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

Sender

Receiver
Slow-start penalty

Sender

Syn

Receiver
Slow-start penalty

Sender

Receiver

Syn

Syn/Ack
Slow-start penalty

Sender

Initial Window

Receiver

Syn

Syn/Ack
Slow-start penalty

Sender

Syn

Syn/Ack

Receiver

Initial Window

Data doesn’t fit
Slow-start penalty

Sender

Initial Window

2nd Window

Receiver

Syn

Syn/Ack

Data doesn’t fit
Slow-start penalty

Sender

Receiver

Initial Window

2nd Window

Syn

Syn/Ack

Data doesn’t fit

Transfer pays second RTT
Cloud workloads

CDF

File Size (KB)
Cloud workloads

54% are too big for default
Cloud workloads

54% are too big for default

85% under 200KB
Global deployments

CDF of RTT (MS)
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

CDF

CWND Size

Default
50
100
150
200
250
Observed windows

CDF windows significantly higher.
Observed windows

CDF windows significantly higher.
Observed windows

CDF

CWND Size

Knee

Default

50

100

150

200

250

CWND windows 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.

Graph showing CDF of transfer time for RTT > 150ms comparing Riptide and Default.
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
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 Graph](image)

- **CDF**: Cumulative Distribution Function
- **CWND Size**: Congestion Window Size
- **Full Traffic**
- **Probe Only**

The graph illustrates the distribution of CWND sizes under full traffic and probe-only conditions.
Traffic matters

Traffic drives up window sizes.