SolarFlare 10GigE Driver For RAMCloud

Behnam Montazeri

Stanford University

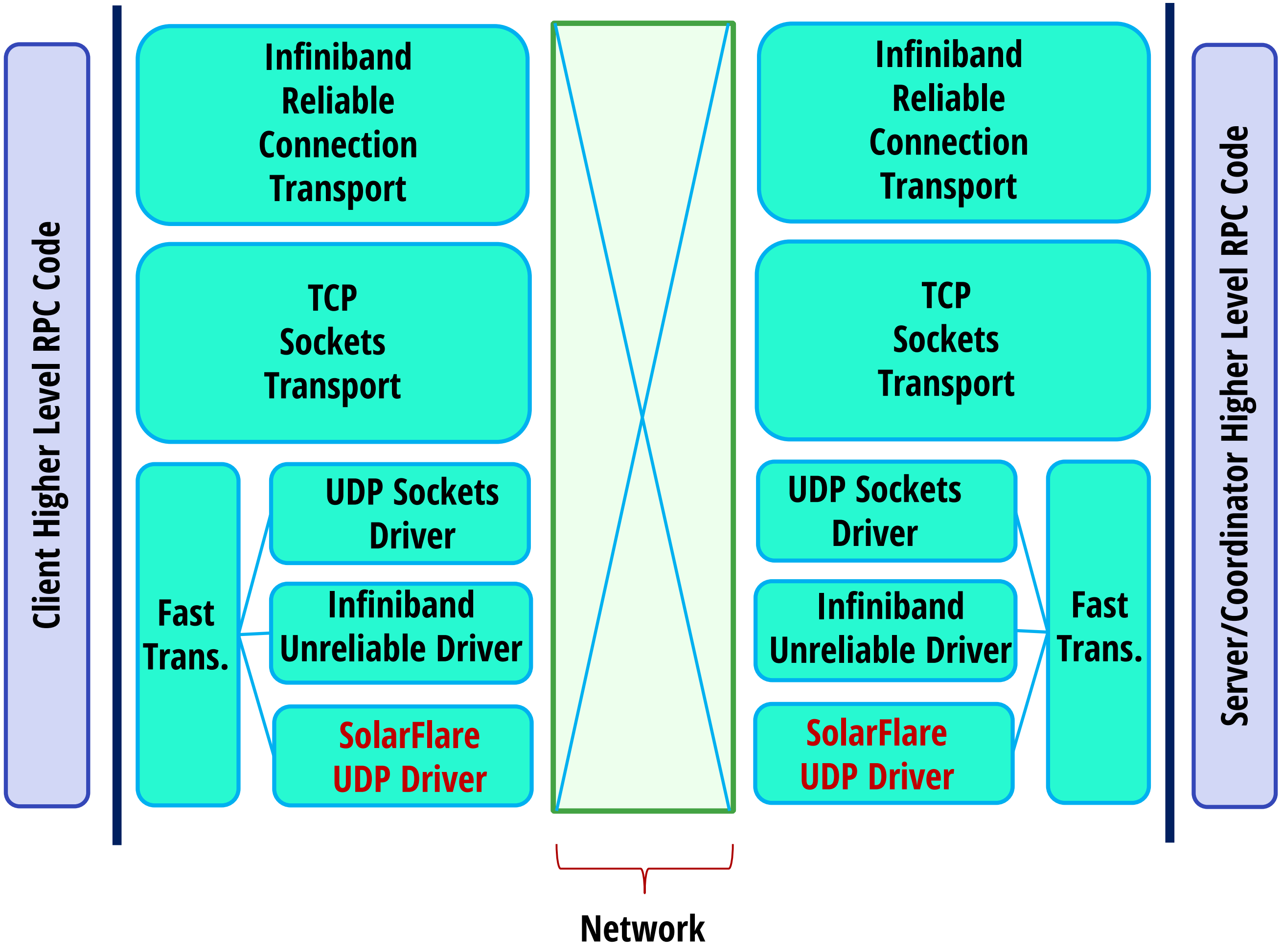
RAMCloud

Overview

- ❑ **10Gb/s Commodity Ethernet Driver For SolarFlare NICs**
  - ✓ Preliminary datagram that hooks into current system
  - ✓ Kernel bypass for minimal latency overhead
  - ✓ Polling based architecture rather than interrupts
  - ✓ Fast user space ARP Cache for Layer 3 Networking
- ❑ **SolarFlare in RAMCloud**
  - ✓ Read latency of 100B object:
    - 9.5us for SolarFlare driver versus 6 us for InfUd driver
    - More than 40us latency if we use Kernel TCP

RPC Transport Layer

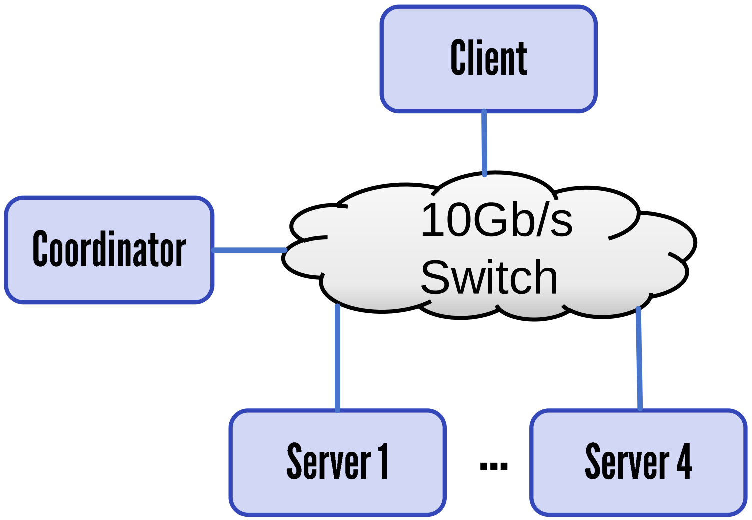
- ❑ Drivers send and receiver datagrams
- ❑ FastTransport Provides Reliable In Order Delivery
- ❑ Driver API:
  - ✓ Connect()
  - ✓ Disconnect()
  - ✓ sendPacket()
- ❑ Driver also provides:
  - ✓ Poller Object
  - ✓ Received Object


Client Higher Level RPC Code
Infiniband Reliable Connection Transport
TCP Sockets Transport
Fast Trans.
UDP Sockets Driver
Infiniband Unreliable Driver
SolarFlare UDP Driver
Infiniband Reliable Connection Transport
TCP Sockets Transport
UDP Sockets Driver
Infiniband Unreliable Driver
SolarFlare UDP Driver
Fast Trans.
Server/Coordinator Higher Level RPC Code
Network

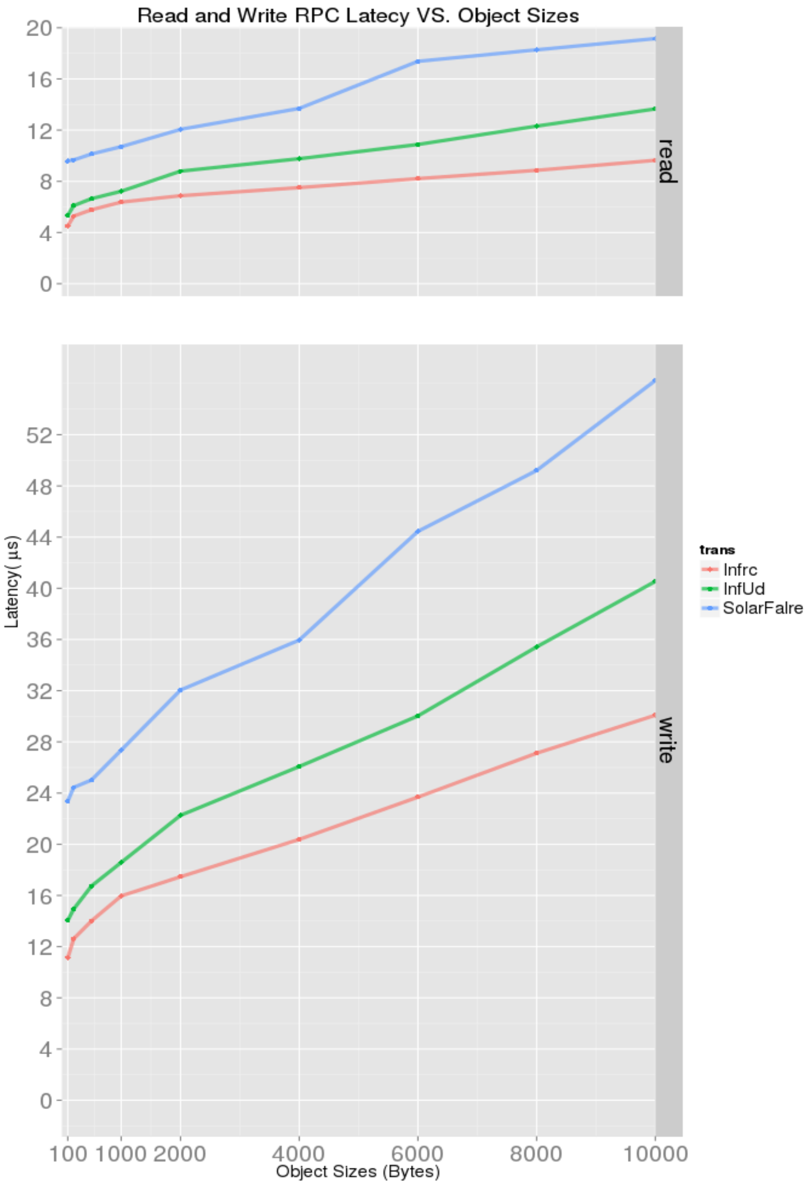
IP-MAC Translations

- □ **Problem:**
  - ✓ We want to send packets based on IP addresses
  - ✓ We need a way to translate IP addresses to MAC addresses
- □ **Solution 1: Use Kernel ARP module**
  - ✓ Need to read kernel route table and ARP table
  - ✓ Involves system calls that are too slow
  - ✓ Sending ARP packets needs root access
- □ **Solution 2: Implement User Space ARP Module**
  - ✓ Keep a cache of selected entries of kernel route table and ARP table
  - ✓ For every IP-MAC translation:
    - Resolve the MAC from the ARP cache, if failed:
      - Resolve from kernel cache or trigger kernel ARP process
      - Update ARP cache

Performance Analysis



- ❑ **RAMCloud with SolarFlare**
  - ✓ **For 100B objects:**
    - **Fast+InfUd read latency is 6us**
    - **Fast+SolarFlare read latency is 9.5us**
    - **The switch accounts for 1.3us**
    - **SolarFlare cluster is 10 to 15% slower than our Infiniband cluster**
    - **The rest of the difference comes from the NIC**



Conclusion

- ❑ **We now have 10Gb/s Ethernet support for RAMCloud**
  - ✓ Developed for SolarFlare NICs
  - ✓ Send and receives packets on layer 3
- ❑ **Faster NICs should get us closer to Infiniband performance**
- ❑ **Lots of room to improve**
  - ✓ A new transport protocol for RPC systems
    - Low latency
    - Highly Scalable: millions of sessions per server
    - General Purpose
  - ✓ Revisiting RPC architecture as a whole