OPENFABRICS ALLIANCE

Interesting Technology Trends (as seen by Red Hat)

11TH ANNUAL

INTERNATIONAL
OPENFABRICS SOFTWARE
DEVELOPERS' WORKSHOP

#OFADevWorkshop

OFIWG and libfabric

- Active, multivendor development environment
- Responsive maintainer
- Designed to be fabric agnostic, unlike libibverbs where different fabrics are hacked in under the skin
- Likely to be “the future” of RDMA support in our products
- libibverbs will still be available, both as a provider to libfabric as well as standalone for maximum performance

Storage technologies

- iSER (Mellanox) / SRP (Bart Van Assche) using LIO kernel target infrastructure
- NFSoRDMA (Oracle)
- Targeting 2u class appliance/20+ disks (we have other products for larger scale file serving, and those products either have RDMA support or are currently working on RDMA support)
- Hardware tested/supported for above items
  - mlx4 - IB/RoCE
  - mlx5 - IB (RoCE when available)
  - Intel - IB (OPA when available)
  - Chelsio - iWARP
  - Emulex - RoCE
  - mthca - Dead. Once it breaks, we will turn it off and not work on it any further

RDMA and Virtualization

- RDMA enabled migration via qemu/libvirt is available today
  - Does not pass RDMA capability to guest
  - Only speeds up migration, does not speed up normal operation
- RDMA passthrough to guest via SRIOV is being integrated
  - Requires changes to kernel/libvirt to enable setting MAC/GUID address of guest device
  - We have some scripts in place to work around this for now, but need to get an official mechanism upstream
  - Migration of guests using RDMA devices can not yet be done live. Either support in the RDMA device itself for live migration, or hot plug support in the application for RDMA devices would need to be in place first. See Liran’s virtualization talk for more details on this.

SRIOV

- Only mlx4 at the moment
- Extremely performant
  - Test platform is 4 socket, 6 core, HT enabled 2.4GHz for a total of 48 cores according to kernel, all PCI busses are attached to CPU sockets 0 and 1, 2 and 3 are memory/compute only nodes, not I/O nodes
  - Guests were defined as 24 core VMs and were pinned to sockets 2 and 3
  - Tests were with PCIe Gen 3 x8 card at 56GBit/s

SRIOV performance numbers

	Host <-> Guest	*Remote host <-> Guest*	*Guest <-> Guest*
TCP BW	*2.8GByte/s 325% CPU Util*	*1.2GByte/s 110% CPU Util*	*2.55GByte/s 150% CPU Util*
UDP BW	*9.8GByte/s 400% CPU Util*	*1.75GByte/s 90% CPU Util*	*3.9GByte/s 140% CPU Util*
RC BW	*5.4GByte/s * 160% CPU Util*	*5.46GByte/s * 90% CPU Util*	*4.25GByte/s 125% CPU Util*
RC Bi-BW	*5.4GByte/s * 190% CPU Util*	*10.3GByte/s 140% CPU Util*	*5.4GByte/s * 160% CPU Util*

*** - PCIe bus limited**

Thank You

OPENFABRICS ALLIANCE

11TH ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP