OPENFABRICS ALLIANCE

Infiniband Monitoring of HPC Clusters using LDMS (Lightweight Distributed Monitoring System)

OpenFabrics Software User Group Workshop

Christopher Beggio

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

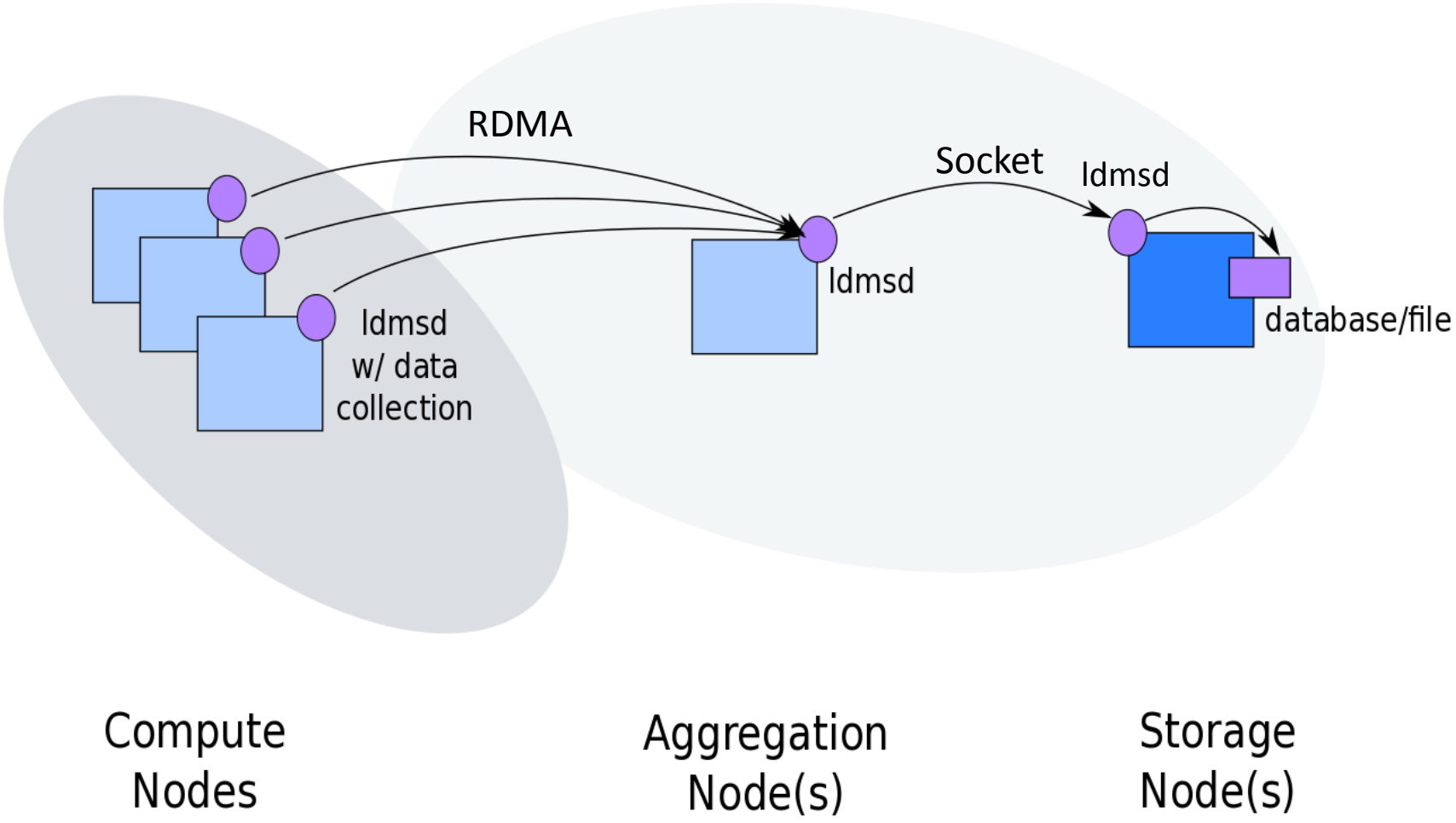
Goals

- Understand how user/applications are utilizing network resources
  - Intra-job communications
  - File system communications
    - PFS
    - NFS
- Understand network congestion
  - Intensity, longevity, extent
  - Drive job scheduling and resource allocation based on:
    - Dynamic state information
    - Knowledge of user/application historic utilization
    - File system locations

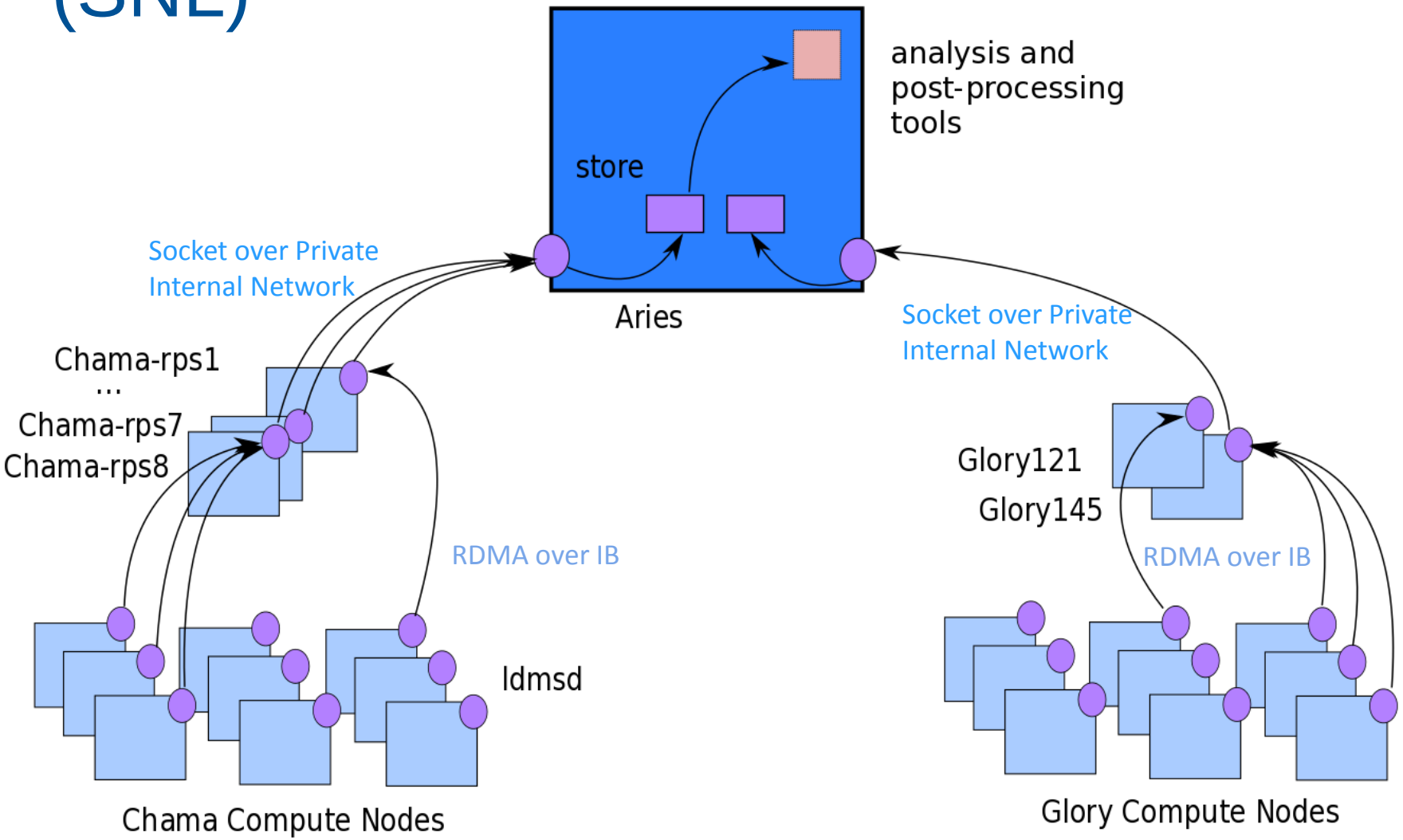
Current Impediments

- Understanding of user/application network needs is occluded by possible congestion both internal and external to application
- Opacity of core switches only lends itself to vague notion of congestion characteristics
- No “snapshot” ability because counter readings aren’t synchronized
- Rapid sampling of core fabric adversely impacts application traffic

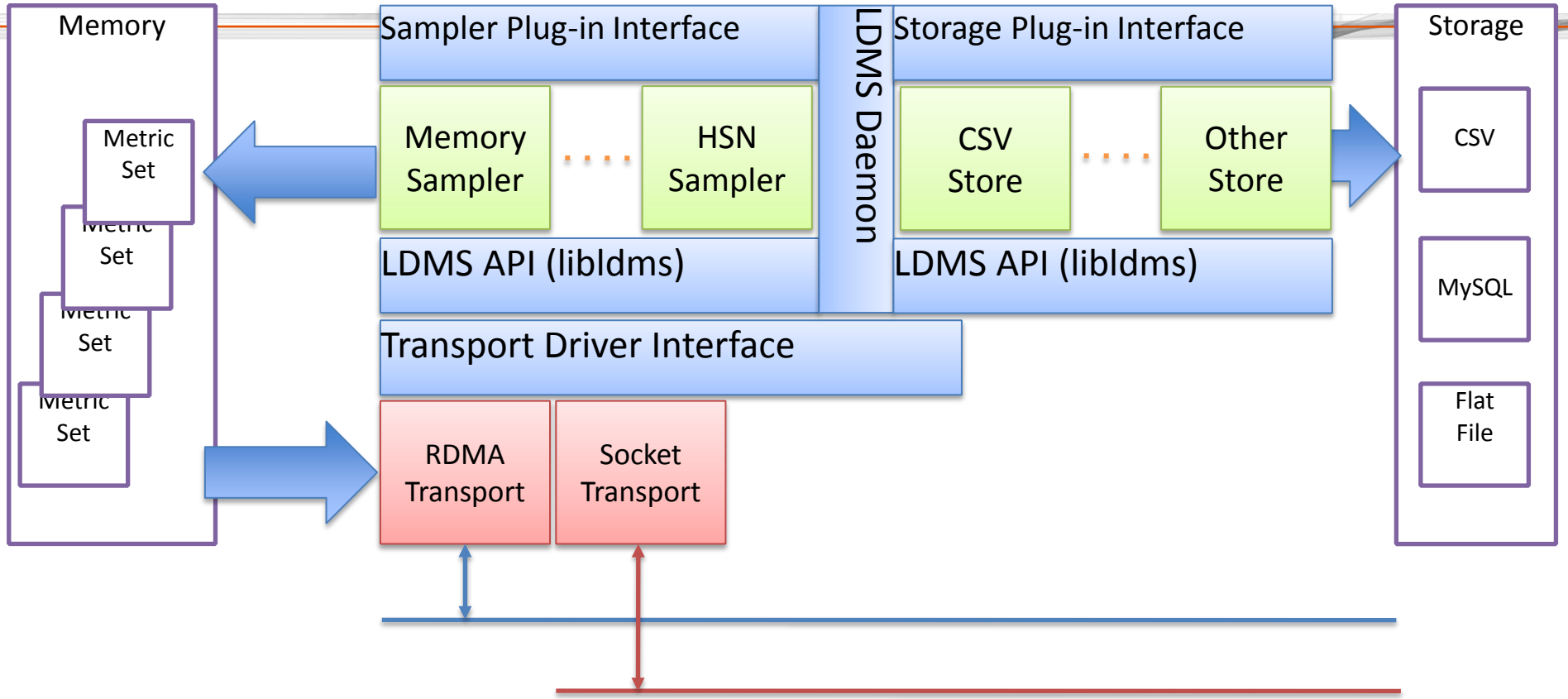
Deployment Configuration: Data Collection, Transport, and Storage



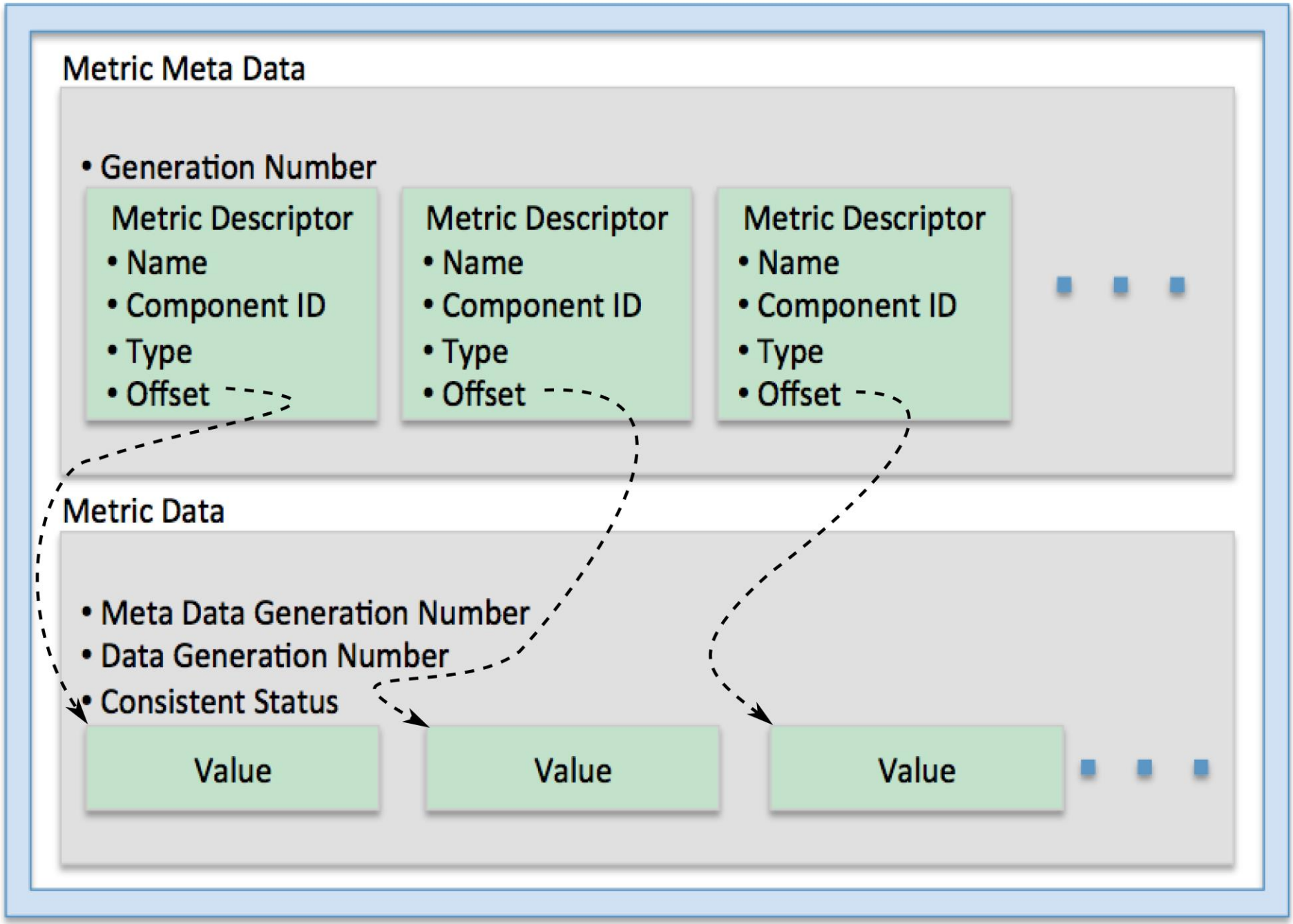
Chama: 1232 node TLCC2 cluster (SNL)



LDMS Architecture



Metric Set Memory



IB Host Based Metric Set

- chama466/sysclassib: consistent, last update: Tue Jan 20 20:36:00 2015 [6129us]
- U64 0 ib.**symbol_error**#qib0.1
- U64 0 ib.**link_error_recovery**#qib0.1
- U64 0 ib.**link_downed**#qib0.1
- U64 0 ib.**port_rcv_errors**#qib0.1
- U64 0 ib.**port_rcv_remote_physical_errors**#qib0.1
- U64 0 ib.**port_rcv_switch_relay_errors**#qib0.1
- U64 0 ib.**port_xmit_discards**#qib0.1
- U64 0 ib.**port_xmit_constraint_errors**#qib0.1
- U64 0 ib.**port_rcv_constraint_errors**#qib0.1
- U64 0 ib.**local_link_integrity_errors**#qib0.1
- U64 0 ib.**excessive_buffer_overrun_errors**#qib0.1
- U64 0 ib.**VL15_dropped**#qib0.1
- U64 5526619018 ib.**port_xmit_data**#qib0.1
- U64 5502039184 ib.**port_rcv_data**#qib0.1
- U64 17515642 ib.**port_xmit_packets**#qib0.1
- U64 17586331 ib.**port_rcv_packets**#qib0.1
- U64 0 ib.**port_xmit_wait**#qib0.1
- U64 3851 ib.**port_unicast_xmit_packets**#qib0.1
- U64 7711 ib.**port_unicast_rcv_packets**#qib0.1
- U64 118 ib.**port_multicast_xmit_packets**#qib0.1
- U64 122831 ib.**port_multicast_rcv_packets**#qib0.1

Host Side Collection
(would also like from every core link)

Errors:
symbol_error
link_error_recovery
port_rcv_errors
port_rcv_remote_physical_errors
port_rcv_switch_relay_errors
port_xmit_constraint_errors
port_rcv_constraint_errors
local_link_integrity_errors
excessive_buffer_overrun_errors

Data Info:
port_xmit_data
port_rcv_data
port_xmit_packets
port_rcv_packets

Misc. Info:
port_xmit_wait
port_xmit_discards
VL15_dropped
link_downed

Other LDMS Metric Set Examples

shuttle-cray.ran.sandia.gov_1/meminfo

- U64 160032 MemFree
- U64 181728 Buffers
- U64 3443332 Cached
- U64 33076 SwapCached
- U64 2987544 Active

shuttle-cray.ran.sandia.gov_1/procstatutil

- U64 1826564 cpu0_user_raw
- U64 699631 cpu0_sys_raw
- U64 663843760 cpu0_idle_raw
- U64 201018 cpu0_iowait_raw

shuttle-cray.ran.sandia.gov_1/vmstat

- U64 40008 nr_free_pages
- U64 122286 nr_interactive_anon
- U64 321902 nr_active_anon
- U64 465532 nr_inactive_file
- U64 424986 nr_active_file

Metric sets:

- (datatype, value, metricname) tuples
- optional per metric user metadata e.g., component id

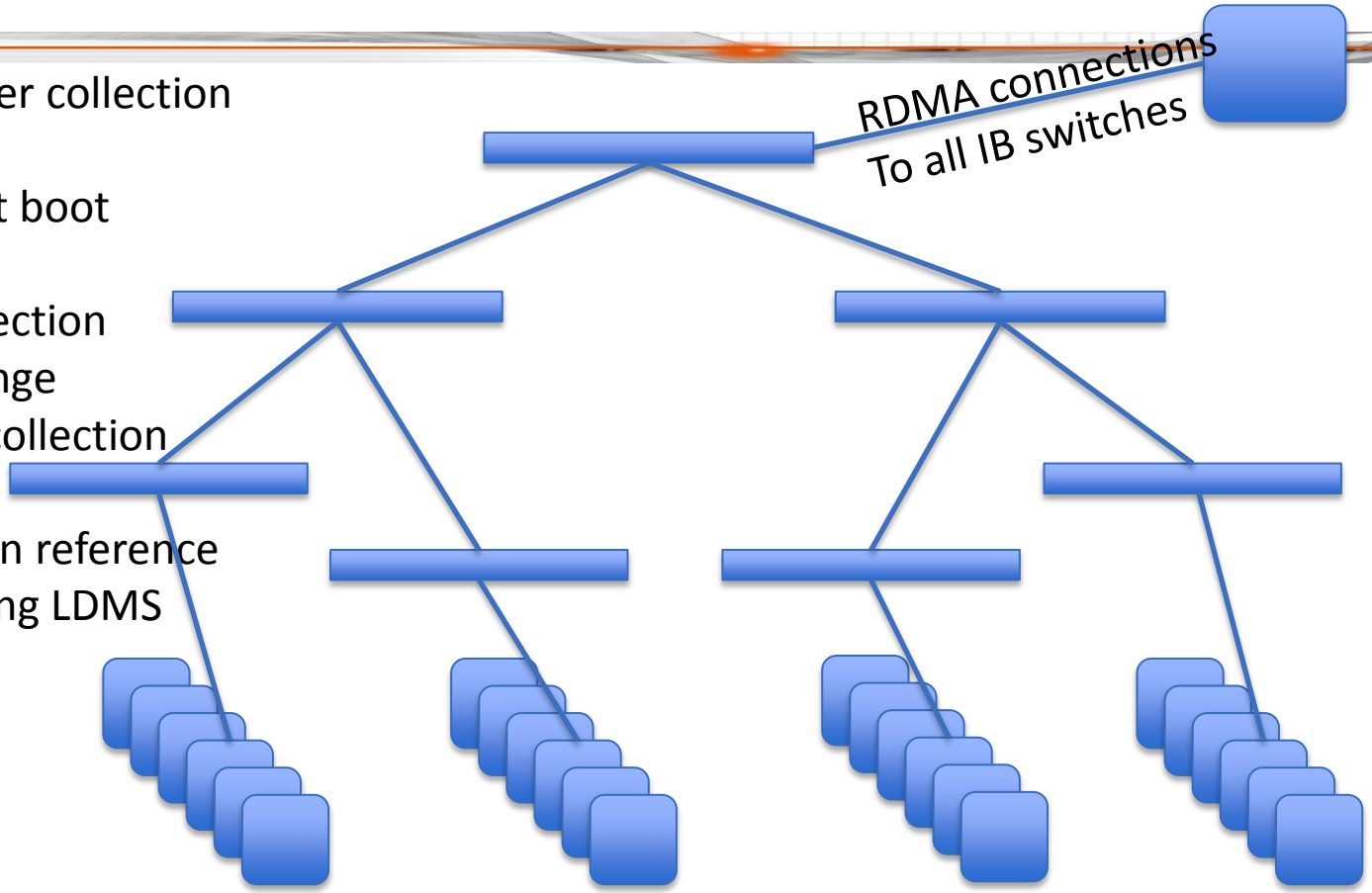
API:

- *ldms_get_set*
- *ldms_get_metric*
- *ldms_get_u64*

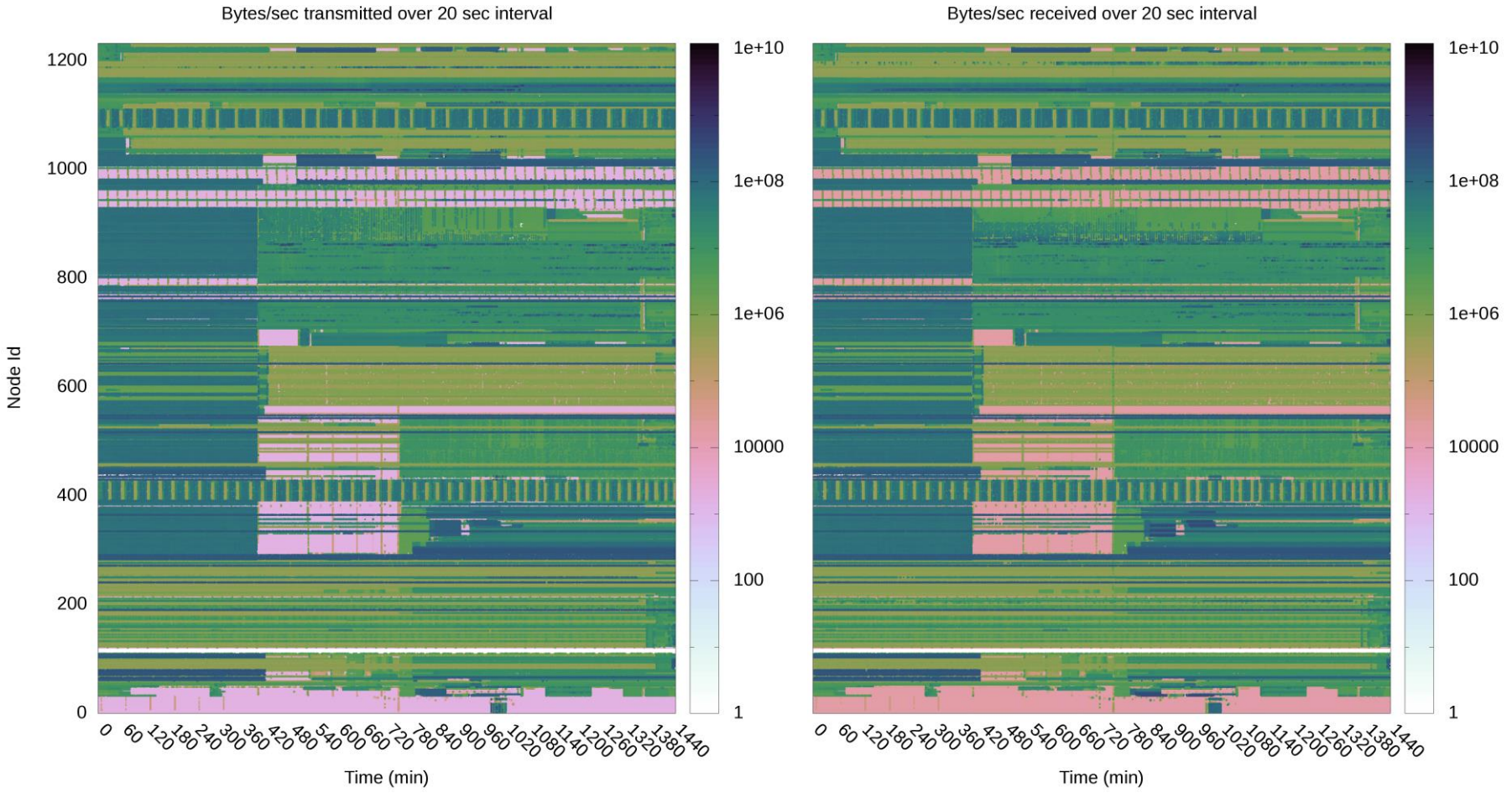
- Same API for on-node and off-node transport

Would Like

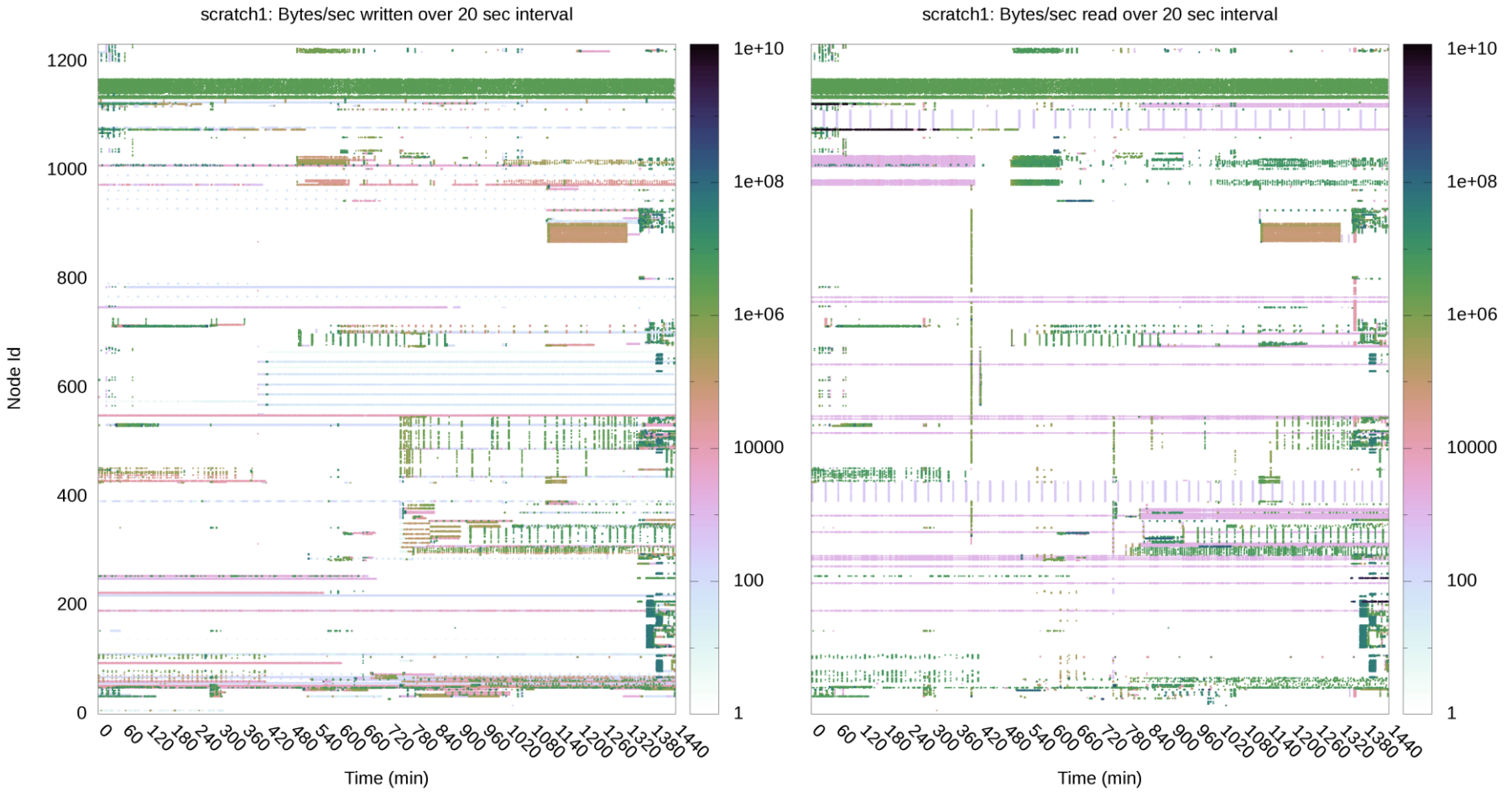
- Synchronized counter collection across all switches
  - Configurable at boot
  - Fixed set
  - On-the-fly collection frequency change
- Offload IB counter collection from computes
- Currently working on reference implementation using LDMS protocol



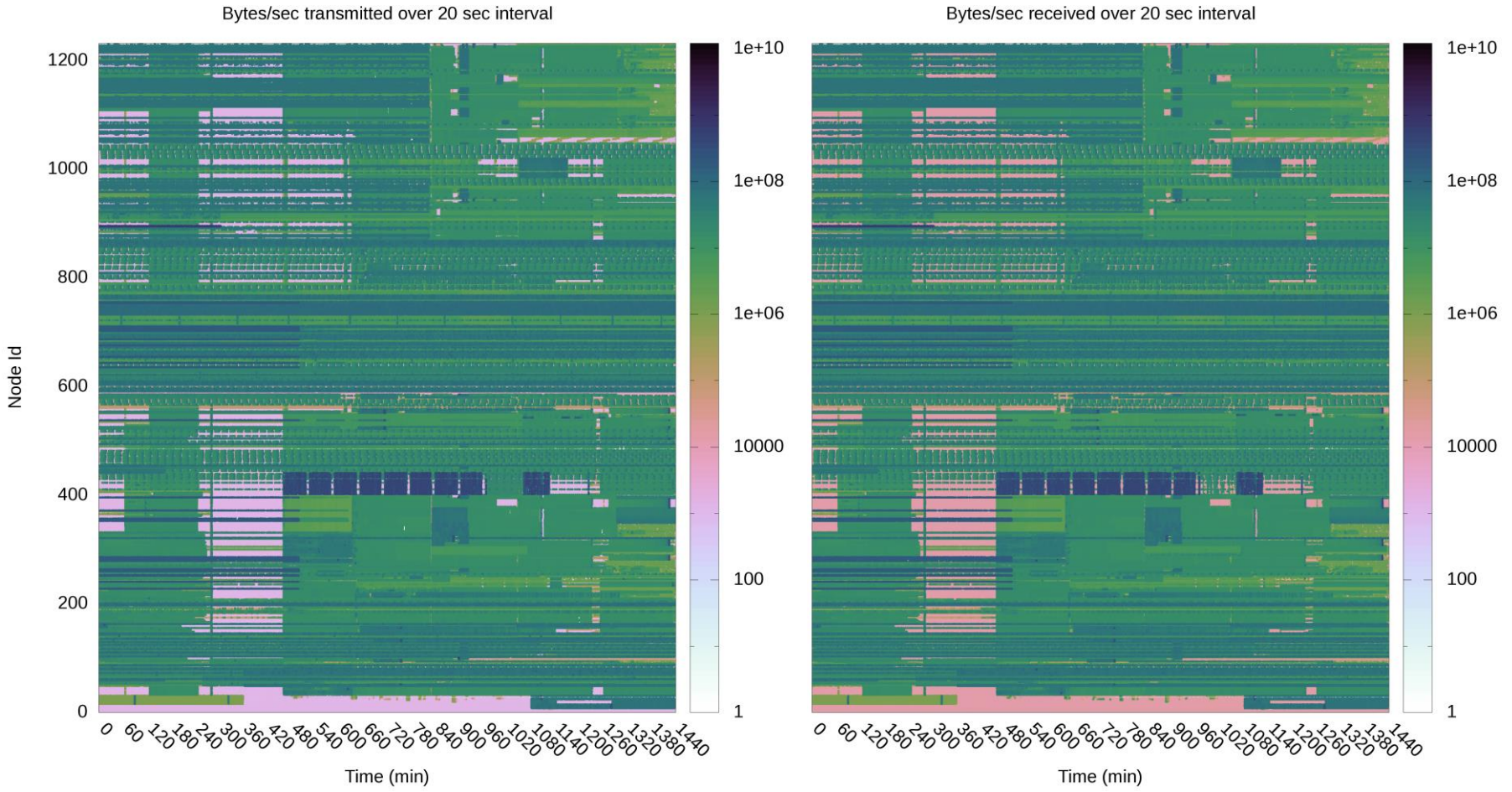
IB Traffic 12/9/2014



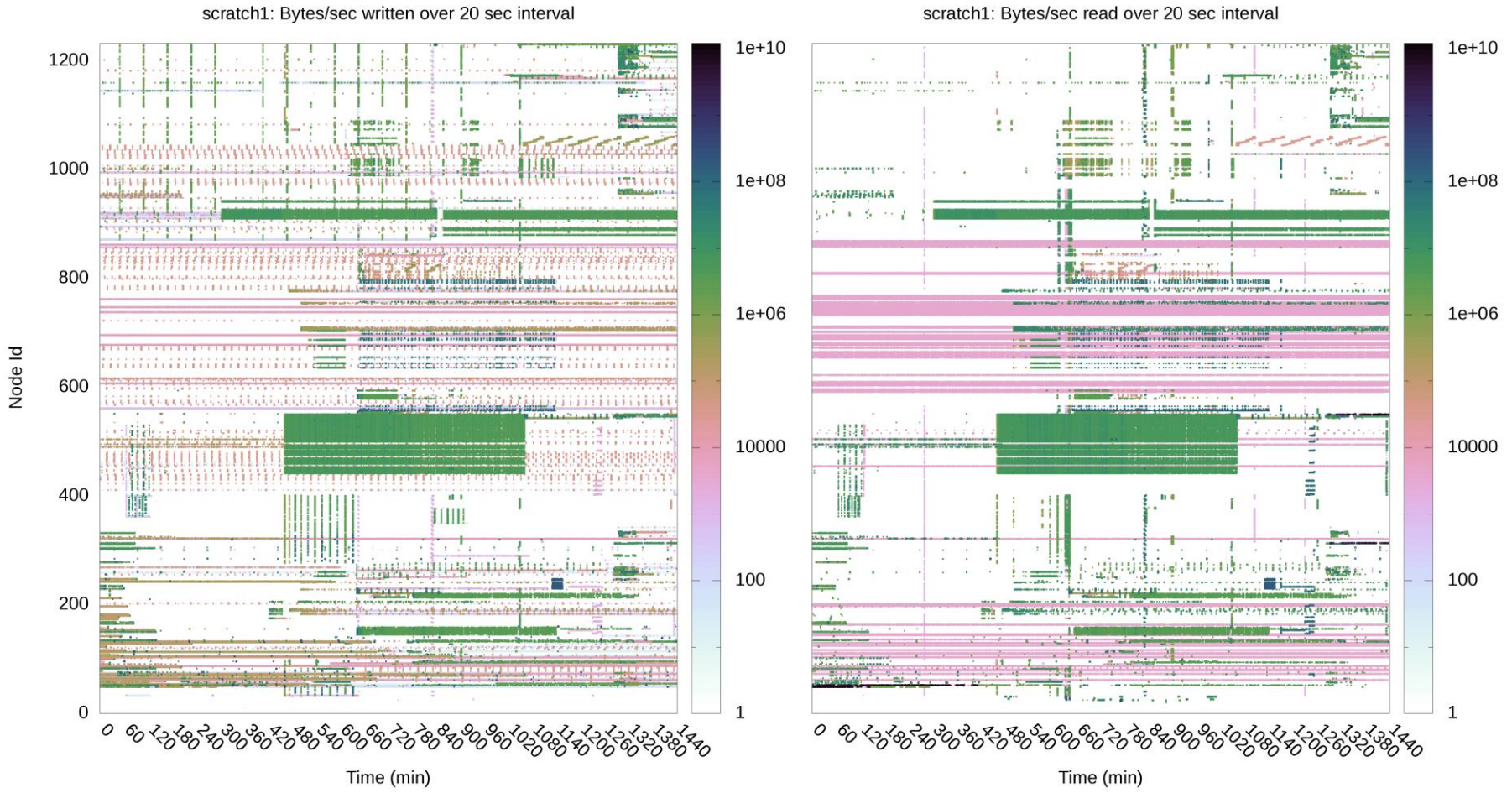
Lustre Traffic 12/9/2014

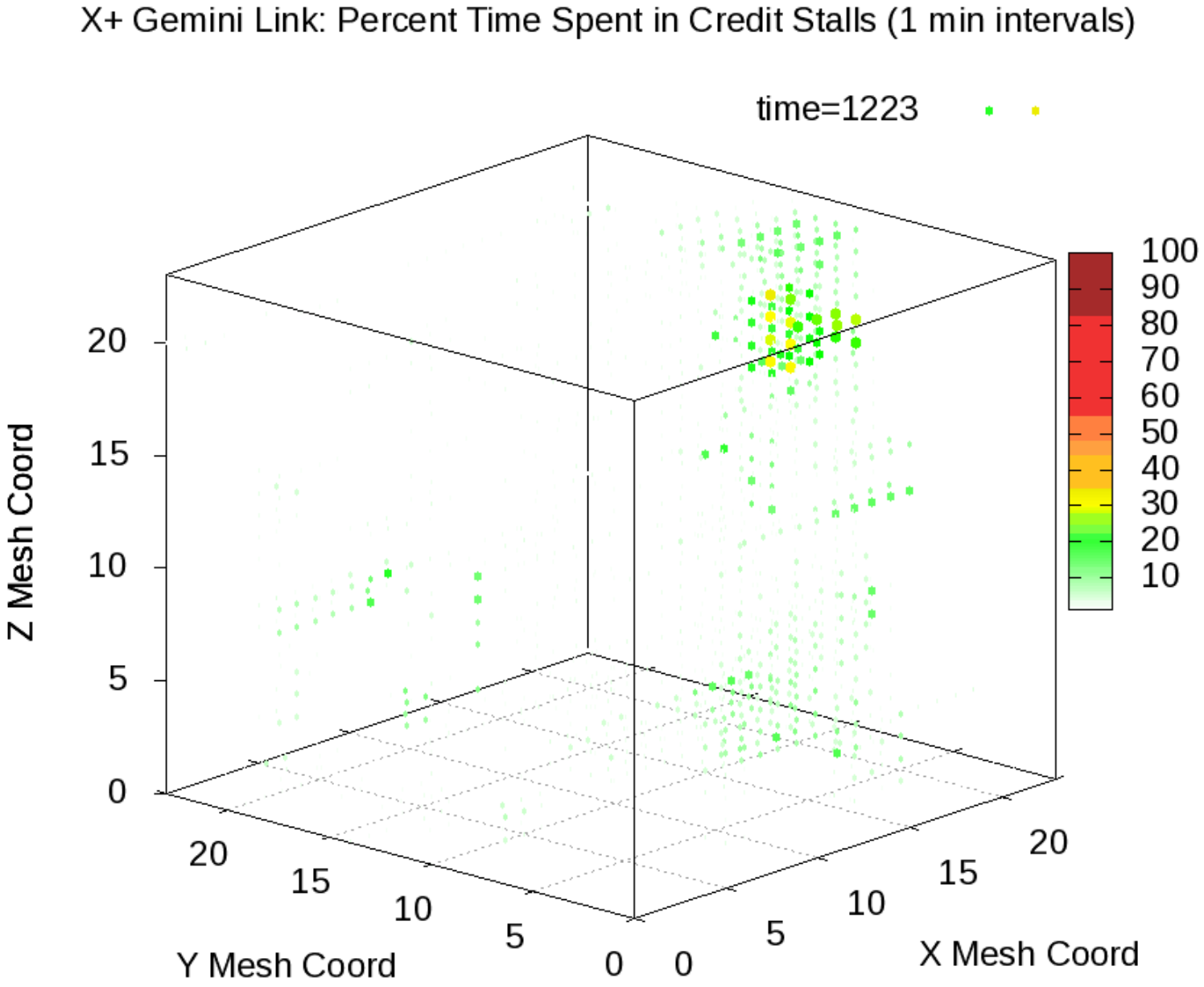


IB Traffic 1/22/2014



Lustre Traffic 1/22/2015

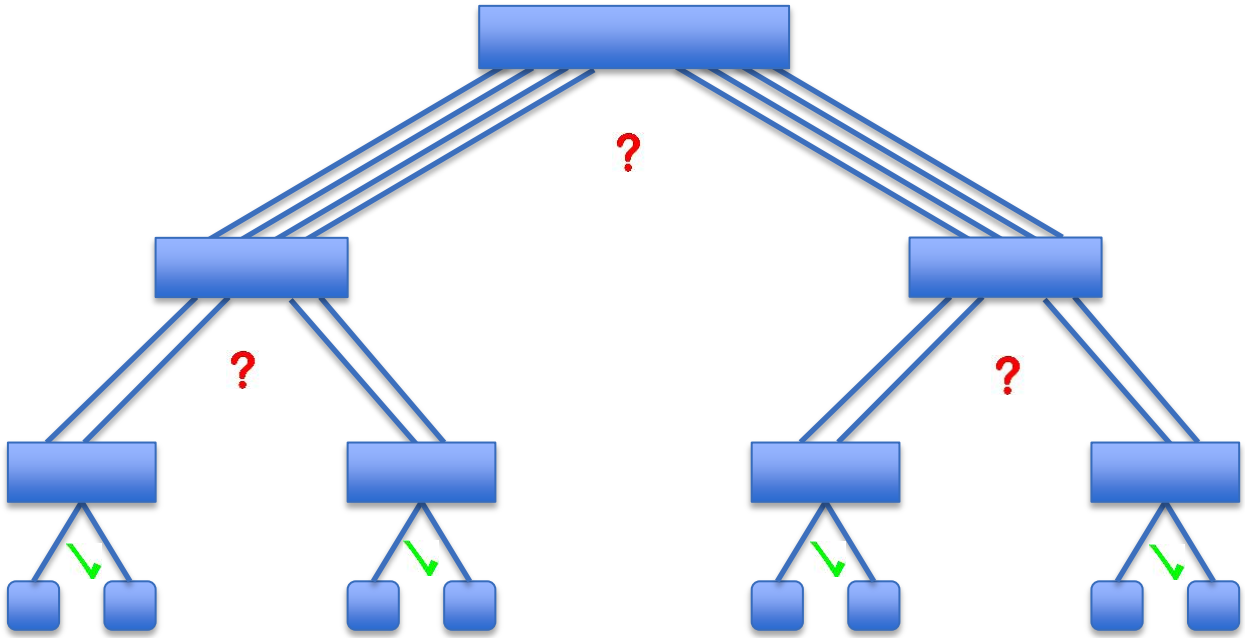



X+ Gemini Link: Percent Time Spent in Credit Stalls (1 min intervals)
time=1223
Z Mesh Coord
0
5
10
15
20
Y Mesh Coord
0
5
10
15
20
X Mesh Coord
0
5
10
15
20
100
90
80
70
60
50
40
30
20
10

IB Traffic Information

? = No information

↘ = Have information



Summary

- Goals
- Impediments
- LDMS Architecture
- IB Host Metrics
- IB Traffic
- Lustre Traffic
- Path forward

Questions?

- https://github.com/ovis-hpc/ovis

Thank You

OPENFABRICS ALLIANCE

OpenFabrics Software
User Group Workshop