

RED HAT :: NASHVILLE :: 2006

SUMMIT



Infiniband and RDMA Technology

Doug Ledford

Top 500 Supercomputers

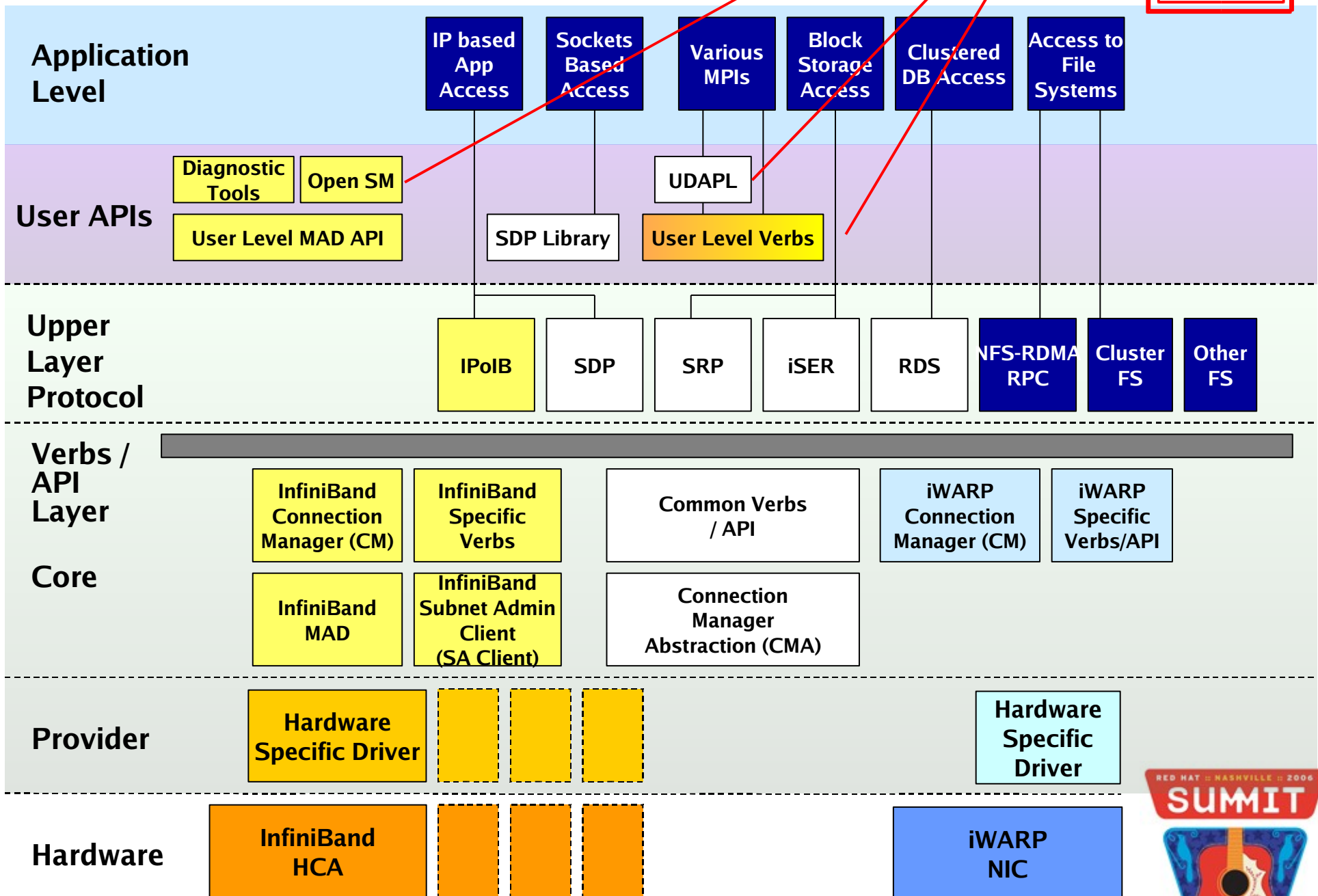
Nov 2005

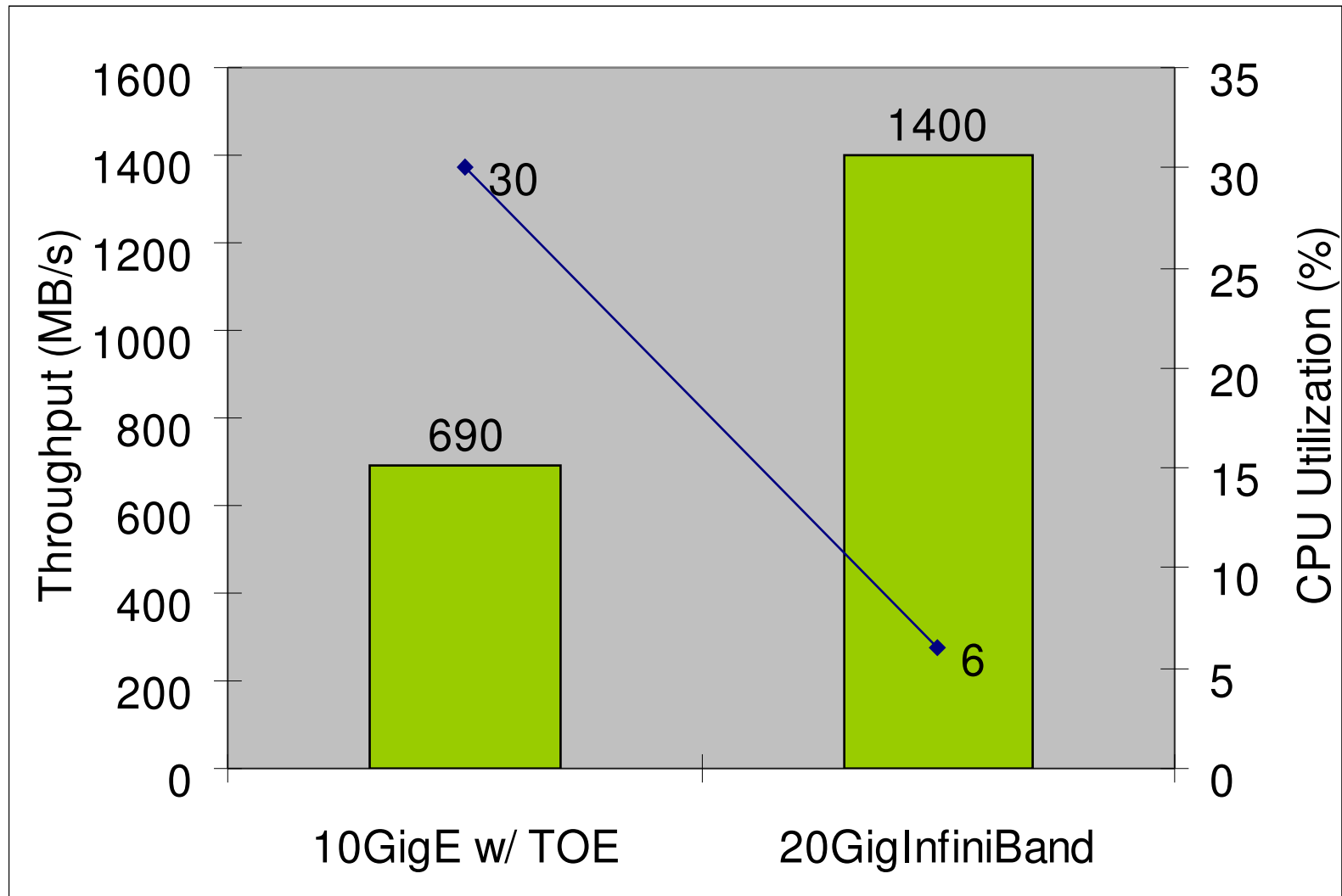
- #5 Sandia National Labs, 4500 machines, 9000 CPUs, 38TFlops, 1 big headache
- Performance great....but....
- Adding new machines problematic due to software interactions
- Diagnosing and locating faults very difficult



OpenFabrics Software Stack

Headache

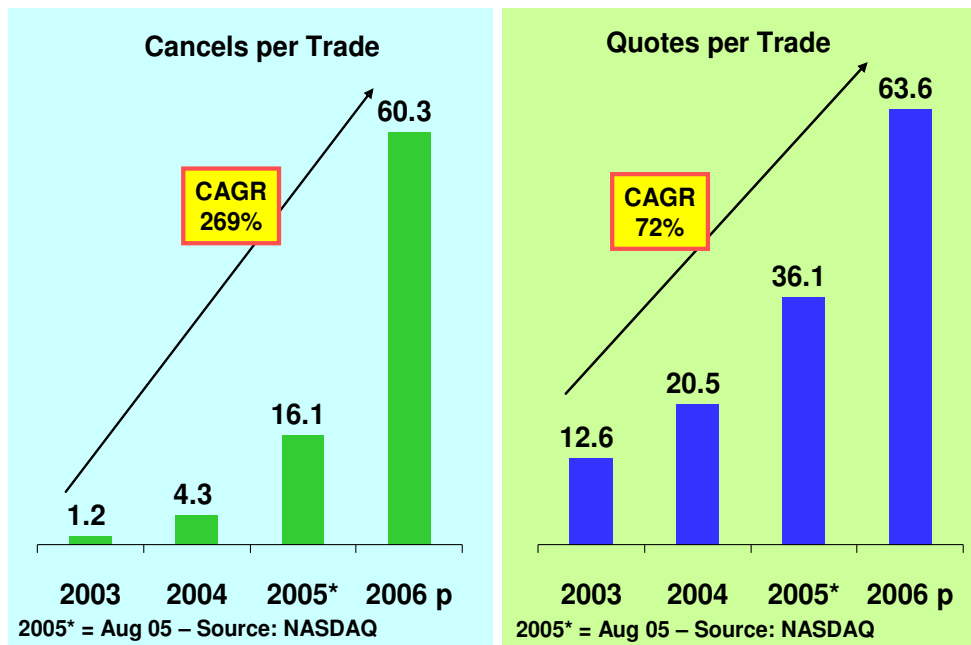




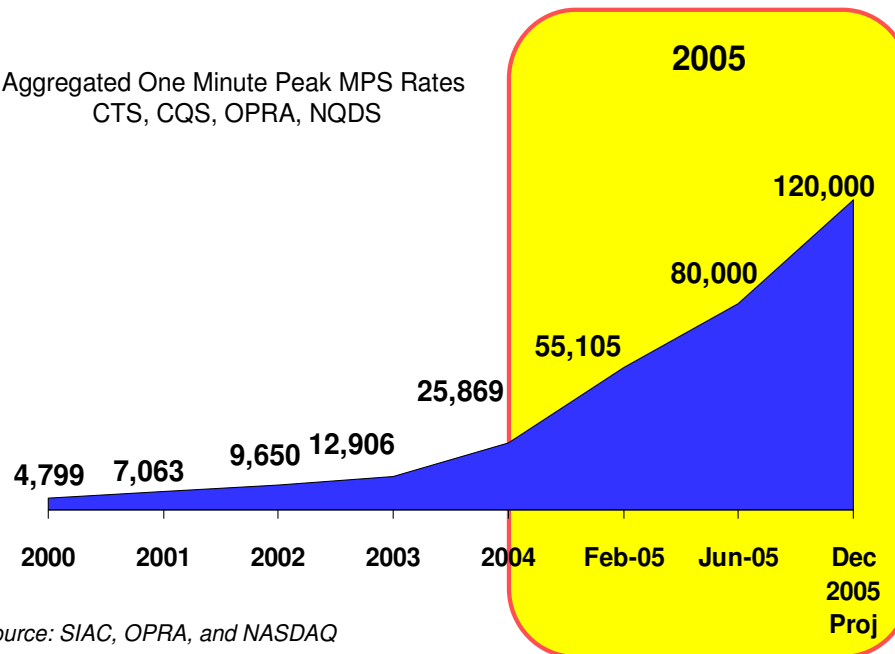
Source: “Head to TOE” from OSU, “InfiniBand and 10-Gigabit Ethernet for I/O in cluster computing” from Sandia National Laboratories, and Mellanox



Wall Street Trading Environment Challenges

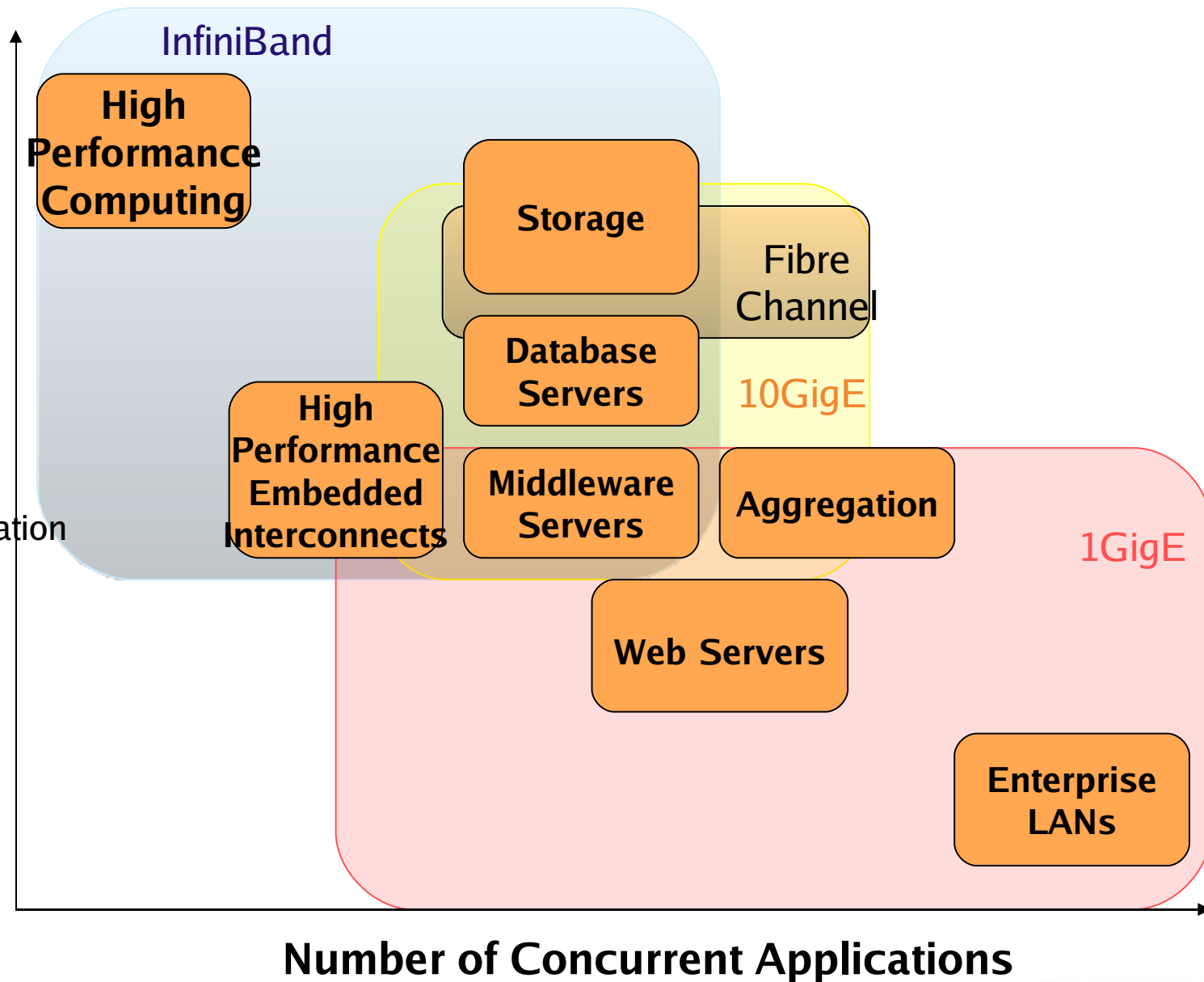


Aggregated One Minute Peak MPS Rates
CTS, CQS, OPRA, NQDS

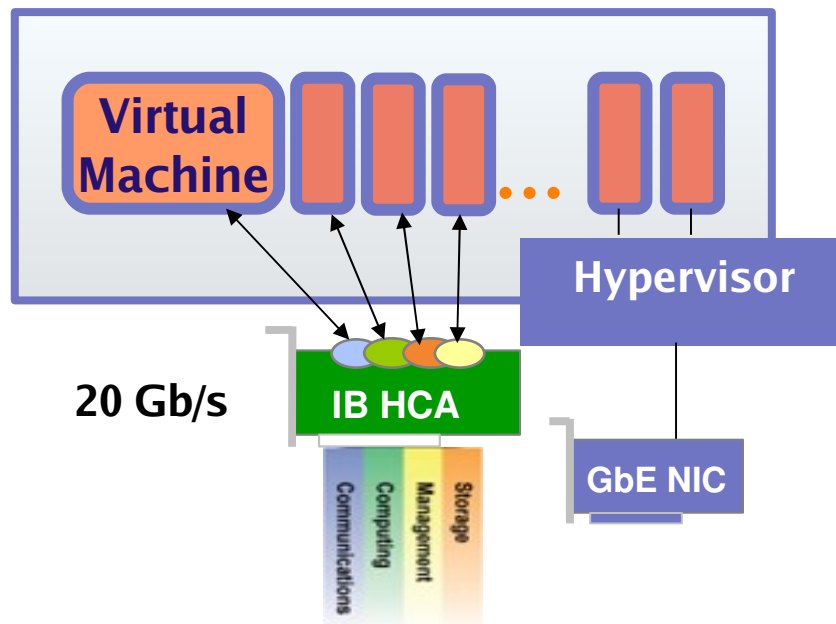


Performance

- Low Latency
- High Bandwidth
- Efficient CPU Utilization
- Reliable Transport

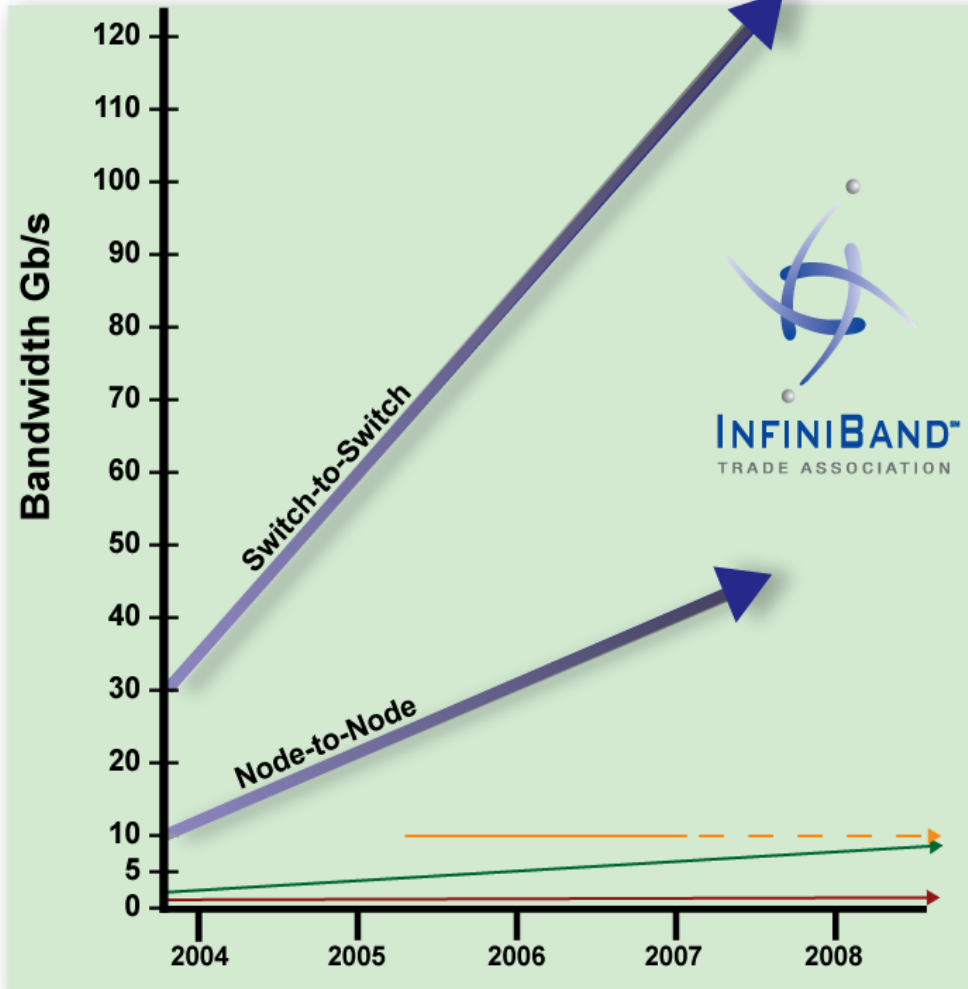


- **Price/Performance**
 - **\$69 (OEM) adapter IC vs. \$500 for similar 10GigE adapter IC solution**
 - **\$200 (OEM) adapter card vs. \$2000 for comparable 10GigE card**
 - **1.4GB/s and 2.7 μ s latency**
- **Virtualization**
 - **Highest utilization of computing and storage resources**
 - **Simplifies adding resources for rapidly expanding data centers**



InfiniBand Roadmap

InfiniBand's roadmap outpaces all proprietary and standard based I/O technologies in both pure performance and price/performance



Number of IB Lanes	Per Lane Bandwidth		
	SDR 2.5Gb/s	DDR 5Gb/s	QDR 10Gb/s
4X	10Gb/s	20Gb/s	40Gb/s
8X	20Gb/s	40Gb/s	80Gb/s
12X	30Gb/s	60Gb/s	120Gb/s

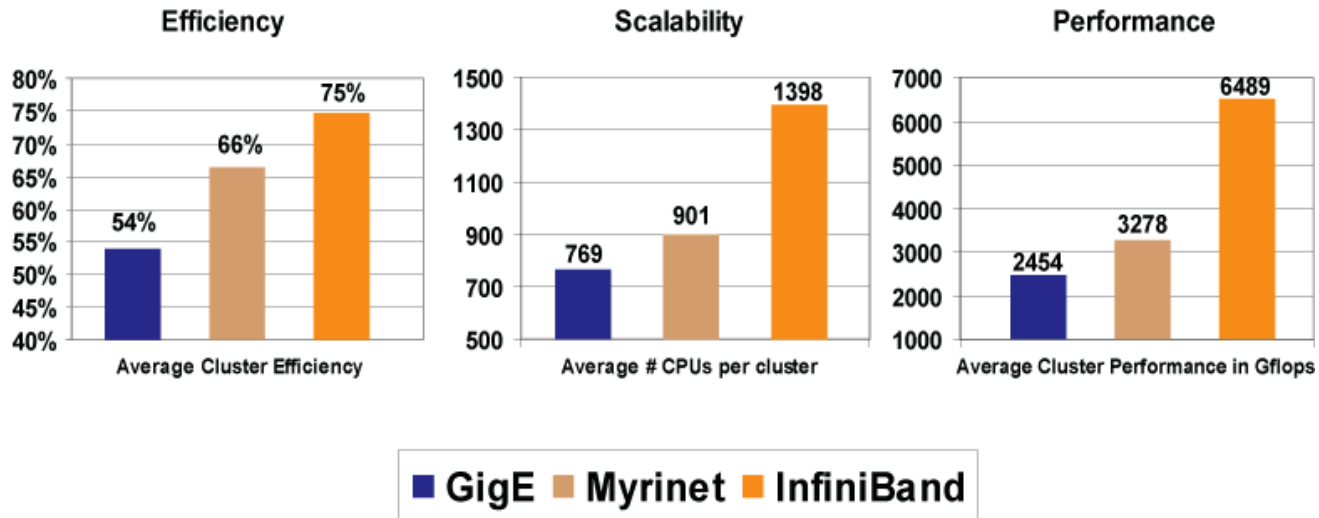
- InfiniBand
- 10GigE, 10G iSCSI, Proprietary
- Fibre Channel*
- GigE, iSCSI

*FCIA estimates 2007/8 for 8Gb/s FC

*Roadmap based on silicon availability

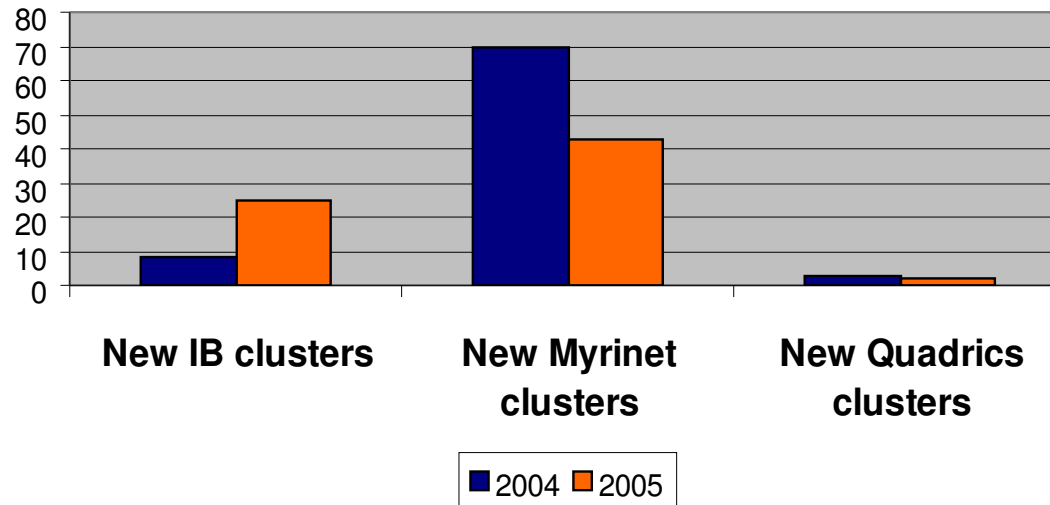


Infiniband/RDMA use climbing rapidly

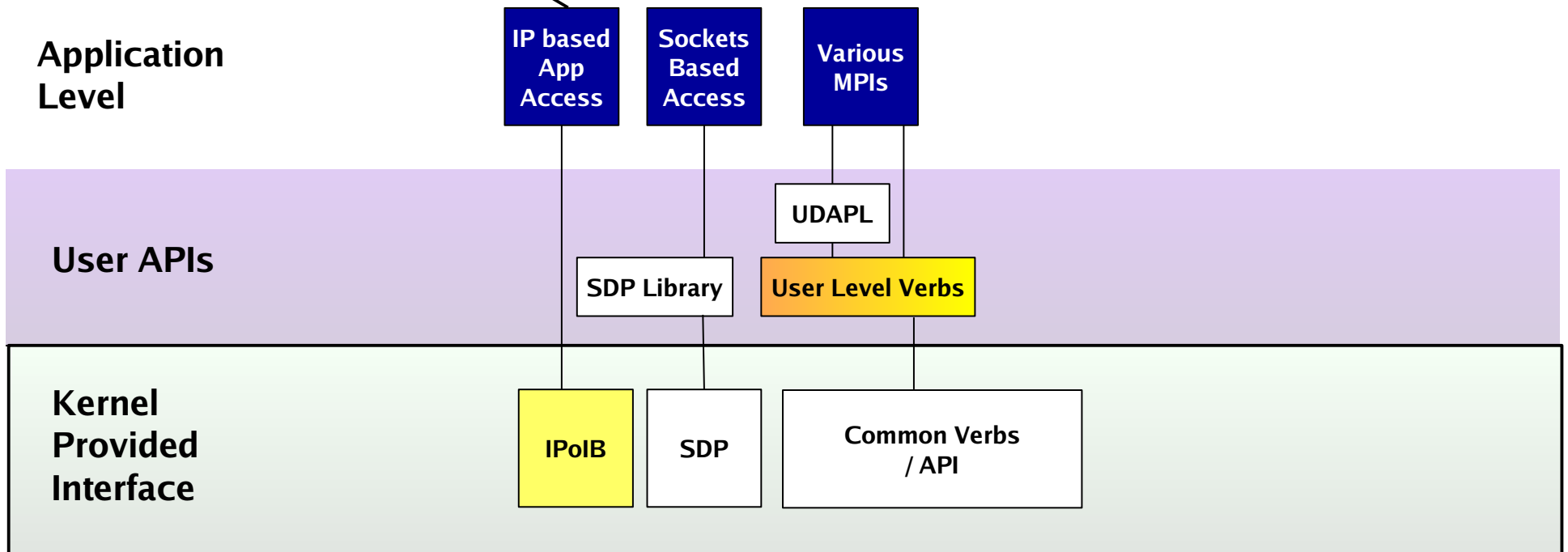


Source: Top500.org

New Top500 Cluster



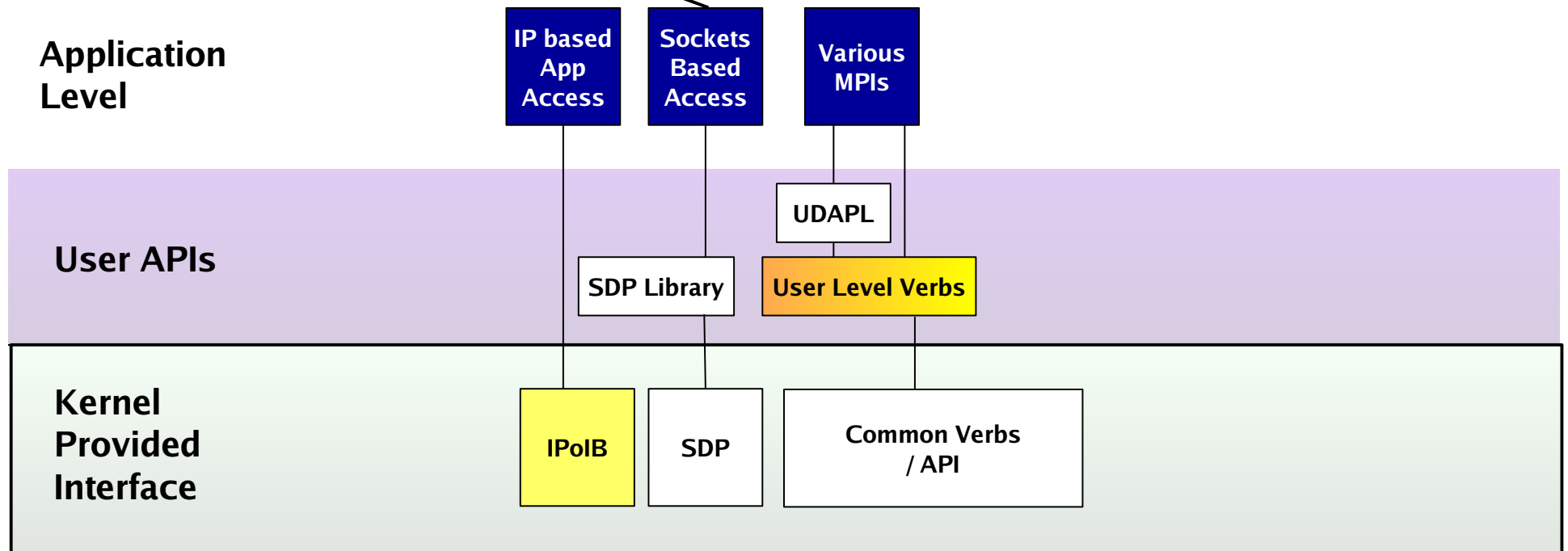
Level 1 - IPoIB



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

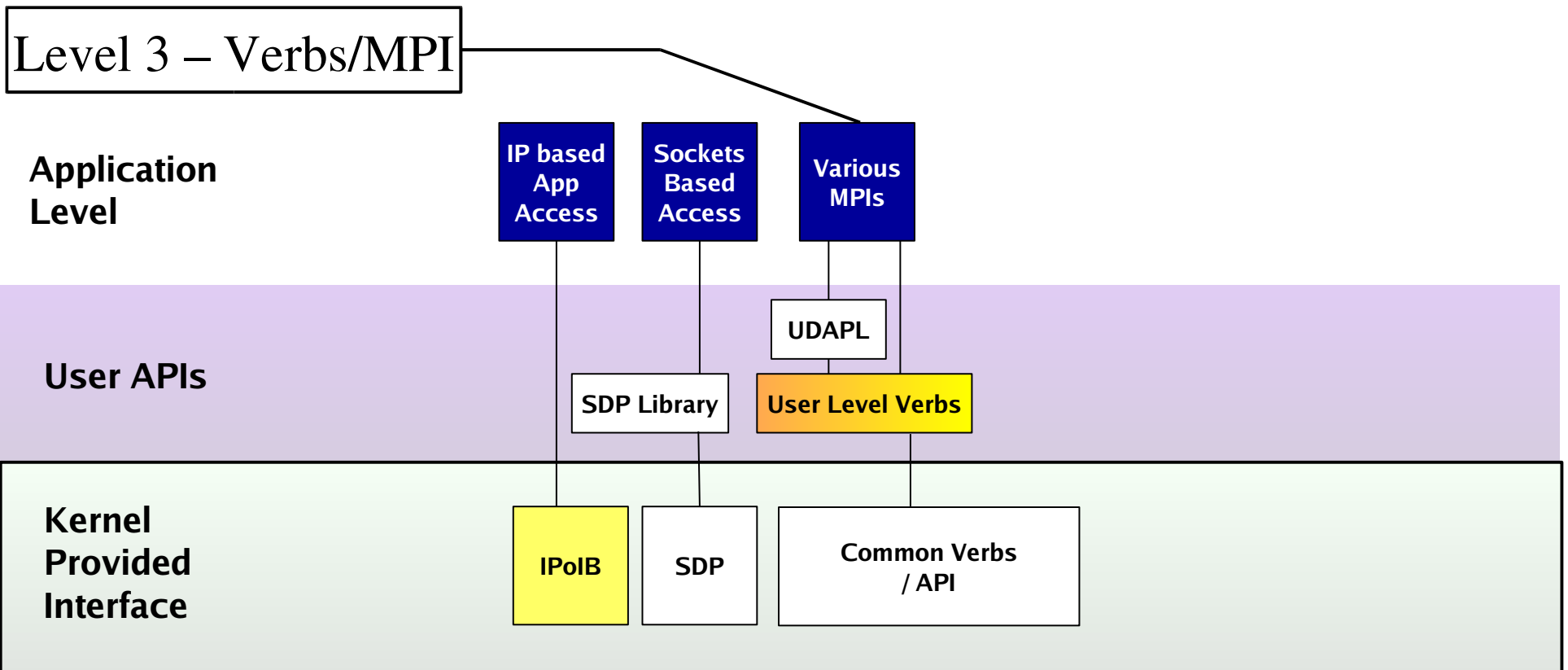


Level 2 – SDP



- 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

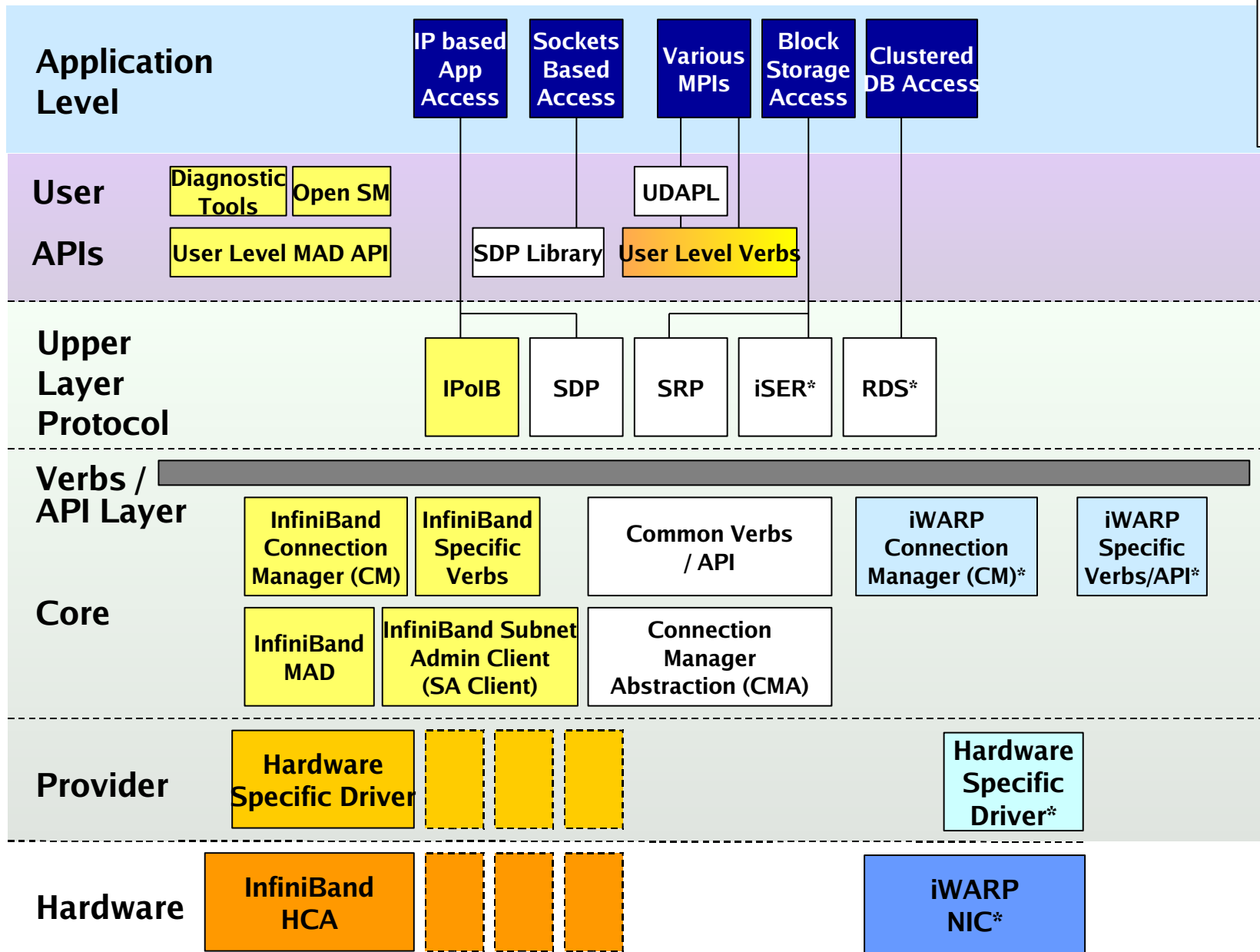




- Code must be written to either the verbs or MPI API
- Code changes are not minimal, and in some cases require rethinking of application design
- This methods gets full benefit of RDMA capabilities



OpenFabrics Software Stack



Key

- Common (White box)
- IB Specific (Yellow box)
- iWARP Specific (Light Blue box)

* - Future

