Agenda

- Introduction to Mellanox Technologies and Products
- Meeting Data Center Server and Storage Connectivity Needs
- InfiniBand in VMware Community Source Development
  - Experiences
  - Enhancements and Optimizations
  - Performance Testing and Results
- End User Benefits
- Release Plan
- Future Enhancements
- Summary
Mellanox Technologies

- A global leader in semiconductor-based server and storage connectivity solutions
- Leading provider of high performance InfiniBand solutions
  - 10Gb/s and 20Gb/s Node-to-Node
  - 30Gb/s and 60Gb/s Switch-to-Switch
  - 2.25us latency
  - 3.5W per HCA (host channel adapter) port
- RDMA and hardware based connection reliability
- Efficient and scalable I/O virtualization
- Price-performance advantages
- Converges clustering, communications, management and storage onto a single link with Quality of Service
InfiniBand Interconnect Building Blocks

Mellanox Technologies

**INFINIBAND ADAPTERS**

- Single-port and dual-port

**INFINIBAND SWITCH SILICON**

- 8-port and 24-port

OEMs

- Servers
- Storage
- Communications Infrastructure Equipment
- Embedded Systems

End Users

- Enterprise Data Centers
- High Performance Computing
- Embedded Systems

*Single-port and dual-port
** 8-port and 24-port
InfiniBand Software & OpenFabrics Alliance

Components are in kernel 2.6.11+

SLES 10 and RHEL 4 Distributions

Microsoft WHQL program
Global Market Penetration

Strong Tier-1 OEM relationships - the key channel to end-users
Connectivity Needs in Server Farms

Front End Servers: Web and other services
- 10/100 Mb/s, 1 Gb/s Ethernet

Application Servers: Business Logic
- 1 Gb/s Ethernet

Back End Servers: Database systems
- 1-10 Gb/s Ethernet, InfiniBand

Storage Servers and Systems
- 1-10 Gb/s FC, InfiniBand, iSCSI

- Server islands → different connectivity needs
- Need for higher resource utilization
- Shared pool of resources
- Maximum flexibility, minimum cost

Uniform connectivity that serves most demanding apps
Why 10Gb/s+ Connectivity?

- Multi Core CPUs
- SAN Adoption
- Shared Resources

More applications per server and I/O
More traffic per I/O with server I/O consolidation
I/O capacity per server dictated by the most demanding apps

Multi-core CPUs mandating 10Gb/s+ connectivity

10Gb/s+ connectivity for all data center servers
Emerging IT Objectives

**IT as a Business Utility**
- Applications just work
- Low maintenance cost
- Do more with less
- Price-performance-power

**IT as a Competitive Advantage**
- Reliable service
- High availability
- Agility and Scalability
- Quicker business results

Server and storage I/O takes on a new role
## Delivering Service Oriented I/O

### End-to-End Quality Of Service
- Congestion control @ source
- Resource allocation

### I/O Consolidation
- Multiple traffic types over one adapter
- Up to 40% power savings

### Optimal Path Management
- Packet drop prevention
- Infrastructure scaling @ wire speed

### Dedicated Virtual Machine Services
- Virtual machine partitioning
- Near native performance

---

**Guaranteed services under adverse conditions**

---

**VMWORLD 2006**
End-to-end Services

- Quality of service
  - More than just 802.1p
  - Multiple link partitions
  - Per traffic class
  - Shaping
  - Resource allocation

- Congestion control
  - L2 based
  - Rate control at source
  - Class based link level flow control
  - Protection against bursts to guarantee no packet drops

Enables E2E traffic differentiation, maintains latency and performance
Scalable E2E Services for VMs

- Scaling to millions of VEPs (virtual functions) per adapter
- Isolation per VEP
- Switching between VEPs
  - Inter VM switching
- QoS, congestion control per VEP
- End-to-end service delivery per VEP

Virtualization Enabler/Intermediary

20 Gb/s
<table>
<thead>
<tr>
<th>Vendor</th>
<th>Bandwidth per port</th>
<th>Power/ port</th>
<th>RDMA</th>
<th>Stateless Offload</th>
<th>Full Offload</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor A</td>
<td>10 or 20Gb/s</td>
<td>~4-6W</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Vendor B</td>
<td>10Gb/s</td>
<td>~20W</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>10Gb/s</td>
<td>12W</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Price-Performance Benefits

End User Price per MB/s
(Adapter+Cable+Switch Port)/ Measured Throughput

Lower is better

Source for GigE and 10GigE iWARP: Published press release adapter and switch pricing, from Chelsio, Dell, Force10

Industry leading price-performance
VMware Community Source Involvement

- One of the first to join Community Source
- Active involvement since
- Motivations
  - Work with virtualization vendor with most market share
  - Spirit of partnership and excellent support
  - Eventual Virtual Infrastructure-based product deployments
- Benefits over Xen
  - Real-world customer success stories
  - Proven multi-OS VM support
  - Go-to-market partnership possibilities

Pioneering role in the Community Source program
The InfiniBand Drivers

- Linux based drivers used as basis
- Device driver, IPoIB and SRP (SCSI RDMA Protocol)
- Storage and Networking functionality
- Subnet Management functionality
- Sourced from OpenFabrics Alliance (www.openfabrics.org)
- Uses latest 2.6.xx kernel API
The Challenges

- ESX Linux API is based on a 2.4 Linux Kernel
  - Not all the 2.4 APIs are implemented
  - Some 2.4 APIs are slightly different in ESX
  - Different memory management
  - New build environment
- Proprietary management for networking and storage
Enhancements or Optimizations

- ESX kernel changes
  - Common spinlock implementation for network and storage drivers
  - Enhancement to VMkernel loader to export Linux-like symbol mechanism
  - New API for network driver to access internal VSwitch data
  - SCSI command with multiple scatter list of 512-byte aligned buffer

- InfiniBand driver changes
  - Abstraction layer to map Linux 2.6 APIs to Linux 2.4 APIs
  - Module heap mechanism to support shared memory between InfiniBand modules
  - Use of new API by network driver for seamless VMotion support
  - Limit one SCSI host and net device per PCI function

Effective collaboration to create compelling solution
InfiniBand with Virtual Infrastructure 3

VM-0
NIC     HBA
InfiniBand Network
Driver (IPoIB)

VM-2
NIC     HBA
InfiniBand Storage
Driver (SRP)

VM-3
NIC     HBA

Virtual Center
NIC     HBA

Console
OS

Network Virtualization
(V-Switch)

SCSI/FS
Virtualization

Hypervisor

InfiniBand Network
Driver (IPoIB)

InfiniBand Storage
Driver (SRP)

IB HCA

Transparent to VMs and Virtual Center
VM Transparent Server I/O Scaling & Consolidation

3X networking, 10X SAN performance

Per adapter performance. Based on comparisons with GigE and 2 Gb/s Fibre Channel
Using Virtual Center Seamlessly with InfiniBand

Storage configuration

vmhba2
Using Virtual Center Seamlessly with InfiniBand

Storage configuration

vmhba2
Using Virtual Center Seamlessly with InfiniBand

Network configuration
vmnic2 (shows as vmhba2:bug)
vSwitch1
Performance Testing Configuration

- **ESX Server:**
  - 2 Intel Dual Core Woodcrest CPUs
  - 4GB Memory
  - InfiniBand 20Gb/s HCA

- **Switch:**
  - Flextronics 20 Gb/s 24 port Switch
  - Native InfiniBand Storage (SAN)
  - 2 Mellanox MTD1000 Targets (Reference Design)
  - [www.mellanox.com](http://www.mellanox.com)
Performance Testing Results Sample - Storage

128KB Read benchmark from one VM

128KB Read benchmarks from two VMs
More Compelling Results

128KB Read benchmarks from four VMs

Same as four dedicated 4Gb/s FC HBAs
Tested Hardware

- Mellanox 10 and 20 Gb/s InfiniBand Adapters
- Cisco and Voltaire InfiniBand Switches
- Cisco Fibre Channel Gateways with EMC back-end storage
- LSI Native InfiniBand Storage Target
Compelling End User Benefits

- **Lower Initial Purchase Cost**
  - Up to 40% savings

- **Lower Per-port Maintenance Cost**
  - Up to 67% savings

- **Lower I/O Power Consumption**
  - Up to 44% savings

- **Seamless**
  - Transparent interfaces to VM apps and Virtual Center Management

Based on $150/port maintenance cost (source: VMware, IDC), end user per port cost (adapter port + cable + switch port) comparisons between 20Gb/s IB HCA ($600/port), GigE NIC ($150/port) and 2Gb/s FC HBA ($725/port) (Source: IDC)

Typical VMware virtual server configuration (source: VMware), 2W power per GigE port, 5W power per FC port, 4-6W power per 20Gb/s IB port (source: Mellanox)

Best of both worlds – seamless + cost/power savings
Compelling End User Benefits (contd.)

- **VI 3 Component Acceleration**
  - Backup, Recovery, Cloning, Virtual Appliance deployment

- **VM Application Acceleration**
  - Database, file system and other storage intensive applications

*Benchmark data not available yet*

**Best of both worlds – seamless + I/O scaling**
Experimental Release Plans

- VMware Virtual Infrastructure (incl. ESX) Experimental Release
- Mellanox InfiniBand drivers installation package
- Targeted for late Q1 2007
- For further details, contact:
  - Junaid Qurashi, Product Management, VMware, jqurashi@vmware.com
  - Sujal Das, Product Management, Mellanox, sujal@mellanox.com
Future Work

- Evaluate experimental release feedback
- GA product plans
- Feature enhancements
  - Based on customer feedback
  - VMotion acceleration over RDMA
  - Networking performance improvements
Call to Action: Evaluate Experimental Release

- InfiniBand: for high performance, reliable server & storage connectivity
- Tier 1 OEM channels and global presence

- Multi-core & virtualization driving 10Gb/s connectivity in the data center
- I/O convergence & IT objectives driving stringent I/O requirements
- InfiniBand delivers “Service Oriented I/O” with service guarantees

- Pioneering role in community source program
- VI 3 with InfiniBand – seamless IT experience with I/O scale-out
- Compelling cost, power and performance benefits

- **Lower Initial Purchase Cost**
  - Up to 40% savings

- **Lower Per-port Maintenance Cost**
  - Up to 67% savings

- **Lower I/O Power Consumption**
  - Up to 44% savings
Please remember to complete your **session evaluation form** and return it to the room monitors as you exit the session.

The presentation for this session can be downloaded at [http://www.vmware.com/vmtn/vmworld/sessions/](http://www.vmware.com/vmtn/vmworld/sessions/)

Enter the following to download (case-sensitive):

- **Username:** cbv_rep
- **Password:** cbvfor9v9r