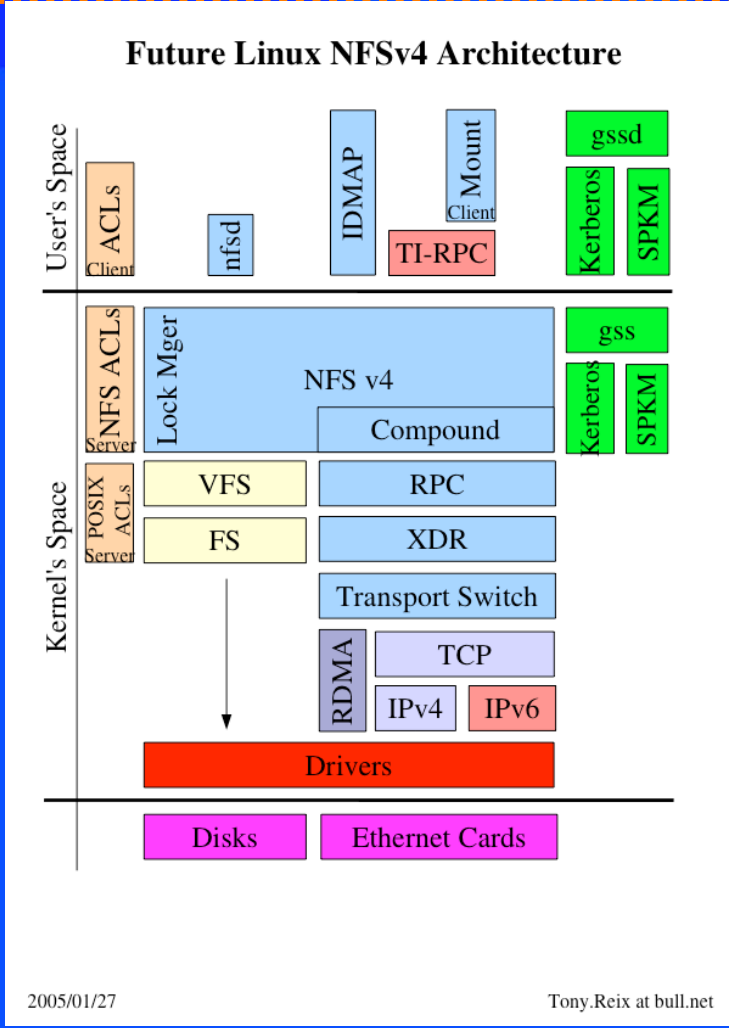
Linux NFS/RDMA Status

Charles J. Antonelli
Center for Information Technology Integration
University of Michigan, Ann Arbor
August 22, 2005

NFS/RDMA

- NFS v2/v3/v4 over RDMA
- Greatly enhanced NFS performance
  - Low overhead
  - Full bandwidth
  - Direct, zero-copy I/O
- Implemented on Linux
  - kDAPL API

NFSv4


Future Linux NFSv4 Architecture
User's Space
Kernel's Space
ACLs
Client
nfsd
IDMAP
Mount
Client
TI-RPC
gssd
Kerberos
SPKM
NFS ACLs
Server
Lock Mger
NFS v4
Compound
gss
Kerberos
SPKM
POSIX ACLs
Server
VFS
RPC
FS
XDR
Transport Switch
RDMA
TCP
IPv4
IPv6
Drivers
Disks
Ethernet Cards
2005/01/27
Tony.Reix at bull.net

Linux NFS/RDMA Server Approach

- RDMA within RPC layer
  - New kernel RPC transport type
  - Direct data transfer between client memory and server buffers
  - Socket-specific NFS kernel code replaced by general interface implemented by socket or RDMA transports
- kDAPL RDMA API (Mellanox stack)
- NFS client code otherwise unchanged
- Transparent to application

kDAPL

- Kernel RDMA via kDAPL
- Very simple subset of kDAPL 1.1 API
  - Connection, connection DTOs
  - Kernel-virtual or physical LMRs, RMRs
  - Small (1KB-4KB typical) send/receive
  - Large RDMA (4KB-64KB typical)
    - All RDMA read/write initiated by server

Linux NFS/RDMA Server

- RPC/RDMA connections implemented
- RPC/RDMA inline requests being implemented
  - Server NFS layer receives requests over RDMA
- NFSv3/v4 RDMA

Linux NFS/RDMA Project

- Linux NFS/RDMA Server
- Demonstrate NFS/RDMA functionality on multiple platforms and network technologies
  - IA64 (SGI Altix)
  - iWarp (Ammasso)
- SC'05
- OpenIB

OpenIB

- Currently developing on proprietary stack
  - … tactical
- Clear direction to OpenIB
  - … strategic
  - … give us kDAPL

CITI

- Developing NFSv4 reference implementation since 1999
  - NFS/RDMA and NFSv4.1 Sessions since 2003
- Funded by Sun, Network Appliance, ASCI, PolyServe, NSF, SGI, Ammasso
- http://www.citi.umich.edu/projects/nfsv4/

Any questions?

http://www.citi.umich.edu/