Low Latency Transport Scheme For Datacenter Networks

***Behnam Montazeri, **Mohammad Alizadeh, *John Ousterhout**

**Stanford University, **Massachusetts Institute of Technology and Cisco*

SECDL/PlatformLab Retreat, May 2015

RAMCloud

Motivations

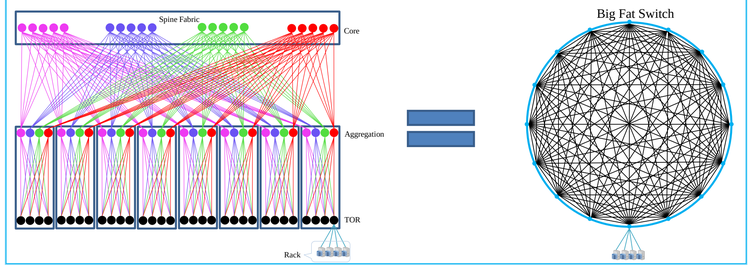
- ❑ **RAMCloud RPC relies on Infiniband reliable transport**
- ❑ **Infiniband has scalability issues and not considered commodity**
- ❑ **We want to achieve low latency over unreliable datagrams**
- ❑ **FastTransport is a primitive transport layer**
  - ✓ Provides reliability for datagram protocols
  - ✓ Lacks congestion control
  - ✓ Not scalable
- ❑ **Designing a new reliable transport protocol**
  - ✓ Fit for datacenter networks
  - ✓ Tailored for RPC systems

Objectives

- ❑ **Low Latency**
  - ✓ As close as possible to hardware limits
  - ✓ Minimal buffer usage
- ❑ **Scalability**
  - ✓ Millions of client connections per server
  - ✓ Minimal per client state
- ❑ **Congestion Control**
  - ✓ Low latency for small request in presence of high network utilization

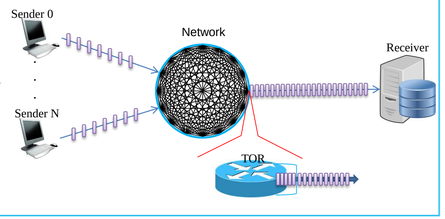
Network Assumptions

- ❑ **Full Bisection Band Width**
- ❑ **Low latency**
- ❑ **Load Balanced**
- ❑ **Switches Provide few priority levels**
- ❑ **Network delays are not fixed**



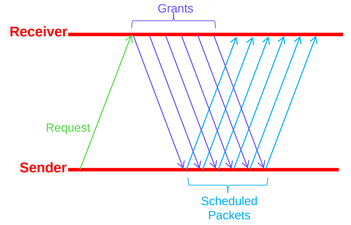
Congestion Primarily At Receiver's TOR

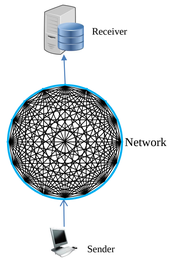
- ❑ **Congestion primarily at receiver's TOR**
- ❑ **Receiver Knows Msg. Sizes**
- ❑ **Receiver's the right place to do Congestions control**



Receiver Side Scheduler

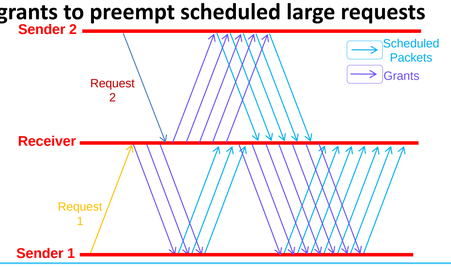
- ❑ **Sender sends request that specifies the message size**
- ❑ **Receiver grants permission for transmission**
- ❑ **Grants are sent in fine grained time intervals**

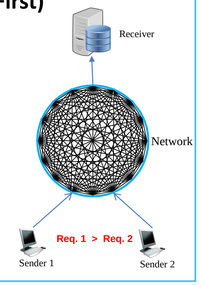




Preemption By Tokens

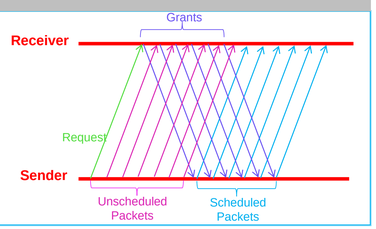
- ❑ **Favor Shortest Request (Shortest Remaining Bytes First)**
- ❑ **Use grants to preempt scheduled large requests**





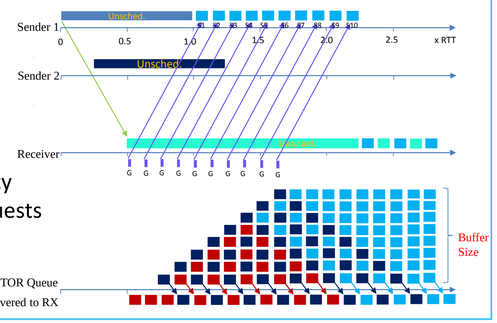
Unscheduled Traffic

- ❑ **Small Unscheduled Traffic covers for one RTT latency overhead**



Problem: Buffer Buildup

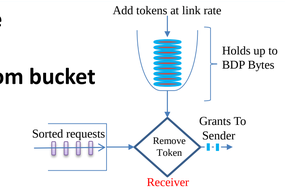
- ❑ **With unscheduled traffic, multiple senders cause buffer build**

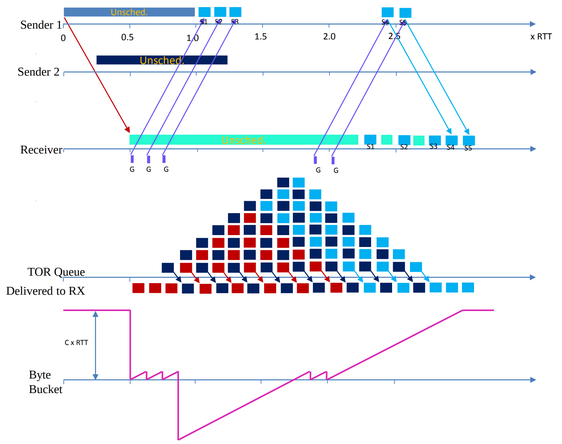


- ✓ Buffer adds latency
- ✓ Buffer limits our ability to preempt large requests for shorter ones

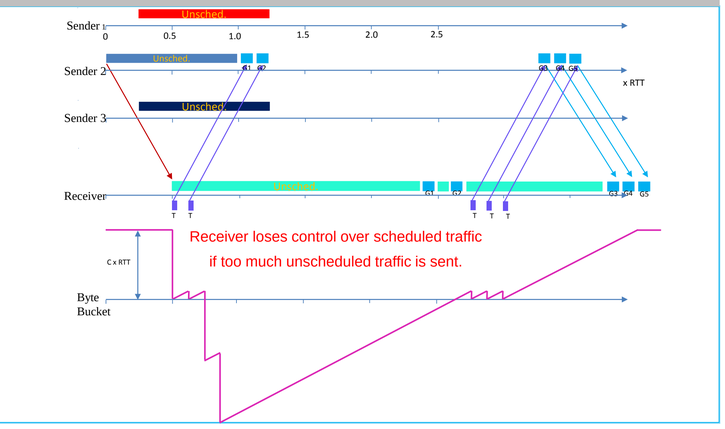
Buffer Buildup: Solution

- ❑ **Bytes are added to the bucket at link rate**
- ❑ **Bucket level is capped at BDP = C x RTT**
- ❑ **Unscheduled traffic will be subtracted from bucket**





Problem: Too Much Unscheduled Traffic



Work Status

- ❑ **Algorithm needs to be polished and finished**
- ❑ **The effect of random delay variations must be taken into account**
- ❑ **Limited number of priorities can be used for preemption**
  - ✓ Higher priority for short requests
  - ✓ Different priority level within unscheduled and/or scheduled traffic
- ❑ **Simulation and implementation of yet to be done**