2006 Sonoma Workshop

OpenIB OpenSM and Diagnostics

Hal Rosenstock, Voltaire

Eitan Zahavi, Mellanox

Agenda

- OpenSM
  - New Features
  - Major Bug Fixes
  - Verification
  - Future Directions
- Diagnostics
  - Status
  - Plans

OpenSM Current Status

- Effort mostly in porting and stabilization
- A number of bugs found and fixed
- Ongoing effort in increasing test coverage
- New Hardware Supported
  - PathScale InfiniPath
  - IBM Galaxy
  - Obsidian Longbow
- OpenSM used at SC05

OpenSM New Features

- GUIDInfoRecord
  - SA support for that query
- Pkeys
  - Trivial P_Key manager enforces default P_Key to exist on all end ports
- Semi-static 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
- SM can exit on fatal network conditions
  - New option that exits on duplicated GUIDs and other fatal conditions
- Options Cache
  - including all non command line
  - Use –c flag to create /var/cache/osm/opensm.opts
- Kill –HUP
  - Forces a new full sweep
- SVN revision of OpenSM indicated
- Primitive console with basic commands

OpenSM Default Changes

- Transaction timeout increased from
  - 100 to 200 milliseconds
- MAXSMPS increased from 1 to 4
- SM priority changed from 0 to 1
- SM key changed from 0 to 1
- All are settable via command line
- Some are settable via interactive console

OpenSM Bug Fixes

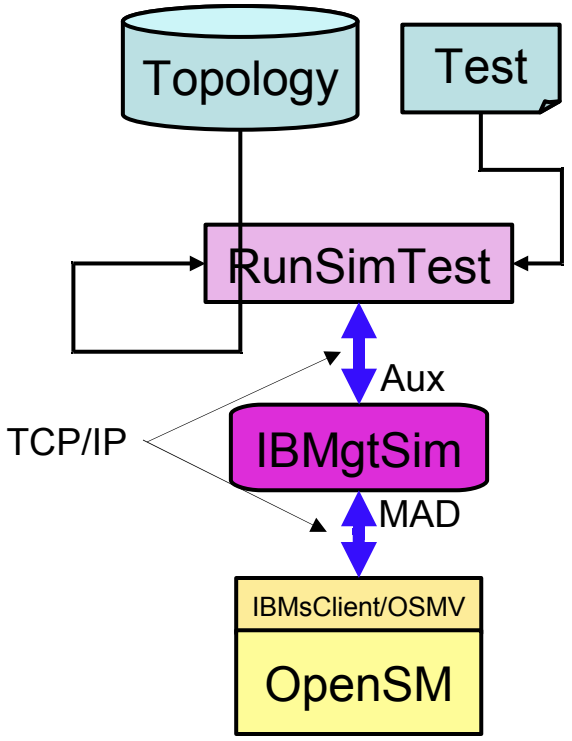
- Fixed a deadlock due to out of order MADs received and bad lock ordering
- Fixed memory overflow found by valgrind
- Fixed complib timer issue when events complete beyond expiration time, which caused consumption of 100% of CPU due to an infinite loop
- Fixed complib timer destruction sequence, which caused crash of OpenSM during exit flow, when ran with “-o”
- Fixed race with transaction matching in vendor library
- Fixed vendor library receiver exit when large umad allocation fails
- OpenSM didn’t complete sweep if driver failed to send a MAD
- OpenSM sometimes hangs during LID assignment phase. Fixed the LID assignment algorithm to support these cases (related to LMC > 0)
- Fixed updating of counters on the number of outstanding mads
- Fixed downing of port to be on NeighborMTU rather than MTUCap change
- Send TrapRepress only if SM is in Master State
- Add support for compliancy statement C14-62.1.1: make sure each node is updated with either 0xFFFF or 0x7FFF pkey in its pkey table
- Fixed PortInfo Record query matching on several fields
- Fixed PathRecord query matching on several fields
- Fixed LinkRecord query matching on several fields
- Fixed MCMemberRecord component mask query issue

OpenSM Verification Flows

- Osmtest – stand alone test application
  - Test — Strength 1-10
    - Multicast — 9
    - Event forwarding — 1
    - Stress testing of RMPP — 9
    - Service Records — 7
    - Comparing Setup to a “gold” one — 9
  - Topologies – real hardware
    - Back to back
    - 2 nodes with a switch
    - > 2 nodes with > 1 switch

OpenSM Verification Flows

- ibmgtsim – IB management simulator
  - Tests — Strength 1-10
    - Stability check 5
    - LID assignment 9
    - LFT Routing (credit loop free) 7
    - Multicast Routing 8
    - P_Key support for SA 8
  - Topologies – simulated fabric
    - 16 / 128 / 512 nodes – fat tree
    - Credit loop partial fat tree



OpenSM Verification Flows

- Tests

	Strength 1-10
• Random CompMask	8
• Stress Queries (SPR)	7
• Trap Gen	7
• Exit Flow	9
• Irresponsive Port	9

  - Topologies – real fabric
    - Back to back
    - 2 nodes with a switch
    - > 2 nodes with > 1 switch

OpenSM Future Directions

- Near Term (Q1)
  - Partition support
  - QoS
  - SA MultiPathRecord support
  - More console support
  - Regression tests and automation
    - Handover
    - NodeInfo, PortInfo, SwitchInfo rule checker
    - Event Forwarding

OpenSM Future Directions

- Longer Term
  - More IBA 1.2 support and 1.2 errata
  - Congestion management
  - Initialization time
  - Pathing algorithms
  - Advanced failover
  - More QoS
  - SNMP MIB support

Partition Support

- Partition configuration
  - Partition name
  - 15 bits of PKey
  - IPoIB broadcast group
  - List of PortGUIDs and whether full or limited member
- When appropriate, OpenSM will update endport Pkey tables as well as those in the adjacent switch leaf ports
- More complete writeup in https://openib.org/svn/gen2/trunk/src/userspace/r

QoS Phase 1 Support

- "Uniform" SL to VL mapping and VLArbitration
  - Separate for switch external ports, switch enhanced port 0, and xCA ports
  - Handle ports which don't support this even if configured
- PathRecord to return SL and path bits
  - By {Source, Destination, QoS key}
    - Current LWG approach to be finalized
    - QoS key could be DSCP byte
  - Partition based (additional partition policy)

Diagnostic Utilities

- management/diags
  - A set of low level utilities (both programs and scripts)
  - Based on management libraries (ibmad, ibumad, ibcommon)
  - Full control and accessibility into the fabric
  - Includes:
    - ibaddr, ibnetdiscover, ibping, ibportstate, ibroute, ibstat, ibsysstat, ibtracert, perfquery, sminfo, smpdump, smpquery
    - Various scripts based on the above
- gen2/utils (https://openib.org/svn/gen2/utils/src/linux-user)
  - Highly integrated diagnostic utilities
  - Based on:
    - OpenSM vendor layer
    - TCL interface for MAD send/receive
  - Includes: ibdiagnet, ibdiagpath

Recent Diagnostics Additions

- Based on SC05 experience in larger subnet
- ibportstate – query, disable, and enable port state/port physical state of an IB switch port
- Additional scripts
  - ibcheckstate: check for non active ports
  - ibcheckwidth: find 1x ports
  - ibcheckerrors: find ports with errors above thresholds
  - discover.pl: use ibnetdiscover topology file and map file of expected nodes to update a previous connectivity file into a current connectivity file (with some annotation)

Upcoming Diagnostics

- Chassis based switches
  - Grouping
  - Logical to Physical Mapping

Diags - ibdiagnet

- Perform a full fabric “rule check”
  - HCA to HCA connectivity (LFT)
  - Multicast groups validity
  - Duplicated LIDs, GUID
  - SMs are alive and only one master
  - Performance counters and packet drops
  - Faulty link statistical isolation
- Optionally given topology definition enables
  - Checks Matching (cabling, width and speed)
  - Used for reporting

Diags – Topology Definition

```
AB16 SW1
        P1 -> XY288 SW2 S2/P1
        P2 -> HCA host2 P1
        P3 -> HCA host1 P2

XY288 SW2 CFG:S4=12X
        S4/P1 -12x->      HCA host3 P1
        S1/P1 -4x-5G->    HCA host4 P1
```

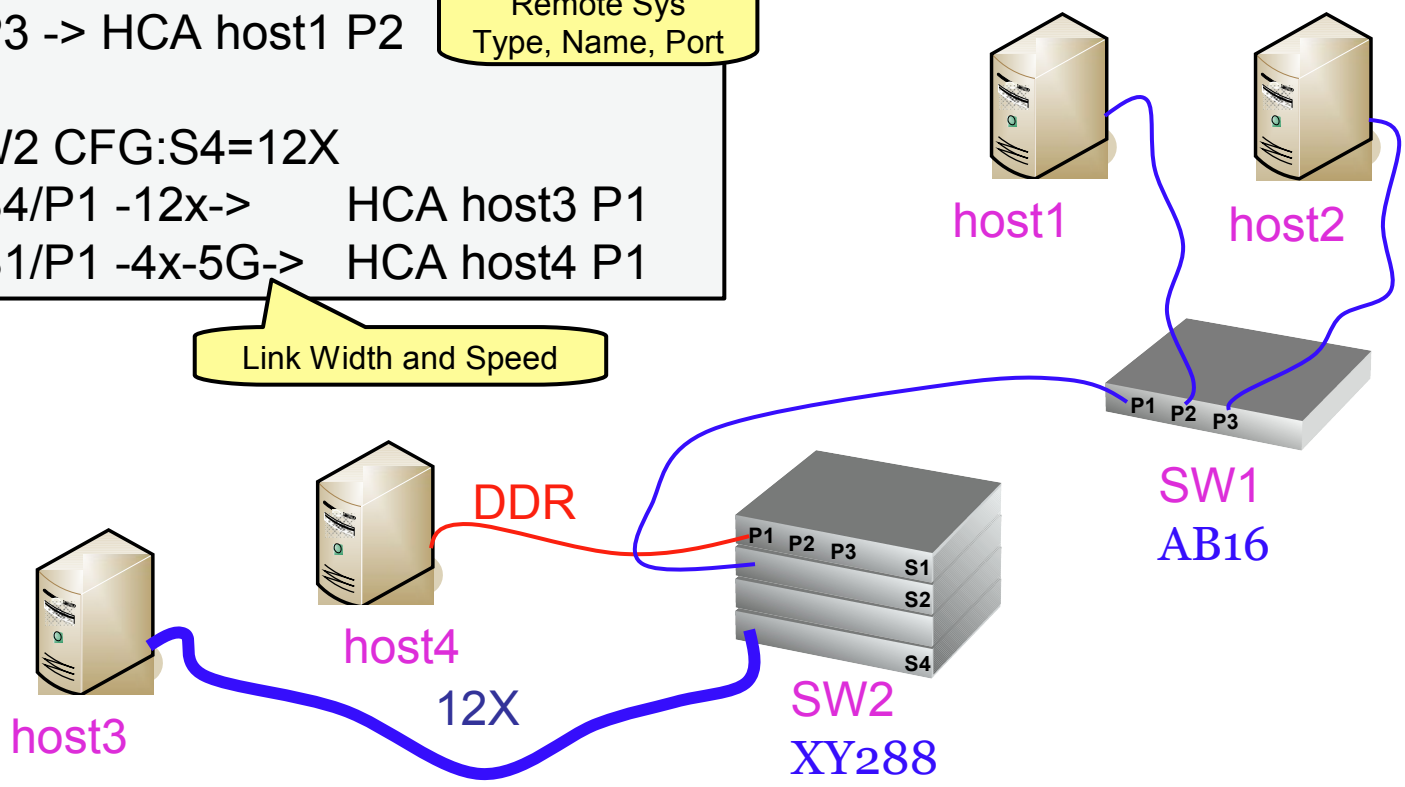
Sys Name

Sys Type

Remote Sys Type, Name, Port

Sys Port Name

Link Width and Speed



Diags - ibdiagnet

- Faulty links

-I- Errors have occurred on the following links
(for errors details, look in log file /tmp/ibdiagnet.log):
Cable: SW1/M/P7(SW1/main/U4/P4) =---= H-7/P2(H-7/U1/P2)
Cable: SW1/M/P5(SW1/main/U4/P6) =---= H-5/P2(H-5/U1/P2)

- Duplicated or Zero LIDs

-E- Device(s) with LID = 0x0000 found in the fabric:
path="1 1 3 5" H-12/U1 PN=2
path="1 1 3 4" H-11/U1 PN=1
path="1 4" H-3/U1 PN=1

- Multiple SMs

SM – master
mtlmd11/P1 priority:15
SM – standby
The Local Device : swcl40/P1 priority:4
swlab178/P1 priority:1

Diags - ibdiagnet

- Mismatched Topology

Missing System:H-7(Cougar)
    Should be connected by cable from port:P2(H-7/U1/P2) to:SW1/M/P7(SW1/U4/P4)

- Multicast Connectivity

-I- Multicast Group:0xC000 has:2 switches and:2 HCAs

-E- Extra switch:SW1/leaf1/U1 in group:0xC000

-E- Extra switch:SW1/main/U4 in group:0xC000

- Credit Loops – violating Up/Down

-E- Potential Credit Loop on Path from:H-1/U1/1 to:H-13/U1/1
    Going:Down from:SW1/main/U1 to:SW1/main/U3
    Going:Up from:SW1/main/U3 to:SW1/main/U1
    Going:Down from:SW1/main/U1 to:SW1/leaf1/U1

Diags - ibdiagpath

- Diagnose a path
- Several Source Destination modes:
  - Directed Routes
  - LIDs
  - Names
- Optional Topology file enables:
  - Matching check
  - Report using topology names

Diags - Future

- ULP level checking
  - IPoIB, MPI, SDP
  - Plug-able benchmark
  - BW and latency report
- Cluster Verification Procedure
  - ibdiagnet
  - Monitoring
  - OS, driver and Firmware versions

2006 Sonoma Workshop

Thank You