14th ANNUAL WORKSHOP 2018

PROACTIVE IDENTIFICATION AND REMEDIATION OF HPC NETWORK SUBSYSTEM FAILURES

Susan Coulter, HPC-Design / Networking

Los Alamos National Laboratory

LA-UR-18-22950

[April 13, 2018]

THE LIFE OF AN HPC NETWORK ADMIN

“ It’s always the network, until it’s not the network. ”

Susan Coulter

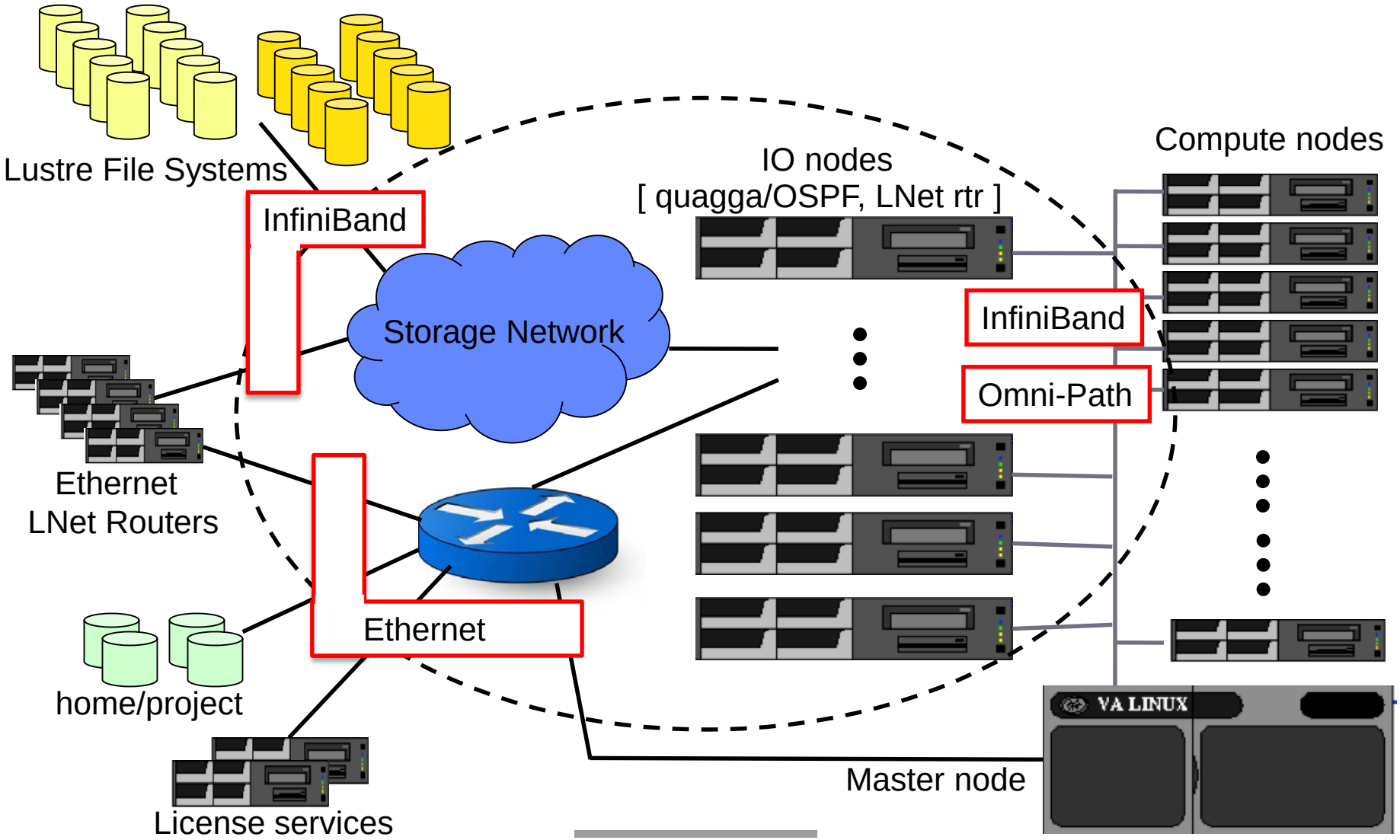
“ Lustre is slow from grizzly – it must be the network. “

“ My job is running 3 times slower this week compared to last week – it must be the network. “

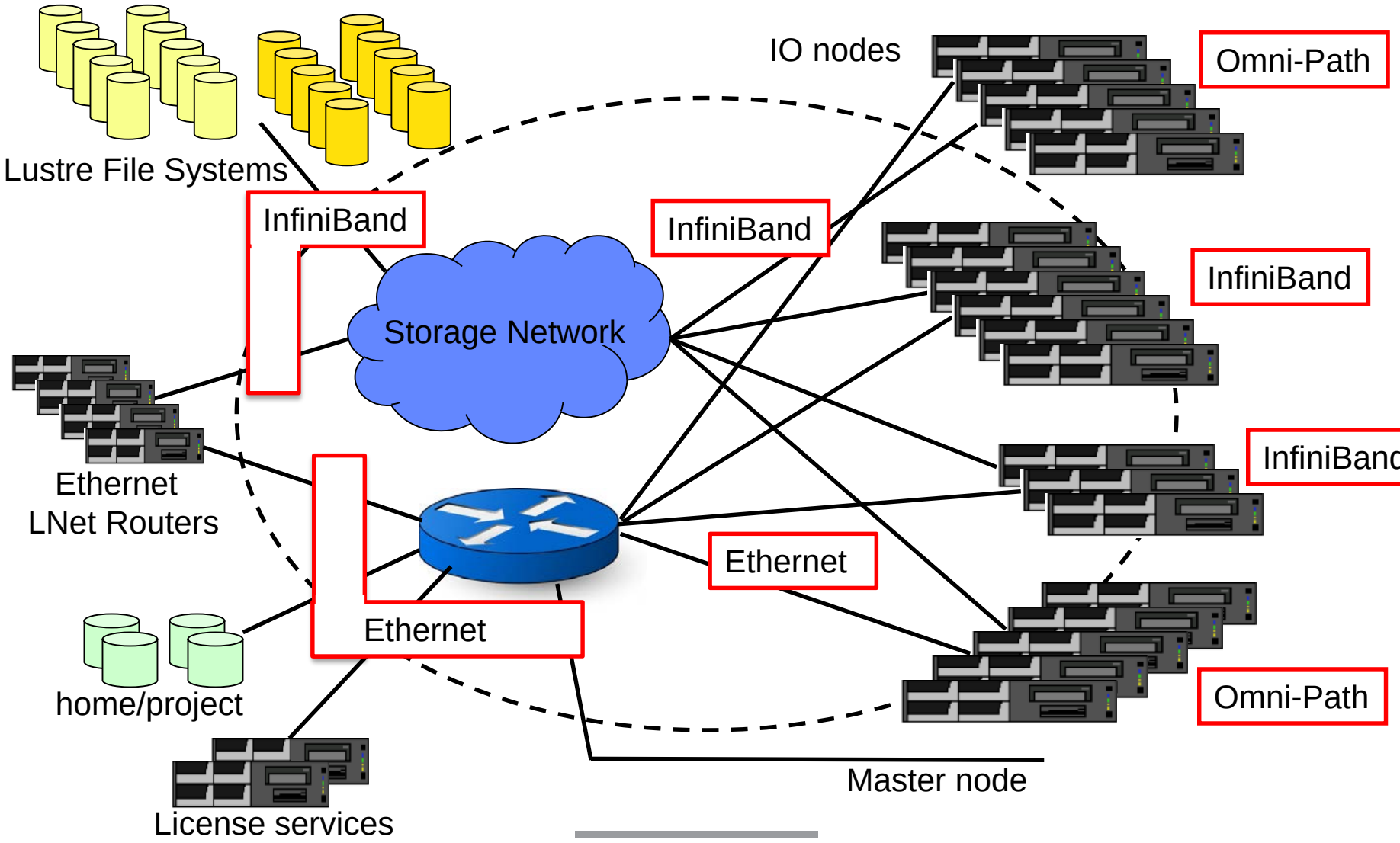
“ NFS mounts to home/project keep dropping and coming back – it must be the network. “

“ My chickens aren’t laying this month – it must be the network. “

TYPICAL LANL HPC IO SUBSYSTEM



TYPICAL LANL HPC IO SUBSYSTEM



PRIMARY LANL SOLUTION - CONCEPT

- **DeadGatewayDetection (DGD)**
  - Monitor the entire IO subsytem
  - Proactively remediate/alleviate network problems when possible
  - Simulate typical network access patterns
  - Be stateful
  - Report status
  - Allow administrators control of the process
  - Allow administrators easy access to status/information

LANL SOLUTION - CRITERIA

- **DeadGatewayDetection (DGD)**
  - Define critical and/or weak points
    - Monitor those points
    - Define tests to be executed
    - Set thresholds for taking action
  - Be transparent to running jobs
  - Portable to all clusters
  - Configuration file driven
  - Use and/or tie into standard logging/monitoring processes
  - Allow repair/replacement of faulty IO nodes without perturbation

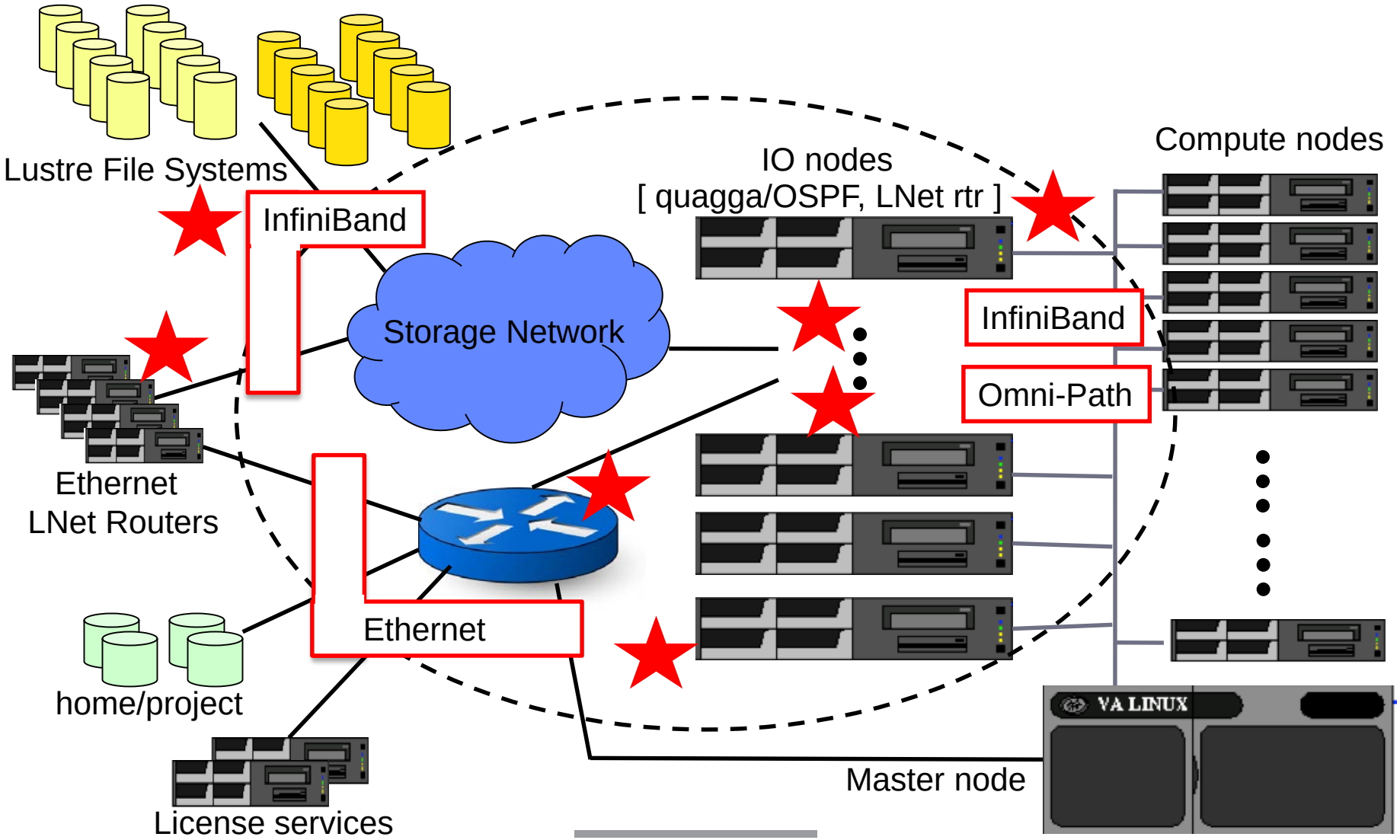
DGD – FUNCTIONAL OVERVIEW

- **Written in Perl**
- **Runs as a daemon on the master node**
- **Read configuration file**
- **Discover and initialize environment**
  - Arrays containing compute node and IO node information
  - Hash table of health status for each test, and each IO node
- **while(0) {**
  - Execute tests
  - Check results
    - Implement mitigations as necessary
  - Sleep
  - Handle signals as appropriate
- **}**

DGD – CURRENT TESTS

- **Functionality of fabric NIC on all IO nodes**
  - Internal to cluster
- **Functionality of ethernet NIC/Bond on all IO nodes**
  - External to cluster
- **Status of OSPFD on IO node (quagga)**
- **Ability to reach ethernet gateway on campus backbone**
- **Identify pertinent messages from IO node syslog**
- **Connectivity of LNet Routers**
- **Functionality of secondary ethernet NICs**
  - Optional
- **Ability to reach secondary ethernet gateway**
  - Optional

PHYSICAL REPRESENTATION OF TEST POINTS



DGD – FUTURE TESTS

- **Results of netstat and/or other status commands**
- **Results of tests launched from IO nodes**
- **Identify pertinent messages from IO dmesg**
- **State of LNet Routers**
  - Correct NID list

DGD – REMEDIATION

- **Compute node ethernet routes modified**
  - Faulty IO node is removed from the ethernet routes
- **LNet Router shut down**
- **OSPFD shut down**

DGD – MULTIPLE FAILURES

- **Thresholds for each cluster**
  - Driven by cluster size and Lustre FGR groups
  - Critical messages
  - Insanity levels

i	Time	Event
>	3/26/18 12:09:03.000 PM	<14>Mar 26 12:09:03 ls-master DeadGatewayDetection[3298]: check_for_state_change: CRITICAL - /etc/INSANE created , Mitigation undone, testing paused, please check the cluster host = ls-master source = tcp:2514 sourcetype = syslog
>	3/26/18 12:08:45.000 PM	<14>Mar 26 12:08:45 ls-master DeadGatewayDetection[3298]: check_for_state_change: CRITICAL - Sanity threshold of 2 exceeded host = ls-master source = tcp:2514 sourcetype = syslog

ADMIN CONTROL – SIGNAL HANDLING

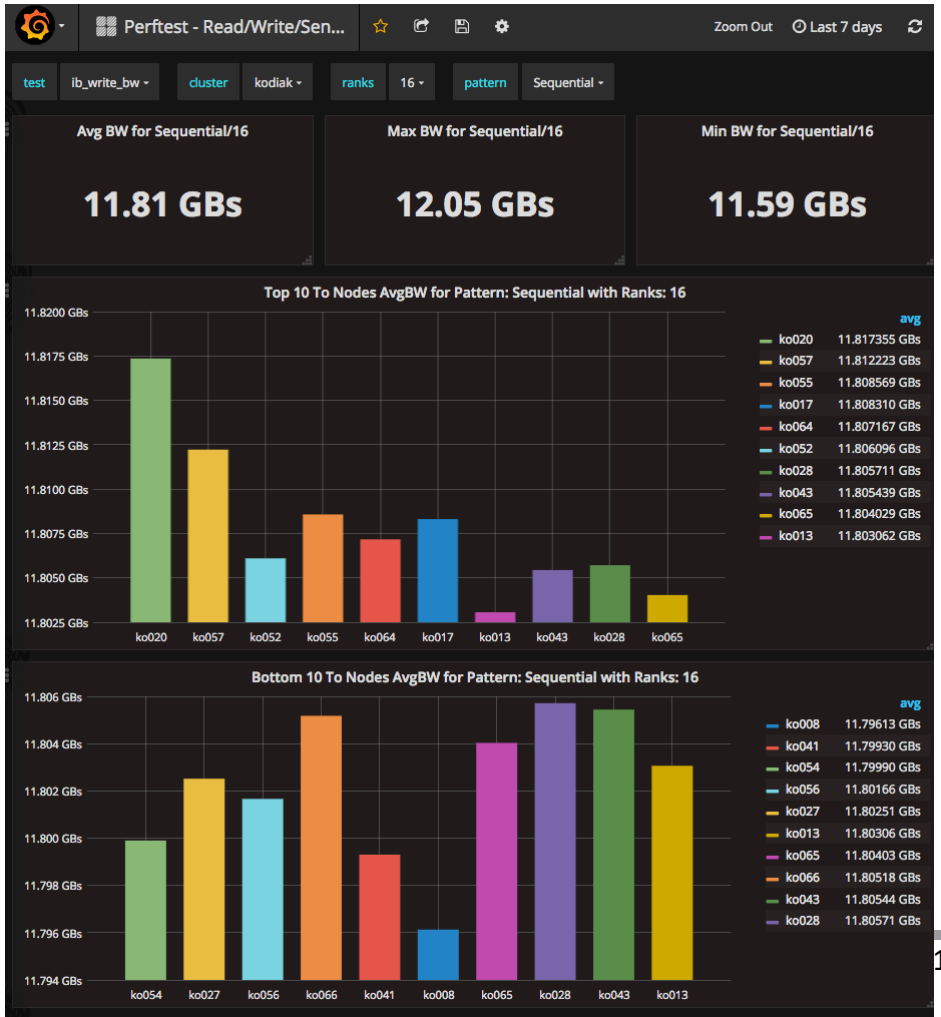
- **start, stop, restart**
- **status**
  - Dumps the current state of the health arrays within DGD
- **wakeup**
  - Only honored when the process is in the sleep portion of the loop
  - Used to minimize the time between an IO node being fixed, and DGD confirming it passes all tests
- **suspend**
  - Used to provide more time to resolve the issue if close to a solution
- **reload**
  - Re-reads the configuration file
    - DEBUG, IOLOG_FILENAME, IOLOG_MSGS, MAX_DEAD_IOS, MAX_FAIL, MAX_PARTIAL_FAIL, PING_SIZE, SANITY_CHECK, SKIP_FILENAME, SLEEP

ODDS & ENDS

- **Built as an RPM**
  - RHEL6 and RHEL7
- **Available on github**
  - Source code
  - Test script
  - spec files
- **Uses randomly selected compute nodes for some tests**
- **Uses arrays of IPs for access to IO and compute nodes**
- **Has SSH timeouts in case nodes are in a wonky state**

ADDITIONAL LANL SOLUTIONS

- **Performance – Baseline performance in the real world**
  - Daily cron jobs using PerfTest via slurm and OpenMPI / IB and OPA

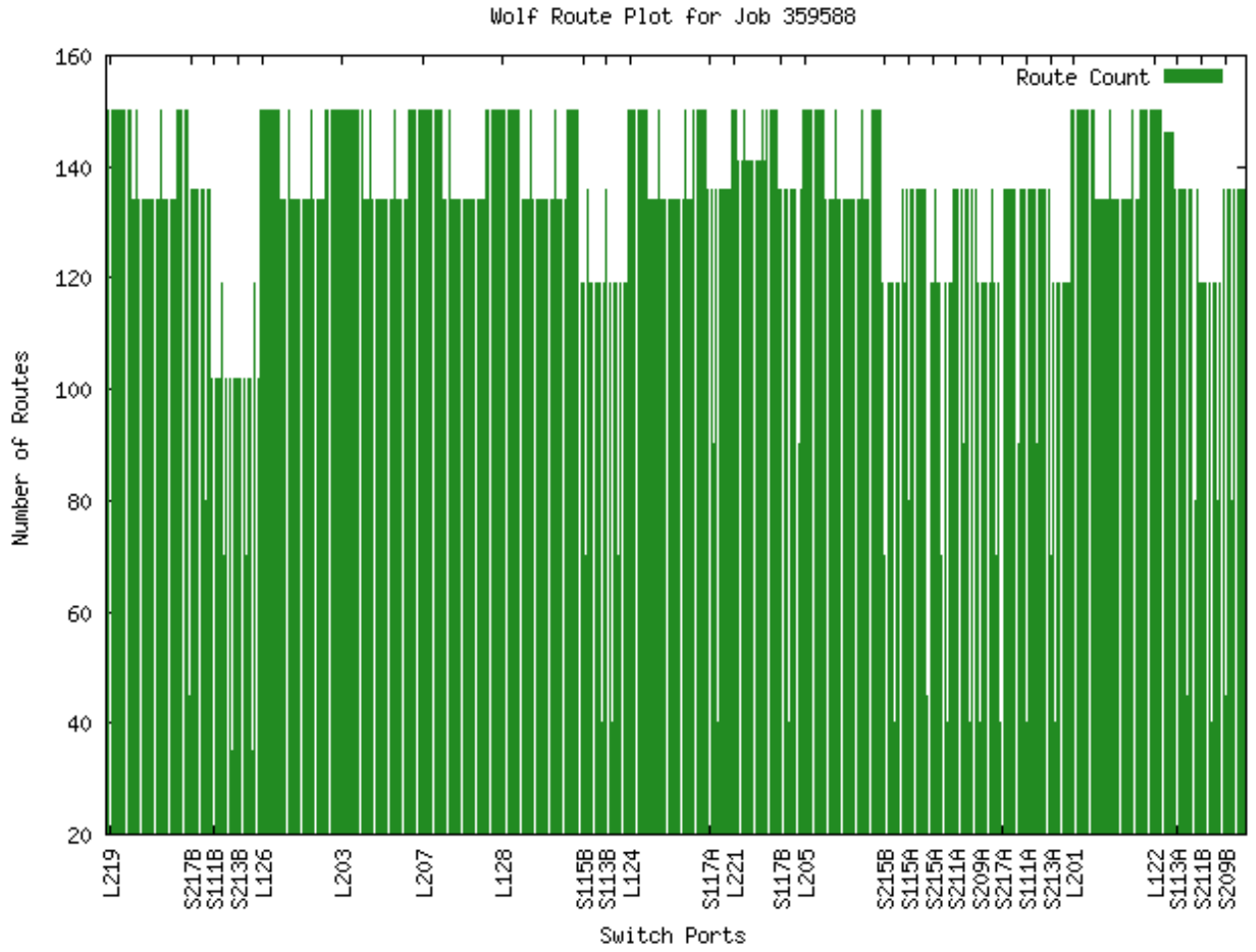


ADDITIONAL LANL SOLUTIONS

- **Performance – Baseline performance in the real world**
  - Daily cron jobs using PerfTest via slurm and OpenMPI / IB and OPA

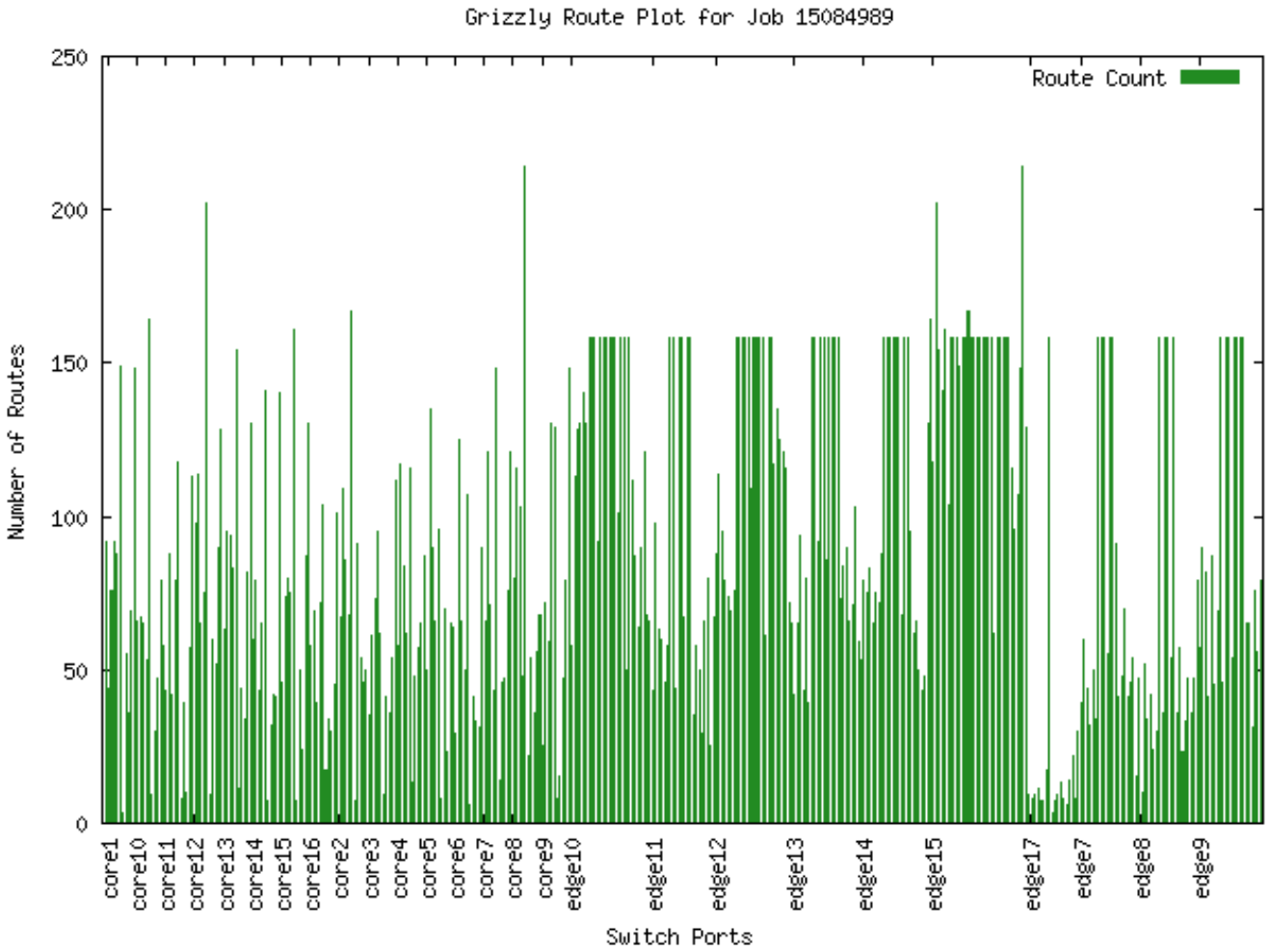
ADDITIONAL LANL SOLUTIONS

- **Routing – Create route maps to illustrate switch port usage**
  - Bash scripts: slurm input, routing tools, create gnuplot / IB and OPA



ADDITIONAL LANL SOLUTIONS

- **Routing – Create route maps to illustrate switch port usage**
  - Bash scripts: slurm input, routing tools, create gnuplot / IB and OPA



14th ANNUAL WORKSHOP 2018

THANK YOU

Susan Coulter, HPC-Design / Networking

github.com/skcoulter

Los Alamos National Laboratory