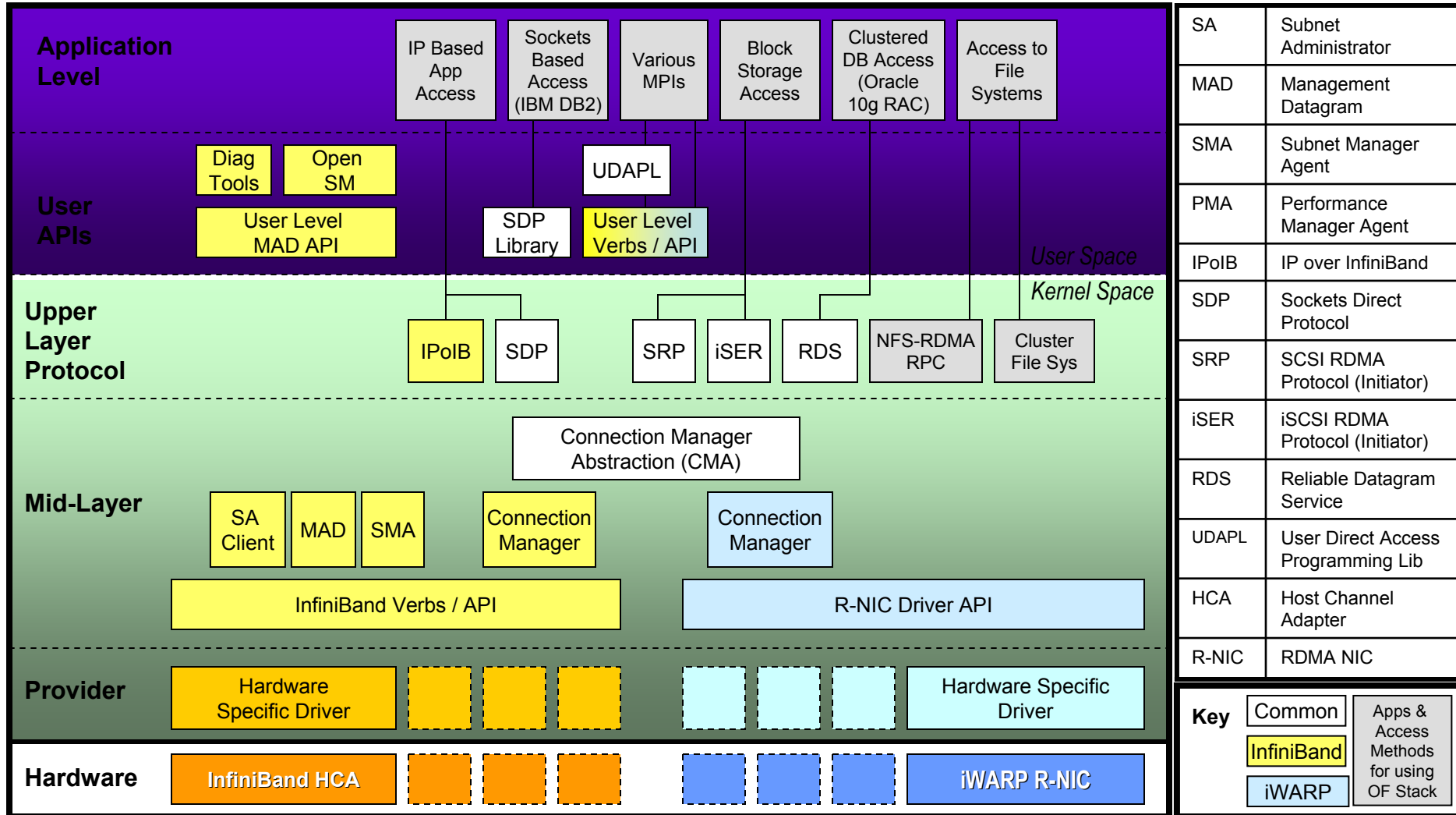


Infiniband Software Architecture

- Volume 1 and 2 -> specify
- really large!
- `/net/hp92/mwolf/IB/OFED-1.0-SRC/openib-1.0/src`
- `/urs/ofed` on polynesia1-6

OpenFabrics Software Stack



SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Basic communication elements

- Work Request
 - Associated with a queue, completion notification, scatter gather memory region, various parameters
- Send Queue
 - Send Buffer, RDMA Read, RDMA Write, Atomic, Mm bind
- Receive Queue; Shared Receive Queue
 - Receive Buffer
- Completion Queue
 - Associate with multiple work queues; notify or poll; completion notification may be turned off; event handlers
- “Connection”
 - Queue pairs, End-to-End contexts

Communication Path

- Hca registers mapped to host (kernel/user space)
- “Slow path” – through kernel and write calls to device
- “Fast path” – bypass kernel and access device registers mmap-ed to userspace
- Kernel keeps track of resources available to a context
- For “fast path” IO memory must be pinned/unpinned (get_user_pages, put_page)
 - (ib_(de)reg_mr)

Verbs

- Interface to HCA
 - configuring and managing the channel adapter, allocating (creating and destroying) queue pairs, configuring QP operation, posting work requests to the QP, getting completion status from the completion queue.
- Directly exported to upper layer applications
- Subset of hardware specific verbs used by kernel services
- Libibverbs – device independent + device dependent kernel and userspace drivers
 - Mellanox - mthca
- `/net/hp92/mwolf/IB/OFED-1.0-SRC/openib-1.0/src/linux-kernel/infiniband/include/rdma`

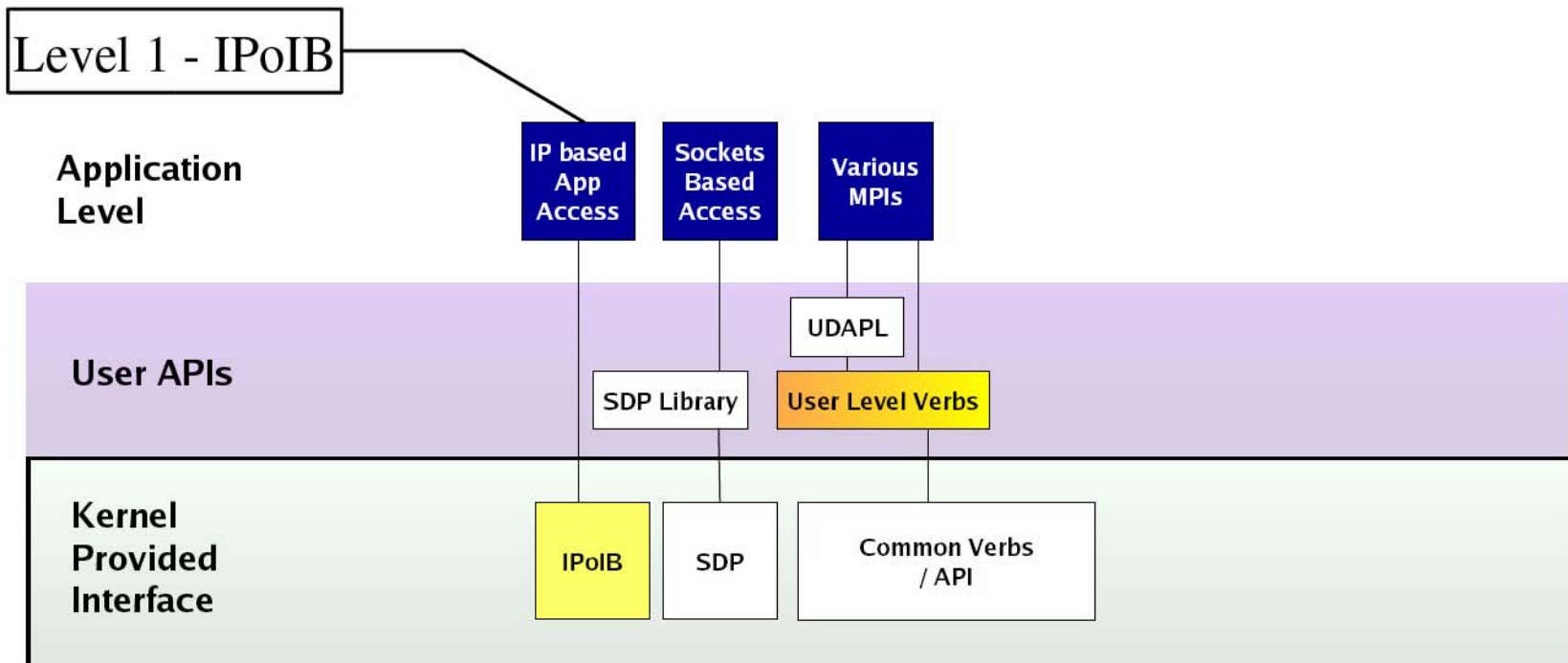
Verb	Mandatory/Optional Classification	Consumer Accessibility
Open HCA	Mandatory	Privileged
Query HCA	Mandatory	Privileged
Modify HCA Attributes	Mandatory	Privileged
Close HCA	Mandatory	Privileged
Allocate Protection Domain	Mandatory	Privileged
Deallocate Protection Domain	Mandatory	Privileged
Allocate Reliable Datagram Domain	RD Service	Privileged
Deallocate Reliable Datagram Domain	RD Service	Privileged
Create Address Handle	Mandatory	User-Level and Privileged
Modify Address Handle	Mandatory	User-Level and Privileged
Query Address Handle	Mandatory	User-Level and Privileged
Destroy Address Handle	Mandatory	User-Level and Privileged
Create Shared Receive Queue	SRQ	Privileged
Modify Shared Receive Queue	SRQ	Privileged
Query Shared Receive Queue	SRQ	Privileged
Destroy Shared Receive Queue	SRQ	Privileged
Create Queue Pair	Mandatory	Privileged
Modify Queue Pair	Mandatory	Privileged
Poll for Completion	Mandatory	User-Level and Privileged
Request Completion Notification	Mandatory	User-Level and Privileged
Set Completion Event Handler	Mandatory	Privileged
Set Asynchronous Event Handler	Mandatory	Privileged

Query Queue Pair	Mandatory	Privileged
Destroy Queue Pair	Mandatory	Privileged
Get Special QP	Mandatory	Privileged
Create Completion Queue	Mandatory	Privileged
Query Completion Queue	Mandatory	Privileged
Resize Completion Queue	Mandatory	Privileged
Destroy Completion Queue	Mandatory	Privileged
Create EE Context	RD Service	Privileged
Modify EE Context Attributes	RD Service	Privileged
Query EE Context	RD Service	Privileged
Destroy EE Context	RD Service	Privileged
Allocate L_Key	Base MM Extensions	Privileged
Register Memory Region	Mandatory	Privileged
Register Physical Memory Region	Mandatory	Privileged
Query Memory Region	Mandatory	Privileged
Deregister Memory Region	Mandatory	Privileged
Reregister Memory Region	Mandatory	Privileged
Reregister Physical Memory Region	Mandatory	Privileged
Register Shared Memory Region	Mandatory	Privileged
Allocate Memory Window	Mandatory	Privileged
Query Memory Window	Mandatory	Privileged
Bind Memory Window	Mandatory	User-Level and Privileged
Deallocate Memory Window	Mandatory	Privileged
Attach QP to Multicast Group	UD Multicast Service	Privileged
Detach QP from Multicast Group	UD Multicast Service	Privileged
Post Send Request	Mandatory	User-Level and Privileged
Post Receive Request	Mandatory	User-Level and Privileged

Alternatives to VAPI

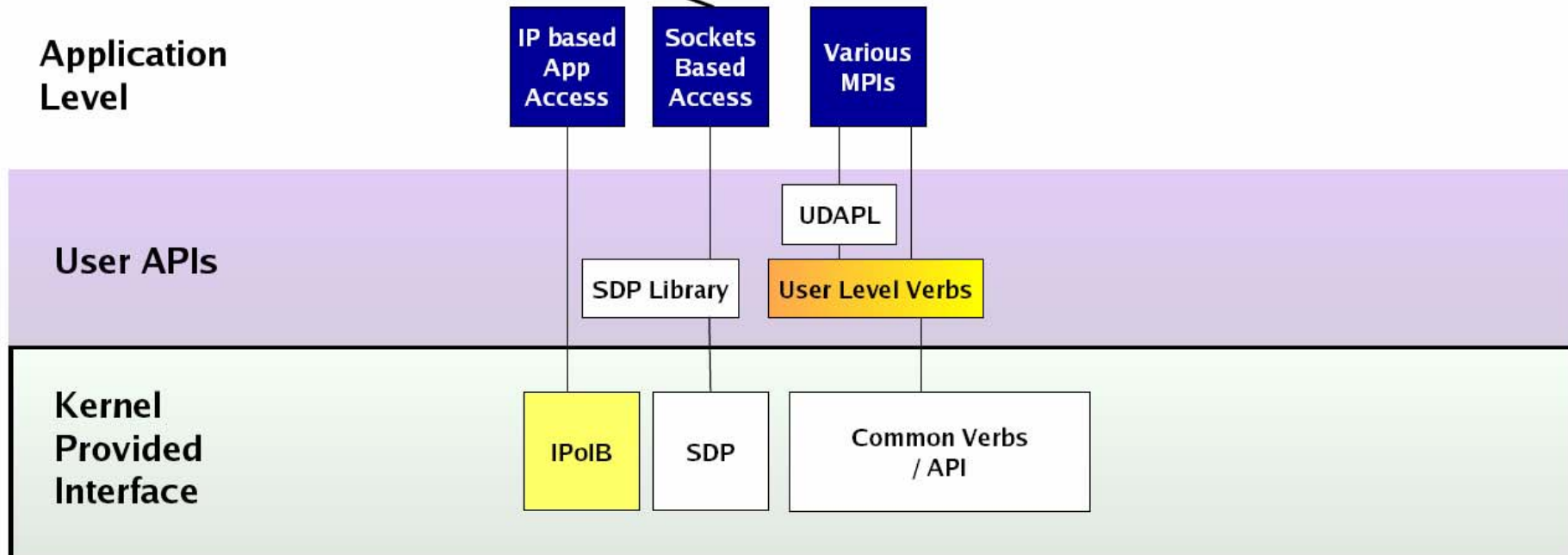
- A range of higher level APIs
- RDMA API
 - slightly above VAPI, semantically closer to familiar APIs such as sockets
 - memory must be registered/locked?
 -/src/userspace/librdmacm/
- uDAPL, kDAPL – Direct Access Programming Library
 - Device independent, transport independent access API which exploits RDMA capabilities of interconnect technologies
- Various MPI libraries

Developing applications

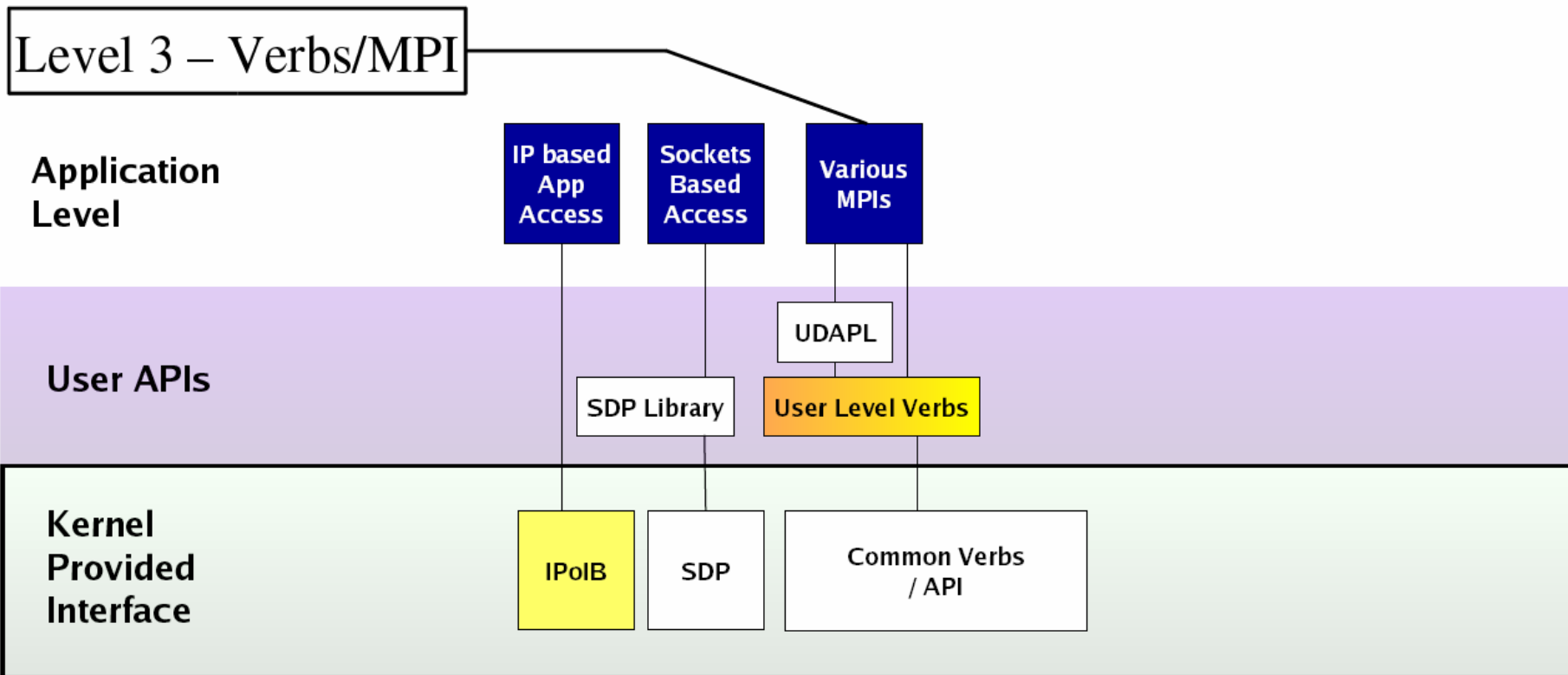


- Easiest to use, requires no modification of applications
- Lowest overall payback

Level 2 – SDP

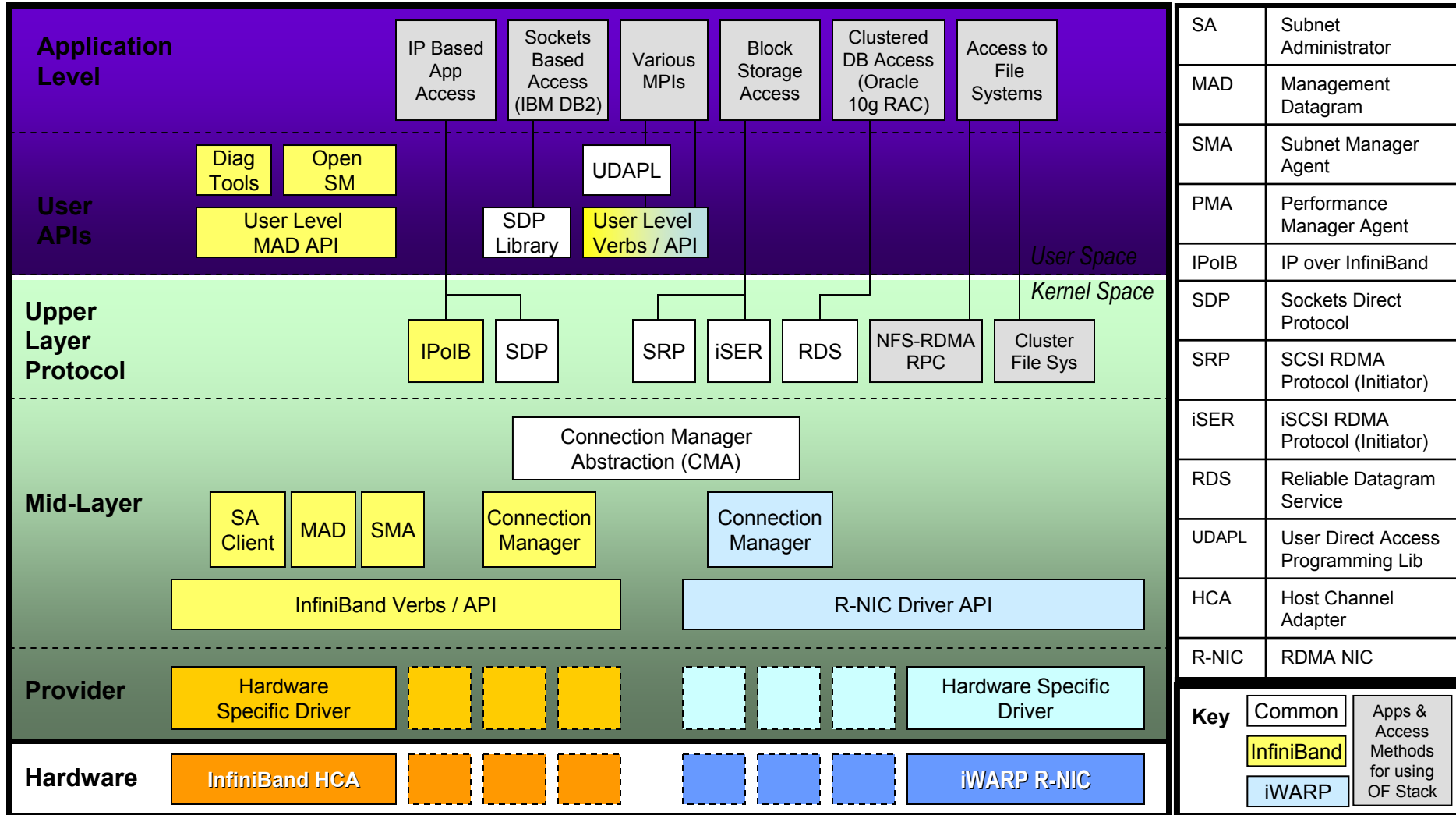


- You *might* be able to use libsdp library to enable SDP in your application without any code changes or recompiles
- If not, the code changes to natively support SDP are very minimal
- This methods gets a good deal of the RDMA benefit



- Code must be written to either the verbs or MPI API
- Code changes are not minimal, and in some cases require rethinking of application design
- This methods gets full benefit of RDMA capabilities

OpenFabrics Software Stack



Connection management

- Transport specific Connection manager
- Connection Management Abstraction – unifying layer
- Create, configure, manage connection, associate events/event handlers

Management tools

- .../userspace/managemet
 - MAD libraries
 - userspace and kernel
 - Diagnostics
 - route, trace, address, stats...
 - Scripts included
 - HCA performance counters mapped in host memory
 - Open Subnet Manager