A Tour of the Linux OpenFabrics Stack

Johann George, QLogic

June 2006

Overview

- Beyond Sockets
  - Provides a common interface that allows applications to take advantage of the RDMA (Remote Direct Memory Access), low latency and high messaging rate capabilities provided by the current generation of networking hardware.
- Standards Compliant
  - Encompass both the InfiniBand and iWARP standards.
- Widely Available
  - Incorporated in the Linux Kernel since 2.6.11.

What's Wrong With Sockets?

- The Berkeley Socket Interface has lasted for more than 20 years and is not going away ... but
  - Heavily oriented towards TCP/IP and UDP.
  - No native support for RDMA.
  - Asynchronous I/O is not naturally built in.
  - Sub-optimal latencies as it does not take advantage of efficiencies provided by modern networking hardware.
  - Hard to provide an interface from the hardware directly to user space.

InfiniBand and iWARP

- With networking bandwidth approaching 10 gigabits per second, two standards evolved in an attempt to fully utilize such capabilities: InfiniBand and iWARP.
- Both defined a wire protocol. Different vendor's hardware could inter-operate.
- Both provide a loosely defined application interface called Verbs.
- Many similarities between the two interfaces.

InfiniBand

- Infinite Bandwidth.
- A whole new networking infrastructure which began in 1999 as a merger of two other technologies: Future I/O and NGIO.
- Originally intended to be within a data center. Maximum copper cable length is still 15 metres (almost 50 feet).
- InfiniBand Verbs is an interface loosely specified by the InfiniBand Trade Association (IBTA) that provides a common interface to devices that support the InfiniBand protocol.

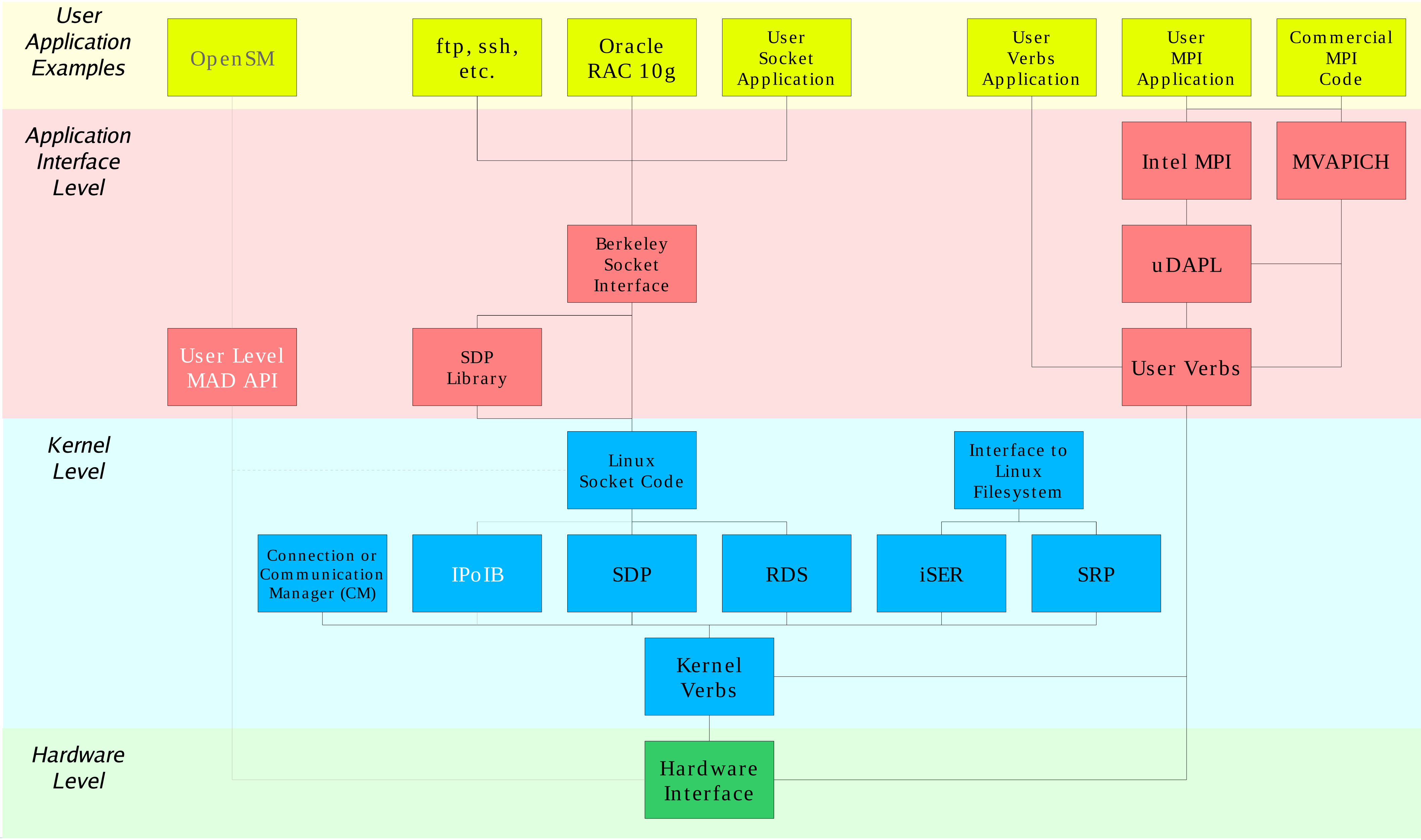
iWARP

- No acronym. ... Internet Wide Area RDMA Protocol.
- Original motivation was to provide an interface that ran over the TCP wire protocol which allowed applications to take advantage of hardware RNICs that provided RDMA features. IETF standard.
- Encompasses the RDMA Verbs and the RDMA over DDP (Direct Data Placement) supported over
  - SCTP
  - TCP wire protocol using MPA (Marker-based PDU (Protocol Data Unit) Aligned Framing) protocols
- Specified by the RDMA Consortium in October, 2002.

Features of the OpenFabrics Stack

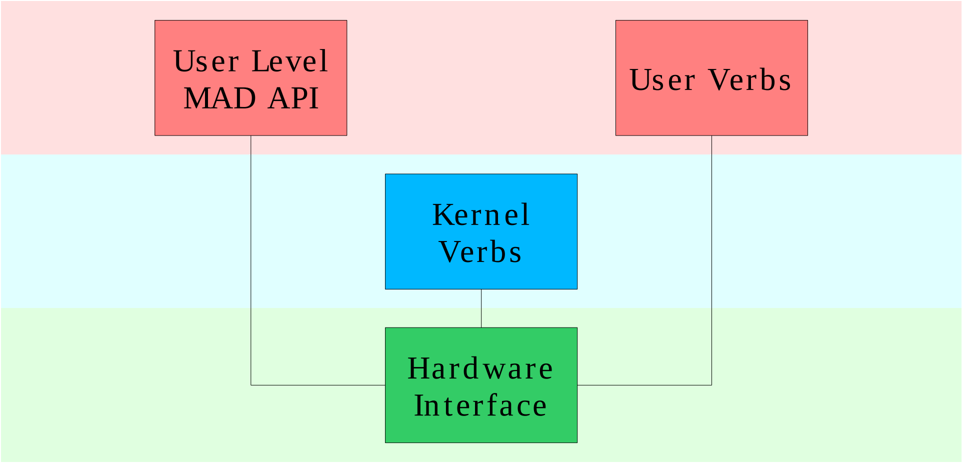
- Provides a common API that can be used whether the underlying transport is InfiniBand or iWARP.
- The same API can also be used on MS-Windows when the underlying transport is InfiniBand.
- Note that the wire protocol between InfiniBand and iWARP is different. An InfiniBand HCA currently cannot inter-operate with an iWARP RNIC. This may be possible in the future with an intelligent bridge.

Linux OpenFabrics Stack



Hardware Interface

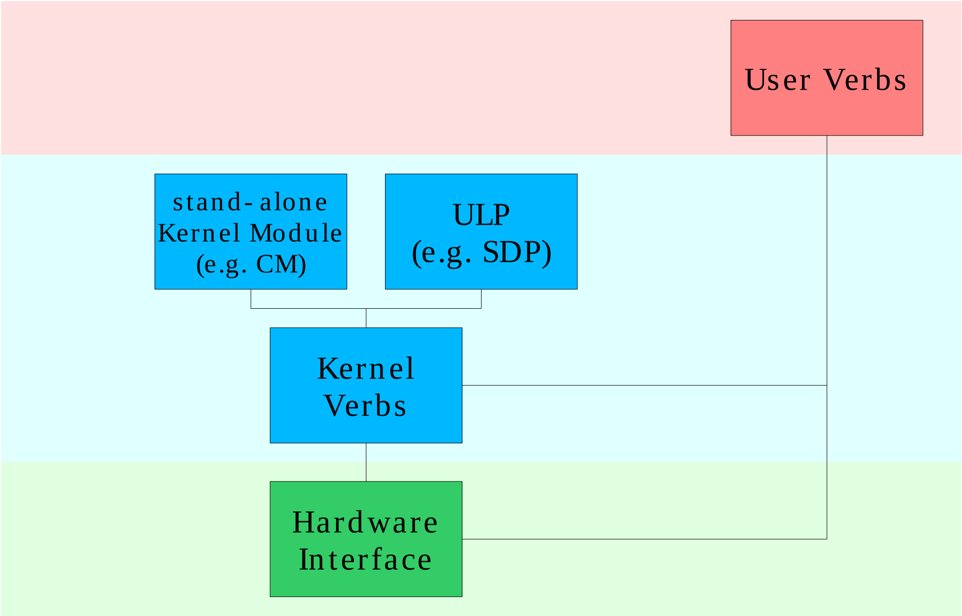
- The hardware specific device driver together with the OpenFabrics stack provides two interfaces to the upper layers: the Kernel Verbs and the User Verbs.
- When the hardware is interfacing to an InfiniBand fabric, a third interface is provided: the User Level MAD API.
- The Hardware Inteface provides the connection to the fabric; either InfiniBand, or in the case of iWARP, TCP/IP to an ethernet fabric.



Kernel Verbs

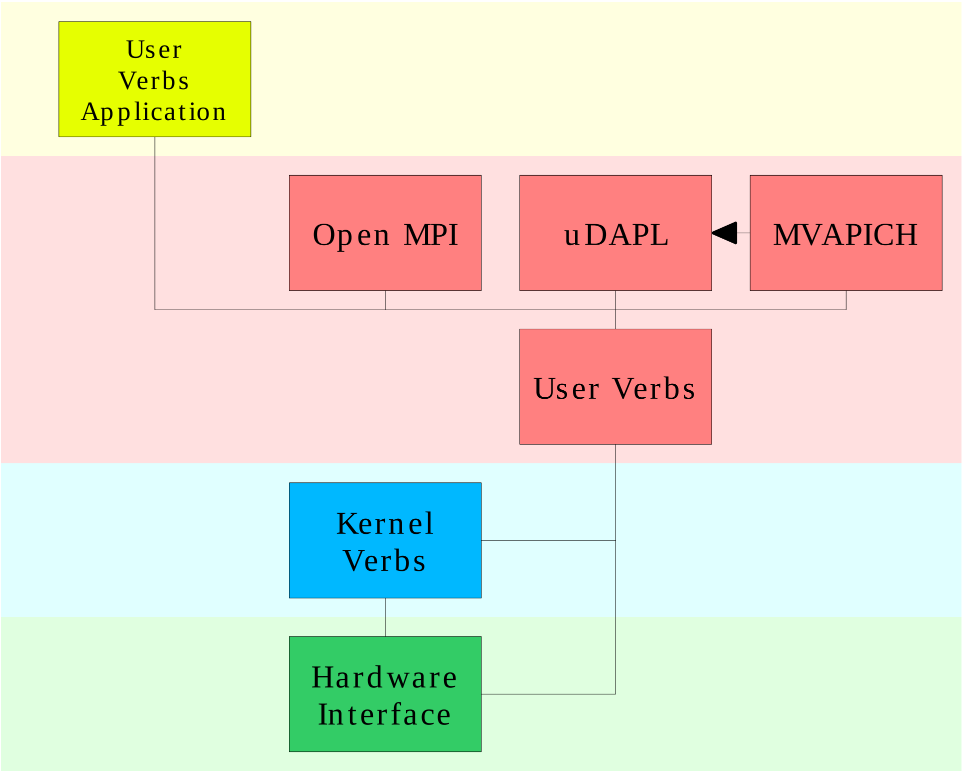
- Modules residing in the kernel.
- Usually used to implement the Upper Level Protocols (ULPs).
- Also used to implement stand-alone kernel modules such as the CM.
- Sometimes used to help implement portions of the User Verbs.

- There are three classes of Kernel Verbs: generic, InfiniBand specific and iWARP specific.



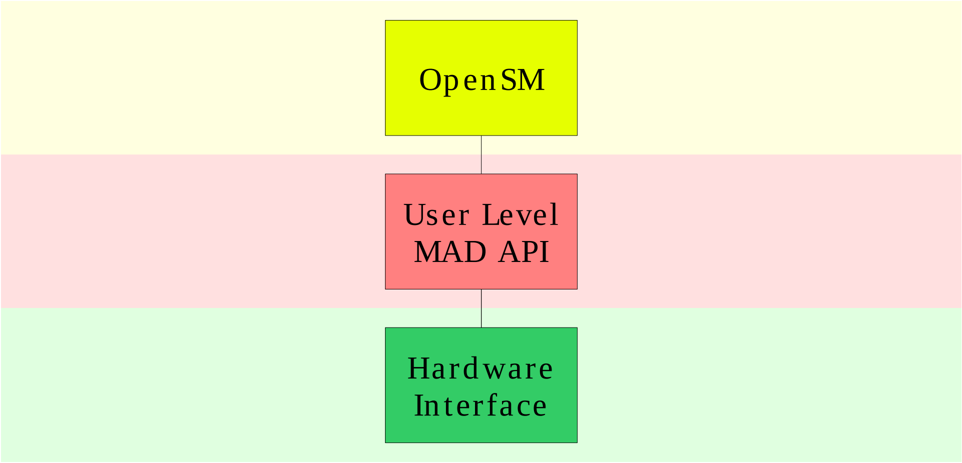
User Verbs

- Used directly by application programs that run in user space.
- Similar to the kernel verbs, they fall into three classes: the generic verbs, those specific to InfiniBand and those specific to iWARP.
- Also used by interfaces such as uDAPL, Open MPI and MVAPICH.
- They provide an intermediate layer that user applications can use.



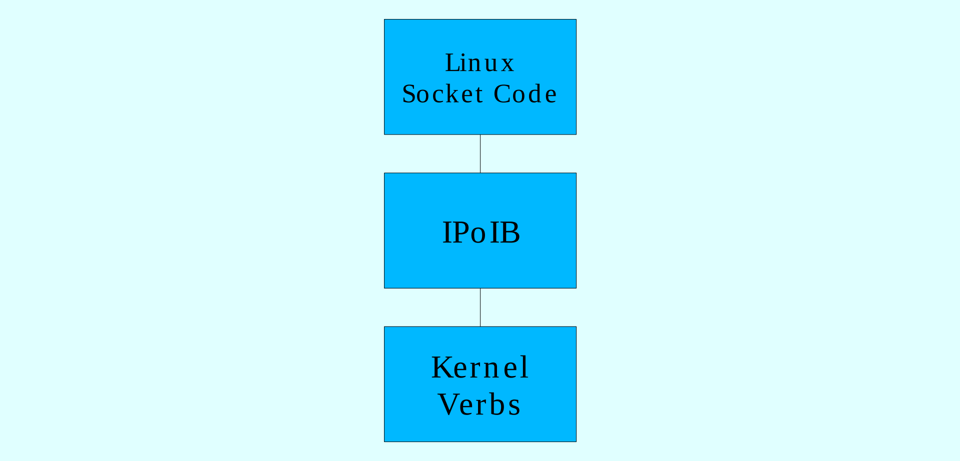
User Level MAD API

- Only provided when running on an InfiniBand fabric.
- Provides an interface for user programs to receive InfiniBand Management Datagrams (MADs).
- Primarily used to support an InfiniBand Subnet Manager (SM) which is used to manage the InfiniBand fabric.



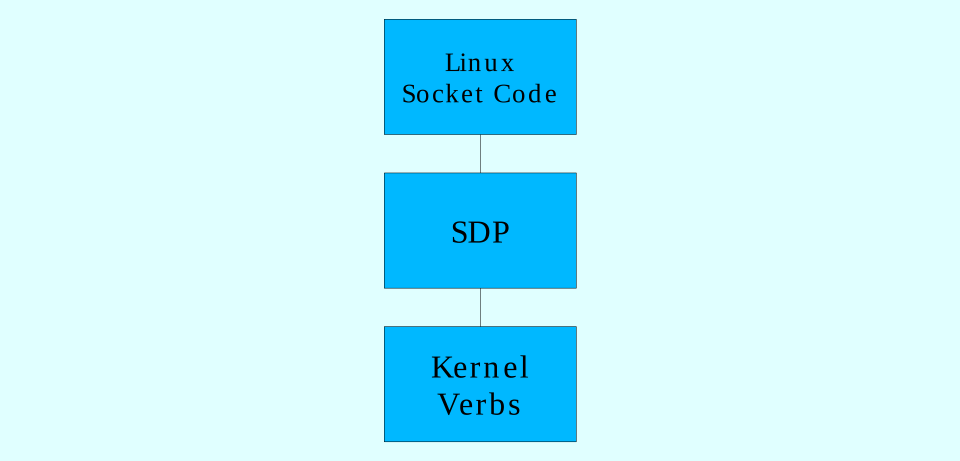
IPoIB

- IP (Internet Protocol) over InfiniBand.
- Provides the TCP and UDP socket interface over InfiniBand.
- Connects into the Linux Kernel socket code.
- Clients are any applications that use sockets such as ssh and ftp.
- Not necessary when using an iWARP NIC since the underlying protocol of iWARP is TCP/IP.



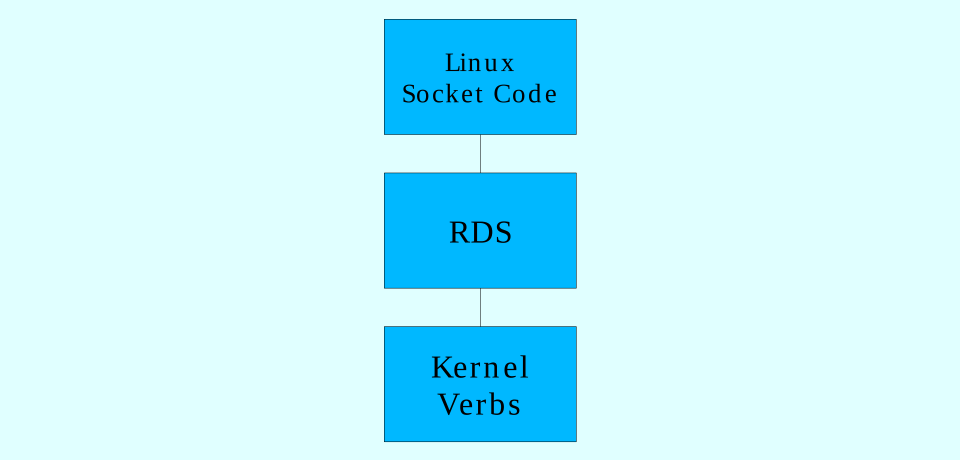
SDP

- Sockets Direct Protocol.
- Defined by IBTA based on initial submission from Microsoft.
- Derived from Sockets Direct.
- Motivated by an attempt to provide a compatible sockets interface that could take advantage of RDMA features that devices provide.
- Utilizes the Berkeley Socket interface.
- Minimal changes required for an application to migrate.



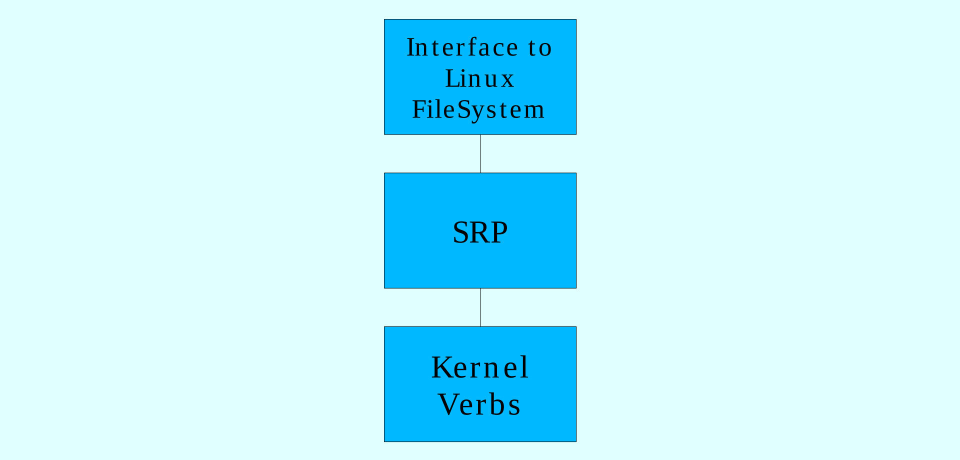
RDS

- Reliable Datagram Sockets.
- Allows messages to be sent reliably to multiple destinations from a single socket.
- User applications can interface with RDS through the Berkeley Socket interface by specifying a different protocol family.
- Useful for connection-less reliable messaging.
- Motivated and defined by Oracle.



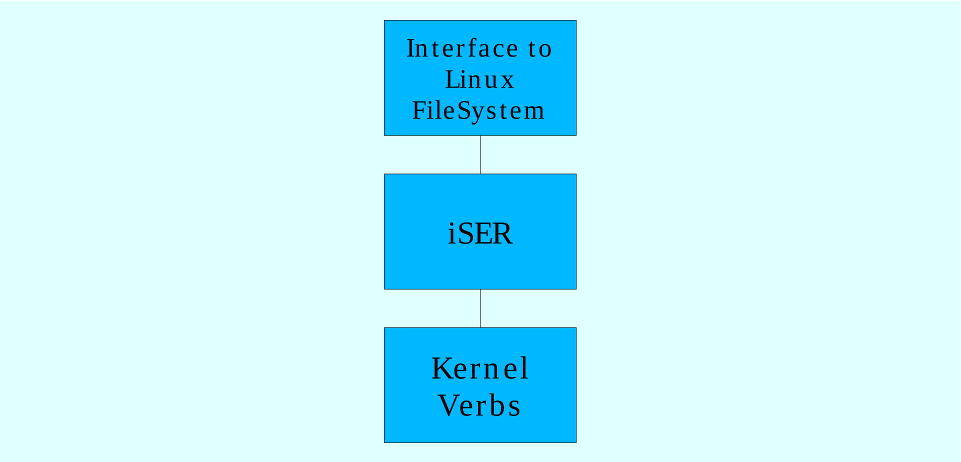
SRP

- SCSI RDMA Protocol.
- Originally intended to allow the SCSI protocol to run over InfiniBand for SAN usage.
- Interfaces directly to the Linux filesystem through the SRP ULP. Users can treat SRP storage as just another device.
- Native SRP devices commercially available that connect over an InfiniBand fabric.
- Can run over iWARP.



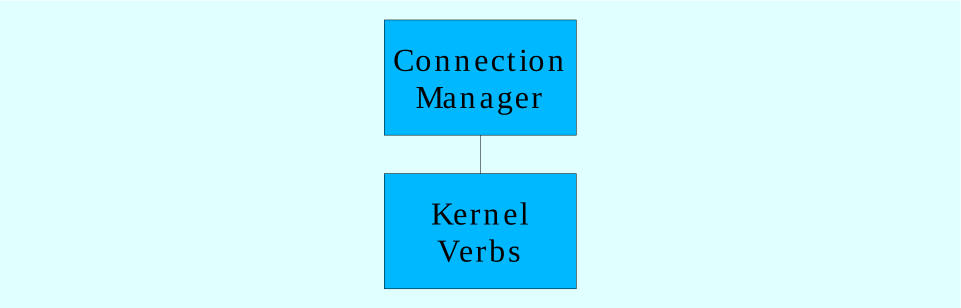
iSER

- iSCSI (Internet SCSI) extensions for RDMA.
- IETF standard.
- Enables iSCSI to take advantage of RDMA. Also simplifies certain iSCSI protocol details such as data integrity management and error recovery.
- Interfaces directly to the Linux filesystem.



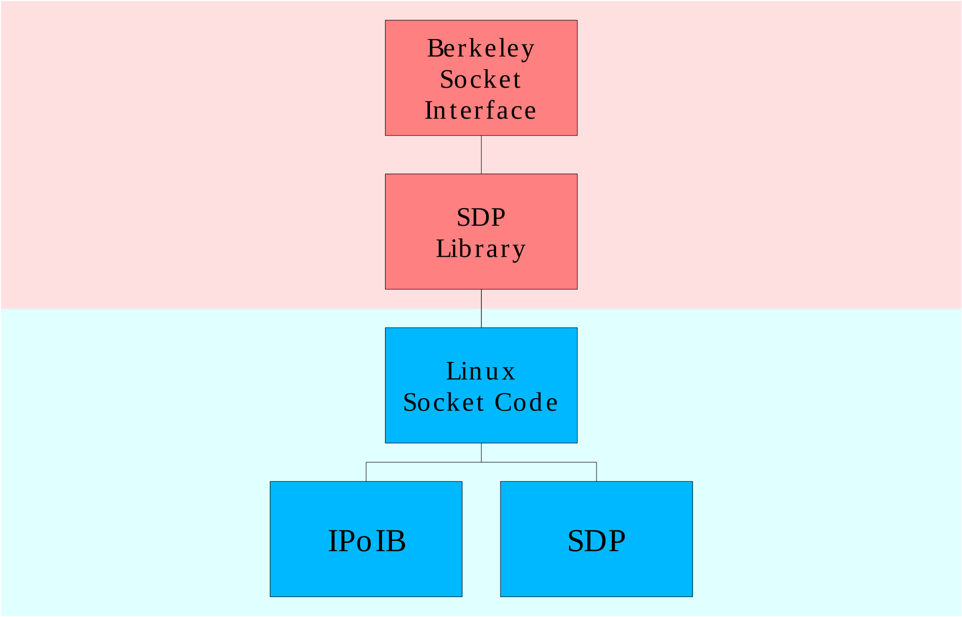
Connection/Communication Manager

- A stand-alone module in the kernel which assists with setting up connections.
- On InfiniBand, it is the Communications Manager, on iWARP, it is the Connection Manager. Common API but separate code bases.
- Referred to as CM by both InfiniBand and iWARP.



SDP Library

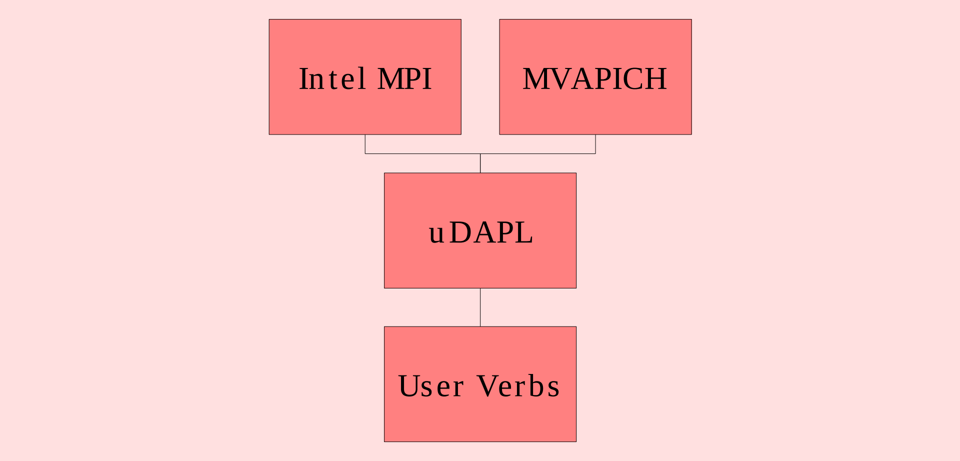
- A shared library that intercepts calls by user applications using the Berkeley Socket Interface intending to use TCP/IP and routing them through SDP.
- Not guaranteed to work on all applications but should work on many.
- Avoids recompilation.
- Allows one to specify sophisticated rules as to which calls to route to SDP.



uDAPL

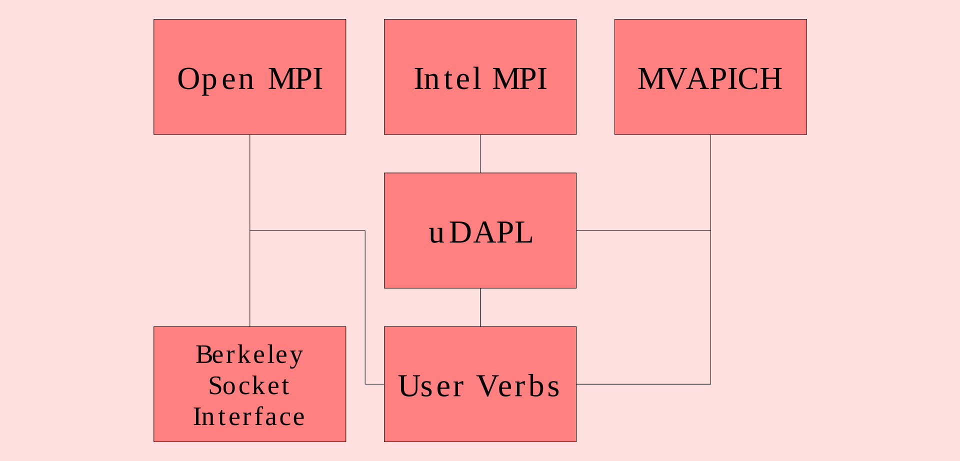
- User Direct Access Provider Library.
- A thin application layer that also interfaces directly to several operating systems such as Linux and Windows as well as the OpenFabrics stack.
- Defined by the DAT (Direct Access Transport) Collaborative.

- Clients:
  - Intel MPI.
  - MVAPICH also provides a uDAPL interface.



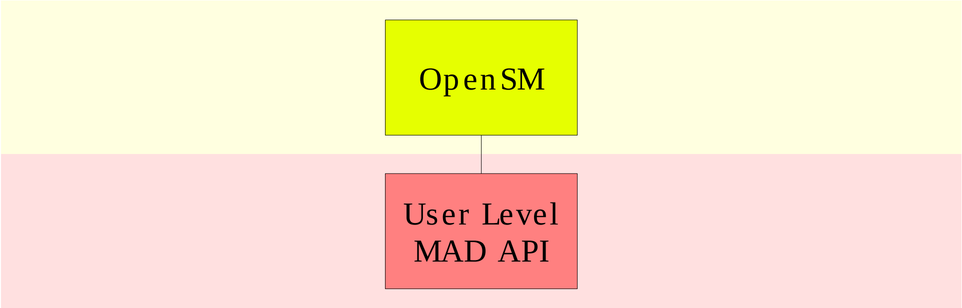
MPI

- Message Passing Interface.
- The primary API used by large cluster applications.
- Provides primitives to allow nodes to farm out computations to other nodes and then synchronize.
- Many variations exists: MPICH, Open MPI, etc.
- MVAPICH and Intel MPI currently run on the OpenFabrics stack.



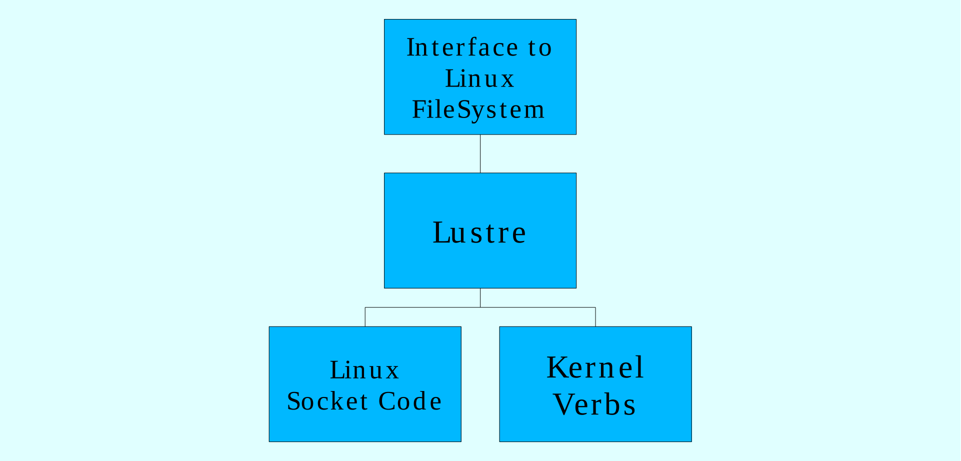
OpenSM

- On an InfiniBand Subnet, there must be one and only one Subnet Manager (SM) running. It may be running on one of the nodes or it may be running inside a switch.
- OpenSM is an Open Source Subnet Manager.
- Interfaces using the User Level MAD API.
- Not needed when running on an iWARP fabric.



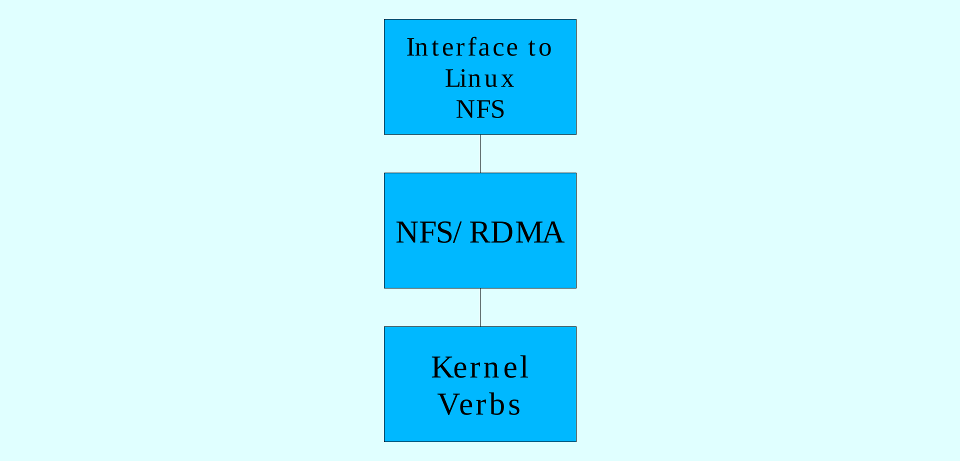
What About Lustre?

- High-performance scalable filesystem developed by Cluster File Systems, Inc.
- Open Source version available for Linux.
- Interfaces to the Linux Socket Code using either TCP/IP or the Kernel Verbs interface.
- Interacts directly with the Linux filesystem.
- Runs entirely in the kernel and considered an OpenFabrics Upper Level Protocol.

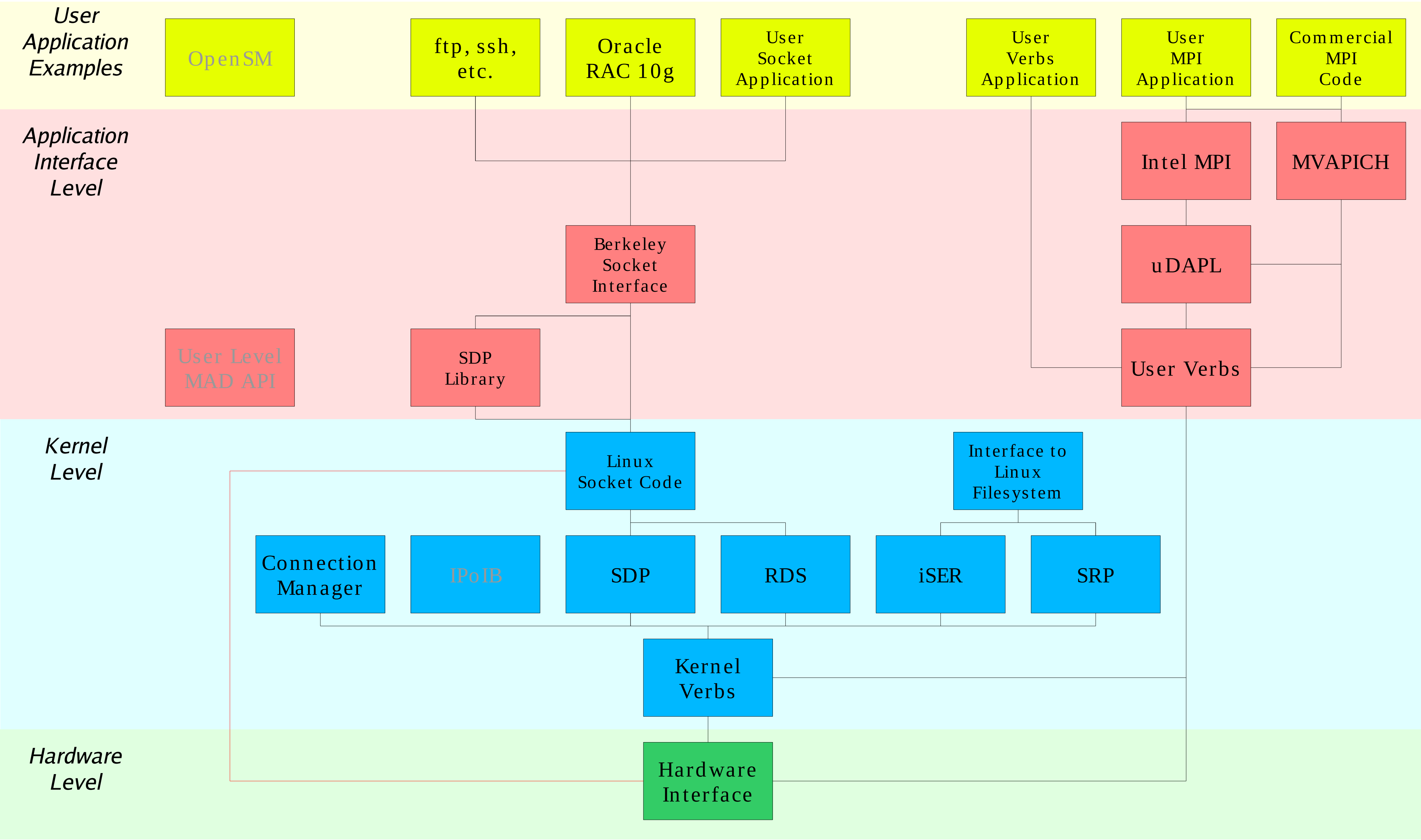


Where Does NFS/RDMA Fit In?

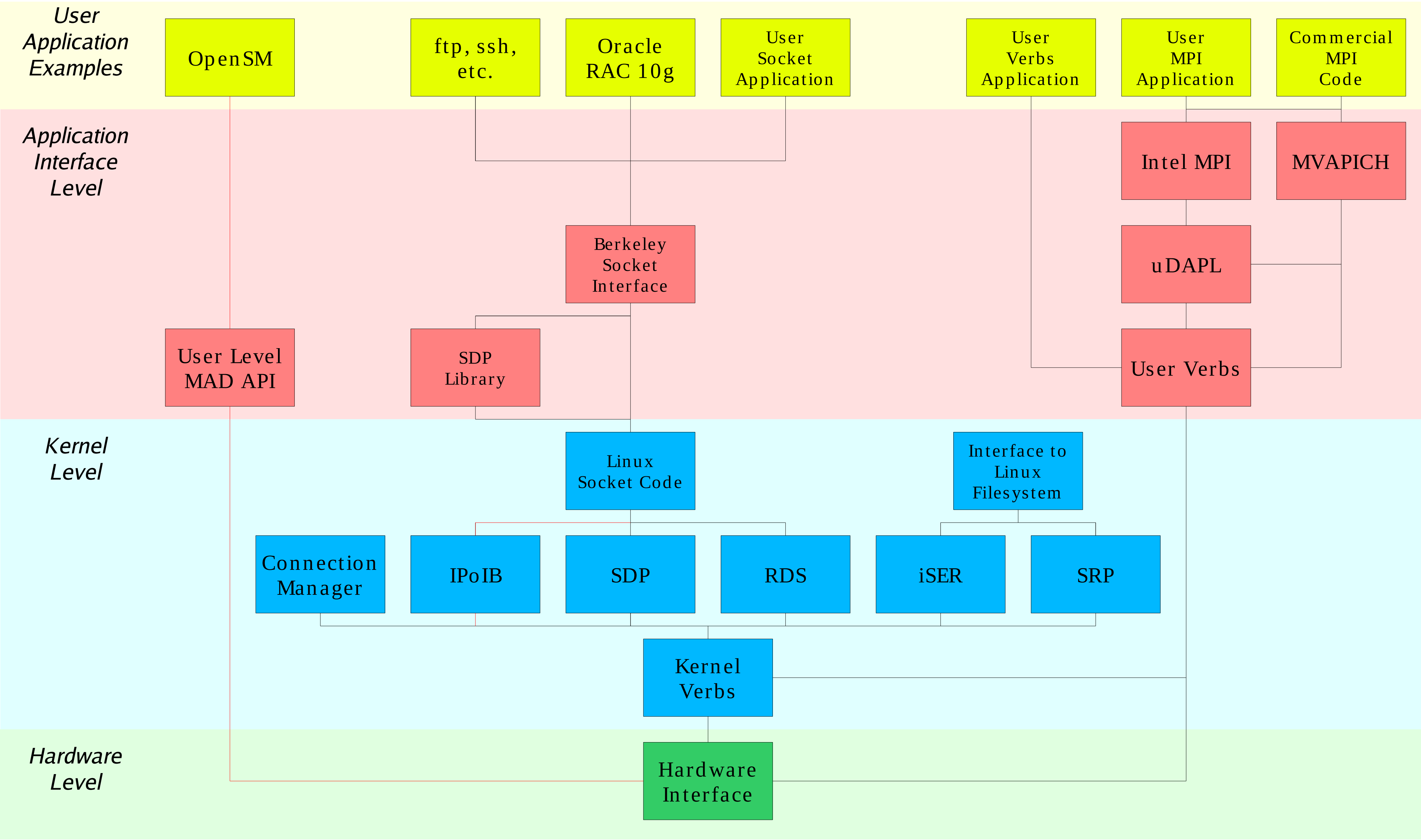
- NFS/RDMA is a RPC-layer protocol that allows NFS to use InfiniBand and iWARP.
- Transparent to users and applications.
- Significant performance boost to clients. Allows NFS to automatically gain the benefits of the OpenFabrics stack.
- Currently available as a set of patches to the Linux kernel. Available from sourceforge.net.



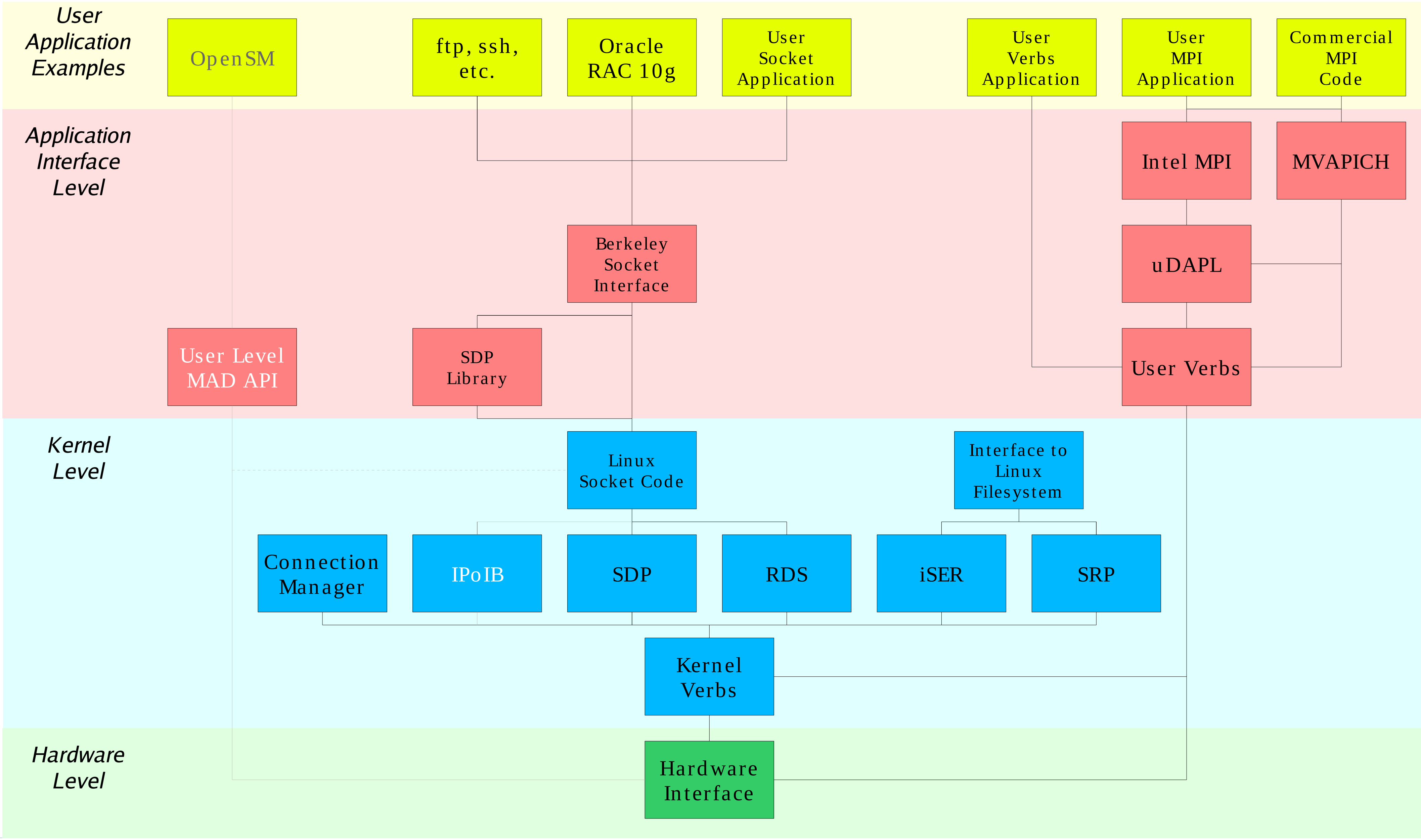
Linux iWARP OpenFabrics Stack



Linux InfiniBand OpenFabrics Stack



Linux OpenFabrics Stack



Open Fabrics Enterprise Distribution

- A snapshot of the OpenFabrics stack that is tested by the community at large.
- Release process similar to that of the Linux kernel. Release candidates are made available until one is approved.
- Release 1.0 made available on June 16, 2006.
- Being picked up by the distributions: RedHat and SuSE.
- New releases planned every few months.

Thank You