IB Monitoring

Florent.Parent@calculquebec.ca

IBUG2014

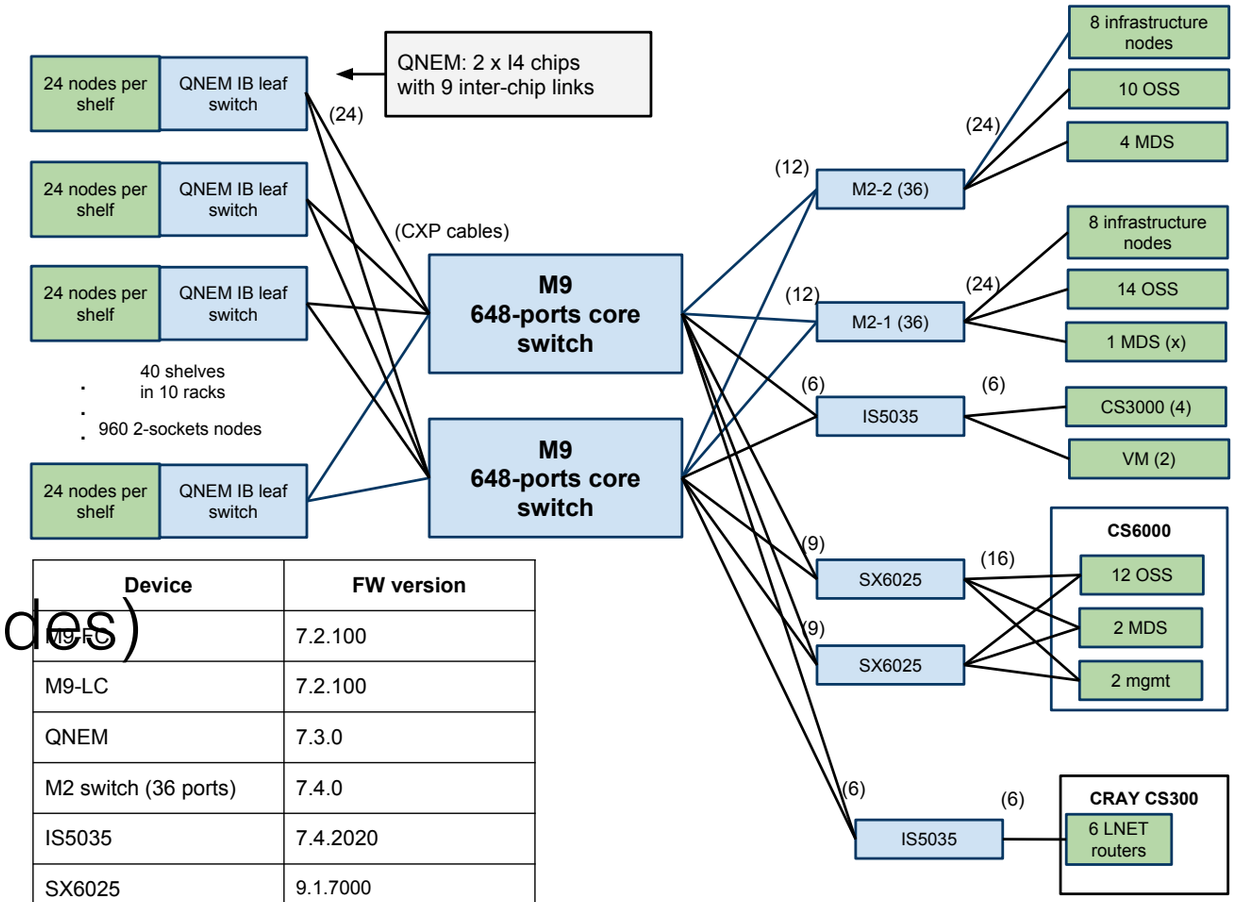
Monterey, CA

Plan

- Recent experiences (adding storage + GPU cluster, OFED/OpenSM upgrade)
- Logging and monitoring IB
- Visualization

Colosse cluster

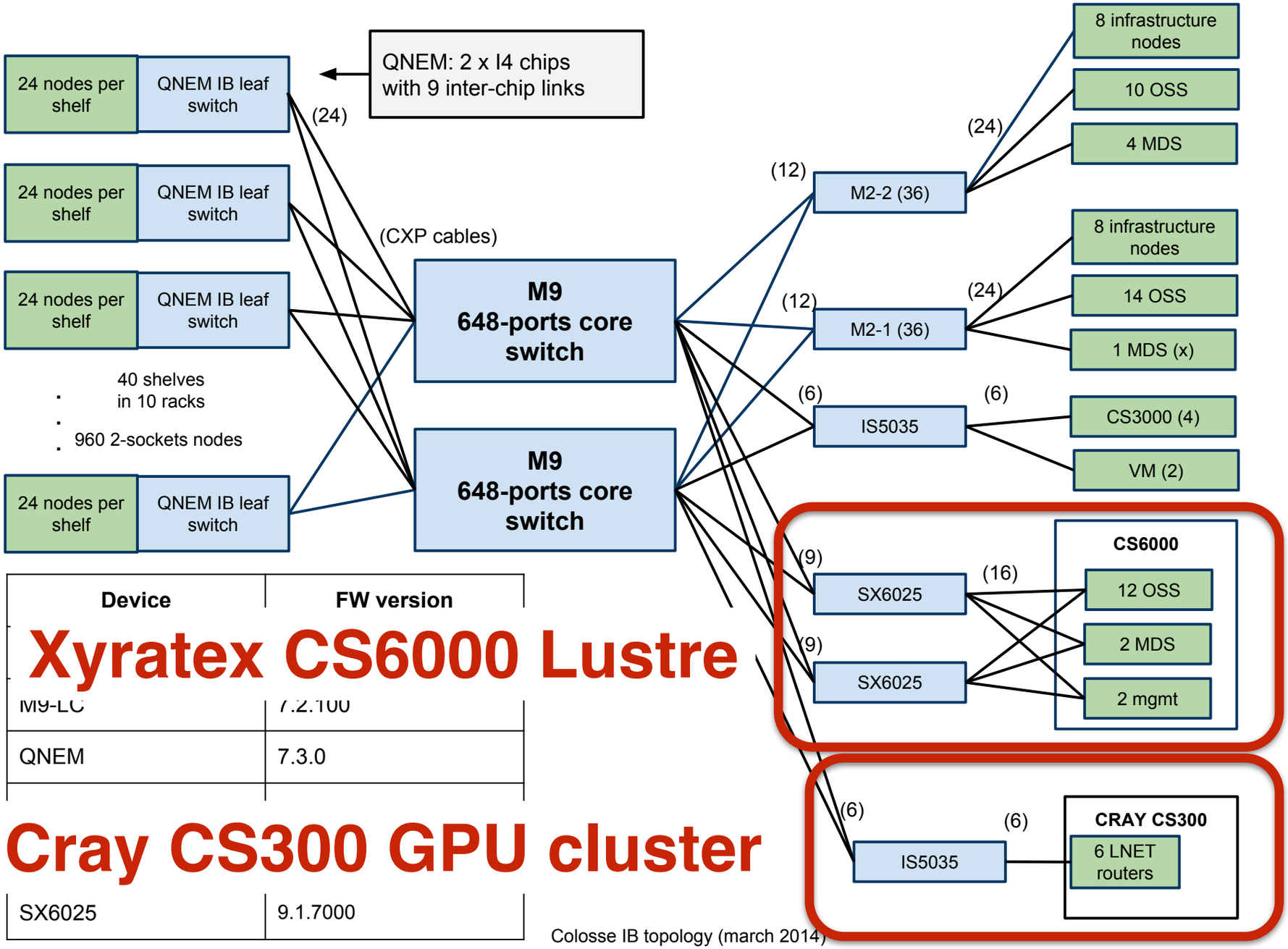
- 960 compute nodes
- 4 Lustre filesystems
- Sun M9 and QNEM QDR switches
- Fat Tree routing (1:1 to compute nodes)
- OpenSM 3.2.6 (until Feb 2014)
- CentOS 5.10
- OFED 1.4.2 -> stock CentOS



Device	FW version
M9-FC	7.2.100
M9-LC	7.2.100
QNEM	7.3.0
M2 switch (36 ports)	7.4.0
IS5035	7.4.2020
SX6025	9.1.7000

Colosse IB topology (march 2014)

New storage and GPU cluster



Device	FW version
M9-LC	7.2.100
QNEM	7.3.0
SX6025	9.1.7000

Xyratex CS6000 Lustre — **Jan ‘14**

Cray CS300 GPU cluster — **Feb ‘14**

Colosse IB topology (march 2014)

Adding new equipment

- Have been holding back against any major changes
  - "If it works, don't fix it", right?
- But adding new storage, GPUs
  - More recent OFED, FDR switches, firmware.
- Do inventory. Plan necessary updates

inventory

node	libibverb	OFED	MLX FW ver
compute	1.1.3	1.5.1	2.7.8100
colosse1	1.1.3	1.5.1	2.7.000
colosse2	1.1.3	1.5.1	2.6.000
MDS lustre1	1.1.7	3.5-2	2.9.1000
OSS lustre1	1.1.7	3.5-2	2.9.1000
MDS lustre2	1.1.6	1.5.4.1	2.6.000
OSS lustre2	1.1.6	1.5.4.1	2.7.000
mgmt1	1.1.2	1.4.2	2.6.000
mgmt2	1.1.2	1.4.2	2.6.000
beast	1.1.3	1.5.1	2.6.000
karma	1.1.3	1.5.1	2.6.000
moabdev	1.1.6	3.5	
analog	1.1.6	3.5	
polaris	1.1.7	3.5-2	2.6.200
redlotus	1.1.7	3.5-2	2.6.000
boot1-5	1.1.3	1.5.1	2.6.200

NOTE:
Have not found an obvious method to figure out the OFED version installed

OpenSM 3.2.6 to 3.3.17

- Motivation: OpenSM getting really out of sync from all other cluster components.

- Initially used stock OpenSM 3.3.15 from CentOS 6.5, but had multiple opensm crashes

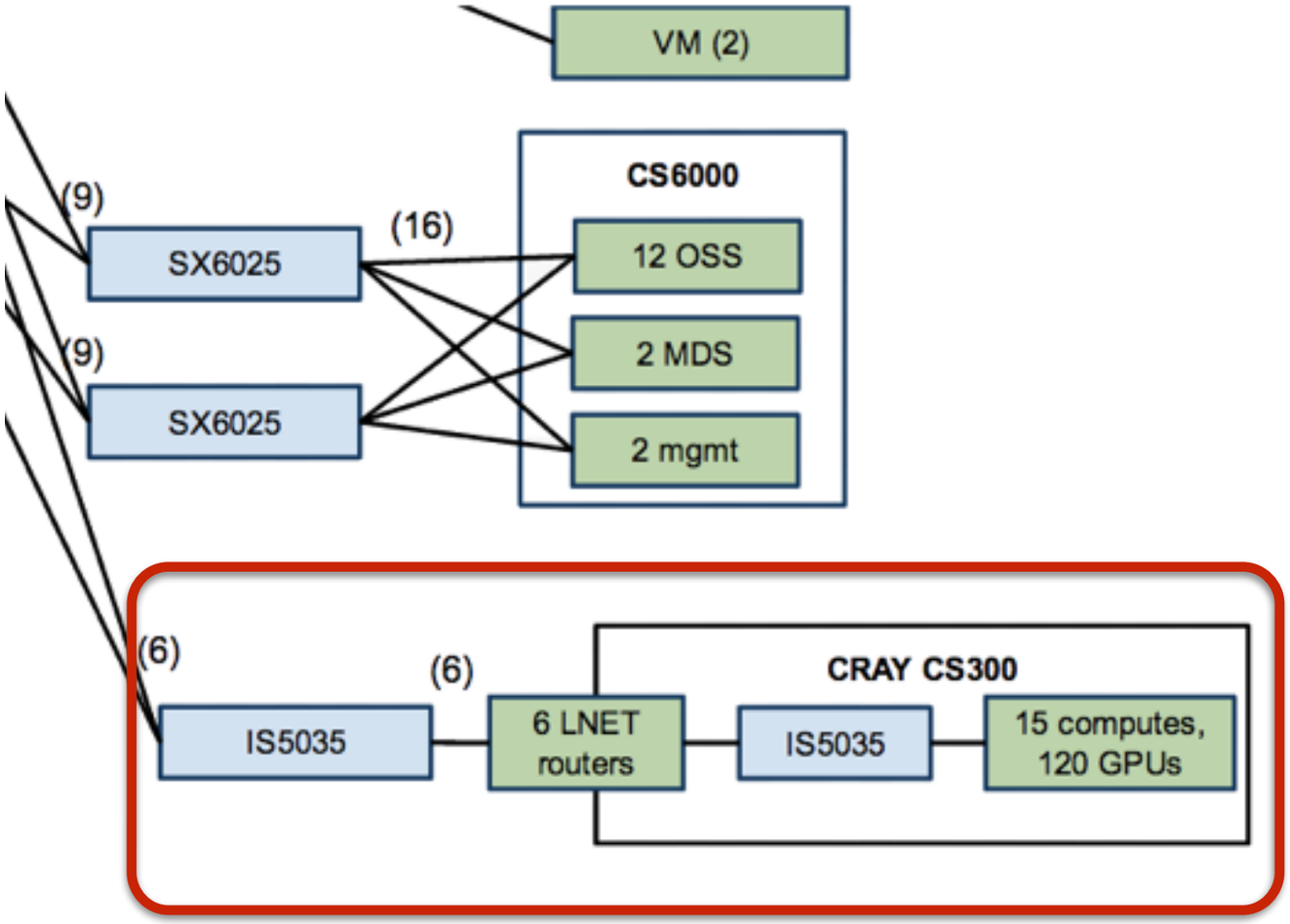
- 3.3.17 has been good so far.

  - Needed “use_mfttop FALSE” to decrease log verbosity (QNEM switch using old MLX firmware)

- Updated to stock CentOS 5 and 6 RDMA stack

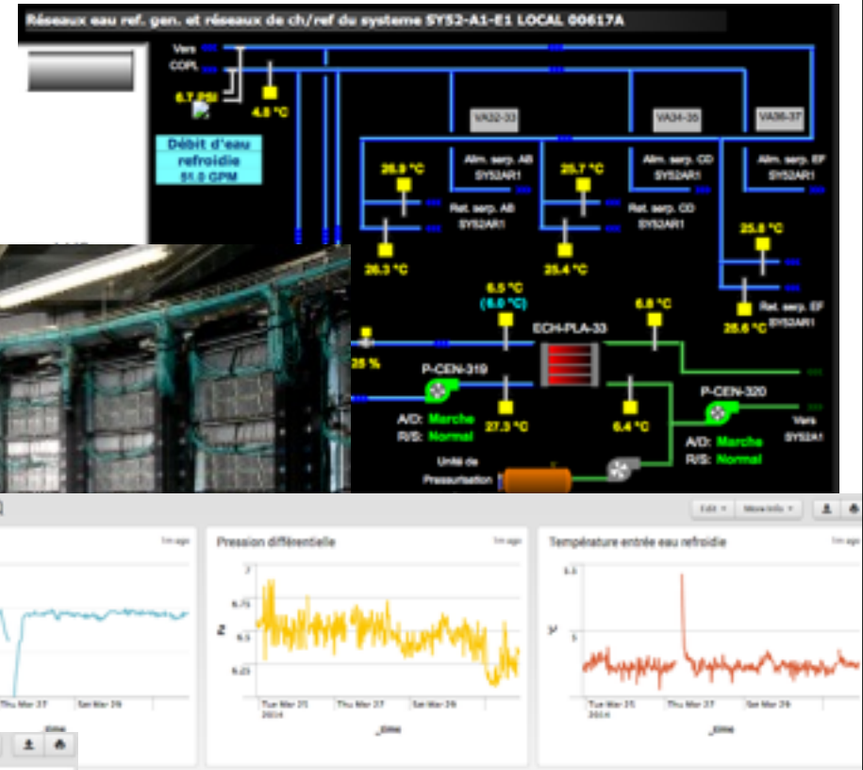
LNET routers

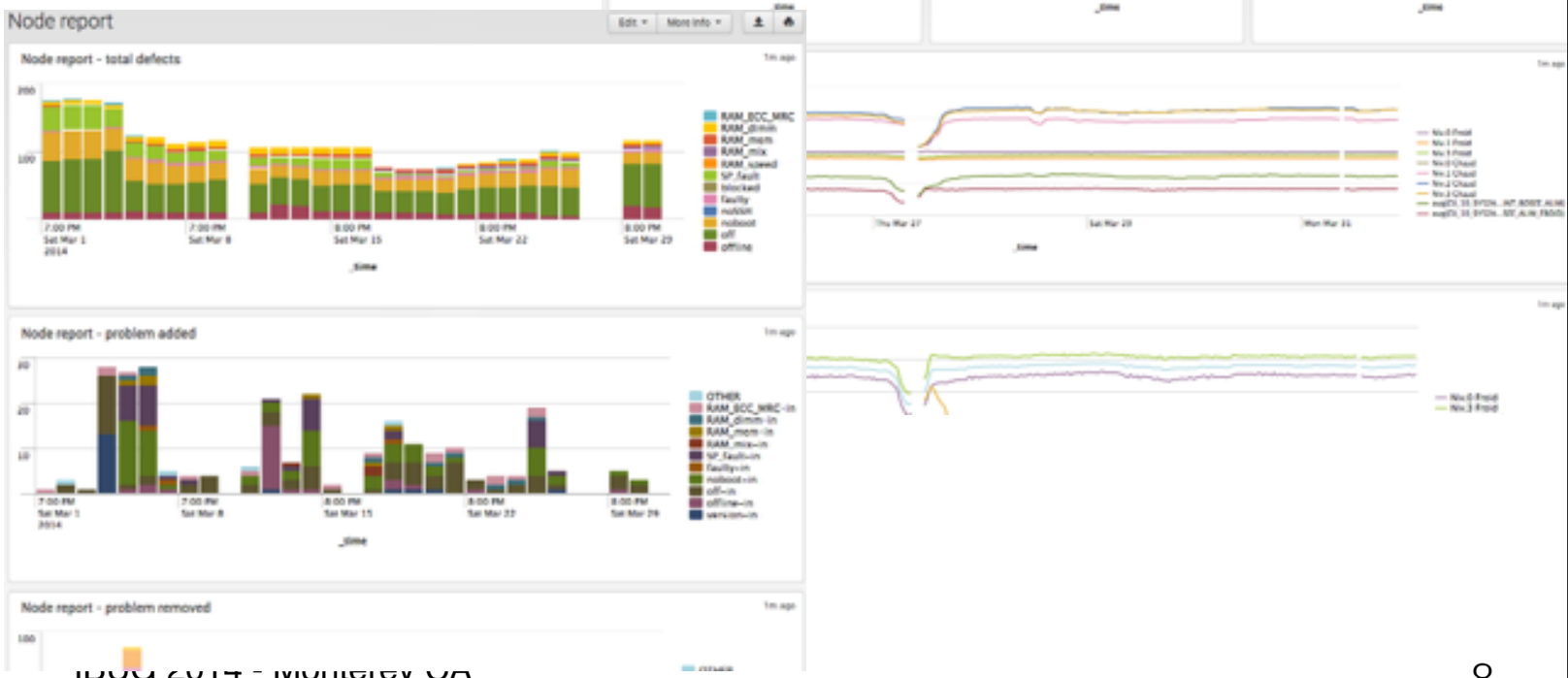
- Motivation for LNET routers:
  - Cray CS300 system
  - Cray “ACE” used for cluster management: integrates an opensm service
  - Preferred to isolate CS300 IB network from existing Colosse IB network
  - Need to mount Lustre filesystems (home, scratch, etc.)



Logging and monitoring

- All logs sent to central server
  - Power
  - Cooling/environmentals
  - Network (ACL, iptables)
  - Scheduler events
  - Node health
  - Infiniband counters
- A lot of unstructured data
- SPLUNK used





Monitoring IB counters

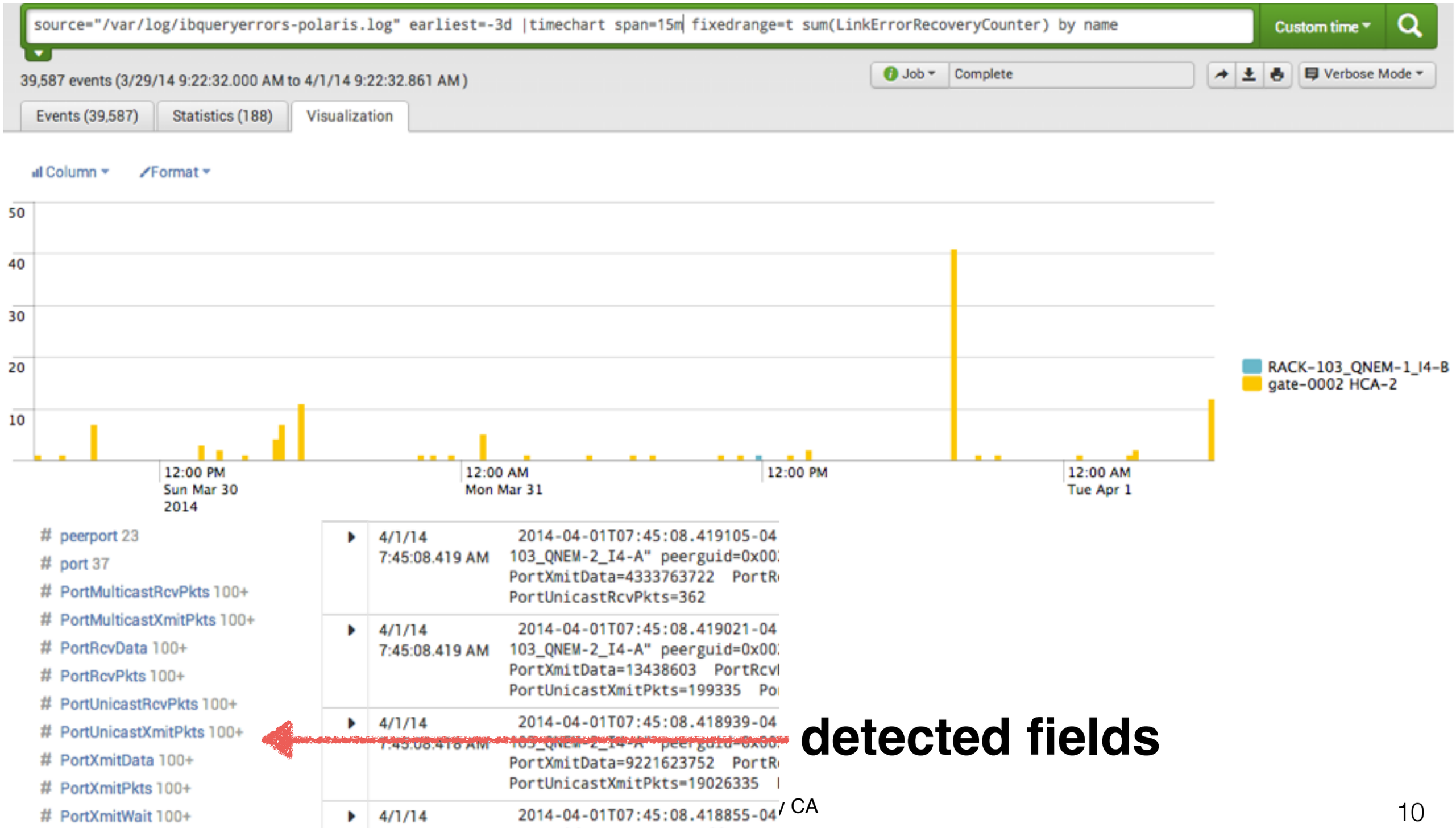
- Using LLNL “ibtrackerror” script (ibqueryerrors)
  - Crontab. Every 15 minutes
- ibtrackerrors -c => Script => syslog

```
Errors for "r106-n4"
   GUID 0x50800200008d6619 port 1: [SymbolErrorCounter == 24869] [PortRcvErrors == 11] [PortXmitData == 360
(1.406KB)] [PortRcvData == 27556 (107.641KB)] [PortXmitPkts == 5 (5.000)] [PortRcvPkts == 648 (648.000)]
       Link info:  626  1[  ] ==( 4X  2.5 Gbps Active/  LinkUp)==>  0x0021283a83c20040  13 18[  ]
"RACK-106_QNEM-1_I4-A" ( Could be 10.0 Gbps)
```

script to group into key/value pairs

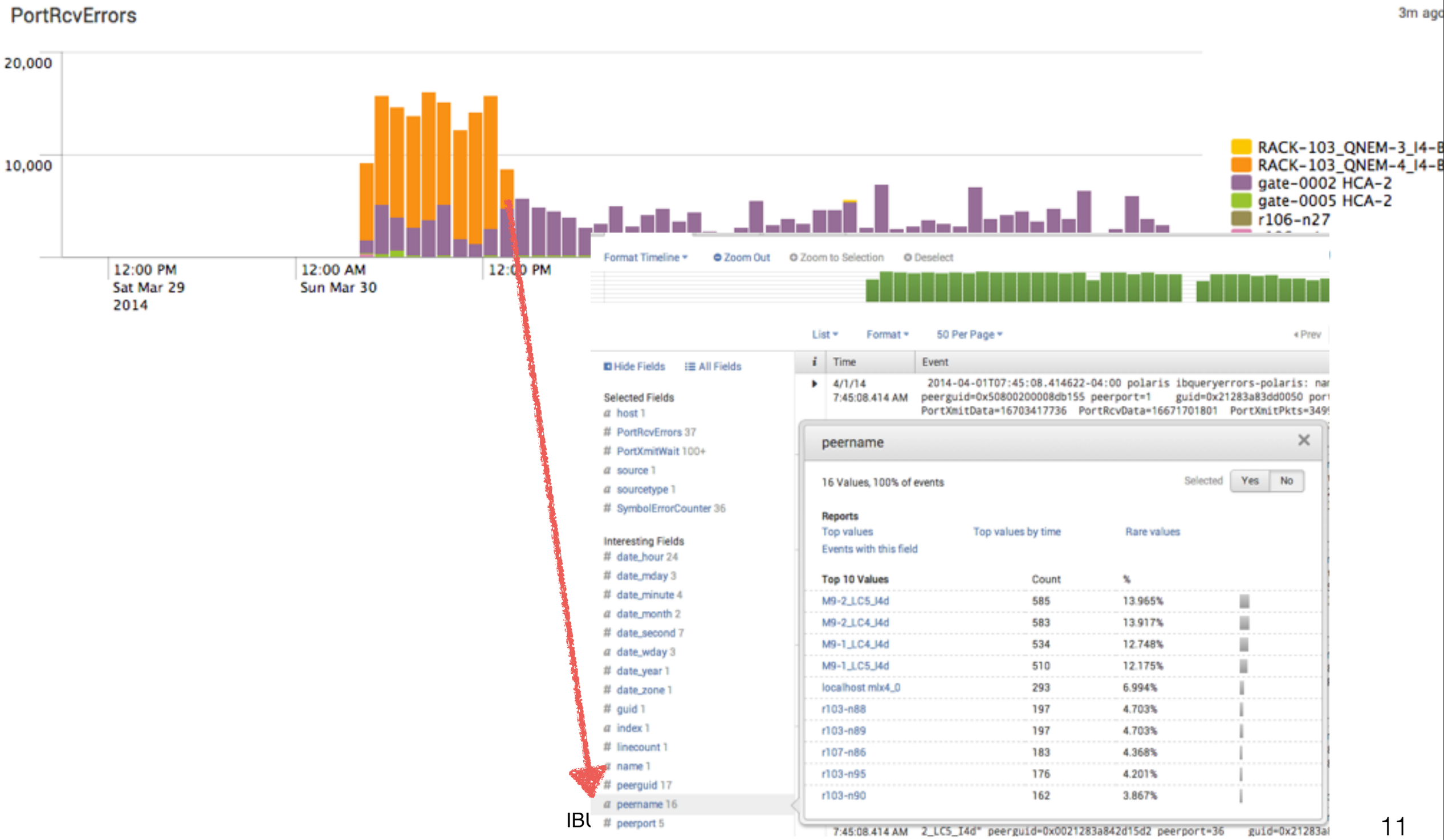
```
2014-03-31T21:45:04.820156-04:00 polaris ibqueryerrors-polaris: name="r106-n4"  peername="RACK-106_QNEM-1_I4-A"
peerguid=0x0021283a83c20040 peerport=18    guid=0x50800200008d6619 port=1  SymbolErrorCounter=24869
PortRcvErrors=11 PortXmitData=360  PortRcvData=27556  PortXmitPkts=5  PortRcvPkts=648
```

SPLUNK Queries

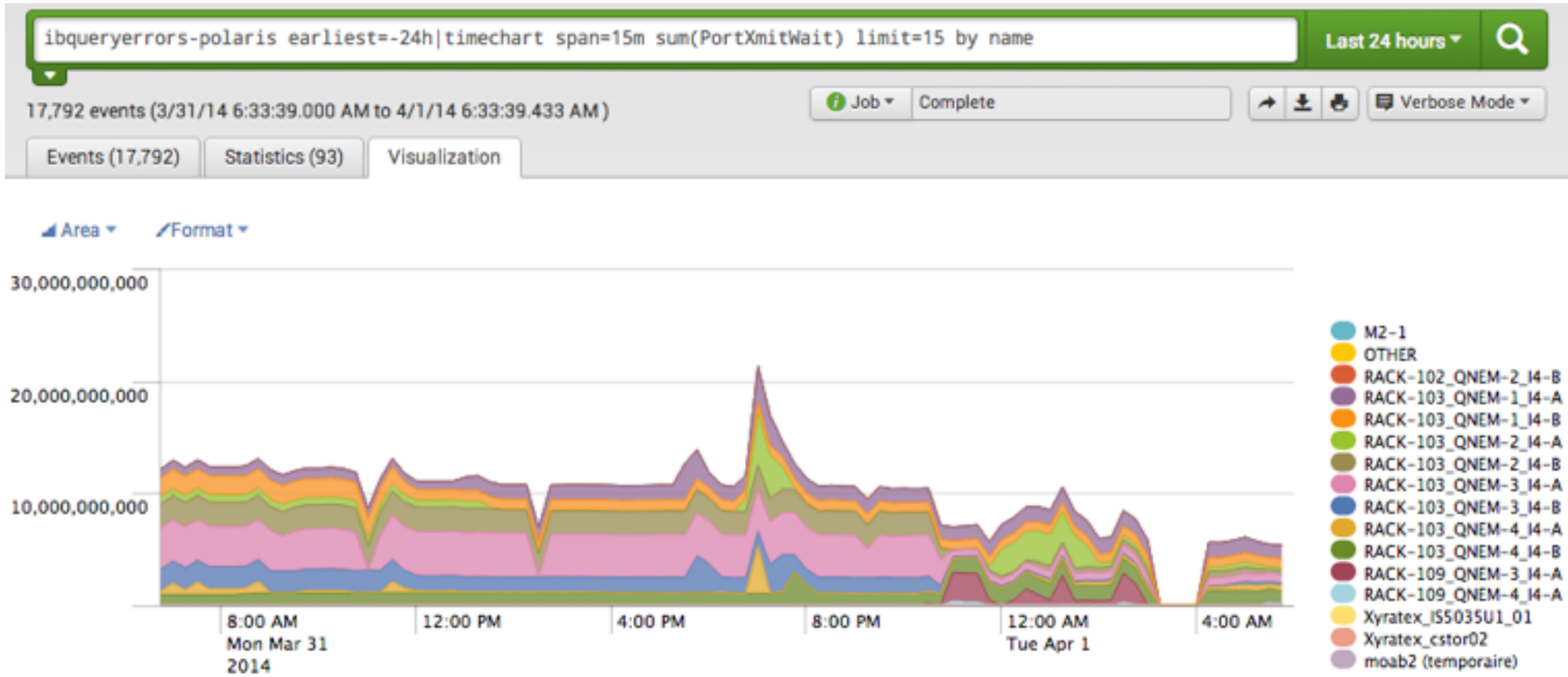


detected fields

IB error counters monitoring



PortXmitWait



- Did not see this counter on our previous OFED 1.4.2 install.
- Starting to monitor this. Need to correlate with other information
  - PortXmitDiscards (severe congestion?)
  - Running jobs (scheduler events)

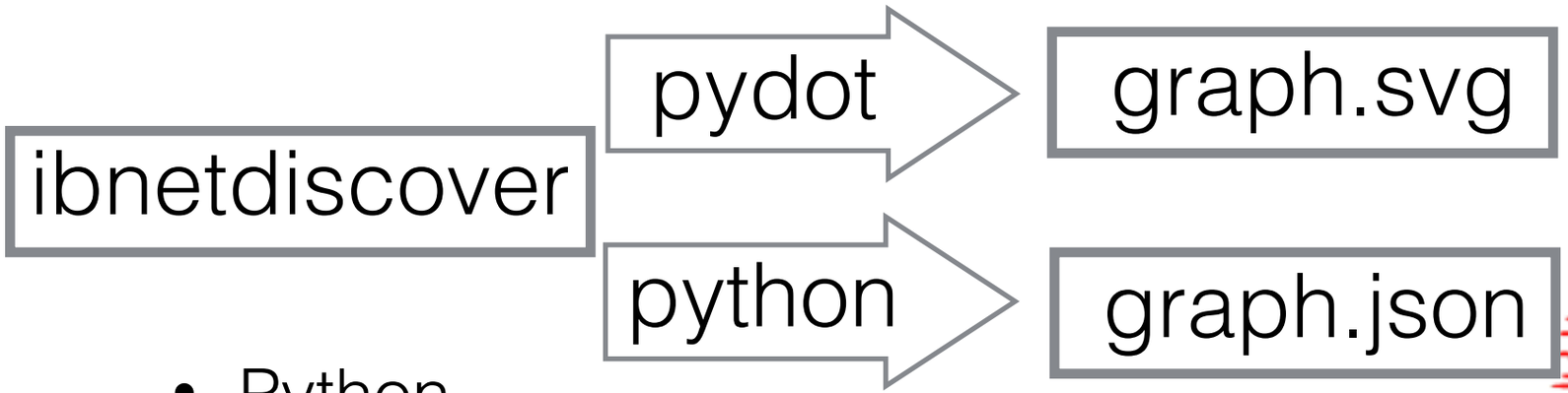
Visualization

- IBUG2013: Blake Caldwell@ORNL presented work in progress with pydot.

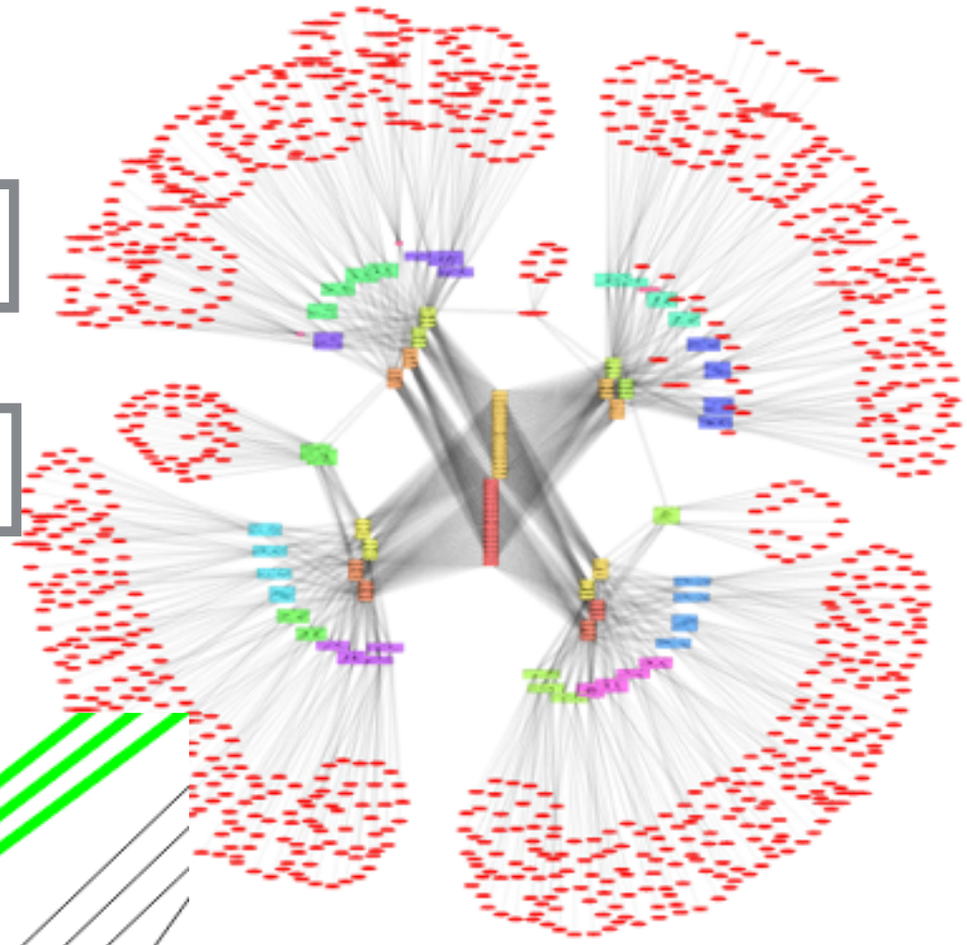
- Student internship project (François Blackburn)

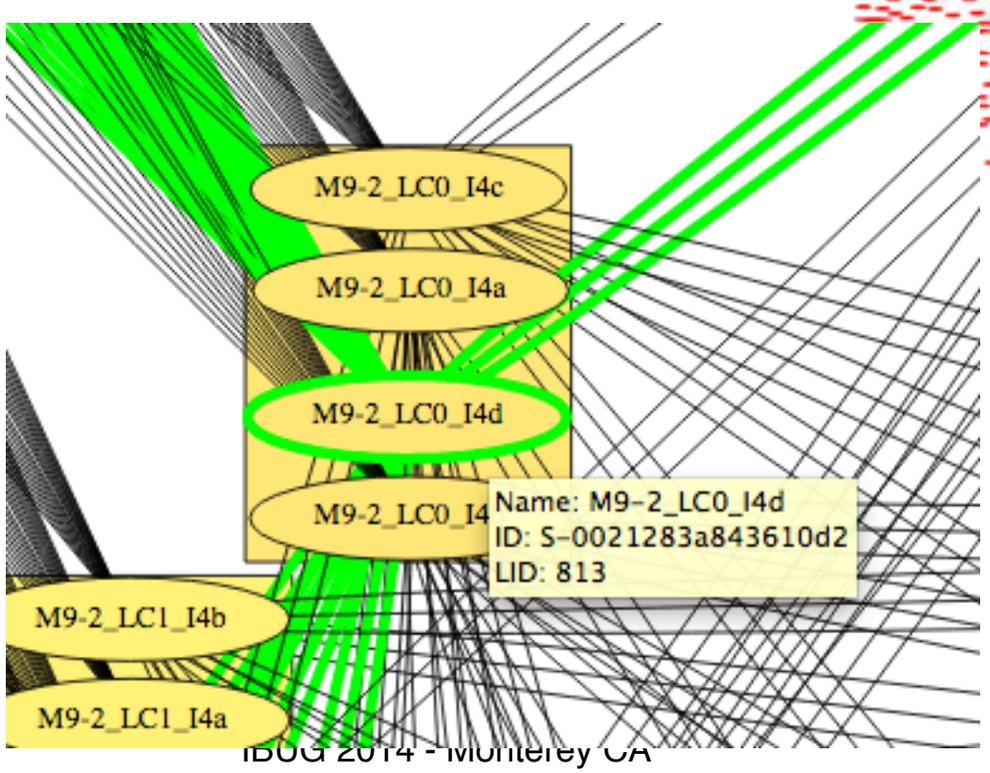
- Explored d3js and graphviz/pydot. Decided on pydot (more control on layout) and JQuery SVG (interactivity)

ibnetdiscover -> pydot

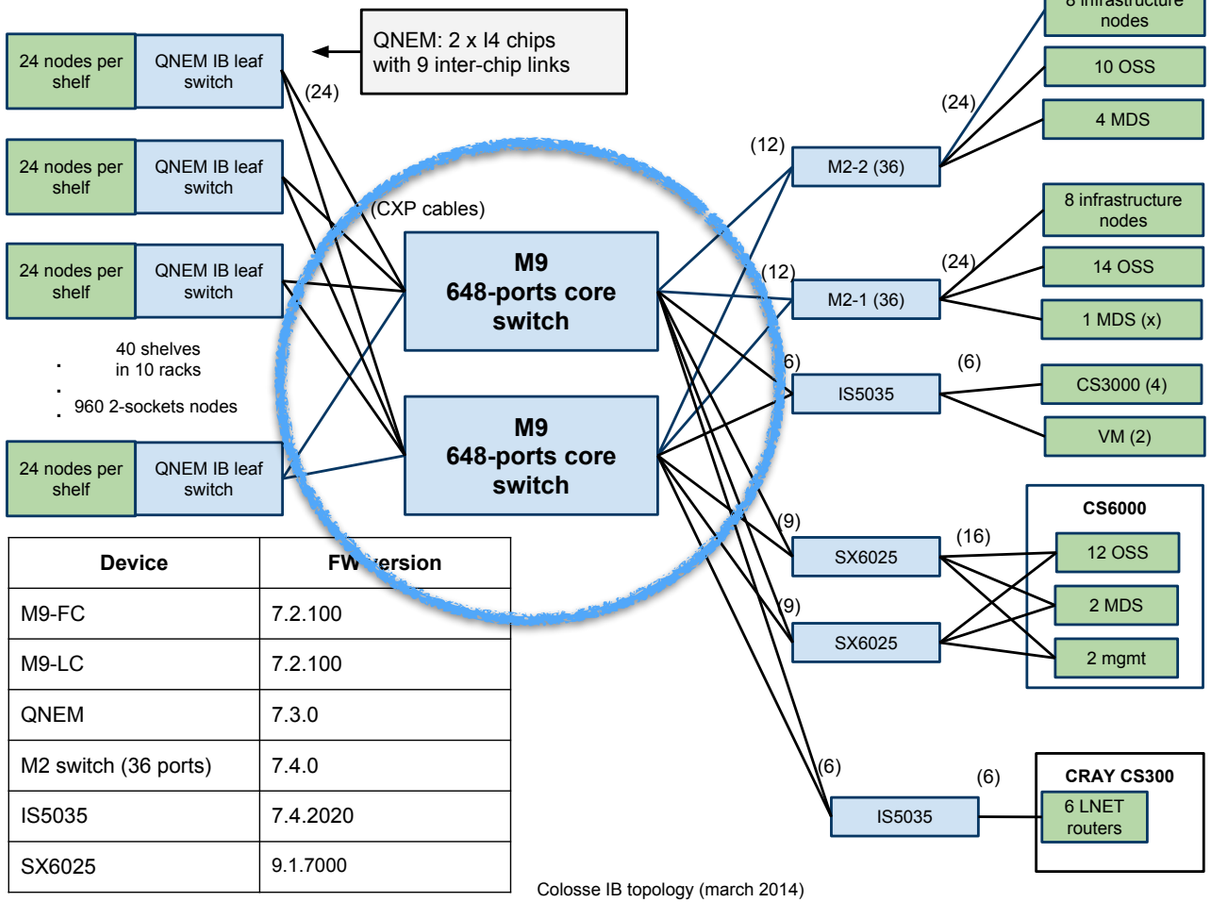


- Python
- pydot
- JQuery SVG



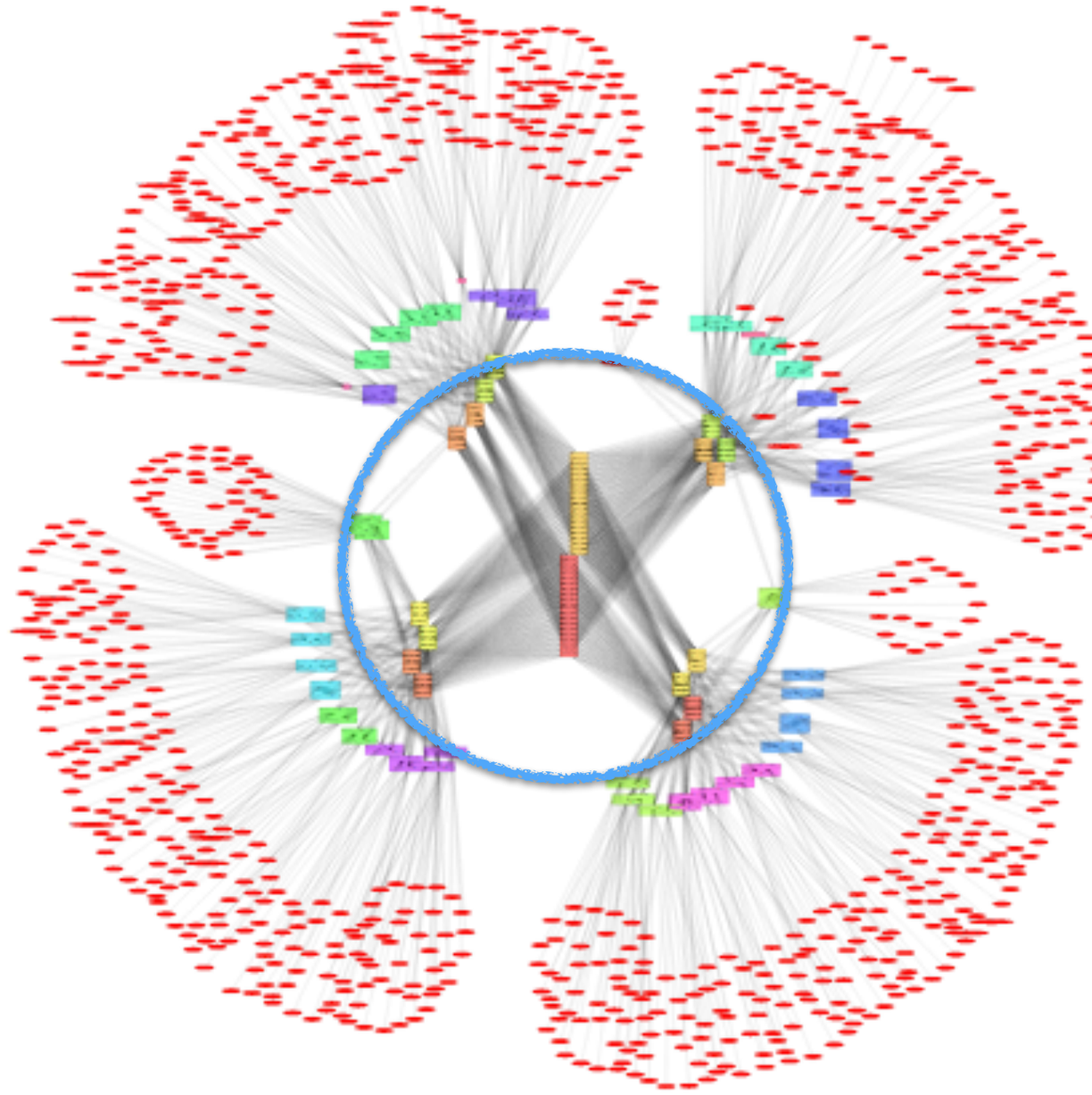


Detailled view of all links



Device	FW version
M9-FC	7.2.100
M9-LC	7.2.100
QNEM	7.3.0
M2 switch (36 ports)	7.4.0
IS5035	7.4.2020
SX6025	9.1.7000

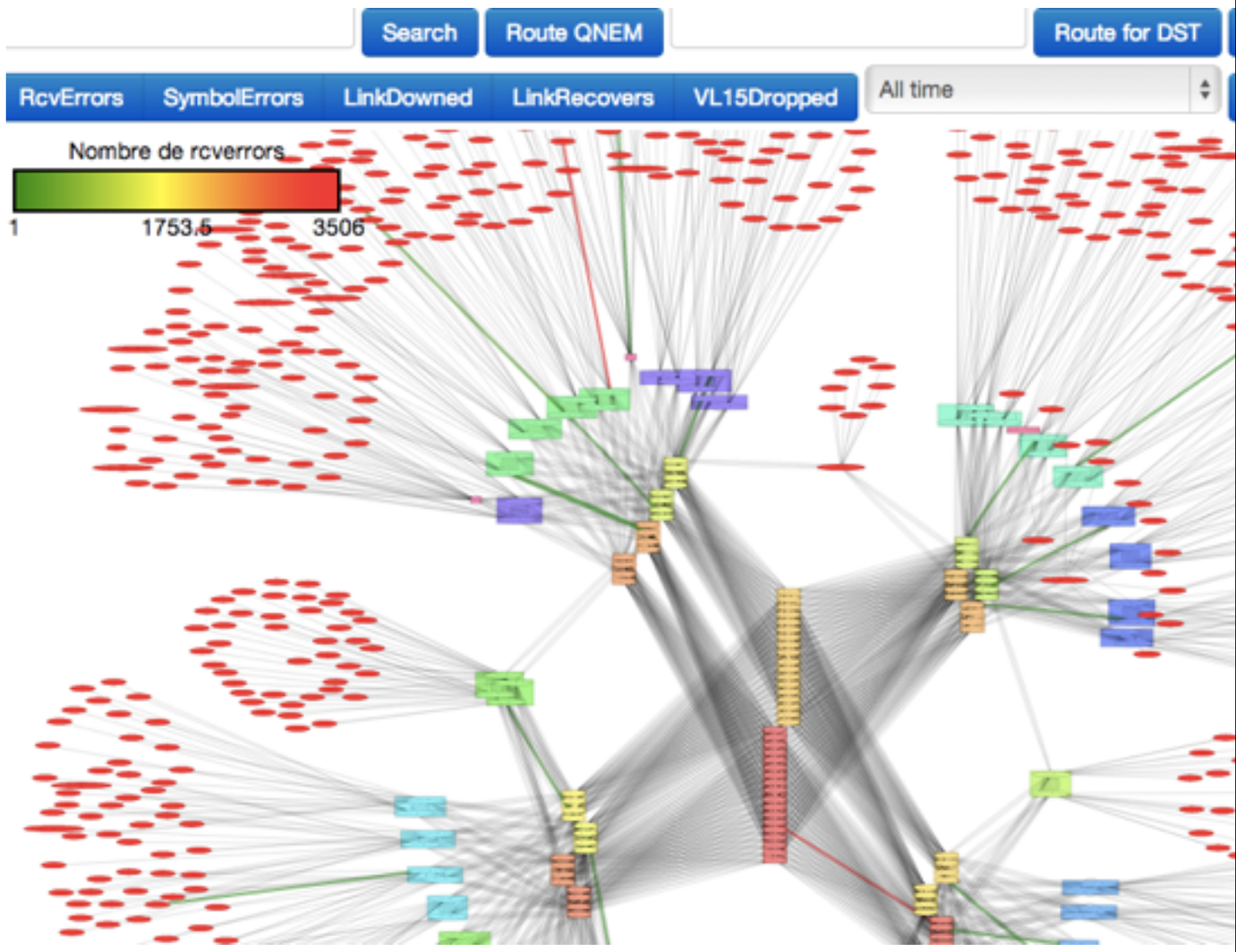
Colosse IB topology (march 2014)



Overlaying error counters values

ibqueryerrors → python → iberrs.json

- Highlight links with specific errors reported
- Useful to see the overall distribution of errors.



To investigate

- Xmit and Rcv counters from all switches
  - How often can we query counters?
    - OpenSM perf. manager?
- Better opensm.log monitoring (syslog?)
- Correlate XmitWait with Discards and job info
- Visualization: Add LFTs to explore routing?

Questions?