Voltaire
The Grid Interconnect Company

Fast I/O for XEN using RDMA Technologies

April 2005
Yaron Haviv, Voltaire, CTO
yaronh@voltaire.com
VMs need to interact efficiently with the rest of the Virtual/Grid world to be successful

Agenda: Ways for I/O virtualization with minimal performance hit
The I/O Virtualization Model

Virtual I/O enable Virtual Environments, and complement XEN

Physical Topology

Management

Servers & Blades

InfiniBand

GbE

Remote Clients and Networks

Multi-Service Switch

FC, iSCSI or RDMA

Logical Topology

Virtual Network

Virtual Servers

Virtual Storage

SAN

GBE

RAID’s

Virtual Routers

Remote Clients and Networks

InfiniBand

GbE
XEN Over InfiniBand Benefits

- Slower I/O
- No isolation
- Multiple cards and fabrics

- Direct HW access for I/O
- One 10Gb/s card for Network, Storage, and IPC
- Scale out using clustering (IPC)

**Better Performance, Isolation, and TCO with InfiniBand**
The problems:
- Context Switches
- Same Receive/Send Queue (I/O Dependencies and starvation)
- Configuration/Security challenges (Multiple VLANs, routing, etc’)

The Solution:
- Dedicate Send/Receive Queue per Virtual NIC
- Bypass Hypervisor and VM0 by using Memory Mapped I/O (separate Mem/IO Pages per VM)
- Packet switching and QoS in hardware
- And, still make sure the VMs are not bound to a specific Hardware
XEN Networking With Direct HW Access

- Slow path done in VM0
- Allow direct access (DA) from Guest to data path operations (with safe DMA through mem registration)
- Can have a generic API (e.g. Datagram Send/Receive), loaded and used by the Ethernet front-end driver
- Direct access code need to be exported from VM0 to guests (like PXE/UNDI)
- Can be used by different hardware/NIC providers (not just for IB), and allow keeping an HW neutral guest
Datamover Architecture (DA) and iSER are optimal for XEN
- One driver and management, support for many different cases
- Centrally managed and provisioned through iSCSI infrastructure (iSNS/SLP)

- Same architecture apply to NICs, RNICs (Ethernet/RDMA). InfiniBand
- Can migrate a virtual machine from IB to Ethernet and keep on working
Automatic discovery of storage
- Automatic selection of Transport (iSER or TCP)
- Same standard management across the entire solution
  - Including IP, IB, and FC storage
XEN Block Storage With Direct Access

- Similar design to DA Networking
- Need definition for direct access API (e.g. send/receive sector)
- Storage front-end driver can detect and use direct access API
- In InfiniBand will map to separate QPs and CQs per VM, same model can apply to other HW
- Migration handled by iSCSI recovery mechanisms (rediscover and connect)
Using XEN For Scale Out

- Low latency and zero copy IPC is critical to clustering
- Need new Front-End drivers for IPC
  - Socket/TOE abstraction (e.g. a BSD Scok_stream), can be used between VMs
  - RDMA abstraction (e.g. DAPL/IT)
IPC and Direct User Access to HW

- Recommend mapping to an industry standard such as DAPL
  - Allow existing RDMA clients such as MPI, Oracle, NFS/RDMA, iSER to work transparently
  - Mapped rather well to IB & Ethernet/RDMA

- Open issues around migration, will initially require application awareness (rediscover and reconnect on failure)

- Open XEN to new real-time, HPC, Media, and transactional markets
InfiniBand and RDMA can solve the I/O problem TODAY
- Dedicated I/O per VM
- Direct access from VM to Hardware through channel architecture
- Protected DMA
- Same card can provide virtual network, storage and IPC (and provide the best performance in each category)

Can build a single infrastructure in XEN for:
- Plain Ethernet, InfiniBand and Ethernet/RDMA
- Allow migration between heterogeneous I/O hardware

Welcome people to participate and in an open source environment
Call for Action

- Establish a direct access I/O working group
  - Open discussion on architecture and implementation alternatives
  - Looking for volunteers to help code missing pieces
- Incorporate required changes in OpenIB (in Linux Kernel)
- Incorporate work into distros (SuSe, RH, ..)
- Work with CPU vendors to optimize memory and interrupt sharing mechanisms