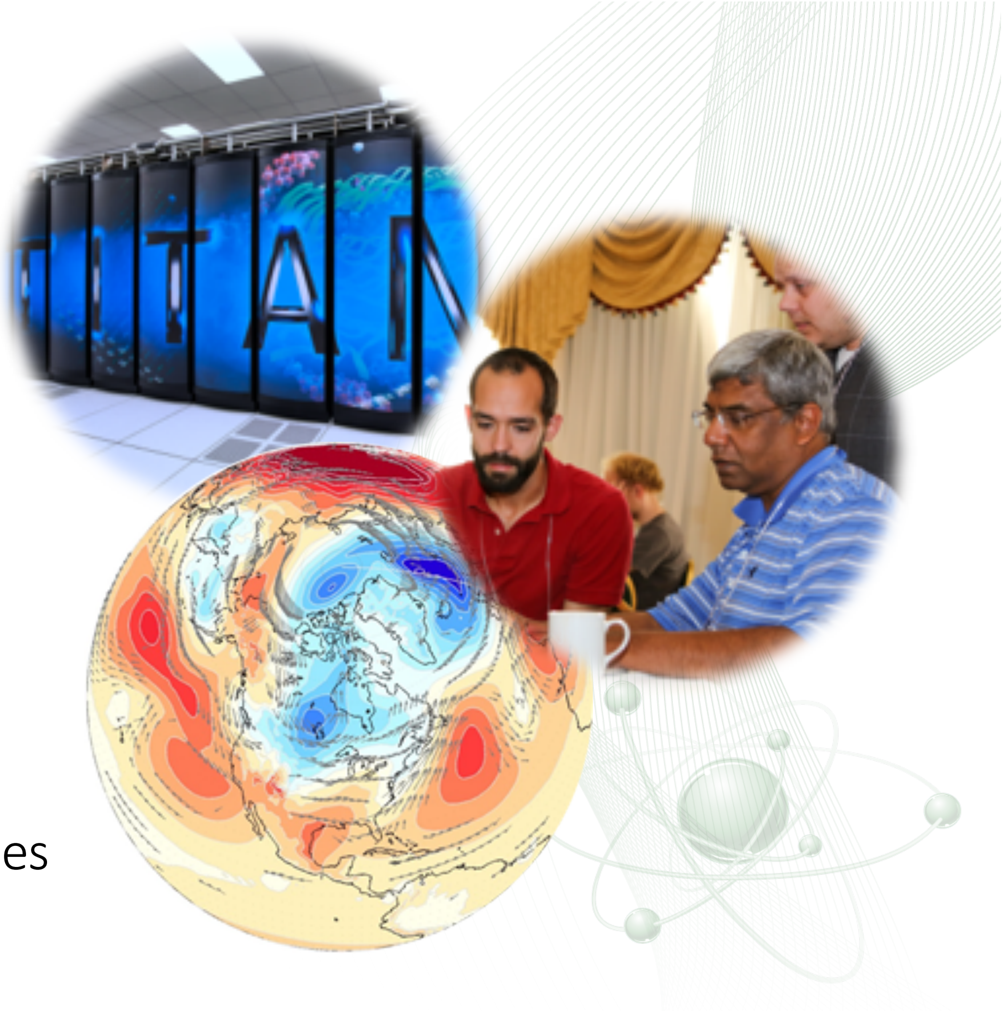
GPU Errors on HPC Systems: Characterization, Quantification, and Implications for Architects and Operations



Jim Rogers

Director, Computing and Facilities
National Center for Computational Sciences
Oak Ridge National Laboratory

Session Description, Session ID S5566

Day: Tuesday, 03/17
Time: 16:30 - 16:55
Location: Room 212B

GPU Errors on HPC Systems: Characterization, Quantification, and Implications for Architects and Operations

Titan, the world's #1 Open Science Supercomputer, contains more than 18,000 GPUs that scientists from domains including astrophysics, fusion, climate, and combustion use routinely to run large-scale simulations. While the performance efficiency of GPUs is well understood, their resilience characteristics in a large-scale computing system have not been fully evaluated. A detailed failure analysis provides a thorough understanding of GPU errors on a large-scale GPU-enabled system. The measurement interval spans up to 22 months and approximately 300M node-hours on the Cray XK7 Titan supercomputer at the Oak Ridge Leadership Computing Facility. We describe representative findings from the field data and discuss the implications of these results relevant to both existing operations and future architectures.

Presenter Overview

Jim Rogers is the Computing and Facilities Director for the National Center for Computational Sciences (NCCS) at Oak Ridge National Laboratory (ORNL). The NCCS provides full facility and operations support for multiple petaFLOP-scale systems including Titan, a 27PF Cray XK7. Jim has a BS in Computer Engineering, and has worked in high performance computing systems acquisition, facilities, integration, and operation for more than 25 years.

Content

- The OLCF’s Cray XK7 Titan
  - Science challenges for the OLCF in the next decade
  - Mechanical packaging for the Cray XK7
    - Cabinet, cage, blade, and node descriptions
  - The NVIDIA GK110 layout
  - NVIDIA Kepler memory architecture
- NVIDIA GPU error assessment
  - Error conditions on the K20x
  - Single Bit Errors – Distribution by row/column
  - Single Bit Errors – Distribution by structure
  - Double Bit Errors
  - Page Retirement Errors
  - [Off the bus] errors –spatial analysis
  - [Off the bus] errors – temporal analysis
- Takeaways…
- A Summit teaser

The OLCF’s Cray XK7 Titan

A Hybrid System with 1:1 AMD Opteron CPU and NVIDIA Kepler GPU

Configuration

- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA K20x (Kepler) GPU
  - 32GB DDR3 + 6 GB GDDR5 memory
- 710TB total system memory
- 512 Service and I/O nodes
- 200 Cabinets

Performance

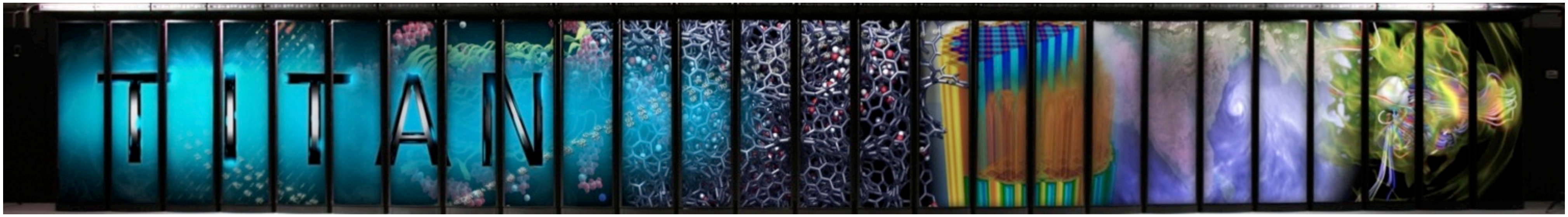
- Peak performance of 27 PF
- Sustained performance of 17.59 PF (HPL)

Utilization

- Delivered 140M node-hours in 2014
- 90% utilization of available hours
- 99.6% scheduled availability

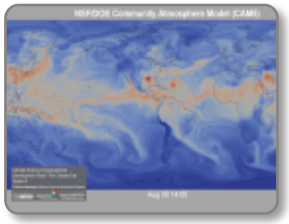
Facilities

- Occupies 4,352 ft2 (404m2)
- Requires four separate 2.5MVA transformers
- Peak electrical consumption of 8.9MW
- Liquid cooled using secondary loop/refrigerant-based system (EcoPhlex)



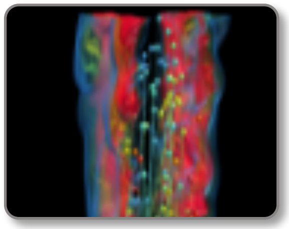
Science challenges for the OLCF in the next decade

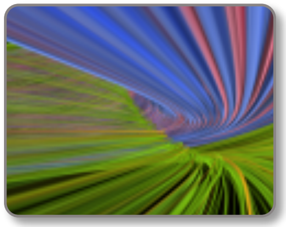
ASCR Mission: "…discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to DOE."



Climate Change Science
Understand the dynamic ecological and chemical evolution of the climate system with uncertainty quantification of impacts on regional and decadal scales.

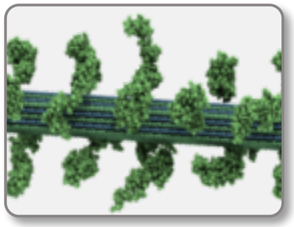
Combustion Science
Increase efficiency by 25%-50% and lower emissions from internal combustion engines using advanced fuels and new, low-temperature combustion concepts.

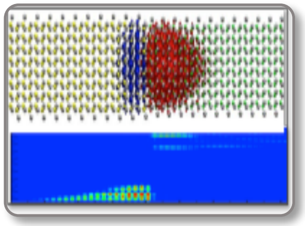




Fusion Energy/ITER
Develop predictive understanding of plasma properties, dynamics, and interactions with surrounding materials.

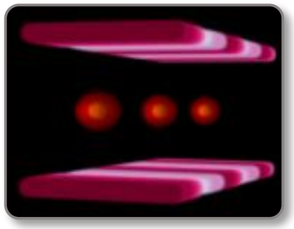
Biomass to Biofuels
Enhance the understanding and production of biofuels for transportation and other bio-products from biomass.





Solar Energy
Improve photovoltaic efficiency and lower cost for organic and inorganic materials.

Globally Optimized Accelerator Designs
Optimize designs as the next generations of accelerators are planned, detailed models will be needed to provide a proof of principle and efficient designs of new light sources.



Mechanical packaging for the Cray XK7

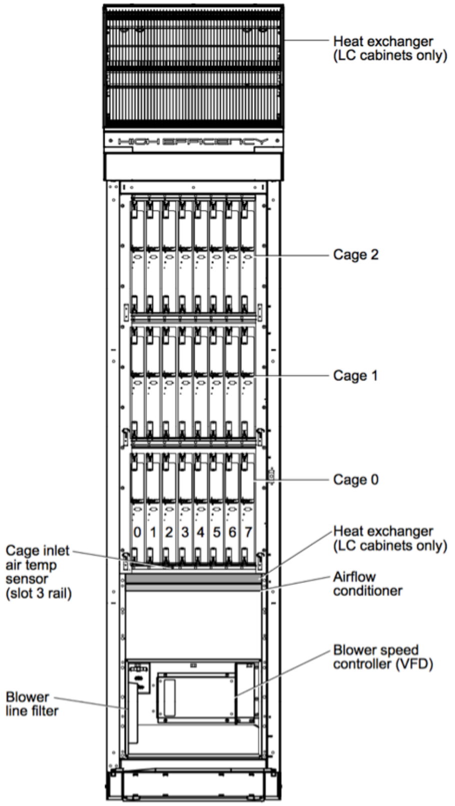


Figure 1. Cray XK7 cabinet schematic

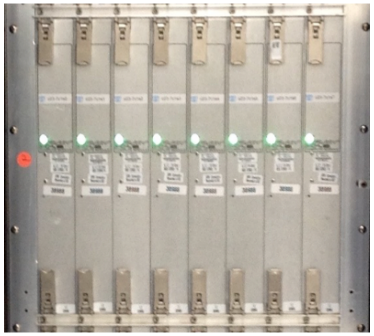


Figure 2. Eight compute blades occupy a single cage

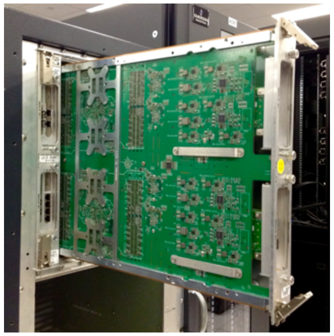


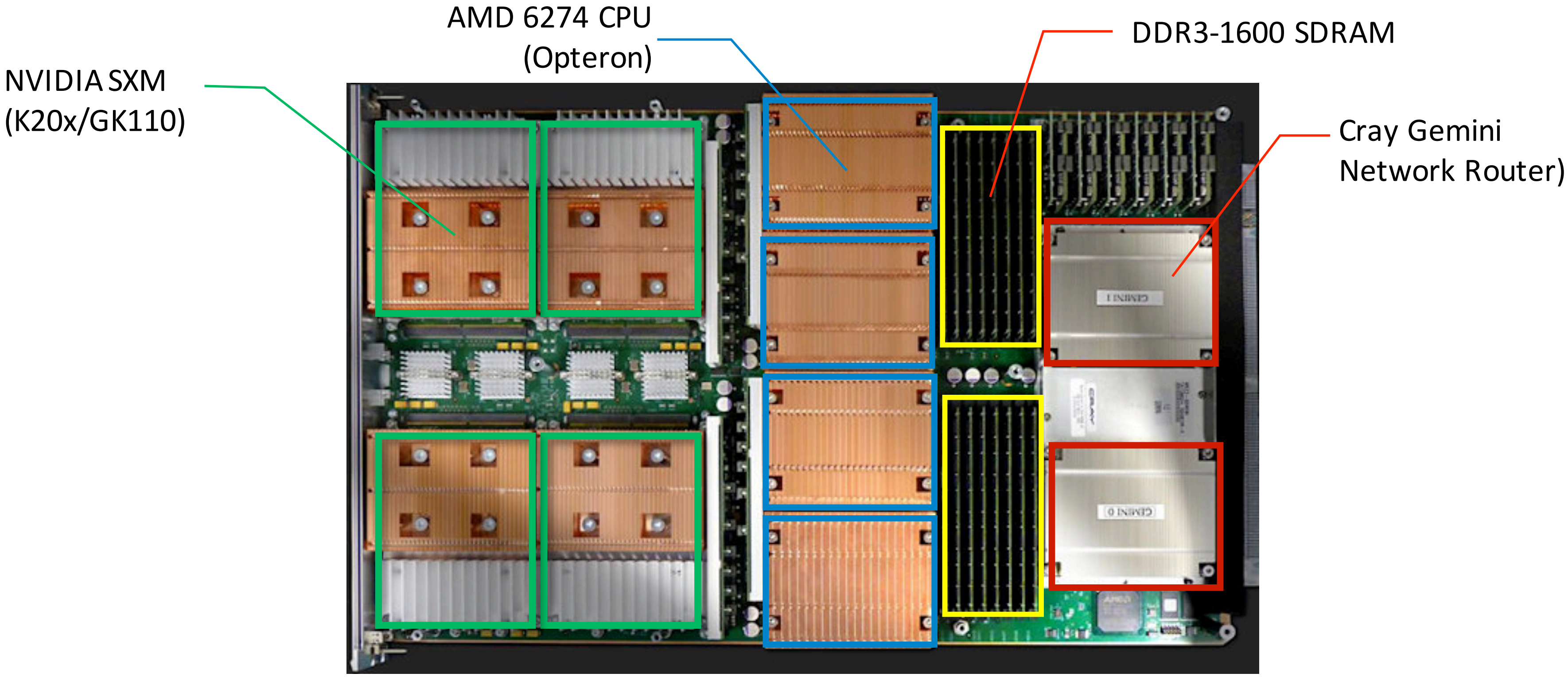
Figure 3. Vertical blade packaging accommodates stacked configuration

A cabinet contains 3 cages. Each cage contains 8 blades. Blades are mounted vertically, and air flows from bottom to top at high velocity. Nominal air temperature (inlet) is 69 ℉ (20.5 ℃). Eject temperature in to the heat exchanger can approach 120℉ (49 ℃). High air flow rates and dense/stacked packaging introduce significant non-uniform temperature distribution issues throughout the cabinet.

ΔT° (in °F) = Watts (cooling) /(.316 x CFM)

ΔT° = 44,500W/(.316*3000) CFM = 47 °F

Mechanical packaging for the Cray XK7



NVIDIA SXM (K20x/GK110)

AMD 6274 CPU (Opteron)

DDR3-1600 SDRAM

Cray Gemini Network Router)

Figure 4. Cray XK7 compute blade. Air flows bottom to top, with an approximate 15 ΔT° (in °F). There are four compute nodes per blade. Each node includes a 16-core AMD Opteron, 32GB DDR3-1600, and an NVIDIA K20X with 6GB GDDR5.

The NVIDIA GK110 layout

PCIe Gen2 Interface

Memory Interface (6 Total)

SMX (15 Total; 14 Yield)

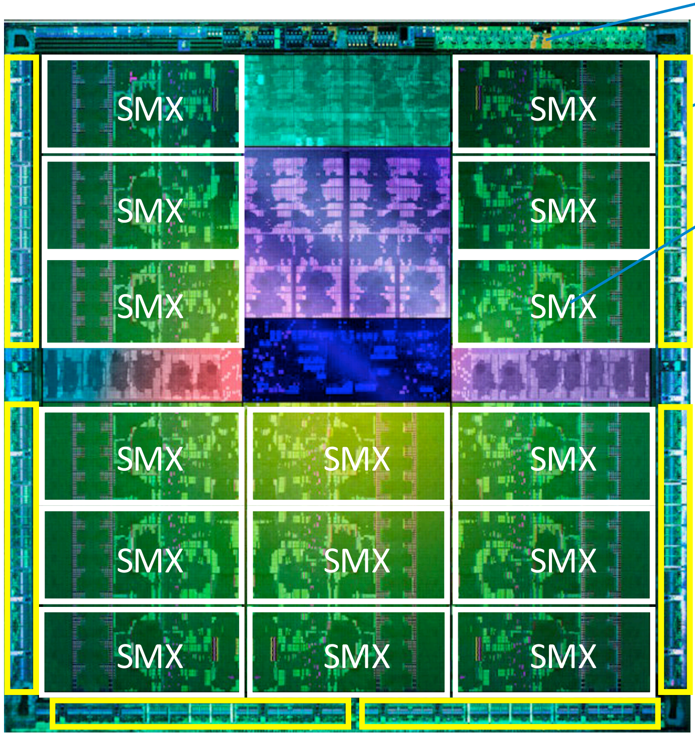


Figure 5. NVIDIA GK110. 7.1B transistors on a 28nm process. PCIe Gen2 interface, 6 Memory Interfaces, 14 SMX streaming multiprocessors. *Die photo courtesy NVIDIA Corporation.*

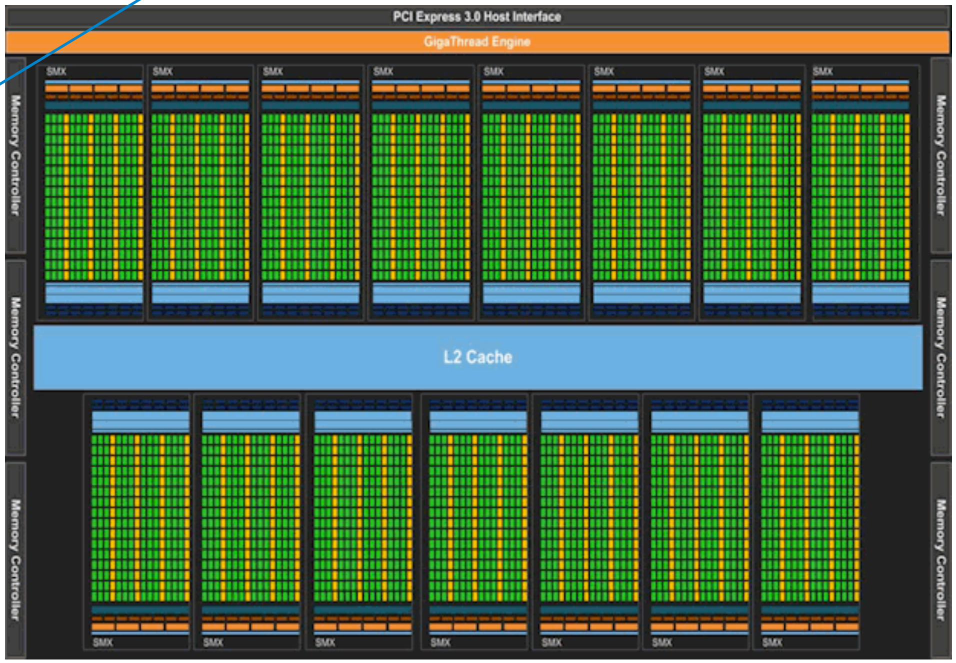


Figure 6. GK110 Logical Layout. On-die memory structures: 1,536KB L2 cache (shared); per-SMX L1 cache (64KB), register file (64Kx32-bit), read-only cache (48KB) *Block diagram courtesy NVIDIA Corporation.*

NVIDIA Kepler memory architecture

Table 1. Memory regions on the NVIDIA K20x.

Memory Region	Unit Size/ Description	Number	Protection Mechanism	Location
Register File Space	65,536 x 32-bit (256KB)	14 (one per SMX)	SECDED ECC	On-die
L1 Cache	64KB	14 (one per SMX)	SECDED ECC	On-die
Read-only Data Cache	48KB	14 (one per SMX)	SEC through parity check	On-die
L2 Cache	1,536KB	1	SECDED ECC	On-die
GDDR5 SGRAM	6GB	24 pieces 64Mx16	SECDED ECC	On-package

Kepler's register files, shared memories, L1 cache, L2 cache and DRAM memory are protected by a Single-Error Correct Double-Error Detect (SECDED) ECC code. In addition, the Read-Only Data Cache supports single-error correction through a parity check; in the event of a parity error, the cache unit automatically invalidates the failed line, forcing a read of the correct data from L2.

Content

- The OLCF’s Cray XK7 Titan
  - Science challenges for the OLCF in the next decade
  - Mechanical packaging for the Cray XK7
    - Cabinet, cage, blade, and node descriptions
  - The NVIDIA GK110 layout
  - NVIDIA Kepler memory architecture

- NVIDIA GPU error assessment
  - Error conditions on the K20x
  - Single Bit Errors – Distribution by row/column
  - Single Bit Errors – Distribution by structure
  - Double Bit Errors
  - Page Retirement Errors
  - [Off the bus] errors –spatial analysis
  - [Off the bus] errors – temporal analysis

- Takeaways…
- A Summit teaser

Error conditions on the K20x

Table 2. GPU Errors on the K20x

Error	XID	Protection	Impact
Single Bit Error (SBE)	-	Corrected by ECC. Silent data corruption (SDC) may occur without ECC support	No application impact
Double Bit Error	48	Detected by ECC. SDC may occur without ECC support.	Application failure
[Off the bus] Error	-	None	Application failure
Display Engine Error	56	None	Application failure
Error Programming Video Memory Interface	57	None	Application failure
Unstable video memory interface detected	58	None	Application failure
Internal micro-controller halt	62	None	Application failure
ECC page retirement error	63,64	None	Application failure
Video processor exception	65	None	Application failure

Single Bit Errors – Distribution by row/column

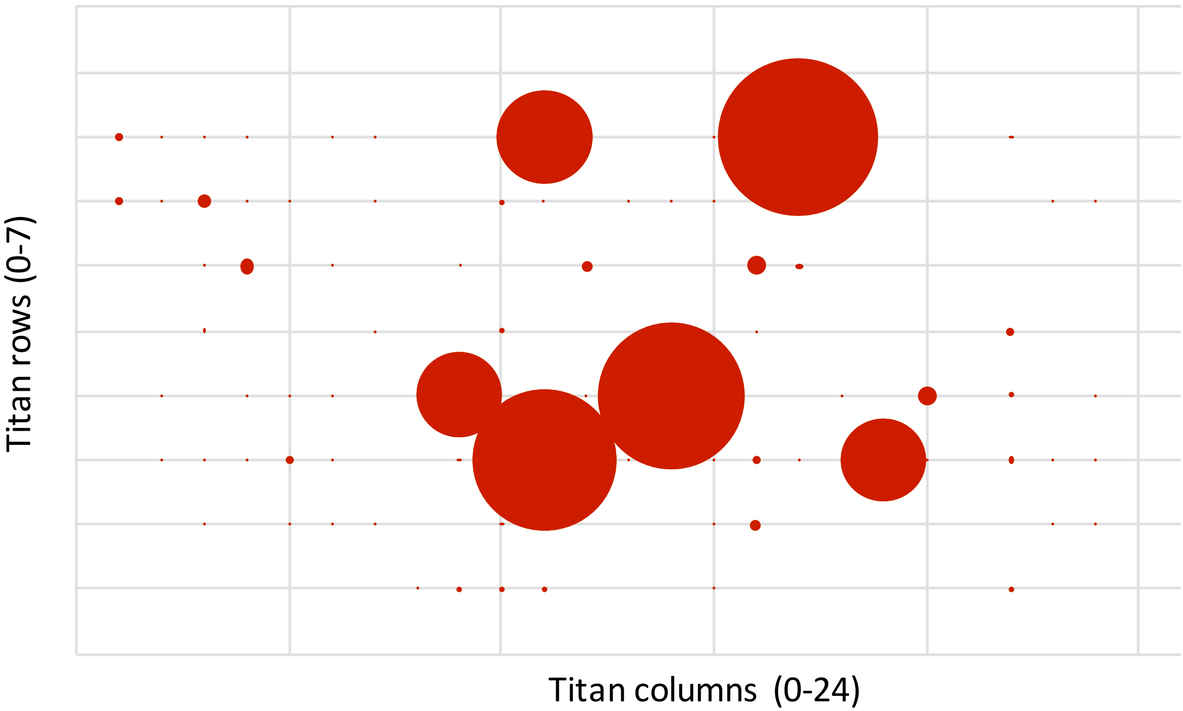


Figure 7. Single Bit Errors on Titan. Measurement Period: the epoch (2012) – August 2014. These counters are never reset.

Distribution of SBEs across Titan, by physical row/column is significantly uneven.

- Total Errors reported: 6,088,374
- Only 899 of 18,688 SXMs reported SBEs.
- 98% of the single bit errors were confined to 10 cards.

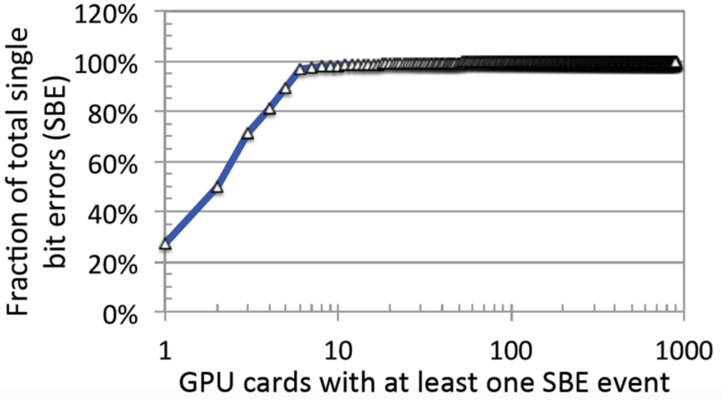


Image courtesy Devesh Tiwari, ORNL

OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY

Single Bit Errors – Distribution by structure

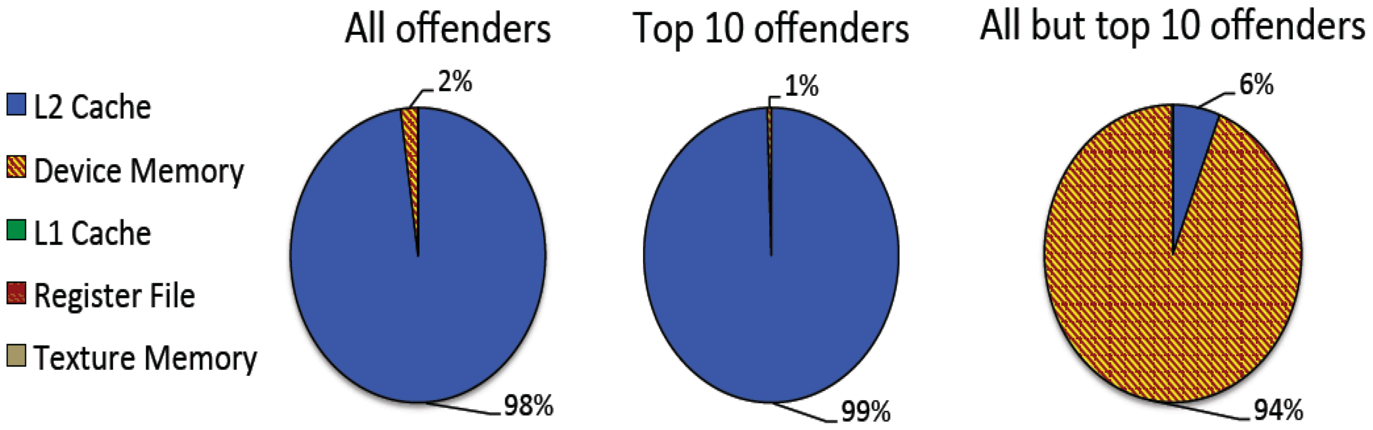


Figure 8. Single Bit Errors on Titan. Distribution by memory structure.
Figure courtesy Devesh Tiwari, ORNL.

Of more than 6M reported SBEs across 899 SXMs, 98% of those SBEs occurred in L2 cache.

Looking at the 10 SXMs that accounted for 98% of all SBEs, 99% of the errors on those 10 SXMs occurred on L2 cache. This is a clear indication of test escapes (5.35%) for the L2 cache.

Removing the 10 worst cards, the distribution of the remainder of the errors is dominated by the device memory (GDDR5).

Note the lack of errors in L1, Register, and read-only data cache.

Double Bit Errors

Distribution of DBEs across Titan, by physical row/column is significantly more evenly distributed than SBEs.

- Measurement Period: June 1, 2013 – Feb 28, 2015
- Total DBEs reported: (just) 91
  - 6 SXMs account for 25% of DBEs
  - MTBF for DBEs on Titan: 7d

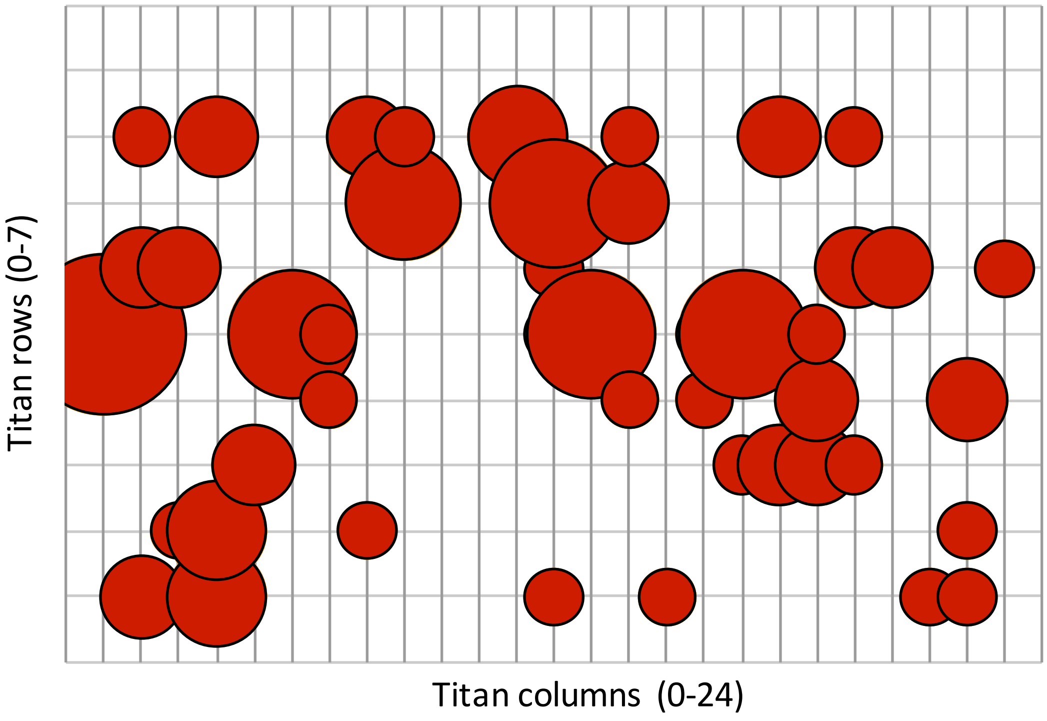


Figure 9. Double Bit Errors on Titan. Distribution by row/column.

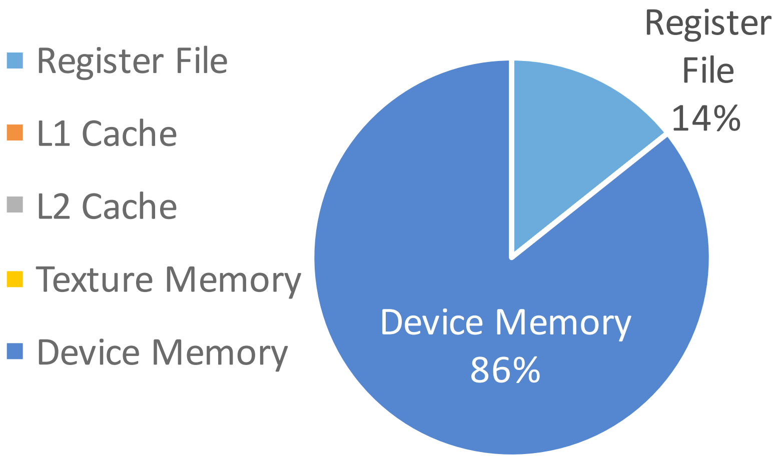


Figure 10. Double Bit Errors on Titan. Distribution by memory structure. 0 reported DBEs in L1, L2, Texture.

Page Retirement Errors

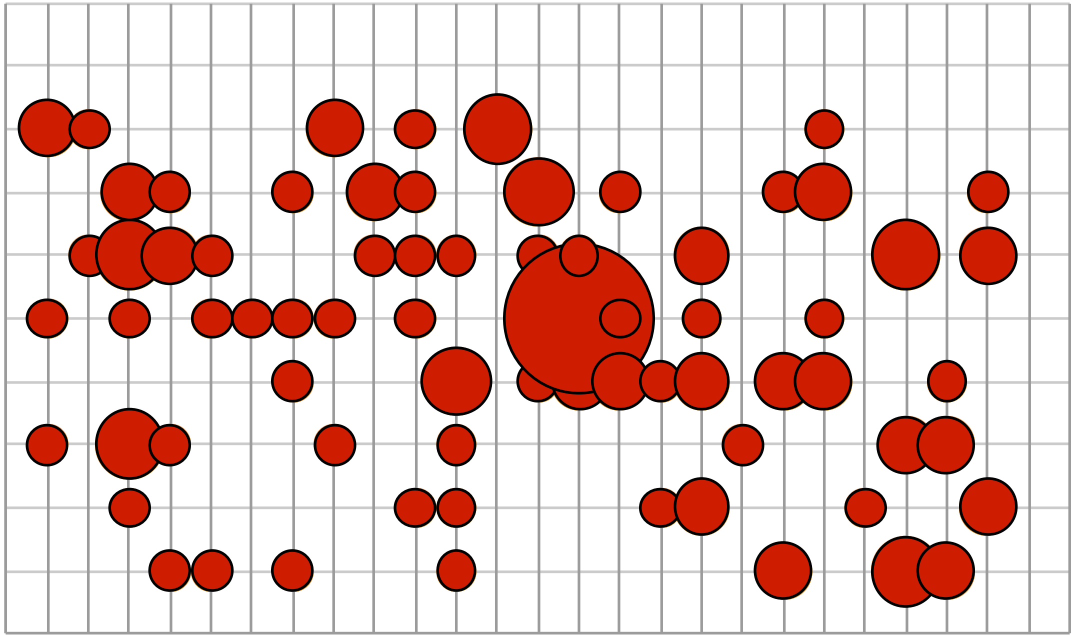


Figure 11. Page Retirement Errors on Titan. Distribution by row/column. A single SXM was responsible for more than 10% of the ECC page retirement errors.

The NVIDIA driver supports "retiring" of bad frame buffer memory cells, by retiring the page the cell belongs to. This dynamic page retirement is done automatically for cells that are degrading in quality.

The NVIDIA driver will retire a page once it has experienced a single DBE or 2 SBE. These addresses are stored in the InfoROM. When the driver loads, it retrieves these addresses from the InfoROM, then has the frame buffer manager set these pages aside.

Ideally, the NVIDIA driver will catch weakening cells at the 2 SBE point and retire the page, before the cell degrades to the point of a DBE and disrupts an application.
A retired page will be stored in the InfoROM for persistence for the life of the board. However, the driver will need to be reloaded for the retirement to take effect

[Off the bus] errors – spatial analysis

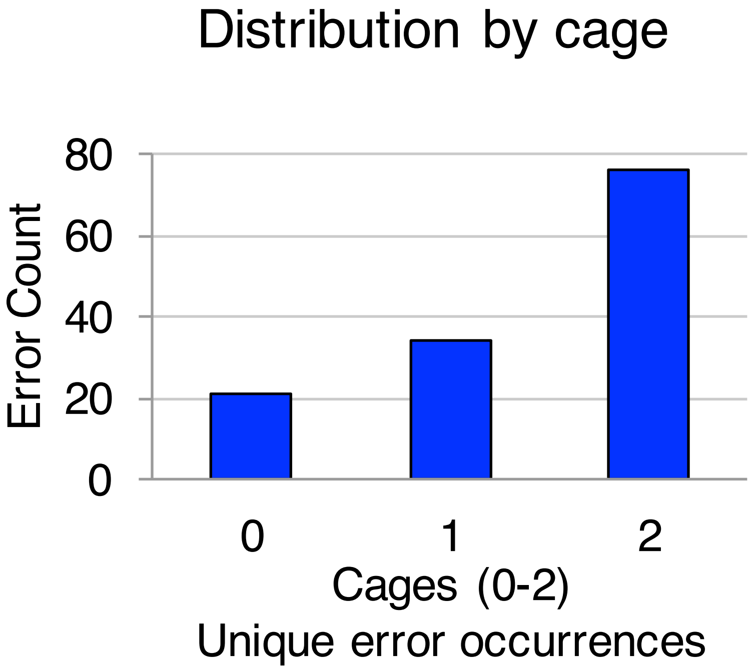


Figure 12. "off the bus" errors, indicating a problem with the PCIe connection from the SXM to the Cray XK7 MB.

PCIe connector issues show specific correlation to temperature, significantly more so than other failure conditions. Cage 2, which can be more than 30 °F warmer than cage 0, shows substantially more errors.

These errors are fatal. The node is lost, and the application fails. Frequently, hardware repair is necessary.

The occurrence of this failure has dropped dramatically since the connector mechanism was reworked to provide more flexibility, reducing susceptibility to heat-related expansion/contraction cycles that caused a hard failure.

[Off the bus] errors – temporal analysis

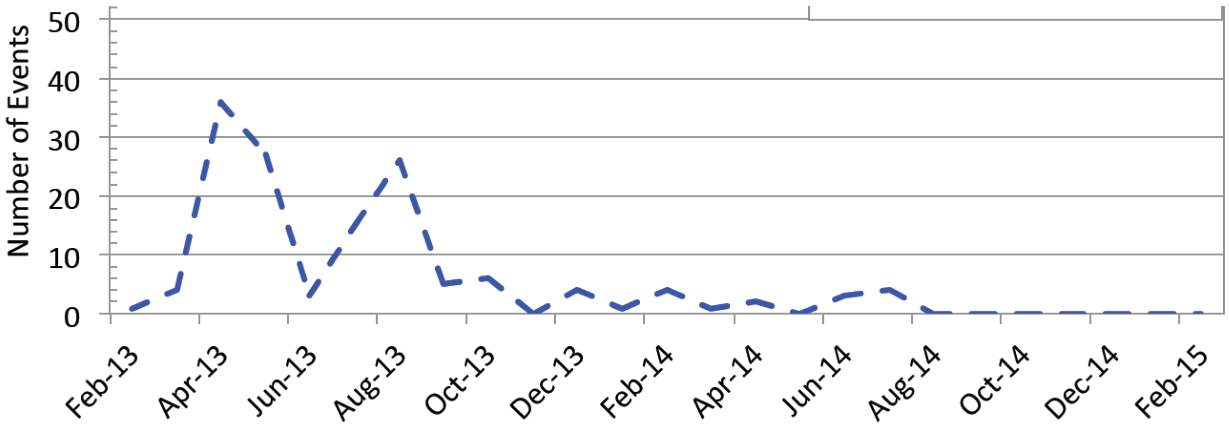


Figure 13. PCI Lane Degrades over time. High failure rates were tracked to stress on the connector that was exacerbated by thermal expansion and contraction. The OEM made significant changes to the physical locking mechanism to reduce the locking forces, and allow appropriate thermal movement.

An off-the-bus error is always followed by an XID 62 (microcontroller halt). This error is fatal to the application.

The technical success of the rework strategy is clear, with these errors effectively eliminated.

Content

- The OLCF’s Cray XK7 Titan
  - Science challenges for the OLCF in the next decade
  - Mechanical packaging for the Cray XK7
    - Cabinet, cage, blade, and node descriptions
  - The NVIDIA GK110 layout
  - NVIDIA Kepler memory architecture
- NVIDIA GPU error assessment
  - Error conditions on the K20x
  - Single Bit Errors – Distribution by row/column
  - Single Bit Errors – Distribution by structure
  - Double Bit Errors
  - Page Retirement Errors
  - [Off the bus] errors –spatial analysis
  - [Off the bus] errors – temporal analysis
- Takeaways…
- A Summit teaser

Takeaways…

- The NVIDIA GK110 and SXM form factor have proven to be very reliable.
  - MTTF for DBEs on 18,688 nodes: 7 days
- Track your SBEs - individual SBEs are not logged by the driver.
  - ORNL snapshots the hardware counters on individual nodes at the end of any job on that node. This provides time stamps per node at scheduler granularity [< 24 hours].
- Consider whether your applications need ECC. This feature can significantly improve reliability/reduce silent data corruption for applications, but at the expense of memory bandwidth.
- NVIDIA provides mechanisms for tracking all other error types (DBE, page retirement, etc) - make sure that you are watching these as indicators of early- and late-life failures on specific cards.
  - Watch FIT/MTBF metrics closely. As electronics age, they will exceed operating margins and fall out of conformance.

2017 OLCF Leadership System

Hybrid CPU/GPU architecture

SUMMIT
OAK RIDGE NATIONAL LABORATORY

Vendors: IBM, NVIDIA, and Mellanox

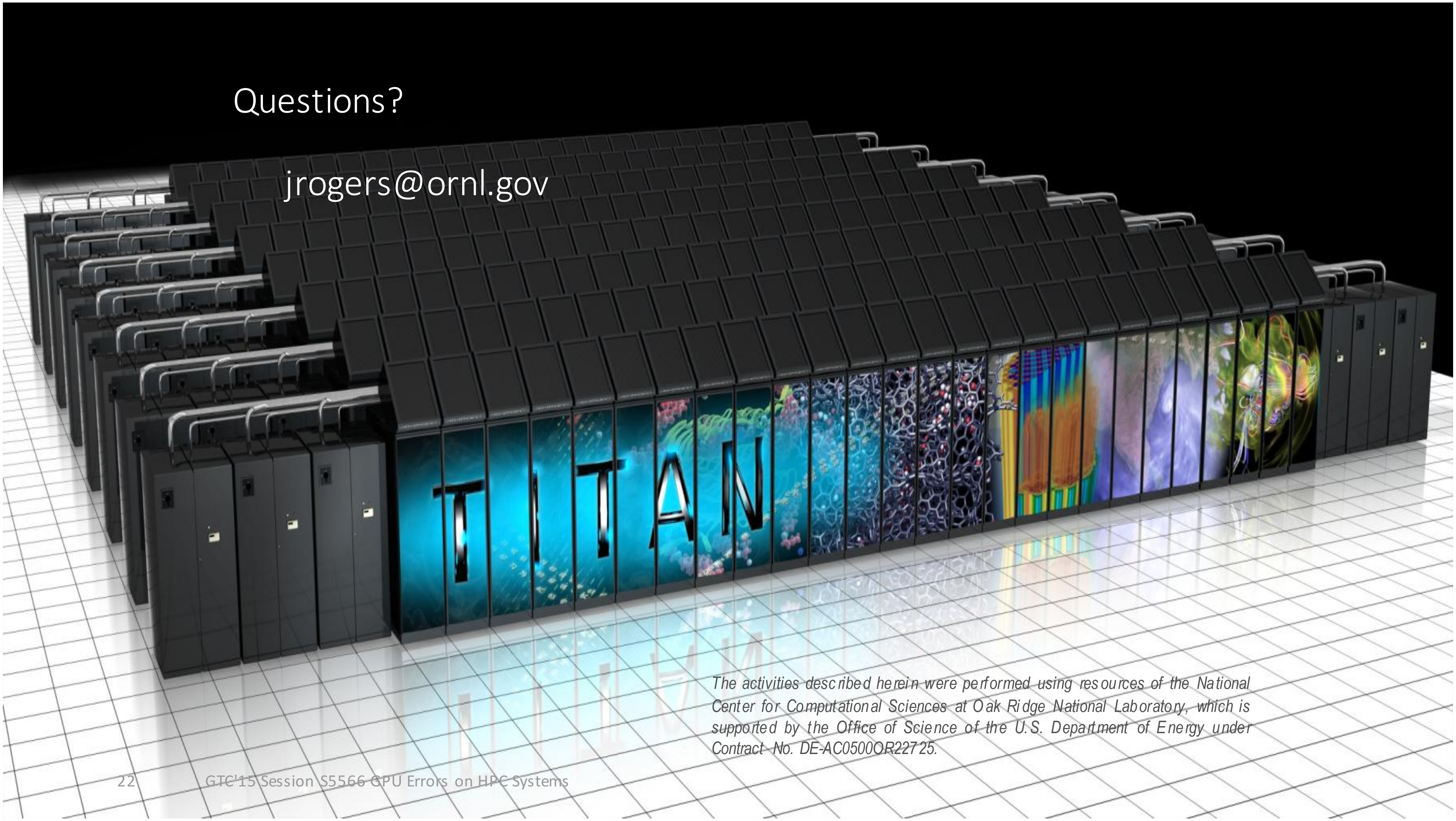
Baseline is 150PF (peak) and 5X Titan's Application Performance

More than 3,400 nodes, each with:

- Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta™ architecture
- CPUs and GPUs completely connected with high speed NVLink™
- Large coherent memory: over 512 GB (HBM + DDR4)
- An additional 800 GB of NVRAM, which can be configured as either a burst buffer or as extended memory

Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect

IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.



Questions?

jrogers@ornl.gov

The activities described herein were performed using resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC0500OR22725.