Peloton Infiniband Experiences

Trent D'Hooge and Ira Weiny

Lawrence Livermore National Lab

OFED Conference

May 1, 2007

UCRL-PRES-230316

Clusters

- Zeus : 288 nodes
- Rhea: 576 nodes
- Atlas: 1152 nodes

Clusters

- Scalable Units
  - 138 Compute nodes
  - 4 router / gateway nodes
  - 1 login node
  - 1 management node
- Compute and login nodes route over IB network to router / gateway nodes to get to lustre and NFS
  - NFS is mounted over IP

IB Details

- Mellanox HCA (Arbel) (4x DDR)
- Voltaire Switches 9024D & 9288 (4X DDR)
  - Running at (SDR)
- OFED 1.1 based stack
  - OSM + diags
  - ibverbs + mthca
  - IPoIB (NFS)
  - Lustre Native IB to IP ”router” ko2iblnd

Overall things went well

- In general OFED and open source worked
- Good support from industry and community
- Working with Mellanox and Voltaire to tune opensm and mvapich variables

Hardware Issues

- Retry Exceeded error ”Code 12”
- HCA Catastrophic errors
- NETDEV Watchdog on IPoIB interface
- HCA command interface hang ”Go Bit”
- Ports negotiating to 1X

Hardware Issues

- Retry exceeded error ”Code 12”
  - Seen on all clusters
  - DDR problems (excessive errs on internal links)
  - Tuning VIADEV options from default for example:
    - VIADEV_DEFAULT_RETRY_COUNT = 7
    - VIADEV_DEFAULT_TIME_OUT = 22
  - Atlas required larger VIADEV_VBUF_TOTAL_SIZE
  - Tuning the SM
    - leaf_vl_stall_count = 0x03

Hardware Issues

- HCA Catastrophic errors
  - Three types
    - Internal
      - Only observed on Atlas. Caused by a certain jobs interacting with one another. FW update seems to have fixed the problem.
    - Parity
      - unknown, not seen very often.
    - Unknown
      - Motherboard / IB card interaction (Noise)
      - IB cards not installed correctly

Hardware Issues

- NETDEV Watchdog on IPoIB interface
  - Seen once on Rhea, and often on Atlas
  - IPoIB not getting priority
  - Applied patch to schedule queues for UD and RC traffic
    - options ib_mthca sched_queue_ud=1
    - Different patch exists in OFED 1.2 (”Work around kernel QP starvation”)

Hardware Issues

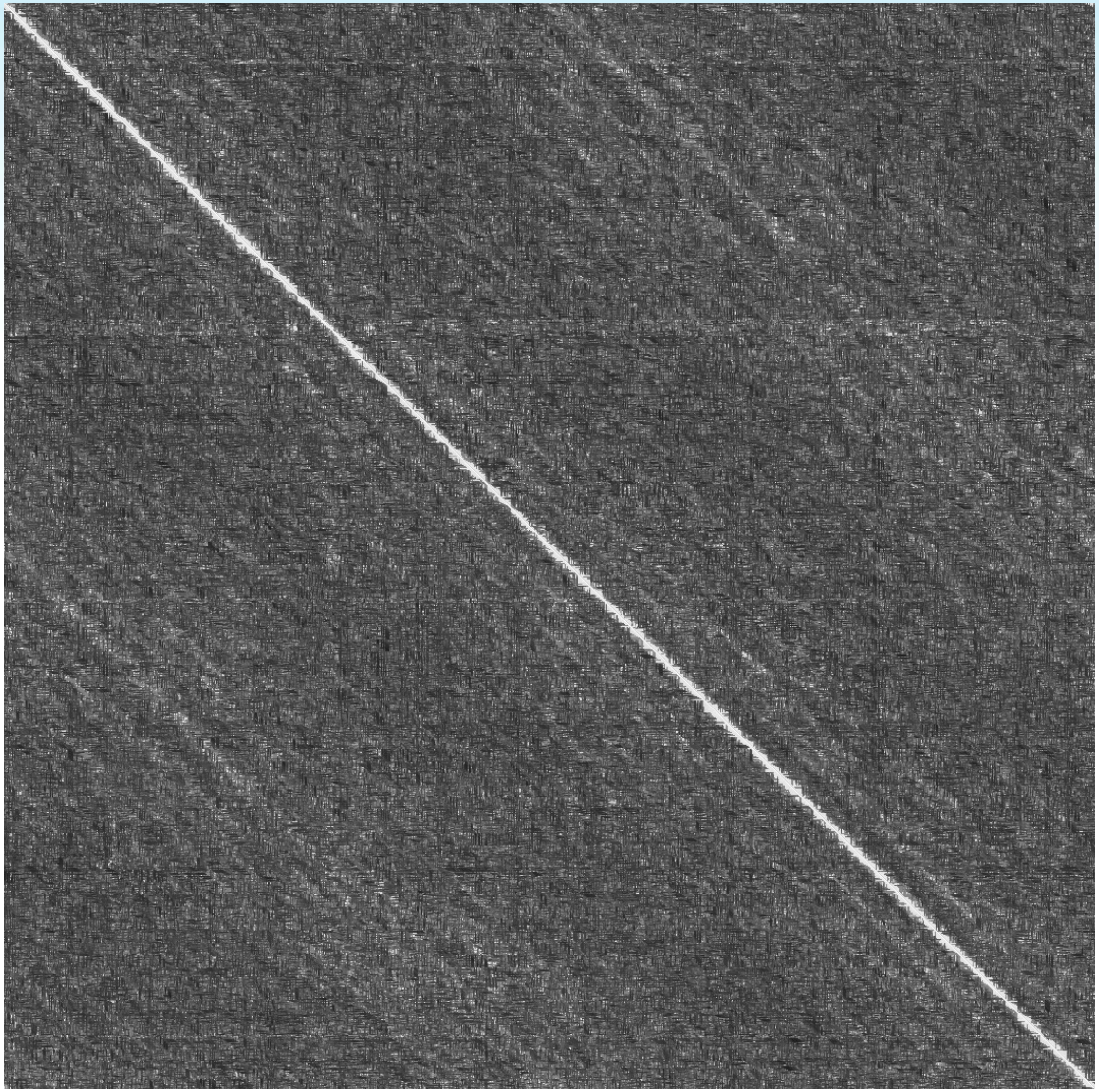
- Go Bit
  - Unknown cause, seen mostly on atlas
  - Might be a PCI issue
- Ports negotiating to 1X
  - On full system reboots some nodes will negotiate at 1X
    - Bouncing link or rebooting node fixes the problem
    - Seen on some internal links early on

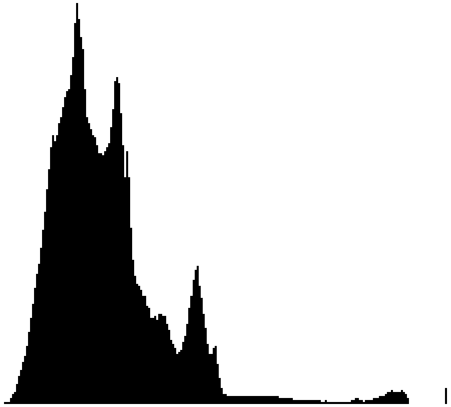
Static Routing results

- Atlas Cluster
- 1152 nodes
- IB 4X SDR
- Single path static routing
- Measured Peak MPI ~ 1.0 GB/s (using MVAPICH)
- Using ”linkcheck”

Static Routing

- Atlas send.bmp (Min: 95.179 Max: 761.987 Avg: 262.764



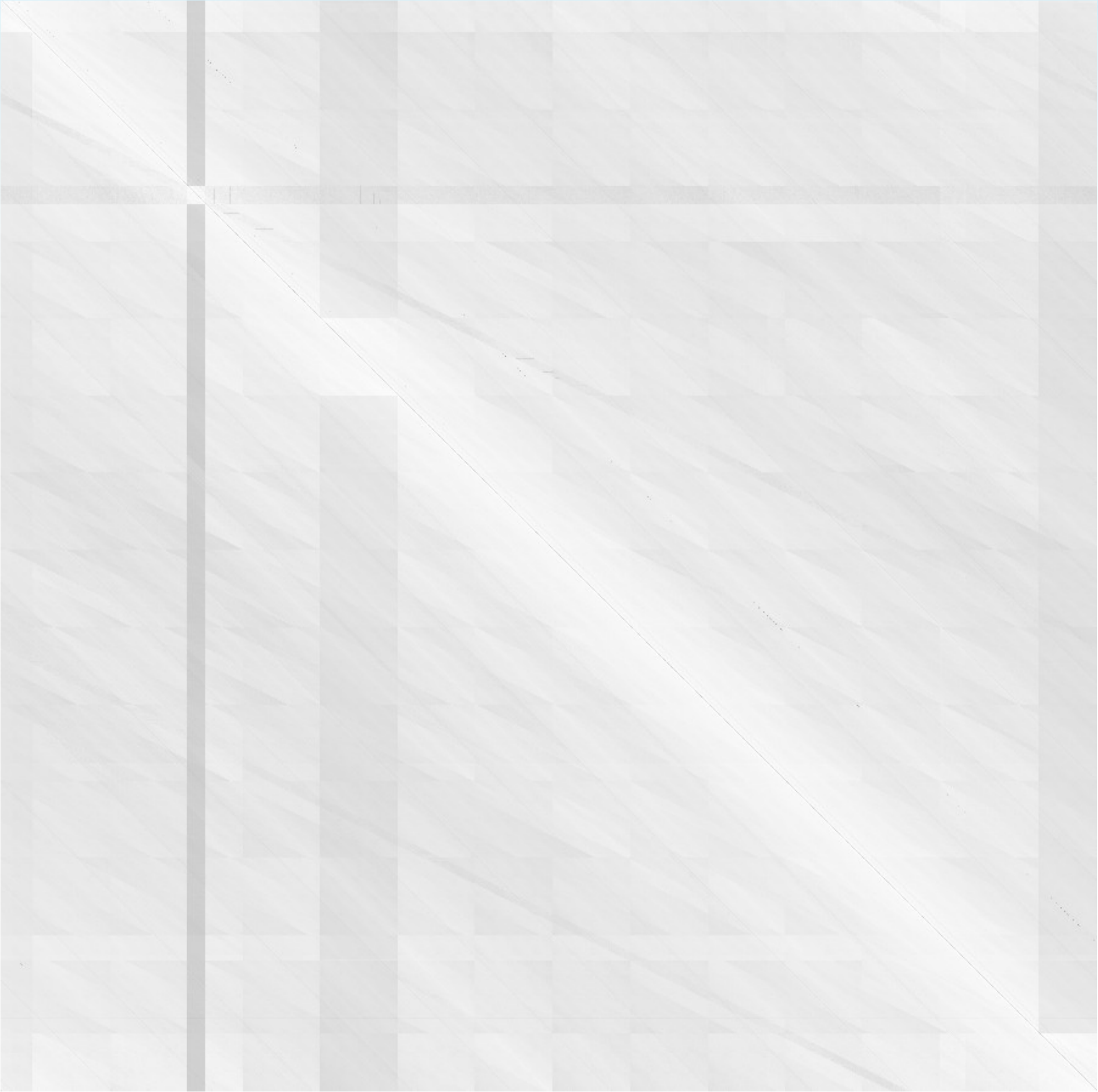


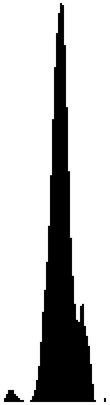
Fully Adaptive Routing

- Thunder Cluster for comparison
- 1024 nodes
- Elan 4 interconnect
- Peak with MPI ~ 900MB/s

Fully Adaptive Routing

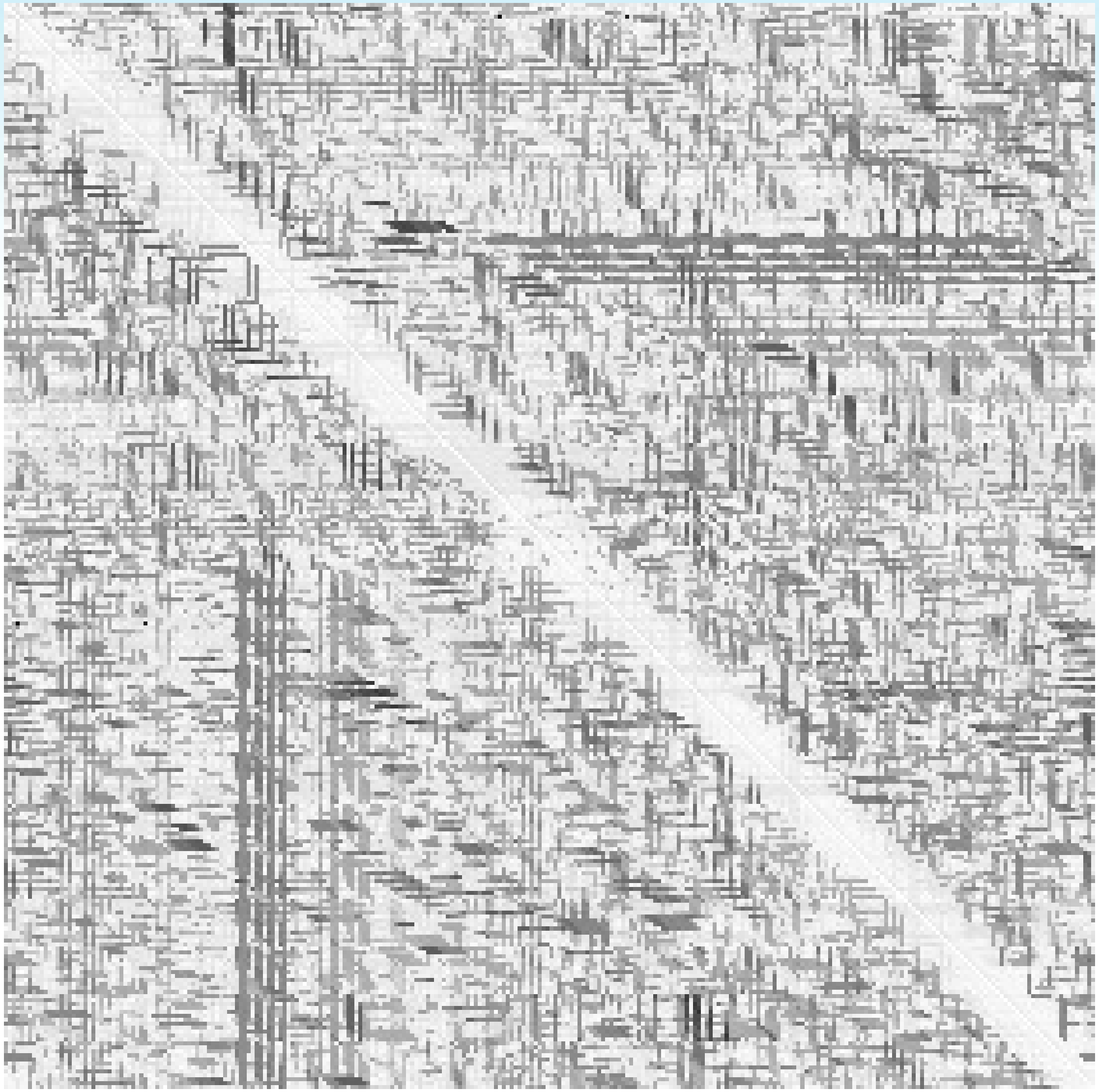
- Thunder send.bmp (Min: 247.87 Max: 402.744 Avg: 368.87)

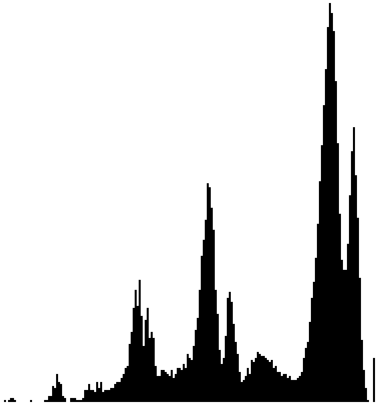




Static Routing (hotspots)

- Zeus (DDR) (Send, Min: 42.047 Max: 734:03 Avg: 579.69)

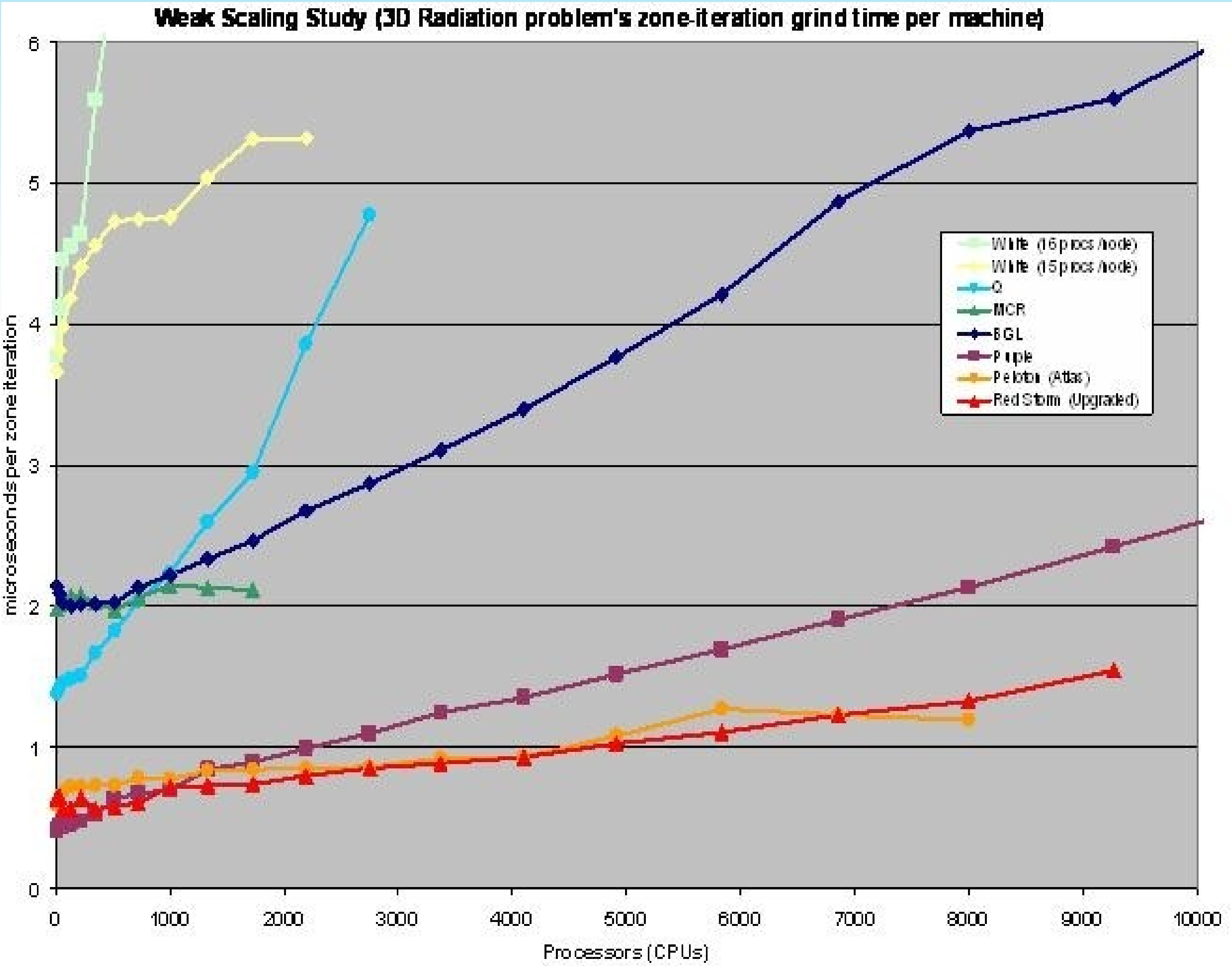




Static Routing (Conclusion)

- ”noise” in the IB images is consistent run after run, it is not noise but contention in the network
- This contention results in a poor average bandwidth
- DDR is more promising but still has ”hot spots”

Scaling comparisons



Software Issues

- Lustre bug caused by incorrect CM private data length (kernel patch supplied)
- Diags lacking
  - Needed tools to understand what was broken and where
  - Learning what errors are real vs. ones that can be ignored
- OSM OFED 1.1 had issues at scale
  - Long routing times on Atlas
  - SA starvation
  - Upgrade to OFED 1.2 (Thanks to Sasha and Hal)

Software Issues

- OpenSM unable to UP/DOWN route Atlas
  - Passing the root switch information to opensm allow the fabric to route properly
- IPoIB stops working
  - Seen on Atlas, any node coming in or out of the IB fabric would cause IPoIB to stop working
    - Race condition in the mcast join was fixed in opensm
- Learning what each mvapich variable is, does, and should be set to.

Where's the code?

- Hard to determine actual source for OFED
  - We require the source as we often have changes which are specific to our site
- ”fixes” patches in kernel code
  - catastrophic error recovery was missed
- src.rpm's used instead of and in addition to code in the release tarball
  - ibutils not in our source release
  - wasted time due to local patch not being used

Where's the code? (cont)

- knowledge of build.sh should not be required to get source
- tarball should have source which matches what can be checked out from git on an OFED X.Y branch
- This is better in 1.2 but kernel is still confusing.
- ”The customer is always right” says Matt ;-)

LLNL OFED improvements

- host name written to node description field
- switch-map support in diags
- diag tools
  - saquery
  - iblinkinfo.pl
  - ibqueryerrors.pl
  - etc.
- opensm console (socket and new cmds)

iblinkinfo.pl

```
17:15:48 > iblinkinfo.pl
Switch 0x0008f10400411b18 ""wopr switch" base":
         2    1[  ]  ==( 4X 5.0 Gbps Active/LinkUp)==>       1[  ] "wopri"
         2    2[  ]  ==( 4X 5.0 Gbps Active/LinkUp)==>       1[  ] "wopr0"
         2    3[  ]  ==( 4X 5.0 Gbps Active/LinkUp)==>       1[  ] "wopr1"
         2    4[  ]  ==( 4X 2.5 Gbps Active/LinkUp)==>       1[  ] "wopr2"
         2    5[  ]  ==( 4X 2.5 Gbps Active/LinkUp)==>       1[  ] "wopr3"
              6[  ]  ==( 4X 2.5 Gbps   Down/Disabled)==>        [  ] ""
         2    7[  ]  ==( 4X 5.0 Gbps Active/LinkUp)==>       1[  ] "wopr5"
              8[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
              9[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             10[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             11[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             12[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             13[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             14[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             15[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             16[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             17[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             18[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             19[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             20[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             21[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             22[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             23[  ]  ==( 4X 2.5 Gbps   Down/Polling)==>        [  ] ""
             24[  ]  ==( 4X 2.5 Gbps   Down/Disabled)==>        [  ] ""
```

ibqueryerrors.pl

```
17:15:48 > ibqueryerrors.pl -r
Errors for 0x0008f10400411b18 ""wopr switch" base"
   1: [XmtDiscards == 386] [RcvSwRelayErrors == 290]
         Link info:     2    1[  ] ==( 4X 5.0 Gbps)==>  0x0002c90200219e64    1[  ] "wopri"
   2: [XmtDiscards == 84] [RcvSwRelayErrors == 58]
         Link info:     2    2[  ] ==( 4X 5.0 Gbps)==>  0x0002c90200219ef0    1[  ] "wopr0"
   3: [XmtDiscards == 4] [RcvSwRelayErrors == 196]
         Link info:     2    3[  ] ==( 4X 5.0 Gbps)==>  0x0002c90200228d34    1[  ] "wopr1"
   4: [XmtDiscards == 3] [RcvSwRelayErrors == 18]
         Link info:     2    4[  ] ==( 4X 2.5 Gbps)==>  0x0002c902002227f0    1[  ] "wopr2"
   5: [XmtDiscards == 4] [RcvSwRelayErrors == 17]
         Link info:     2    5[  ] ==( 4X 2.5 Gbps)==>  0x0002c902002265ec    1[  ] "wopr3"
   7: [RcvSwRelayErrors == 45]
         Link info:     2    7[  ] ==( 4X 5.0 Gbps)==>  0x0002c902002268c4    1[  ] "wopr5"
   12: [SymbolErrors == 65535] [LinkDowned == 1] [RcvErrors == 9] [XmtDiscards == 4]
         Link info:     2   12[  ]  ==( 4X 2.5 Gbps)==>     (Disconnected)
   16: [XmtDiscards == 2]
         Link info:     2   16[  ]  ==( 4X 2.5 Gbps)==>     (Disconnected)
   24: [SymbolErrors == 65535] [XmtDiscards == 12]
         Link info:     2   24[  ]  ==( 4X 2.5 Gbps)==>     (Disconnected)
```

OpenSM console

```
OpenSM $ help
Supported commands and syntax:
help [<command>]
quit (not valid in local mode; use ctl-c)
loglevel [<log-level>]
priority [<sm-priority>]
resweep [heavy|light]
status [loop]
logflush -- flush the osm.log file
portstatus [ca|switch|router]
OpenSM $ status
   OpenSM Version       : OpenSM Rev:openib-3.1.0
   SM State/Mgr State : Master/Idle
   SA State             : Ready
   Routing Engine       : updn

   MAD stats
   ---------
   QP0 MADS outstanding           : 0
   QP0 MADS outstanding (on wire) : 0
   QP0 MADS rcvd                  : 198
   QP0 MADS sent                  : 198
   QP0 unicasts sent              : 1
   QP1 MADS outstanding           : 0
   QP1 MADS rcvd                  : 57
   QP1 MADS sent                  : 0
```

<etc>

LLNL local improvements

- LLNL specific tools
  - ibtrackerrors (cron job runs every 4 hours)
  - ibcheckfabric
  - ibnodeinmcast

ibcheckfabric

```
16:48:37 > ibcheckfabric
Collecting port information...
Switch Port Stats:
   9 down port(s)
   2 disabled port(s)
      0x0008f104003f15c2 ""          17[  ]  ==( 4X 2.5 Gbps
Down/Disabled)==>                         [  ] ""
      0x0008f104003f15d9 ""          10[  ]  ==( 4X 2.5 Gbps
Down/Disabled)==>                         [  ] ""
   5760 port(s) at 4X
   5751 port(s) at 2.5 Gbps (SDR)  [Active]
```

ibnodesinmcast

```
17:16:39 > ibnodesinmcast -m 0xc000
1 host(s) up but not in mcast group: wopr4
```

Future

- Interesting improvements in hardware
- OSM code clean up
- Congestion monitoring
- Alternate routing algorithms (See Matt)
- Time stamping each error on the fabric
  - Would allow you find out what was going on when a nodes failures (Performance Manager)

Thanks to

- Hal Rosenstock (Voltaire)
- Sasha Khopyonsky (Voltaire)
- Adam Moody (LLNL)
- Todd Wilde (Mellanox)
- Chris Perreault (Voltaire)
- Appro

VIADEV variables

```
VIADEV_DEFAULT_RETRY_COUNT|=7
VIADEV_DEFAULT_TIME_OUT|=22
VIADEV_NUM_RDMA_BUFFER|=4
VIADEV_ADAPTIVE_RDMA_LIMIT|=2
VIADEV_SQ_SIZE_MAX|=64
VIADEV_DEFAULT_MAX_SG_LIST|=1
VIADEV_MAX_INLINE_SIZE|=80
VIADEV_SRQ_SIZE|=2048
VIADEV_VBUF_TOTAL_SIZE|=9216
VIADEV_VBUF_POOL_SIZE|=512
VIADEV_VBUF_SECONDARY_POOL_SIZE|=128
DISABLE_RDMA_ALLTOALL|=1
DISABLE_RDMA_ALLGATHER|=1
DISABLE_RDMA_BARRIER|=1
```