

RESEARCH EFFICIENCY: GAMIFYING RESOURCE MANAGEMENT ON DCC

Duke | Code<+>

NCShare

Aaron Dotson, Alex Li, Chloe Tang, Ian Wagner, Kelvin Quaicoe, Kyre Williams, Theo Olubodon



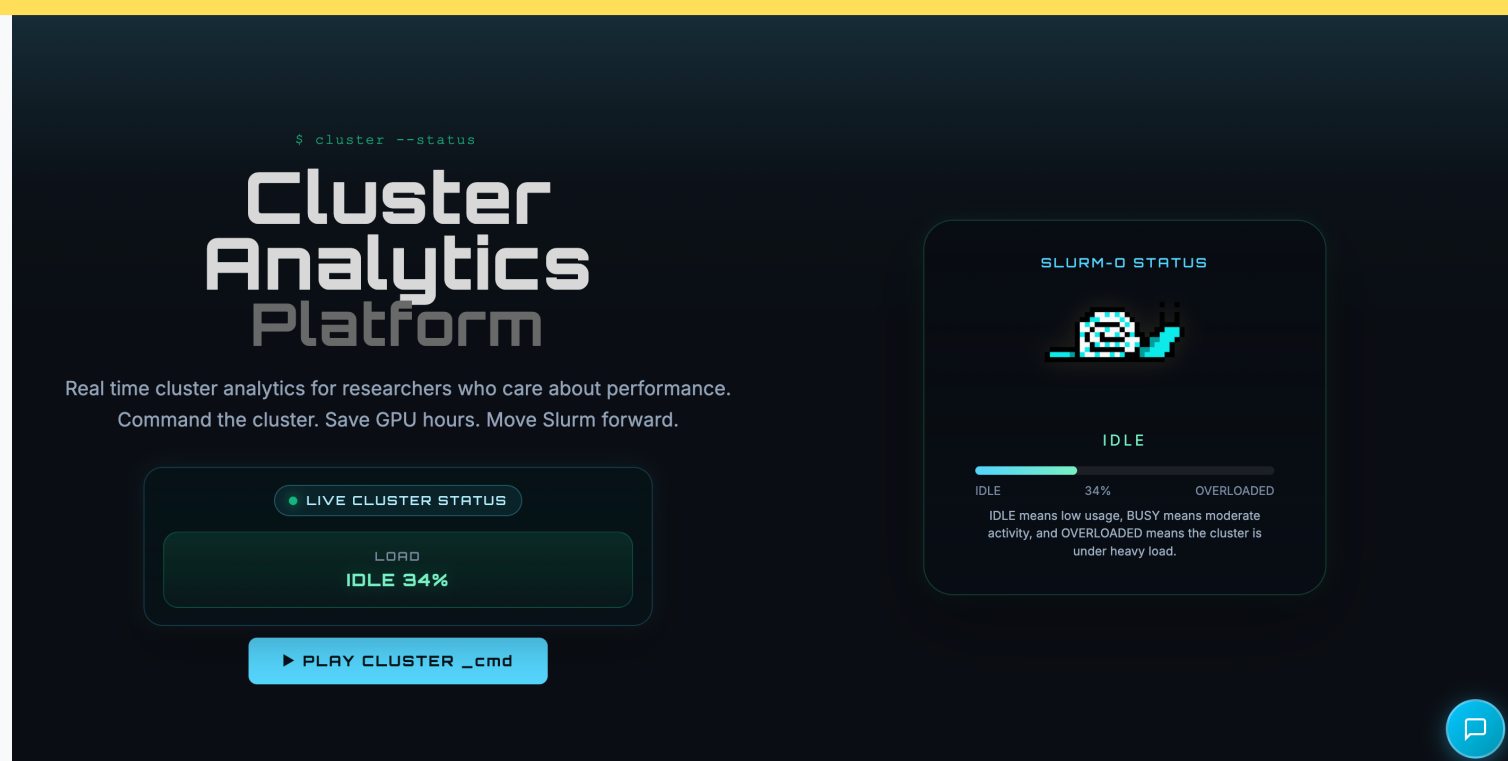
PROBLEM + BACKGROUND

HPC Researchers rely on the **Duke Compute Cluster**.

- Many **over-allocate** CPU, GPU, and Memory resources.
- Leading to **Wasted computing power, Higher energy use, and Fewer resources available for others.**



OUR SOLUTION



re currently free, with no pending jobs waiting for that GPU type. You also do not currently have any running GPU jobs or idle GPU allocations, so there is nothing you need to release before submitting this one.

That said, based on your recent usage, an A5000 may be more GPU memory than this workload usually needs. Your recent average GPU compute utilization is about 28.9%, and your average GPU memory utilization is about 24.8%. Those are two separate signals: the memory number suggests you are using only a fraction of the 24 GB available on an A5000 – roughly 6 GB if this 'animals' workload is representative – while the compute number suggests the GPU is also not being kept very busy. Moving to a larger or faster GPU would not help much unless the code is changed to feed the GPU more work.

Your recent completed 'animals' jobs also ran very quickly: job '11898478' completed in '00:03:49', and job '11910842' completed in '00:03:43', both on 1 A5000 with 24 GB RAM. For that pattern, a 2-hour time limit is much longer than needed unless this new run is substantially larger.

I also noticed several zero-second failures in the last 14 days, including '11884316' and '11884317' for 'animals' on A5000, as well as earlier 'mnist' attempts on 'rtx_5000' and 'rtx_2080'. Because these failed with '00:00:00' elapsed, they look more like launch, environment, path, module, or script issues than GPU capacity or wall-time problems. The later successful 'animals' jobs suggest you may already have fixed that, but if this new submission is similar to the failed ones, it is worth checking the batch script and stderr before scaling up.

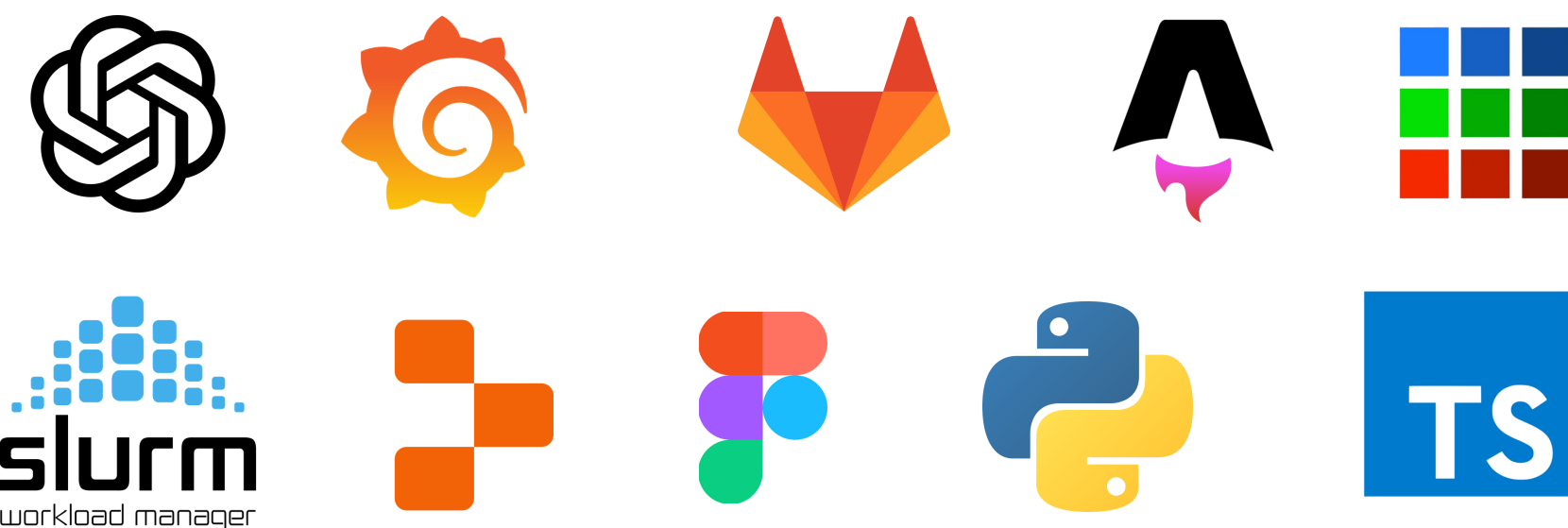
For future runs, I would consider routing smaller tests to the 11-12 GB cards first. Right now there are 20 free 'rtx_2080' GPUs with 11 GB VRAM, 32 free 'titanv' GPUs with 12 GB VRAM, and 19 free 'p100' GPUs with 12 GB VRAM, all with no pending jobs. If your job stays below about 9-10 GB of GPU memory, those are better fits than a 24 GB A5000. Keep using 'a5000' when you expect to exceed that range, need the A5000 performance/features specifically, or are doing a production run where the faster card is justified.

Concrete asks:

- Submit this 'animals' job on 'a5000' if this is a production run or if you expect it to need more than about 10 GB of GPU memory.
- For small/debug runs, try '--gres=gpu:rtx_2080:1', '--gres=gpu:titanv:1', or '--gres=gpu:p100:1' instead.
- If the workload is still similar to jobs '11898478' and '11910842', reduce the wall time from '02:00:00' to something like '00:15:00' or '00:30:00'.
- If you see another zero-second failure, check the job stderr/stdout first; it is likely a startup or environment problem rather than a GPU sizing issue.



TOOLS



ACKNOWLEDGEMENTS

NC SHARE Institution

Team Leads: Joe Shamblin & Uthpala Herath

Code+ Admins: Sheri Tibbs & Jen Vizas

Stakeholders: Dr. Charley Kniefel & Dr. Alejandro Cardenas-Avendaño



OBJECTIVE

Creating an **Gamified** Website & that displays a Dashboard that helps **HPC / NC SHARE** researchers understand and optimize their computing resource usage.



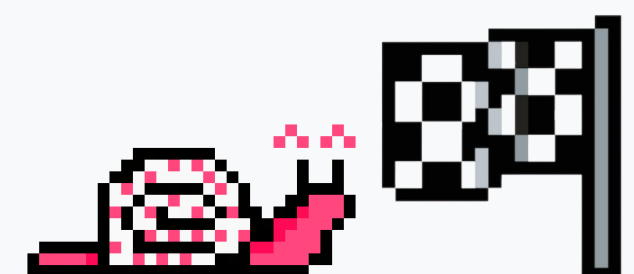
KEY FEATURES

- EXPRESSIVE AI CHATBOT:** Prompts Slurm, DCC documentation & Website leaderboard real-time data.
- WEBSITE & GAME:** Teaches researchers how the Cluster Operates / Information displayed on leaderboards, & User's Section.
- AI TERMINAL LLM:** In-Depth SLURM commands & GPU Feedback.
- BACKEND:** Grafana, Python & TypeScript.



FUTURE STEPS

- Improvements** to the Website Game, AI chatbot, & AI Terminal LLM
- Showcase across the **NC SHARE** Institution.



OUR GIT REPOS

SLURM AI TERMINAL: <https://tinyurl.com/3t9cr7j8>

WEBSITE: <https://tinyurl.com/y822cxfz>