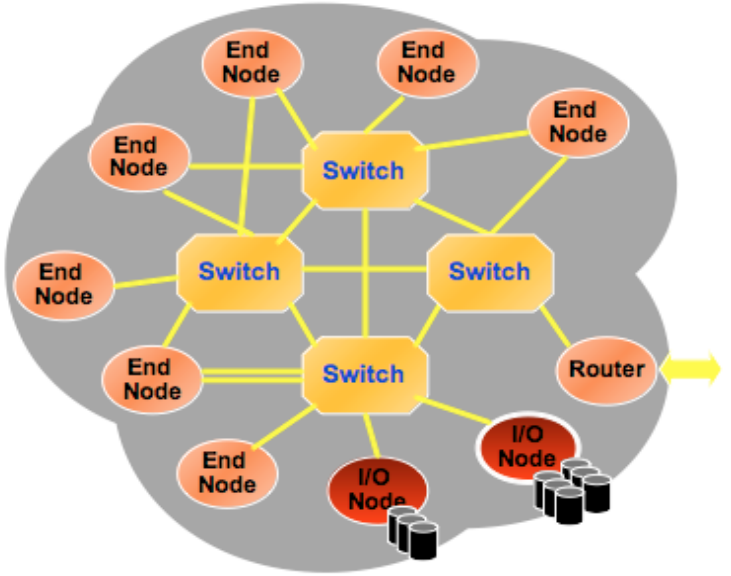
SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2011

The role of a InfiniBand and automated data tiering in achieving extreme storage performance

Cynthia Mcquire
Oracle

Introduction to InfiniBand

- Standard interconnect
  - Defined by the InfiniBand Trade Association
- Defined to facilitate low cost and high performance implementation
- Channel based I/O
- I/O consolidation
  - Communication, computation, management and storage over a single fabric



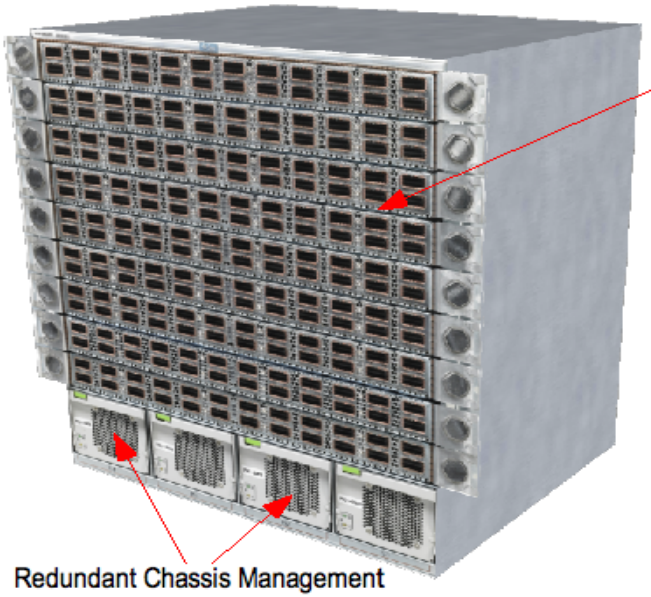
Number of IB Lanes	Per Lane Bandwidth		
	SDR 2.5Gb/s	DDR 5Gb/s	QDR 10Gb/s
4X	10Gb/s	20Gb/s	40Gb/s
8X	20Gb/s	40Gb/s	80Gb/s
12X	30Gb/s	60Gb/s	120Gb/s

InfiniBand Feature Highlights

- Switch fabric links
  - 10Gb/s to 40Gb/s HCA links
  - Up to 120Gb/s switch-switch
- HW transport protocol
  - Reliable and unreliable
  - Connected and datagram
- Kernel bypass
  - Memory translation and protection tables
- Memory exposed to remote node
  - RDMA-read and RDMA-write
- Quality Of service
  - IO channels at the adapter level
  - Virtual lane at the link level
- Scalability and flexibility
  - Up to 48K nodes in subnet, up to 2^{128} in network
- Network partitioning
  - Multiple networks on a single wire
- Reliable, lossless, self-managing fabric
  - End to end flow control
  - Link level flow control
  - Multicast support
  - Congestion control
  - Automatic path migration
- I/O Virtualization with channel architecture
  - Dedicated services to guests OSes
  - HW assisted protection and inter-process isolation
  - Enable I/O consolidation

QDR Product Portfolio

M9: Up to 648 Quad Data Rate ports in less than 12U Up to 54 InfiniScale IV switch chips per chassis

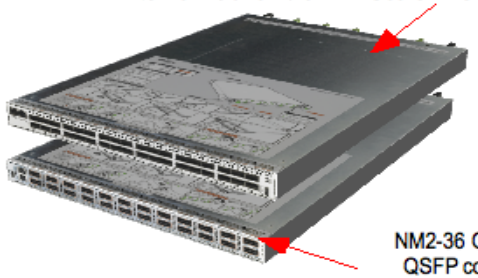


9 Horizontal Line Cards

SDR, DDR, QDR

Redundant Chassis Management

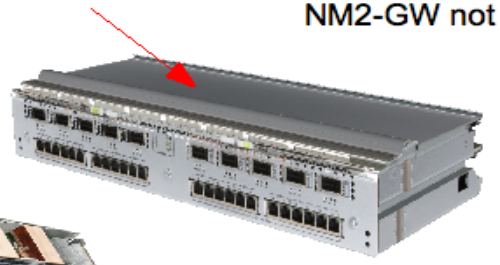
NM2-72: 72 QDR IB ports with 24 iPASS connectors Internal Fabric w/ 6 InfiniScale IV switch chips

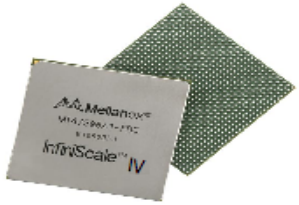


NM2-36 QDR IB ports with 36 QSFP connectors Single InfiniScale IV Switch Chip meshed

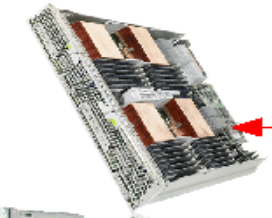
QDR NEM: 10 CXP connectors with 2 fully internal InfiniScale IV switches

NM2-GW not shown



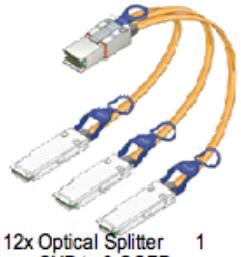


Mellanox 36 Port QDR Switch



X6275 Blade: 2 Node each with QDR HCAs

ConnectX Gen2 8x / 4x QDR



12x Optical Splitter 1 CXP to 3 QSFP

12x CXP to CXP Optical

12x CXP to CXP Copper

PCIe Gen 2.0 QDR HCA ExpressModule

PCIe Gen 2.0 QDR HCA LP PCI card

PCIe Gen 2.0 DDR HCA FEM card

Applications

- Exadata
- Exalogic
- Database – Oracle RAC
- Low latency messaging
- ZFS Storage Appliances

Exadata
Extreme Database Performance

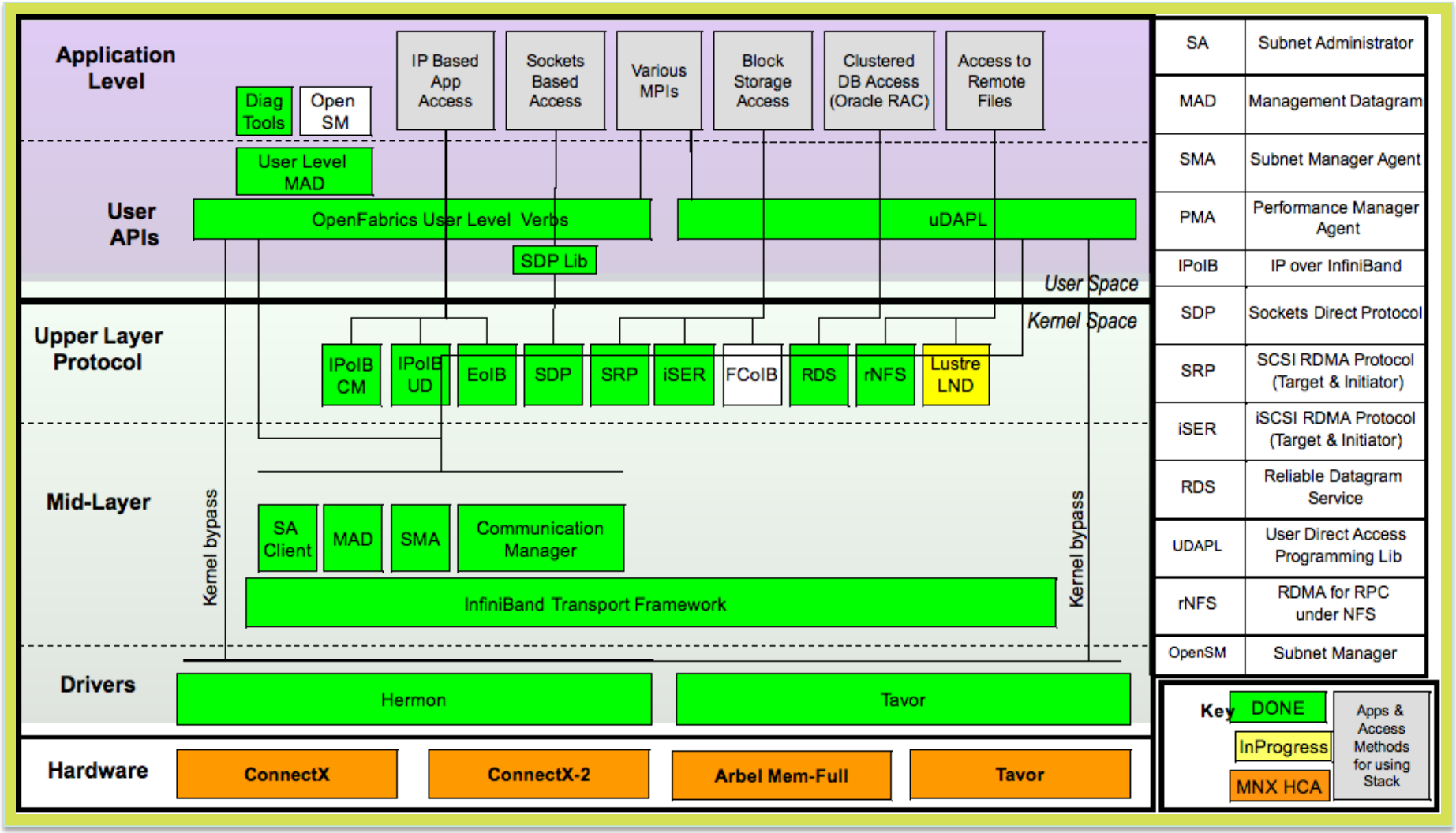
InfiniBand Network

- Redundant 40Gb/s switches
- Unified server & storage network

ZFS Storage Appliances
Extreme Database Performance

Solaris InfiniBand Overview

Solaris InfiniBand Stack



InfiniBand Stack: HCA Drivers

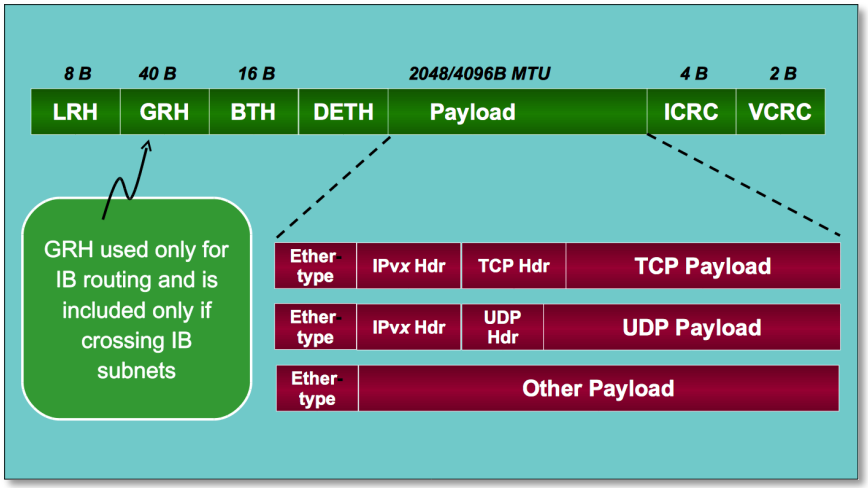
- Primary vendor of HCA silicon is Mellanox
- Two drivers supported:
  - tavor(7D) – for InfiniHost (PCI-X), InfiniHost III Ex (PCIe), up to DDR rates
  - hermon(7D) – for ConnectX[-2] (PCIe gen 1 & 2), up to QDR rates
- Hermon (7D) offers additional functional and performance features
  - 256 MSIx vectors – used to balance CPU loading
  - Offloads – LSO (network), FRWR (memory registration)
  - Capabilities – Relaxed ordering support, HCA hot-plug
- Firmware update via fwflash(1m)
- Integration with Solaris FMA
- Show picture of HCA from Mellanox PRM, qps. cqs

InfiniBand Stack: Transport (IBTF)

- Enables multiple services over IB transport
  - Compliant with 1.2 specification
  - Utilizes network services to map IP address to IB address
- Exports Transport Interface (TI) upward to service drivers
  - The "Verbs" interface
- Requests services from HCA Driver
  - Channel Interface (CI) downward to HCAs
- Provides Management Datagram (MAD) interface to components
  - Communication manager establishes connections through fabric
  - Device manager puts attached I/O devices in Solaris devfs
  - Agents (SMA, PMA, etc), Apps (OpenSM) can access fabric components in standard ways

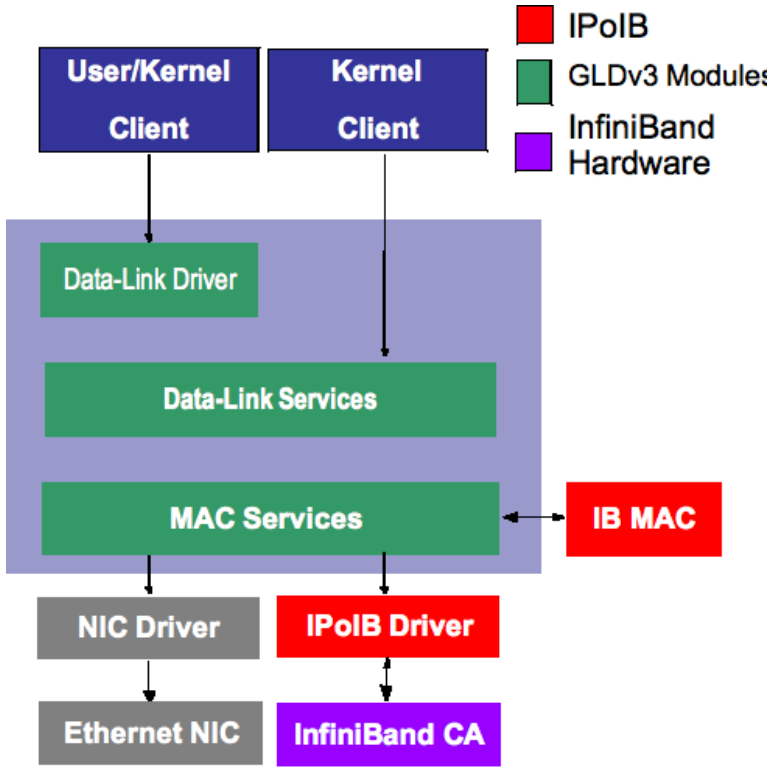
InfiniBand Stack: IP over IB

- ETF standard compliant
  - 4391 – IP transmission over IB
  - 4755 - IP transmission over the RC and UC modes of IBA to support large MTUs
  - 4390 – DHCP over IB
- 20 byte link layer address
  - Most common "porting" issue
  - Does not persist across reboot
- Maps IP subnet to IB partition
- Enables IP addressing for other IB ULPs



InfiniBand Stack: IPoIB (continued)

- TCP checksum offload and LSO, interrupt moderation supported
- dladm(1m) administration
  - "link-mode" property to specify "cm" or "ud" data transfer mode
  - IB "vnics" through new *-part subcommands
- IPMP for high availability
- Interoperable with Linux and Windows
- Observability
  - Snoop and Wireshark for IB



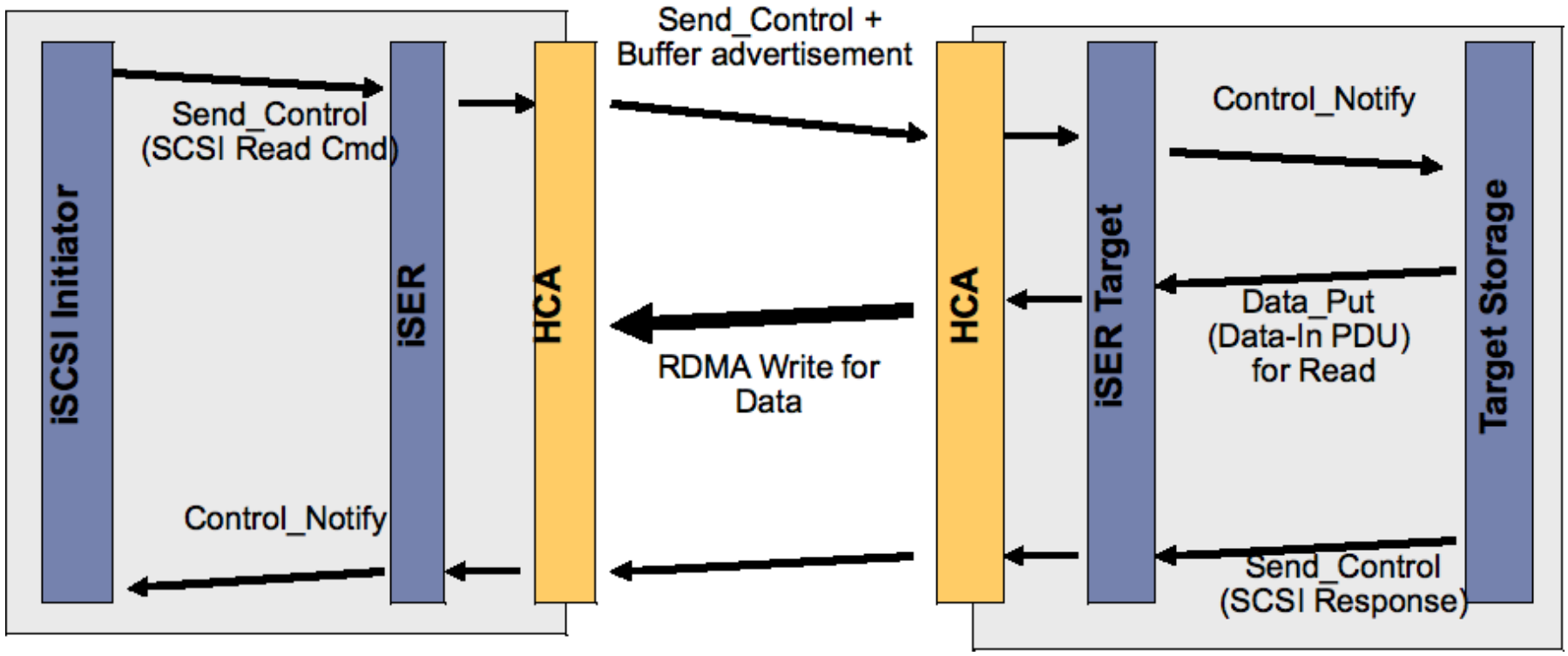
InfiniBand Stack: Encapsulation

- Ethernet over IB (EoIB)
  - Protocol developed by Mellanox and used with the Sun NanoMagnum Gateway (NM2-GW) for bridging IB and Ethernet networks
    - Protocol defines the procedure for tunneling both unicast and multicast Ethernet packets over IB using the Unreliable Datagram (UD) transport
      - The Ethernet packet includes the standard header, VLAN tags, if appropriate, and the payload.
    - Protocol defines control messages used by Gateway management software to associate IB connected servers with Ethernet ports on Gateway
  - EoIB protocol server-side endpoint on Solaris is implemented as standard network driver (NIC)
    - EoIB NICs behave like regular Ethernet NICs
      - Networking stack works seemlessly over EoIB NICs

IB Connected Storage: iSER

- iSCSI extensions to RDMA
  - Defined by the IETF as RFC 5046 with extensions defined by IBTA for InfiniBand
  - iSER extends the data transfer model of iSCSI
    - Reliable protocol and data integrity by IB
    - Zero copy using RDMA
    - Minimal CPU overhead per IO because transport is IB
  - iSER still maintains compliance with iSCSI
    - Uses exactly the same iSCSI PDUs
    - Utilizes existing iSCSI infrastructure (e.g. bootstrapping, MIB, negotiation, naming and discovery, and security)
    - Utilizes an iSCSI mechanism, "login key negotiation" to determine whether to use the iSCSI or iSER data transfer models
    - Add a 12 Byte iSER header before iSCSI PDU (for control and buffer advertisement)

IB Connected Storage: iSER



Example SCSI Read

- Initiator Send Command PDU (Protocol data unit) to Target
- Target return data using RDMA Write
- Target send Response PDU back when completed transaction
- Initiator receives Response and complete SCSI operation

IB Connected Storage

- iSER: iSCSI Extensions for RDMA
  - itadm(1m) and iscsiadm(1m) used to setup initiator
  - Target support in Solaris 11 Express
  - Except for data movement, iSER interoperates as an iSCSI Initiator and Target
    - If both the iSER initiator and target are unable to negotiate an RDMA channel, or the channel fails and it can not be re-established, data movement regress back to TCP/IP
- SRP: SCSI RDMA Protocol
  - ANSI T-10 standard
  - Uses RDMA for high BW, low latency, low overhead
  - First IB storage protocol implemented
  - Considered the IB industry standard
    - Initiator supported on VMware, Linux and Windows
  - Target support in Solaris 11 Express
  - Uses IB addressing and device management model

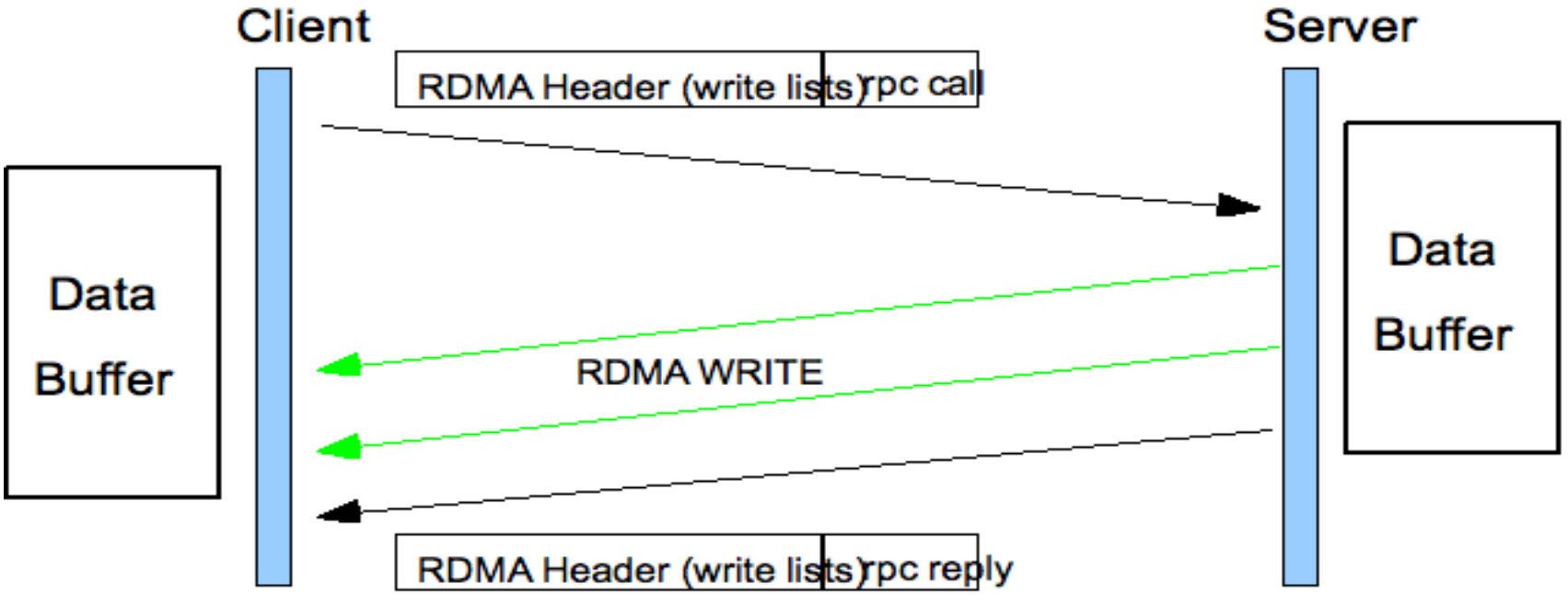
IB Connected Storage: NFS and IB

- NFS and InfiniBand (NFS/RDMA)
- Semantically, NFS is NFS
  - Applications see no difference in semantics
  - RDMA helps in throughput, latency, efficiency
- RDMA under NFSv3 and NFSv4 protocols
- Use Reliable Connected
- NFS runs on standard TCP and UDP networks
  - NFS-over-IPoIB requires no special setup
  - NFS-over-IPoIB does benefit from faster media
- Remote DMA (RDMA) on Infiniband
  - RDMA is transparently negotiated if available
- mount(1M) used to set "proto=tcp" or "proto=rdma"
  - Default mode will negotiate
- RDMA interop tested continuously at Connectathon / Bakeathon events
- Tested to interoperate with Linux
  - Requires at least 2.6.31 kernel

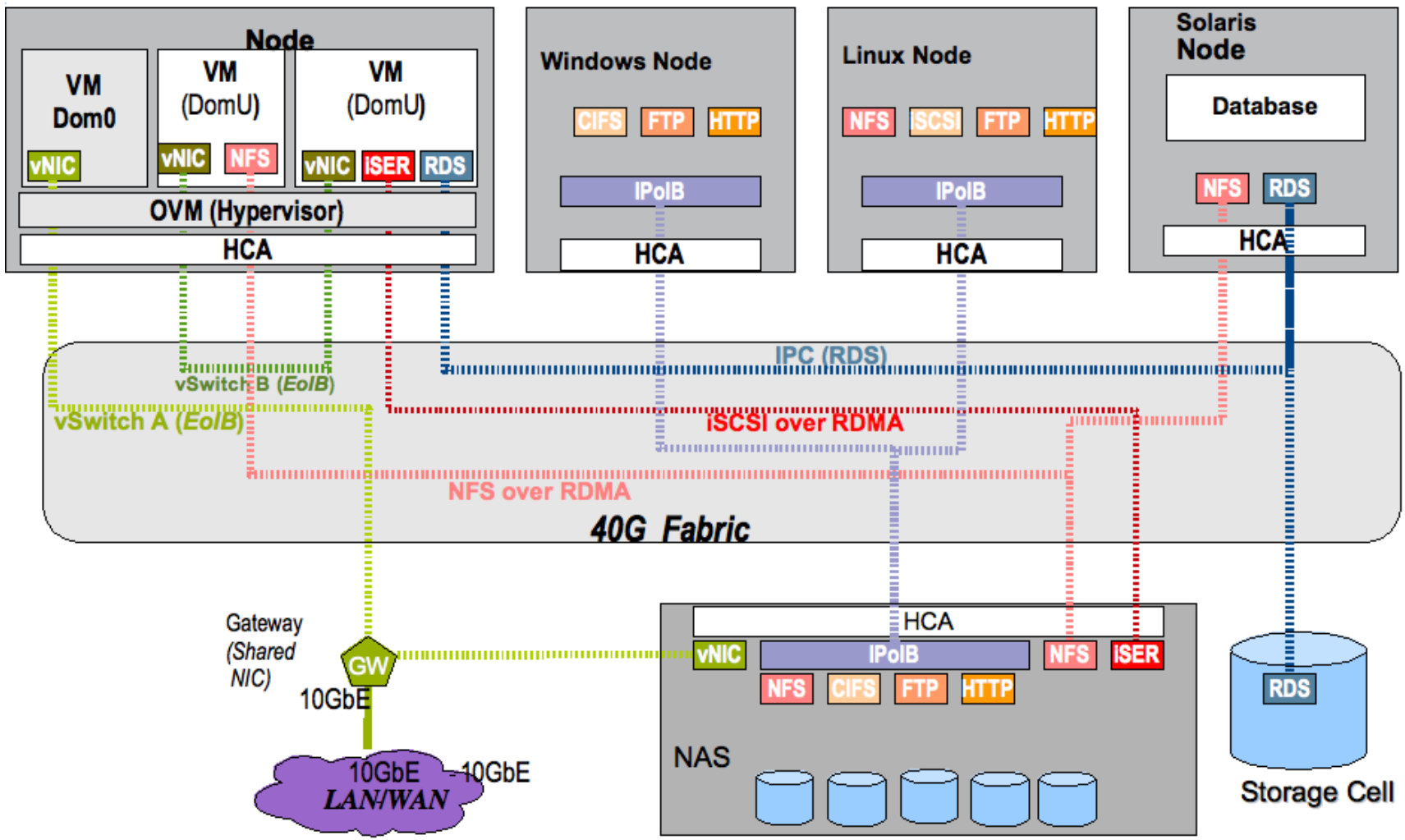
NFS/RDMA (continued)

NFS reads

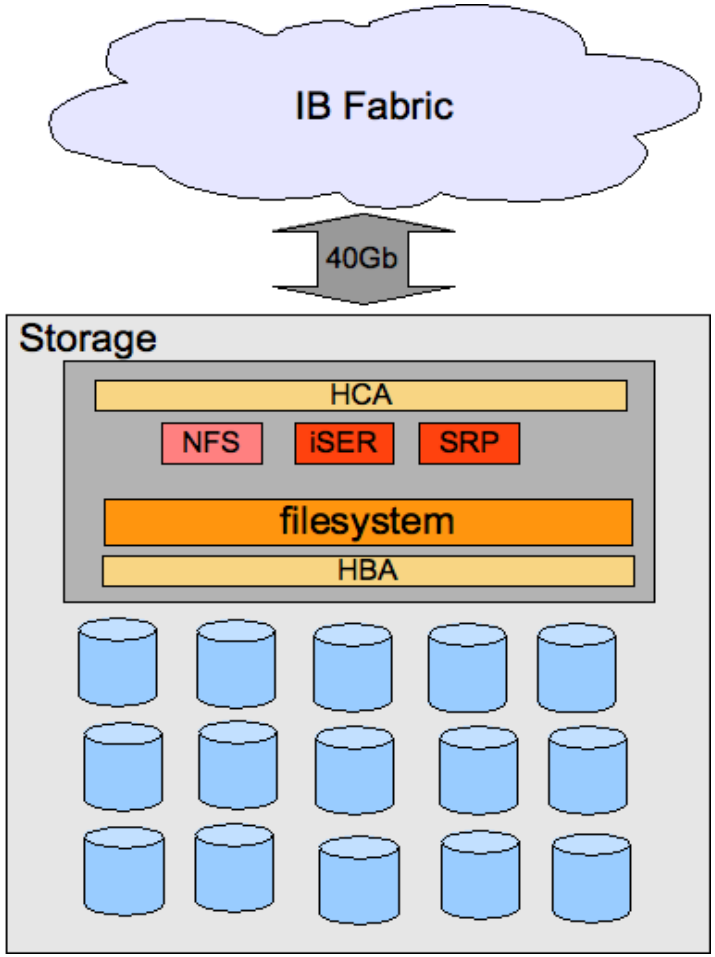
- Client posts a write chunk list
- Server transfers the data to the client using RDMA_WRITE
- Server notifies the client with inline reply



The Converged Fabric Stack

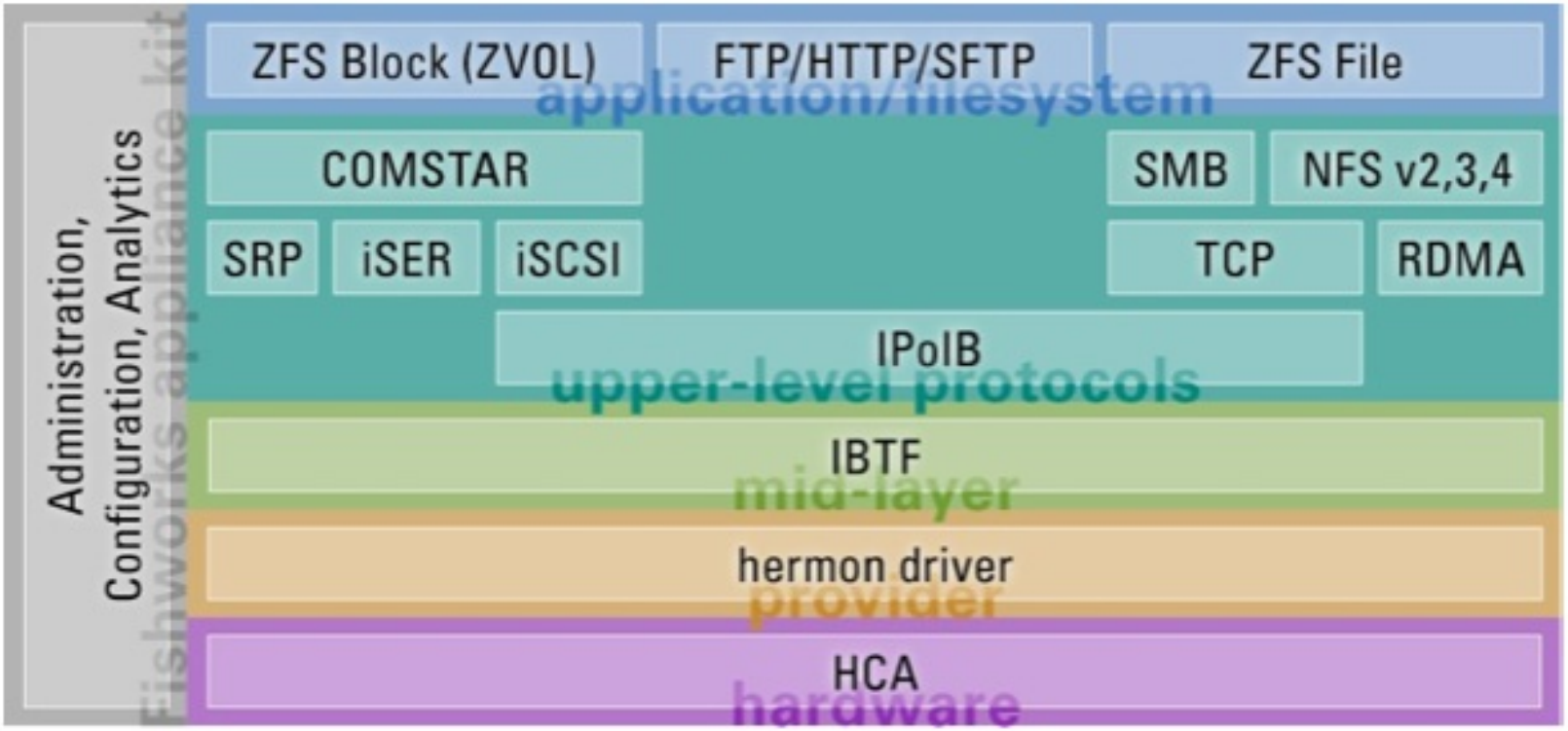


Not there yet…



- **Benefits of InfiniBand negated by bottlenecks of the back-end**
  - Filesystem
  - IO bus
  - HBS
  - Storage fabric
  - Disks

Storage Stack for High Performance



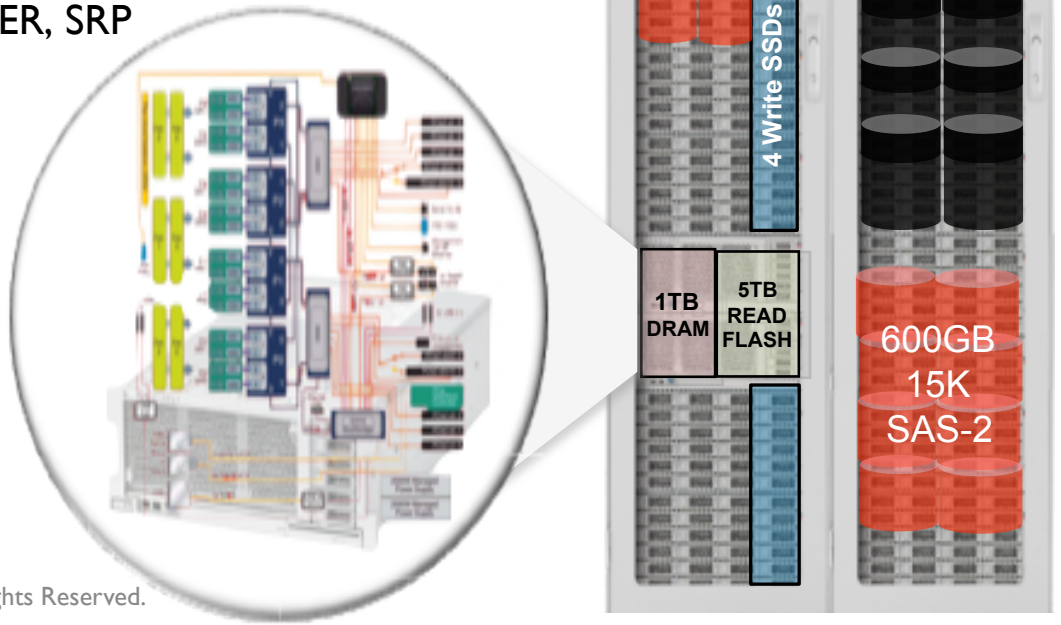
Storage Software for High Performance

- Kernel level Common SCSI Target – COMSTAR
- Reduced copy between filesystem and protocol layers
  - Shared buffers with ZFS and NFS/iSCSI/iSER
- ZFS - Pooled storage: converged filesystems and volumes
  - Pipelined I/O
  - Dynamic striping across
    - Distributes load across devices
  - Intelligent prefetch
    - Multiple independent prefetch streams
  - Variable block sizes
    - Large blocks: less metadata, higher bandwidth
    - Small blocks: more space-efficient for small objects
    - Record-structured file (e.g. databases) sizes matched to filesystem match to avoid read/modify/write

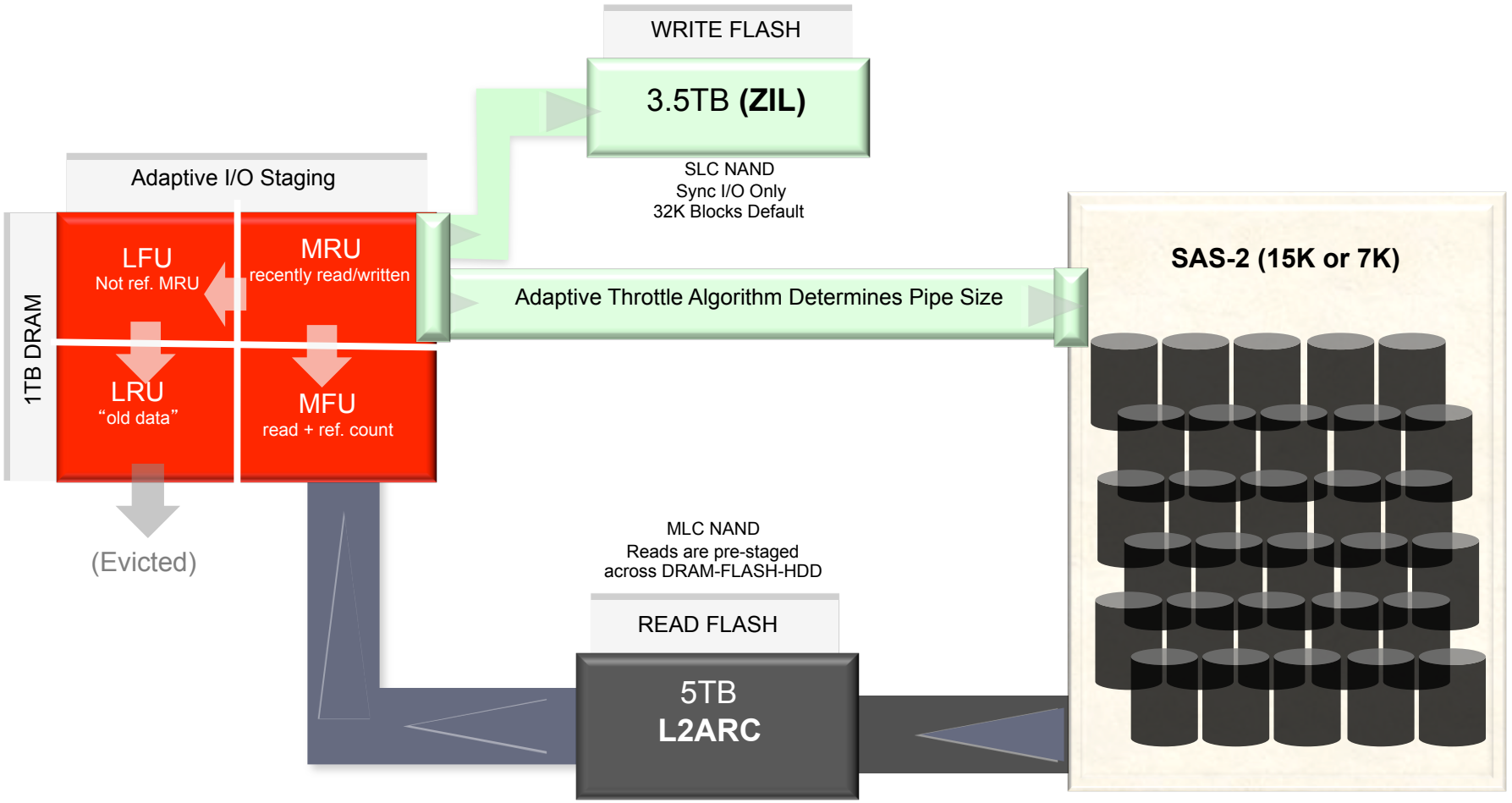
Storage Hardware for High Performance

Using Commodity Components

- SAS-2 architecture
- Gen-2 PCI Express
  - Up to 6 QDR HCAs for high bandwidth
- 40 cores
- Multi-tiered automation
  - RDMA to main memory: NFS, iSER, SRP
  - 1TB DRAM
  - Read Flash
  - Write Flash
  - 15K SAS-2
  - 7.2K SAS-2



Hybrid Storage Pool – ZFS



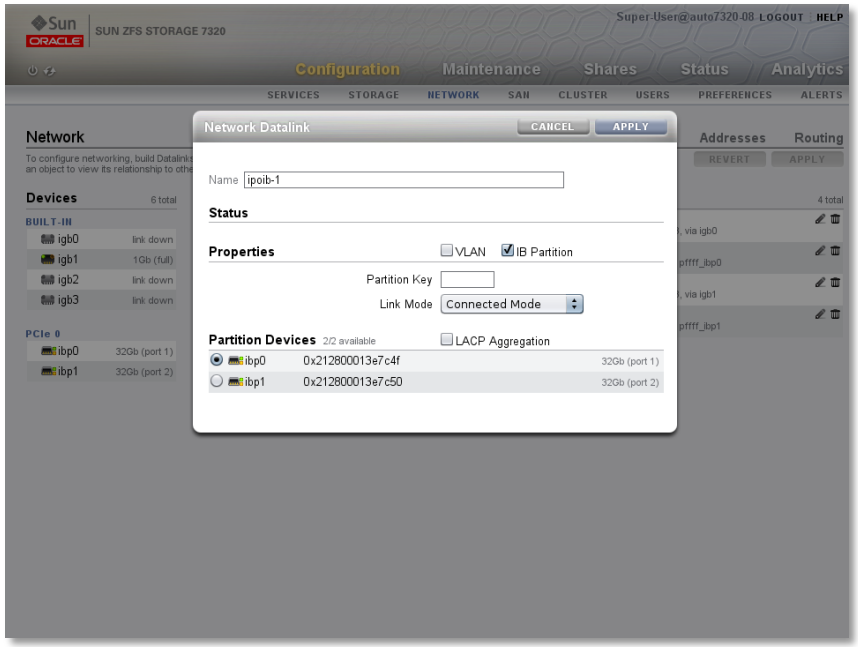
Hybrid Storage Pool – Components

- Disks for high capacity, high bandwidth, low power
  - Great aggregate throughput
- DRAM as primary ZFS cache
  - Used as very fast adaptive replacement cache (ARC)
- Logzilla: write-optimized flash
  - Used as the ZFS intent log (ZIL)
  - Synchronous writes don't wait for disk; acknowledged as soon as they reach the logzilla
  - Data asynchornously streamed to disk
- Readzilla: read optimized flash
  - Used as second tier cache (L2ARC)

Administrative Nightmare?

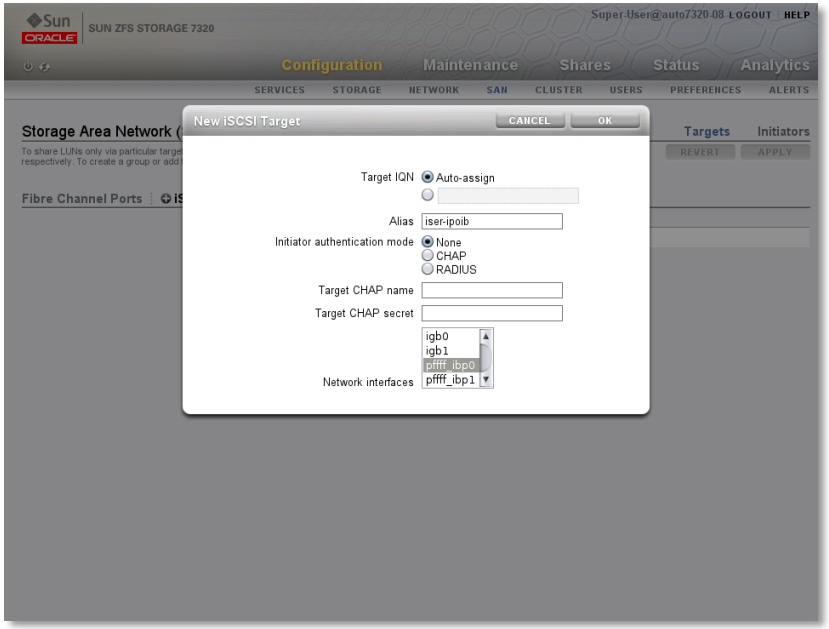
- **Solaris**
  - Integrated into existing command set: ifconfig, dladm, itadm, iscsiadm
  - OFVA verbs library to 'roll your' diagnostic tools
  - Wireshark, snoop
- **Linux**
  - Wireshark
  - OFED stack
  - Commands, libraries, diagnostic, performance tool

Administrative Ease



- Configured just like IP datalinks and interfaces over an IB port
- Specify the partition key and done!

- Configured just like iSCSI targets over IB network interface



Administrative Ease

- Observability and diagnosability
  - CPU usage
  - Protocol IOPS
  - IPoIB and RDMA
    - Throughput
    - Latency

search.oracle.com

ZFS Storage Appliances