OPENFABRICS ALLIANCE

OpenSM Logging

Hal Rosenstock
Mellanox Technologies

Agenda

- OpenSM Update
- Per Module Logging Feature
- Log File Walkthru

OpenSM “Notable” Releases

- Releases nominally every 6-9 months
  - Independent of OFED
- FDR and FDR-10 support (OpenSM 3.3.11 – Aug 2011)
  - FDR (and EDR) are IBTA standards
  - FDR-10 is MLNX proprietary
- SRIOV support (OpenSM 3.3.14 – May 2012)
  - Additional GUIDs for virtual machines
    - Bug fixes beyond 3.3.14

Quick OpenSM Update

- Last release 3.3.17 – Feb 2014
  - Also included in OFED 3.12 which is now in process (@ RC1)
  - Previous release 3.3.16 – Feb 2013
- Mainly bug fixes beyond that but no new features so far
  - ~40 commits past 3.3.17 right now
- Regression tests being run against latest master

OpenSM Upcoming Features

- Event reporting scalability
- Bad hardware
- Heldback switches
- Multicast improvements
- Routing chains
- Credit-loop free UC and MC routing for UPDN/ FTREE
- Multithreaded updn/minhop/dor
- QFT

OpenSM Logging Related Command Line Options

- -f, --log_file <file name>
  This option defines the log to be the given file. By default, the log goes to /var/log/opensm.log. For the log to go to standard output use -f stdout.
- -L, --log_limit <size in MB>
  This option defines maximal log file size in MB. When specified the log file will be truncated upon reaching this limit.
- -e, --erase_log_file
  This option will cause deletion of the log file (if it previously exists). By default, the log file is accumulative.

OpenSM Logging Related Config File Options


```
# Force flush of the log file after each log message
force_log_flush FALSE

# Log file to be used
log_file /var/log/opensm.log

# Limit the size of the log file in MB. If overrun, log is restarted
log_max_size 0

# If TRUE will accumulate the log over multiple OpenSM sessions
accum_log_file TRUE

# Per module logging configuration file
# Each line in config file contains <module_name><separator><log_flags>
# where module_name is file name including .c
# separator is either = , space, or tab
# log_flags is the same flags as used in the coarse/overall logging
per_module_logging_file /usr/local/etc/opensm/per-module-logging.conf
```

OpenSM Log Levels

- Overall log verbosity level
  - log_flags config file option
  - Related command line options
    - -D <value>
    - -V
    - -v, --verbose
    - -d, --debug <value>

-D <value>

- This option sets the log verbosity level. A flags field must follow the -D option. A bit set/clear in the flags enables/disables a specific log level as follows:

BIT	LOG LEVEL ENABLED
0x01	ERROR (error messages)
0x02	INFO (basic messages, low volume)
0x04	VERBOSE (interesting stuff, moderate volume)
0x08	DEBUG (diagnostic, high volume)
0x10	FUNCS (function entry/exit, very high volume)
0x20	FRAMES (dumps all SMP and GMP frames)
0x40	ROUTING (dump FDB routing information)
0x80	SYS (syslog at LOG_INFO level in addition to OpenSM logging)

-D <value>

- Without -D, OpenSM defaults to ERROR + INFO (0x3).
- Specifying -D 0 disables all messages.
- Specifying -D 0xFF enables all messages (see -V). High verbosity levels may require increasing the transaction timeout with the -t option.

Other Related Log Level Command Line Options

- -v, --verbose
  This option increases the log verbosity level. The -v option may be specified multiple times to further increase the verbosity level. See the -D option for more information about log verbosity.
- -V
  This option sets the maximum verbosity level and forces log flushing. The -V option is equivalent to -D 0xFF -d 2. See the -D option for more information about log verbosity.
- -d, --debug <value>
  -d2 - Force log flushing after each log message

Issues with OpenSM Logging

- Coarseness of log level
  - One level for all of OpenSM
  - Too many log messages as increase verbosity/log level
- Somewhat “cryptic” nature of messages logged

Per Module Logging (PML)

- Log level per “module”
  - Module is a source code file
- Introduced so can keep “overall” level low but dial up level in specific modules/files
  - Need to have idea of which modules/files to dial up
- PML can change on the “fly” with SIGHUP
- Added to upstream master git tree in June/July 2012
- Part of OpenSM 3.3.15 and beyond

Per Module Logging (PML)

- Enable via per_module_logging_file option in options file – set to PML config file name
  - Disable by setting per_module_logging_file to (null) in options file
- Per module logging config file format

  Set of lines with module name and logging level as follows:
  <module name><separator><logging level>
  where:
  <module name> is the file name including .c
  <separator> is either = , space, or tab
  <logging level> is the same levels as used in the coarse/overall logging
- Module names may vary between releases
  - 3.3.16 and beyond have all modules listed
  - 3.3.15 has one less module (no osm_congestion_control.c)

Per Module Logging (PML) Module Names Based on Latest Upstream Master

From opensm/osm_subnet.c:

```
static const char *module_name_str[] = {
        "main.c",
        "osm_console.c",
        "osm_console_io.c",
        "osm_db_files.c",
        "osm_db_pack.c",
        "osm_drop_mgr.c",
        "osm_dump.c",
        "osm_event_plugin.c",
        "osm_guid_info_rcv.c",
        "osm_guid_mgr.c",
        "osm_helper.c",
        "osm_inform.c",
        "osm_lid_mgr.c",
        "osm_lin_fwd_rcv.c",
        "osm_link_mgr.c",
        "osm_log.c",
        "osm_mad_pool.c",
        "osm_mcast_fwd_rcv.c",
        "osm_mcast_mgr.c",
        "osm_mcast_tbl.c",
        "osm_mcm_port.c",
```

Per Module Logging (PML) Module Names Based on Latest Upstream Master

"osm_mesh.c",
"osm_mlnx_ext_port_info_rcv.c",
"osm_mtree.c",
"osm_multicast.c",
"osm_node.c",
"osm_node_desc_rcv.c",
"osm_node_info_rcv.c",
"osm_opensm.c",
"osm_perfmgr.c",
"osm_perfmgr_db.c",
"osm_pkey.c",
"osm_pkey_mgr.c",
"osm_pkey_rcv.c",
"osm_port.c",
"osm_port_info_rcv.c",
"osm_prtn.c",
"osm_prtn_config.c",
"osm_qos.c",
"osm_qos_parser_l.c",
"osm_qos_parser_y.c",
"osm_qos_policy.c",
"osm_remote_sm.c",
"osm_req.c",
"osm_resp.c",
"osm_router.c",

Per Module Logging (PML) Module Names Based on Latest Upstream Master

"osm_sa.c",
"osm_sa_class_port_info.c",
"osm_sa_guidinfo_record.c",
"osm_sa_informinfo.c",
"osm_sa_lft_record.c",
"osm_sa_link_record.c",
"osm_sa_mad_ctrl.c",
"osm_sa_mcmember_record.c",
"osm_sa_mft_record.c",
"osm_sa_multipath_record.c",
"osm_sa_node_record.c",
"osm_sa_path_record.c",
"osm_sa_pkey_record.c",
"osm_sa_portinfo_record.c",
"osm_sa_service_record.c",
"osm_sa_slvl_record.c",
"osm_sa_sminfo_record.c",
"osm_sa_sw_info_record.c",
"osm_sa_vlarb_record.c",
"osm_service.c",
"osm_slvl_map_rcv.c",
"osm_sm.c",
"osm_sminfo_rcv.c",
"osm_sm_mad_ctrl.c",
"osm_sm_state_mgr.c",
"osm_state_mgr.c",

Per Module Logging (PML) Module Names Based on Latest Upstream Master


```
        "osm_subnet.c",
        "osm_sw_info_rcv.c",
        "osm_switch.c",
        "osm_torus.c",
        "osm_trap_rcv.c",
        "osm_ucast_cache.c",
        "osm_ucast_dnup.c",
        "osm_ucast_file.c",
        "osm_ucast_ftree.c",
        "osm_ucast_lash.c",
        "osm_ucast_mgr.c",
        "osm_ucast_updn.c",
        "osm_vendor_ibumad.c",
        "osm_vl15intf.c",
        "osm_vl_arb_rcv.c",
        "st.c",
        "osm_ucast_dfsssp.c",
        "osm_congestion_control.c",
        /* Add new module names here ... */
        /* FILE_ID define in those modules must be identical to index here */
        /* last FILE_ID is currently 89 */
};
```

OpenSM Log Messages Overview

- Format: date time [thread ID] log level
  Feb 19 12:40:45 897693 [91A48700] 0x01 ->
- ERR number if in message is unique
- Having OpenSM sources helps
  - Tracking error number in source module shows where generated and can read code and comments
- IBA spec knowledge is helpful
  - Primarily volume 1 IB management related chapters

OpenSM Log File Walkthru

- pi_rcv_check_and_fix_lid: ERR 0F04: Got invalid base LID 65535 from the network. Corrected to 0
  - SM queried for PortInfo for some end port and received base LID 0xffff
  - IBA spec is mute on what LID to use when port is not yet configured by SM
    - Some SMAs use 0 and other use 0xffff
  - This really shouldn't be "error"
    - Log message is really debug info

OpenSM Log File Walkthru

- subn_validate_neighbor: ERR 7518: neighbor does not point back at us (guid: 0x0002c902002a0669, port 1)
- subn_validate_neighbor: ERR 7518: neighbor does not point back at us (guid: 0x0005ad0007042dcf, port 4)
  - /var/cache/neighbors introduced by mkey support
  - Peer port GUID and port number
    - 0x0008f105006002d4:9 0x0008f105007002fe:31
  - Entries are paired
    - 0x0008f105007002fe:31 0x0008f105006002d4:9
  - Message indicates that the reverse entry does not match the forward one
    - Benign error
    - Probably due to some offline topology change or subnet instability
    - Need to investigate further

OpenSM Log File Walkthru

- osm_get_port_by_mad_addr: ERR 7504: Lid is out of range: 860
  - LID requested by some lookup is not currently known by SM
- osm_pr_rcv_process: ERR 1F16: Cannot find requester physical port
  - Port that requested SA PathRecord is not currently known by SM
  - Causes query to timeout at end port stack as SM has no way to respond
- Both errors above are indicative of queries during “changing” subnet
  - Should be benign as long as end port stack retries

OpenSM Log File Walkthru

- SA PathRecord Query Handling
  - End port stack (SA client) issues SA PathRecord query
  - SM walks path from end to end before returning response
  - Various errors on path walk
    - Should be transient due to subnet “changes”
  - SM either returns SA ERROR_NO_RECORDS if SubAdmGet or 0 records if SubnAdmGetTable if possible

Some SA PathRecord Query related log messages

- pr_rcv_get_path_parms: ERR 1F05: Can't find remote phys port of ibsw-1 (GUID: 0x0002c90300908780) port 3 while routing to LID 342
  - Destination GUID not currently known by SM
- pr_rcv_get_path_parms: ERR 1F07: Dead end path on switch ibsw-1 (GUID: 0x00066a00e30029b5) to LID 342
  - Routing issue
- pr_rcv_get_port_pair_paths: ERR 1F21:Obtained destination LID of 0. No such LID possible (client-1 mlx4_1 port 1)
  - SM has not yet configured LID for destination port

OpenSM Log File Walkthru

- IB multicast is setup via SA MCMemberRecord
  - No broadcast by default as with “LANs”
- Used by IPoIB and EoIB
  - librdmacm and other multicast libraries also
- Single attribute used for both group creation and port joining (as well as port leaving and group deletion)

Most Common SA MCMemberRecord Query Errors

- Most Common SA MCMemberRecord Query Errors
  - Group creation
    - Insufficient parameters to create MC group
      - To create MC group, need additional parameters like rate, MTU, etc.
      - Common with IPv4 “router” multicast groups like 224.0.0.x
  - Port join
    - Port has lower rate or MTU than MC group
      - Possible workaround is to reduce rate or MTU of entire MC group to allow this port to join
        - » Administrator decision

SA MCMemberRecord Query related log messages

- mcmr_rcv_join_mgrp: ERR 1B11: method = SubnAdmSet, scope_state = 0x1, component mask = 0x0000000000010083, expected comp mask = 0x0000000000130c7, MGID: ff12:601b:ffff:: 2 from port 0x0002c9030014d081 (client-1 mlx4_1)
- mcmr_rcv_join_mgrp: ERR 1B11: Port 0x50800200008d9339 (MT25408 ConnectX Mellanox Technologies) failed to join non-existing multicast group with MGID ff12:401b:ffff::16, insufficient components specified for implicit create (comp_mask 0x10083
  - Log message improved (latter is 3.3.17 message)
  - MC Group is IPv6 (0x601b) based
  - This is due to insufficient parameters being supplied
  - Can preconfigure this group in partitions.conf to eliminate this log message

SA MCMemberRecord Query related log messages

- mcmr_rcv_join_mgrp: ERR 1B12: validate_more_comp_fields, validate_port_caps, or JoinState = 0 failed for MGID: ff12:a01b:fe80::d00:0:0 port 0x0002c903001c5621 (MT25408 ConnectX Mellanox Technologies), sending IB_SA_MAD_STATUS_REQ_INVALID
  - Port capabilities (rate, MTU) insufficient for group is most likely cause
  - Turn up log level if not (perhaps with PML)

OpenSM Log File Walkthru

- drop_mgr_remove_port: Removed port with GUID:0x002590ffff192171 LID range [389, 389] of node:cja241 HCA-1
  - OpenSM has drop manager which deals with removing nodes and ports when the subnet changes
  - Informational message just indicating that port was removed at SM

OpenSM Log File Walkthru

- log_trap_info: Received Generic Notice type:1 num: 131 (Flow Control Update watchdog timer expired) Producer:2 (Switch) from LID:364 Port 14 TID: 0x0000000000000f8
  - Switch SMA issued urgent trap
    - Flow Control Update watchdog timer expired at <LIDADDR><PORTNO>
  - Flow control update errors
    - For each VL active in the current port configuration, except VL 15 there shall be a watchdog timer monitoring the arrival of flow control updates. If the timer expires without receiving an update, a flow control update error has occurred. The period of the watchdog timer shall be 400,000 +3%/-51% symbol times. This timer shall only run when PortState = Arm or Active. When PortState = ActiveD, this timer shall be reset. When PortState = Initialize or when a flow control packet is received, the timer shall be reset.
  - Likely due to mismatch in OperationalVLs on peer ports

OpenSM Log File Walkthru

- log_rcv_cb_error: ERR 3111: Received MAD with error status = 0x1C
  SubnGetResp(SwitchInfo), attr_mod 0x0, TID 0x1014000d
  Initial path: 0,1,7,1,2,25 Return path: 0,10,31,13,1,12
  - Status 0x1c (status 7) is SMA rejection of SwitchInfo MAD
    - Likely due to issue with new MulticastFDBTop field option
  - Initial path is outgoing path from SM to switch indicating error
  - Check firmware version of indicated switches
    - Update if old and if possible
  - Another alternative is to disable this at SM
    - # Use SwitchInfo:MulticastFDBTop if advertised in PortInfo:CapabilityMask
      use_mfttop TRUE

OpenSM Log File Walkthru

- sm_mad_ctrl_send_err_cb: ERR 3120 Timeout while getting attribute 0x11 (NodeInfo); Possible mis-set mkey?
  - SM did not receive response to NodeInfo query from SMA
    - First query from SM to node
  - Should be transient error
    - Should add path or LID to SMA to error message so can debug
  - Check VL15 dropped counter
  - Likely subnet changing issue
  - Could also be mkey issue

OpenSM Log File Walkthru

- vl15_send_mad: ERR 3E03: MAD send failed (IB_UNKNOWN_ERROR)
  - SM class MADs (SMPs) are sent on VL15
  - Indicates osm_vendor_send failure
    - See next slide

OpenSM Log File Walkthru – OpenSM vendor layer

- OpenSM vendor layer
  - libvendor/osm_vendor_ibumad.c
  - Uses libibumad (and user_mad and mad kernel modules) for QP0 (SM class) and QP1 (GS class) sending/receiving
- osm_vendor_send: ERR 5430: Send p_madw = 0x7fd66404c320 of size 256 TID 0x3000098626 failed -5 (Invalid argument)
  - Error from kernel - multiple reasons return EINVAL

OpenSM Log File Walkthru

- sm_mad_ctrl_send_err_cb: ERR 3120 Timeout while getting attribute 0xFF90 (MLNXExtendedPortInfo); Possible mis-set mkey?
  - Mellanox proprietary SM MAD for FDR10 support
  - May indicate “old” Mellanox firmware
    - Check version and update if possible
  - Alternative is to shut off FDR10 support in SM via fdr10 option in opensm config file
    - # FDR10 on ports on devices that support FDR10
Values are:
0: don't use fdr10 (no MLNX ExtendedPortInfo MADs)
Default 1: enable fdr10 when supported
2: disable fdr10 when supported
fdr10 0
  - General note on any SM MAD timeout indicated in log message
    - Could be unresponsive node
      - Link Up but SMA not responding

OpenSM Log File Walkthru - PerfMgr

- perfmgr_mad_send_err_callback: ERR 5402: cja241 HCA-1 (0x2590ffff192170) port 1 LID 389 TID 0x58d5d3c
- perfmgr_mad_send_err_callback: ERR 5402: cja241 HCA-1 (0x2590ffff192170) port 1 LID 389 TID 0x58d7171
  - PMA indicated did not respond to PerfMgr query (get or set)
- perfmgr_send_mad: ERR 54FF: PM was NOT in Suspended state???
  - When sending PerfMgt MAD to PMA, PerfMgr was not in “suspended” state which is what was expected
    - Not sure why – need to investigate code for this

OpenSM Log File Walkthru - PerfMgr

- log_send_error: ERR 5410: Send completed with error (IB_TIMEOUT) – dropping
  - PerfMgr did not receive response to PMA query
  - In latest source, this message has changed and indicates the failure is on PerfMgt ClassPortInfo query and node and port to which is was directed to
- log_send_error: ERR 5411: DR SMP Send completed with error (IB_TIMEOUT) – dropping
  - Inconsistency with this error number and latest sources
  - Now indicates PerfMgr failed to clear counters for node/port

“Cryptic” Log Messages

- Always looking to improve wording of log messages
  - Suggestions are welcome!
- Most common laments about messages are related to SA multicast and SM MAD timeouts/rejections

Thank You

OPENFABRICS ALLIANCE