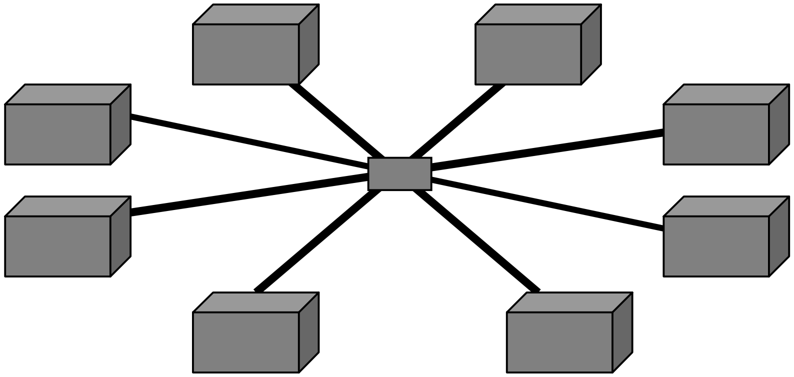
NFS over RDMA

Brent Callaghan

brent@eng.sun.com

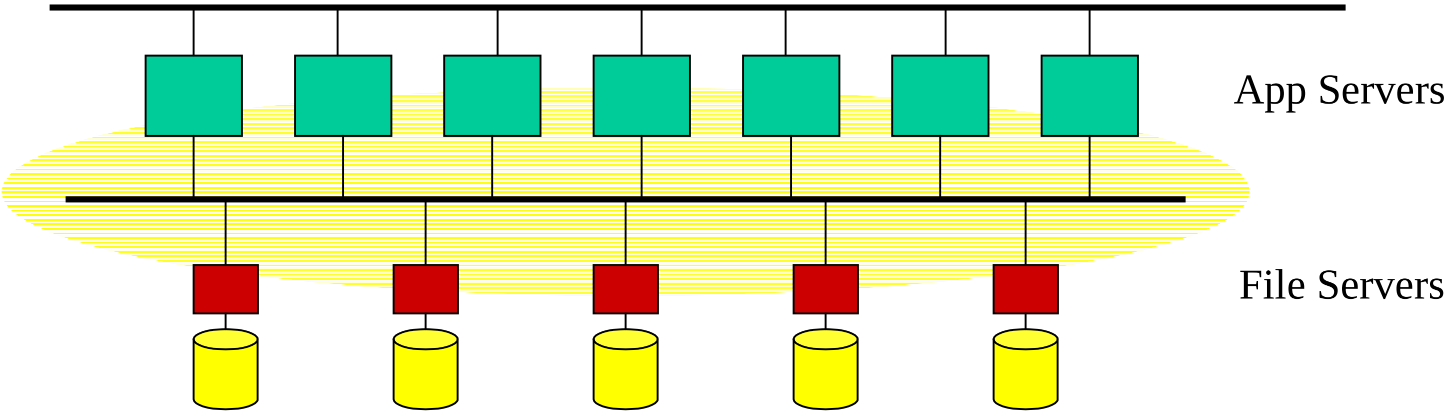
Clusters and Interconnects

- Multiple hosts
- Loosely or tightly coupled
- High speed interconnect using RDMA
  - Myrinet, Giganet, Servernet, Infiniband, ...



NAS: Network Attached Storage

- Transaction processing clusters
  - Web, Mail, ASPs, eCommerce
- Separate servers from storage
- “Room area” - not “Wide Area”

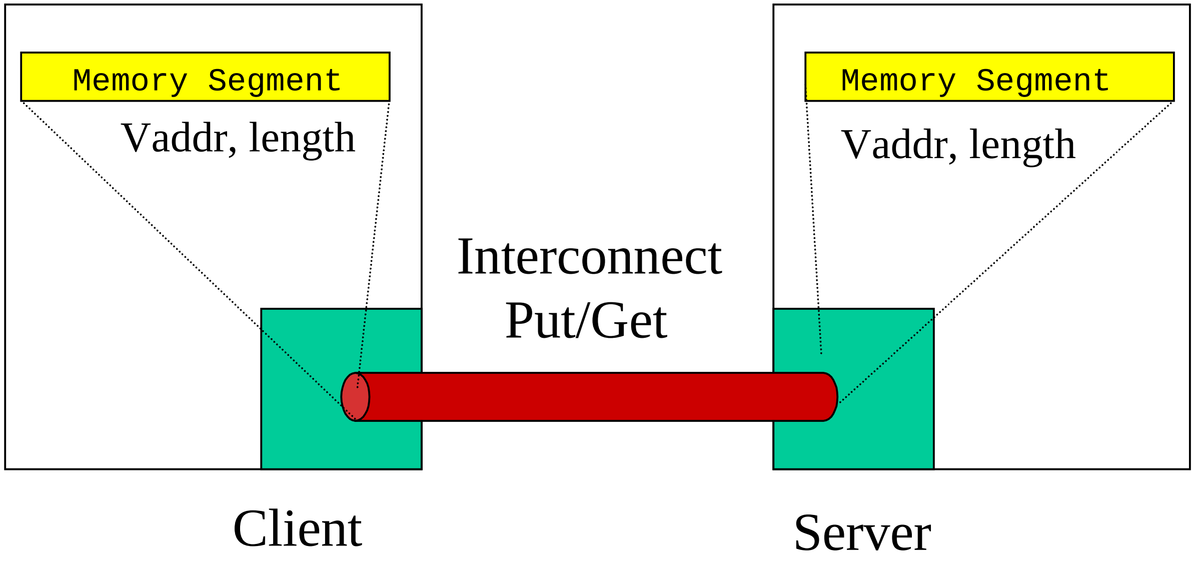


In-room Environment

- Low latency - nodes are meters apart
- High bandwidth - runs are short, cheap
- Low error rate
- Simple network
- Physical security
- Tightly configured & controlled
- TCP/IP is baggage

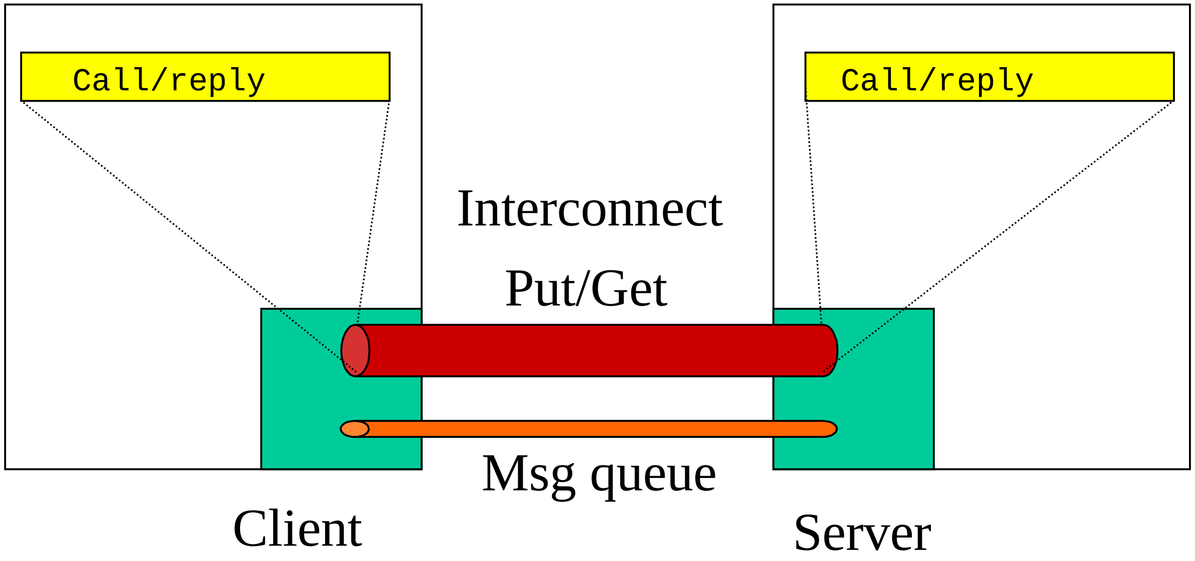
What is RDMA?

- DMA: Direct Memory Access
- RDMA: *Remote* Direct Memory Access



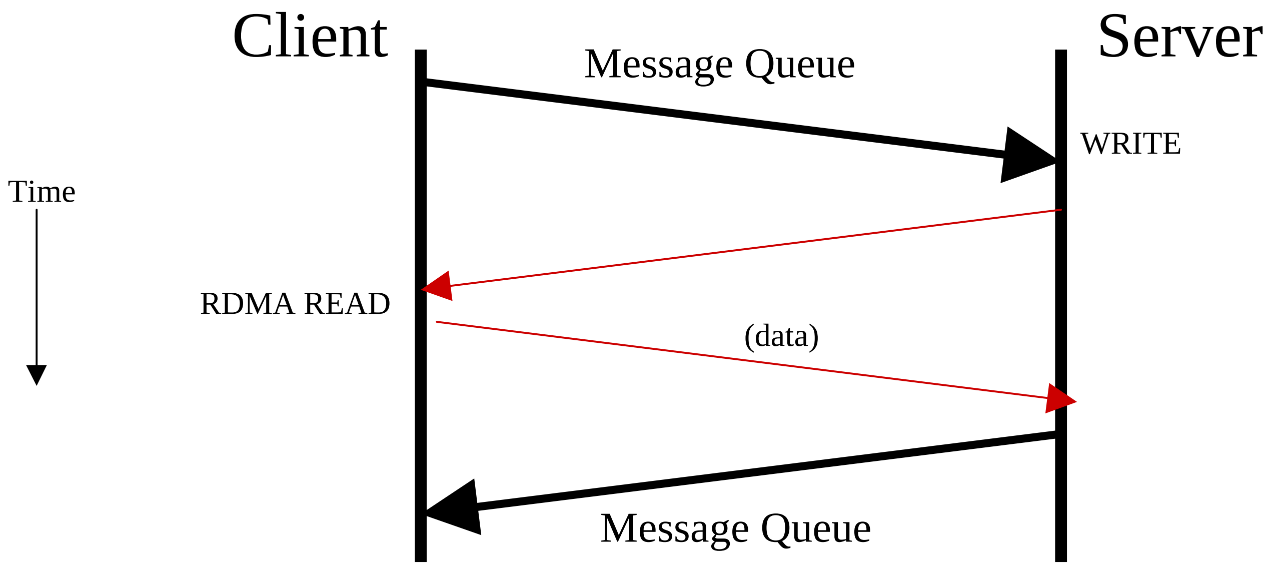
RPC via RDMA

- RPC call/reply transferred in a memory segment
- Notification via message queue

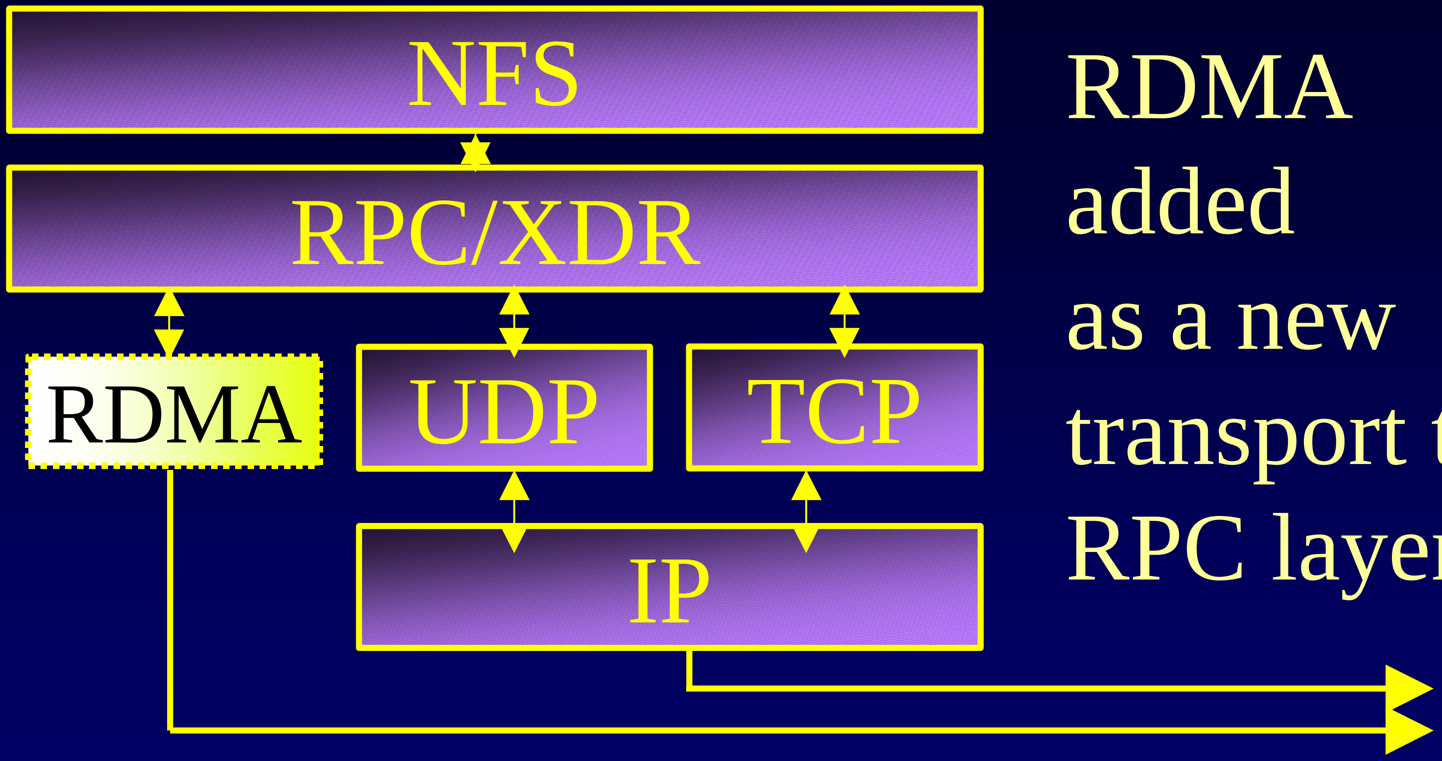


RPC via RDMA

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



Adapting NFS to RDMA



RDMA added as a new transport to RPC layer

It’s Just Another Transport

- No changes to NFS layer
  - NFS v2, v3, v4 just work
- Supports other protocols: NLM, ACL
- Invisible to users, developers

What are the Numbers?

- We have a prototype
  - Using Solaris cluster interconnect
  - RSMPI & SCI
- Aiming for “wire speed”
- Without pegging the CPU!
- Will publish numbers when ready

Questions ?