Datacenter Fabric Workshop

OpenIB OpenSM

Hal Rosenstock, Voltaire

Eitan Zahavi, Mellanox

Agenda

- Current Status
- OpenSM 1.8.0
  - New Features
  - Major Bug Fixes
- Future Directions

Current Status of OpenIB OpenSM

- Primarily based on Mellanox Gold 1.6.1/1.7.0
  - OpenIB vendor layer
  - Autotools build
- Some feature additions for Solaris 10 and 11 interoperability
    - PathRecord support
    - MCMemberRecord support
- Some selected bug fixes and feature enhancements from subsequent Mellanox Gold OpenSMs
  - Partial DDR support (LinkSpeedActive)
  - Only update MFTs if changed

Current Status of OpenIB OpenSM

- Some other bug fixes
  - Mostly involving MCMemberRecord
  - Handover/failover changes
    - SM GUID comparison endian issue
    - Default polling interval
- Upgrade to include Mellanox Gold 1.8.0 in process
  - Also, osmtest being ported
  - Mellanox is leading this effort

New Features of OpenSM 1.8.0

- Semistatic LID assignment
  - No LID change on SM restart or node reboot
  - Critical for IPoIB to avoid communication loss
- Irresponsive port scan during light sweep
  - No response but Link state not down
- Switch ports with HCA neighbor have lower HOQLife
  - Faster drain so bad HCA not impact subnet
- Pkeys
  - Not reordered
  - Default values not set
- DDR and QDR support
- Options Cache
  - including all non command line
  - Use –c flag to create /var/cache/osm/opensm.opts
- Kill –HUP
  - Forces a new full sweep

Major Bug Fixes of OpenSM 1.8.0

- Overflow on SA queries (now drops them if overflow)
- Multicast tree build took forever on large clusters
- MTU and Rate selectors ignored during MCMemberRecord Query
- Deleted multicast groups existing until deferred deletion
- Crashed on any zero Port or Node GUID
- SMInfo with a non default PKey was dropped
- DDR and QDR rates were not calculated correctly
- Fail to error Service Record delete of non-existing record
- Memory leak in SA Client code
- Multicast Join did not check for ‘JoinState != 0’
- PortInfo SA query fail if base_lid component used
- OpenSM runs out of MLIDs even though some groups were deleted
- Complib race in Passive Lock caused a deadlock (now use rwlock)

Many more less severe bugs fixed

Future Directions

- Mellanox
  - Partition support – Q3
  - Initialization time – Q3
  - Regression tests and automation – Q3
  - Simulator – Q3
  - QoS – Q4
- SA MultiPathRecord support
- Pathing algorithms
- Advanced failover

Datacenter Fabric Workshop

Thank You