OPENFABRICS ALLIANCE

OFV-WG: InfiniBand tools overview

April 14th, 2015

Alex Netes

Agenda

- infiniband-diags
- ibutils
- Libibprof
- ibdump

Infiniband-diags

- Package included in OFED
- Set of single operation tools
  - smpquery/smpdump – tool to send SMPs
  - saquery – tool to send SA queries
  - iblinkinfo - report link info for all links in the fabric
  - ibnetdiscover - discover InfiniBand topology
  - ibqueryerrors - query and report IB port counters
  - ibtracert - trace InfiniBand path
  - **There are many more…**

Infiniband-diags - examples

```
# smpquery ni -D 0
# Node info: DR path slid 65535; dlid 65535; 0
BaseVers:.....................1
ClassVers:....................1
NodeType:.....................Channel
Adapter
NumPorts:.....................1
SystemGuid:...................
0x0002c9030004e93b
Guid:.........................
0x0002c9030004e938
PortGuid:.....................
0x0002c9030004e939
PartCap:......................128
DevId:........................0x673c
Revision:.....................0x000000a0
LocalPort:....................1
VendorId:.....................0x0002c9
```

```
# saquery -g
MCMemberRecord group dump:

MGID..................ff12:401b:ffff::1
               Mlid..................0xC001
               Mtu...................0x84
               pkey..................0xFFFF
               Rate..................0x83
               SL....................0x0
MCMemberRecord group dump:

MGID..................ff12:401b:ffff::ffff:ffff
               Mlid..................0xC000
               Mtu...................0x84
               pkey..................0xFFFF
               Rate..................0x83
               SL....................0x0
MCMemberRecord group dump:

MGID..................ff12:601b:ffff::1
               Mlid..................0xC004
               Mtu...................0x84
               pkey..................0xFFFF
               Rate..................0x83
               SL....................0x0
```

Infiniband-diags - examples

```
# iblinkinfo
CA: r-ufm96 HCA-1:
      0x0002c90300ff6980     12    1[  ] ==( 4X       14.0625 Gbps Active/  LinkUp)==>     268    5[  ]
"SwitchIB Mellanox Technologies" ( )
CA: r-ufm102 mlx5_0:
      0xe41d2d03005cf25c     10    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>     174    6[  ]
"MF0;switch-de779e:SX6012/U1" ( Could be 14.0625 Gbps)
CA: r-ufm102 mlx5_1:
      0xe41d2d03005cf25d      4    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>     174    4[  ]
"MF0;switch-de779e:SX6012/U1" ( Could be 14.0625 Gbps)
CA: r-ufm111 HCA-1:
      0x0002c903003421b1     14    1[  ] ==( 4X       14.0625 Gbps Active/  LinkUp)==>     174    7[  ]
"MF0;switch-de779e:SX6012/U1" ( )
CA: r-ufm101 HCA-2:
      0x0002c9030006ba5b      6    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>     174    3[  ]
"MF0;switch-de779e:SX6012/U1" ( )
CA: r-ufm102 mlx4_0:
      0x0002c9030010c6f1      2    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>     174    1[  ]
"MF0;switch-de779e:SX6012/U1" ( )
Switch: 0xe41d2d030003e470 SwitchIB Mellanox Technologies:
         268    1[  ] ==(                Down/ Polling)==>             [  ] "" ( )
         268    2[  ] ==(                Down/ Polling)==>             [  ] "" ( )
         268    3[  ] ==(                Down/ Polling)==>             [  ] "" ( )
         268    4[  ] ==( 4X       14.0625 Gbps Active/  LinkUp)==>     174    5[  ] "MF0;switch-
de779e:SX6012/U1" ( )
         268    5[  ] ==( 4X       14.0625 Gbps Active/  LinkUp)==>      12    1[  ] "r-ufm96 HCA-1" ( )
```

Infiniband-diags - examples

```
#
# Topology file: generated on Tue Apr 14 18:51:10 2015
#
# Initiated from node 0002c9030004e938 port 0002c9030004e939

vendid=0x2c9
devid=0xcb20
sysimgguid=0xe41d2d030003e470
switchguid=0xe41d2d030003e470(e41d2d030003e470)
Switch  36 "S-e41d2d030003e470"         # "SwitchIB Mellanox Technologies" base port 0 lid 268 lmc 0
[4]     "S-f4521403005764b0"[5]         # "MF0;switch-de779e:SX6012/U1" lid 174 4xFDR
[5]     "H-0002c90300ff6980"[1](2c90300ff6980)          # "r-ufm96 HCA-1" lid 12 4xFDR
[31]    "S-f4521403005764b0"[9]         # "MF0;switch-de779e:SX6012/U1" lid 174 4xFDR
[32]    "S-e41d2d030003e470"[33]                # "SwitchIB Mellanox Technologies" lid 268 4xEDR
[33]    "S-e41d2d030003e470"[32]                # "SwitchIB Mellanox Technologies" lid 268 4xEDR

vendid=0x2c9
devid=0xc738
sysimgguid=0xf4521403005764b0
switchguid=0xf4521403005764b0(f4521403005764b0)
Switch  12 "S-f4521403005764b0"         # "MF0;switch-de779e:SX6012/U1" enhanced port 0 lid 174 lmc 0
[5]     "S-e41d2d030003e470"[4]         # "SwitchIB Mellanox Technologies" lid 268 4xFDR
[6]     "H-e41d2d03005cf25c"[1](e41d2d03005cf25c)               # "r-ufm102 mlx5_0" lid 10 4xQDR
[9]     "S-e41d2d030003e470"[31]                # "SwitchIB Mellanox Technologies" lid 268 4xFDR

vendid=0x2c9
devid=0x1011
sysimgguid=0x2c90300ff6980
caguid=0x2c90300ff6980
Ca      2 "H-0002c90300ff6980"          # "r-ufm96 HCA-1"
[1](2c90300ff6980)      "S-e41d2d030003e470"[5]         # lid 12 lmc 0 "SwitchIB Mellanox Technologies" lid 268 4xFDR

vendid=0x2c9
devid=0x1013
sysimgguid=0xe41d2d03005cf25c
caguid=0xe41d2d03005cf25c
Ca      1 "H-e41d2d03005cf25c"          # "r-ufm102 mlx5_0"
[1](e41d2d03005cf25c)   "S-f4521403005764b0"[6]         # lid 10 lmc 0 "MF0;switch-de779e:SX6012/U1" lid 174 4xQDR
```

Infiniband-diags - examples

```
# ibqueryerrors
Errors for "r-ufm96 HCA-1"
   GUID 0x2c90300ff6980 port 1: [LinkDownedCounter == 56] [PortRcvErrors == 53114] [PortRcvSwitchRelayErrors == 52555]
[PortXmitDiscards == 5] [LocalLinkIntegrityErrors == 15]
Errors for "r-ufm102 mlx5_0"
   GUID 0xe41d2d03005cf25c port 1: [LinkDownedCounter == 46] [PortRcvErrors == 2057] [PortXmitDiscards == 104]
Errors for "r-ufm102 mlx5_1"
   GUID 0xe41d2d03005cf25d port 1: [LinkDownedCounter == 67] [PortRcvErrors == 2948] [PortXmitDiscards == 106]
Errors for "r-ufm111 HCA-1"
   GUID 0x2c903003421b1 port 1: [PortXmitWait == 2]
Errors for "r-ufm101 HCA-2"
   GUID 0x2c9030006ba5b port 1: [LinkDownedCounter == 16] [PortRcvErrors == 5] [PortXmitDiscards == 32]
Errors for "r-ufm102 mlx4_0"
   GUID 0x2c9030010c6f1 port 1: [LinkDownedCounter == 16] [PortRcvErrors == 719] [PortXmitDiscards == 32]
Errors for 0xe41d2d030003e470 "SwitchIB Mellanox Technologies"
Errors for 0xf4521403005764b0 "MF0;switch-de779e:SX6012/U1"
   GUID 0xf4521403005764b0 port ALL: [LinkDownedCounter == 30] [PortRcvSwitchRelayErrors == 229] [PortXmitDiscards == 1404]
[PortXmitWait == 47985256]
   GUID 0xf4521403005764b0 port 0: [PortXmitWait == 47985256]
   GUID 0xf4521403005764b0 port 1: [LinkDownedCounter == 2] [PortRcvSwitchRelayErrors == 174] [PortXmitDiscards == 67]
   GUID 0xf4521403005764b0 port 2: [LinkDownedCounter == 13] [PortRcvSwitchRelayErrors == 32] [PortXmitDiscards == 182]
   GUID 0xf4521403005764b0 port 3: [LinkDownedCounter == 12] [PortRcvSwitchRelayErrors == 3] [PortXmitDiscards == 1]
   GUID 0xf4521403005764b0 port 4: [LinkDownedCounter == 1] [PortXmitDiscards == 176]
   GUID 0xf4521403005764b0 port 5: [PortXmitDiscards == 900]
   GUID 0xf4521403005764b0 port 6: [LinkDownedCounter == 1] [PortRcvSwitchRelayErrors == 1] [PortXmitDiscards == 78]
   GUID 0xf4521403005764b0 port 7: [LinkDownedCounter == 1] [PortRcvSwitchRelayErrors == 19]
Errors for "r-ufm101 HCA-1"
   GUID 0x2c9030004e939 port 1: [LinkDownedCounter == 16] [PortRcvErrors == 89] [PortXmitDiscards == 32]

## Summary: 9 nodes checked, 9 bad nodes found
##          56 ports checked, 28 ports have errors beyond threshold
## Thresholds:
## Suppressed:
```

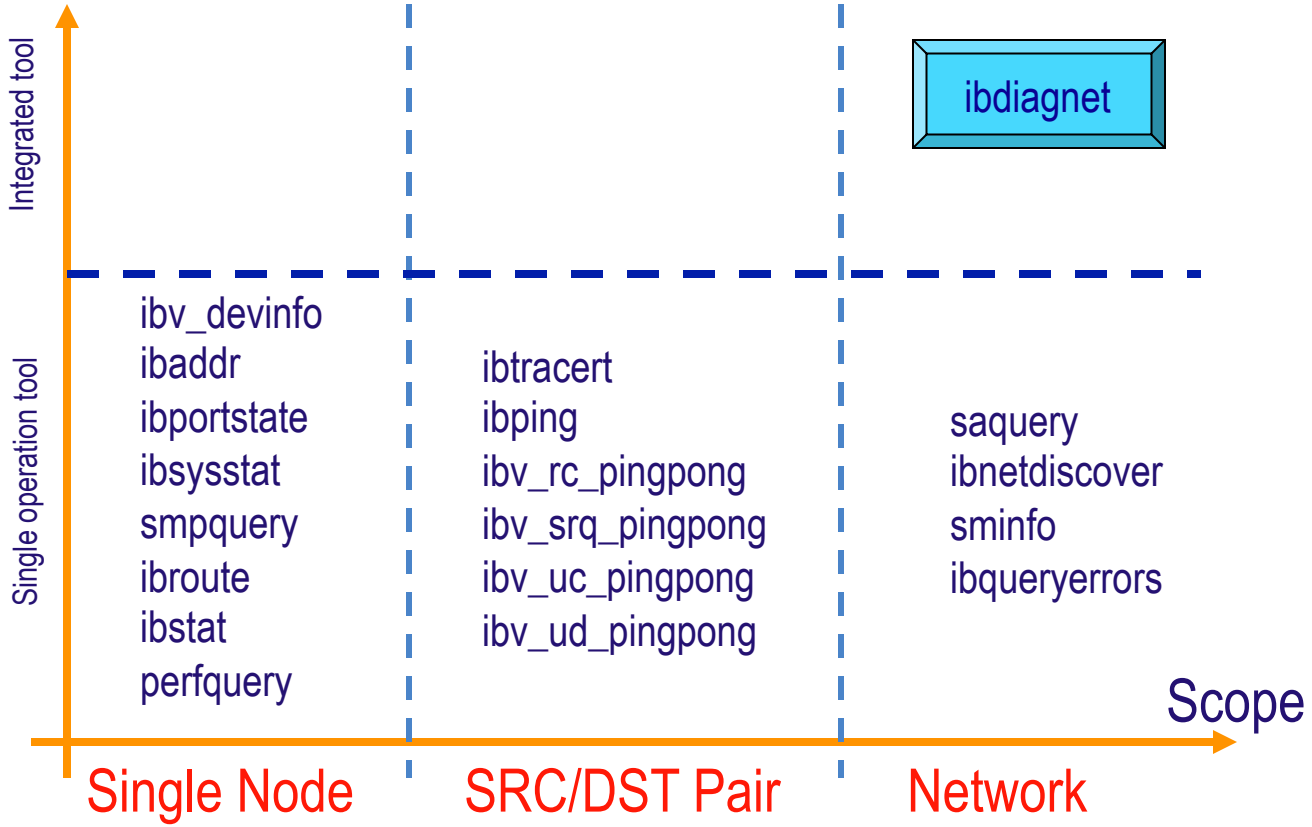
Infiniband-diags - examples

```
# ibtracert 25 24

From switch {0x0021283a8620b0f0} portnum 0
lid 25-25 "SwitchIB Mellanox Technologies"

[1] -> switch port {0x0021283a8620b0c0}[9]
lid 22-22 "SwitchIB Mellanox Technologies"

[2] -> switch port {0x0021283a8620b0e0}[8]
lid 24-24 "SwitchIB Mellanox Technologies"

To switch {0x0021283a8620b0e0} portnum 0 lid
24-24 "SwitchIB Mellanox Technologies"
```

ibutils

- ibdiagnet – IB network diagnostic tool
- ibcongest – Static congestion analysis
- ibdmchk – Network offline checker
- ibtoopdiff - Fabric Topology Matcher
- ibmgtsim – IB management simulator

ibdiagnet and other OFA tools



ibdiagnet tool

- Scans the fabric using directed / lid route packets
  - Extracts all the available information regarding its connectivity and devices.
- Checks errors in ports, nodes, links and cluster scopes and reports them.

ibdiagnet checks and features

- Fabric discovery
- Duplicated GUIDs detection – check node / port GUIDs
- Links check –
  - No bad link
  - All links are in logical state active
- Lids check –
  - No duplicated lids
  - No zero lids
- SM check –
  - There is one master subnet manger in the fabric
  - The master subnet manger is the correct one
- Speed / Width checks –
  - Actual Speed/Width is according to maximum supported values or a given command line values

ibdiagnet checks and features

- Port counters –
  - No overflowed for error counters
  - A threshold can be given for each counter via command line
  - 2 Sample of counters during run and check each error counter don't exceed its default threshold. The time between the 2 samples can be given via command line

- Routing Checks –
  - Check partition configuration
  - Check IPoIB subnets configuration
  - Check QoS Configuration (SL2VL)
  - Check all HCA to HCA routes –
    - CA to CA : LFT ROUTE HOP HISTOGRAM
    - LMC BASED ROTING :COMMON NODES HISTOGRAM
    - LMC BASED ROTING :COMMON SYSTEMS HISTOGRAM
    - LFT CA to CA : SWITCH OUT PORT - NUM PATHS HISTOGRAM
    - LFT CA to CA : SWITCH OUT PORT - NUM DLIDS HISTOGRAM
  - Credit loops

- Topology matching –
  - reports any difference between given topo and discovered fabric

Ibdiagnet2

- Mellanox OFED tool
- Extended features
  - Supports extended speeds
  - Supports additional counters
    - Extended speeds
    - Counters per lane
    - FEC extended speeds counters
  - Supports scope feature
  - Dumps FEC / retransmission info in “iblinkinfo” format
  - Extended routing checks
    - Adaptive routing
    - Multicast
  - Alias GUIDs checks
  - Dumps TOPO and IBNL files
  - Plugin interface

Ibdiagnet2 - example

```
Discovery
-I- Fabric Discover finished successfully
-I- Discovered 112 nodes (48 Switches & 64 CA-s).
-I- Port Info Extended finished successfully
-I- Duplicated GUIDs detection finished successfully
-I- Switch Info retrieving finished successfully

-------------------------------------------
Topology
-I- Parsing topology definition:/opt/wd/ibmgtsim.20150411_165112_28544/topology_mismatch_host_name
-I- Defined 112/112 systems/nodes
-I- Topology Matching
-W- Found mismatches between the topology defined in /opt/wd/ibmgtsim.20150411_165112_28544/topology_mismatch_host_name and the discovered fabric.
-I- Performing Topology Matching ...
-----------------------------------------------------------------
-I- Stage 1: Pair nodes by their topo-name and NodeDesc...
-I- Matched 110 nodes by name.
-----------------------------------------------------------------
-I- Stage 2.1 Matching more nodes if connected to previously matched nodes ...
-I- Successfuly matched 2 more nodes
-I- Stage 2.2 Matching more nodes if connected to previously matched nodes ...
-----------------------------------------------------------------
-I- Stage 3: Final reports ....
-W- Total: 2 mismatched names discovered
-W- Matching nodes have different names. Expected node name: Host-2/U1, Discovered node name: H-35/U1
-W- Matching nodes have different names. Expected node name: Host-1/U1, Discovered node name: H-7/U1
-----------------------------------------------------------------
-I- Total: 384 fully matching ports found
-----------------------------------------------------------------
-I- Topo match map written to:/opt/wd/ibmgtsim.20150411_165112_28544/mismatching_host_names/topodiff.names

-------------------------------------------
Summary
-I- Stage                    Warnings   Errors     Comment
-I- Discovery                4          0
-I- Topology                 3          0

-I- You can find detailed errors/warnings in: /opt/wd/ibmgtsim.20150411_165112_28544/mismatching_host_names/ibdiagnet2.log

-I- ibdiagnet database file   : /opt/wd/ibmgtsim.20150411_165112_28544/mismatching_host_names/ibdiagnet2.db_csv
-I- LST file                  : /opt/wd/ibmgtsim.20150411_165112_28544/mismatching_host_names/ibdiagnet2.lst
-I- Network dump file         : /opt/wd/ibmgtsim.20150411_165112_28544/mismatching_host_names/ibdiagnet2.net_dump
```

ibdiagnet2 dump files

- **ibdiagnet2.log** - A log file with detailed information.
- **ibdiagnet2.db_csv** - A dump of the internal tool database.
- **ibdiagnet2.lst** - A list of all the nodes, ports and links in the fabric.
- **ibdiagnet2.pm** - A dump of all the nodes PM counters.
- **ibdiagnet2.mlnx_cntrs** - A dump of all the nodes Mellanox diagnostic counters.
- **ibdiagnet2.net_dump** - A dump of all the links and their features.
- **ibdiagnet2.pkey** - A list of all pkeys found in the fabric.
- **ibdiagnet2.aguid** - A list of all alias GUIDs found in the fabric.
- **ibdiagnet2.sm** - A dump of all the SM (state and priority) in the fabric.
- **ibdiagnet2.fdbs** - A dump of unicast forwarding tables of the fabric switches.
- **ibdiagnet2.mcfdbs** - A dump of multicast forwarding tables of the fabric switches.
- **ibdiagnet2.slvl** - A dump of SLVL tables of the fabric switches.
- **ibdiagnet2.nodes_info** - A dump of all the nodes vendor specific general information for nodes who supports it.
- **ibdiagnet2.plft** - A dump of Private LFT Mapping of the fabric switches.
- **ibdiagnet2.ar** - A dump of Adaptive Routing configuration of the fabric switches.
- **ibdiagnet2.vl2vl** - A dump of VL to VL configuration of the fabric switches.

ibcongest

- Analyzes congestion for a traffic schedule provided in a "schedule-file" or use an automatically generated schedule of all-to-all-shift. The schedule file may define multiple stages of communication - each one of them is considered as infinite and the traffic switches together to the new stage.
- Calculates routing for a given topology (topo-mode) or use extracted lst/fdb files (lst-mode).
- The utility provides both Host-Spot-Degree analysis, which is the count of number of flows per link, and bandwidth per-flow calculation.
- For HSD this tool provides a histogram of number of links vs. HSD over all stages and a histogram of number of stages vs. the worst link HSD per stage.
- Similarly for flow bandwidth a histogram of flows vs. bandwidth and stages vs. worst bandwidth per stage. Bandwidth can be calculated assuming perfect congestion or without any congestion control.
- Detailed log files can be obtained by providing specific flags.

ibcongest - example

```
------------------ NUM ALTERNATE PORTS TO CA HISTOGRAM --------------------
Describes how many out ports on every switch have the same Min Hop to each
target CA. Or in other words how many alternate routes are possible at the
switch level. This is useful to show the symmetry of the cluster.

OUT-PORTS NUM-SW-LID-PAIRS
   1   8
---------------------- CA to CA : MIN HOP HISTOGRAM ----------------------
The number of CA pairs that are in each number of hops distance.
The data is based on topology only - even before any routing is run.

HOPS NUM-CA-CA-PAIRS
  2   56
-------------------------------------------------------------------------
-I- Found worst min hops:2 at node:SW-1-0/U1 to node:H-1/U1
-------------------------- TRACE PATH BY MIN HOPS -------------------------
-I- Tracing by Min Hops from lid:2 to lid:1
[  0] FROM Host:SW-1-0 Plug:P1
          Node:SW-1-0/U1 Port:1
      TO   Host:H-1 Plug:U1/P1
          Node:H-1/U1 Port:1
-------------------------------------------------------------------------
-I- Using standard OpenSM Routing
-I- Verifying all CA to CA paths ...
---------------------- CA to CA : LFT ROUTE HOP HISTOGRAM ----------------
The number of CA pairs that are in each number of hops distance.
This data is based on the result of the routing algorithm.

HOPS NUM-CA-CA-PAIRS
  2   56
```

```
---------- LFT CA to CA : SWITCH OUT PORT - NUM DLIDS HISTOGRAM -----------
Number of actual Destination LIDs going through each switch out port
considering
all the CA to CA paths. Ports driving CAs are ignored (as they must
have = Nca - 1). If the fabric is routed correctly the histogram
should be narrow for all ports on same level of the tree.
A detailed report is provided in /tmp/ibdmchk.sw_out_port_num_dlids.

NUM-DLIDS NUM-SWITCH-PORTS
-------------------------------------------------------------------------


-I- Scanned:56 CA to CA paths
-------------------------------------------------------------------------


-I- Analyzing Congestion ...
-I- Simulating all-to-all traffic for 8 nodes
-I- Final Congestion Report
-------------------------------------------------------------------------
-I- Traced total:56 paths
-I- Worst link over subscrition:1 port:SW-1-0/P1
---------------------- TOTAL CONGESTION HISTOGRAM ----------------------
Describes distribution of oversubscription of paths per port.
NUM-PATHS NUM-OUT-PORTS
   0   56
   1   56
-------------------------------------------------------------------------

---------------------- STAGE CONGESTION HISTOGRAM ----------------------
Describes distribution of worst oversubscription of paths per stage.
WORST-CONG NUM-STAGES
   1   7
```

Libibprof – profiling library

- Profiling library for:
  - OFA verbs.
  - Mellanox MXM/HCOLL libraries.
- Plugin injection into verb library API to simulate network failures.
- Free under BSD license available https://github.com/mellanox-hpc/libibprof.

- Portable profiling infrastructure for parallel codes.
- Provides low-overhead performance summary of the calls from libibverbs.so, libmxm.so, libhcoll.so libraries.

Libibprof – example

```
% LD_PRELOAD=libibprof.so ibv_devinfo

================================================================================================

libibverbs                      :      count    total(ms)      avg(ms)      max(ms)      min(ms)
================================================================================================
ibv_get_device_list             :          1       0.8897       0.8897       0.8897       0.8897
ibv_free_device_list            :          1       0.0001       0.0001       0.0001       0.0001
ibv_open_device                 :          2       2.7479       1.3739       2.4120       0.3358
ibv_close_device                :          2       1.7194       0.8597       1.6812       0.0382
ibv_query_port                  :          4       1.8440       0.4610       0.8201       0.1702
ibv_exp_query_device            :          2       0.3972       0.1986       0.2835       0.1137
================================================================================================
total                           :                  7.5982
================================================================================================
wall time (%)                   :                  0.0010 %
================================================================================================
```

Libibprof – example

```
% LD_PRELOAD=libibprof.so mpirun -x LD_PRELOAD -np 8 mpi_hello

Hello from task 1 on jenkins01!
Hello from task 3 on jenkins01!
Hello from task 5 on jenkins01!
Hello from task 7 on jenkins01!
Hello from task 2 on jenkins01!
Hello from task 6 on jenkins01!
Hello from task 0 on jenkins01!
Hello from task 4 on jenkins01!

================================================================================================

libibverbs                     :      count    total(ms)      avg(ms)      max(ms)      min(ms)
================================================================================================
ibv_get_device_list            :          1       2.6830       2.6830       2.6830       2.6830
ibv_free_device_list           :          1       0.0003       0.0003       0.0003       0.0003
ibv_open_device                :          1       0.4915       0.4915       0.4915       0.4915
ibv_close_device               :          1       0.0357       0.0357       0.0357       0.0357
ibv_create_comp_channel        :          2       0.0093       0.0047       0.0050       0.0044
ibv_destroy_comp_channel       :          2       0.0074       0.0037       0.0038       0.0037
ibv_query_port                 :          3       1.3320       0.4440       0.6961       0.1486
ibv_alloc_pd                   :          1       0.0041       0.0041       0.0041       0.0041
ibv_dealloc_pd                 :          1       0.0042       0.0042       0.0042       0.0042
ibv_dereg_mr                   :          3       2.2027       0.7342       1.4345       0.3286
ibv_create_cq                  :          3      12.4912       4.1637       6.6688       2.4448
ibv_poll_cq                    :        350       0.1092       0.0003       0.0021       0.0000

================================================================================================
```

Libibprof – system variables

- IBPROF_TEST_MASK
  - 0 - fatal
  - 1 - error
  - 2 - warn
  - 3 - info
  - 4 - trace
- Example: export IBPROF_TEST_MASK=0xFF

Libibprof – system variables

- IBPROF_MODE
  - Modules:
    - USE_IBV - libibverbs (set as default)
    - USE_HCOL - libhcoll
    - USE_MXM - libmxm
  - Values:
    - 0 - none (transparent mode)
    - 1 - time profiling
    - 2 - error injection:
      - IBPROF_ERR_SEED - value for random generator
      - IBPROF_ERR_PERCENT - % of failures
    - 3 - verbose
      - IBPROF_TEST_MASK - 5th bit should be ON
- Example: export IBPROF_MODE=USE_IBV=0,USE_HCOL=1

Libibprof – system variables

- IBPROF_DUMP_FILE
  - Special symbols in file name:
    - %J - Job ID
    - %H - Host name
    - %T - Process ID (rank)
- IBPROF_OUTPUT_PREFIX=1
  - Prefix output from each process with it's hostname and pid.
- IBPROF_FORMAT=xml
  - Output format as xml.

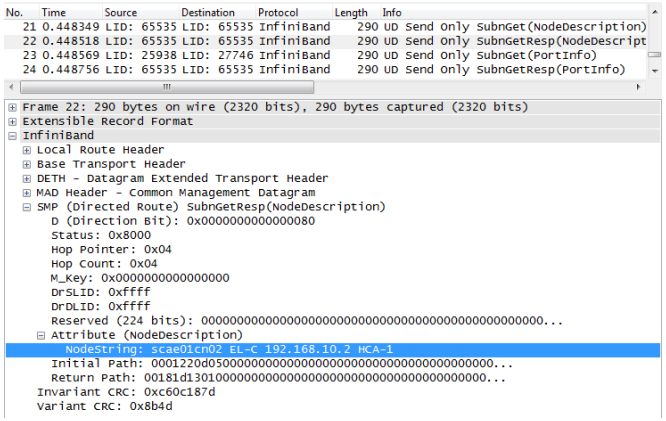
Libibprof – measure an arbitrary interval of code flow

```
/*
 * callid - Unique ID for measurement.
 * name - String name associated with measurement.
 */
ibprof_interval_start(int callid, char *name)
ibprof_interval_end(int callid)
```

ibdump

- Mellanox OFED tool
- Dump Infiniband traffic from Mellanox Technologies HCAs

```
# ibdump
Initiating resources ...
searching for IB devices in host
Port active_mtu=4096
MR was registered with addr=0x1ad6970, lkey=0x2100, rkey=0x2100, flags=0x1
 ---------------------------------------------
 Device                          : "mlx4_1"
 Physical port                   : 1
 Link layer                      : Infiniband
 Dump file                       : sniffer.pcap
 Sniffer WQEs (max burst size)   : 4096
 ---------------------------------------------


Ready to capture (Press ^c to stop):
Captured:         156 packets,         135712 bytes
```



Thank You