OPENFABRICS ALLIANCE

Scaling out the datacenter

Author: Wim Coekaerts

Date: 4/5/2011

Infiniband in IT

- still fairly unknown yet it has been around
  - it is an HPC thing
  - super computers, labs, R&D
  - assumption IB is not for more standard server deployments in a datacenter
  - it actually is different - bus vs network
- most of company IT built on ethernet
  - lack of knowledge of IB in general
  - reluctance to deploy a new infrastructure
  - considered difficult and costly to maintain both

Infiniband in IT

- how to make our database faster
- Oracle database had a need to scale up/out
  - database cluster nodes
    - Distributed Lock Manager
    - move database blocks from server to server zero-copy/rdma
  - create a special communications protocol in IB : RDS
    - very low latency, high bandwidth, very low overhead
  - now the ability to create database clusters with IB/ RDS
- address storage smarts
  - move from standard storage (SAN/NFS/iSCSI) files or devices to smart storage (storage cells)

Our definition of HPC

- Engineered system : EXADATA
  - complete system
  - hardware
  - software
  - infiniband
- Blazingly fast
- one unit pre-wired
  - ethernet out
  - IT doesn't see IB
- Add in datacenter

Our definition of HPC

- Added sparc supercluster and exalogic
  - make sure systems can co-exist in datacenters
  - IB integrated into the rack(s) and management
  - plug exadata racks together to create a larger system
  - plug an exalogic system into an exadata system with IB directly
- the ethernet network stops at the rack(s)
  - less resistance from the system admins
  - allows us to introduce IB much more easily as it is mostly hidden or integrated

Concerns and focus areas for us

- integrate with virtualization solutions (and cloud)
  - for the most part IB gives us a very fixed bare metal server environments
  - we need to easily move virtual machines around yet keep performance for network and storage inside the VM
  - SRIOV appears great for performance but no flexibility for migration between servers - not transparent to the VMs and the environment
  - EoIB / paravirt IB / pass QPs
  - do IB in hypervisor/lower stack and expose virtual ethernet and virtual disks

Concerns and focus areas for us

- easy integration/compatibility with ethernet in datacenters (IPoIB, EoIB)
- interoperability between the IB vendors is critical
  - IB stack in OS
  - IB cards, switches
- configuration of an IB setup is very complex
  - make management easier
  - look like network

Concerns and focus areas for us

- get ofed fully upstream in Linux
  - it’s great to have a stack that builds on so many versions but there’s a lot of luggage to carry forward
- stability
  - we break all OS environments - also the ofed stack
  - need to see how we can do better QA for ofed releases as it’s related to our work
  - we spent a lot of time making RDS scale/stable but protocols like SDP are not at that level
- NUMA performance
  - every server even a 2 socket is now a NUMA system
  - scaling all this on an 8 socket is a real challenge