NFS INDUSTRY CONFERENCE

NFS at 10Gb/sec

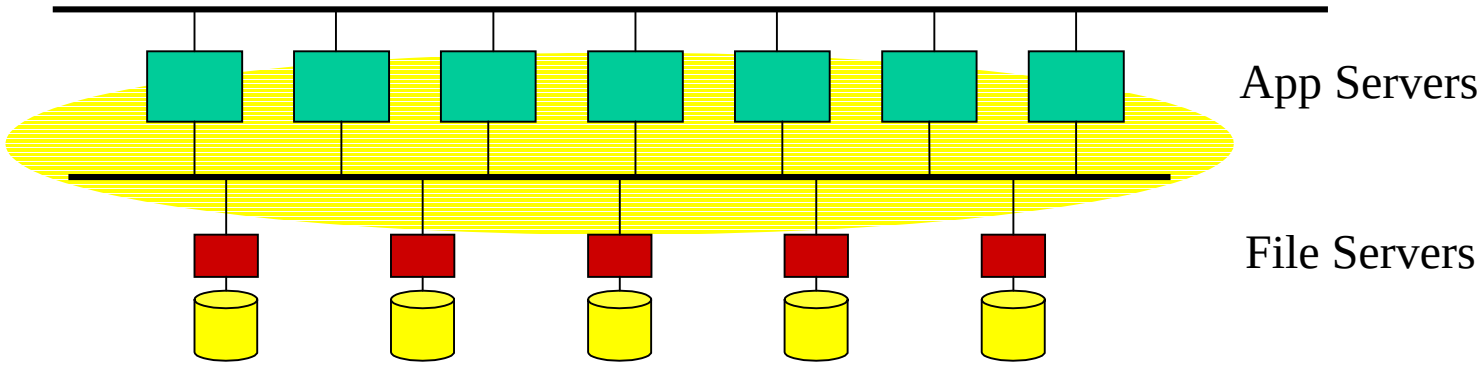
Brent Callaghan
Sun Microsystems, Inc.

October 22-23, 2002

NFS INDUSTRY CONFERENCE

Data Center NFS

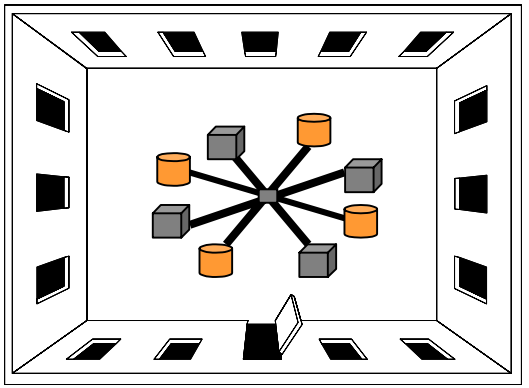
- Transaction processing clusters
  - Database, Web, Email, eCommerce
- Dataless application servers
- Apps & Data are meters apart



NFS INDUSTRY CONFERENCE

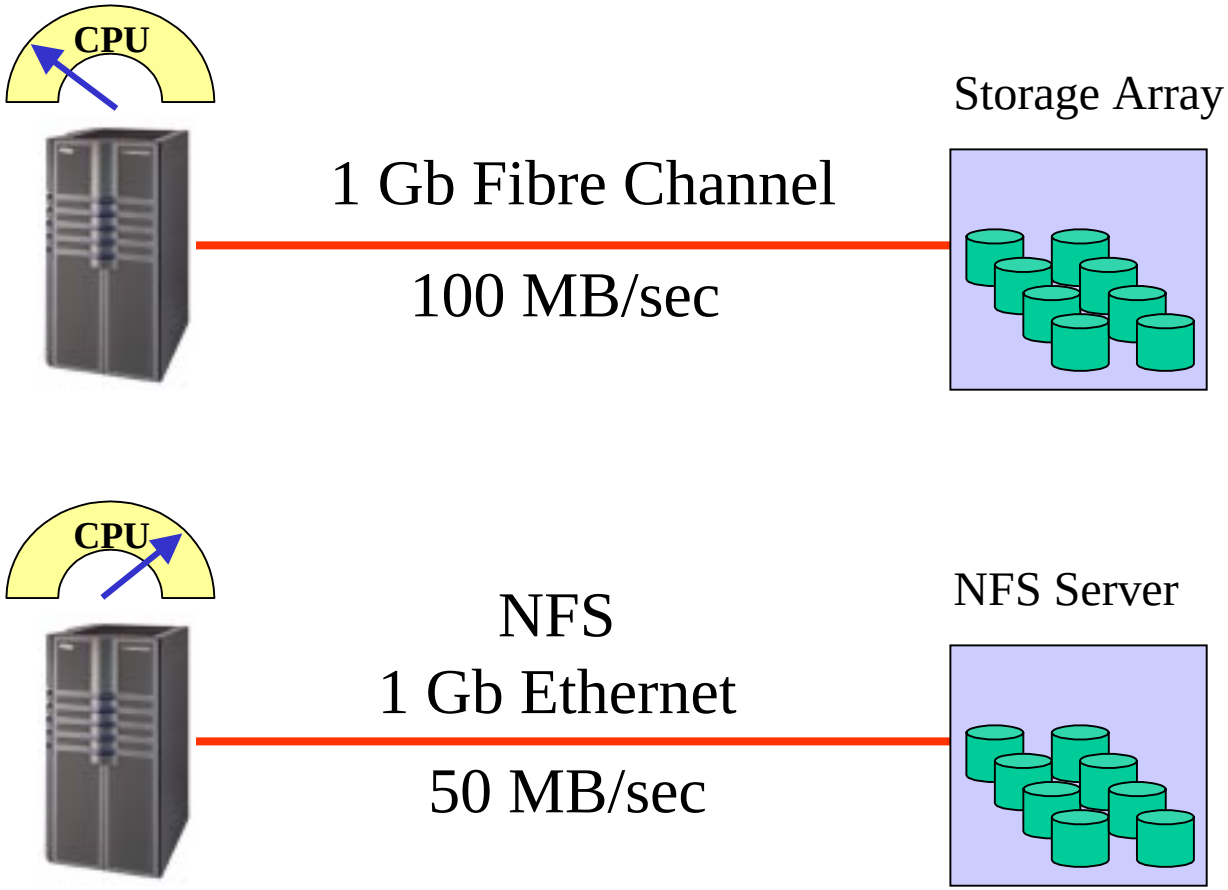
Data Center Network

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

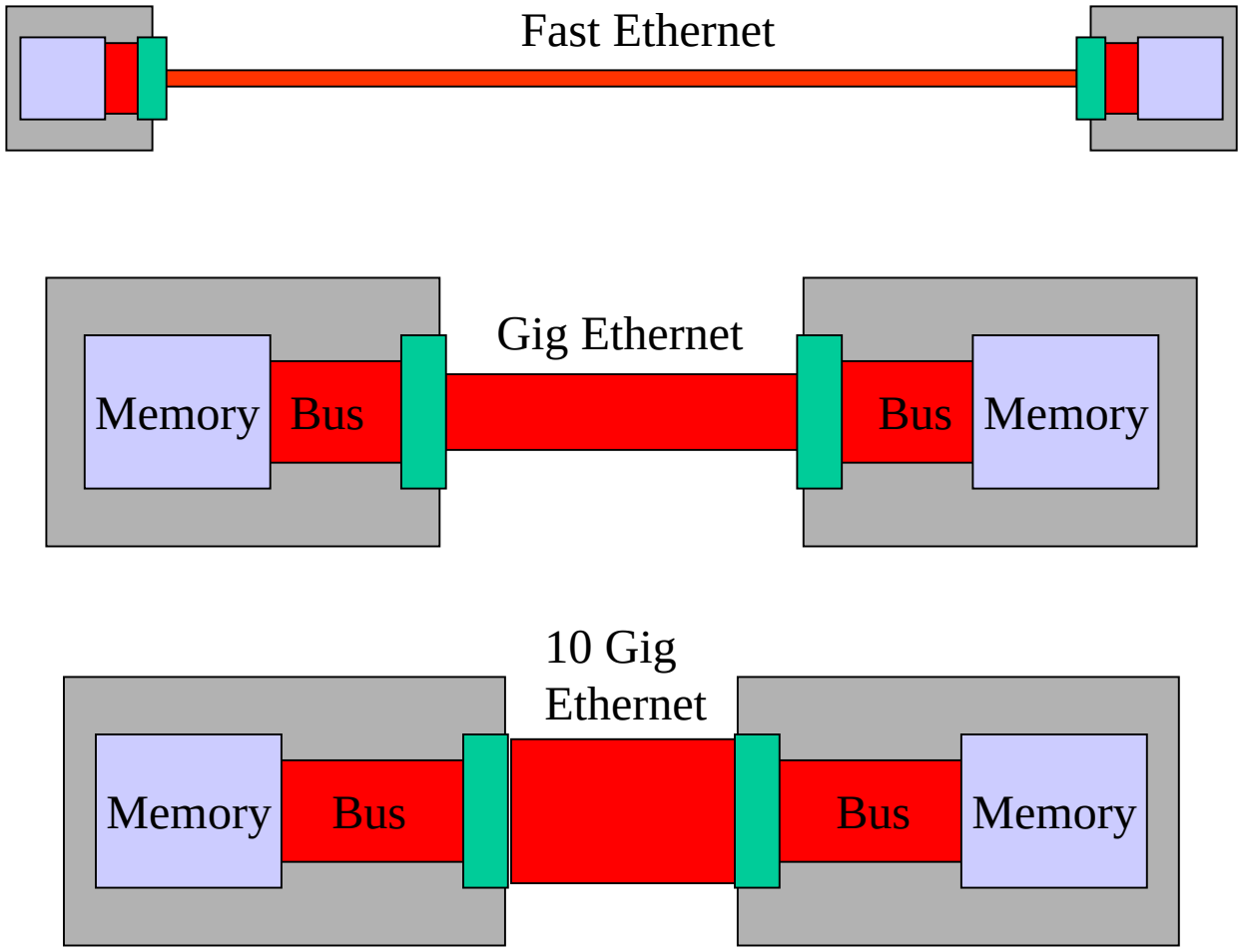


NFS INDUSTRY CONFERENCE

A Problem: Data Center Performance

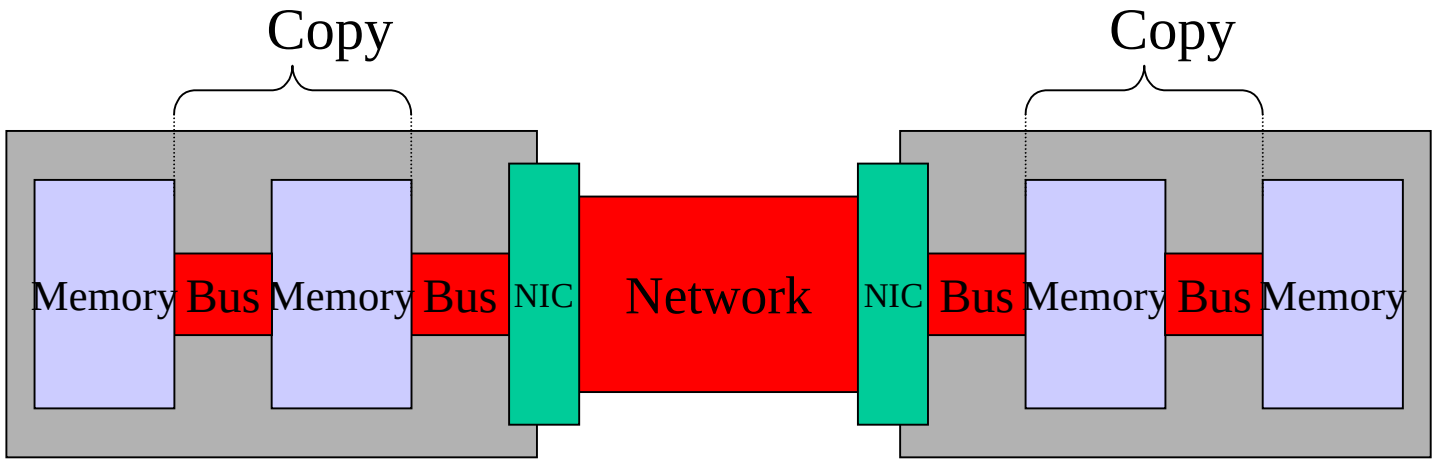


Network vs Bus Performance



NFS INDUSTRY CONFERENCE

Network vs Bus Performance

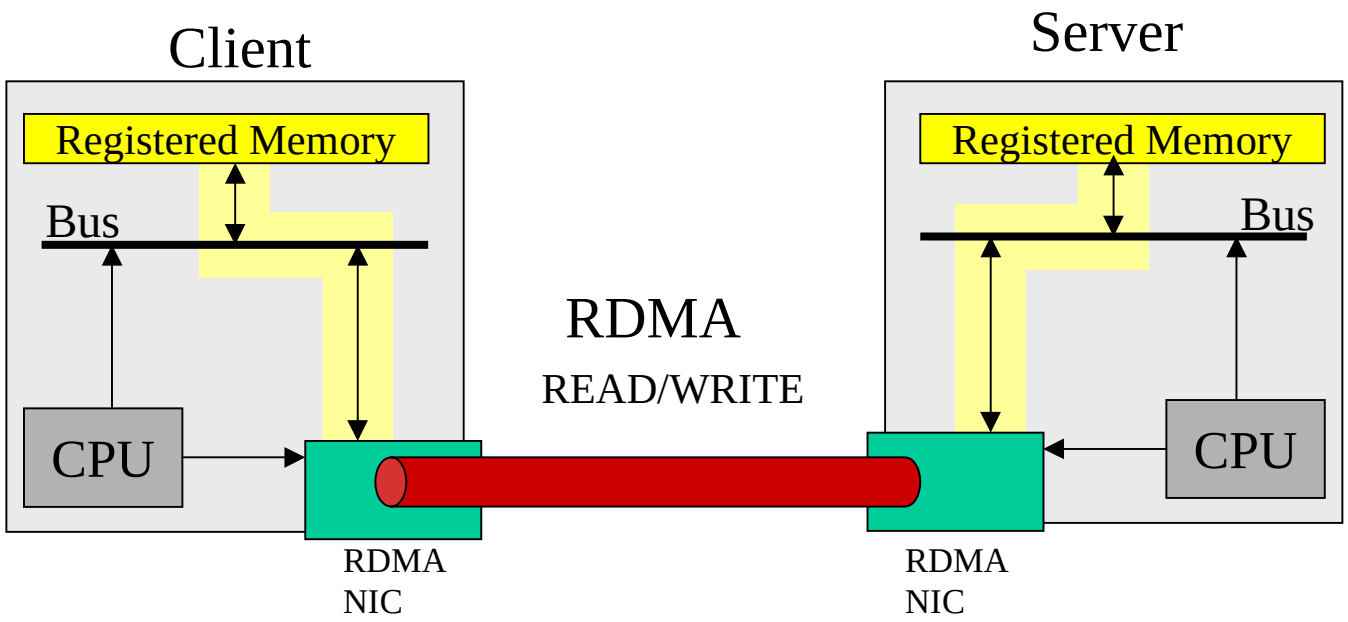


Latency gets worse with each memory copy which further loads the CPU.

NFS INDUSTRY CONFERENCE

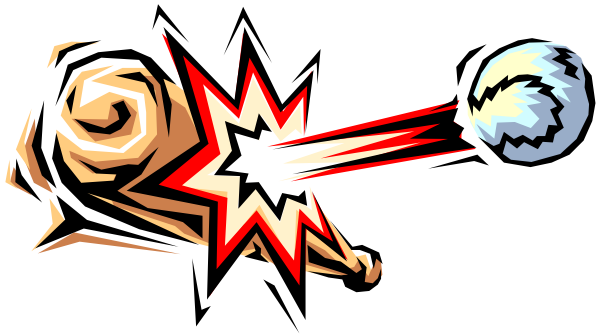
What is RDMA ?

- DMA: Direct Memory Access
- RDMA: *Remote* Direct Memory Access
- Supports Direct Placement
- Networking offload for CPU



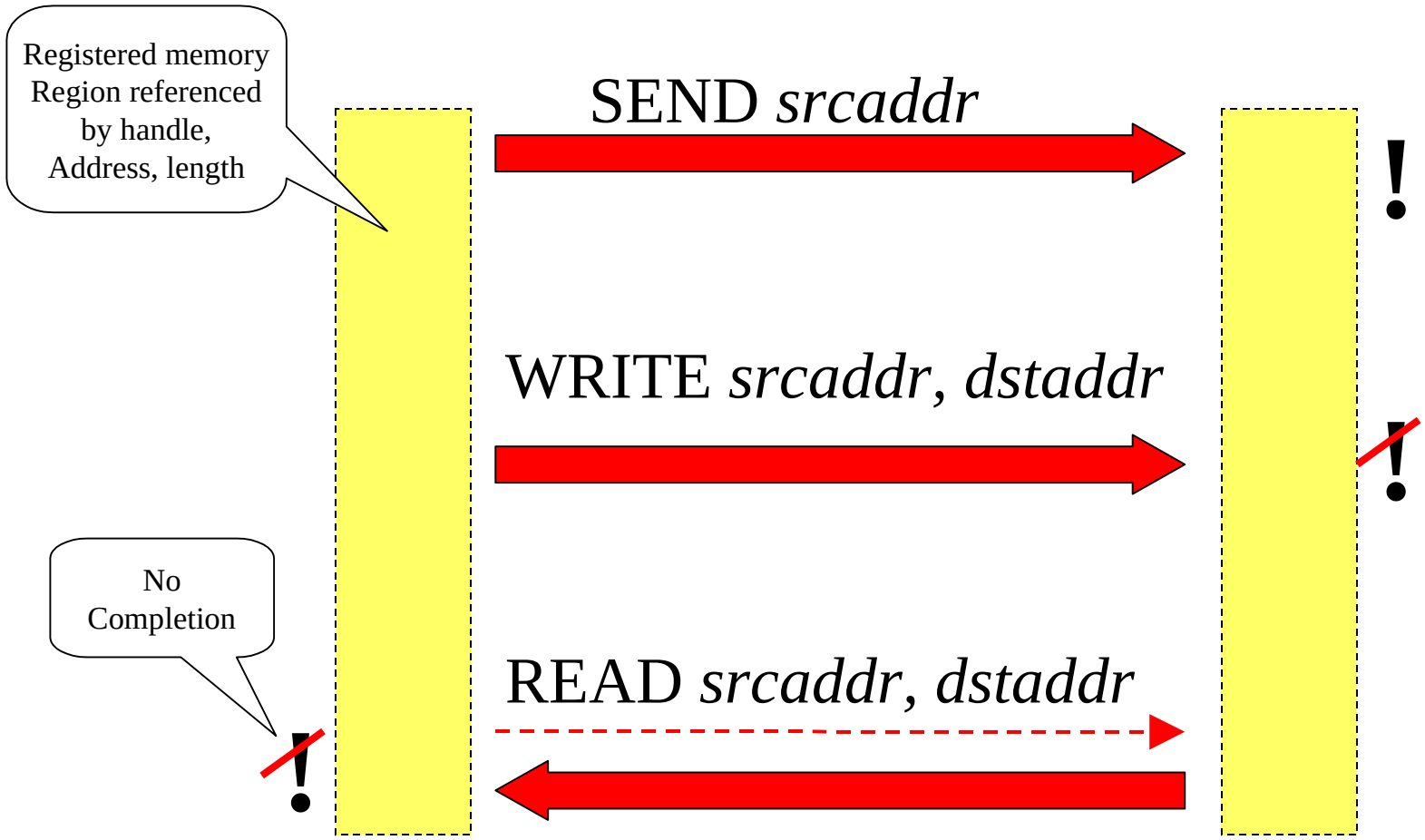
NFS INDUSTRY CONFERENCE

RDMA Sweet Spot



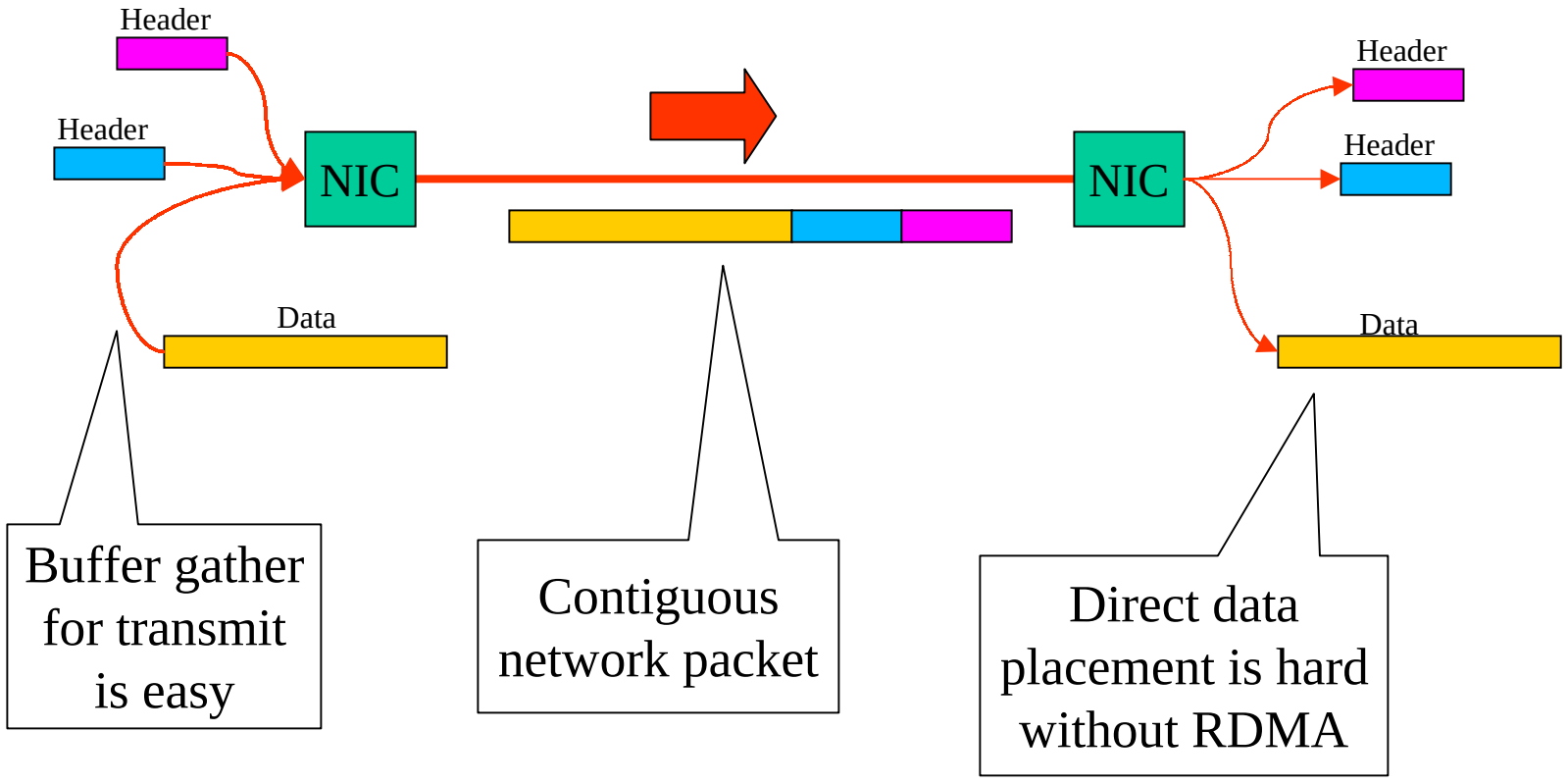
- High bandwidth
  - > 1Gb links
- Big chunks of data
  - More than 1KB
- Short distance (low latency)
  - 10's of Meters
- Busy CPU
- Protocols
  - NFS, replication, database, backup

RDMA Operations



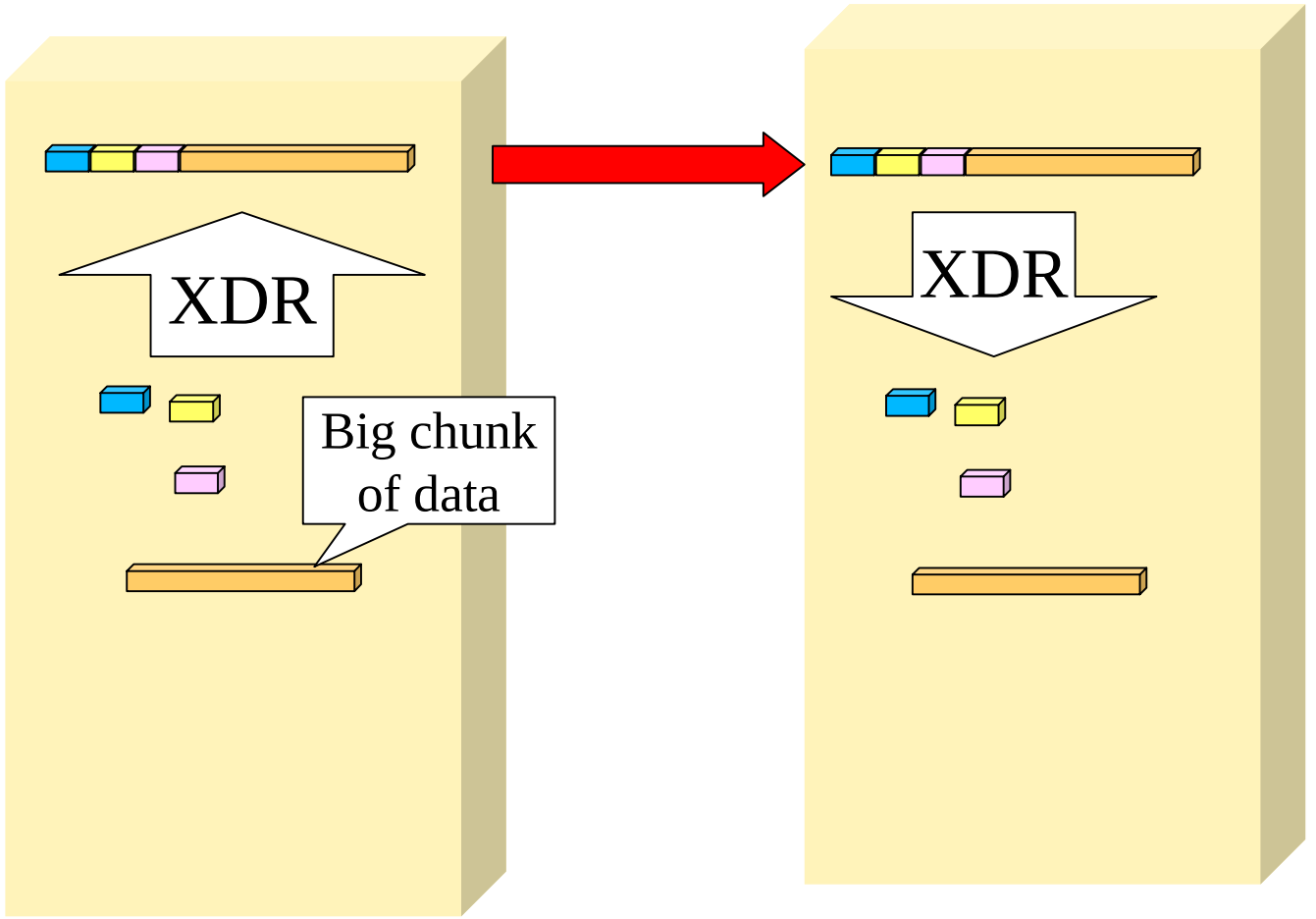
NFS INDUSTRY CONFERENCE

RDMA Provides Direct Data Placement

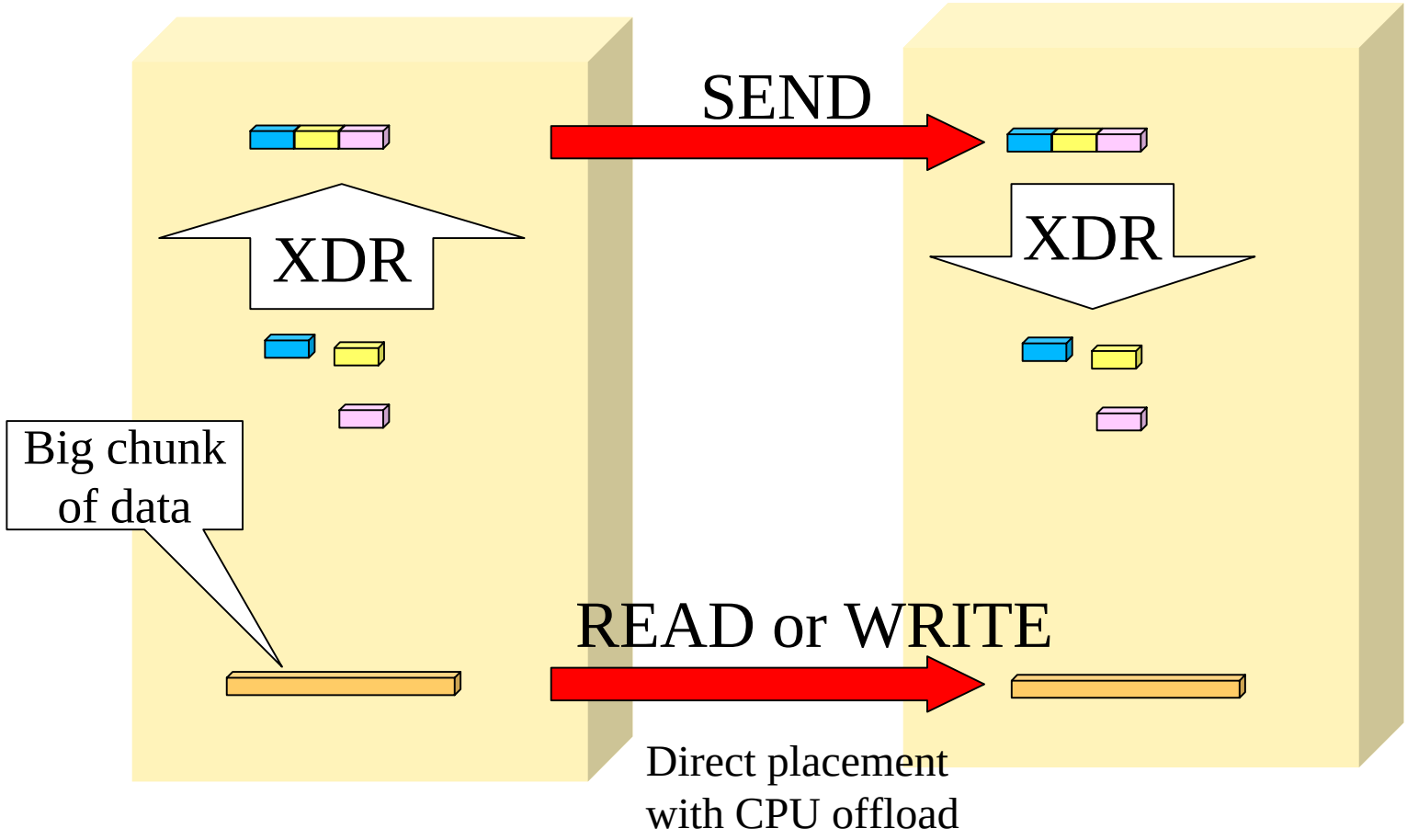


NFS INDUSTRY CONFERENCE

Conventional RPC Data Movement



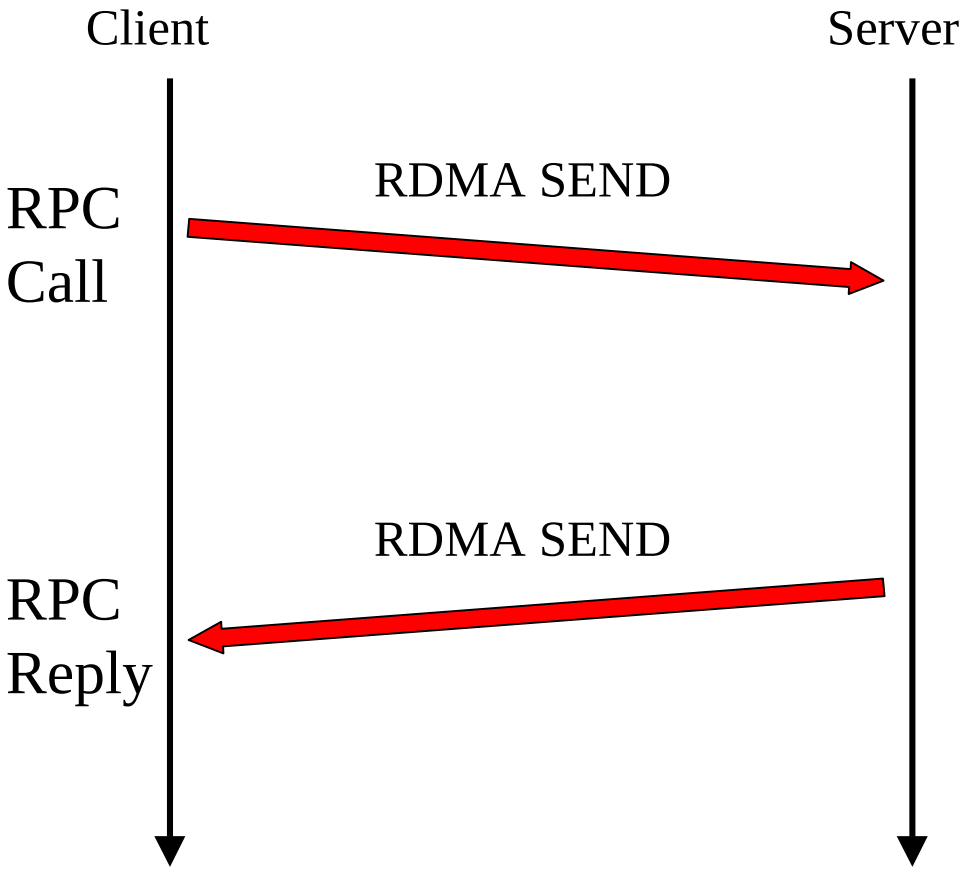
RDMA RPC Data Movement



NFS INDUSTRY CONFERENCE

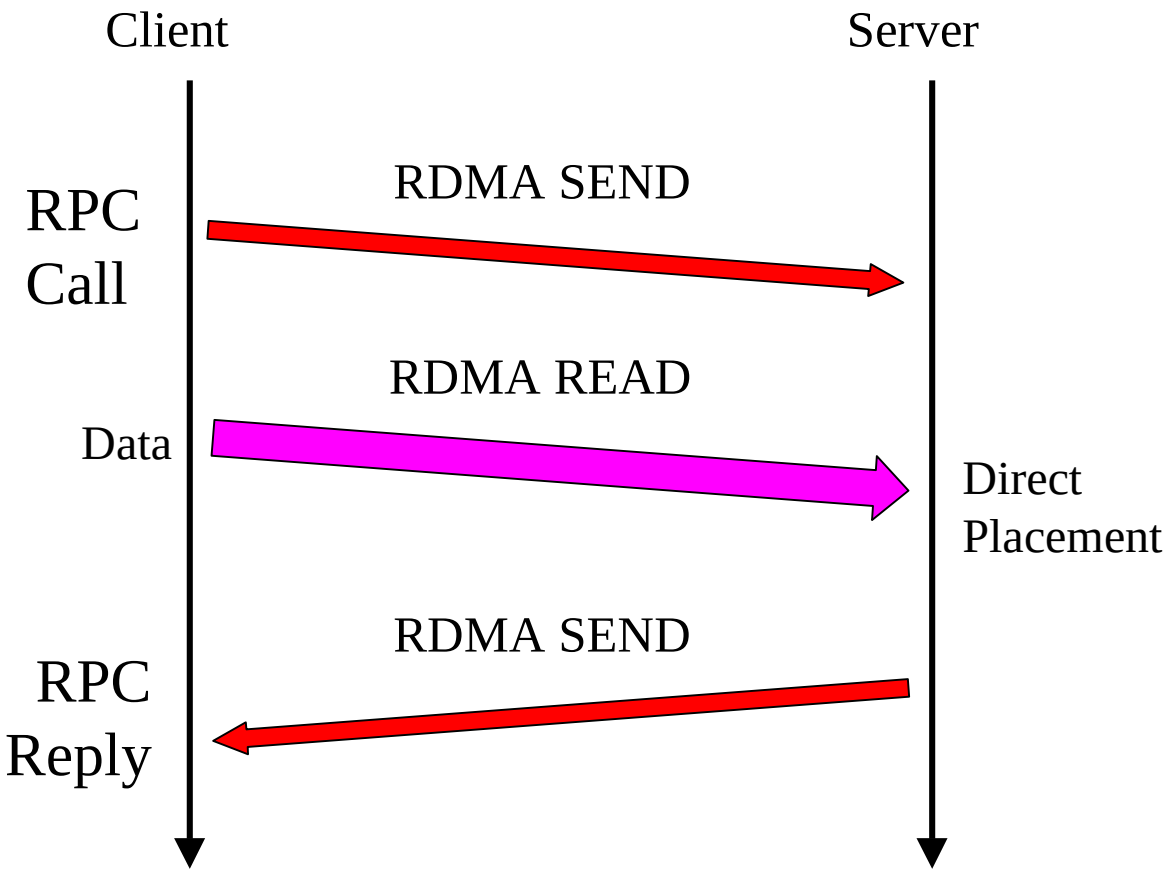
Small RPC Messages

Most RPC Messages are small.
Examples:
- LOOKUP
- GETATTR



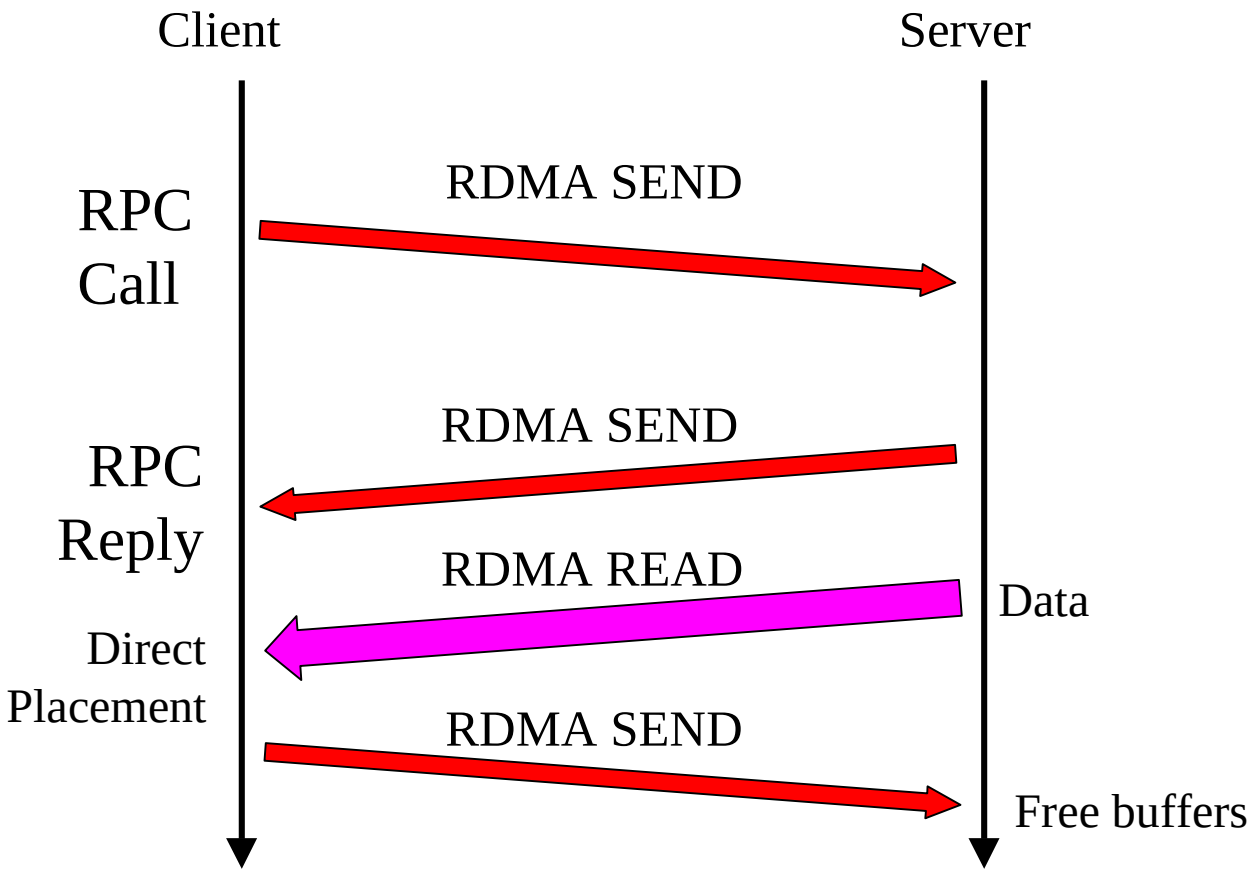
Big RPC Call

Example:
WRITE

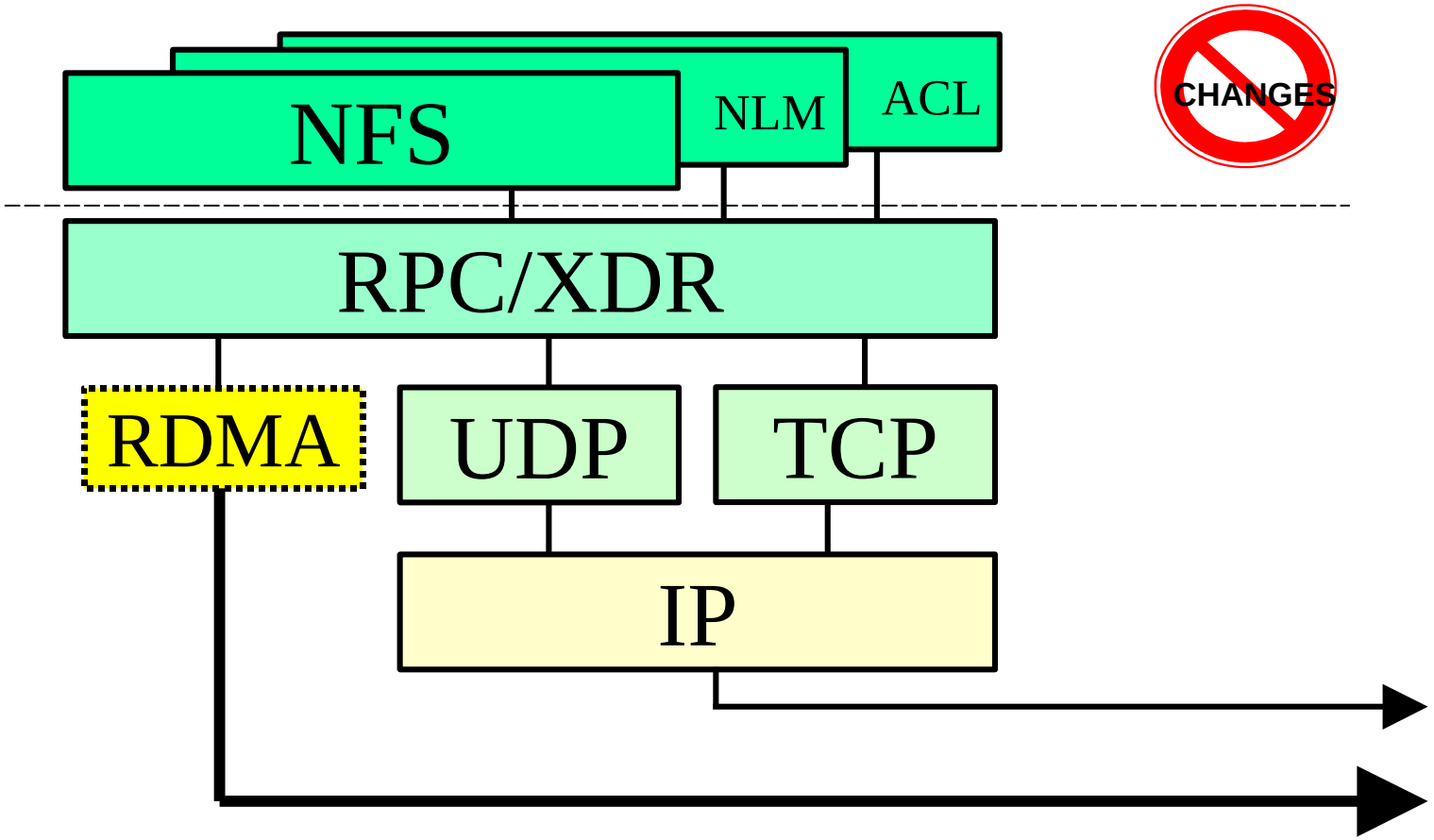


Big RPC Reply

Example: READ



Adapting NFS to RDMA



NFS INDUSTRY CONFERENCE

RDMA Flavors

- On Ethernet
  - Emulex GN9000/VI - VI/TCP via 1Gb fibre
  - IETF RDDP Working Group
    - An interoperable, Internet standard
    - RDMA via a Direct Data Placement (DDP) layer
    - Defined for SCTP and TCP
- On Infiniband
  - Supported natively by all IB hardware
- Other
  - Myrinet, Fibre channel, CLan, ...

NFS INDUSTRY CONFERENCE

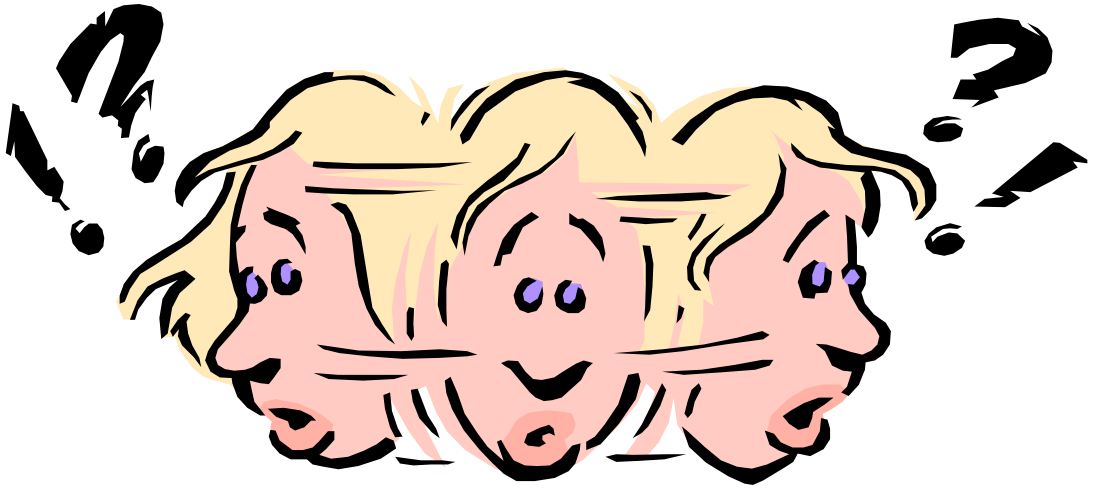
Solaris Prototype

- Extension to Solaris kernel RPC
  - Supports all NFS versions, 2, 3 & 4
  - All kernel RPC: Lock Manager, NFS_ACL
- Behaves like a normal NFS mount
  - only a LOT faster!
- Supporting two RDMA flavors
  - kVIPL with Emulex GN9000/VI over Gigabit Ethernet
  - Infiniband

NFS INDUSTRY CONFERENCE

NFS/RDMA Standard

- NFS will continue to be an open, interoperable protocol!
- Need to publish standards for doing NFS/RDMA via Ethernet & Infiniband
- Storage Network Industry Association (SNIA)
  - NFS/RDMA Technical working group
- Abstract NFS/RDMA protocol
  - based on generic RDMA operations & features
- Transport mappings
  - for NFS/RDDP & NFS/Infiniband



Questions & Answers