## Runtime collection and analysis of system metrics for production monitoring of Trinity Phase II


*DeConinck, Nam, Morton, Bonnie, Lueninghoener (LANL)*
*Brandt, **Gentile**, Pedretti, Agelastos, Vaughan, Hammond, Allan (SNL)*
*Davis, Repik (Cray)*

SAND2017-5067 C

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 1

## Outline

- Motivation
- Trinity Phase 2
- Monitoring Setup + Enabling analysis
- Examples of Analysis
- Next steps

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 2

## Motivation

- SC15 BoF – Community's biggest question: Why does my application performance vary so much?
- Cray SMWG – Biggest desires from monitoring: Want to understand performance and utilization of the HSN and IO?
- SC16 BoF – Sites weren't monitoring because they don't know what they would look for!
- This talk:
  - What we are trying to find out?
  - What are we analyzing and how are we analyzing it?
  - How are we enabling analysis?
  - Where are we going from here?

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 3

## Trinity Phase II

- 9,984 nodes Cray XC KNL + 346 Service nodes
- 48 cabinets = 24 electrical groups
- Trinity Phase I: ~9,000 Haswell + additional service nodes
- Phase I & II integration scheduled in June

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 4

## Monitoring Setup

- Data sources:
  - LLM forwarding
  - Off-SMW ERD endpoint with PMDB and forwarding (e.g., 1Hz power and SEDC)
  - LDMS for node level (e.g., network, snx open/close/read/write, load/cpu, memory, perf, 10Hz power etc)
    - Network counters are exposed on-node via gpcd. Information about traffic, stalls between various interfaces, etc. (S-0045-20)
- Analysis and Storage Targets
  - Monitoring cluster
  - Short term active working set and storage
  - Long term archive and retrieval
  - Downstream consumers of analysis

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 5

## Enabling Analysis

- Goal: Determine actionable metrics that we can associate with performance impact and system issues
- Enabling Analysis
  - Streaming analysis at the monitoring cluster before insertion into a database and other storage.
    - Functional forms of data: rates, aggregations and ratios (e.g., stall/flit)
  - Integrate numeric out-of-band, numeric in-band, log
  - On-node streaming computations/transforms
  - SOS database
    - Binary formatted; supporting rolling-off and reloading segments; on-the-fly, multiple, flexible-indexing
- Tools:
  - Time-series numerical and log analysis and viz (e.g., graphana on SOS, Splunk,)
  - Domain-specific analysis and viz development
  - Baler – log patterns and numeric associations. Associative rule-mining.
    - 50,000 message processing inserts/sec single SOS instance. Supports parallel instances.

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 6

## DAT

- Goal:
  - Overhead testing of LDMS
  - ***Controlled scenarios to determine actionable indicators for monitoring***
- Applications:
  - CTH. Memory bandwidth bound. Nearest neighbor communication. Test problem designed for a consistent amount of work for each time step.
  - SPARC.
  - Partisn. Run with corespec. All application data structures placed exclusively in MCDRAM. Writes out a wall-clock timestamp to the output file on every cycle (100).

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 7

| WK1 | Mode | Group | W/Mon 1 | W/Mon 2 | Baseline 1 | Baseline 2 |
|---|---|---|---|---|---|---|
| CTH1 (1024) | Quad cache | 31 - 33 | 1282 | 1275 | 1297 | 1285 |
| CTH2 | Quad cache | 34 - 36 | 1298 | 1319 | 1326 | 1339 |
| CTH3 | Quad cache | 37 - 39 | 1270 | 1277 | 1291 | 1291 |
| CTH4 | Quad cache | 40 - 42 | 1452 | 1456 | 1465 | 1439 |
| SPARC1 | Quad flat | 43 - 45 | 1355 | 1359 | 1346 | 1347 |
| SPARC2 | Quad flat | 46 - 48 | 1354 | 1360 | 1347 | 1348 |
| SPARC3 (2048) | Quad flat | 49 - 54 | 1482 | 1485 | 1485 | 1487 |
| **WK2** | **Mode** | **Group** | **W/Mon 1** | **W/Mon 2** | **Baseline 1** | **Baseline 2** |
| CTH1 (1024) | Quad cache | 31 - 33 | 1206 | 1205 | 1223 | 1209 |
| CTH2 | Quad cache | 34 - 36 | 1249 | 1255 | 1236 | 1240 |
| CTH3 | Quad cache | 37 - 39 | 1203 | 1204 | 1183 | 1204 |
| CTH4 | Quad cache | 40 - 42 | 1403 | 1417 | 1413 | 1437 |
| PARTISN1 (1024) | Quad flat | 43 - 45 | 947 | 1027 | 1080 | 1033 |
| PARTISN2 | Quad flat | 46 - 48 | 998 | 888 | 978 | 1015 |
| PARTISN3 | Quad flat | 49 - 51 | 1216 | 978 | 1000 | 992 |
| PARTISN4 | Quad flat | 52 - 54 | 960 | 1033 | 1303 | 1025 |

Group 30 unallocated

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 8

## Power

| | Nodes | Runtime (s) | Total Energy (J) | Avg Power Per Node (W) | Avg CPU Power Per Node (W) | Avg Mem Power Per Node (W) | Number of Nodes Throttled |
|---|---|---|---|---|---|---|---|
| CTH 1 | 1024 | 1236 | 264790220 | 209.21 | 149.87 | 12.30 | 1 |
| CTH 2 | 1024 | 1249 | 264785934 | 207.03 | 148.33 | 11.95 | 0 |
| CTH 3 | 1024 | 1196 | 254463798 | 207.78 | 149.72 | 12.23 | 0 |
| CTH 4 | 1024 | **1424** | 299234211 | 205.21 | 147.28 | 12.12 | 1 |
| PARTISN 1 | 1024 | 1120 | 233431163 | 203.54 | 146.73 | 11.57 | 5 |
| PARTISN 2 | 1024 | 1019 | 210674495 | 201.90 | 146.32 | 10.53 | 2 |
| PARTISN 3 | 1024 | 1039 | 215450107 | 202.50 | 146.94 | 10.87 | 3 |
| PARTISN 4 | 1024 | **1343** | 278823301 | 202.75 | 145.69 | 11.81 | 4 |

PARTISN 2 – 1019 sec. Memory Power Per Node 10.53. (Run times from RUR)
PARTISN 4 – 1343 sec. Memory Power Per Node 11.81.
Baseline runs

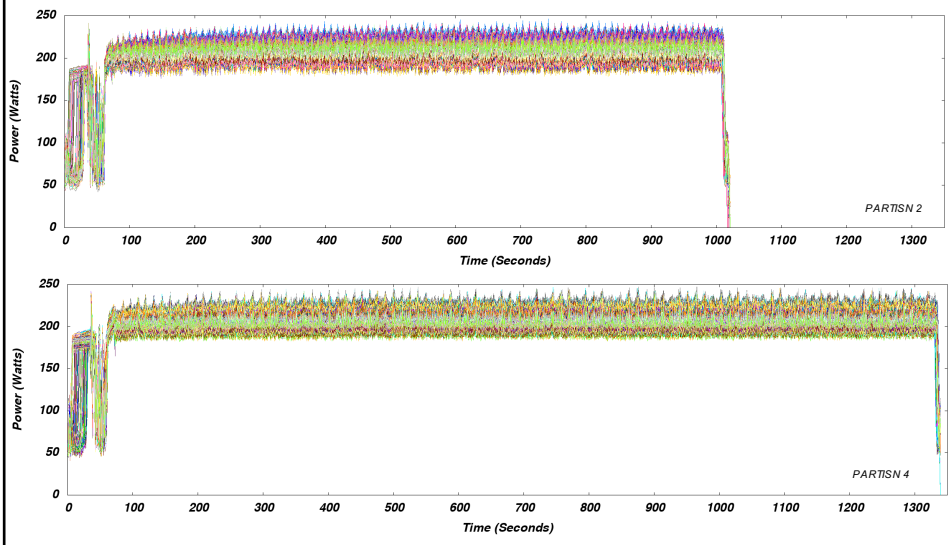
- Avg power per node similar
- Memory power %12 difference, but PARTISN running out of on-package MCDRAM, which is counted in the CPU energy rather than the memory energy
- Part-to-part differences?

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 9

## Node Power

- Spikes may correspond to the cycles (100)
- No centralized slowdown phase

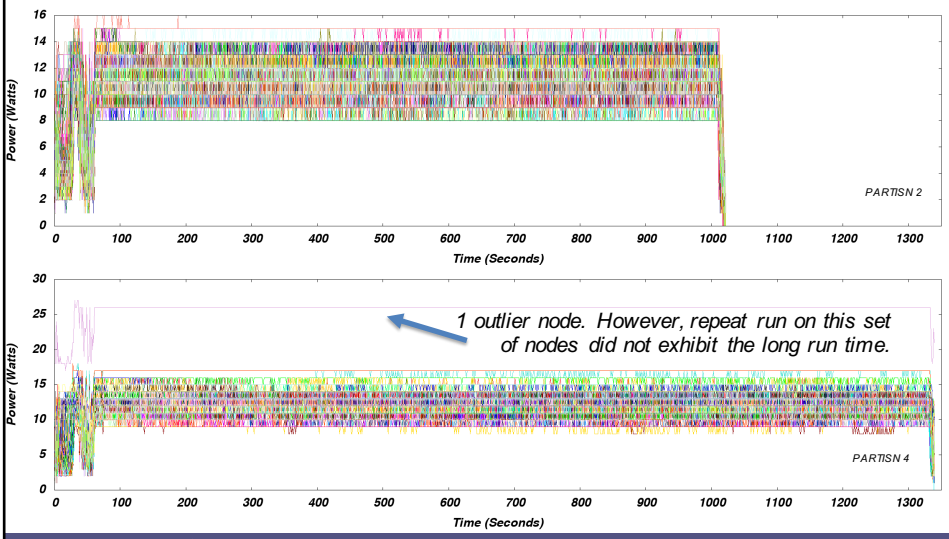



Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 10

## Memory Power

*Is the power spread significant wrt performance indicators?*




Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 11

## Cray XC Aries Dragonfly Configuration

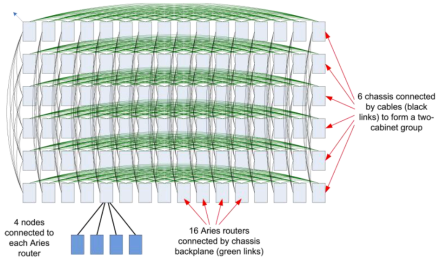



Figure 4. Structure of a Cascade electrical group. Each row represents the 16 Aries in a chassis, with 4 nodes attached to each, and connected by the chassis backplane (green links). Each column represents an Aries in one of the six chassis of a two-cabinet group, connected by electrical cables (black links).

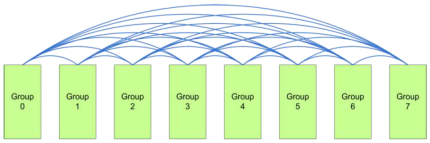
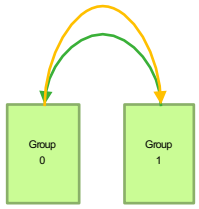
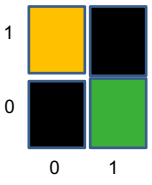



Figure 5. The global (blue) links connect Dragonfly groups together. In a large system these links are active optical cables.




From: *Cray Cascade: a Scalable HPC System based on a Dragonfly Network*, Faanes et. al., SC 2012

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 12

# Aries: Dynamic Routing

- Packets are generally routed *adaptively* along either minimal or non-minimal paths but can also be routed *deterministically*. *(Note: configurable routing parameters which can influence the routes that can be taken)*
- Adaptive routing:
  - Four possible routes are chosen at random, 2 minimal and 2 non-minimal.
  - The load on each of the 4 selected paths is compared and the path with the lightest load is selected.
- Minimal routing:
  - Minimal routing within an electrical group will take at most one Green and one Black hop
  - Minimal routing between groups will route minimally in both the source and the target groups and will take exactly one Blue link.
- Non-minimal routing:
  - Used to avoid congestion and to spread non-uniform traffic evenly over the set of available links in the system.
  - Non-minimal routes within a group can take up to two Green hops and two Black hops.
  - A global non-minimal path routes to an intermediate Aries router and then to the destination node using a minimal path. Maximum is 10 hops.
- Deterministic routing:
  - Used when runtime requires ordered packet delivery
  - Route is selected using a deterministic hash computed on fields from the packet header

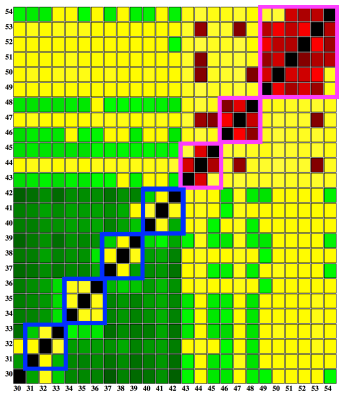
From: *Cray Cascade: a Scalable HPC System based on a Dragonfly Network*, Faanes et. al., SC 2012

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

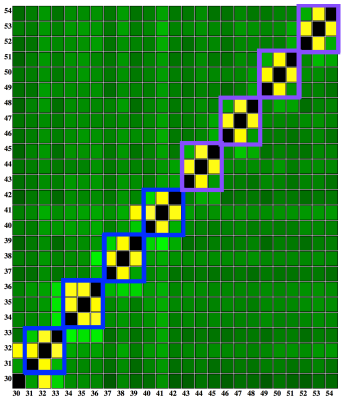
Slide 13

# HSN

*Workload 1*



*Workload 2*



*Max Red 3e8 Flits/sec*
*Mid Yellow 2e8 Flits/sec*
*Mid Green 1e8 Flits/sec*

- *Compression going into the network*
- *Uncompressed coming out of the network*
- *Flits are different sizes at different points in the network (Aries Hardware Counters S-00450-20)*

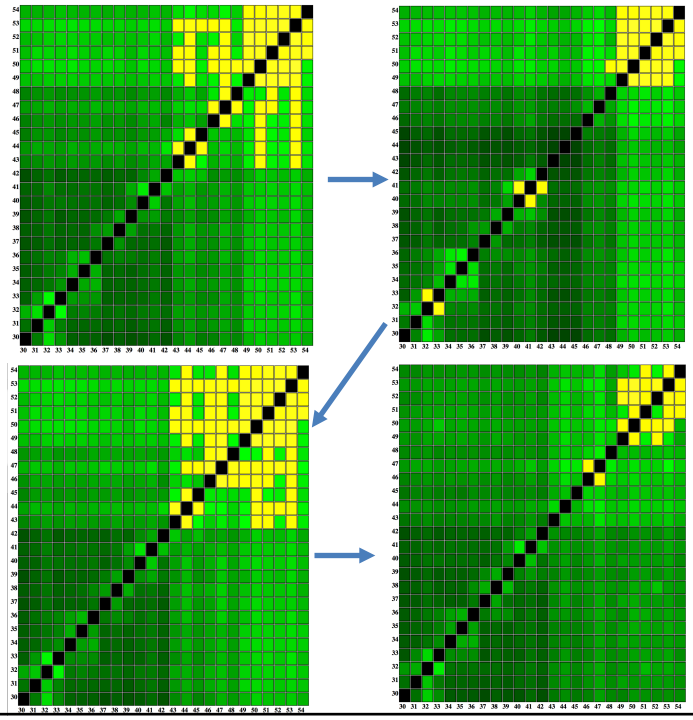
- Interest in assessing when an application's traffic is potentially affecting other applications.
  - In our allocations, this would be traffic sent outside the indicated boxes
- Plots: Max incoming Flits (summed over the VC) per sec over the workload over the 8 blue links between workgroups
  - *Not* a snapshot in time
- Traffic being sent into all electrical groups from all electrical groups.

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

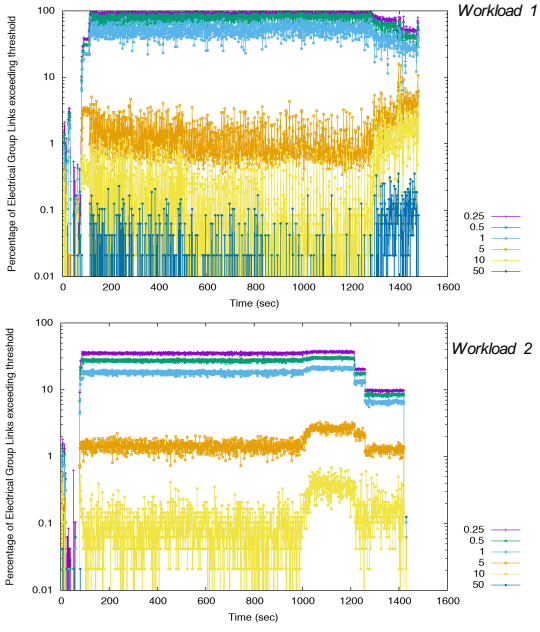
Slide 14

## Timeseries (30 sec intervals)

- Workload 1
- Flux in intensity of communications
- Temporal variations



## Congestion Measures




Stall/flit (flits to be forwarded to another link or a local NIC):

AR_RTR_x_y_INQ_PRF_INCOMING_FLIT_VCz

AR_RTR_x_y)_INQ_PRF_ROWBUS_STALL_CNT

- Stall counter increments when a ready to forward flit is prevented by backpressure from doing so (S-0045-20)
- 0.25 has been suggested to be "normal"
- Stall/flit increment at equal rates, then flits are crossing the interface at half the rate they would be with no stalls (S-0045-20)
- %age blue links with stall/flit ratio over threshold vs time

## Other Backpressures

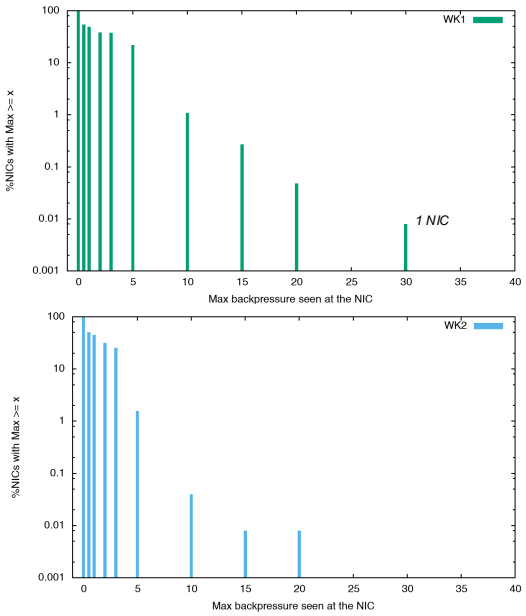
- Other interfaces
- "Node backpressure" ~ request stalls/flits to NIC n

  AR_NL_PRF_REQ_PTILES_TO_NIC_n_FLITS

  AR_NL_PRF_REQ_PTILES_TO_NIC_n_STALLED

- c0-7c0s14 NIC 3 has highest max. Also continuous through time.

*Node allocated In the outlier CTH 4 runs in both workloads*




## Log + Numerical Analysis

- Baler discovers patterns from data with no user guidance
  - Dictionary
  - Meta-clustering of similar patterns
  - Numeric data can be made into a pattern via value ranges (X < Temp < Y)

Deterministic patterns enable comparison cross-system, across time:

* - - Node * interrupt *=*, *=*, *=* *[*]: * * * Processor Hot

* * - - Node * power budget exceeded! Power=*, Limit=*, * Correction Time=*

Loglines:

bcsysd 2080 - - Node 2 interrupt IREQ=0x20000, USRA=0x0, USRB=0x80 USRB[7]: C0_PROCHOT CPU 0 Processor Hot

bcpmd 2140 - - Node 2 power budget exceeded! Power=340, Limit=322, Max Correction Time=6

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 18

## Discovering significant messages

- Augmented dictionary with 100 words (e.g., bios, linktune),
- Weighted 50 words (e.g., fail, throttle)
- 5 months: 4.5 billion lines (w/o job data) -> 497K patterns -> 11K meta-patterns ->1350 weighted meta-patterns
- Can associate log patterns with numeric patterns. Can query for patterns by components, times, etc

```
==============================
Weighted Matches: 5 (3/11252)
==============================
(W=5)   312     nlrd • found_critical_aries_error: handling failed PCI-• link on •-• (node •-•)
(W=5)   330     nlrd • found_critical_aries_error: handling failed • •-•
(W=5)   5290    * HWERR[•]:•:Uncorrectable AER_COMPLETION_TIMEOUT Error:•=•:•=•:•=•:•=•:•=•:•=•:•=•:•=•

==============================
Weighted Matches: 4.75 (1/11252)
==============================
(W=4.75)        316     nlrd • set_warm_swap_err: appending warm swap error text: * * **aborted due to hardware failure during * *

==============================
Weighted Matches: 4.5 (5/11252)
==============================
(W=4.5) 317     nlrd • Calling user exit script /opt/cray/hss/default/•/•.• due to link recovery failure
(W=4.5) 318     nlrd • Error string was: * * **aborted due to hardware failure during * *
(W=4.5) 459     controllermessages • • - - do_node_*: *; • • resiliency_quiesce_active • orb_•_mask •
(W=4.5) 7571    * pbs_mom - •-• - LOG_ERROR::•_err, prolog/epilog failed, file: /•/spool/torque/mom_•/*, exit: -•, prolog/epilog timeout occurred, * * *
(W=4.5) 8150    messages aprun • •-• [alps_•@•] •=none, Error, user=•, batch_id=•.•-•, • prolog failed for batch_id •.•-•, • •, • •; application launch aborted

==============================
Weighted Matches: 4 (8/11252)
==============================
(W=4)   3       * HWERR[**]**:•:Correctable AER_BAD_* Error:•=•:•=•:•=•:•=•:•=•:•=•:•=•:•=•
(W=4)   9       • HWERR[•-•][•]:•:sequential crc error & !ignore seqbadcrc see NUM SEQ BAD CRC MAX Error:•=•:•=•:•=•:•=•:•=•:•=•
```

## Correlating Log and Numerical Data

- Baler meta-pattern:

  • HWERR[•-•][•]:•:The pcie had • link width change or • speed change (•, •, or • speed)

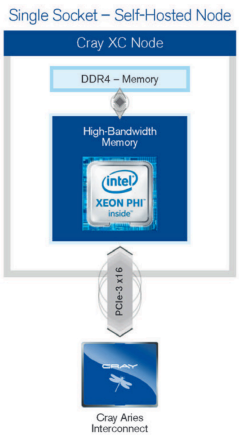
  Log message:

  c0-7c0s14a0n3 hwerrlog HWERR[c0-7c0s14a0n3][162]:0x5b09:The pcie had a link width change or a speed change (Gen1, 2, or 3 speed)

| c0-7c0s14a0n3* | 6 |
|---|---|
| c1-5c1s14a0n2 | 1 |
| c1-9c1s9a0n0 | 1 |
| c5-7c1s5a0n2 | 1 |
| c11-5c1s2a0n1 | 6 |

*Happens near the reboot at the morning of the DAT and near the reboot after

Figure from: http://www.cray.com/sites/default/files/resources/CrayXC_IntelXeonPhiPDC.pdf




*Is this related to the high backpressure at the same NIC ? And to the performance of CTH4?*

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 20

## Next steps

- Questions:
  - How to best present and analyze large numbers and large dimensions of data for exploration?
    - What dimensional reductions result in meaningful results?
  - How can we get the domain knowledge we need to guide analysis and to distinguish significant associations vs coincidences?
    - Many "unsupervised" procedures return meaningless results
- Solidifying the Trinity monitoring data paths after the system integration in June
- Integration of Application with System data
- Streaming analysis advancements:
  - On-node transforms
  - Minerva test site

Los Alamos NATIONAL LABORATORY EST. 1943 | Sandia National Laboratories | NNSA National Nuclear Security Administration

Slide 21