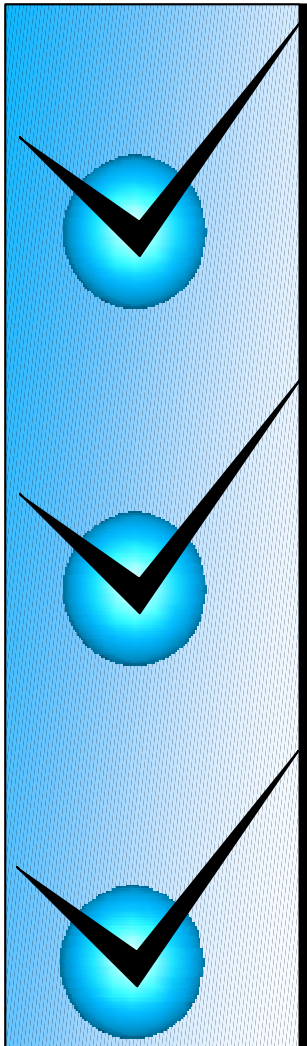


NFS over RDMA

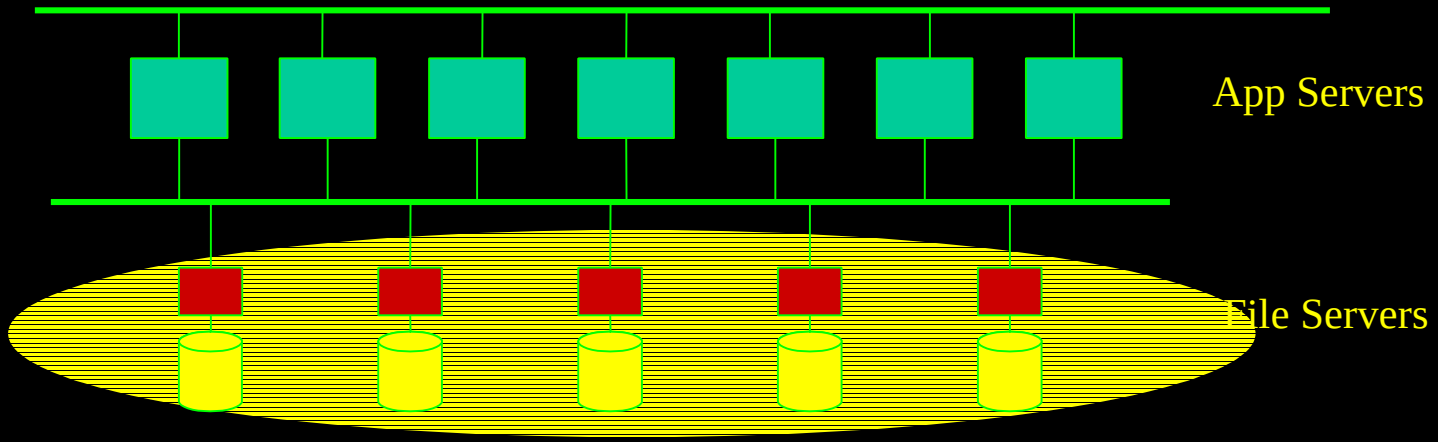
Alex Chiu
Staff Engineer
Sun Microsystems
alex.chiu@sun.com

Agenda

- Data Center In-Room Networks
- What is RDMA?
- NFS over RDMA
- RPC via RDMA
- Status

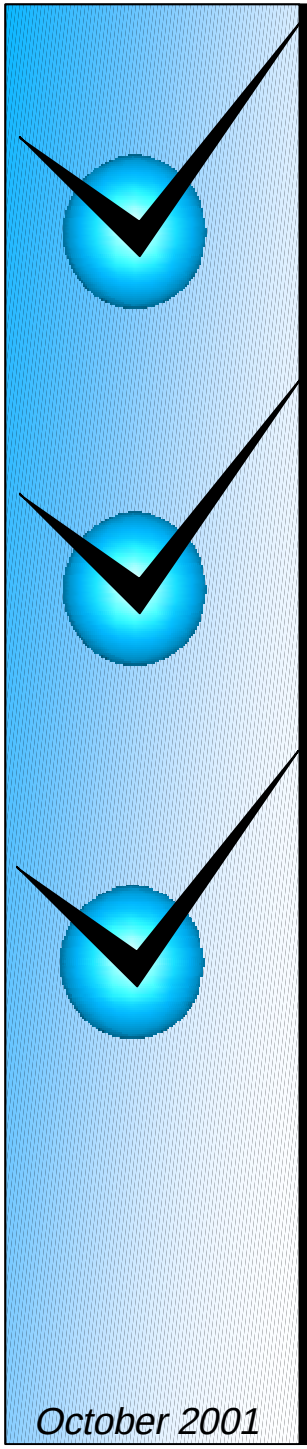
In-Room Networks

- Transaction processing
  - Web, Mail, ASPs, eCommerce
- Separate servers from storage
- "Room area" - not "Wide area"



In-Room Networks

- Networked with high speed interconnects (Virtual Interface, InfiniBand, ...)
  - Fast data access
  - Reduced CPU overhead
  - Memory registration
    - Reuse registered memory buffers
    - Avoid memory locking and virtual-to-physical address translation



In-Room Networks

- Features
  - Low latency
  - High bandwidth
  - Simple network
  - Low error rate
  - Physical security
  - Tightly configured & controlled

What is RDMA?

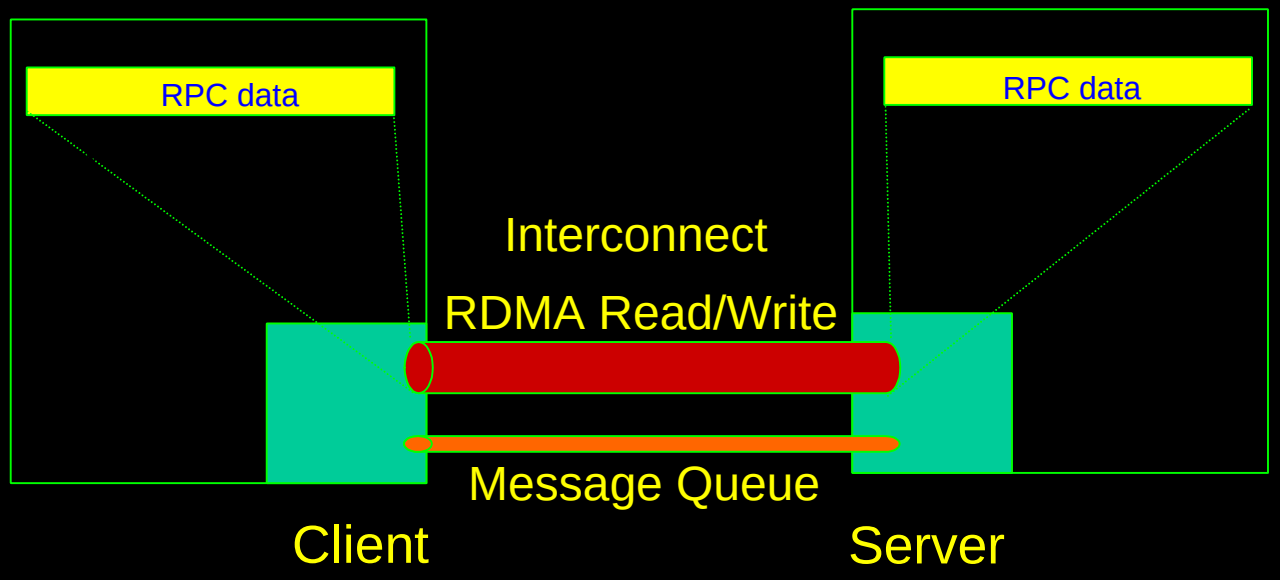
- RDMA: Remote Direct Memory Access
  Remote data transfer to and from memory directly without CPU intervention

What is NFS over RDMA?

- NFS for data center ***in-room*** networks with RDMA-capable high speed interconnects
- RDMA as a new RPC transport

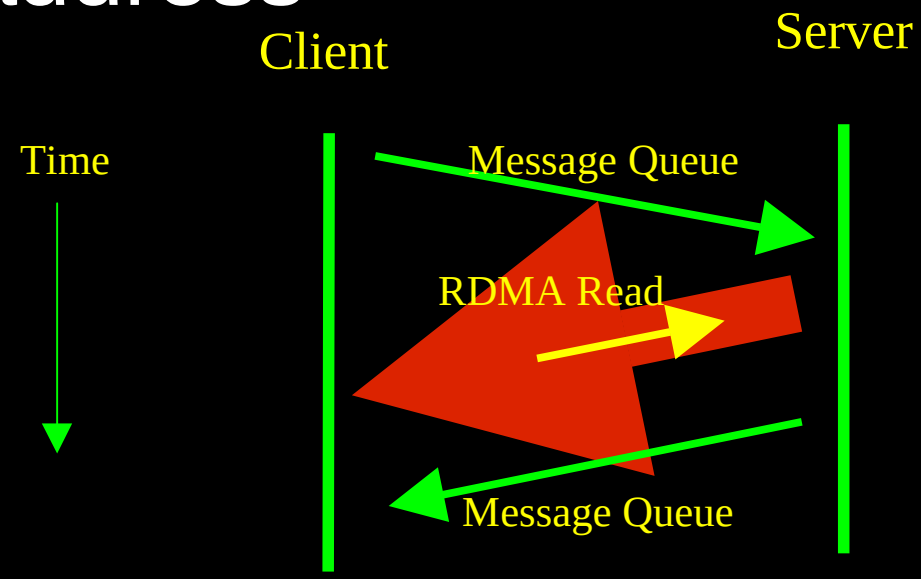
RPC via RDMA

- RPC data in a memory segment
- Notification via message queue



RPC via RDMA

- RDMA Reads initiated by receiver
- Data buffers referenced by virtual address



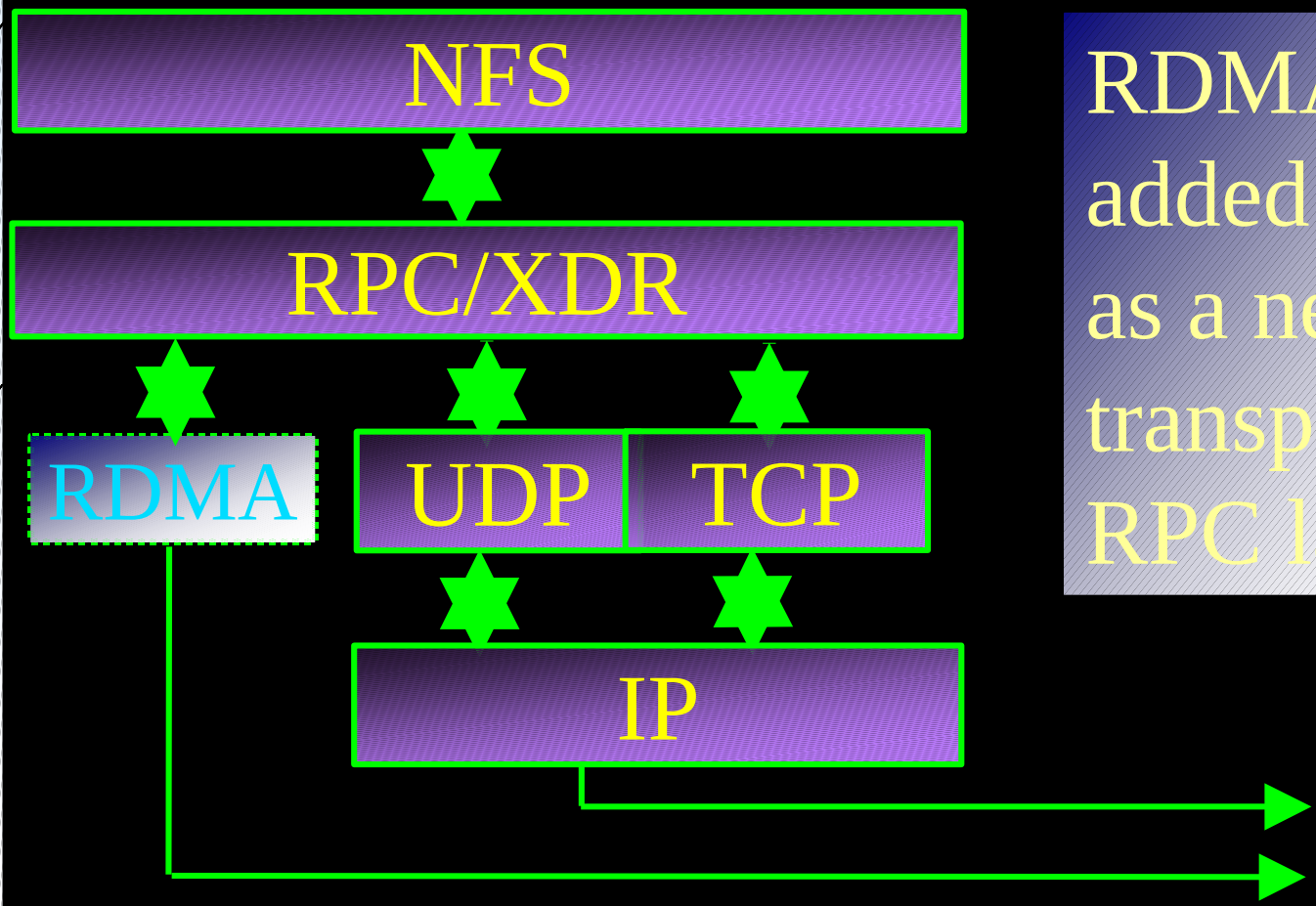
Node Addressing

- Consequences of bypassing TCP/IP
  - Name services, IP addressing unavailable
  - How to establish interconnects?

Node Addressing

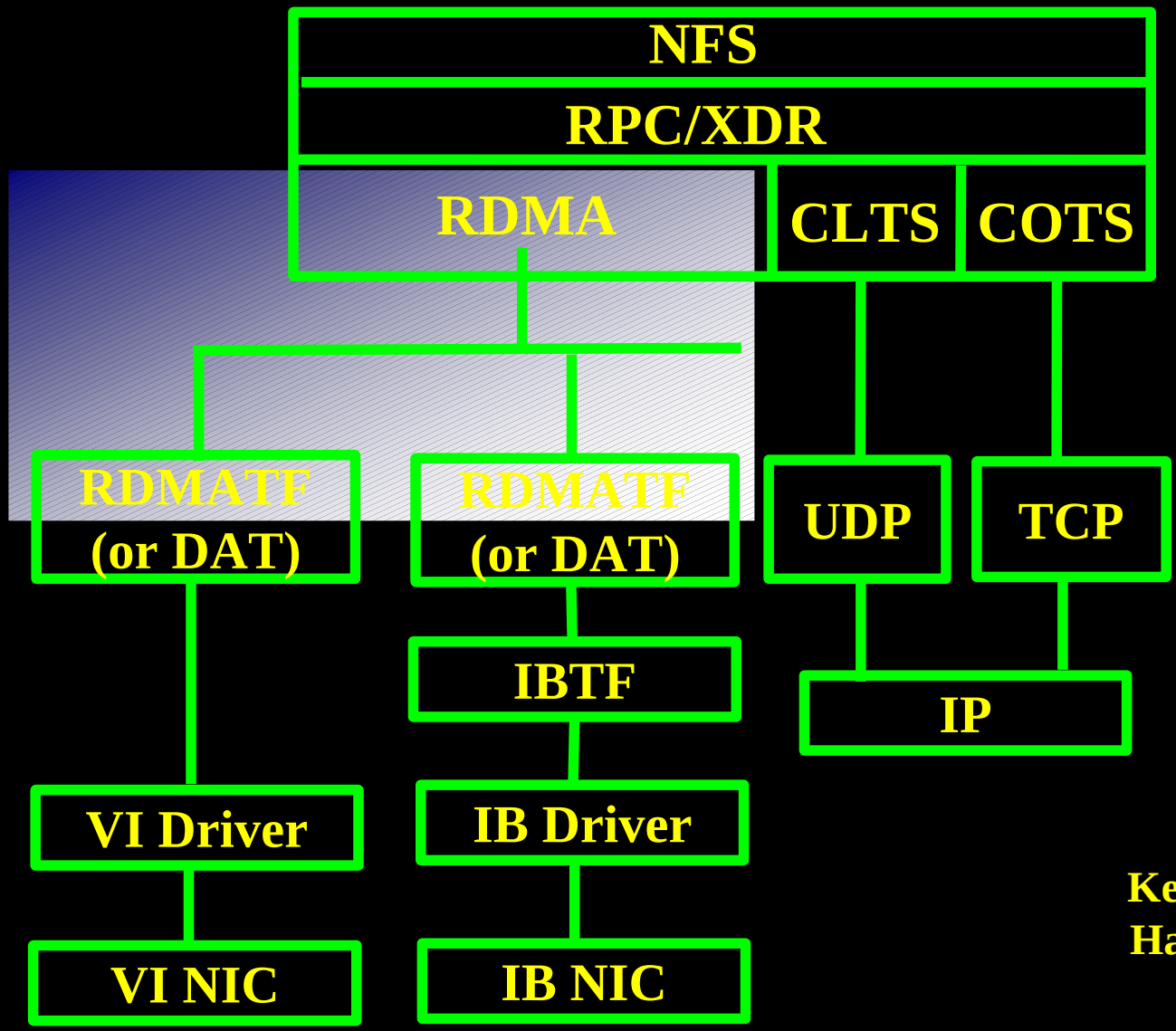
- Interim solution
  - establish a TCP connection
  - exchange over the TCP connection the parameters necessary for establishing an interconnect
  - establish the interconnect

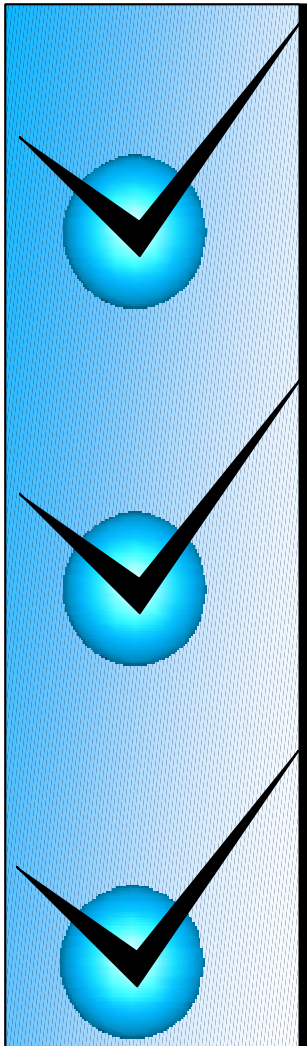
Adapting RPC to RDMA



RDMA added as a new transport to RPC layer

Adapting RPC to RDMA

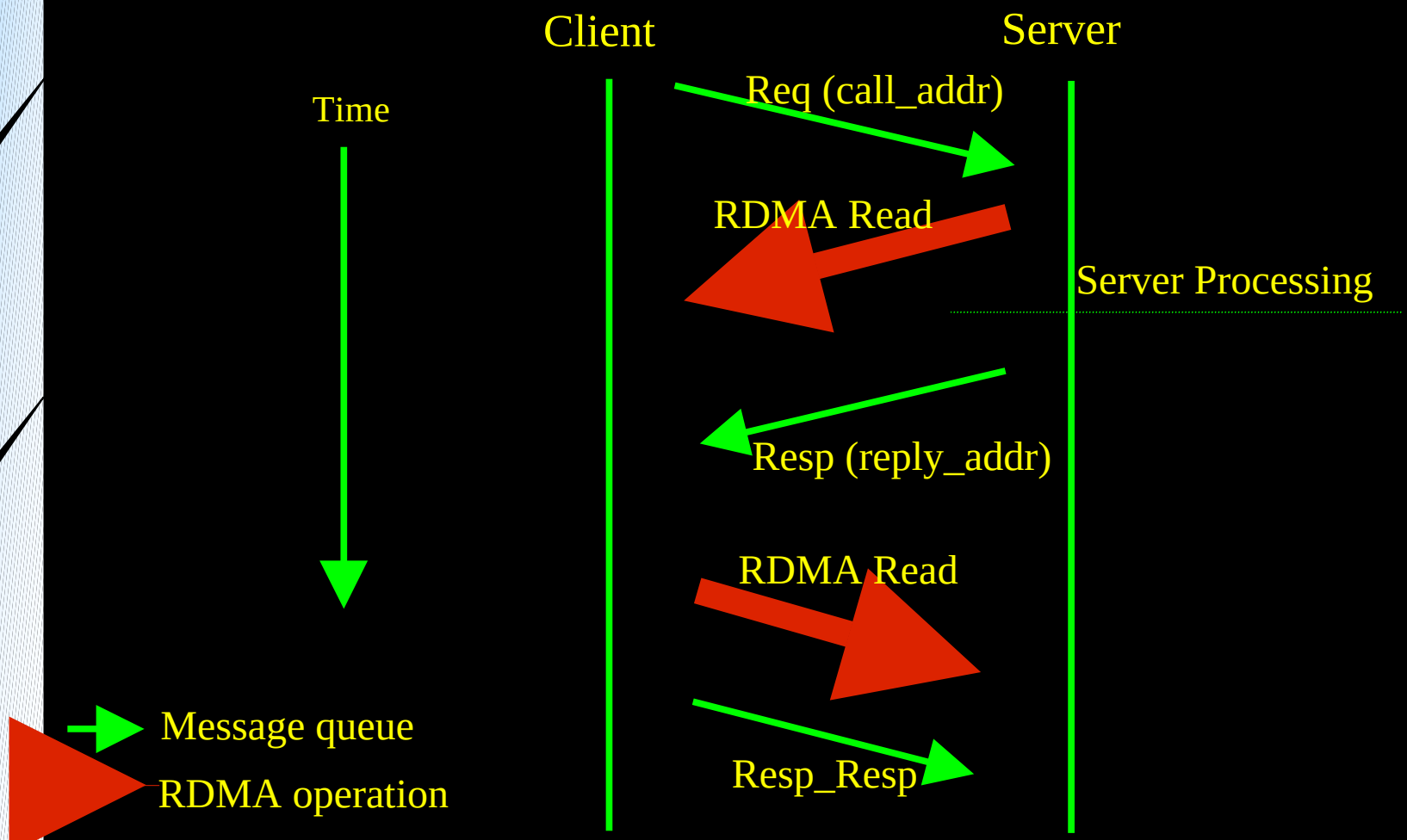




RPC RDMA Protocols

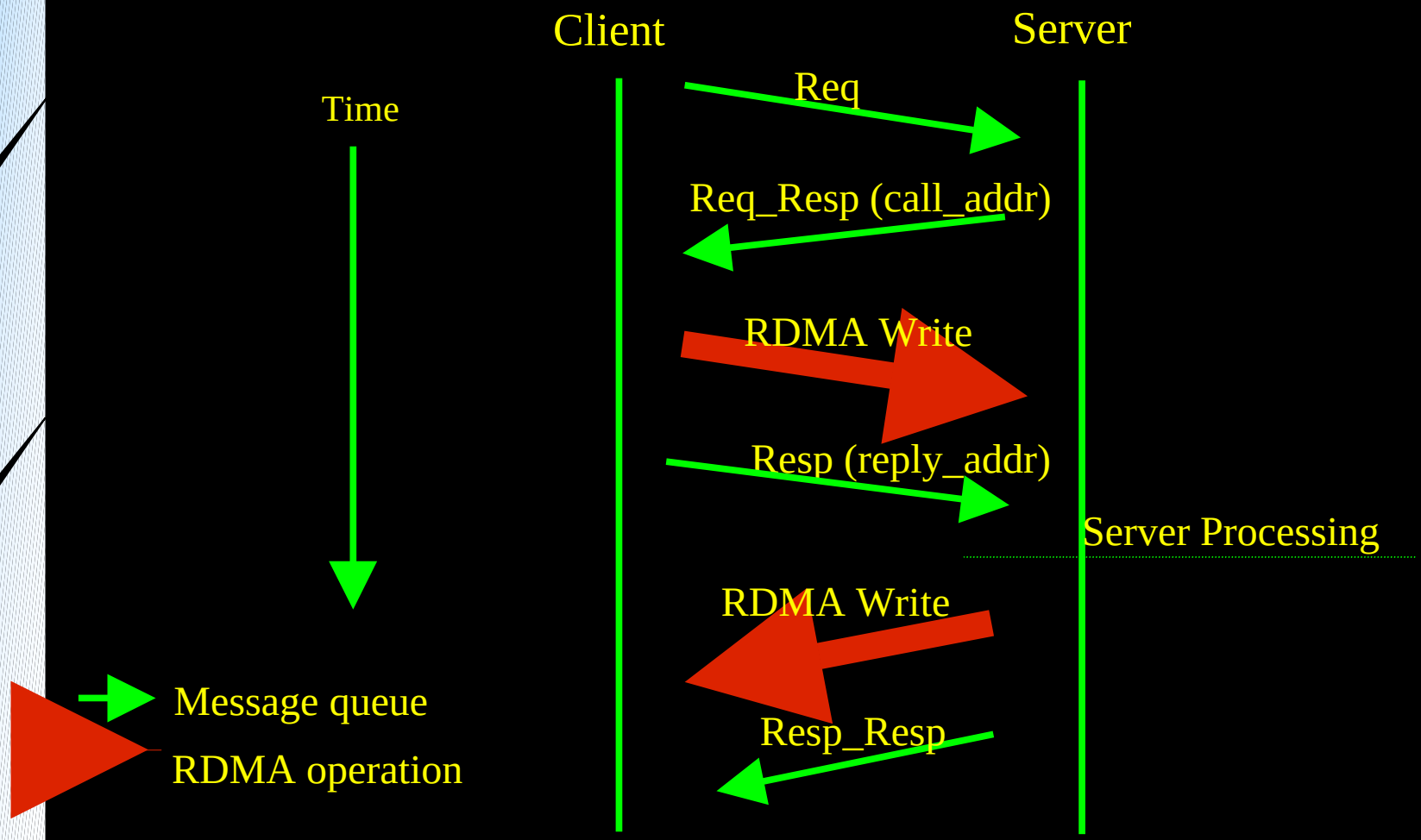
- RPC RDMA Read Only
- RPC RDMA Write Only
- RPC RDMA Read & Write

RPC RDMA Protocols



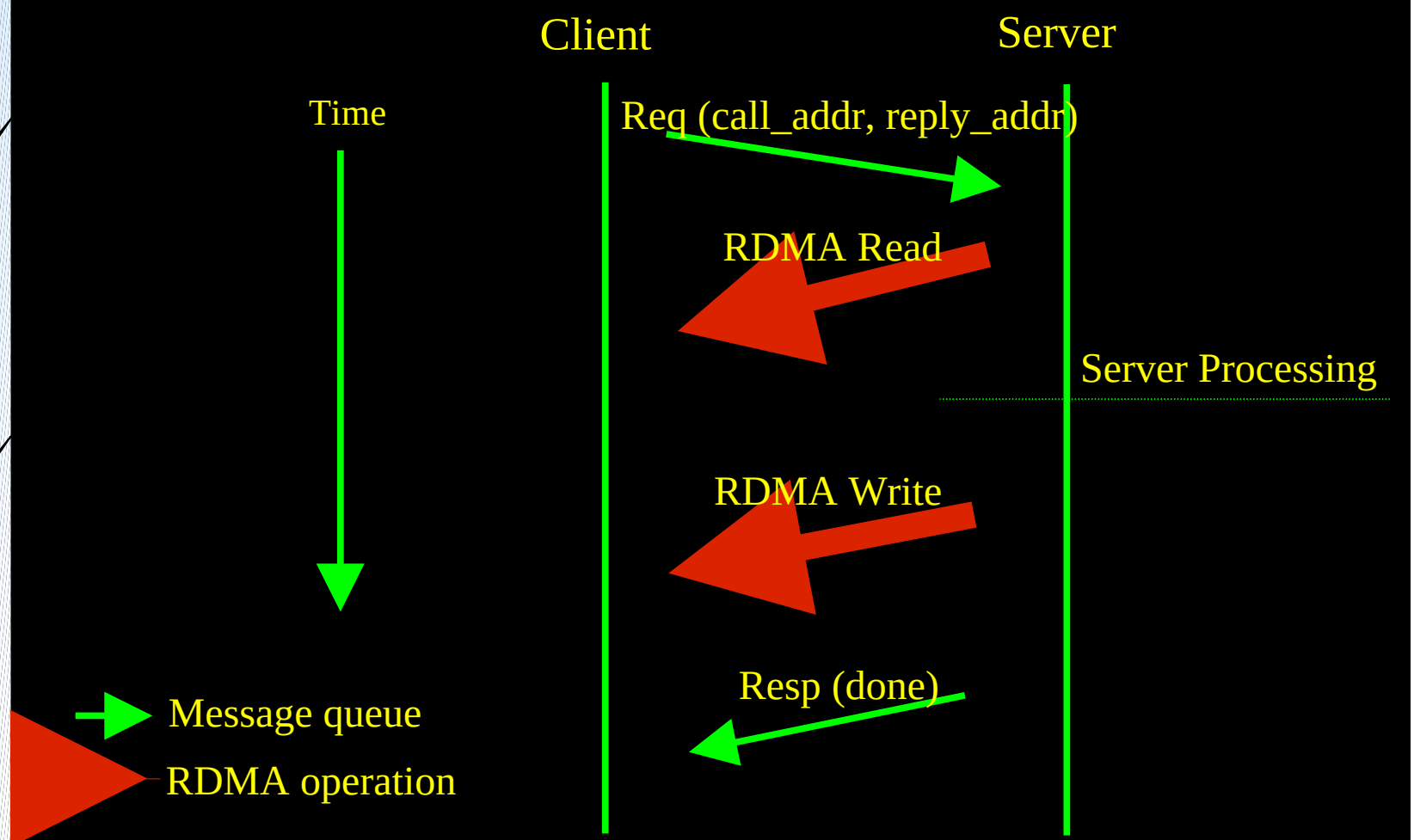
RPC RDMA Read

RPC RDMA Protocols

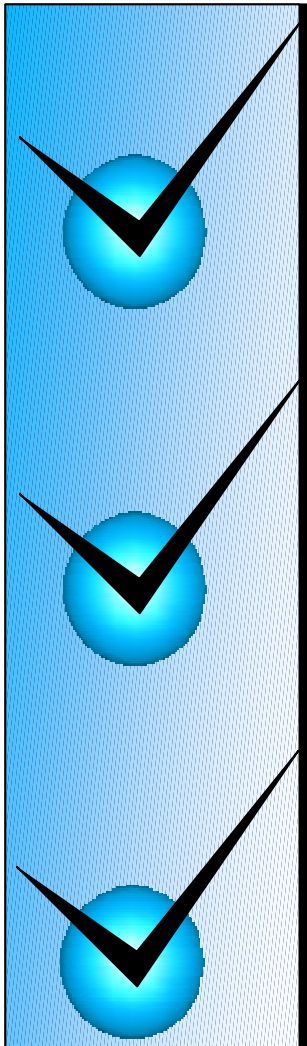


RPC RDMA Write

RPC RDMA Protocols



RPC RDMA Read/Write

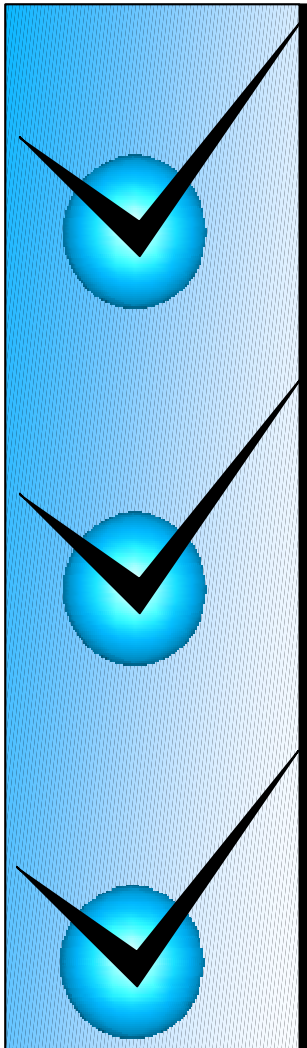


It's Just Another Transport

- No changes to NFS administration or protocols
  - NFS v2, v3, v4
  - Other RPC-based protocols
- Invisible to users, application developers

Proof of Concept

- We have a Remote Shared Memory based prototype on Sun Machines
- What we have learned for in-room file sharing
  - Use of high speed interconnects
  - No TCP/IP
  - Without pegging the CPU!



Status

- Dependencies
  - IB: IB RDMA API, IBTF
  - VI: Emulex, QLogic, Troika
- In progress:
  - Design, development
  - Standards: SNIA, IETF