Sandia National Laboratories

Exceptional service in the national interest

A User Perspective on HPC Challenges and Diagnostics

a case study

M. Scot Swan, David Sirajuddin, Keith Cartwright, Chris Moore

CCE MMAI Meeting

2021-08-19

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. ...ation under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

NNSA

HPC Case Study using EMPIRE

EMPIRE is Sandia's next-generation plasma simulation tool. It is able to simulate plasmas over a broad density range, with PIC dominating at low densities, fluid at high densities, and a hybrid approach in the middle.

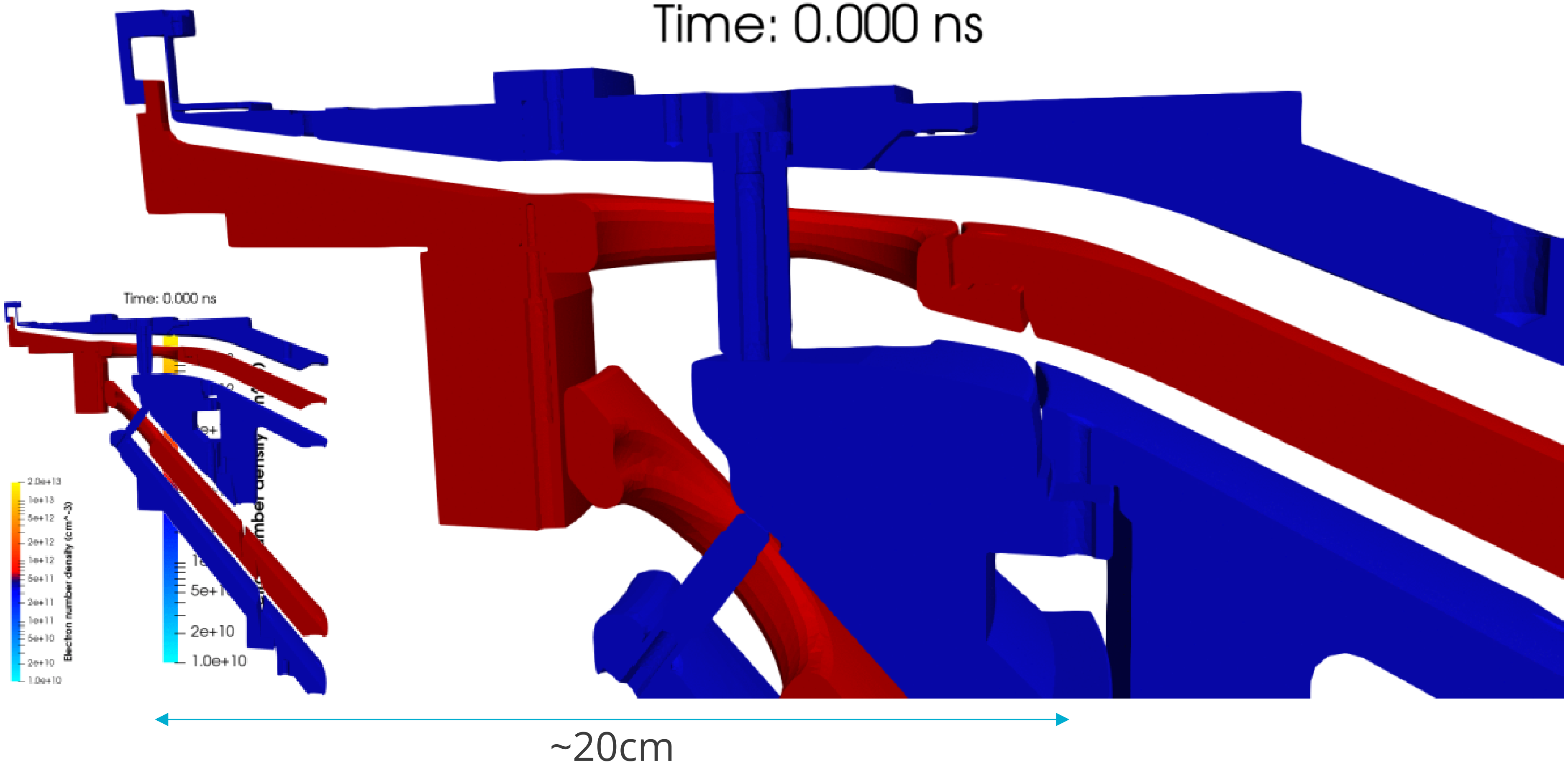
During the summer of 2021, we had an opportunity to do 6 runs that were allocated 7 days and 10,440 cores each (290 nodes, ~20% of the cluster).

Highlights:

- Allowed us to run 3 types of simulations (PF18A, BDot, RKA) with 2 runs each
- Allowed us to run to later times
- Allowed us to run at higher fidelities (less stochastic noise from PIC)
- Gave users experience scaling up simulations
- Exercised new HPC diagnostic tools
- Found out how better diagnostic tools can help developers and users

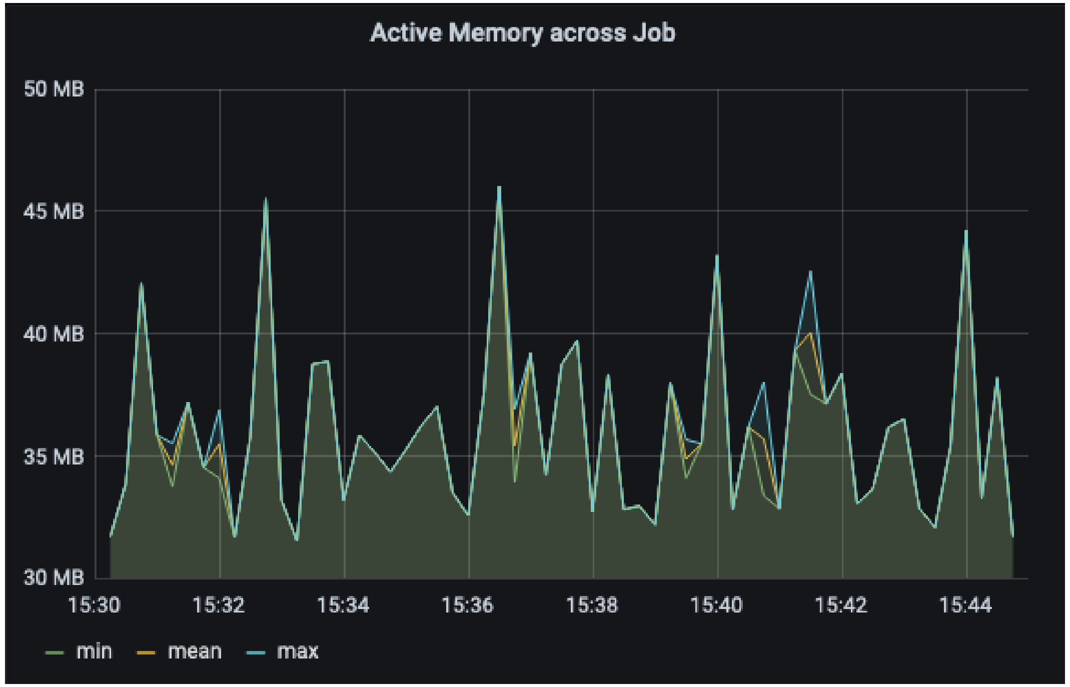
PowerFlow 18A (PF18A)

The PF18A simulation is of the convolute (convergence point) of the Z-Machine. It investigates the power loss due to plasma generation while up to 26 MA move through it to the load.



PF18A Debrief

The PF18A simulations were the first set of large simulations run this summer and they were the most uneventful. There were some growing pains where initial resource estimates for the refined simulation were not accurate, as well as finding the optimal balance between MPI ranks and number of threads.



Screenshot of LDMS Grafana memory dashboard showing a recent small, balanced run.

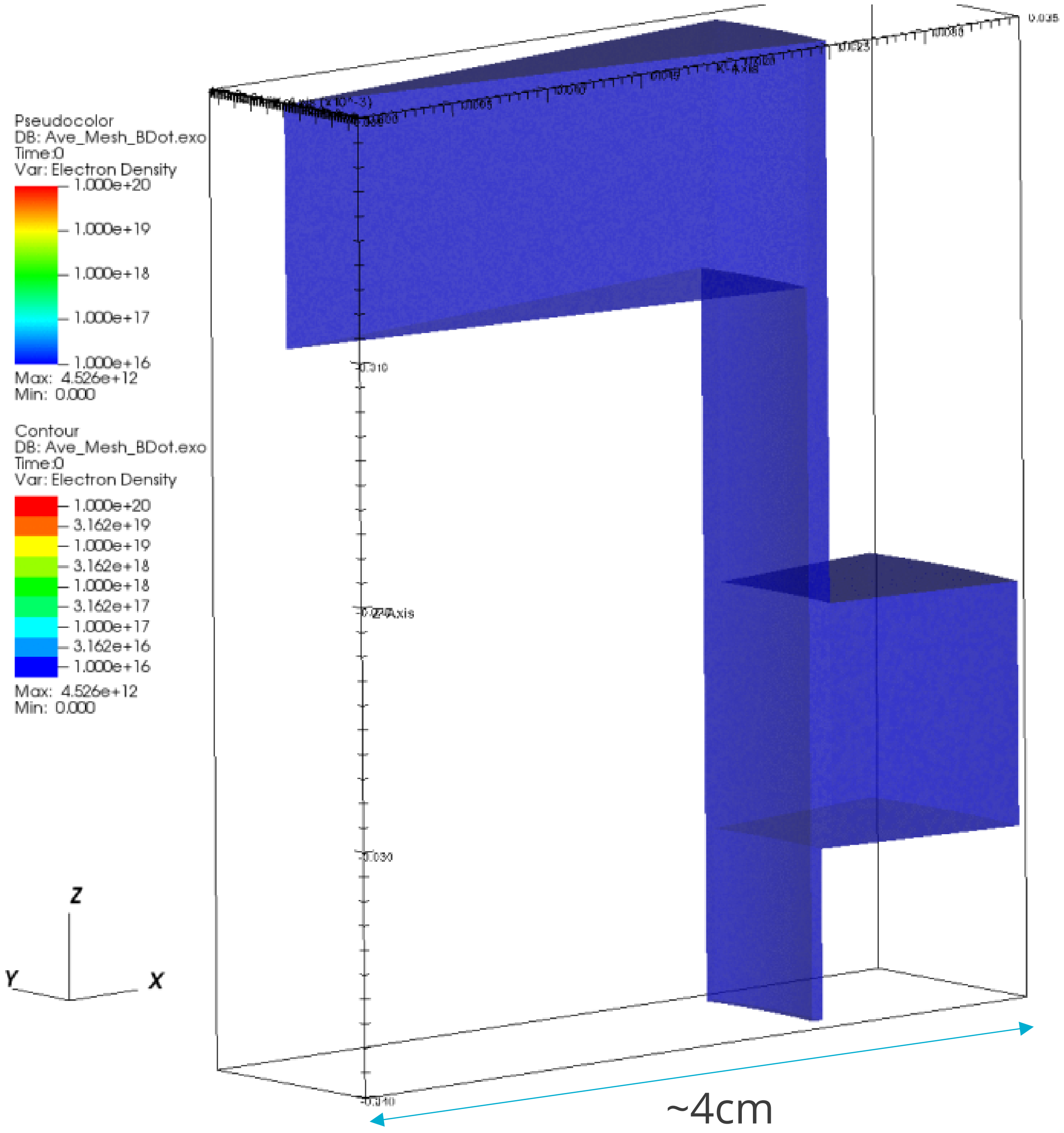
🧮 HPC Success Story

The LDMS real-time diagnostics allowed us to view memory usage on a per-node basis and to determine actual memory requirements. This enabled us to account for unbalanced resource usage and to quickly and correctly size our problem to fit within 290 nodes.

BDot

An X-ray source illuminates from above (through a thin metal foil) and causes electrons to launch upwards. Electrons are absorbed by the foil and then create a clockwise current, generating a magnetic field in the lower-right cavity where the B-Dot current sensor is located.

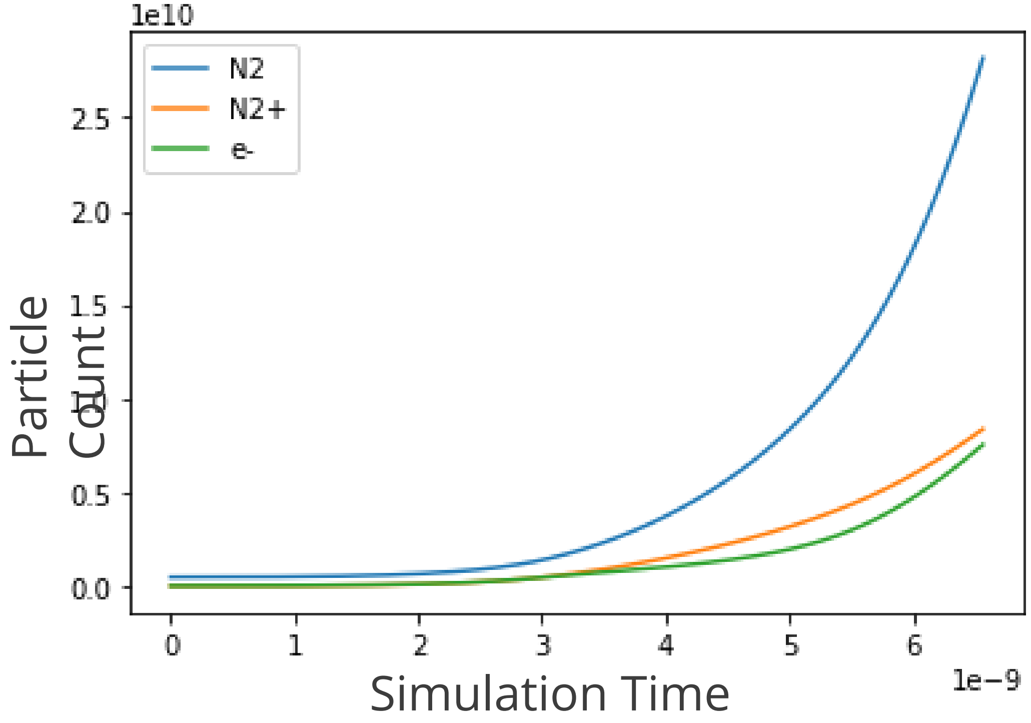
Cavity is pre-filled with low pressure nitrogen gas. At peak current, about 7,000 A are crossing the gap.



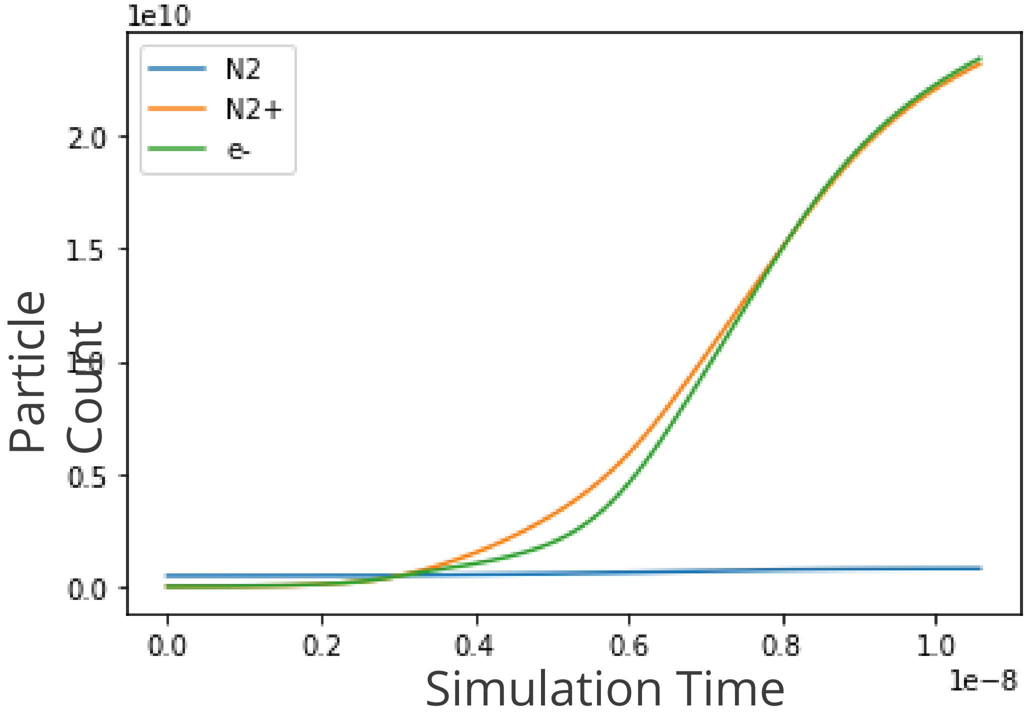
BDot Debrief

The BDot simulations were the second set of large simulations run this summer and they were the most eventful. Essentially the first week-long allocation was used entirely by managing scaling issues internal to EMPIRE.

Without Neutral Merge

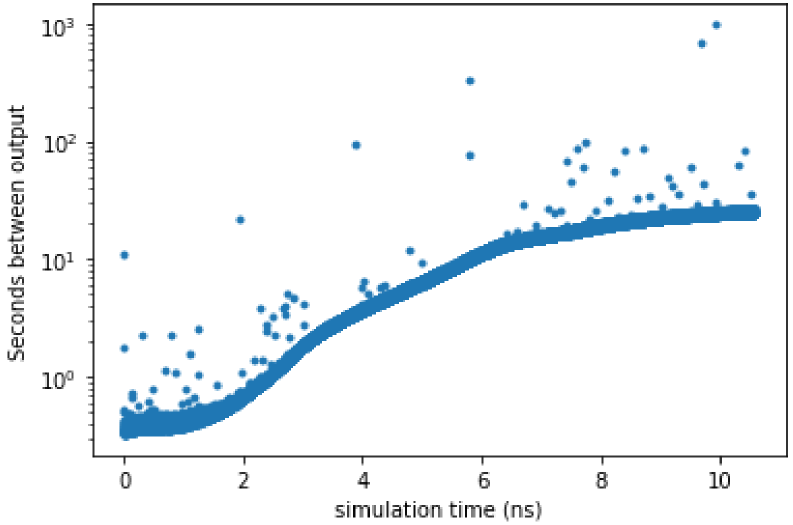


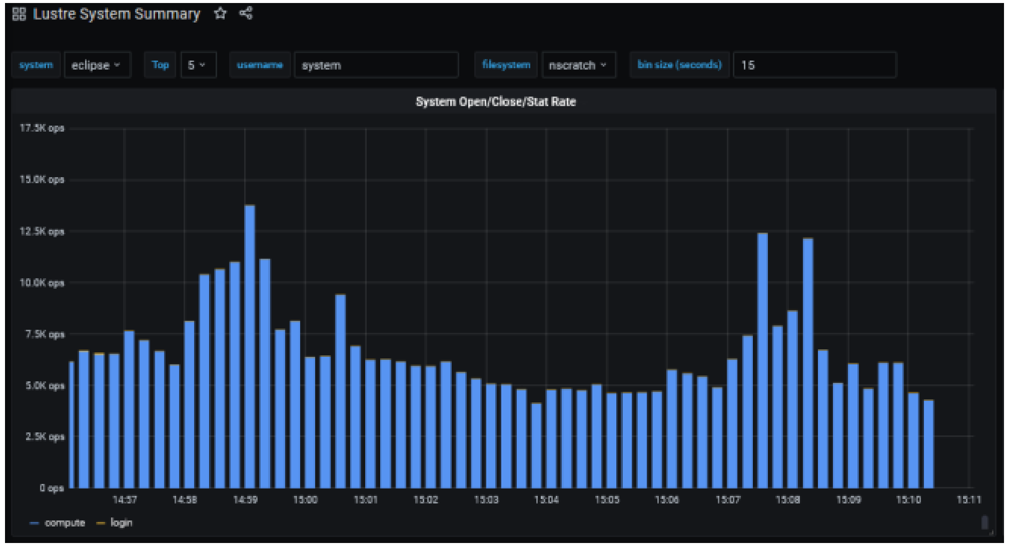
With Neutral Merge



BDot Debrief (continued)

With the second week-long allocation, we finally got a BDot simulation to run, but it didn't get as far as we had hoped or expected. Particle creation was still occurring, but not exponentially. Towards the end of the simulation we were averaging 1ns simulated for 24h wall time.





Lustre System Summary

system eclipse

Top 5

username system

filesystem nscratch

bin size (seconds) 15

System Open/Close/Stat Rate

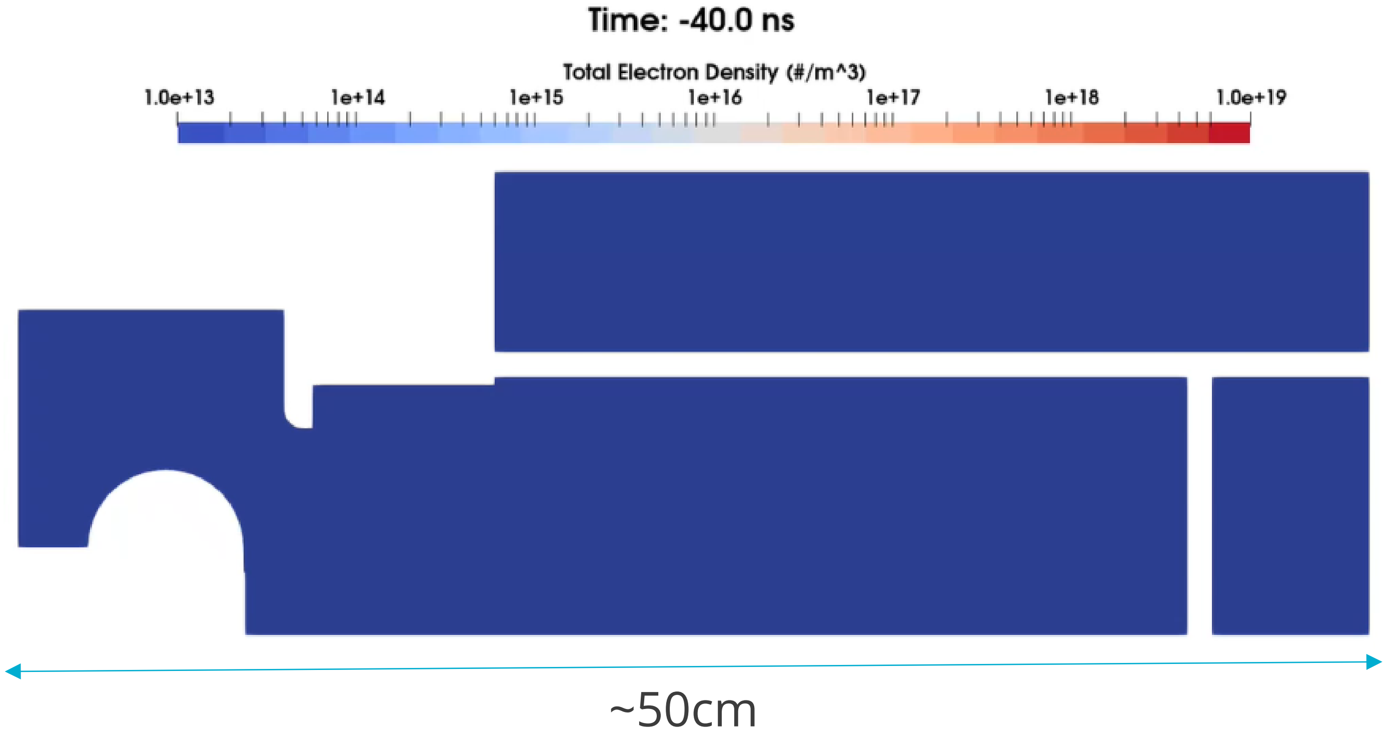
Screenshot of LDMS Grafana IO dashboard

HPC Success Story

The LDMS real-time diagnostics helped us figure out why some time steps took up to 20 minutes while most were only 20 seconds. We were able to determine that scratch drive congestion was a contributor (the last checkpoint/restart dump was 4TB and it takes a while to write that to disk)

Relativistic Klystron Amplifier (RKA)

The last of the 3 simulations. The RKA set up was as easy as the PF18A. This simulation is of an electron beam propagating in Argon. Electrons emit from the rounded object on the left, eventually being absorbed by the walls and producing a counter-clockwise current back to the emitter. Kinetic energies vary up to ~500 kV.



RKA Debrief

The first simulation ran without incident, other than not quite getting as far as we had hoped. The second simulation ran until about 70ns (102h wall time) and then crashed due to a “No space left on device” error on the 14PB scratch drive (our 3TB simulation probably didn’t cause the failure). It took us about 79 hours to get a viable job in the queue as we had to wait for the scratch drive to stabilize and work through inexplicable executable troubles.

🧮 HPC Horror Story

We were already behind schedule to finish the RKA runs and these hardware issues pushed us even further behind. In the end, we didn’t finish our RKA priority runs until 8 days after when we originally scheduled them to be finished with the HPC admins.

Soapbox

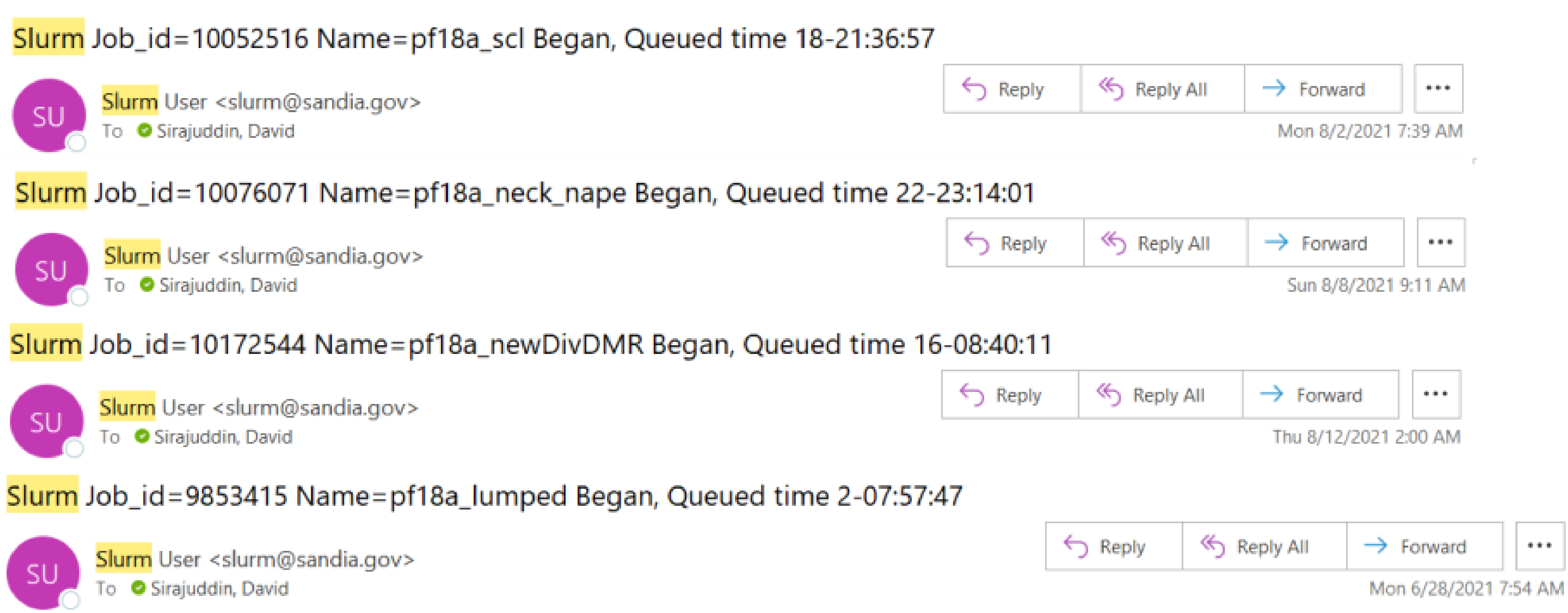
I asked the EMPIRE team what concerns they would like me to bring up and here they are:

1. Cluster environments need to mirror each other between SRN and SCN.
2. Having a modern version of python with frequently used packages everywhere is vital.
3. Build times on CTS1 are astronomical because of the linker we're forced to use.
4. We need a solution for getting large jobs running faster. Queues for 290-node jobs can be as high as 18-22 days even when the job is short. That's one reason why we couldn't debug our at-scale jobs without using our priority time.
5. Filesystem and node connection failures happen far too frequently and often ruin longer-running production runs.

What questions do you have for me?

Backup: Long Queue Times Without Priority

These submissions are for 290 nodes requesting 30-36 hours.



Slurm Job_id=10052516 Name=pf18a_scl Began, Queued time 18-21:36:57

Slurm User <slurm@sandia.gov>
To Sirajuddin, David
Mon 8/2/2021 7:39 AM

Slurm Job_id=10076071 Name=pf18a_neck_nape Began, Queued time 22-23:14:01

Slurm User <slurm@sandia.gov>
To Sirajuddin, David
Sun 8/8/2021 9:11 AM

Slurm Job_id=10172544 Name=pf18a_newDivDMR Began, Queued time 16-08:40:11

Slurm User <slurm@sandia.gov>
To Sirajuddin, David
Thu 8/12/2021 2:00 AM

Slurm Job_id=9853415 Name=pf18a_lumped Began, Queued time 2-07:57:47

Slurm User <slurm@sandia.gov>
To Sirajuddin, David
Mon 6/28/2021 7:54 AM