Annual Report FY24-25

CRCD is always working to improve our service and capabilities, and to innovate. During 2024, we distributed a survey specifically for MPI users and a wide-ranging survey for all users. We closely studied the feedback from our users and worked to correct the problems they encounter. Our performance in 2025 and beyond reflects us listening to our users and responding to existing concerns, but also our team's exploration of the forefront of computing such as generative AI.

ChatCRCD

One of the most significant new offerings was ChatCRCD, an internally deployed large language model (LLM) to augment our support site. Built upon open-source foundational LLMs, we use AI techniques such as RAG (Retrieval-Augmented Generation), LoRA (Low-Rank Adaptation of Large Language Models), and Fine Tuning to provide CRCD-specific context to the LLMs.

ChatCRCD is becoming an integral part of CRCD's support services. It is trained on data from the CRCD context, including help tickets, and is accessible via a portal with a UI interface. For example, users could ask ChatCRCD how to create a job submission script for MATLAB, and it would be able to provide a usable job submission script. Essentially, users can ask any question that would go into a help ticket.

Our aim is that ChatCRCD will be able to provide not a generalized answer, but an answer that aligns with the specific setup of the CRCD ecosystem as a co-pilot to augment our user support site.

Teach Cluster

CRCD offers computing resources for coursework as well as for research. After hearing feedback from instructors and students who sometimes have difficulty accessing resources for term projects, we have created a dedicated teaching cluster with access to CPUs and GPU from a new JupyterHub portal, https://jupyter.crc.pitt.edu. At midnight when students need computing resources to finish homework, CRCD has them covered.

Improved User Manual

One significant comment we received in the user survey was that our User Manual and Documentation was hard to use. Users pointed out that information is sometimes outdated because we have changed cluster settings. We responded to this feedback in a

number of ways. The first is a new streamlined site, which will prune away many of the items that have accumulated over the life of CRCD.

Replacing Technology via the Tick-Tock-Chime Cadence

CRCD has created a plan where each technology is replaced over five years—with two separate purchases for each technology. In MPI for example we make a new purchase in year one. That's tick. Then in year three we start a new purchase. That's tock. At the end of five years of usage, the hardware is retired from the research cluster and repurposed for the teach cluster for an additional two years of usage. This is chime.

In this way, CRC straddles the gap between old and new technology. Rather than making one big purchase in year one and running the technology for five years, when the hardware starts showing its age, we stagger the purchases so that we have new hardware in years one and three.

MPI and HTC Clusters Expansion

- 2.5X cores in new MPI vs old MPI cluster
    - 3-4x more performant cores
- Doubling the capacity of HTC

CRCD Hosted NVIDIA Workshops and Events

"5 Ways to Get Started with GPUs," Feb. 28

An introduction to acceleration of computationally intensive code using GPUs that was presented by Mike O'Keeffe, Senior Solutions Architect, NVIDIA

"Using NVIDIA GPUs with Python," March 6

Hands-on experience accelerating Python codes with NVIDIA GPUs presented by Zoe Ryan, Solutions Architect, NVIDIA

Post-NVIDIA GTC 2025 Highlights, April 4

Highlights from Jensen Huang's GTC 2025 keynote address, with takeaways presented by Patty Delafuente, Sr. Solution Architect, NVIDIA and the CRCD Team that attended GTC 2025

Grant Funding to PIs that Depend on CRCD Resources for their Research

- Supported 280 grants with a total expenditure of $106, 842,083

Research Products

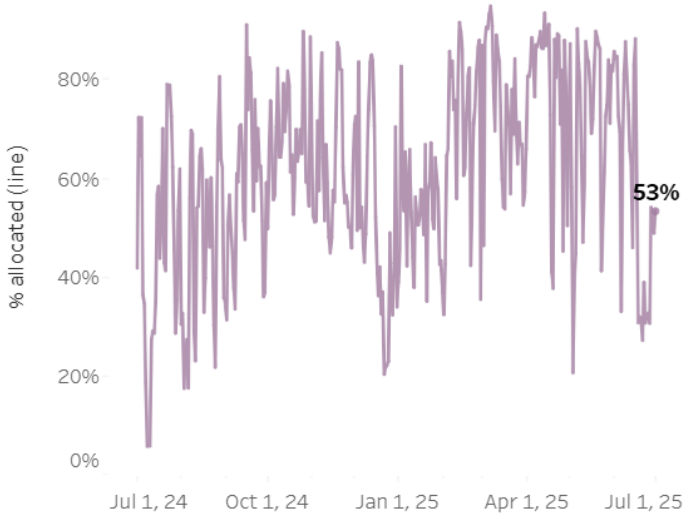
- Enabled 390 journal publications/presentations

Users/Jobs on all CRC cluster(s): July 1, 2024 to July 1, 2025

1,321	5	128	706	182	99	589	6	1
All users across **113** departments used **100956.9K** core-hours	**Applicant** users across **5** departments used **14.4K** core-hours	**Faculty** users across **56** departments used **12595.5K** core-hours	**Grad/Post-doc** users across **91** departments used **57826.0K** core-hours	**Sponsored Account** users across **47** departments used **19265.0K** core-hours	**Staff** users across **35** departments used **2971.6K** core-hours	**Undergrad** users across **89** departments used **29215.6K** core-hours	**Univ of P. Physicians Faculty** users across **4** departments used **10.3K** core-hours	**Volunteer (without ER)** users across **1** departments used **3.4K** core-hours

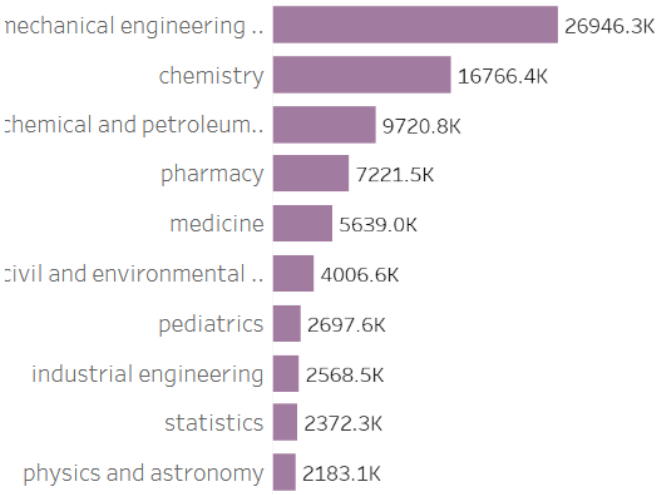
Utilization over Time: All cluster

% **core-hours** allocated by cluster
over July 01, 2024 - July 01, 2025



Top 10 departments

by **allocated core-hours** on the **all** cluster
from July 1, 2024 to July 1, 2025
(out of a total **113** departments)



Computing Cluster Average Annual Usage

- All clusters: 60.85 percent
- MPI: 55.68 percent
- SMP: 71.81 percent
- HTC: 60.93 percent
- GPU: 52.27 percent

Computing Cluster Peak Usage

- All clusters: 95 percent (03_07_25)
- gpu: 86 percent (03_18 and 06_15_25)
- htc: 94 percent (03_7_25)
- mpi: 99 percent (07_4 and 10 _11_24 and 03_05_25)
- smp: 96 percent (05_28_25)