

InfiniBand Management



CONFIDENTIAL

OFED Cluster Based Tools



CONFIDENTIAL

Single Node

ibv_devinfo

ibstat

ibportstate

ibroute

smpquery

perfquery

SRC/DST Pair

lbdiagpath

ibtracert

ibv_rc_pingpong

ibv_srq_pingpong

ibv_ud_pingpong

ib_send_bw

ib_write_bw

Network

lbdiagnet

ibnetdiscover

ibhosts

ibswitches

saquery

sminfo

smpdump

- Open source Linux tools
- pdsh allows to run same command on multiple machines
 - Example
 - ‘pdsh -R ssh -w ibc0[01-10] ls’ will run ls command on ibc001 through ibc010
- dshbak formats output of pdsh into more readable form
 - -c flag will make nodes with identical output be grouped in one listing
 - Example
 - pdsh -w ibd0[02-32] ‘ibstat | grep State’ | dshbak -c
 - ibd[002-032]
 - -----State: Initializing
State: Down

■ Run performance tests

- /usr/bin/ib_write_bw
- /usr/bin/ib_write_lat
- /usr/bin/ib_read_bw
- /usr/bin/ib_read_lat
- /usr/bin/ib_send_bw
- /usr/bin/ib_send_lat

■ Usage

- Server: <test name> <options>
- Client: <test name> <options> <server IP address>

Note: Same options must be passed to both server and client. Use -h for all options.

■ ibswitches

- Lists all switches in cluster

■ ibhosts

- Lists all HCAs in cluster

■ ibtracert

- Shows path between two lids

```
– [root@ibd001 mft-2.5.0]# ibtracert -G 0x0002c90300001481 0x0002c90300001489
  From ca {0x0002c90300001480} portnum 1 lid 12-12 "ibd017 HCA-1"
  [1] -> switch port {0x000b8cffff002772}[5] lid 39-39 "MT47396 Infiniscale-III Mellanox Technologies"
  [6] -> ca port {0x0002c90300001489}[1] lid 15-15 "ibd012 HCA-1"
  To ca {0x0002c90300001488} portnum 1 lid 15-15 "ibd012 HCA-1"
```

- Reports a complete topology of cluster
- Shows all interconnect connections reporting:
 - Port LIDs
 - Port GUIDs
 - Host names
 - Link Speed
- GUID to name file can be used for more readable topology in regards to switch devices

- Simple usage is: `ibnetdiscover --node-name-map <guid to name file>`

```
root@mtilab32:~  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]# ibnetdiscover --node-name-map node_name_map  
#  
# Topology file: generated on Mon May 25 11:57:29 2009  
#  
# Max of 2 hops discovered  
# Initiated from node 0002c90300000148 port 0002c90300000149  
  
vendid=0x2c9  
devid=0xbd36  
sysimgguid=0x2c9020040525b  
switchguid=0x2c90200405258(2c90200405258)  
Switch 36 "S-0002c90200405258" # "SWITCH-1" enhanced port 0 lid 8 lmc 0  
[18] "H-0002c9030000057c"[1](2c9030000057d) # "mtilab31 HCA-1" lid 17 4xDDR  
[32] "H-0002c90300000148"[1](2c90300000149) # "mtilab32 HCA-1" lid 13 4xDDR  
  
vendid=0x2c9  
devid=0x634a  
sysimgguid=0x2c9030000057f  
caguid=0x2c9030000057c  
Ca 2 "H-0002c9030000057c" # "mtilab31 HCA-1"  
[1](2c9030000057d) "S-0002c90200405258"[18] # lid 17 lmc 0 "SWITCH-1" lid 8 4xDDR  
  
vendid=0x2c9  
devid=0x634a  
sysimgguid=0x2c9030000014b  
caguid=0x2c90300000148  
Ca 2 "H-0002c90300000148" # "mtilab32 HCA-1"  
[1](2c90300000149) "S-0002c90200405258"[32] # lid 13 lmc 0 "SWITCH-1" lid 8 4xDDR  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#  
[root@mtilab32 ~]#
```

- **SymbolErrors**
 - **Total number of minor link errors. Usually an 8b/10b error due to a bit error**
- **Link Recovers**
 - **Total number of times the Port Training state machine has successfully completed the link error recovery process.**
- **LinkDowned**
 - **Total number of times the Port Training state machine has failed the link error recovery process and downed the link.**
- **RcvErrors**
 - **Total number of packets containing an error that were receive on the port. Usually due to a CRC error caused by a bit error within the packet.**
- **RcvSwRelayErrors**
 - **Total number of packets received on the port that were discarded because they could not be forwarded by the switch relay. This counter should typically be ignored since Anafa-II has a bug that counts these when it gets a multicast packet on a port where that port also belongs to the multicast group of the packet.**
- **XmtDiscards**
 - **Total number of outbound packets discarded by the port because the port is down or congested. Usually due to the output port HOQ lifetime being exceeded.**
- **VL15Dropped**
 - **Number of incoming VL15 packets dropped due to resource limitations (e.g., lack of buffers) in the port**
- **XmtData,RcvData**
 - **Total number of 32-bit data words transmitted and received on the port.**
- **XmtPkts,RcvPkts**
 - **Total number of data packets transmitted and received on the port.**

■ Integrated diagnostic tools

- Queries cluster topology and indicates any port errors, link width, or link speed mismatch.
- Automates calls to many “low level” operations

■ Easy to use

- Similar flags, logs and reports for both tools
- Report using meaningful names when topology file is provided

- **-i <dev-index> -p <port-num>**
 - Device index (0..N) and port number connected to the network
- **-o <out-dir>**
 - Directory to output the reports to
- **-lw <1x|4x|12x> -ls <2.5|5|10>**
 - Link speed and width checked on every port on the network
- **-pm -pc**
 - Perform error counters extensive check or clear counters respectively
- **-r**
 - Extensive additional checks performed.
- **-P**
 - Sets threshold for error levels. Also checks for errors of counters based on absolute value of the error counter. When not using **-P** flag, error thresholds are only triggered based on how many errors were incremented DURING the ibdiagnet run.
- **-c**
 - Packets to be sent on each link for error level checking
- **-h -V -v**
 - Help, Verbosity and Revision flags respectively

- Ibdiagnet is particularly useful in finding misconfigured links (speed/width, topology mismatches, and marginal link/cable issues).
- Typical usage:
 - Clear all port counters using 'ibdiagnet -pc'
 - Stress the cluster
 - Check cluster using 'ibdiagnet -lw 4x -ls 5 -P all=1'
 - Checks for link speed, link width, and port error counters greater than 1

```
root@mtlab32:~  
-----  
-I- PM Counters Info  
-----  
-I- No illegal PM counters values were found  
-----  
-I- Links With links width != 4x (as set by -lw option)  
-----  
-I- No unmatched Links (with width != 4x) were found  
-----  
-I- Links With links speed != 5 (as set by -ls option)  
-----  
-I- No unmatched Links (with speed != 5) were found  
-----  
-I- Fabric Partitions Report (see ibdiagnet,pkey for a full hosts list)  
-----  
-I- PKey:0x7fff Hosts:2 full:2 partial:0  
-----  
-I- IPoIB Subnets Check  
-----  
-I- Subnet: IPv4 PKey:0x7fff QKey:0x00000b1b MTU:2048Byte rate:10Gbps SL:0x00  
-W- Suboptimal rate for group. Lowest member rate:20Gbps > group-rate:10Gbps  
-----  
-I- Bad Links Info  
-----  
-I- No bad link were found  
-----  
-I- Stages Status Report:  
-----  
STAGE Errors Warnings  
Bad GUIDs/LIDs Check 0 0  
Link State Active Check 0 0  
Performance Counters Report 0 0  
Specific Link Width Check 0 0  
Specific Link Speed Check 0 0  
Partitions Check 0 0  
IPoIB Subnets Check 0 1  
-----  
Please see /tmp/ibdiagnet.log for complete log  
-----  
-I- Done. Run time was 1 seconds.  
[root@mtlab32 ~]#
```

- Plan your cluster
- Install all switches, Nodes and Cables
- Install IB SW stack
- Make sure all FW are up to date
- Run SM and make sure all nodes are active
- Check topology matching
- Check link level errors while cluster is in idle
 - Approximately 2 error per min allowed or 100 an hour
 - Fix problematic links until stable
- Stress the fabric (for example with MPI)
 - Approximately 2 error per min allowed or 100 an hour
 - Fix problematic links until stable

FabricIT Mellanox Management



CONFIDENTIAL

- **Chassis Management – ships with all switch systems that have CPU Modules**
 - System monitoring
 - RS232 console, 10/100/1000 Eth, IPoIB management
 - CLI / Web Interface / SNMP communication protocols
- **Fabric Management – FabricIT-EFM**
 - Subnet management, cluster diagnostics
 - IPoIB, CLI / Web Interface / SNMP communication protocols

■ Hardware monitoring

- Monitor and configure system parameters
- CPU / Memory / File System resources
- Port management
- Power supply management
- LED status
- Voltage, temperature status
- System reset

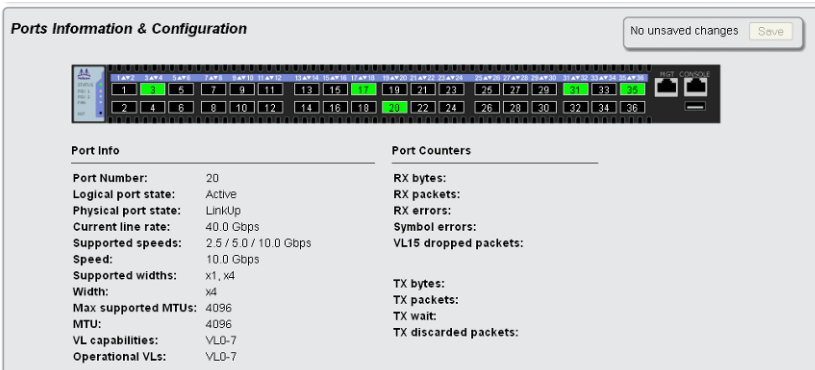
■ Error and Event Logs on the Switch

■ SNMP support

- Get, Traps
- Standard MIBs

■ Easy to use communication protocols

- CLI & Web interface
- Secure login and access with ACLs (Telnet/SSH and Secure HTTP)
 - Authentication And Authorization (AAA) : RADIUS, TACACS+
- IPoIB



Ports Information & Configuration

No unsaved changes Save

Port	1	5	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	
1	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36

Port Info

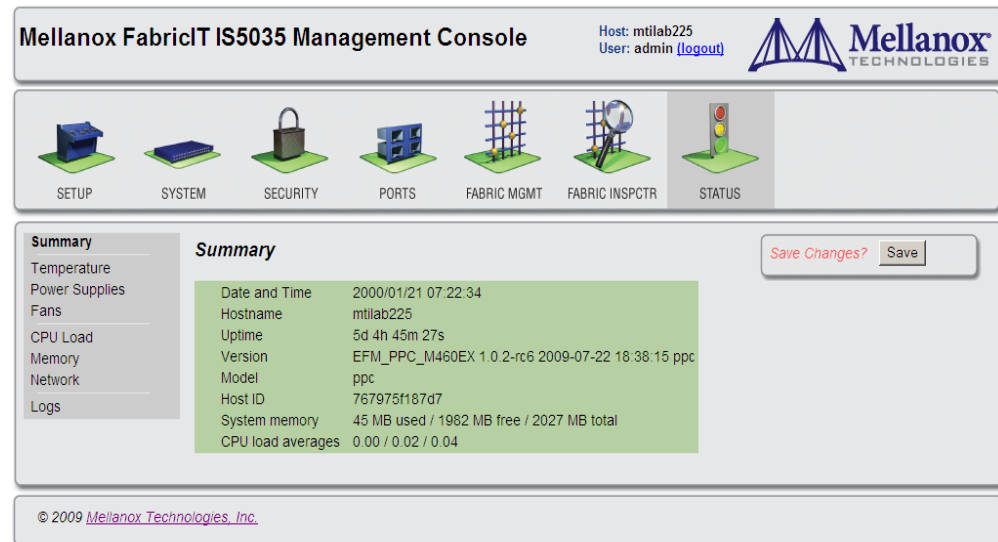
Port Number: 20
Logical port state: Active
Physical port state: LinkUp
Current line rate: 40.0 Gbps
Supported speeds: 2.5 / 5.0 / 10.0 Gbps
Speed: 10.0 Gbps
Supported widths: x1, x4
Width: x4
Max supported MTUs: 4096
MTU: 4096
VL capabilities: VL0-7
Operational VLs: VL0-7

Port Counters

RX bytes:
RX packets:
RX errors:
Symbol errors:
VL15 dropped packets:
TX bytes:
TX packets:
TX wait:
TX discarded packets:

- **Fabric Subnet Manager**
 - Subnet Manager and Subnet Administrator
 - Fabric initialization
 - Routing algorithm
 - Execution on boot-up or manually
 - Error logs and Debug Information
- **Advanced features**
 - QoS manager
 - Fabric Inspector cluster management
- **Fabric Inspector**
 - SM status, location, route checks
 - Duplicate GUID/LID's checks
 - Simple and intuitive interface for bring-up and maintenance
- **Additional Mellanox Tools**
 - Switch device Information
 - Switch Firmware upgrades
 - Port status
 - Error logs and Debug Information

- **Easy to use communication protocols**
 - CLI & Web Interface
 - Secure login and access with ACLs (Telnet/SSH and Secure HTTP)
 - Authentication And Authorization (AAA) : RADIUS, TACACS+
 - SNMP Agent
 - 3rd Party management (IBM Tivoli, HP OpenView, packet sniffer) tool interface
 - IPoIB



Mellanox FabricIT IS5035 Management Console

Host: mtilab225
User: admin (logout)

Mellanox TECHNOLOGIES

SETUP SYSTEM SECURITY PORTS FABRIC MGMT FABRIC INSPCTR STATUS

Summary

Temperature	
Power Supplies	
Fans	
CPU Load	
Memory	
Network	
Logs	

Summary

Date and Time	2000/01/21 07:22:34
Hostname	mtilab225
Uptime	5d 4h 45m 27s
Version	EFM_PPC_M460EX 1.0.2-rc6 2009-07-22 18:38:15 ppc
Model	ppc
Host ID	767975f187d7
System memory	45 MB used / 1982 MB free / 2027 MB total
CPU load averages	0.00 / 0.02 / 0.04


Save Changes? Save

© 2009 Mellanox Technologies, Inc.

Manager User Interfaces

Mellanox FabricIT MTS3610 Management Console

Host: switch-112082
User: admin (logout)



SETUP SYSTEM SECURITY PORTS FABRIC MGMT FABRIC INSPCTR STATUS

Summary

Summary	Summary
Temperature	Date and Time 2009/05/21 21:01:46
Power Supplies	Hostname switch-112082
Fans	Uptime 3h 29m 58.380s
CPU Load	Version EFM_PPC 1.0.0 2009-05-21 13:41:05 ppc
Memory	Model ppc
Network	Host ID ecaa8794f376
Logs	System memory 46 MB used / 458 MB free / 504 MB total
	CPU load averages 0.51 / 0.61 / 0.62

Unsaved changes Save

Web Interface

Familiar CLI

```
- PuTTY
debug      Debugging commands
demo      Set demo constant
echo      Set echo daemon configuration
email     Configure email and event notification via email
exit     Leave configuration mode
file     Manipulate files on disk
ftp-server Configure FTP server settings
help     View description of the interactive help system
hostname Set the system's hostname
image    Manipulate system software images
interface Configure network interfaces
tb8 (config) # interface
ether1 ether2 lo
tb8 (config) # interface
ether1 ether2 lo
tb8 (config) # interface ether1
alias      comment      ip          speed
bond       dhcp          mtu         zeroconf
bridge-group duplex      shutdown
tb8 (config) # interface ether1
alias      comment      ip          speed
bond       dhcp          mtu         zeroconf
bridge-group duplex      shutdown
tb8 (config) #
```

FabricIT Demo



CONFIDENTIAL

Thank You

www.mellanox.com

