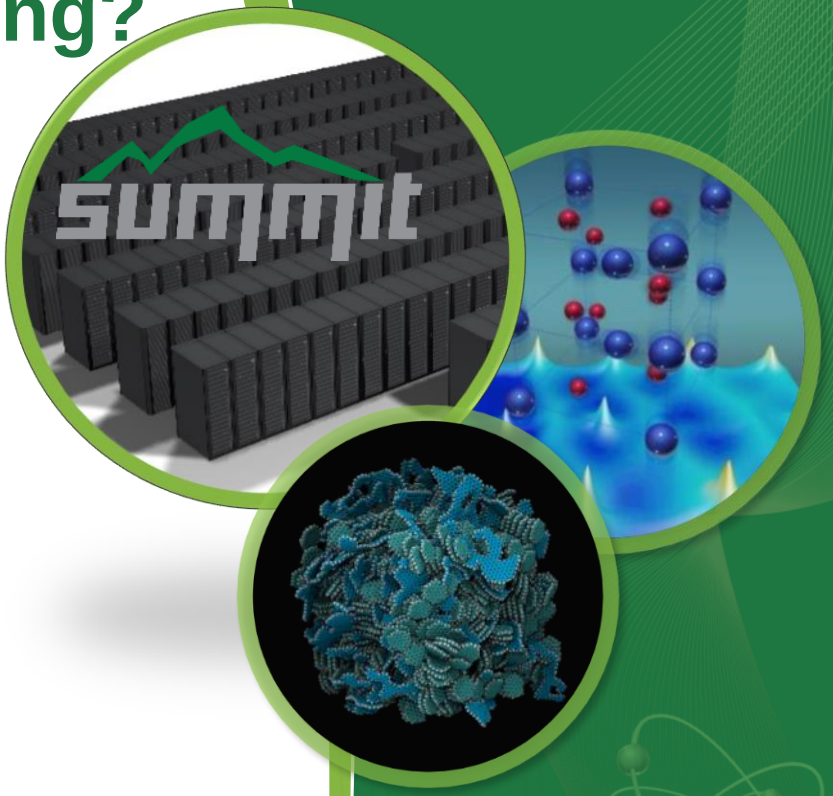
Lessons Learned in 40 years of HPC
or
How can anything possibly go wrong?



Energy Efficient HPC Working Group
SC'17 Workshop
Denver, Colorado
November 12, 2017

Buddy Bland
Program Director
Oak Ridge Leadership Computing Facility

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

How has HPC changed in 65 years at ORNL

800x cost
14 *trillion* times faster!

1953: ORACLE

- 14 KFLOPS – world's fastest
- 40 bit word
- 2,048 words of electrostatic memory

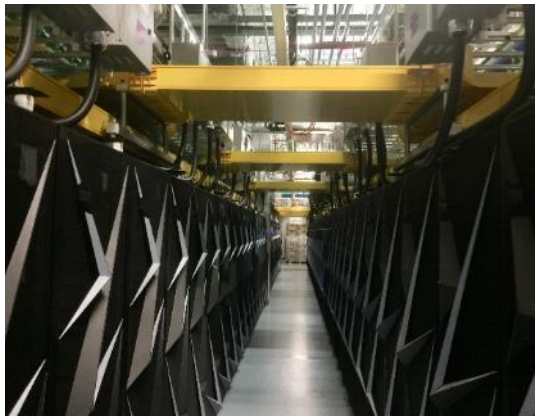
UNION CARBIDE PHOTO No. 15296-216

The ORACLE (Oak Ridge Automatic Computer and Logical Engine) in use at Oak Ridge National Laboratory, showing the arithmetic unit, electro-static memory, and the power-control cabinet. It has been estimated that the ORACLE can accomplish 100 man-years of computing in a standard eight-hour day.

- 2,560,000 word magnetic tape
- 75 kW power
- 5,000 vacuum tubes
- $250,000

2018: Summit – IBM AC922

- ~ 200 PFLOPS ~ $200M
- 4,608 nodes each with:
  - 2 POWER9 processors
  - 6 NVIDIA V100 GPUs
  - 512 GB DDR4
  - 96 GB HBM2
  - 1,600 GB NVMe
  - NVLink2 cache coherent link between POWER9 and V100 processors



- ~ 13 MW power
- Total memory: 24 PB DDR3, 400 TB HBM2, 7.3 PB NVMe
- 250 PB disks

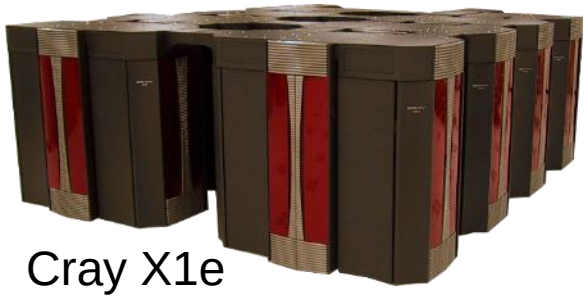
Machines I have worked on over my career (1980-2017)

Cray 1S

Cray X-MP

IBM SP-2

IBM SP3


Cray X1e

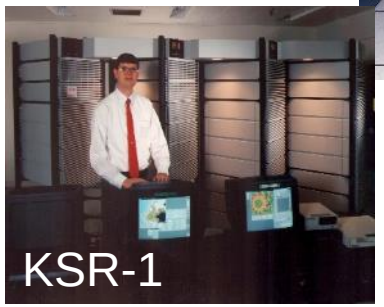
IBM BG/P

Compaq Alphaserver

SGI Altix

Cray XT3

CDC 7600


KSR-1

SRC-6

Cray XT5

Cray XT4

Cray 2

Intel Paragon

Cray XK7

IBM SP4

Installation at OLCF

Cabinets installed on concrete floor with power, cooling water, and network cabling overhead

180 miles of fiber optic Infiniband cable linking 4,608 compute nodes and 30,000 disks via 18 EDR Director switches and 500+ in-rack switches

Today's leadership systems are so complex that we will always have something broken

- It is rare that we have all 18,688 nodes running on Titan
- But in some ways, it has always been this way
  - Every SC conference has had talks on system resiliency and reliability – this conference has 8 workshops, papers, panels, and BOFs that address these subjects
- Despite our best efforts, sometimes problems get through
- In this talk I will give you some examples of some problems that I have seen over my 40 years working in HPC and perhaps a few lessons we have learned along the way.
- I would like this to be a discussion. Tell us what you have seen and the lessons you have learned.

We all know about software problems

- 512 single processor nodes
- Lights on the front panel that showed processor and interconnect usage
- First time we booted the machine, it was a serial boot. We watched the nodes come to life
  0, 1, 2, …, 253, 254, 255, 0
- The boot node counter was an 8-bit variable so only half of the nodes booted
- Intel didn’t have the capability to build out the machine so it had never been tested on more than 4 cabinets

Intel Paragon XP/S 35

But hardware problems are much more difficult to correct

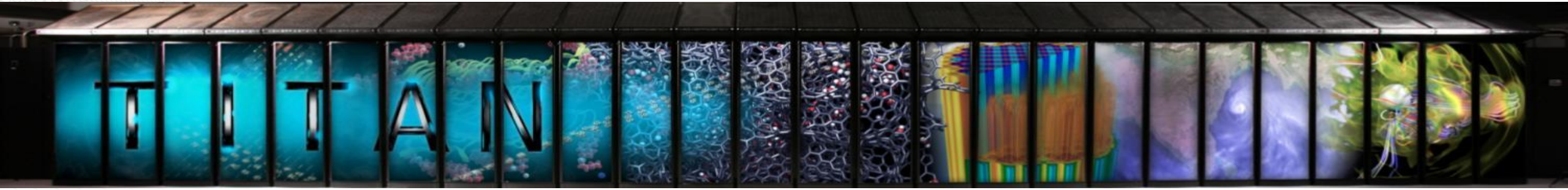
- When we built our first large Lustre file system, we found anomalies in the performance.
  - Some of the object storage targets were running slower than others.
  - In a parallel file system, the slowest OST gates the performance of the entire file system.
- We ended up testing every OST and found some of the disks were slower than others. Ultimately the vendor replaced the slow drives so that we could get the contracted performance.

Spider file system
10,000 one-terabyte disk drives

Upgrades are often risky business

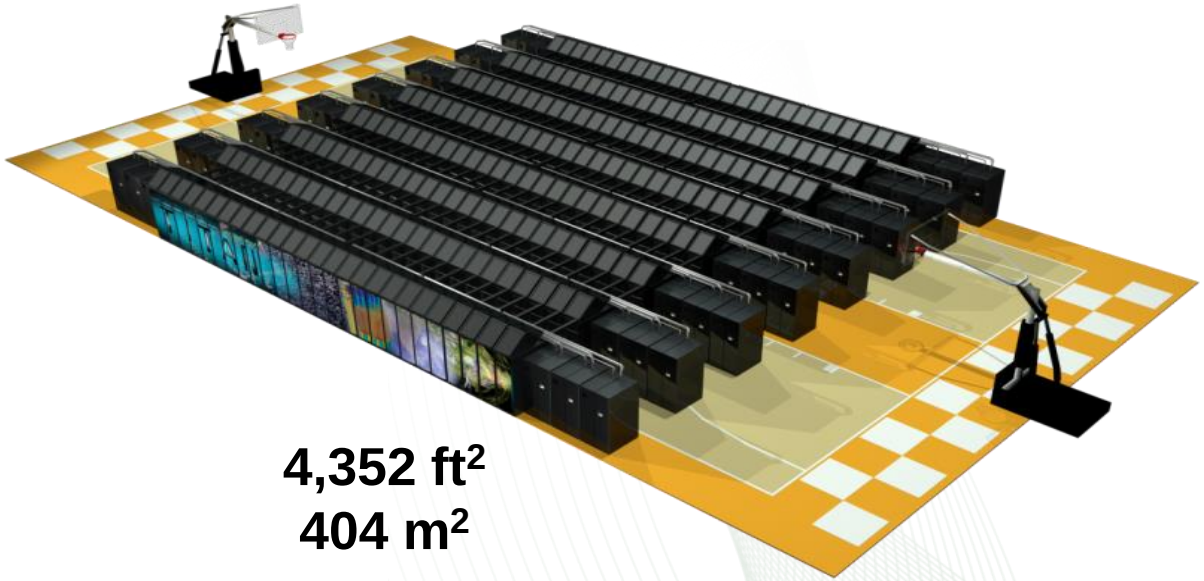
- Eagle: IBM Power3 upgrade to Power3+
  - This was intended to be a simple processor upgrade
  - The Power3+ processors had a bug that had to be fixed . . . Twice!
  - We had to replace every processor in the machine three times, only one of which was planned.

ORNL's "Titan" Hybrid System: Cray XK7 with 18,688 AMD Opteron and NVIDIA Tesla processors



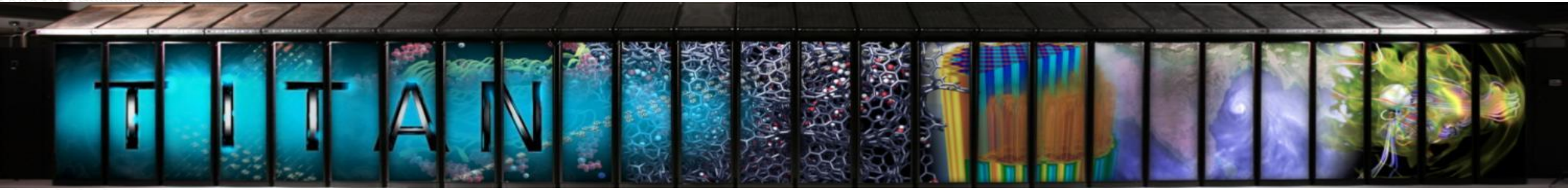
Titan was an upgrade of Jaguar

- Reused the cabinets, power supplies, cooling, backplanes, and interconnect
- Replaced node boards, fans, and added power supplies
- Upgrade was done on one-third of the machine at a time while users ran on the other two-thirds
- We did a rigorous acceptance test after each segment was upgraded.
- And yet ….



4,352 ft^2
404 m^2

ORNL's "Titan" Hybrid System: Cray XK7 with 18,688 AMD Opteron and NVIDIA Tesla processors



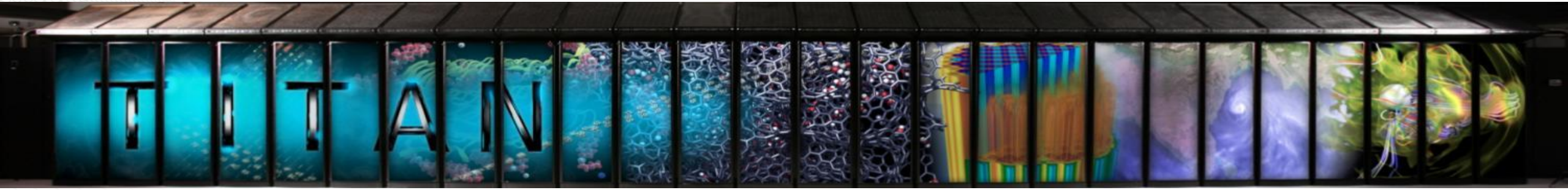
Titan was an upgrade of Jaguar

- Reused the cabinets, power supplies, cooling, backplanes, and interconnect
- Replaced node boards, fans, and added power supplies
- Upgrade was done on one-third of the machine at a time while users ran on the other two-thirds
- We did a rigorous acceptance test after each segment was upgraded.
- And yet ….

After all of the system was upgraded, we started seeing a few applications giving unrepeatable answers when running on the CPUs

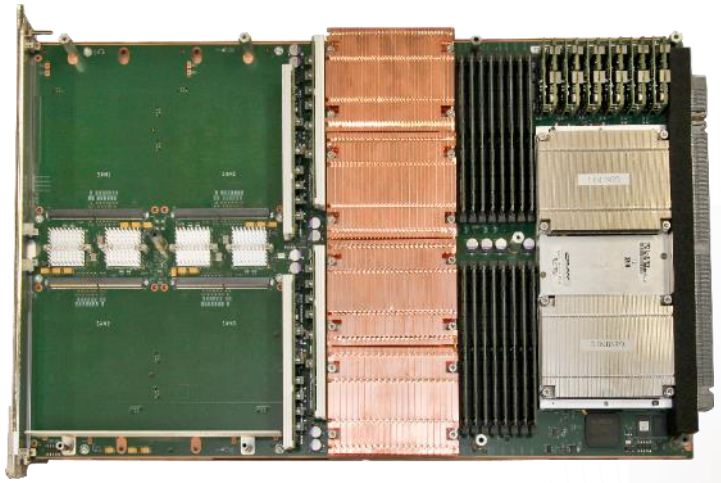
- After extensive testing and lots of discussion with Cray and AMD, we were able to demonstrate that there was a bug in the Opteron processor.
- We were the only ones to find the bug because it happened so rarely that you needed to test tens of thousands of processors to see it.
- The bug turned out to be a test escape that was already corrected before we found the problem but we still had to replace 18,688 processors with new ones.

ORNL's "Titan" Hybrid System: Cray XK7 with 18,688 AMD Opteron and NVIDIA Tesla processors

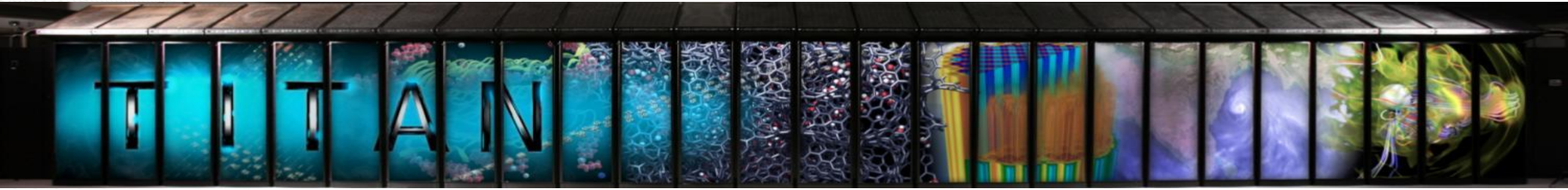


Some problems are mechanical

- The GPUs in Titan are on a daughter card that connects to the mother board via an "L" shaped connector.
- The daughter card is then attached with screws on all four corners to the motherboard.
- Over time, we started seeing GPU modules drop off of the PCIe bus.
- Analysis showed that the solder of the "L" connector was cracked.
- Cray sent a team to ORNL to resolder all of the connectors.



ORNL's "Titan" Hybrid System: Cray XK7 with 18,688 AMD Opteron and NVIDIA Tesla processors



Some problems are mechanical

- The GPUs in Titan are on a daughter card that connects to the mother board via an "L" shaped connector.
- The daughter card is then attached with screws on all four corners to the motherboard.
- Over time, we started seeing GPU modules drop off of the PCIe bus.
- Analysis showed that the solder of the "L" connector was cracked.
- Cray sent a team to ORNL to resolder all of the connectors.

But this didn't fix the problem!

- Further analysis showed that the solder was cracking over time.
- The GPUs are 250 watt chips. As they heated up with heavy usage, the daughter card would expand, putting pressure on the connector and breaking the solder.
- The fix was simple. Instead of anchoring the card on all four corners with screws, a plastic mount was used on the end away from the connector. The mount would flex, eliminating the pressure that broke the solder.

Not every issue is caused by the system

Some are power, cooling, or other facility problems

IBM Power4 - Cheetah

- When first installed the computer room had just enough air for the system.
- We found that some cabinets were overheating because the air distribution was not uniform.
- Our analysis found that cool air was getting short circuited back to the AHU. Our computer center manager used a piece of cardboard and some duct tape to block the airflow path to allow the machine to run until we could make the required changes to the computer room.

Air flow can be a problem in other ways

- We needed to put our Cray XT4 system in a room that had a low ceiling.
- The airflow in the XT4 cabinet is bottom to top, but we only had 18 inches of height above the cabinet to the suspended ceiling.
- Since we had so many cabinets, we modeled the airflow and found that we would get back pressure from the ceiling.

- We removed the acoustic ceiling tiles and replaced it with expanded metal to allow the air room to flow.

How much room is under your floor?

- This system was first installed in a 1950s era computer room. 8 inch raised floor, 60 lbs/ft^2 floor loading
- The power connected to the bottom of the machine. We had to lift a 4,000 pound system about 4 inches to be able to make the connections, then lower it once all the power was connected.
- To fit the system in the elevator, we had to remove the elevator doors. A workman sat on top of the elevator cab to override the door lockout switch.
- On larger systems, even a 1 meter floor may not be enough if you have large insulated pipes that need to cross each other.

Cray X1

Remember the Paragon with the 8-bit boot counter …

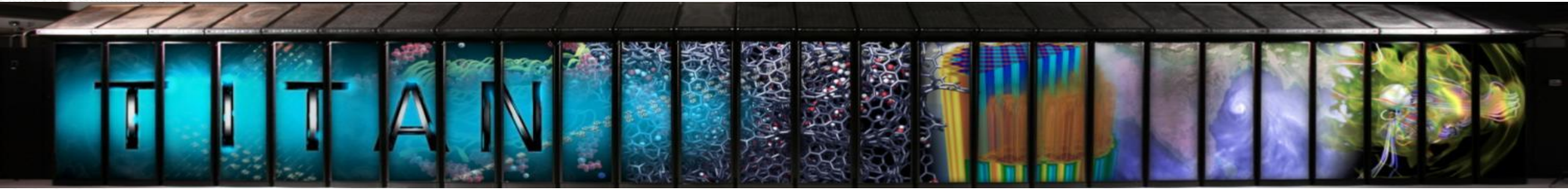
Intel Paragon XP/S 35

- Once we got it running we found another problem
- The room kept getting hotter
- In just a few minutes, the room temperature got to 95F (35C) and the machine shut itself down
- The 40 year old pipes from the chillers to the AHUs were sized correctly, but had so much scale in them that the water flow was too low to cool the room
- We had to get new piping installed in about 2 weeks

But wait! There's more

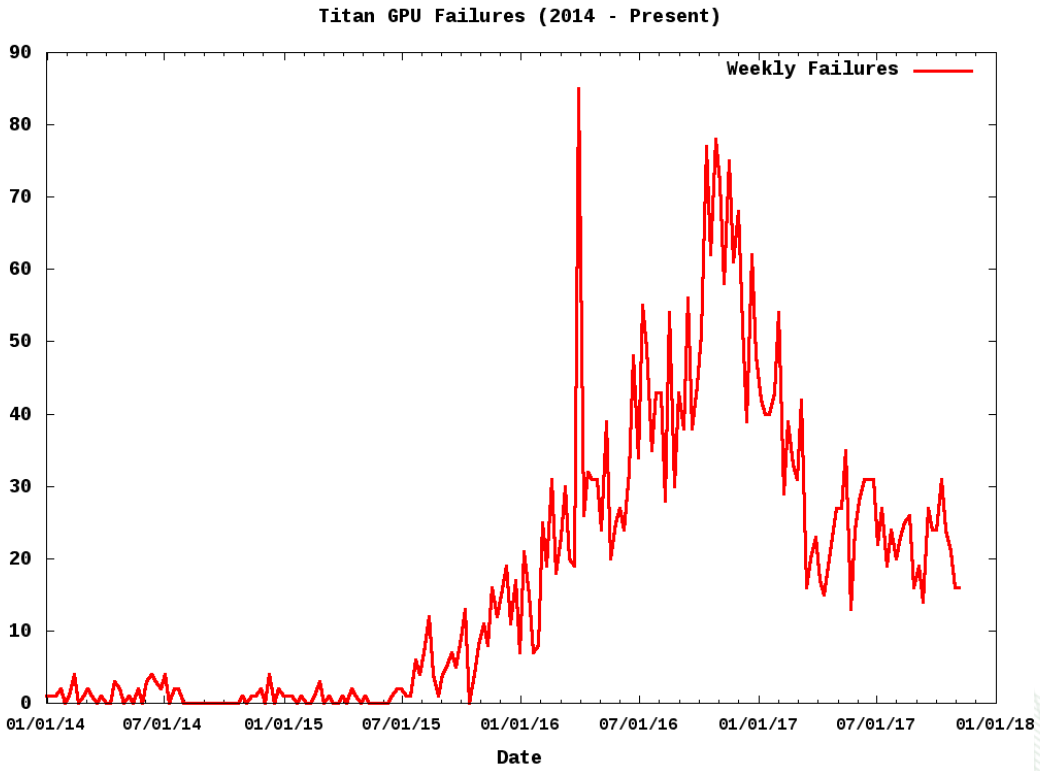
- When did you last check your fault current limits?
- After the Paragon had been running for many months we found a bigger problem
- A circuit board had an internal defect that eventually caused a short circuit
- The cabinet had a 100 amp circuit breaker, but even a dead short didn't draw 100 amps
- The circuit board smoked and burned a hole through the board with black soot coating the inside of the cabinet
- In the pre-VESDA days, we were lucky that the operator smelled the smoke and shut down the machine

ORNL's "Titan" Hybrid System: Cray XK7 with 18,688 AMD Opteron and NVIDIA Tesla processors

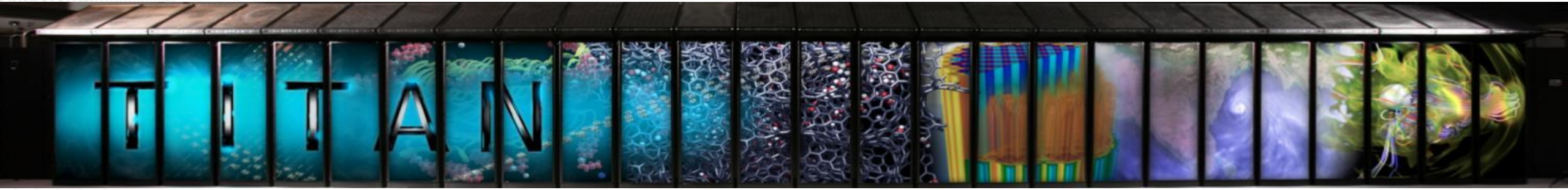


Not all problems happen quickly

- Over three years after Titan was put into service, it started reporting double bit memory errors on the GDDR5 memory.
- The failure rate increased very slowly but eventually began to get to the point that it was noticeable by users who had trouble getting long-running, large node count jobs to complete.
- Cray and NVIDIA did extensive analysis and found a resister on the daughter card was failing, causing the GDDR5 memory to stop working.



ORNL's "Titan" Hybrid System: Cray XK7 with 18,688 AMD Opteron and NVIDIA Tesla processors



Not all problems happen quickly

- Over three years after Titan was put into service, it started reporting double bit memory errors on the GDDR5 memory.
- The failure rate increased very slowly but eventually began to get to the point that it was noticeable by users who had trouble getting long-running, large node count jobs to complete.
- Cray and NVIDIA did extensive analysis and found a resister on the daughter card was failing, causing the GDDR5 memory to stop working.

Silver in the resister reacted with sulfur causing the resistance to change (NOT a GPU problem). But where was the sulfur coming from?

- The assumption was it was environmental sulfur in our computer room, but that didn't pan out.
- Our materials science group at ORNL analyzed both failing and good parts in much more detail than either NVIDIA or Cray could. No smoking gun, but it appeared that the sulfur contamination came from the card assembly location.
- We have replaced over half of the GPU cards with new versions that have sulfur resistant resistors.
- NVIDIA had to manufacture new cards since no one had enough spares to meet the need.

Some problems are not found until it's time to remove the machine

- When it was time to decommission this system, our property office sold the $10M machine to a local scrap dealer for a few hundred dollars.
- When he came to get the machine, one of the property people didn't understand that old computers have little or no value.
- He laid down in the street to block the truck from leaving with the system for several hours while the misunderstanding was resolved.

Cray X-MP

What are the key lessons from all of this?

- Don’t assume that your system has been tested at the factory at full scale.
- Plan for time to debug large systems after they have been assembled at your facility. This is especially true for software.
- When planning an upgrade, plan for what you will do if the machine doesn’t come up on schedule.
- Many problems will never be seen at small scale. They happen so rarely that you won’t see it until you have a very large system.
- Don’t assume that your acceptance testing will find all the problems. Many are rare and it takes a long time for them to show up, happen often enough to be noticed, replicated, and debugged.
- Keep checking results periodically after acceptance and try to track down every glitch. (very hard)

What are the key lessons from all of this?

- Big machines can use as much power and cooling water as a small town.
- An expert electrical power engineer is a must.
- Same for an expert mechanical engineer. This is especially true if you want a low PUE.
- Sometimes the first problem you find isn't the root cause. It takes thorough analysis to find the problem.
- Don't just assume that the vendor is always right in debugging problems. It's really nice to have your own X-ray photoelectron spectroscopy and scanning electron microscopy systems and a team of material scientists to call on when you need it! ☺
- Do you know that your vendor has enough spare parts for the projected life of the system, or contracts to provide them from a supplier? What if something bad happens?

Most importantly, It takes a team to build leadership class, first-of-a-kind systems

- The senior managers of the customer and vendors must talk frequently. If you wait until there is a problem, it will be hard to fix if it is the first time the leadership have spoken.
  - ORNL's lab director and associate lab director for computing talk with our vendors every other week and have for 12 years.
- The technical teams for each of the subsystems talk with their vendor counterparts weekly to insure that there are no misunderstandings.
- When there are issues, they are quickly elevated to people with the power to fix problems. Don't let problems fester. You never know what problems will end up taking a long time to fix or how many problems you have left to find and fix.
- Customer and Vendor must treat each other with respect

Questions? BlandAS@ORNL.GOV We are always hiring