Linux Foundation

OPENFABRICS ALLIANCE

Collaboration Summit, HPC Track

Bill Boas,

bboas@systemfabricworks.com

VP, System Fabric Works: Exec. Dir OFA

What is the Alliance?

OPENFABRICS ALLIANCE

- An industry wide community committed to the development, distribution and promotion of open-source software for data center fabrics for high-performance, low latency high availability server and storage connectivity
  - Component, software & system vendors
  - Academic, enterprise & government end-users

AMD APPRO Chelsio Communications CISCO DataDirect NETWORKS DELL Endace

FLEXTRONICS hp IBM intel BERKELEY LAB LSI LOGIC Los Alamos NATIONAL LABORATORY EST. 1943

Mellanox TECHNOLOGIES NetApp Neterion ORACLE QLOGIC RNA Networks

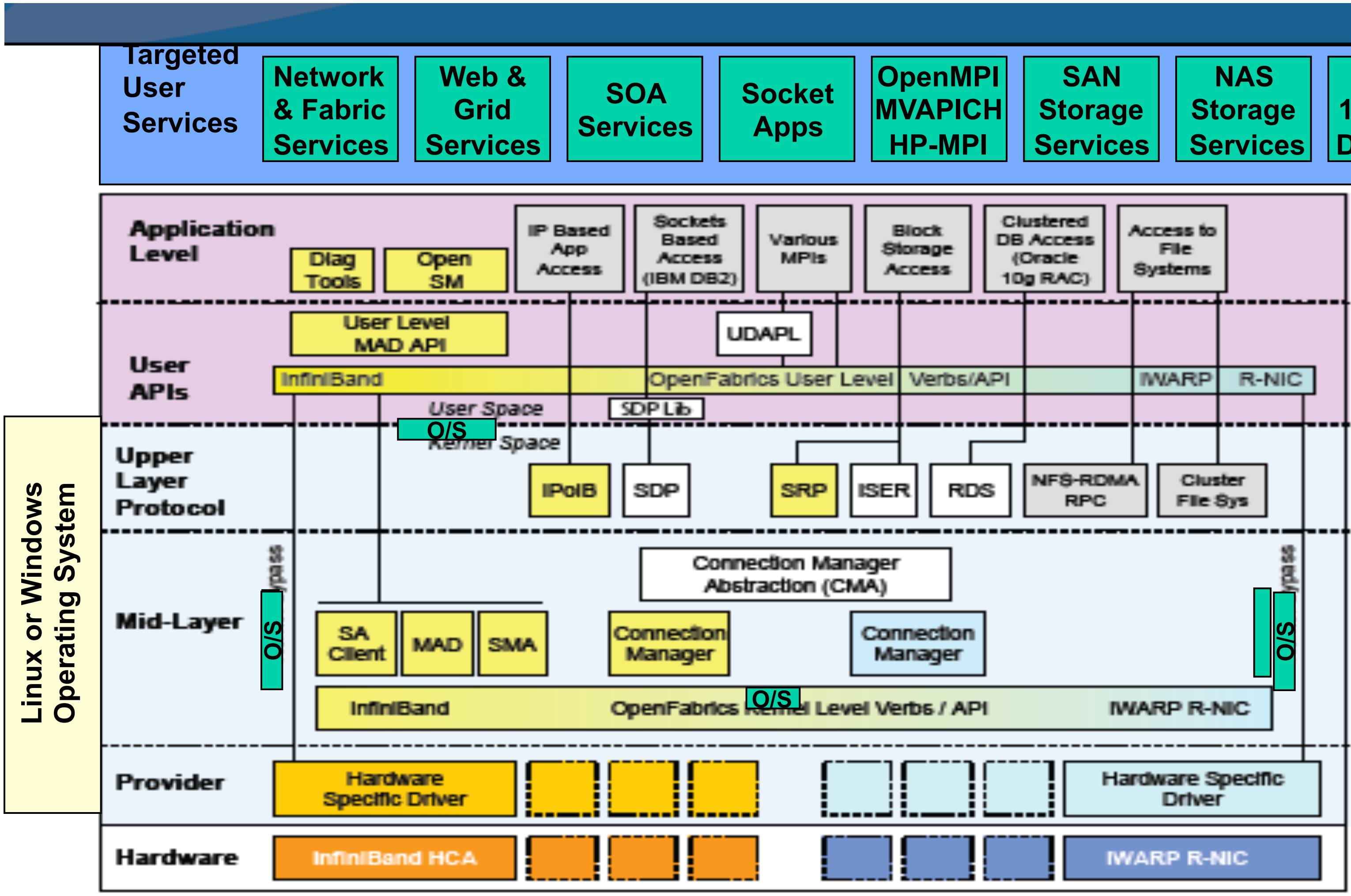
Sandia National Laboratories sgi Sun microsystems System Fabric Works Voltaire XSIGO SYSTEMS

Latest roster at www.openfabrics.org

Mission Statement

- Unify the cohesive development of a single open-source, RDMA-enabled, transport independent software stack that is architected for high-performance, low-latency and maximized efficiency
- Promote industry awareness, acceptance, and benefits of these solutions for server and storage clustering and connectivity applications
- Manage the interoperability testing and certification of the software running on different hardware solutions

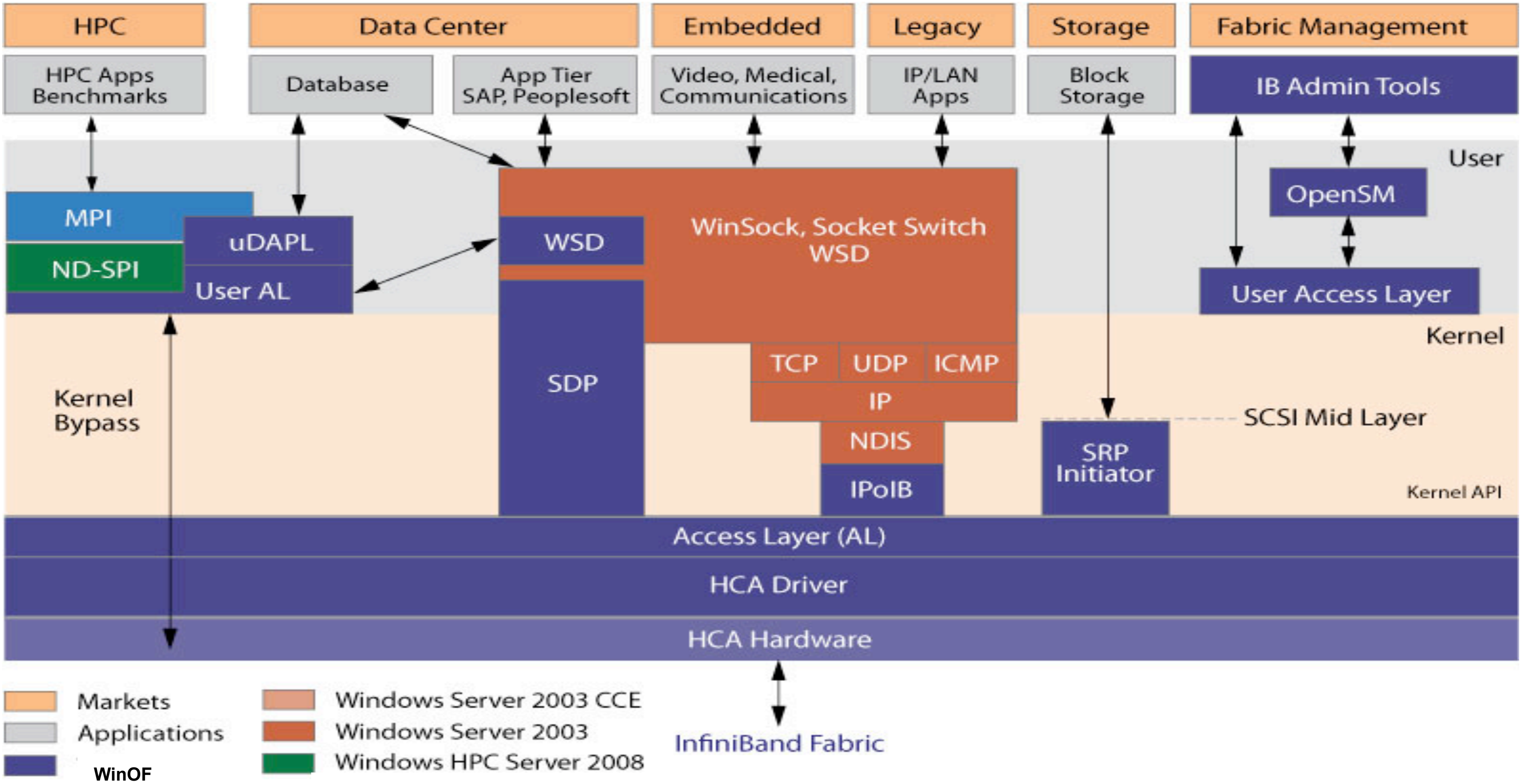
OpenFabrics Software Stack



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

Key
Common
InfiniBand
IWARP
Apps & Access Methods for using OF Stack

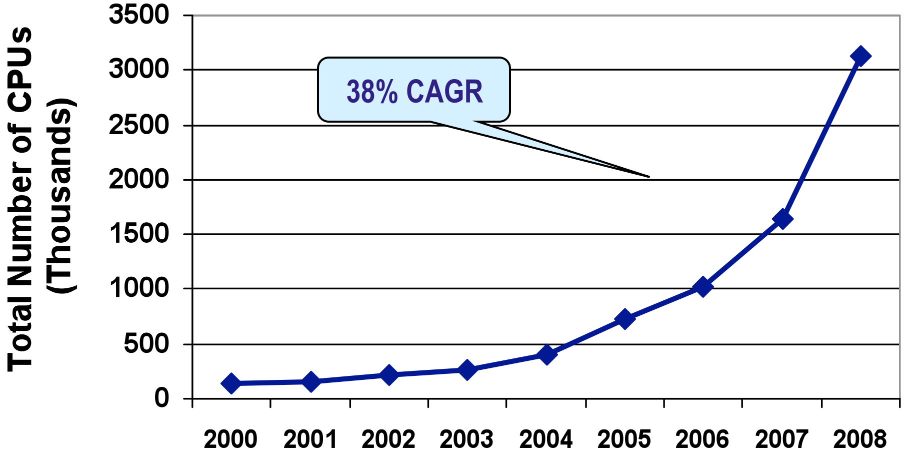
Windows OpenFabrics (WinOF)



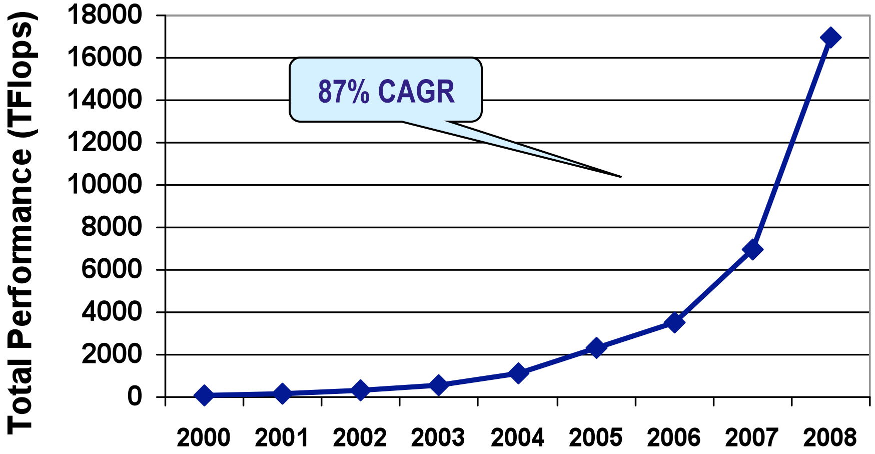
- Supported Platforms
  - x86, x86_64, IA64, XP 32&64, Server 2003 – WHQL, CCS 2003 – WHQL, Server 2008, HPC Server 2008

Top500 Performance Trends

Total # of CPUs on the Top500

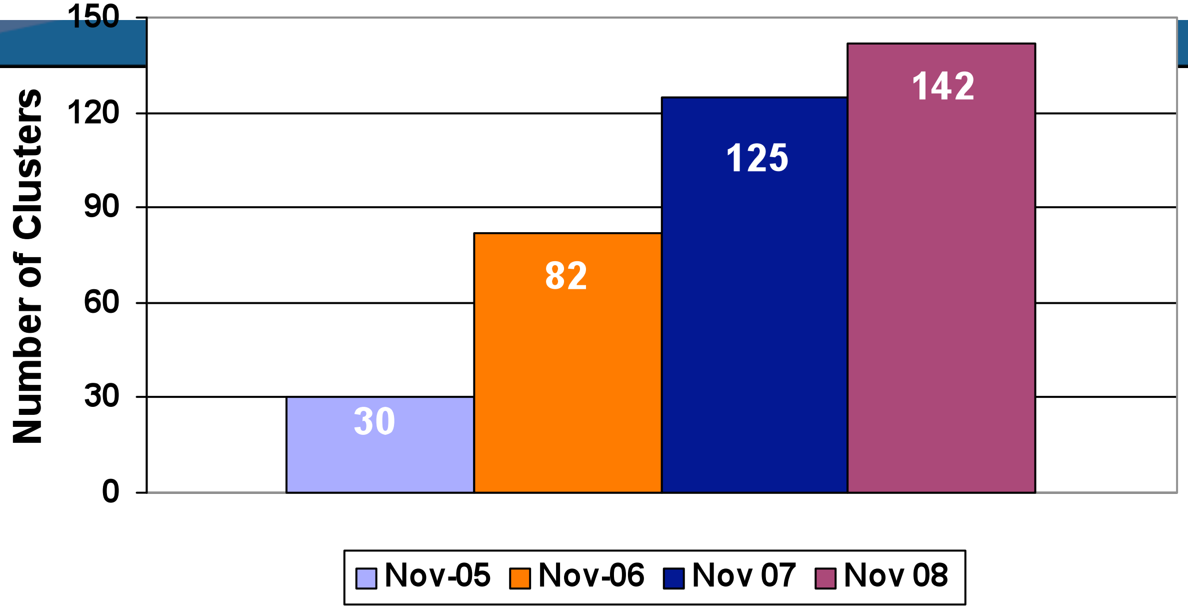


Total Performance of the Top500



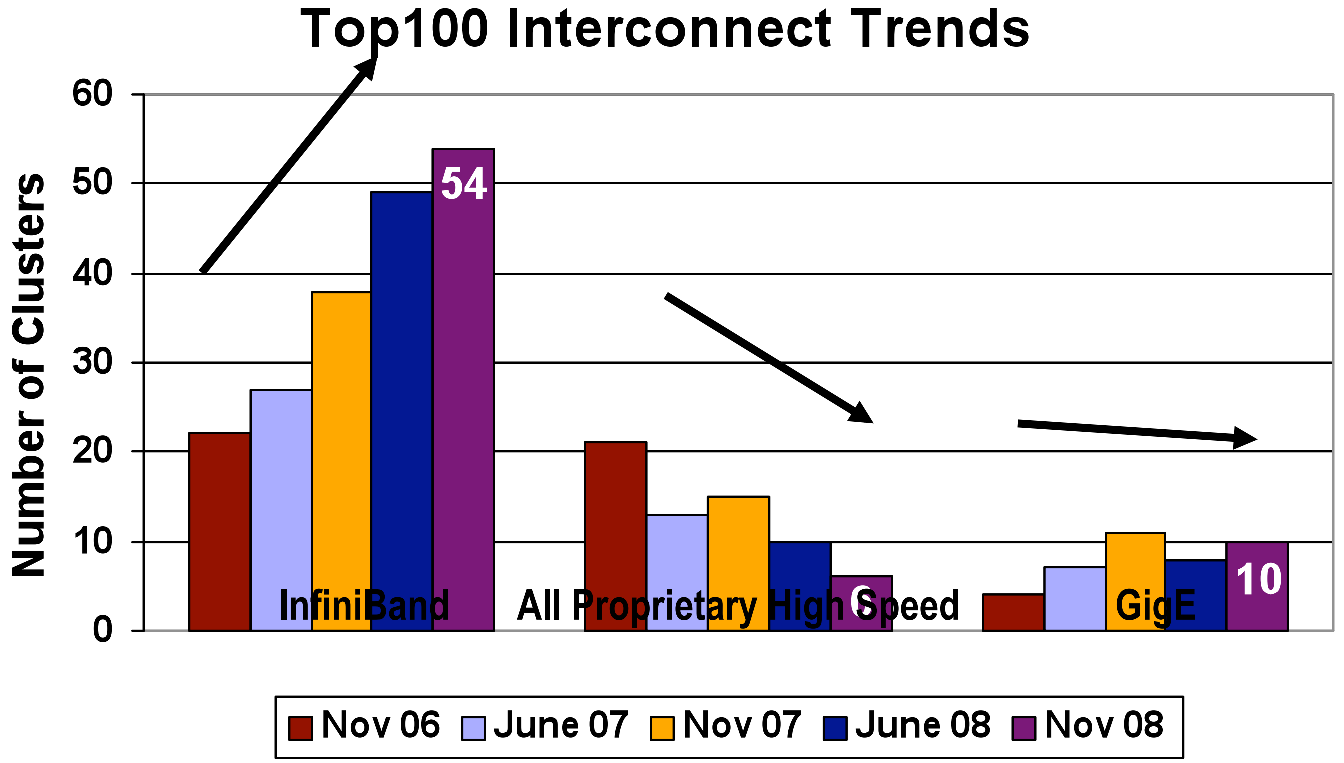
- Explosive computing market growth
- Clusters continue to dominate with 82% of the Top500 list
- Petaflop barrier shattered with the appearance of LANL Roadrunner cluster
  - Interconnect is IB DDR and OpenFabrics software

Top500 InfiniBand Trends



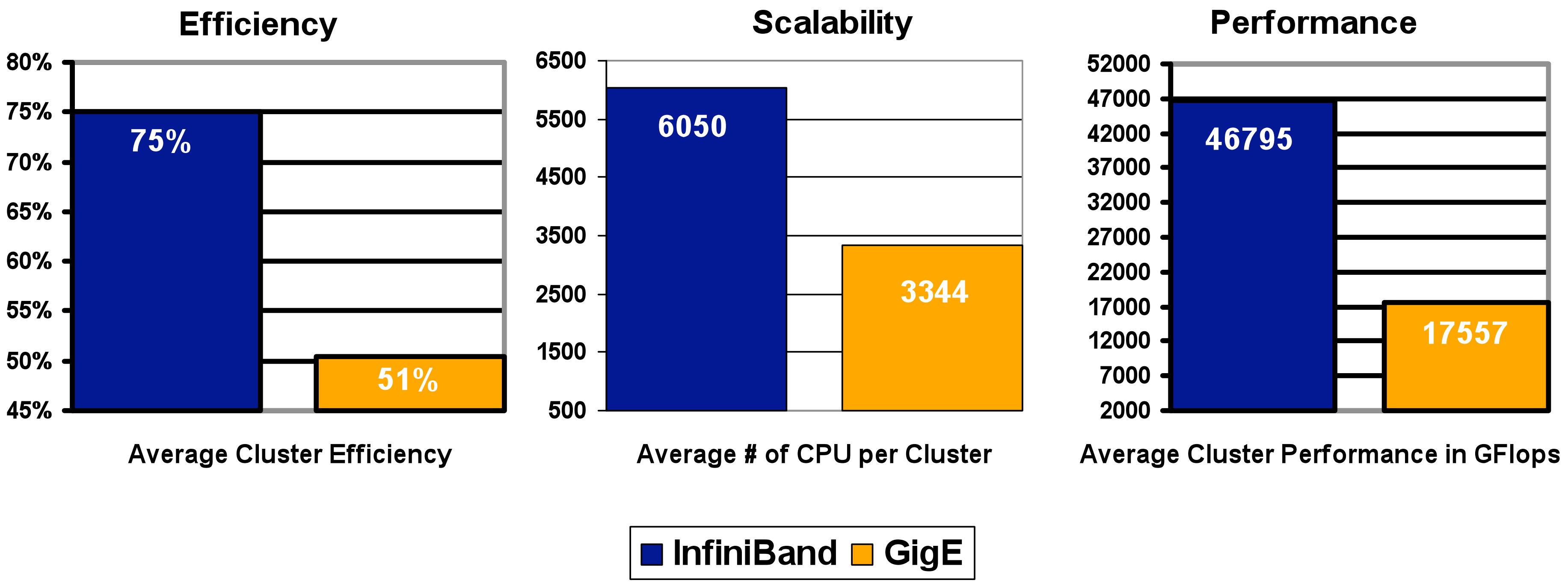
- IB + OFED is the only growing standard interconnect technology
  - 142 clusters, 16% increase versus June 2008 list
  - GigE and proprietary interconnects shows decline, no 10GigE clusters on the list
- IB+OFED makes the most powerful clusters - Top10
  - 4 of the top 10 (#1, #3, #6, #10), both Linux based and Windows based
- The most used interconnect in the Top200
  - 54% of the Top100, 37% of the Top200
- IB+OFED clusters responsible to 35% of the total Top500 performance and these are the most power efficient clusters

Interconnect Trends



- InfiniBand is the only growing high speed interconnect in the Top100
    - 54 clusters, 42% higher than Nov 07 list
    - More than 5X higher than GigE, 9X higher than all proprietary high speed interconnects

Scalabilty, Power and Efficiency



IB + OFED maximizes the cluster’s compute power

Other Industry-Wide Usage

- Financial
- Virtualization
- Database
  - OLTP
  - Data Warehousing
- High Performance Computing
  - Government & Research
  - Commercial
- Hosting Services
  - Cloud Computing
- Web 2.0
- …and many more

29 West Inc GIGASPACES REUTERS TANGOSOL TIBCO The Power of Now® NYSE Euronext Advanced Trading Solutions

- Reduce latency up to 10X
- Predictable data delivery
- 600K → 10M Messages per second
- Algorithmic trading, market making, quotes, arbitrage

Comparison of IB vs. GigE

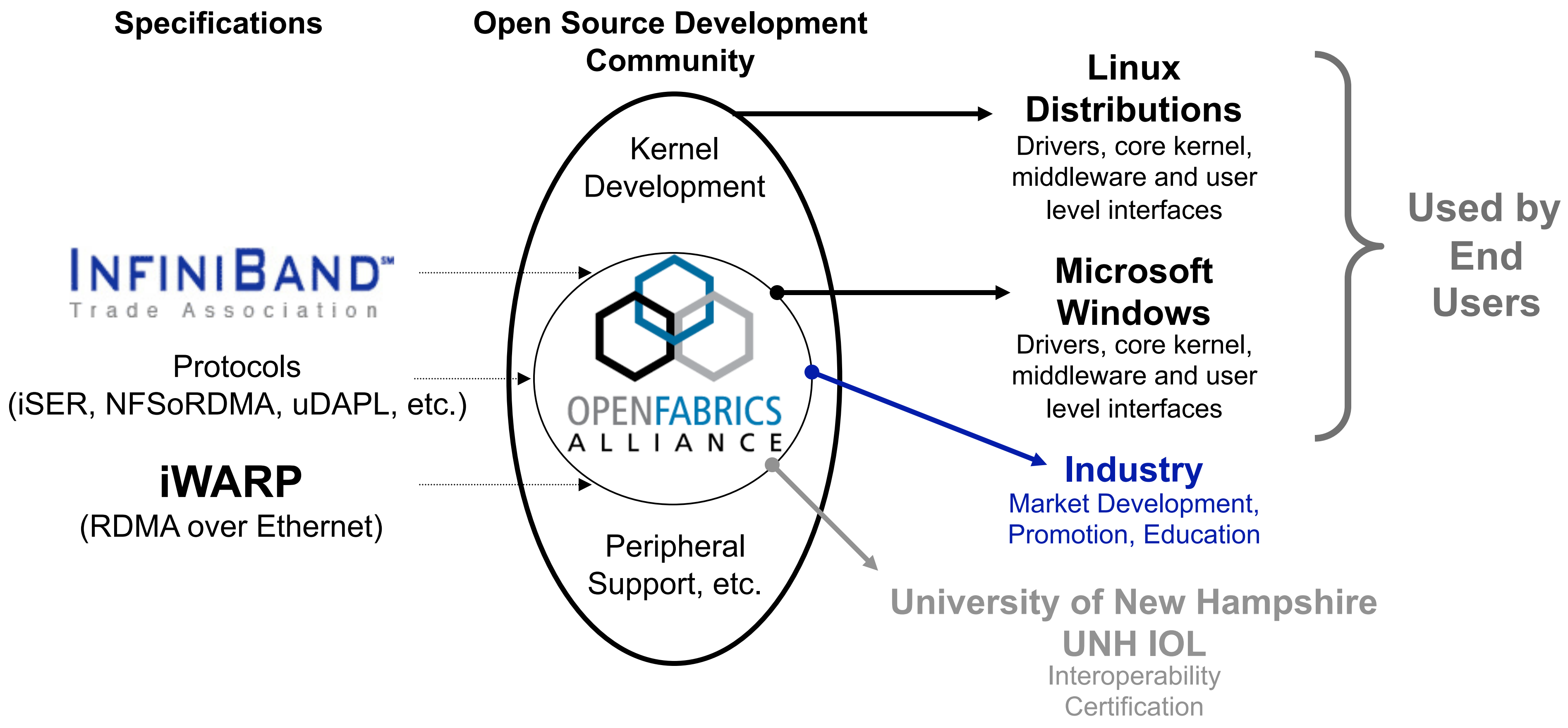
- InfiniBand grid for mission-critical global risk management systems
- 15% to 70% increased HW utilization
- Reduced TCO ($10M/year)

Source: Wall Street & Technology May, 2006

How does the Alliance Work?

- Developers contribute open-source code
  - Often sponsored by vendors or end users
  - In their interest to collaborate on a single robust & high performance stack
- Elected Officers and Working Group volunteers
  - Chairman, Vice Chairman, Treasurer, Secretary and Working Group Chairs
- Open contributions and participation from the industry (both technical and marketing)
- Marketing and promotion through industry events, tradeshows, press releases and end-user interaction

OpenFabrics Alliance (OFA) Role



Specifications

INFINIBAND℠ Trade Association

Protocols (iSER, NFSoRDMA, uDAPL, etc.)

iWARP
(RDMA over Ethernet)

Open Source Development Community

Kernel Development

OPENFABRICS ALLIANCE

Peripheral Support, etc.

Linux Distributions
Drivers, core kernel, middleware and user level interfaces

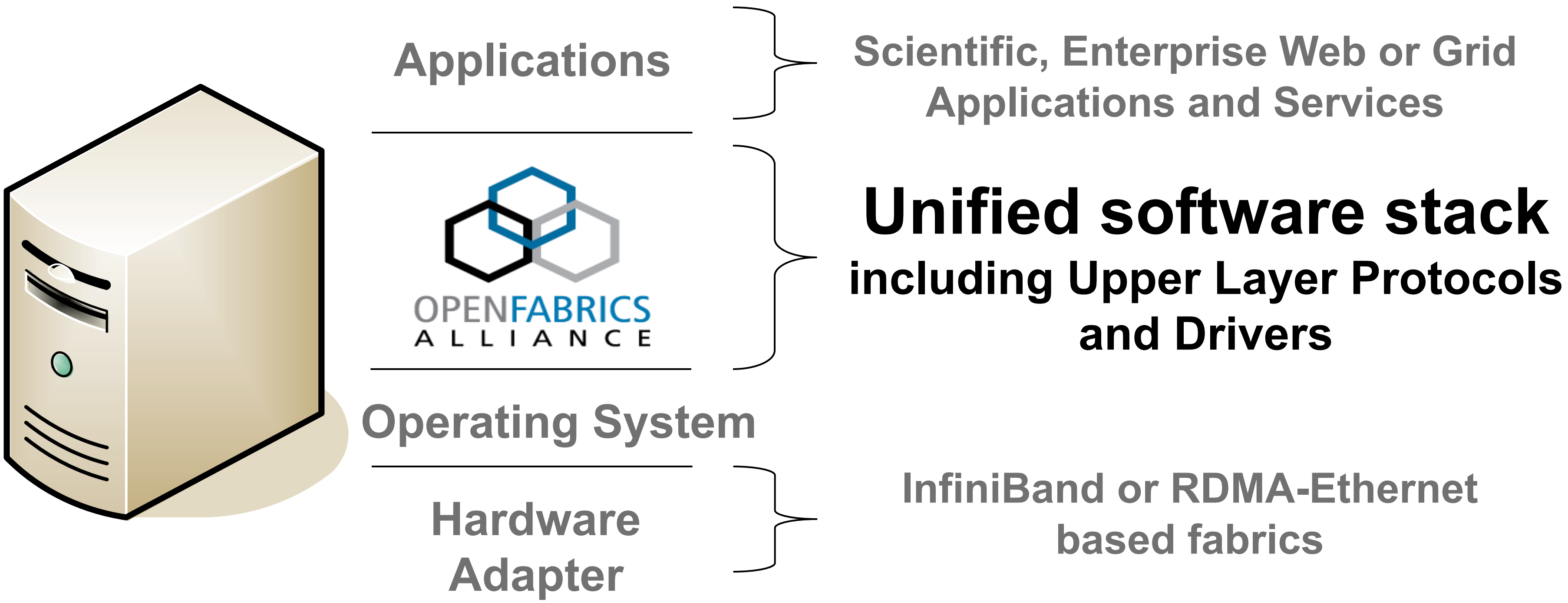
Microsoft Windows
Drivers, core kernel, middleware and user level interfaces

Used by End Users

Industry
Market Development, Promotion, Education

University of New Hampshire UNH IOL
Interoperability Certification

Transport Independence



Applications

Scientific, Enterprise Web or Grid Applications and Services

OPENFABRICS ALLIANCE

Unified software stack

including Upper Layer Protocols and Drivers

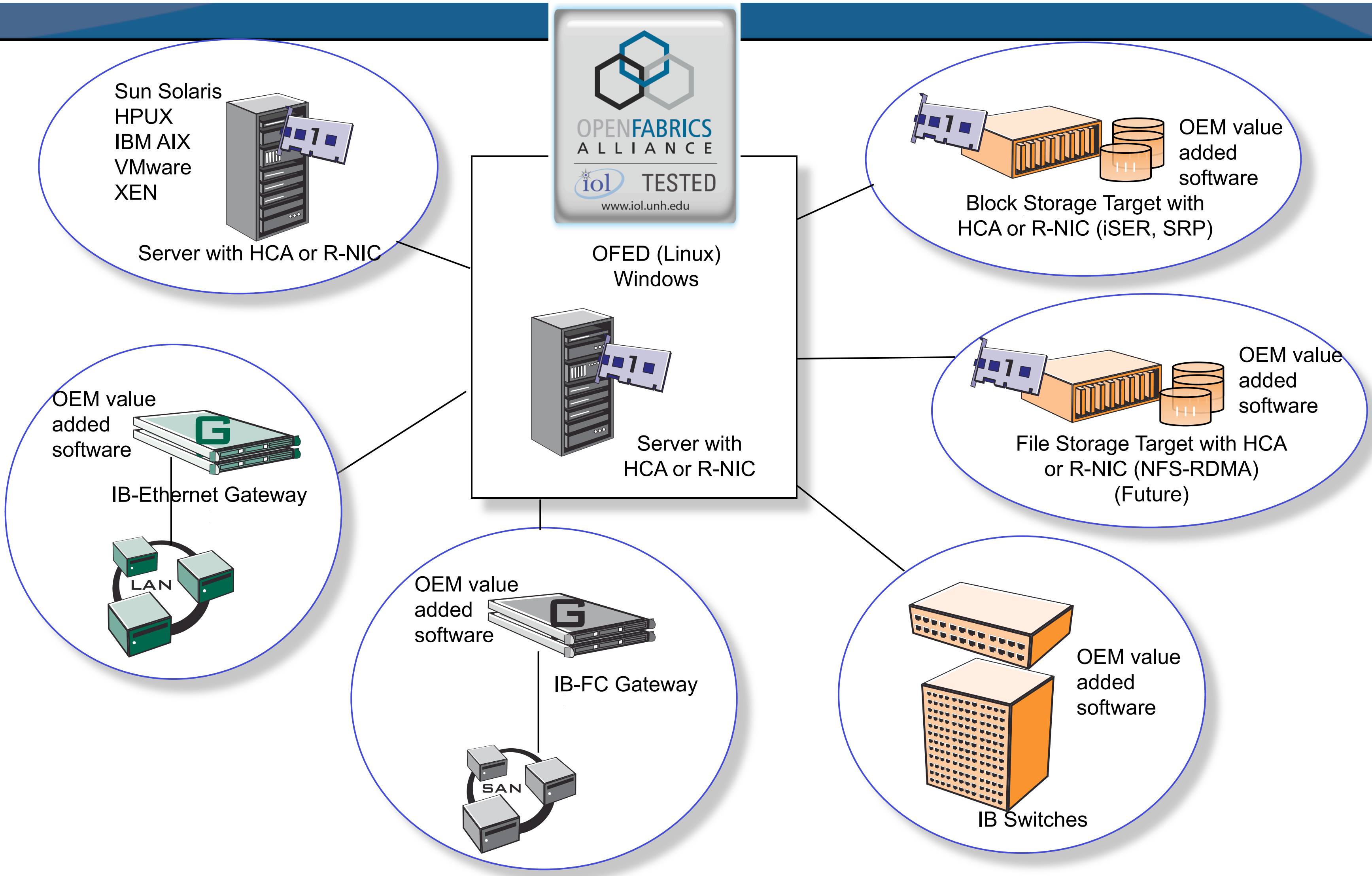
Operating System

Hardware Adapter

InfiniBand or RDMA-Ethernet based fabrics

- Leveraging a single software stack, developers and end-users have the freedom to chose a fabric solution
- Allows operating systems and applications to maximize performance and efficiency when interacting with the fabric

Logo Interoperability Program



OPENFABRICS ALLIANCE

iol TESTED

www.iol.unh.edu

OFED (Linux)
Windows

Server with
HCA or R-NIC

Sun Solaris
HPUX
IBM AIX
VMware
XEN

Server with HCA or R-NIC

OEM value
added
software

Block Storage Target with
HCA or R-NIC (iSER, SRP)

OEM value
added
software

File Storage Target with HCA
or R-NIC (NFS-RDMA)
(Future)

OEM value
added
software

IB-Ethernet Gateway

LAN

OEM value
added
software

IB-FC Gateway

SAN

OEM value
added
software

IB Switches

History

- JUN 2004 – Founded as OpenIB.org w/ Focus on IB + Linux
  - Funding from the U.S. Department of Energy
- APR 2005 – Added Windows Development
- NOV 2005 – Hosted IB SCinet at SC|05, 30+ Vendors
- MAR 2006 – Expanded Charter to include iWARP and changed name to OpenFabrics.org
- JUN 2006 – First OFA Enterprise Distribution release (IB)
- NOV 2006 – Hosted InfiniBand & iWARP SCinet at SC|06
- Need to add Windows Release

Working Groups

- ➢➢ Working Groups are subset of members who do work!
  - Each group is led by an appointed Chair and Vice-Chair
  - Any OpenFabrics member is free to participate and contribute
- Executive (XWG): Delegated to run OFA
- Developers (DWG): Code creation and maintenance
- Enterprise (EWG): Qualified and tested distribution of code
  Interoperability (IWG): works with UNH-IOL to validate and certify
- Legal (LWG): Code contribution and licensing
- Marketing (MWG): Recruiting and promotion
- User (UWG) and HSIR (High Speed Interconnect Roundtable): End-user requirements, including Wall Street

Licensing and Development

- OFA serves as the code repository
- Dual-license allows for inclusion in both open-source and non-open source operating system environments
  - Code checked in under GPL **AND** BSD
  - Code checked out under GPL **OR** BSD
- Current development focus
  - InfiniBand and iWARP (RDMA-over-Ethernet) interconnect technology
  - Linux and Microsoft Windows operating systems
  - Xen virtualization

OFED Status and Futures

- Linux OFED components
- Releases done in last year:
  - OFED 1.3.1
  - OFED 1.4
- 2009 Plans:
  - OFED 1.4.1
  - OFED 1.5
- How to contribute

Linux OFED Components

OFA Development

- HCA/NIC Drivers
  - IB: IBM, Mellanox, QLogic
  - iWARP: Chelsio, Intel
- Core: Verbs, mad, SMA, CM, CMA
- IPoIB
- SDP
- SRP and SRP Target
- iSER and iSER Target
- RDS
- NFS-RDMA
- Qlogic_VNIC
- uDAPL
- OSM
- Diagnostic tools

Add on

- Bonding module
- Open iSCSI
- MPI Components
  - MVAPICH
  - Open MPI
  - MVAPICH2
  - Benchmark tests

Tested with

- Proprietary MPIs: Intel, HP
- Proprietary SMs: Cisco, Voltaire, Qlogic

2008 Look Back

- Linux OFED components
- Releases done in last year:
  - OFED 1.3.1
  - OFED 1.4
- 2009 Plans:
  - OFED 1.4.1
  - OFED 1.5
- How to contribute

OFED 1.3.1

- OFED 1.3.1 release on June 3, 2008
  - Added support for RedHat EL 5.2 and SLES 10 SP2
  - Fixed several critical bugs
- Distro integration:
  - Red Hat AS 4.7 and RHEL 5.2, SLES10 SP2
- Used in Intel ® Cluster Ready Solutions
- Passed Oracle 11g certification with RDS

OFED 1.4

- General Info
  - Released in December 10, 2008
  - Passed in the Interoperability event in Nov 2008
  - Added support for CentOS and OEL (Oracle Enterprise Linux)
  - Kernel base 2.6.27
- Distro integration:
  - SLES 11
  - RHEL 4.8, 5.4 (not released yet)
- Used in Intel ® Cluster Ready Solutions

OFED 1.4 Features

- New: BMME verbs (fast memory thru send queue (FRWR); Local invalidate send work requests; Read with invalidate)
- New: iSer Target
- New: NFS-RDMA – as technology preview
- New: VPI support – Eth and IB for ConnectX

OFED 1.4 Features – Cont.

- IPoIB:
  - LRO and LSO for Datagram mode
  - Improved Bonding failover response time
- uDAPL:
  - Socket CM for scalability and interop with Windows
  - UD extensions
- Qlogic_vnic:
  - Hot swap of EVIC and dynamic update of existing connections with QLogic dynamic update daemon.
  - Performance improvements of Ethernet broadcast & multicast traffic.

OFED 1.4 Features - Cont.

- New management package (ver 3.2):
  - OpenSM
    - Cashed routing
    - Multi lid routing balancing for updn/minhop routing algorithms
    - Preserve base lid routes when LMC > 0
    - OpenSM configuration unification
    - IPv6 Solicited Node Multicast addresses consolidation
    - Routing Chaining
    - Failover/Handover improvements: Query remote SMs during light sweep
    - Ordered routing paths balancing
  - ibutils:
    - Report created in CSV format
    - Congestion Control in ibutils
  - Diagnostic tools: ibnetdiscover library - to accelerate another tools

OFED 1.4 Features - Cont.

- New MPI versions:
  - OSU MVAPICH 1.1.0
  - Open MPI 1.2.8
  - OSU MVAPICH2 1.2p1
  - Tests: Updated IMB 3.1

OFED 1.4 OS Matrix

- kernel.org: **kernel 2.6.26 and 2.6.27**
- Novell
  - SLES 10
  - SLES 10 SP1 (up1)
  - SLES 10 SP2
- Redhat
  - RHEL 4 (up4, up5, up6, **up7**)
  - RHEL 5 (up1, up2)
- OEL
  - **OEL 5**
- Free distros (with limited QA):
  - Open SuSE 10.3
  - **Fedora Core 9**
  - Ubuntu 6.06 (with RPM package installed)
  - **CentOS 5.2**

** new for OFED 1.4 in bold*

2009 OFED 1.4.1

- Support for RHEL 5.3 and SLES 11
- NFS/RDMA in beta
  - OSes: RHEL 5.2, 5.3 and SLES 10 SP2
- Open MPI 1.3.1
- RDS with iWARP support in beta
- VPI ConnectX IB/Eth port sensing
- Critical bug fixes

OFED 1.4.1

- Schedule:
  - RC1 - Mar 4 - done
  - RC2 - Mar 19 - done
  - RC3 - Apr 2
  - GA - Apr 20

OFED 1.5 Features Plans

- Kernel base: 2.6.30
- Add support for RedHat EL 5.4 and EL 4.8
- Kernel verbs: Multiple Event Queues to support Multi-core CPUs
- NFS/RDMA – GA
- RDS from the kernel; support for iWarp – GA
- SDP – Performance improvements: small and medium messages BW, reduced jitter, GA quality
- Support for Mellanox vNIC (EoIB) and FCoIB with BridgeX device
- New MPI features – details in MPI session
- More features according to requirements that will be raised here …

OFED 1.5 – Management Features

- Unify API with Windows
- OSM:
  - Fat-tree enhancements:
    - Connect roots
    - Credit loop-free multicast routing with managed switches
  - SM handover – enable SM on every node
  - Shadow SA DB
  - M_Key management
- More details in OpenSM Update

OFED 1.5 OS Matrix

- kernel.org: **kernel 2.6.29 and 2.6.30**
- Novell
  - SLES 10
  - SLES 10 SP1 (up1)
  - SLES 10 SP2
  - SLES 11
- Redhat
  - RHEL 4 (up4, up5, up6, up7, **u8**)
  - RHEL5 no updates, up1
  - RHEL 5 (up2, u3, **up4**),
- OEL
  - OEL 5
- Free distros (with limited QA):
  - Open SuSE 10.3
  - Fedora Core 9
  - Ubuntu 7 (with RPM package installed)
  - CentOS 5.2, 5.3

- *new for OFED 1.5 in **bold***
- *drop support for items in blue*

OFED 1.5 Schedule

- Preliminary Schedule
  - Development tree opened when 2.6.30-rc1is available
    - People can start development now against 2.6.29 Linux kernel
  - Feature Freeze: May 7, 09
  - Alpha Release: May 12, 09
  - Beta Release: Jun 9, 09
  - RC1: Jun 25, 09
  - RC2-RCx: About every 2 weeks as needed
    - We usually have ~6 RCs
  - Release: Sep 15, 09

What is an RC?

- RC = Release candidate – something pretty close to what we'd like to release.
- An early RC will be sent for interoperability testing.
- Not the time to complete your new feature!
- This is the opportunity for testing and fixing bugs.

How to contribute?

- Developing new code and features
- Bug fixes
- Performance tuning
- Contribute backports for new OSes
- Doing QA and testing
- Sending patches and comments to the mailing lists:
  - ewg@lists.openfabrics.org **– OFED specific only**
  - general@lists.openfabrics.org **– General development**
- Opening bugs in Bugzilla (https://bugs.openfabrics.org/)
  - When opening a new bug you should choose OpenFabrics Linux
  - Old bugs must be tested with new releases and updated on bugzilla
- Participate in EWG bi-weekly meetings
  - Meeting minutes on the web: http://www.openfabrics.org/txt/documentation/linux/EWG_meeting_minutes/

Benefits of Membership

- Understand latest development status and schedules
- Influence the development of capabilities and features you need recognized and prioritized
- Association with marketing efforts
  - Press releases, tradeshows, speaking opportunities, workshops
- Interaction with industry thought leaders
- **If your organization is using or is interested in using RDMA-enabled fabric technology, please talk to me after**

Four Membership Levels

- **Promoters ($5000/year, $3000 initiation)**
  - Organizations and enterprises that wish to strongly influence the process and features in software created and the accompanying promotional activities to enhance the code they use or provide
- **Adopters ($3000/year, $3000 initiation)**
  - Organizations and enterprises that wish to contribute to and participate in the processes and work of the promoters but do not feel the need to strongly affect the outcomes
- **Supporters ($1000/year, $3000 initiation)**
  - Organizations and enterprises that wish to use the OpenFabrics software, leverage the promotional activities, be tied into the work of the Alliance but not necessarily contribute
- **Consulting (Free)**
  - Organizations and individuals that the Alliance selects for honorary membership on an annual basis based on the perceived value of their membership to the Alliance

- All members agree to understand the Bylaws and Membership agreements and to work within the Alliance processes and rules described therein

Join Today!

- Key Contacts
  - Jim Ryan Chair – jim.ryan@intel.com
  - Bill Boas, Vice Chair – bboas@systemfabricworks.com
  - Johann George, Treasurer – johann@georgex.org
  - Wayne Augsberger – Marketing Chair – wayne@mellanox.com
- To join the Alliance review Bylaws and sign Membership Agreement
  - Available for download at www.openfabrics.org
  - Return agreement to the Chair
- Pay membership fee to the Treasurer
- Start attending monthly promoters meetings and working group meetings and contribute as appropriate