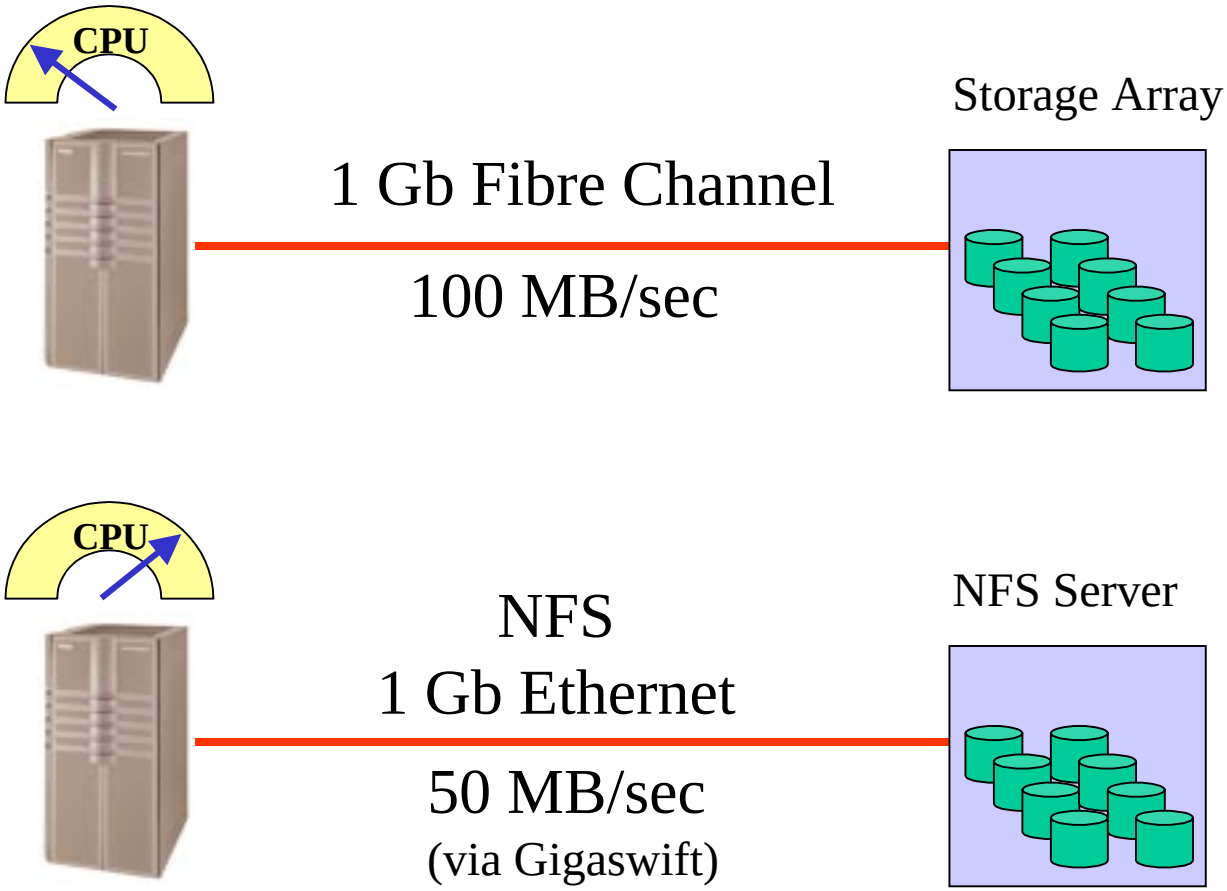
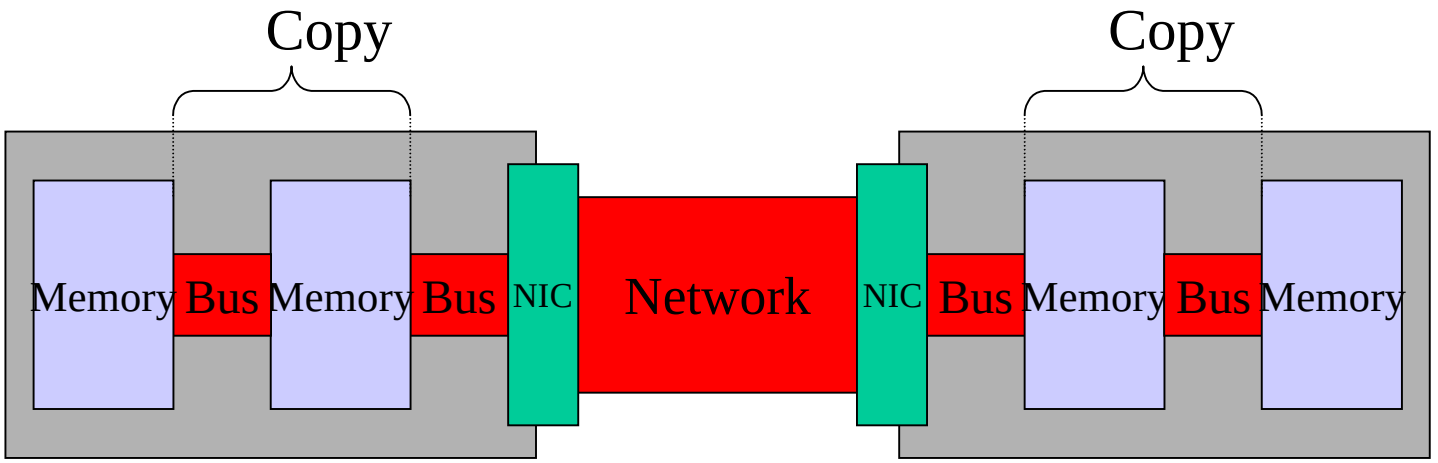
NFS over RDMA

Brent Callaghan
Sun Microsystems, Inc.
brent@eng.sun.com

A Problem:
Data Center Performance



Network vs Bus Performance

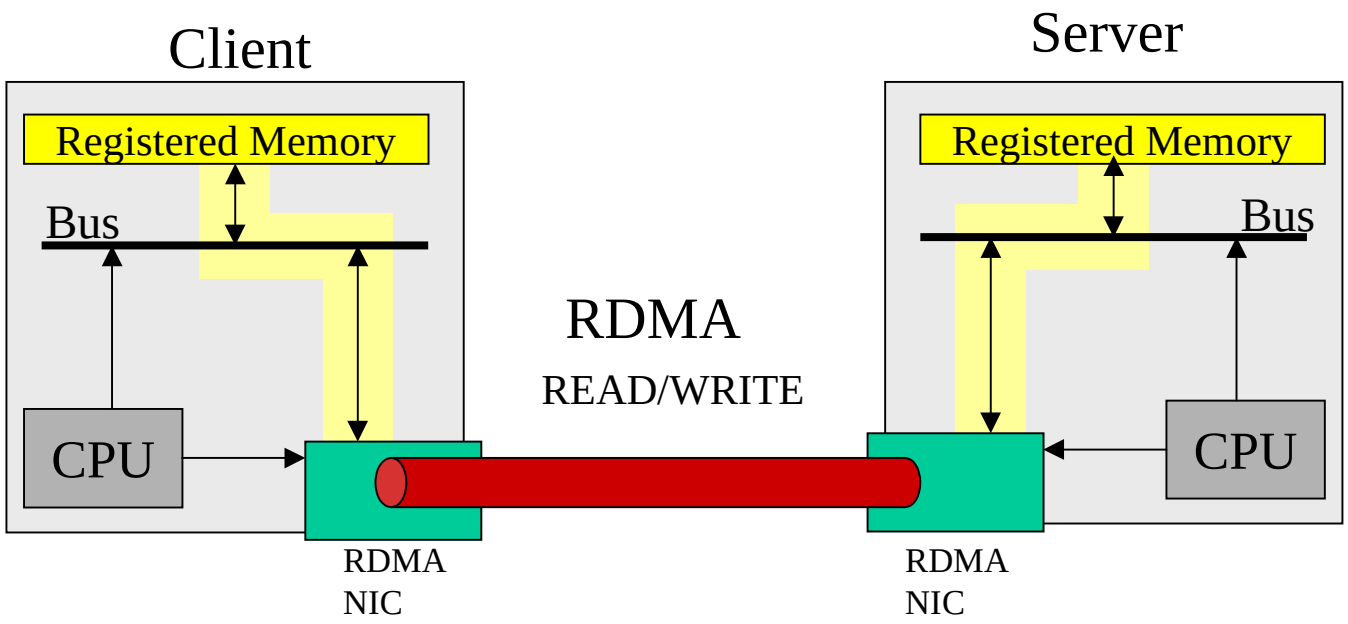


Latency gets worse with each memory copy

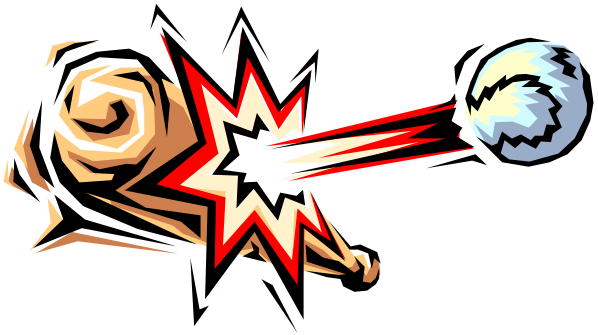
which further loads the CPU.

What is RDMA ?

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

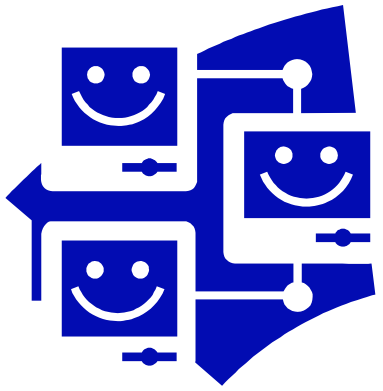


RDMA Sweet Spot



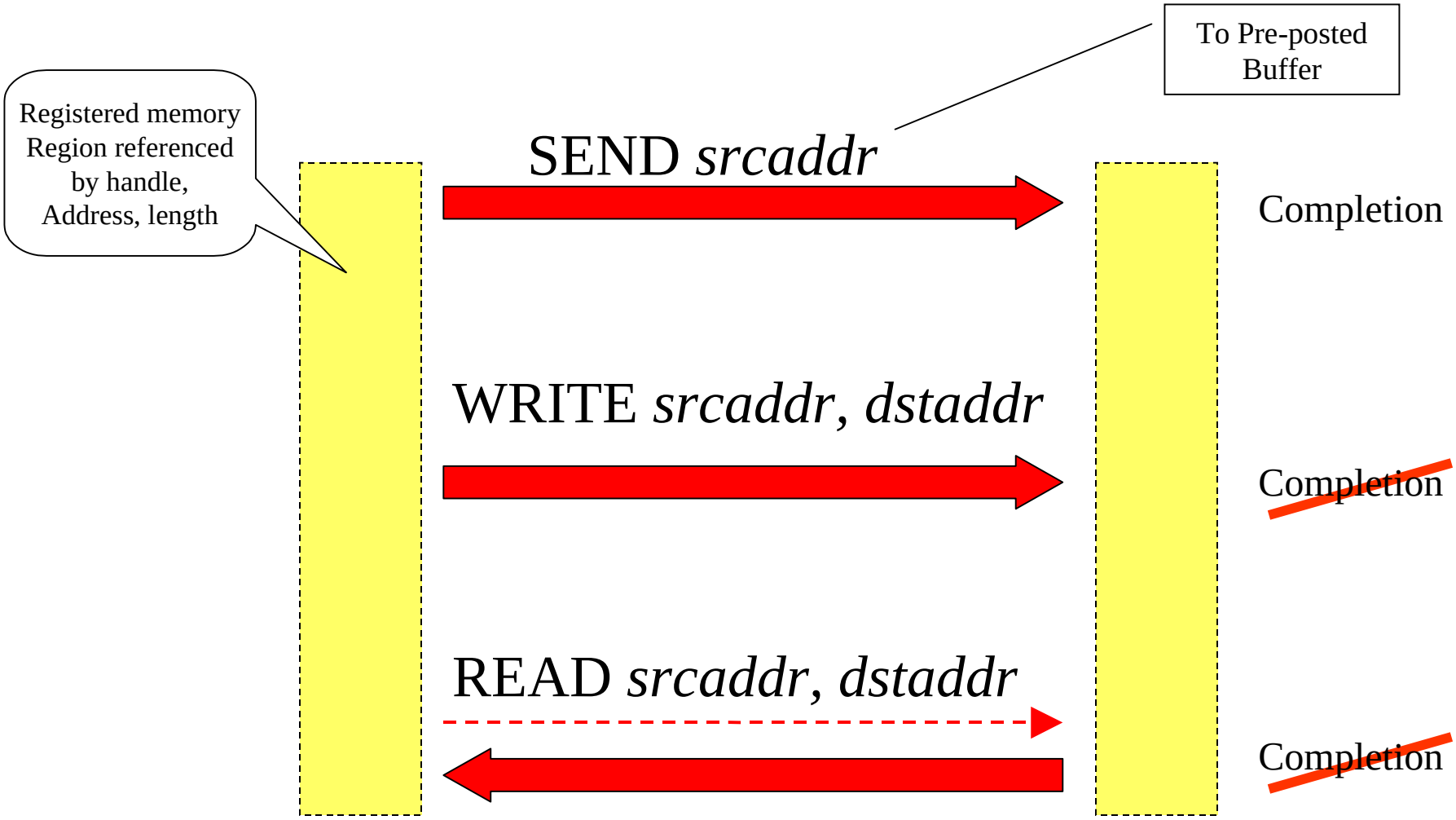
- High bandwidth
  - > 1Gb links
- Big chunks of data
  - More than 1KB
- Short distance (low latency)
  - 10’s of Meters
- Busy CPU

Protocols Using RDMA

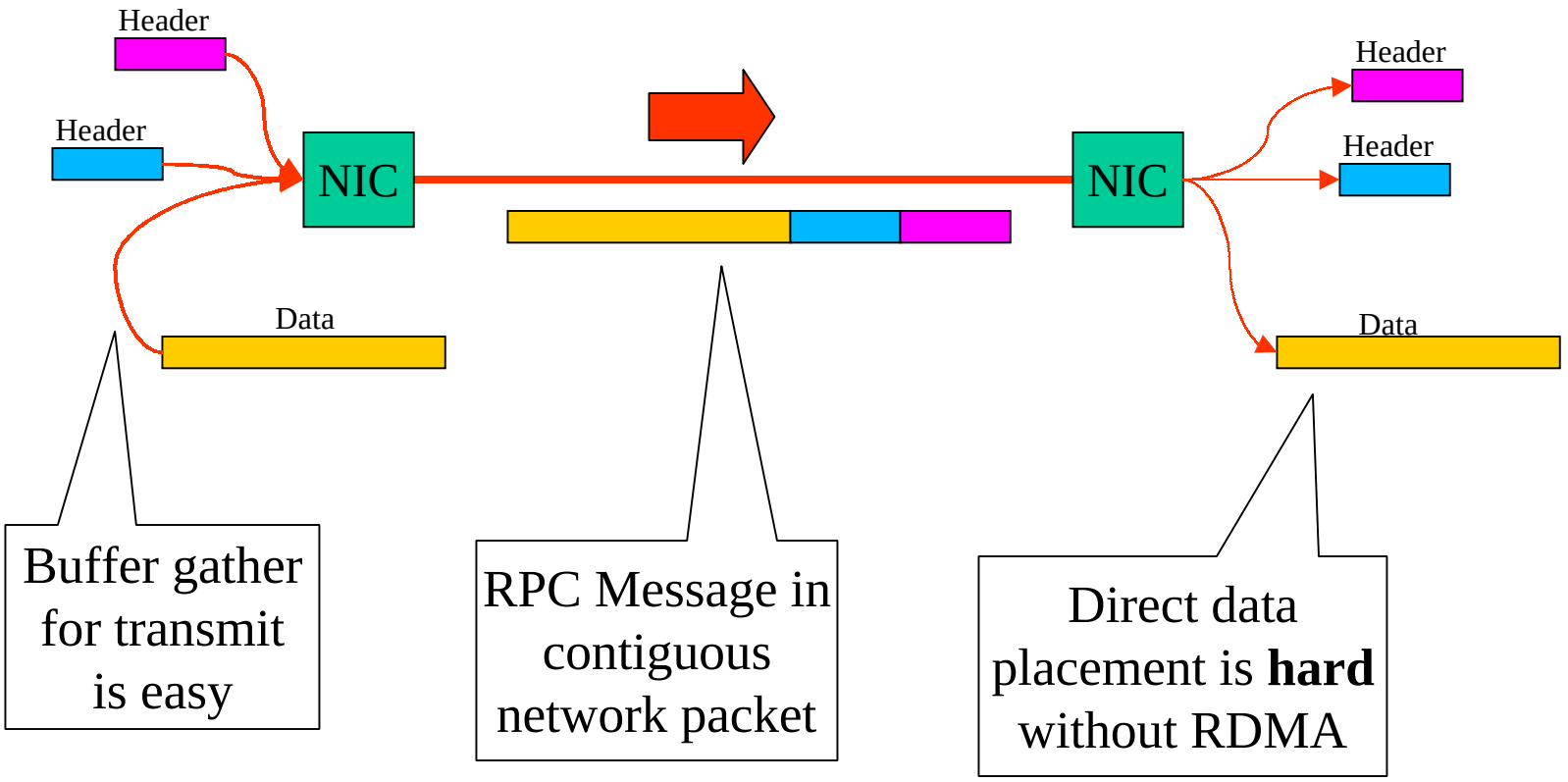


- NFS/RDMA
  - Benefits all RPC based applications
- Sockets Direct Protocol (SDP)
  - Make RDMA accessible to sockets based apps
- SCSI RDMA Protocol (SRP)
  - Needed for IB. Competition for iSCSI ?
- Direct Access File System (DAFS)
  - Based on NFS version 4
  - DAFS Database Accelerator: TPC-C results are “best price/performance for a UNIX based system.”

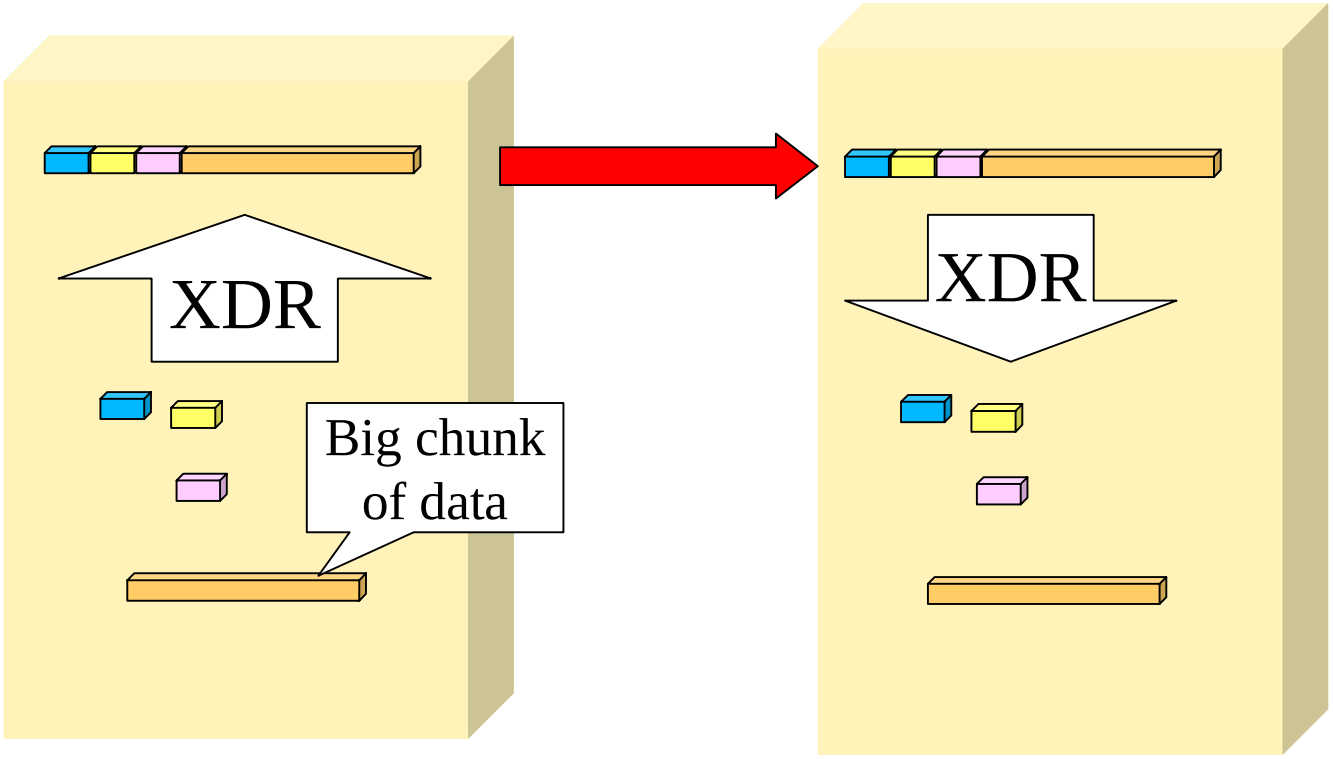
RDMA Operations



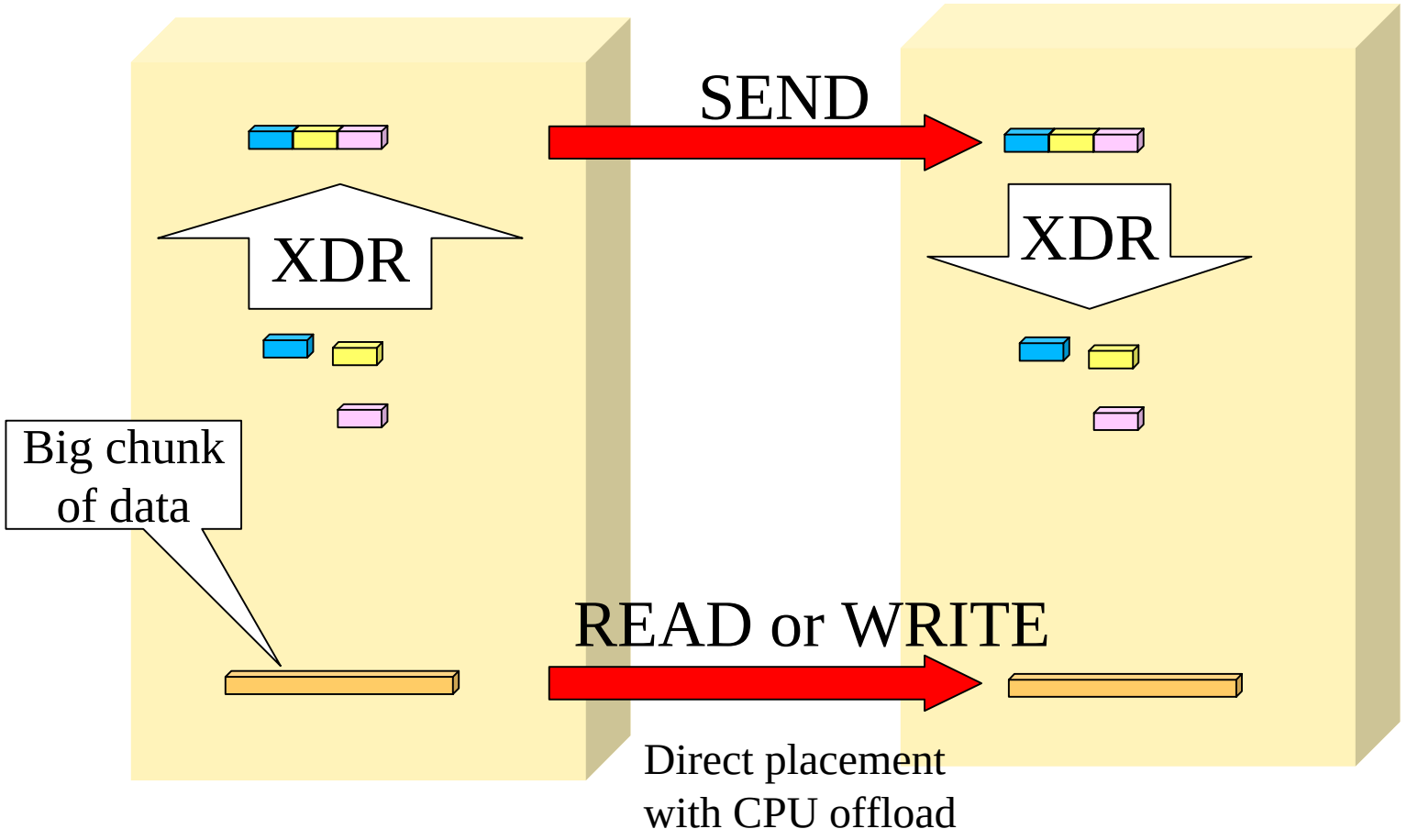
RDMA Provides Direct Data Placement



Conventional RPC Data Movement

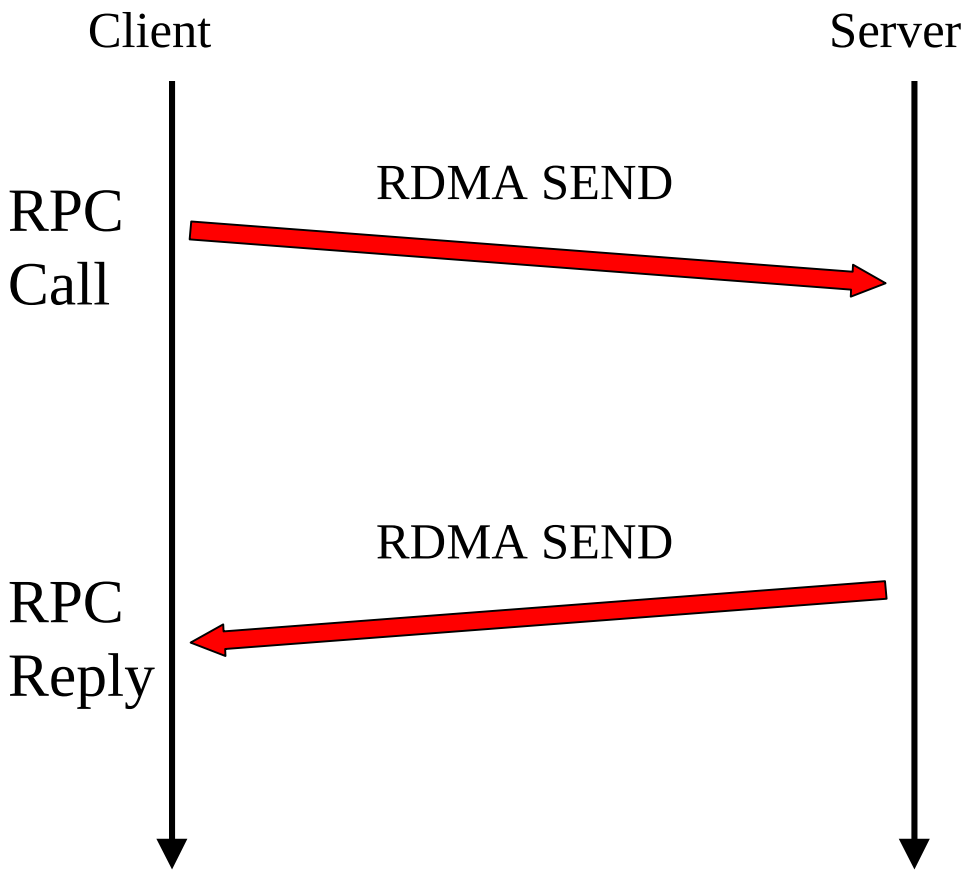


RDMA RPC Data Movement



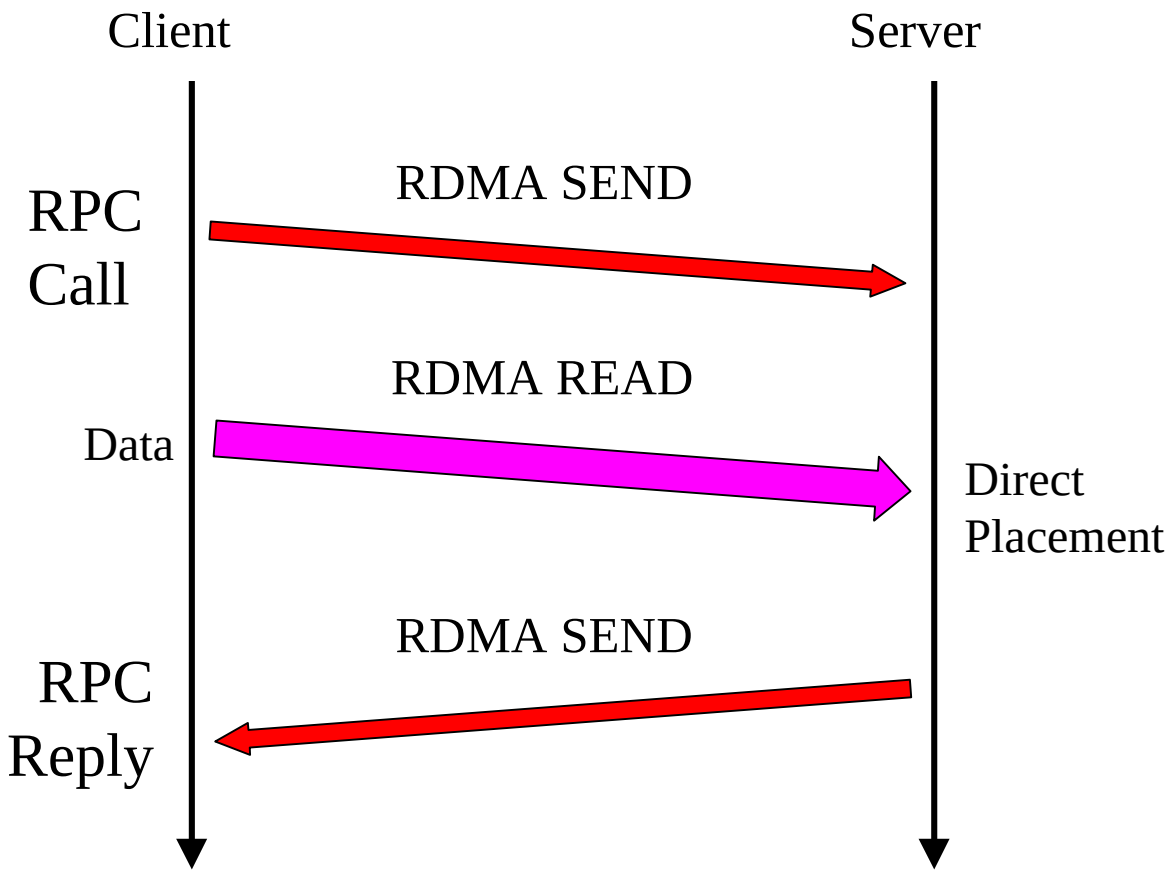
Small RPC Messages

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



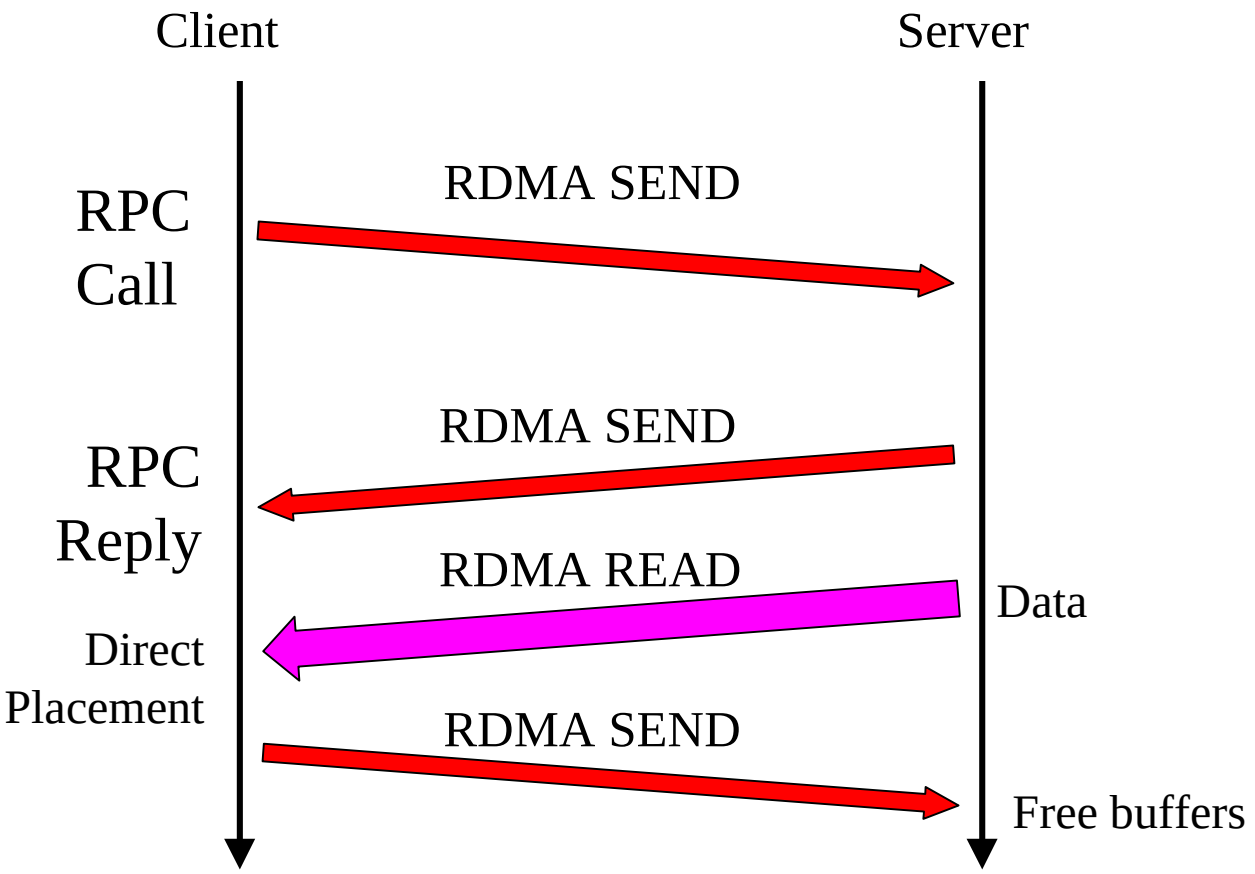
Big RPC Call

Example:
NFS WRITE

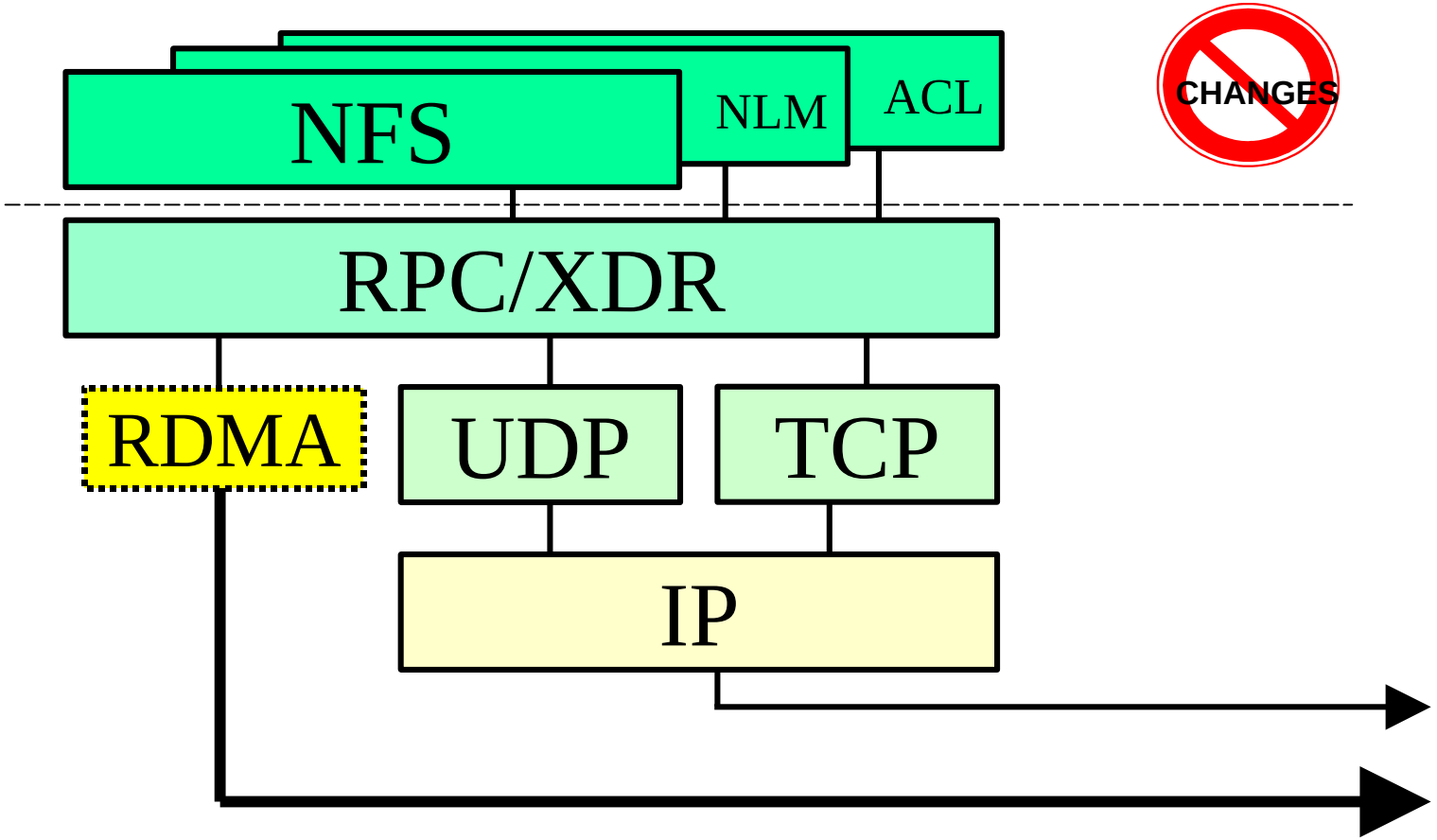


Big RPC Reply

Example:
NFS READ



Adapting NFS to RDMA



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 more efficient!
- Supporting two RDMA flavors
  - kVIPL with Emulex GN9000/VI over Gigabit Ethernet
    - File copy: 103 MB/sec vs 60 MB/sec over Gigaswift
  - Infiniband
    - Now testing with Mellanox Tavor cards

RDMA Flavors

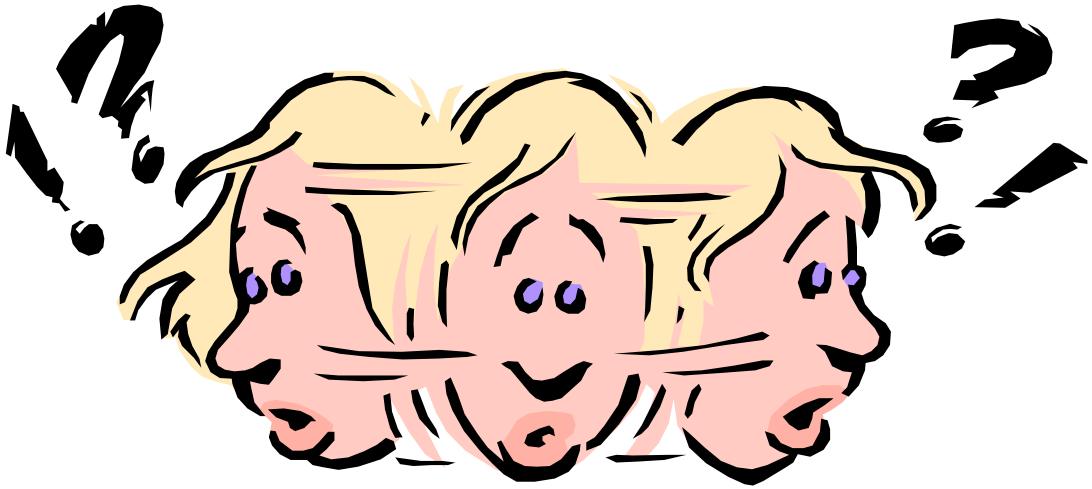
- On Ethernet
  - Emulex GN9000/VI - VI/TCP via 1Gb fibre
  - RDMA Consortium
    - RDMA over TCP
  - IETF RDDP Working Group
    - An interoperable, Internet standard
    - Defined for SCTP and TCP
- On Infiniband
  - Supported natively by all IB hardware
- Other
  - Myrinet, Fibre channel, CLan, ...

RDMA API Standardization

- Windows Sockets Direct
  - Extensions to Winsock for RDMA
  - Uses SDP
- UNIX Sockets Extensions
  - Open Group's ICSC WG, Sockets API Extensions
- Direct Access Transport (DAT)
  - Replaces VI API
  - uDAPL & kDAPL
- Interconnect Transport API (ICSC/Open Group)
  - Based on DAPL

NFS/RDMA Standard

- NFS will continue to be an open, interoperable protocol!
- Need to publish standards for doing NFS/RDMA via Ethernet & Infiniband
- IETF has a role in RDMA protocol development
  - RDDP Working Group
- IETF hosts standards for ONC RPC & NFS
- Proposal to extend NFSv4 charter to include NFS/RDMA protocol standard.



Questions & Answers