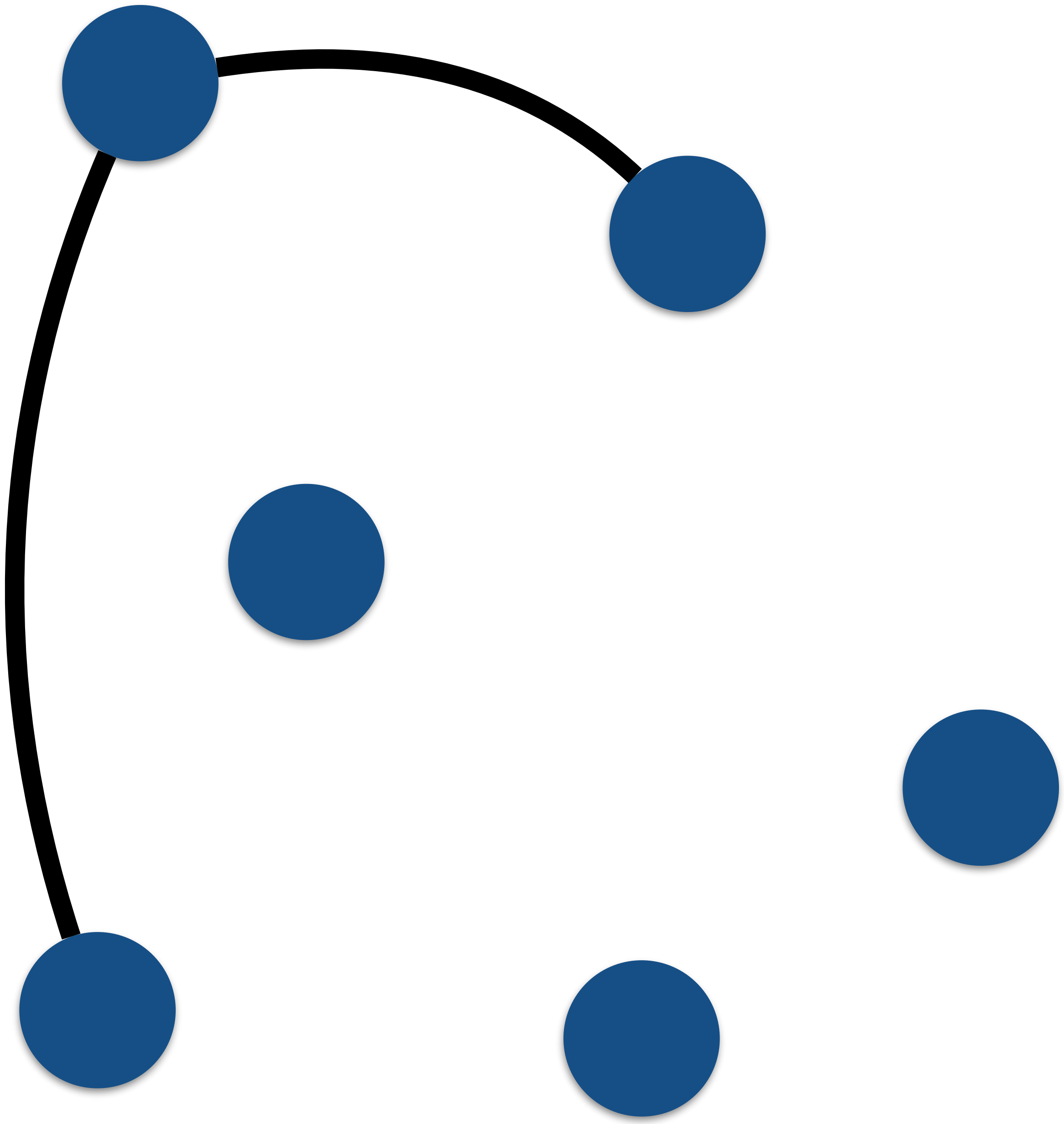
Cloud systems

- Frequent opening of connections between PoPs.
- In a perfect world, would have a mesh.
- Application and resource constraints limit reuse.



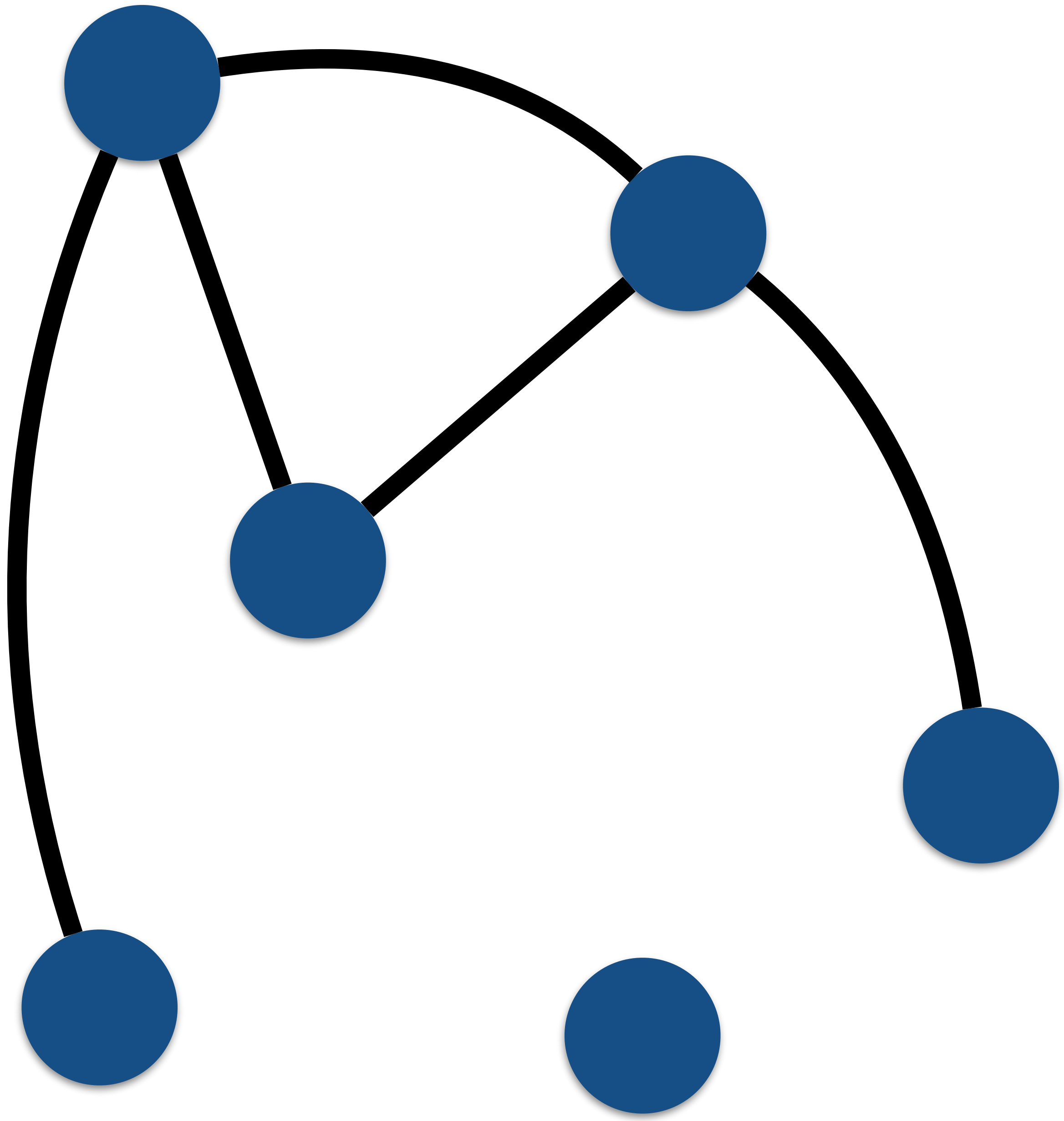
Cloud systems

- Frequent opening of connections between PoPs.
- In a perfect world, would have a mesh.
- Application and resource constraints limit reuse.



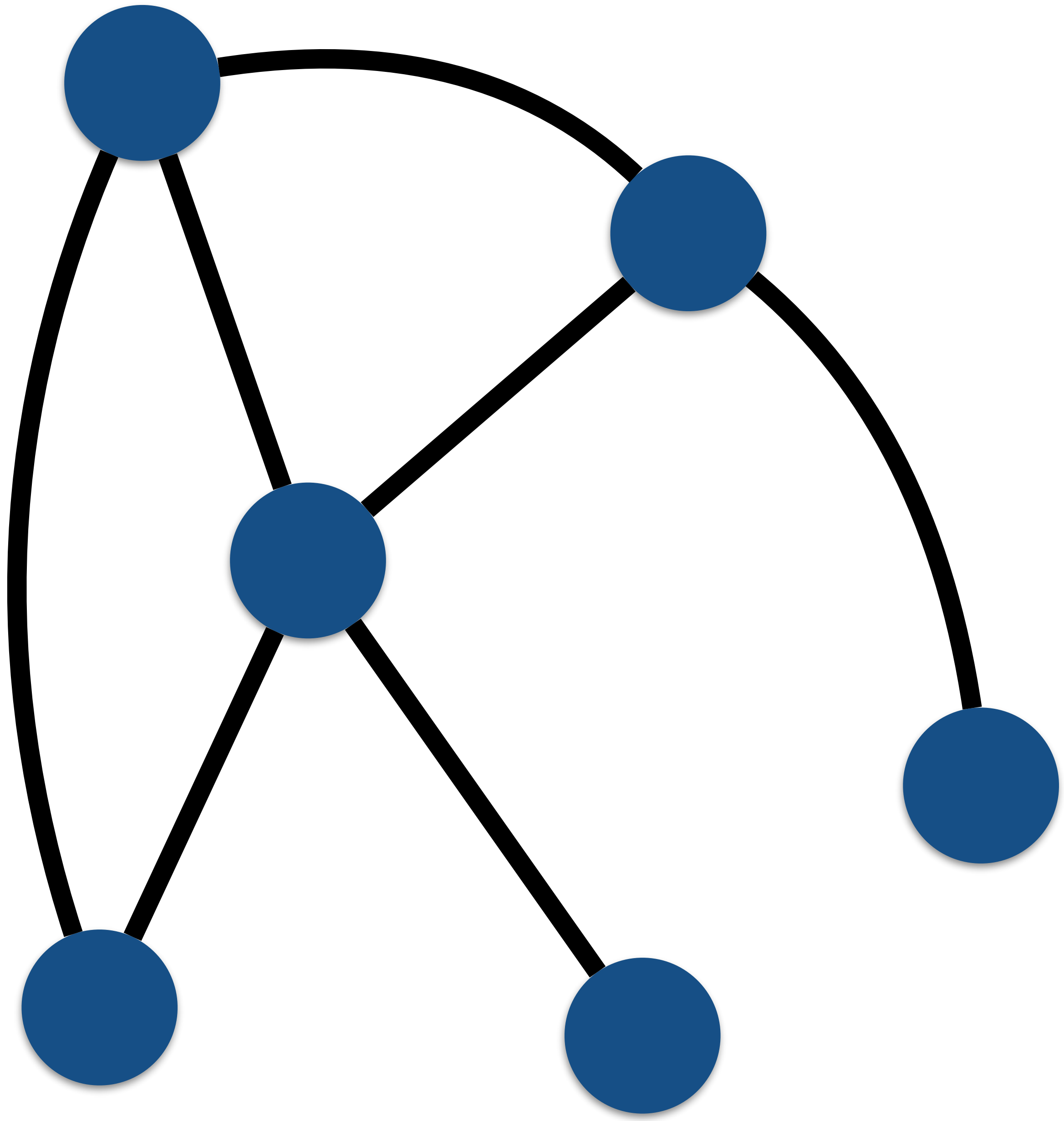
Cloud systems

- Frequent opening of connections between PoPs.
- In a perfect world, would have a mesh.
- Application and resource constraints limit reuse.



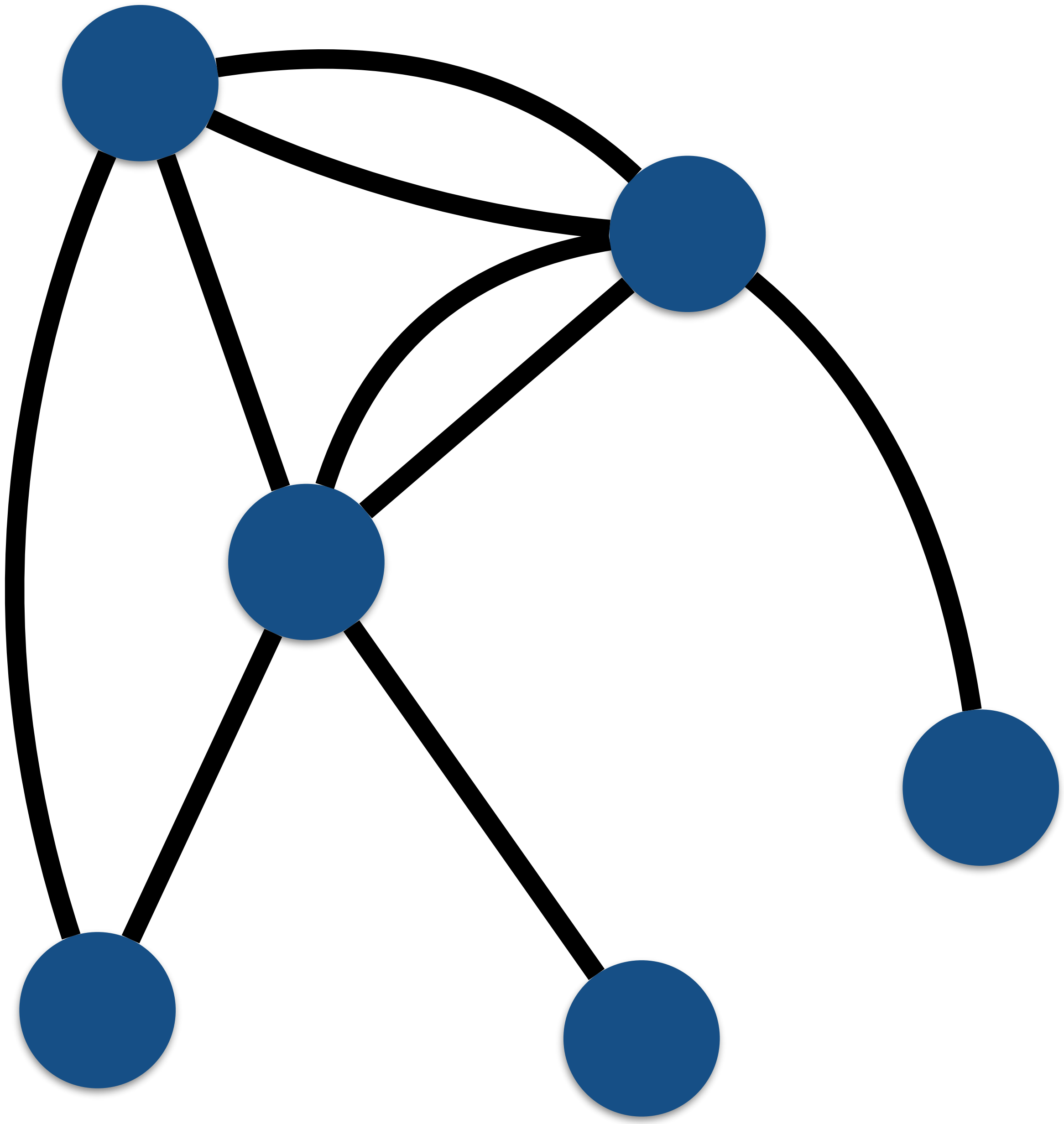
Cloud systems

- Frequent opening of connections between PoPs.
- In a perfect world, would have a mesh.
- Application and resource constraints limit reuse.



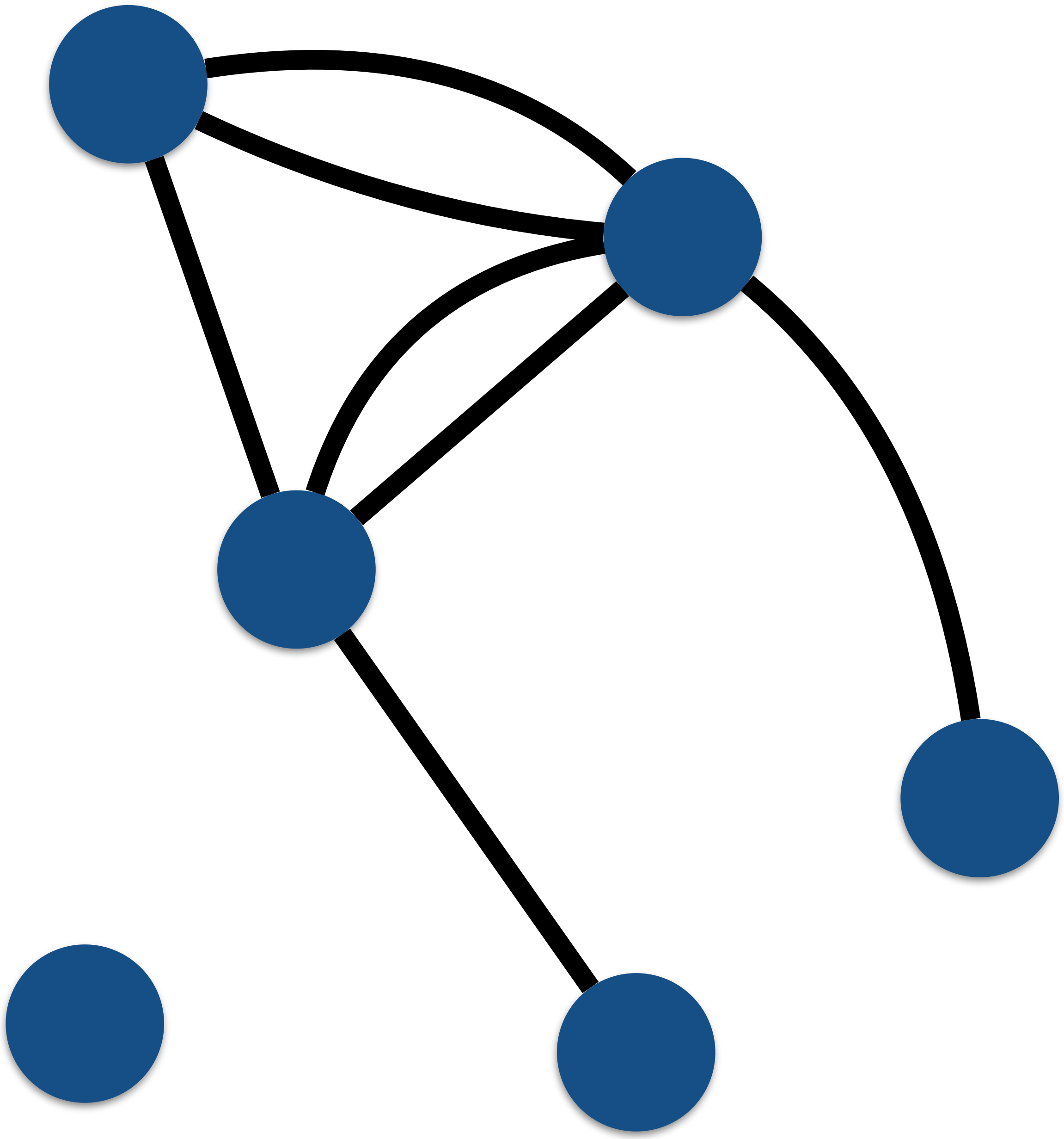
Cloud systems

- Frequent opening of connections between PoPs.
- In a perfect world, would have a mesh.
- Application and resource constraints limit reuse.

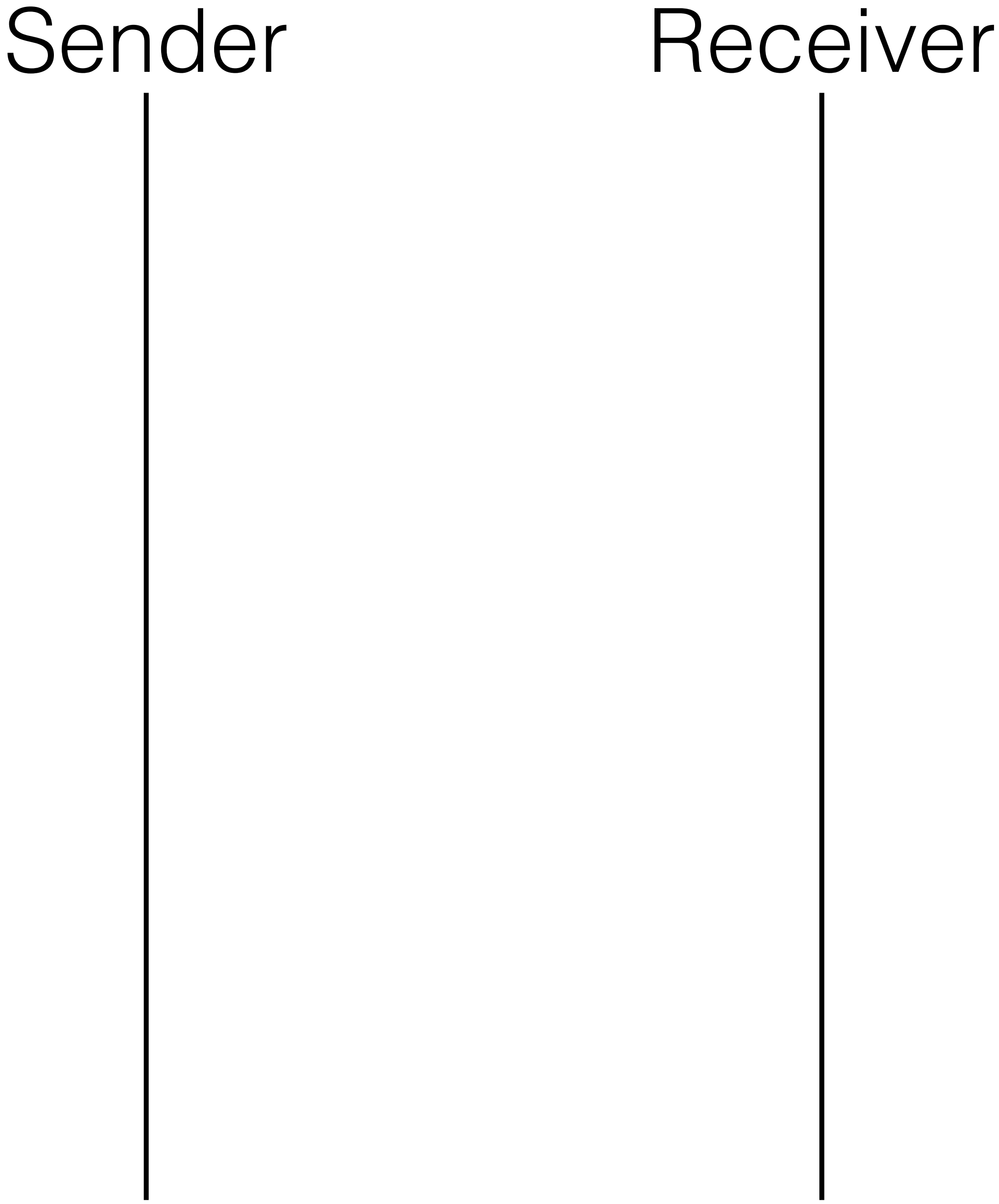


Cloud systems

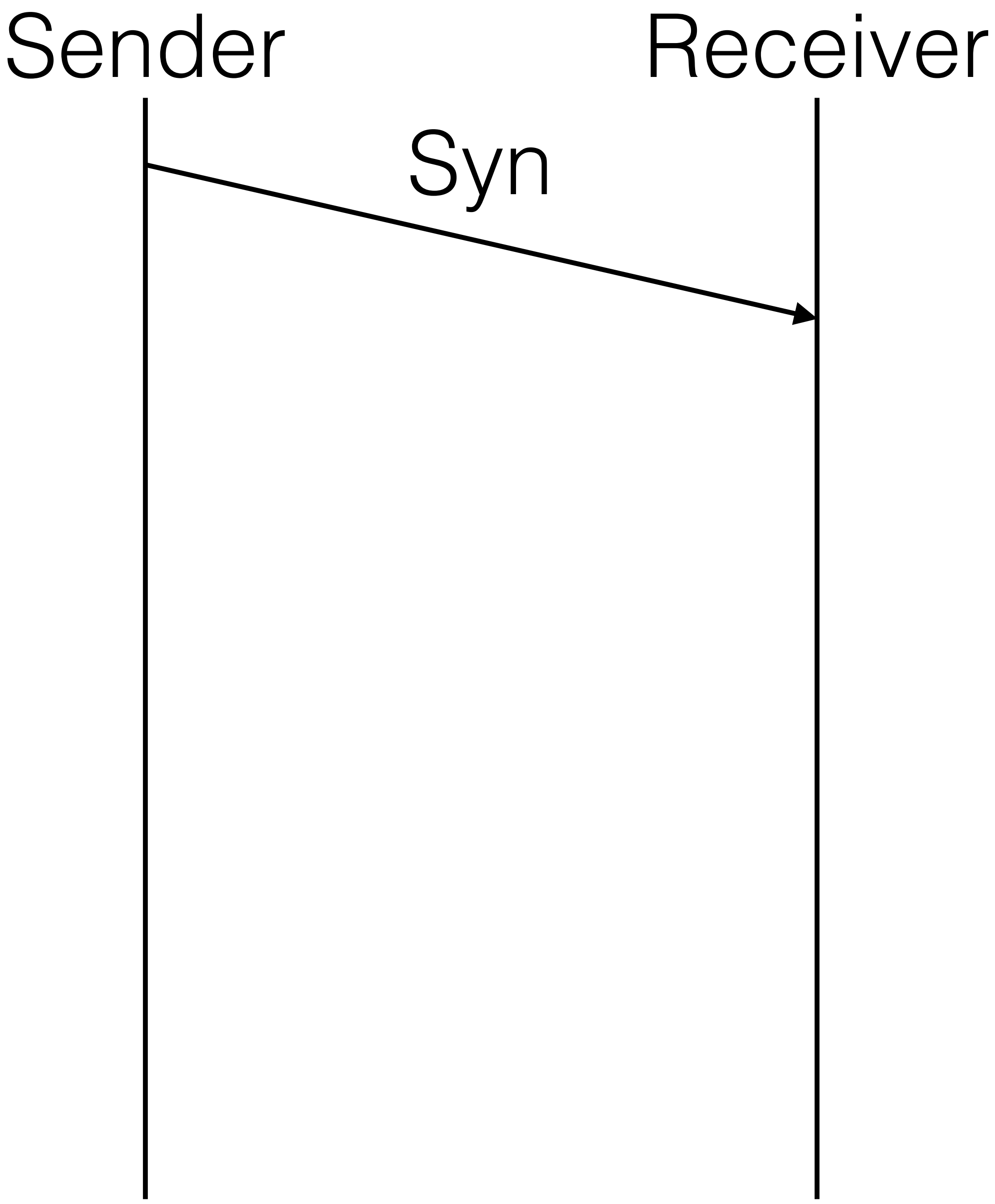
- Frequent opening of connections between PoPs.
- In a perfect world, would have a mesh.
- Application and resource constraints limit reuse.



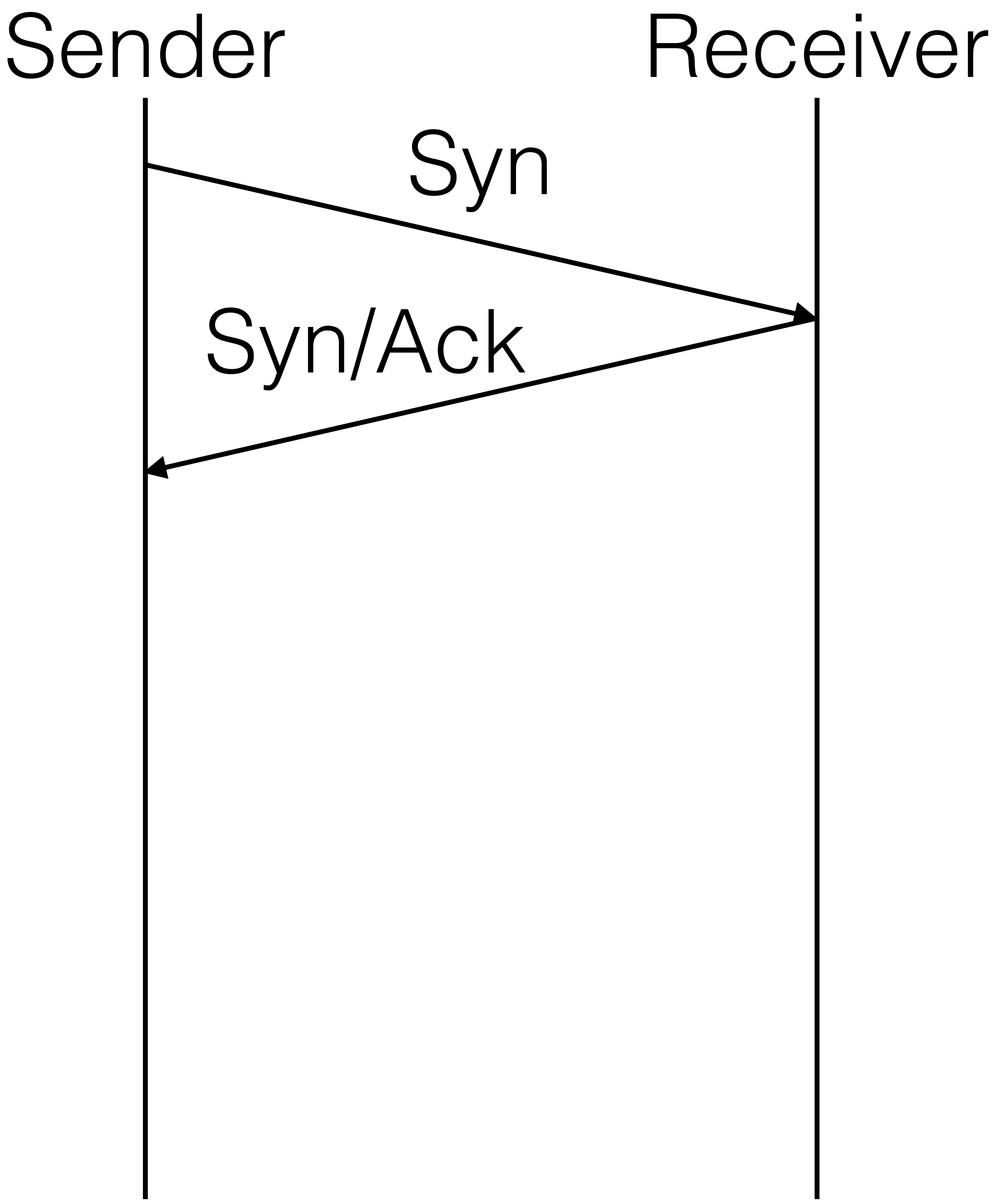
Slow-start penalty



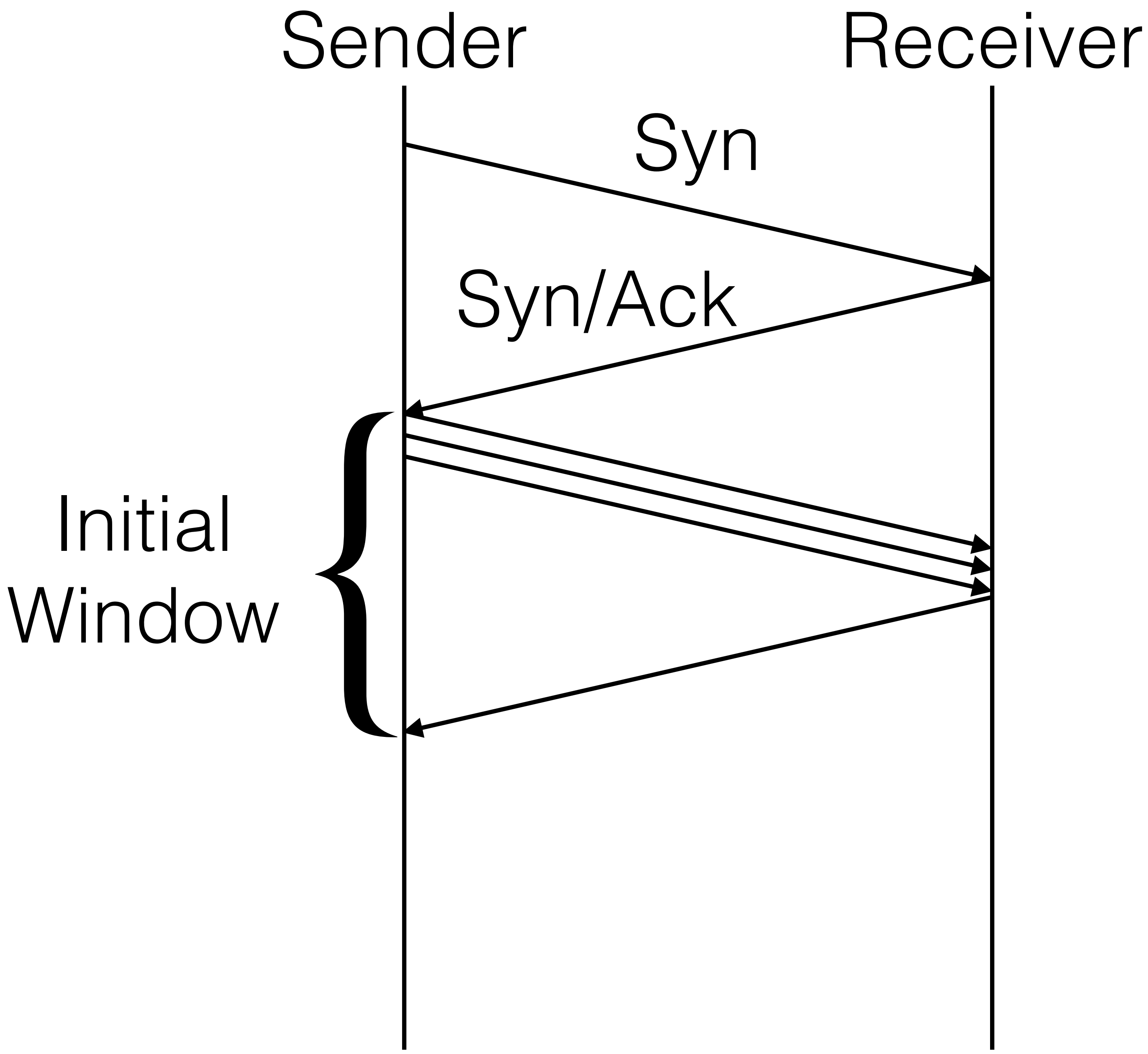
Slow-start penalty



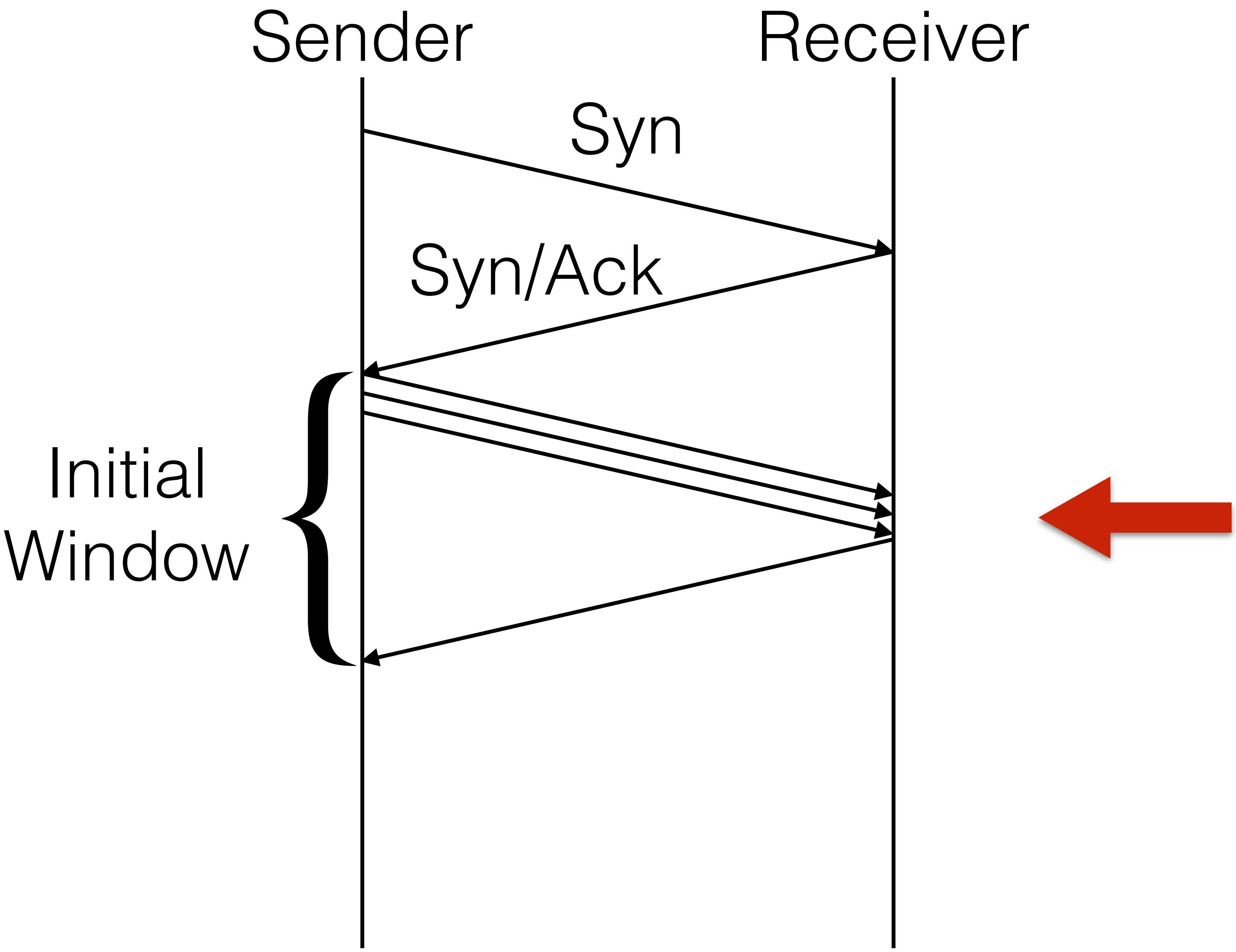
Slow-start penalty



Slow-start penalty

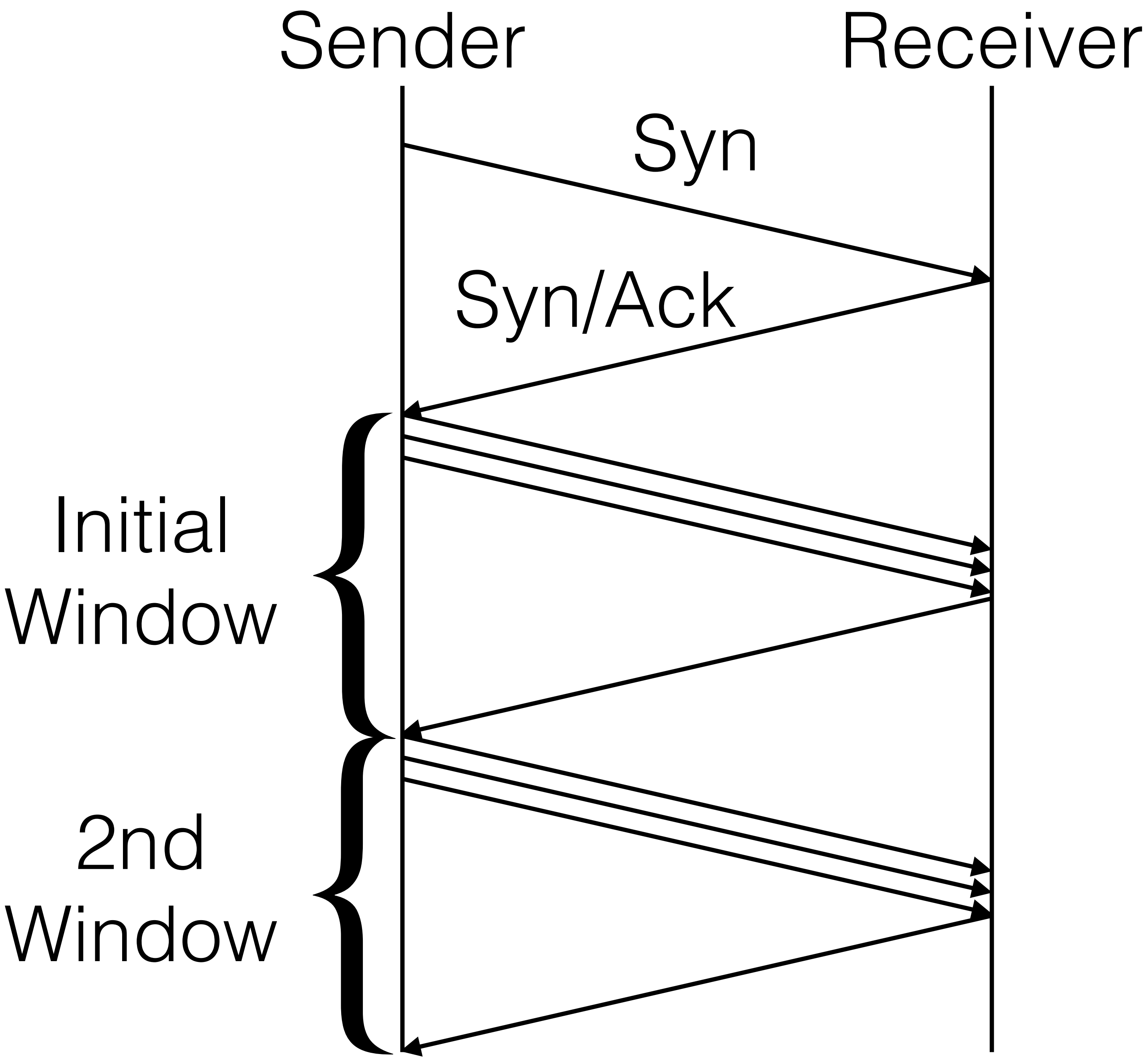


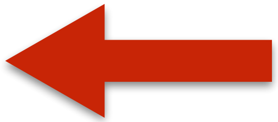
Slow-start penalty



Data doesn’t fit

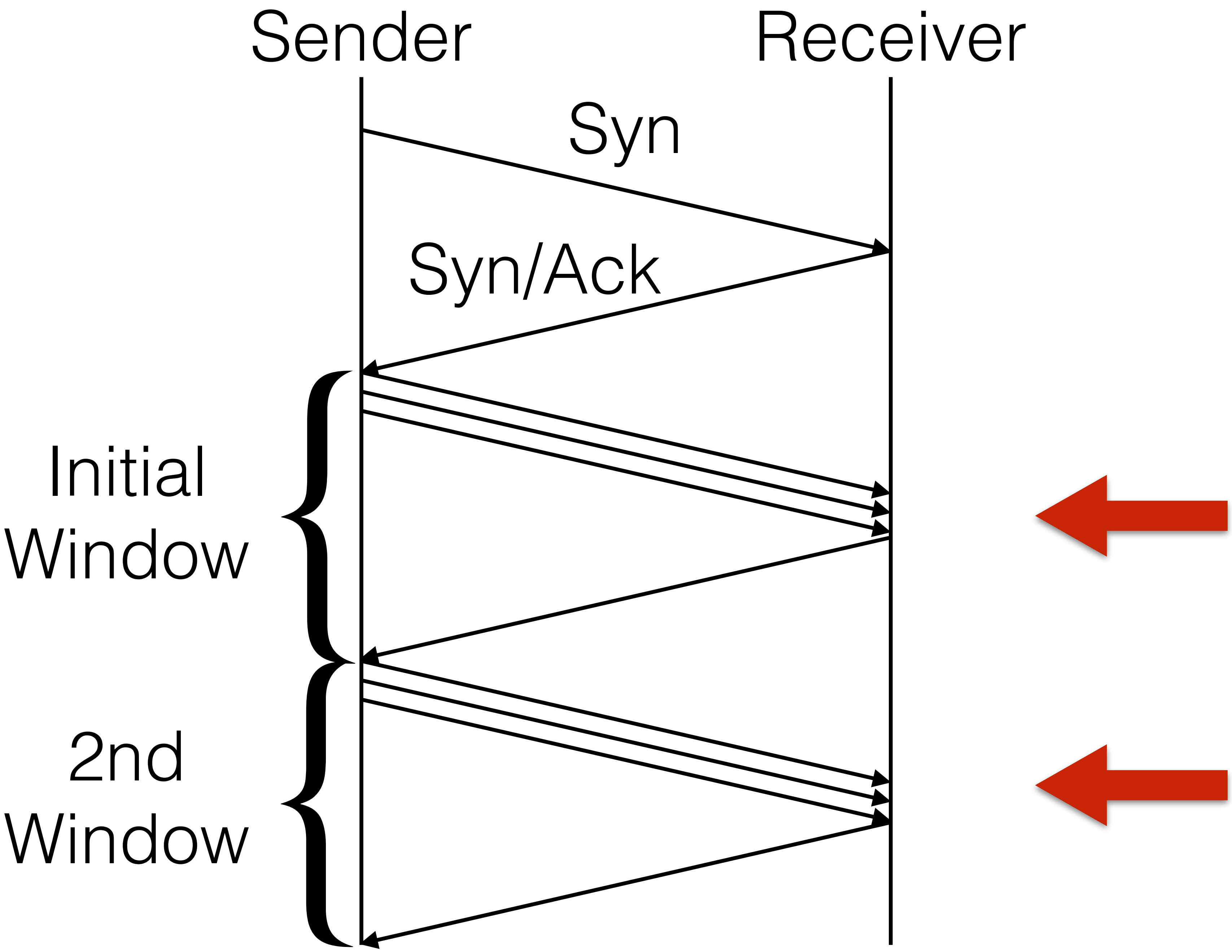
Slow-start penalty





Data doesn’t fit

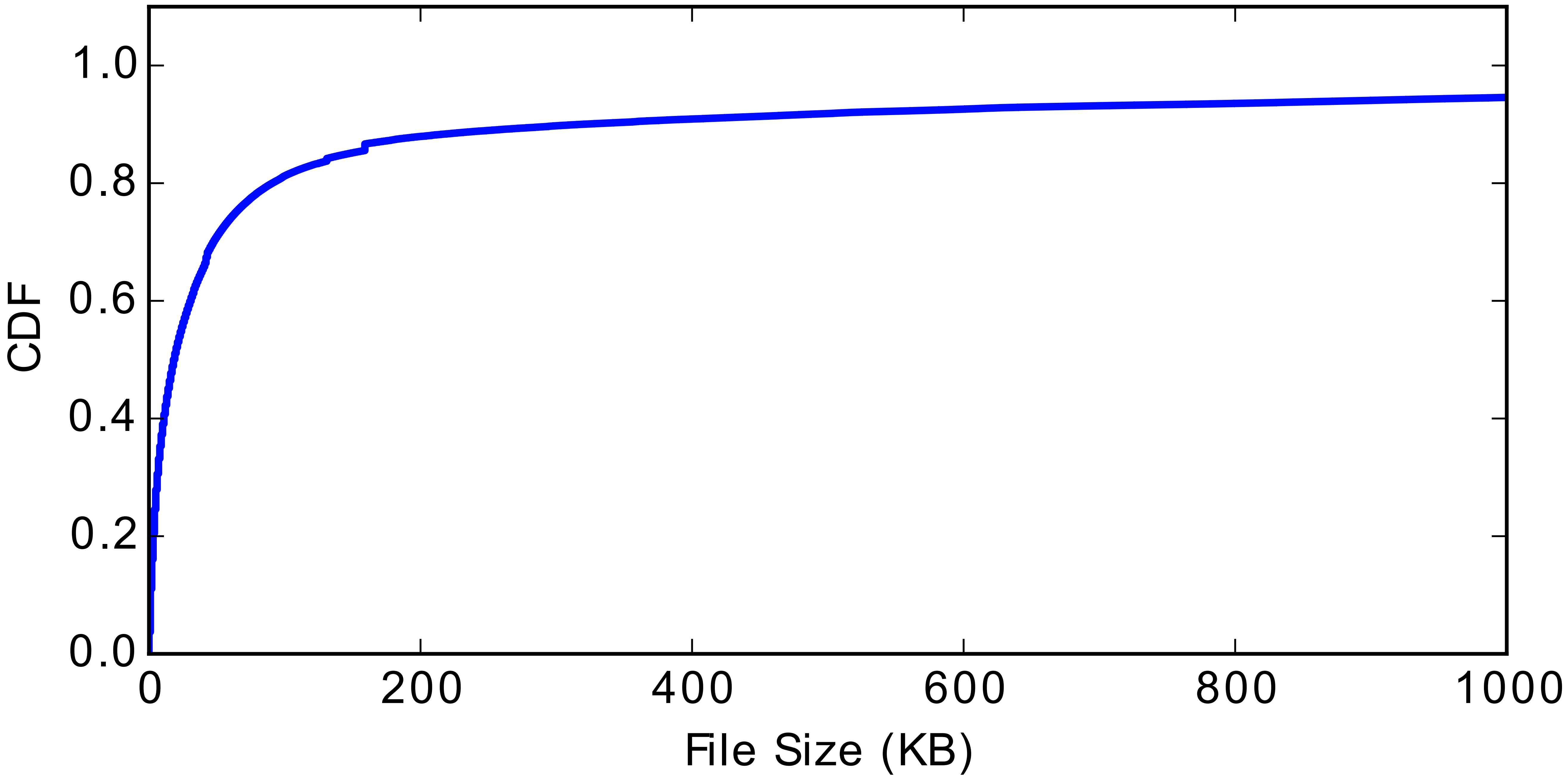
Slow-start penalty



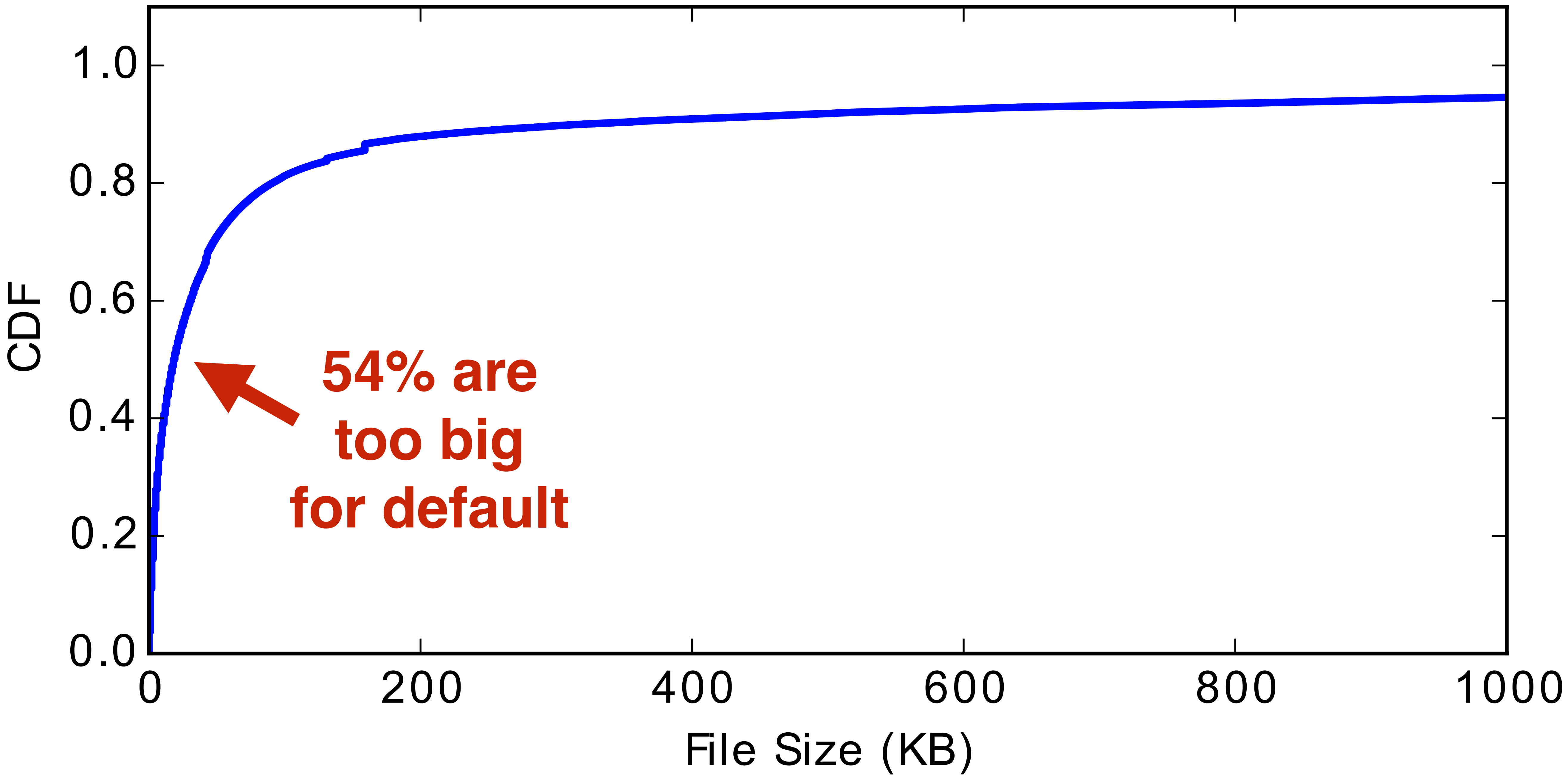
Data doesn’t fit

Transfer pays second RTT

Cloud workloads

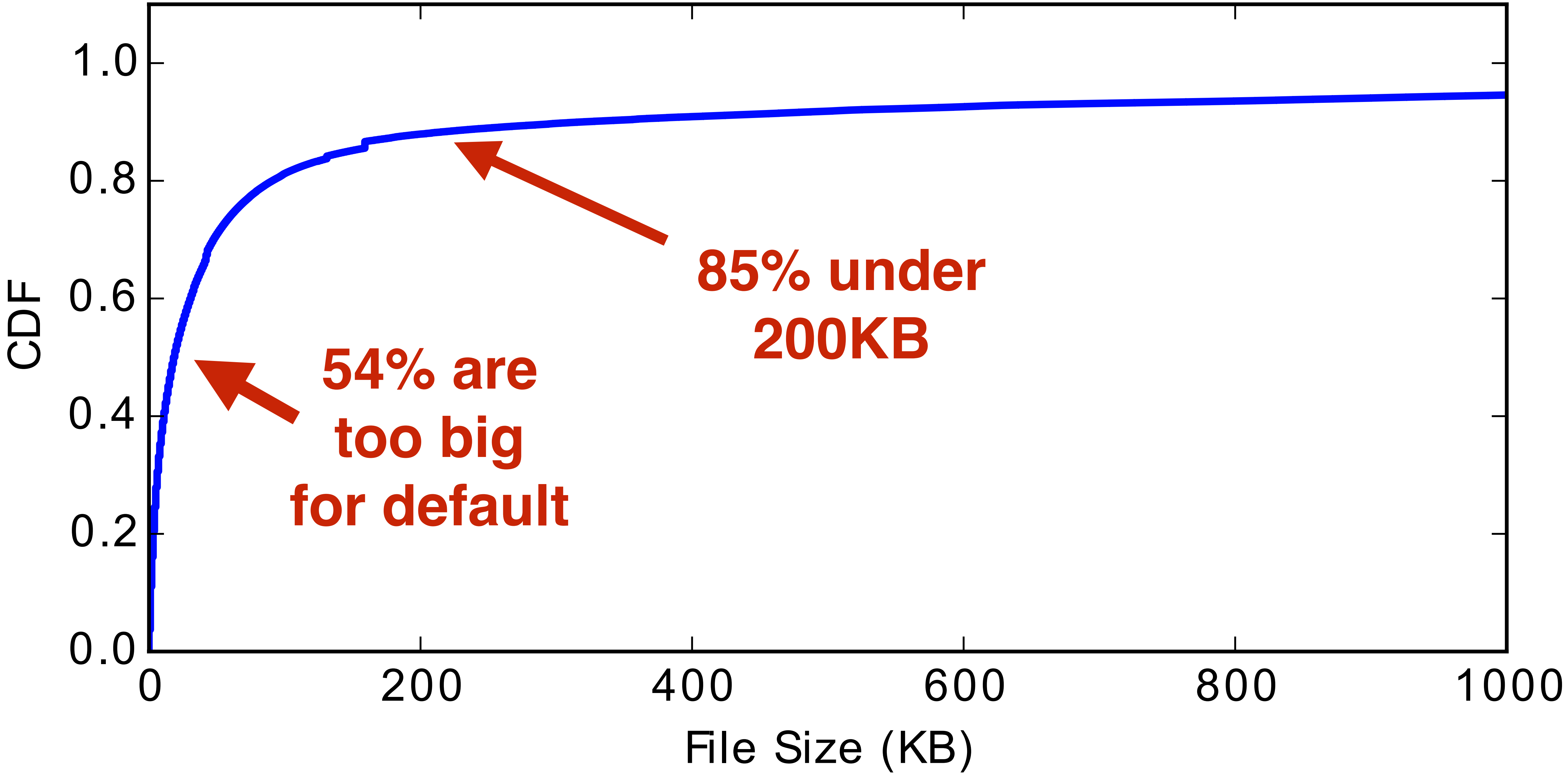


Cloud workloads



54% are too big for default

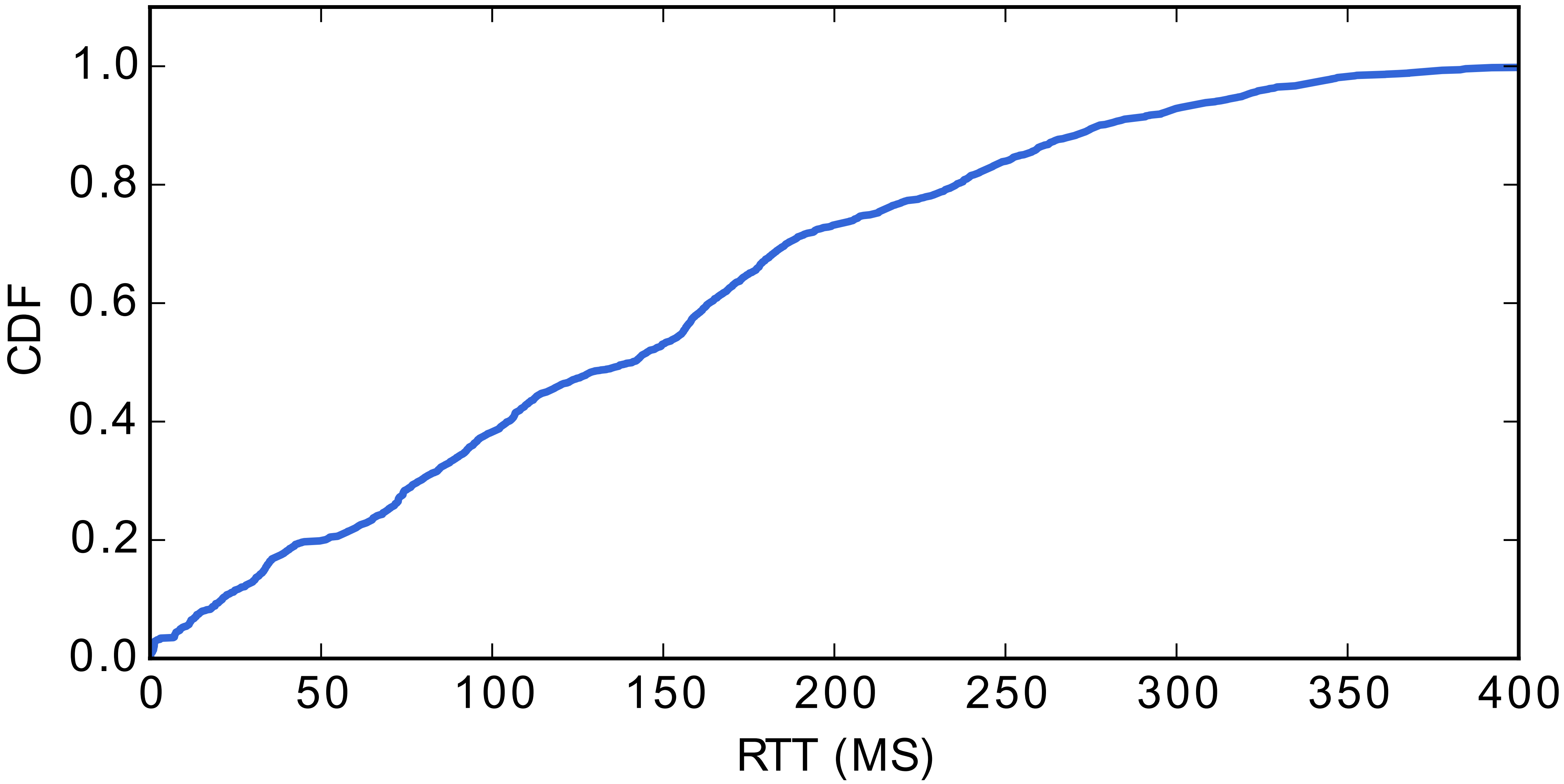
Cloud workloads



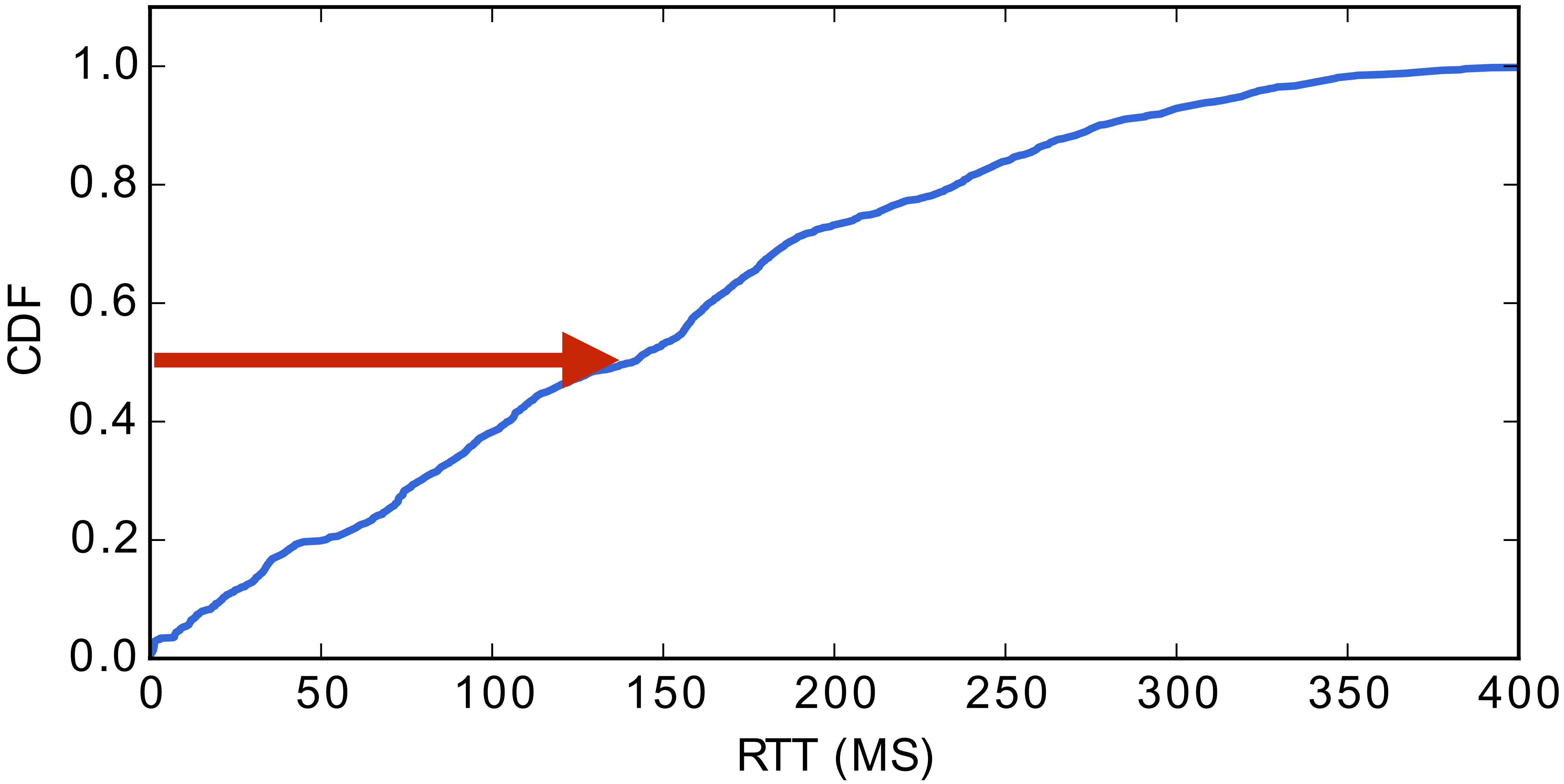
85% under 200KB

54% are too big for default

Global deployments



Global deployments



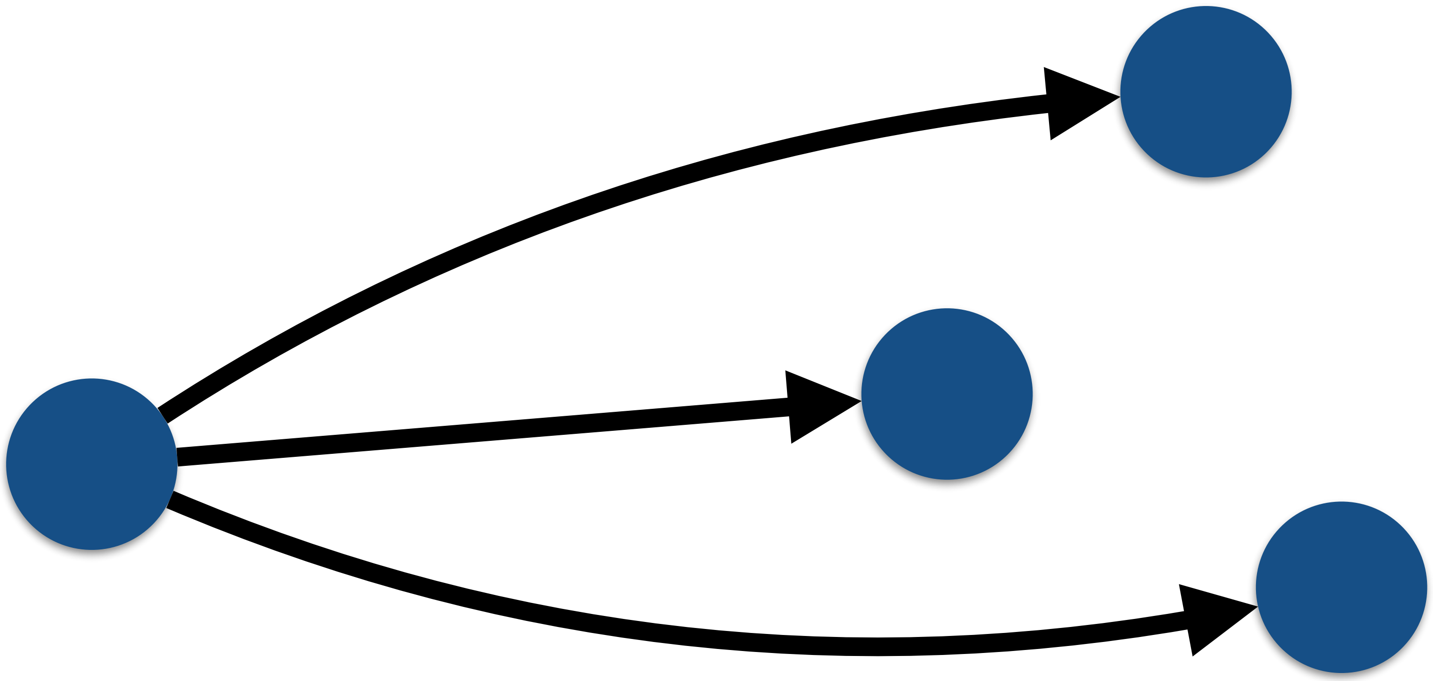
Median RTT is over 125 ms

Global deployments

- Can’t just blindly increase the congestion window on a global deployment.
    - Would risk significant loss.

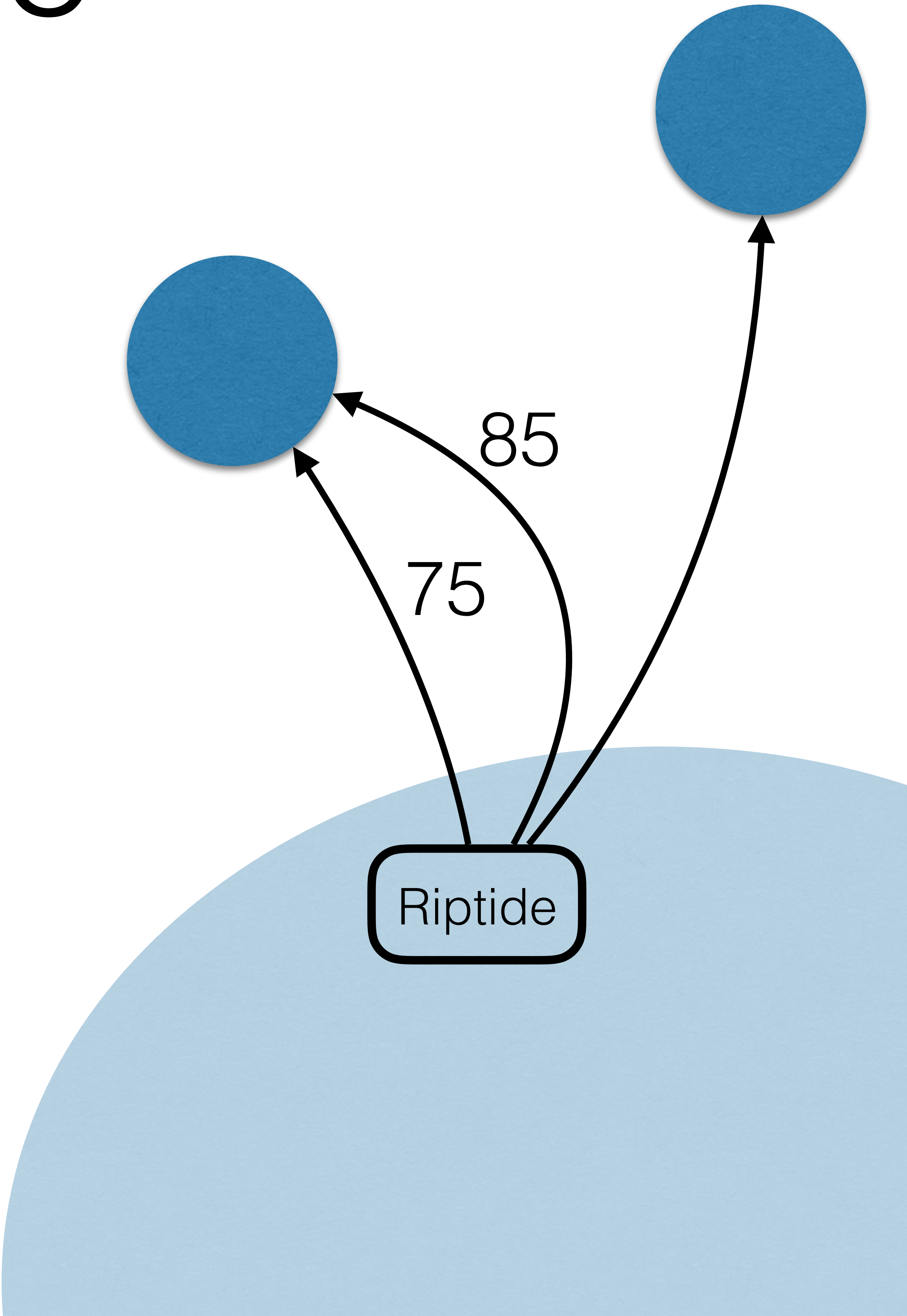
Riptide

- Observes current congestion windows.
- New connections set initial window to a known-safe level.
- Operates in a totally standalone manner.



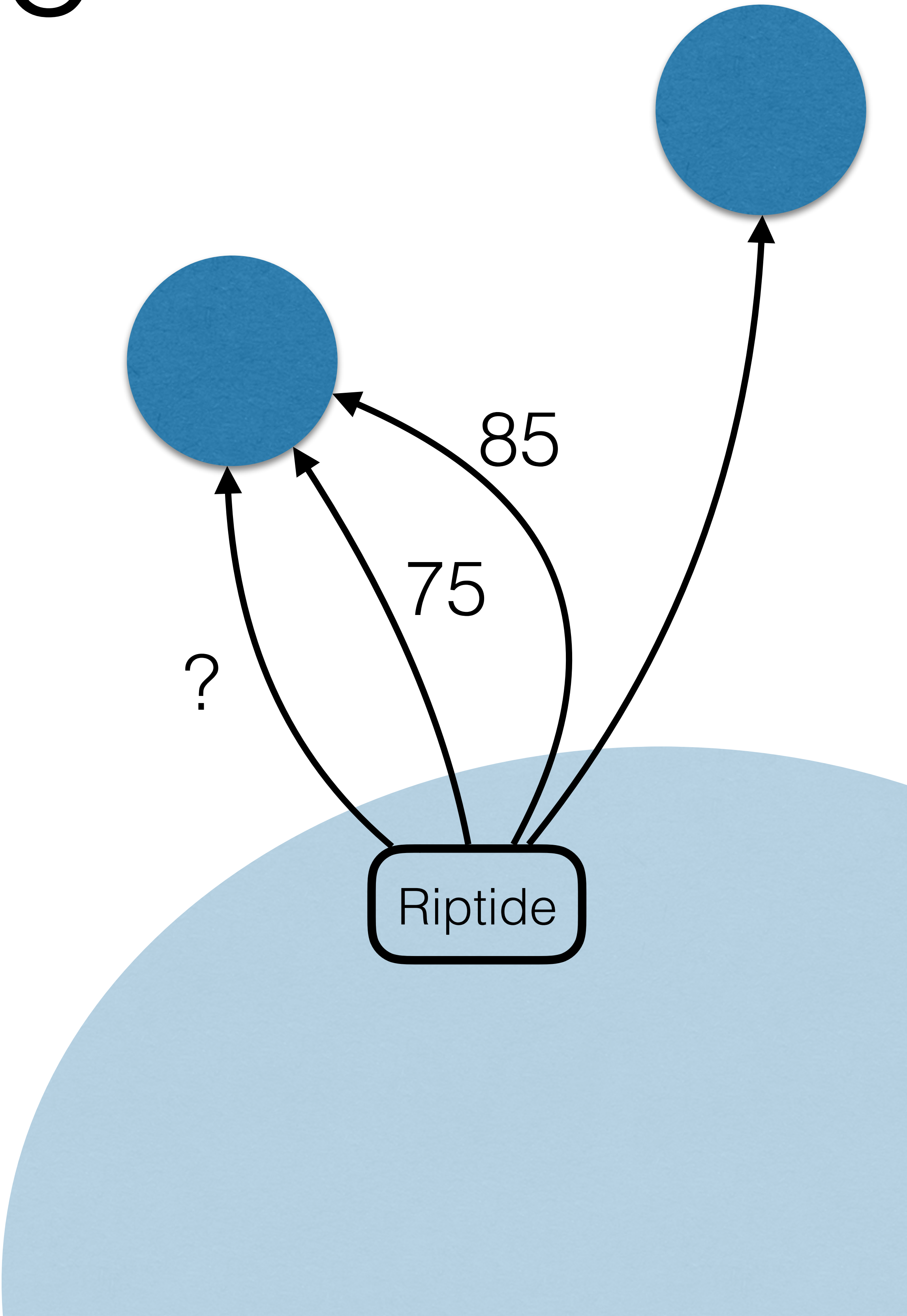
Riptide

- Riptide observes CWND for all open connections to a destination.
- New connections will be opened with INIT_CWND set to the average of existing windows.



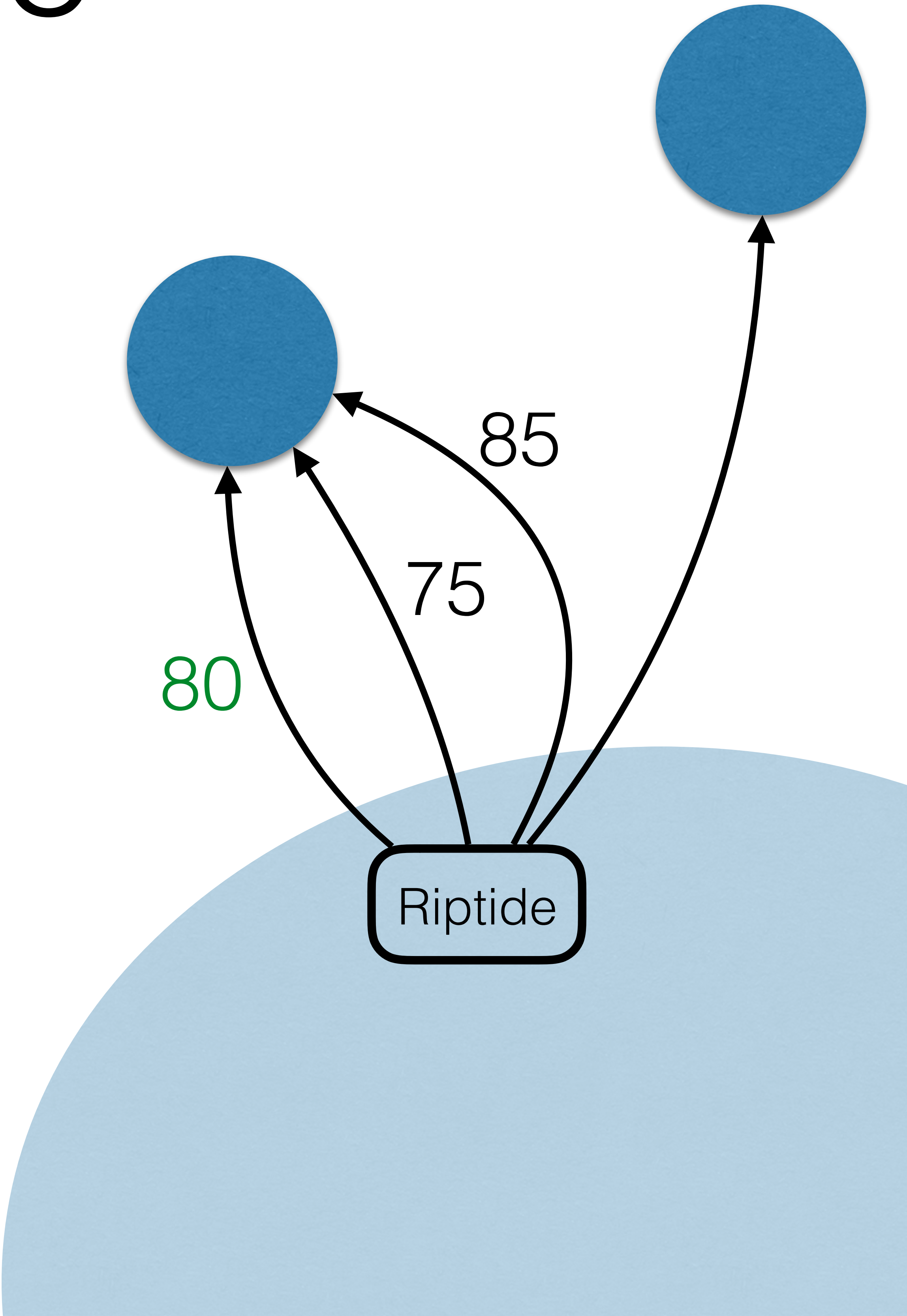
Riptide

- Riptide observes CWND for all open connections to a destination.
- New connections will be opened with INIT_CWND set to the average of existing windows.



Riptide

- Riptide observes CWND for all open connections to a destination.
- New connections will be opened with INIT_CWND set to the average of existing windows.

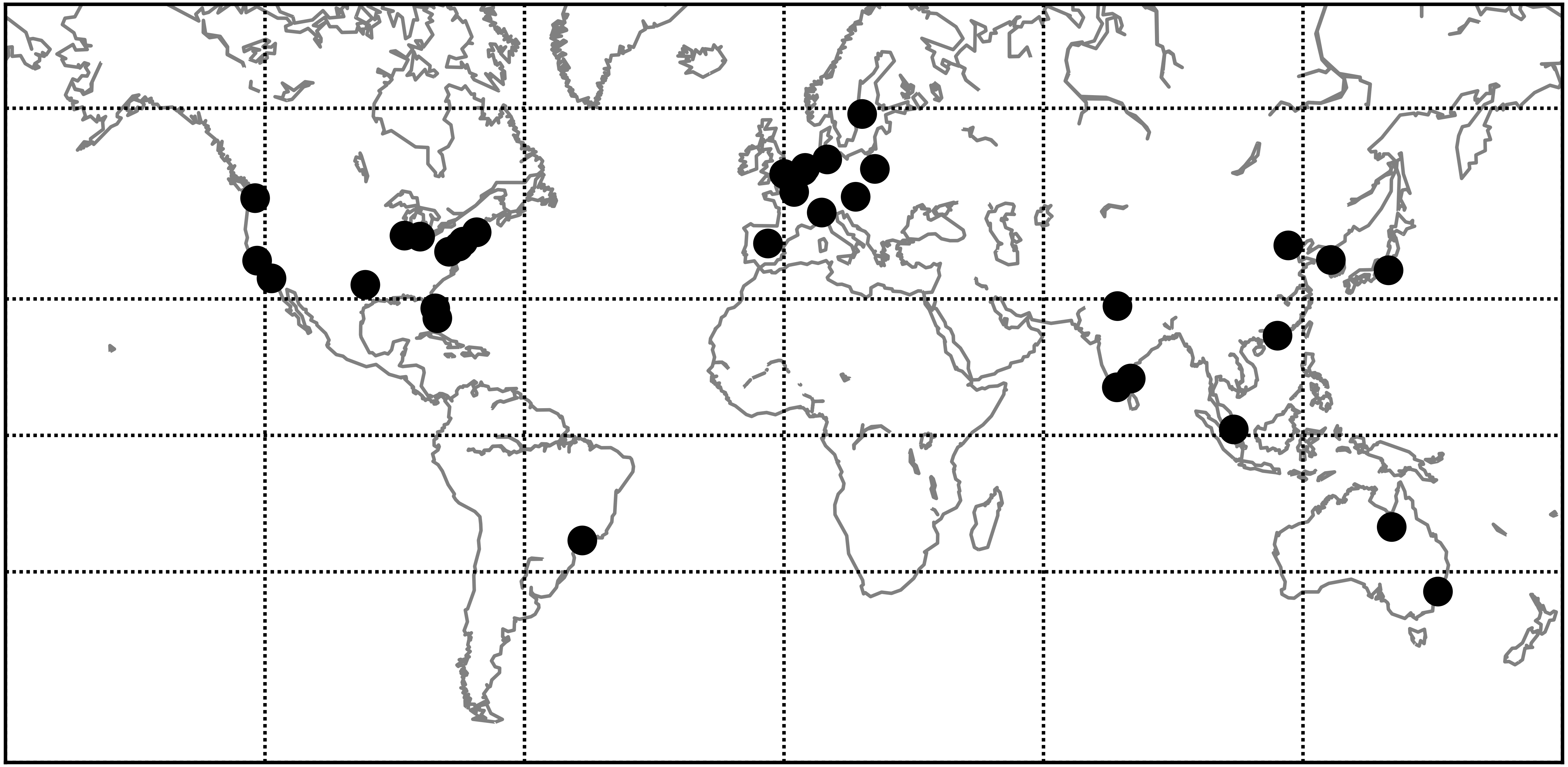


Implementation

- Implemented as a Python script in user space.
- Use the `ss` tool to observe existing windows.
- Polls current connections once per second.
- Sets new windows via `ip route` interface.

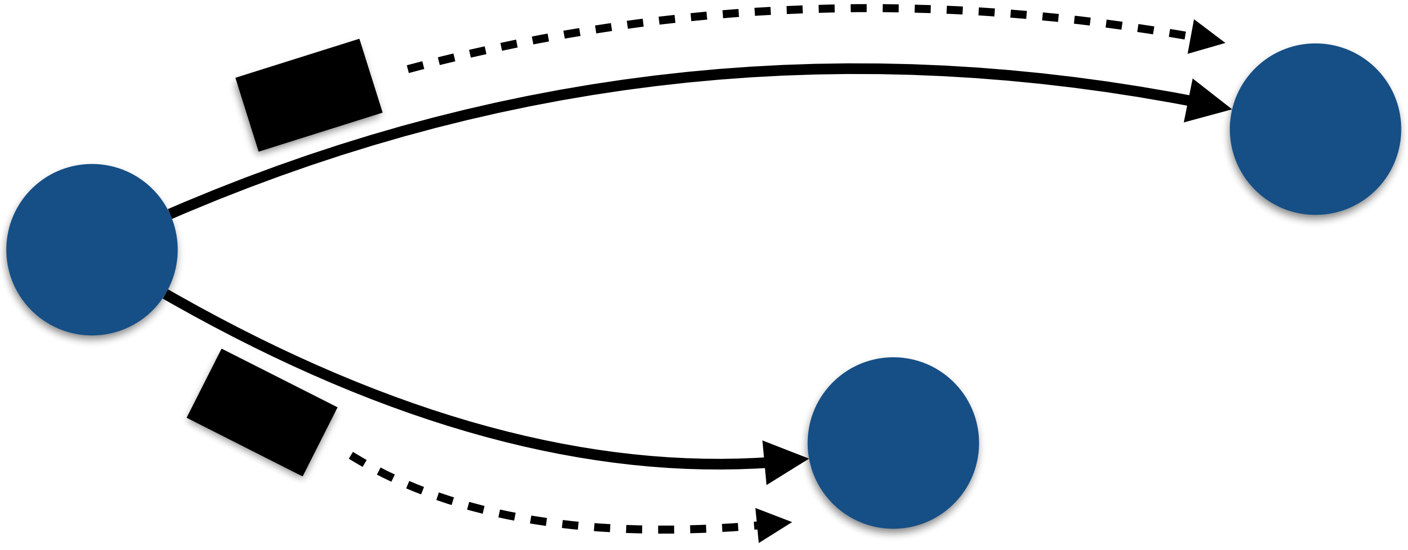
```
ip route add 10.0.0.127 dev eth0 proto \\
     static initcwnd 80 via 10.0.0.1
```

Riptide Deployment

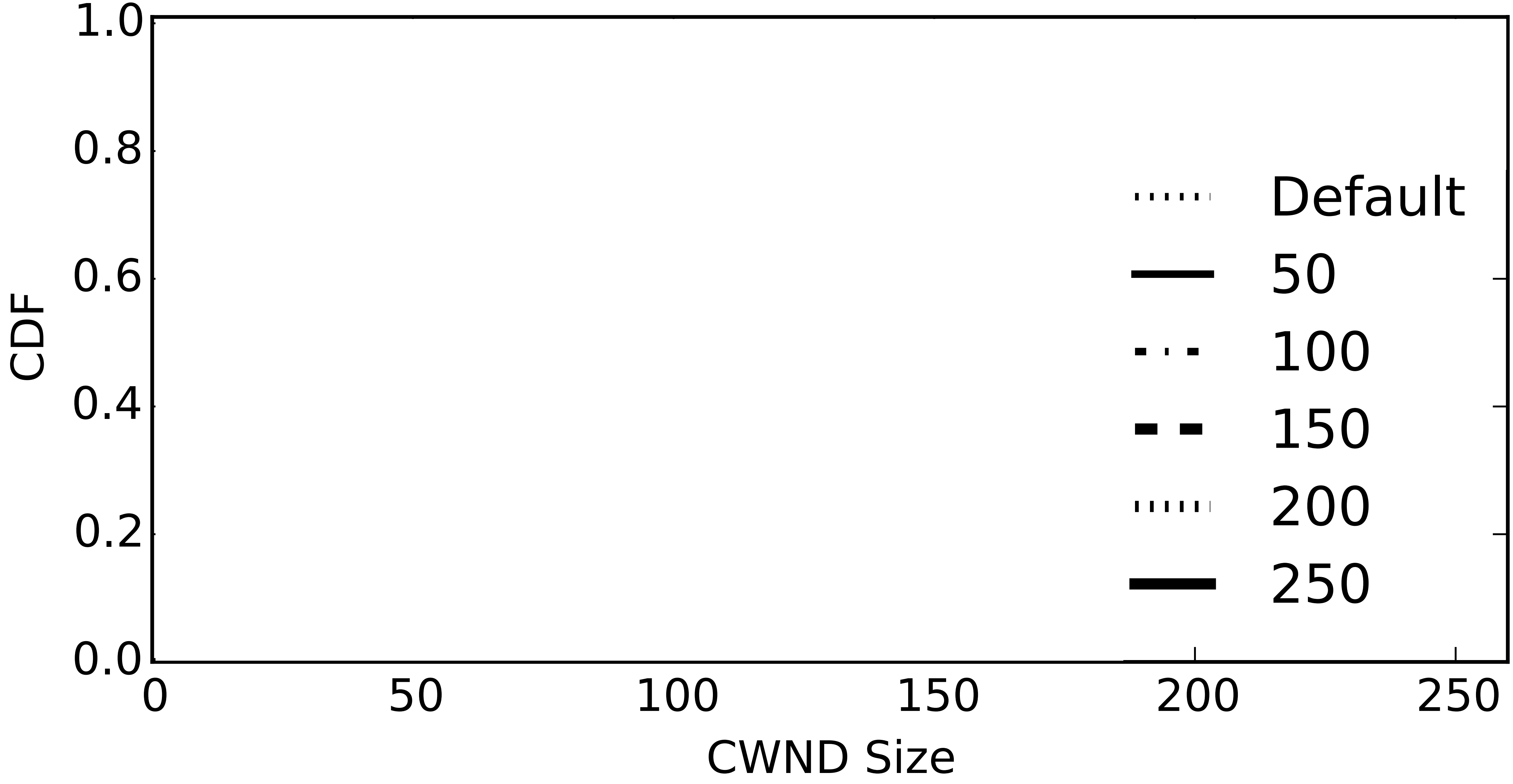


Probes

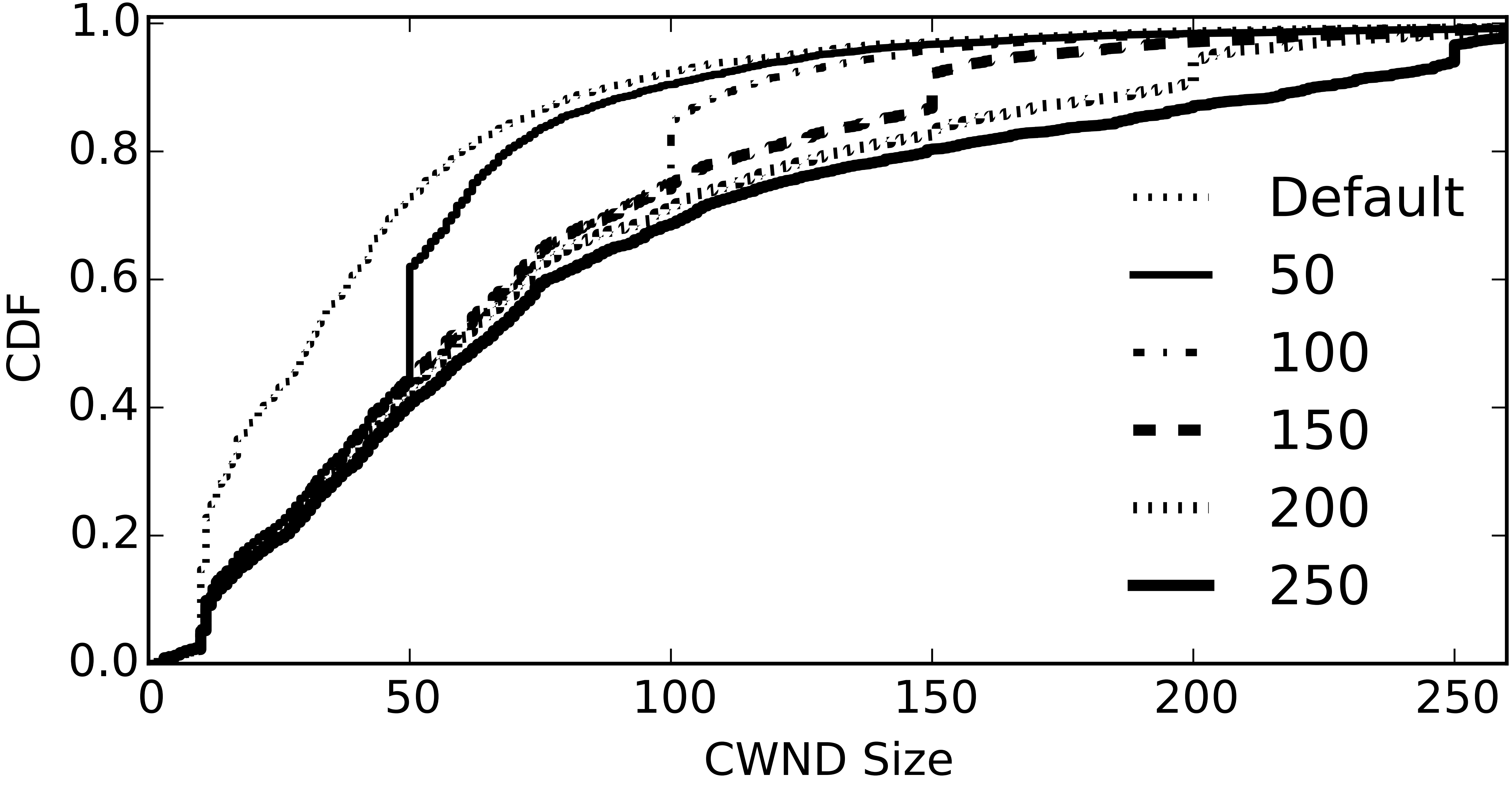
- To test the current state of the network, send hourly probes between PoPs.
- Currently employ 10K, 50K, 100K probes.



Observed windows

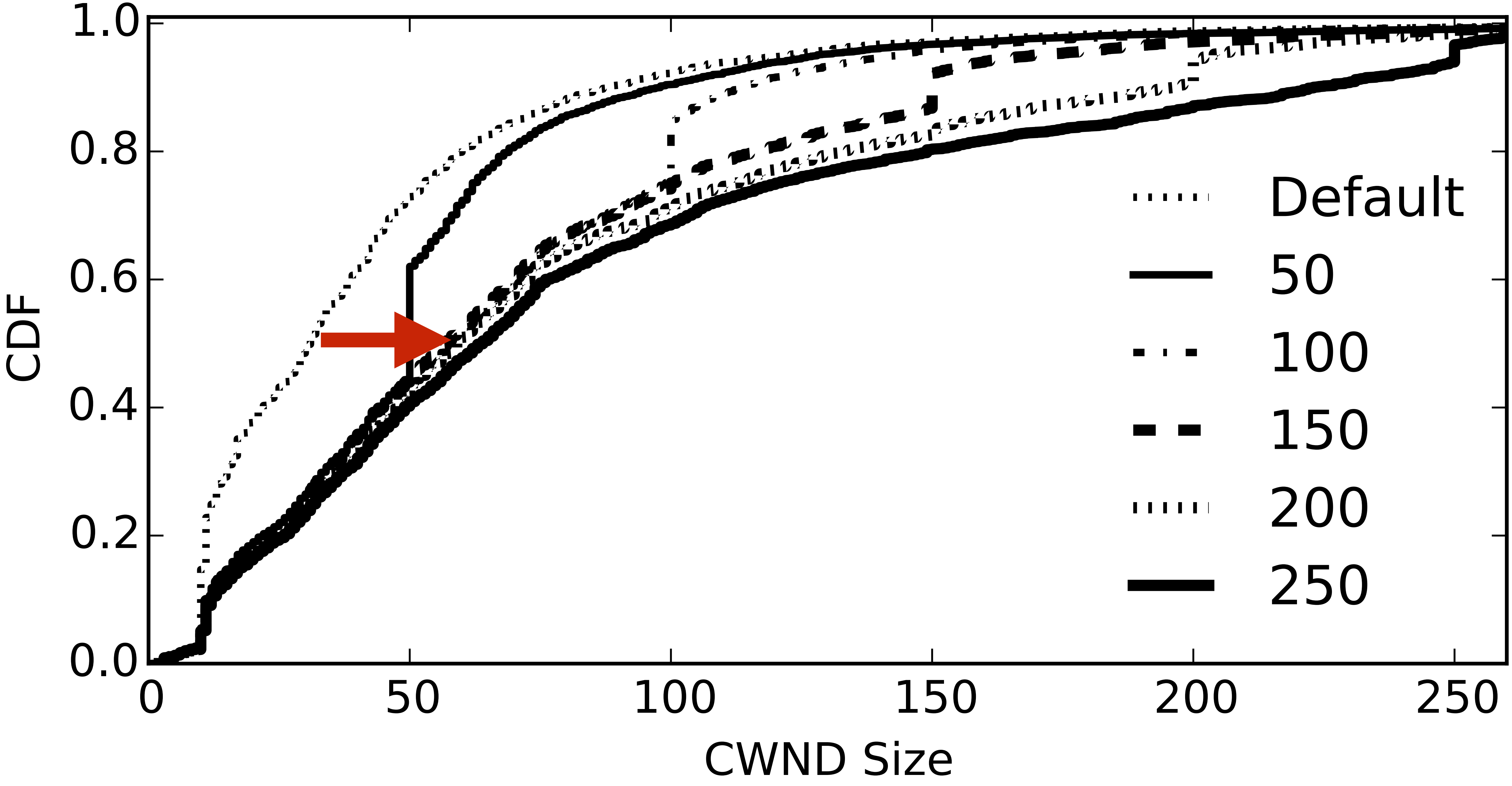


Observed windows



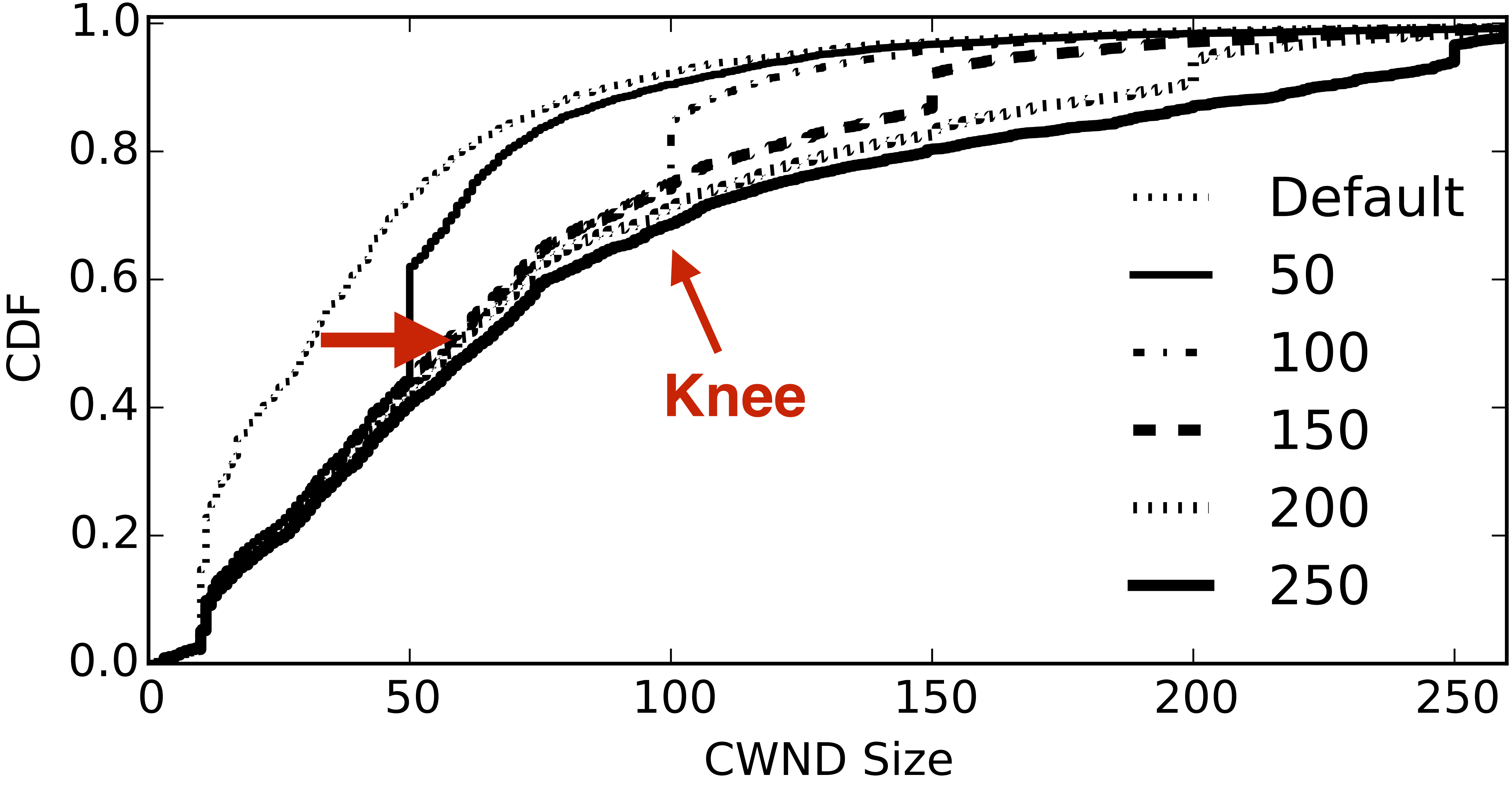
CWND windows significantly higher.

Observed windows



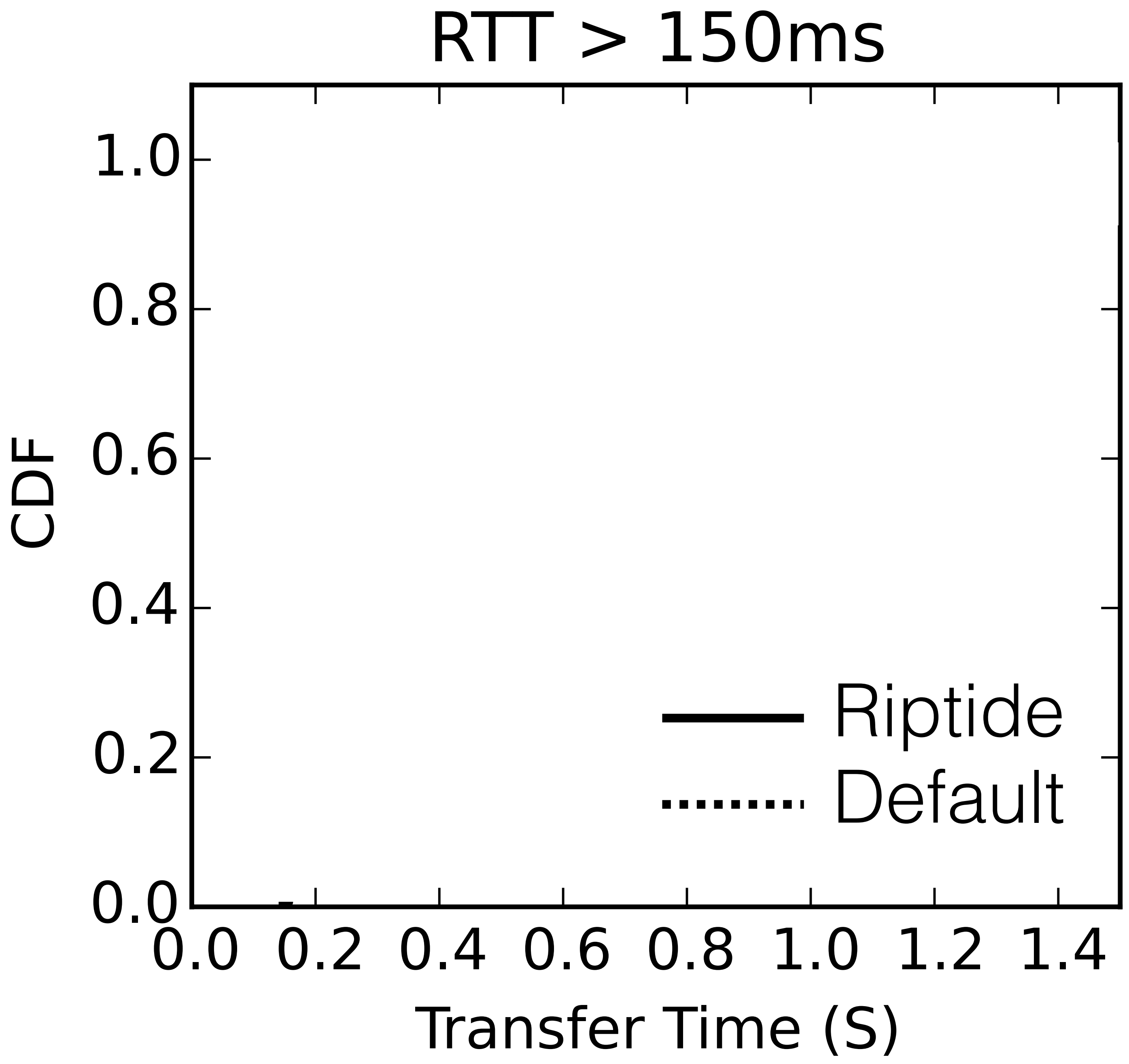
CWND windows significantly higher.

Observed windows



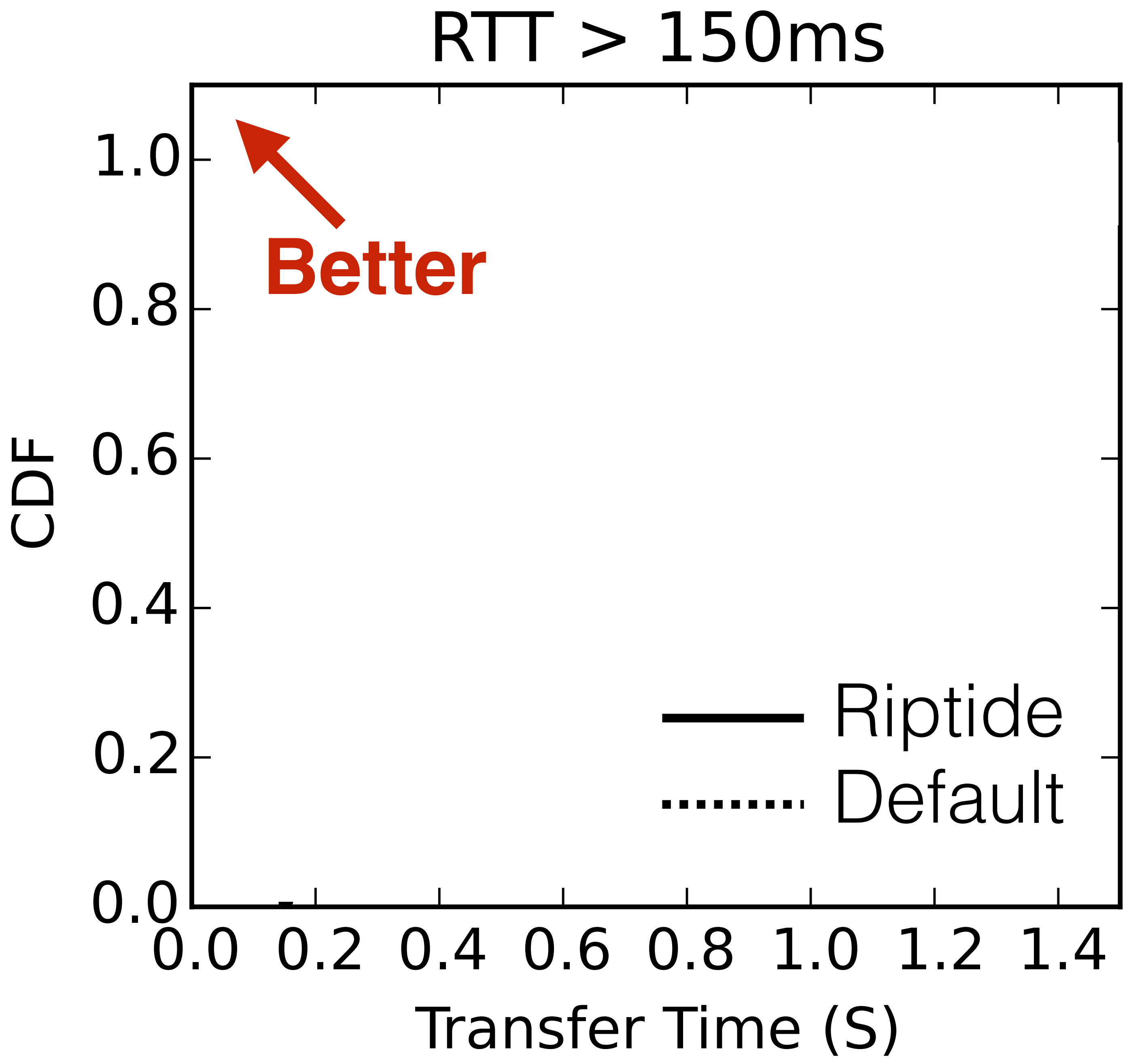
CWND windows significantly higher.

Probe completion times



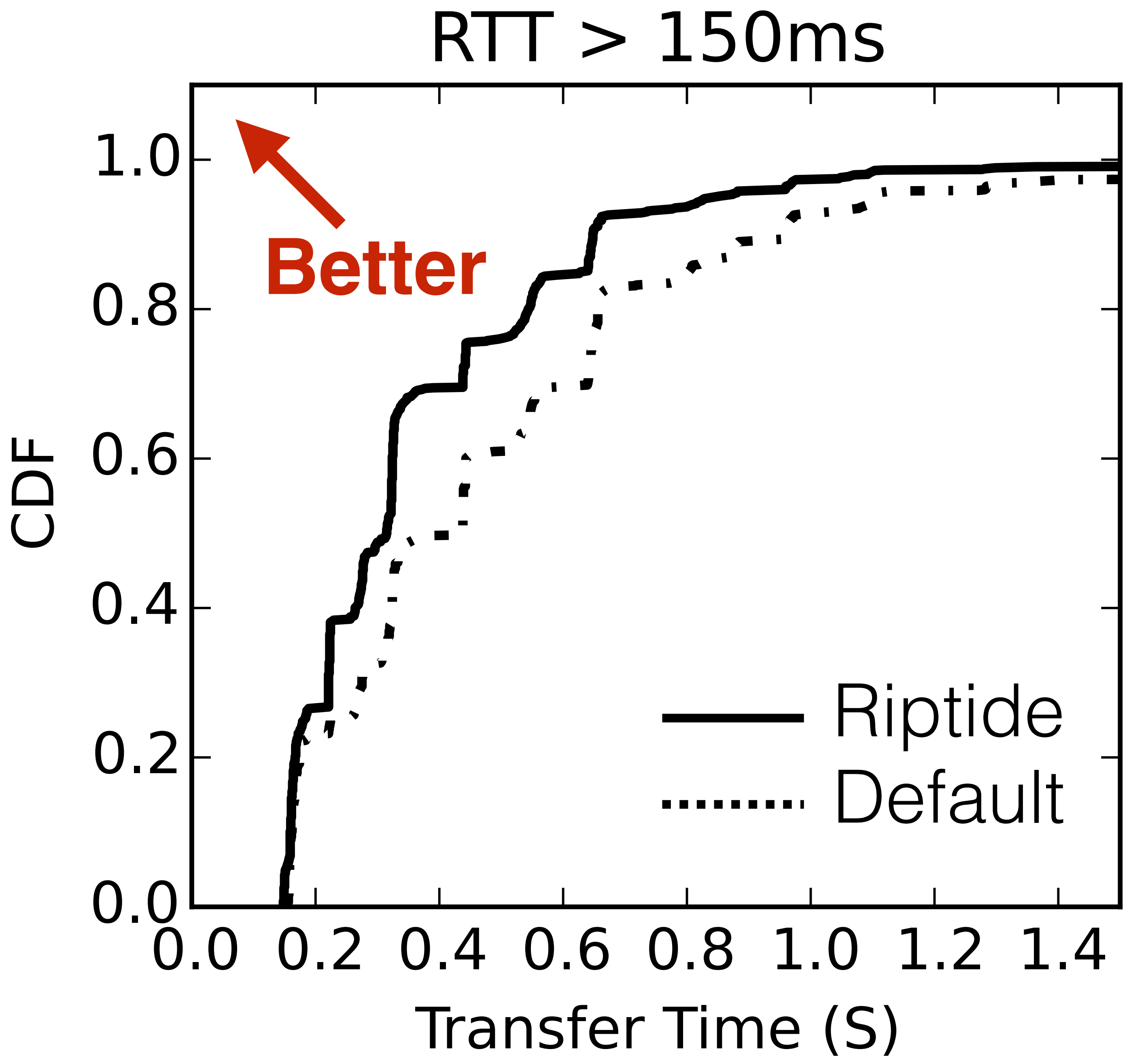
- Clients are able to complete the probe transfers in fewer round trips.
- Reduces total transfer time.

Probe completion times



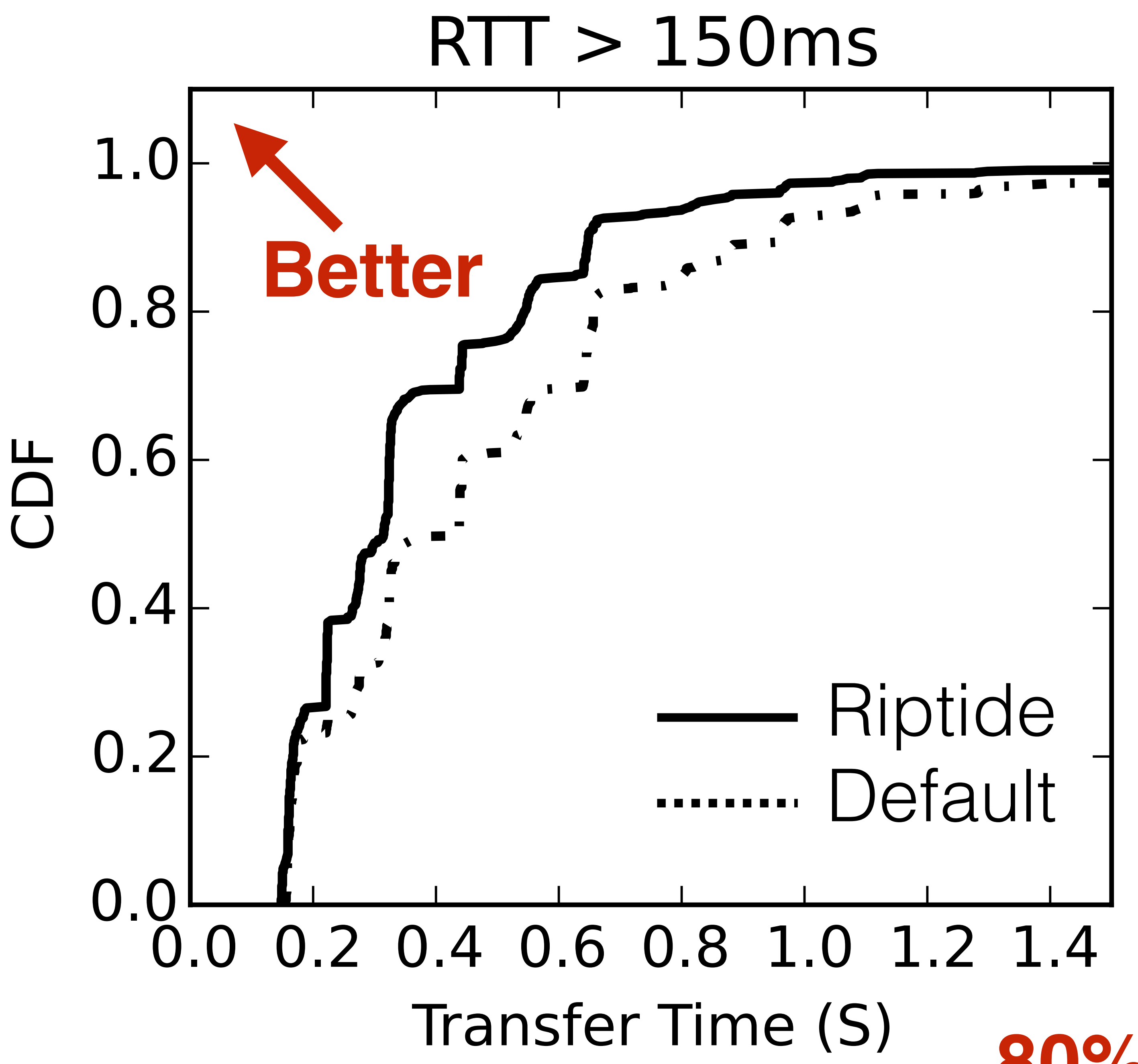
- Clients are able to complete the probe transfers in fewer round trips.
- Reduces total transfer time.

Probe completion times



- Clients are able to complete the probe transfers in fewer round trips.
- Reduces total transfer time.

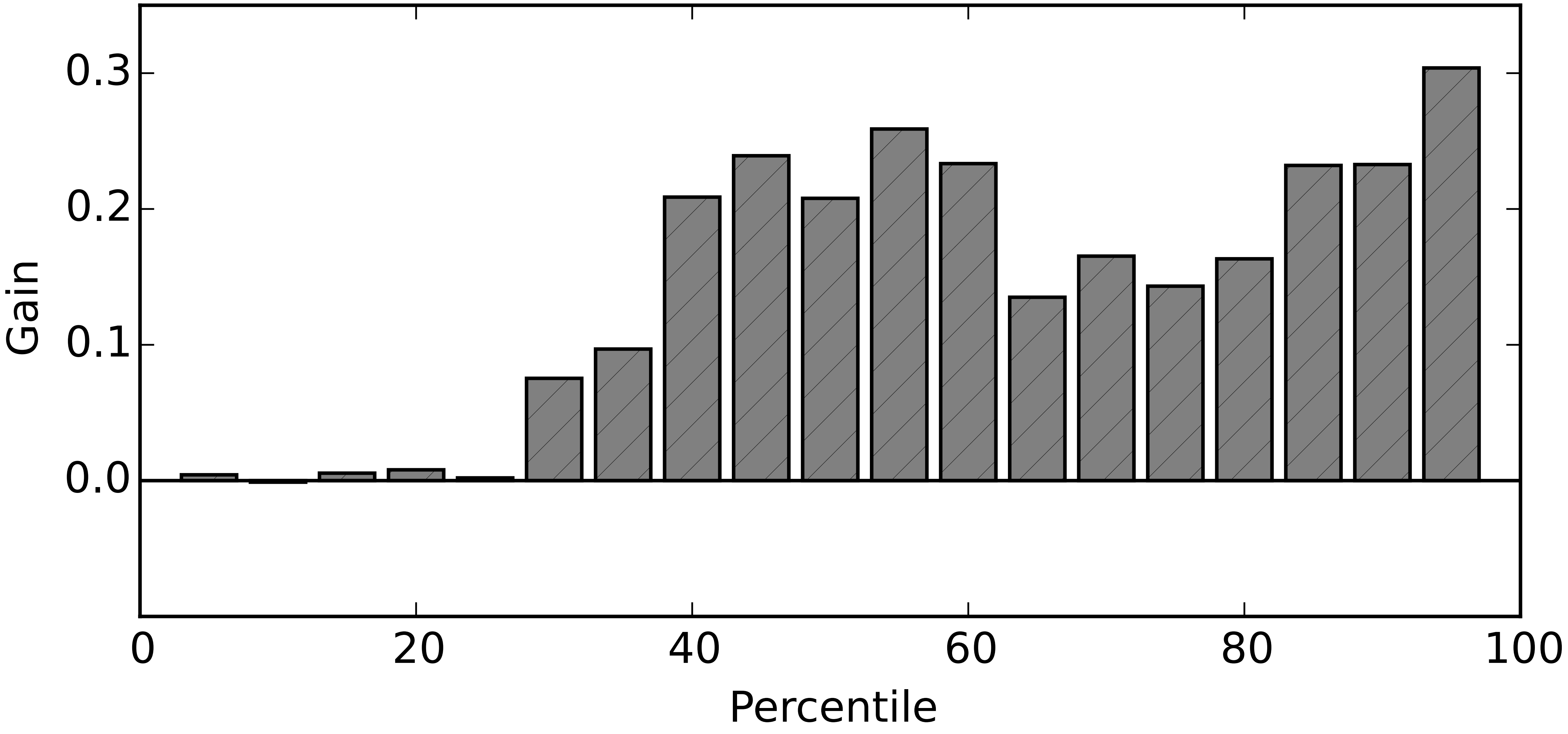
Probe completion times



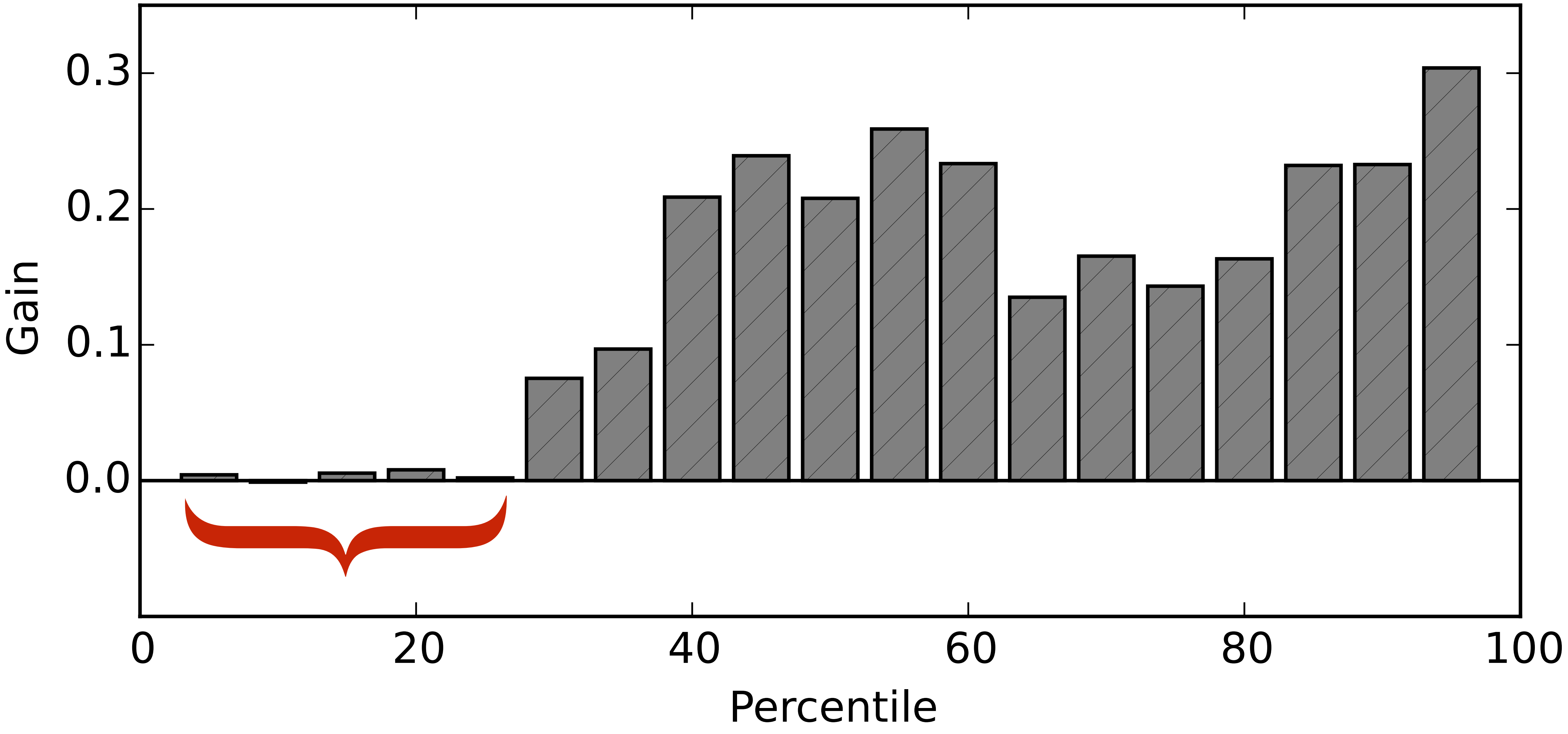
- Clients are able to complete the probe transfers in fewer round trips.
- Reduces total transfer time.

80% of transfers were faster

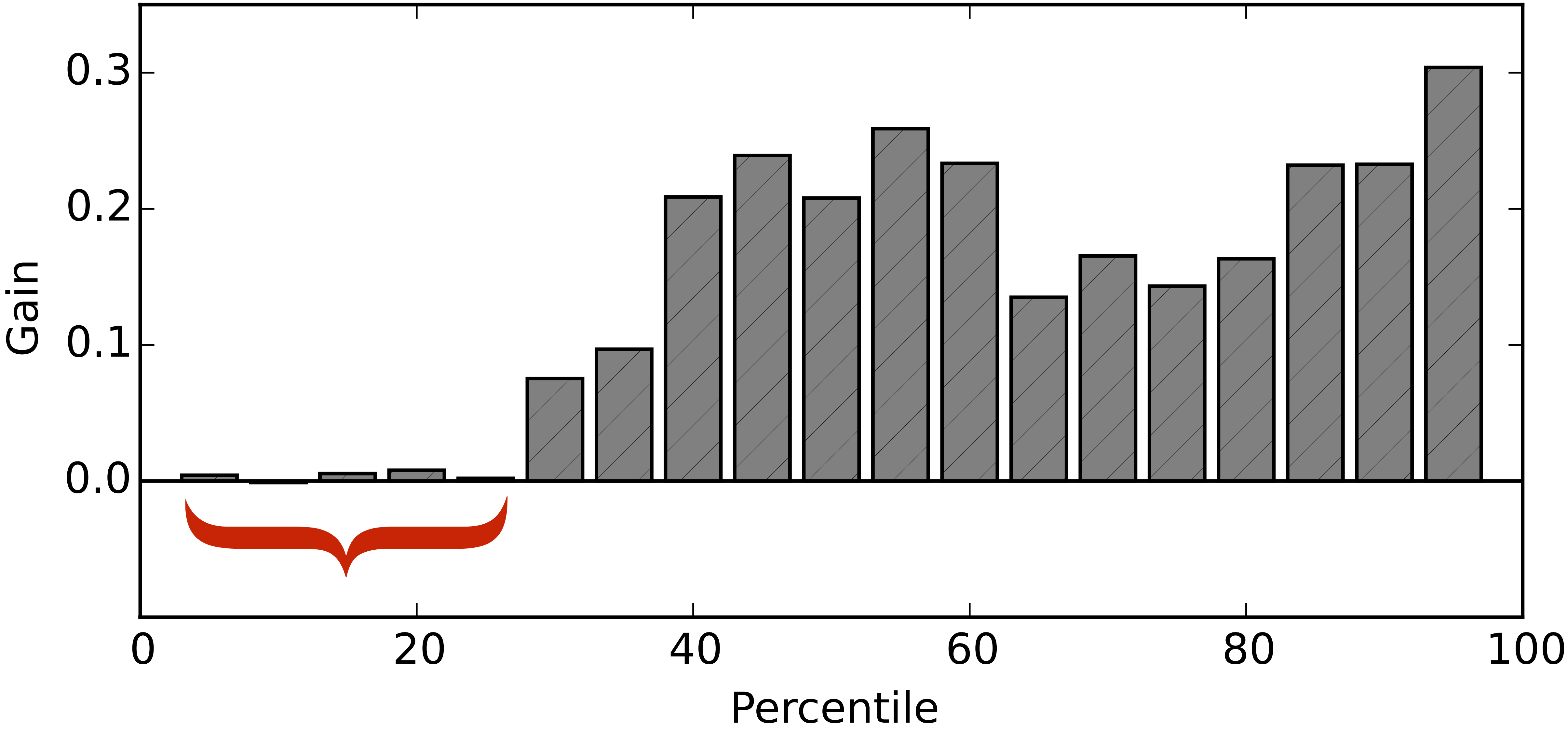
Gain Percentile



Gain Percentile



Gain Percentile



Gains were highest at upper percentiles.

Conclusion

- Demonstrated design and implementation of a simple tool to observe and adjust congestion windows.
- Deployed the system in a production network.
- Achieved significant increase in average congestion window.
- Demonstrated improvements in completion time, reducing slow-start penalty

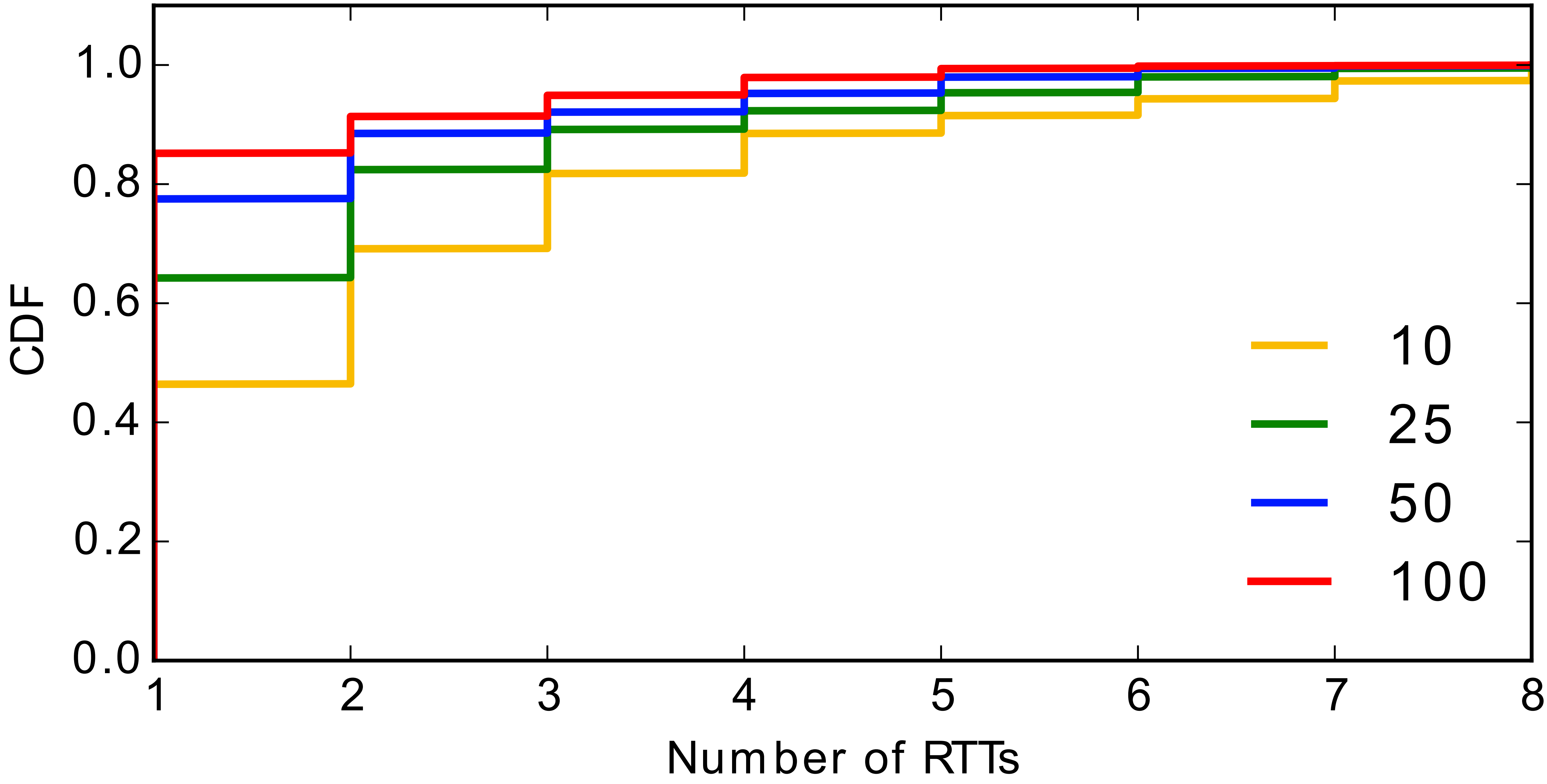
Thank you!

Extras

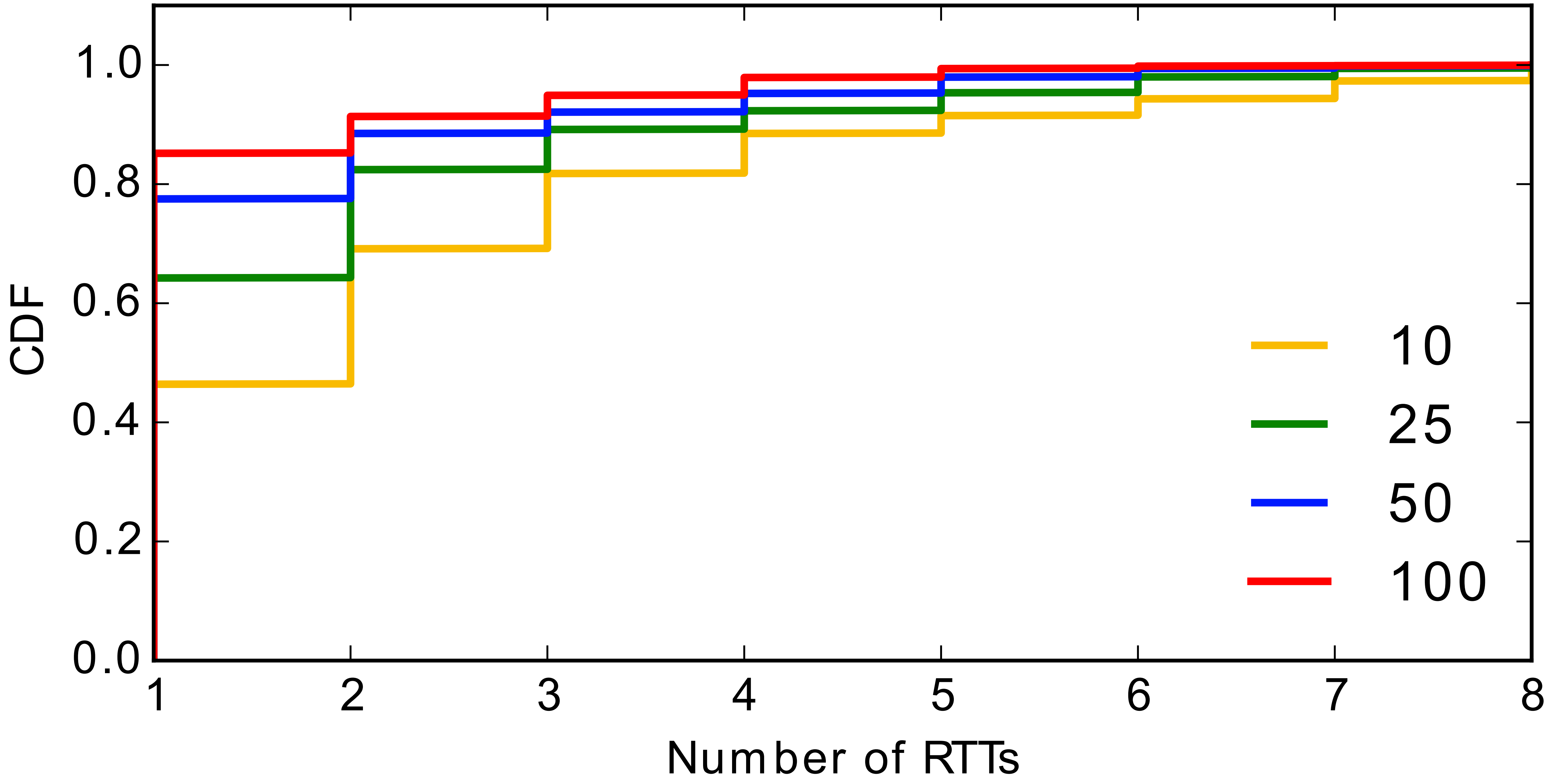
Cloud systems

- Complexity means node-level resource constraints
- Frequent connections between Points-of-Presence (PoPs).
- In many cases dominated by small file transactions.

Cloud workloads

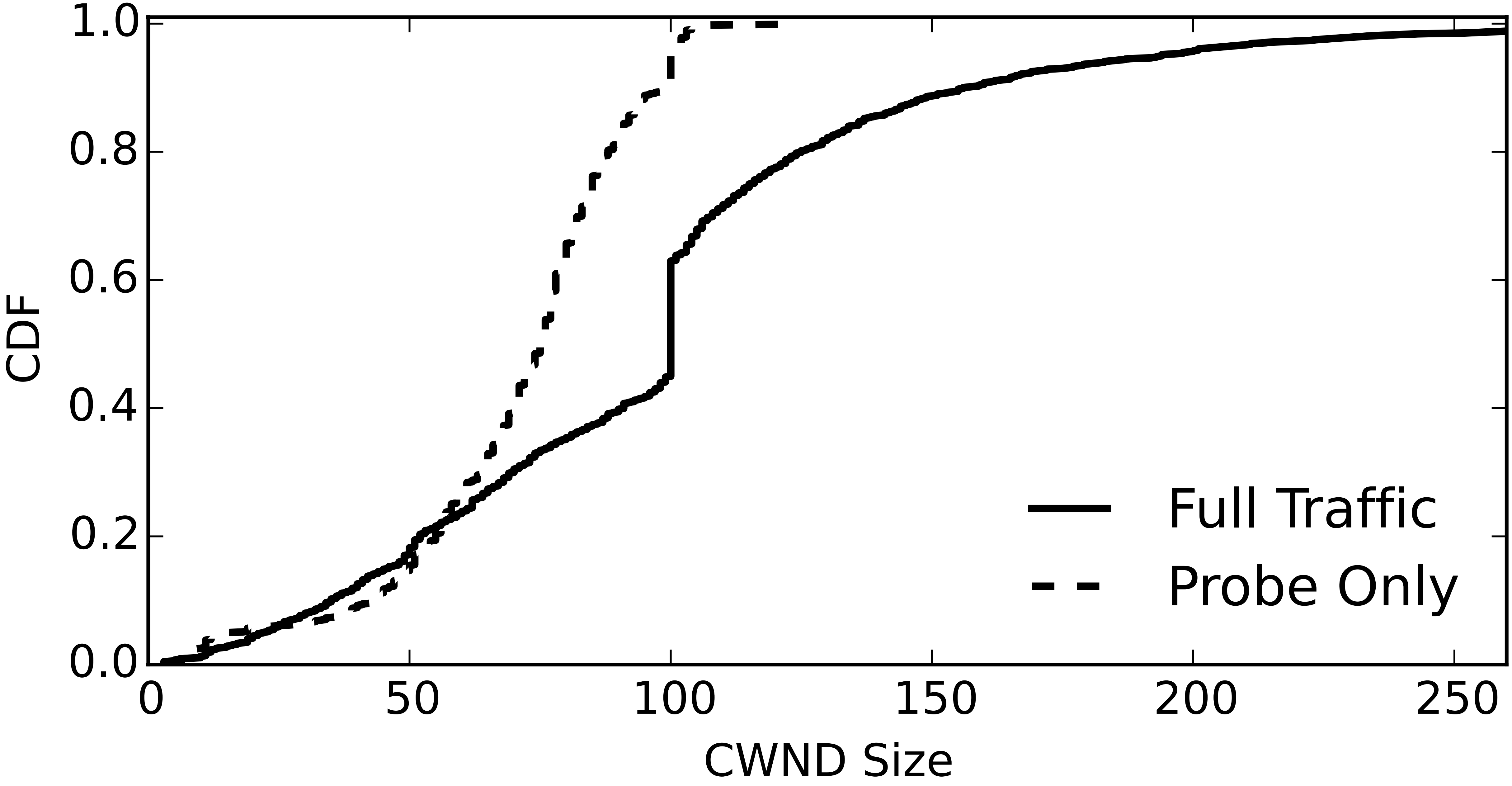


Cloud workloads

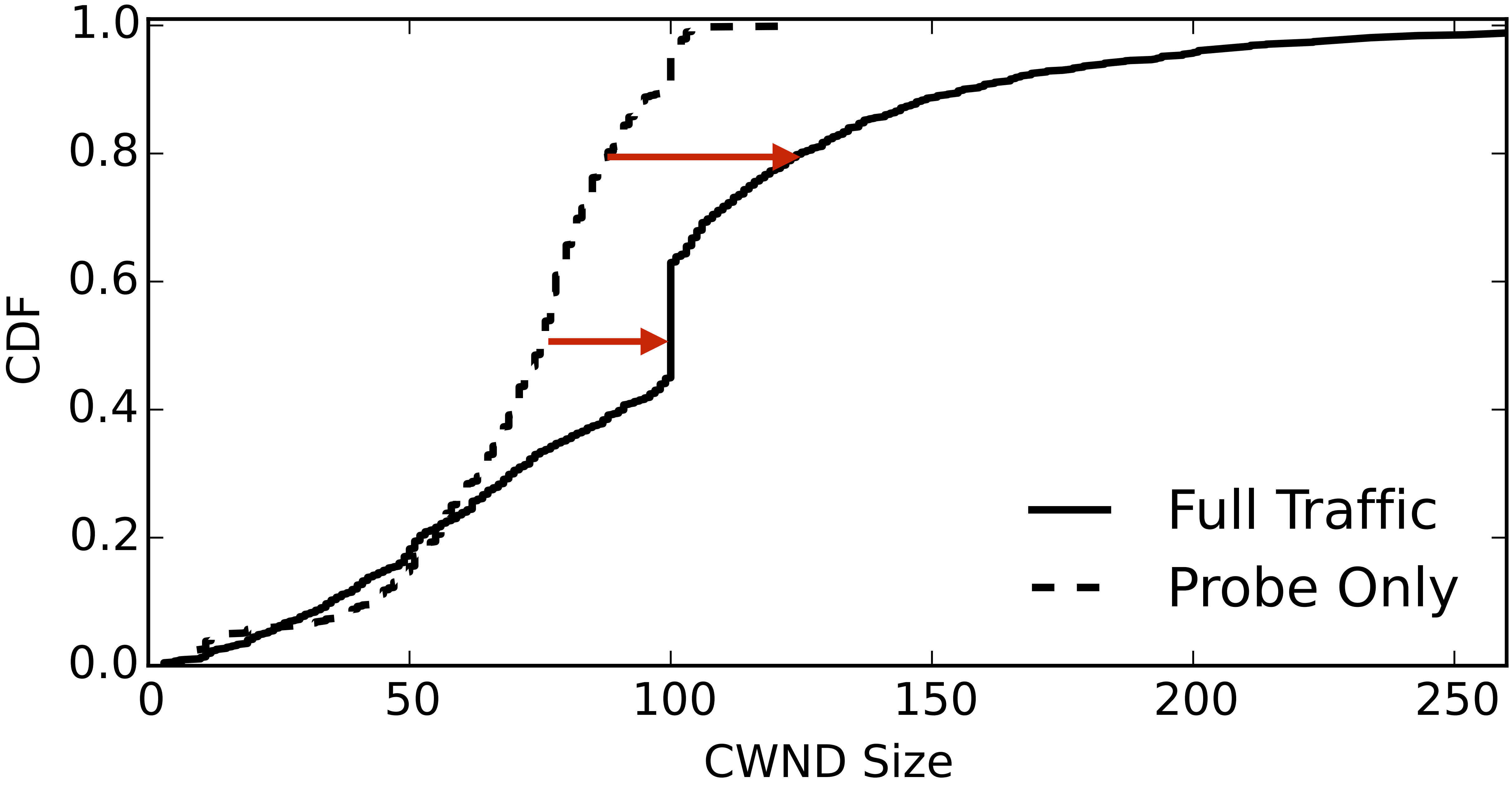


Larger windows reduce RTTs

Traffic matters



Traffic matters



Traffic drives up window sizes.