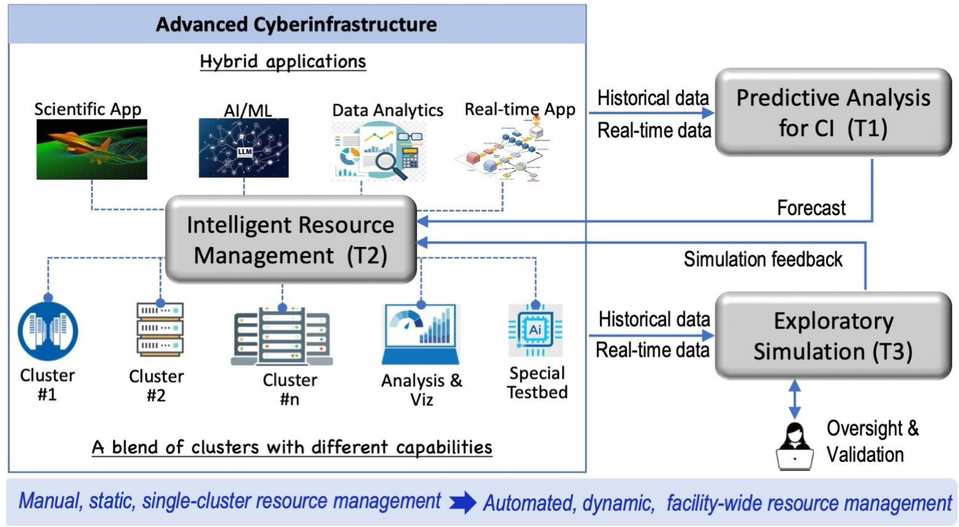
OAC Core: AIMCI: Artificial Intelligence for Managing Cyberinfrastructure

Zhiling Lan and Michael E. Papka
University of Illinois Chicago (UIC) and Argonne National Laboratory

ABSTRACT

Advanced cyberinfrastructure (CI) is undergoing disruptive changes in system architectures and application workloads. The landscape of cyberinfrastructure workloads is rapidly expanding beyond traditional computational simulations to include a hybrid mix of applications. CI facilities now host diverse high-performance systems with heterogeneous configurations, leading to a complex mix of computing, memory, and storage components. Existing CI management methods, which are heavily heuristic or manual-based, struggle with these evolving challenges. This project addresses the complex challenges of CI resource management by integrating artificial intelligence (AI) technologies with human expertise. The proposed AIMCI framework transitions from managing isolated single clusters to coordinating facility-wide management, orchestrating the entire facility as a unified pool of diverse resources for a broad spectrum of applications with various resource requirements.

HIGHLIGHT

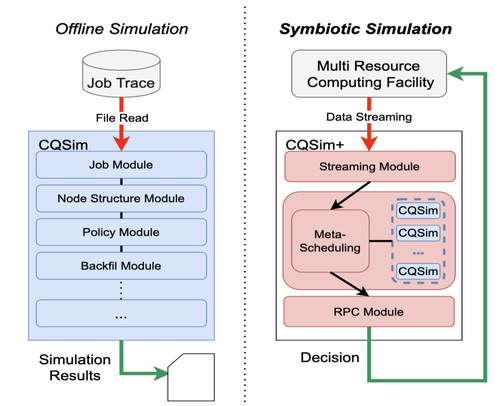


"Computers are incredibly fast, accurate, and stupid; humans are incredibly slow, inaccurate, and brilliant. Together, they are powerful beyond imagination." — Unknown

Similarly, integrating AI technologies with human expertise offers a powerful strategy for tackling the complexities of resource management in advanced cyberinfrastructure.

Results

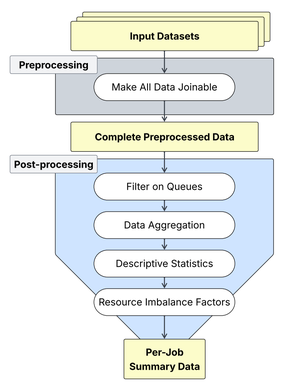
Symbiotic simulation for facility-wide, multi-resource scheduling

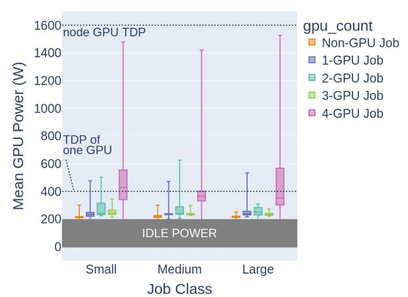


Original System	System Selected by CQSim+	Number of Jobs	Wait Time (s) (siloed)	Wait Time (s) (CQSim+)	% Impv.
Polaris	Polaris	18,439	18,707	16,898	9.7%
Theta	Polaris	4,943	10,363	5,439	47.5%
Polaris	Theta	7,827	44,061	2,107	95.2%
Theta	Theta	3,338	27,441	29,510	-7.5%

Job Wait Time Comparison: Siloed System Management vs. Facility-Wide Management Using CQSim+

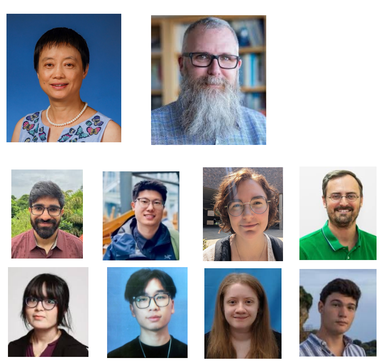
From Logs to Action: Navigating Power-Saving Opportunities





GPU Power Draw by Job Class on Polaris in 2024

MEMBERS



Team: UIC SPEAR Members
In collaboration with ALCF Operations Team

INTELLECTUAL MERIT

- Explores AI/ML to provide predictive insights and optimize decision-making in the constantly evolving CI landscape, allowing humans to focus on critical tasks of oversight and validation.
- Shifts from isolated single-cluster management to coordinated facility-wide management, where the entire facility is orchestrated as a unified pool of diverse resources for a mix of applications.

Broader Impacts

- Release the resulting AI models and software as open source to the community on GitHub.
- Strengthen diversity-focused programs at UIC (a Minority-Serving Institution) through an integrated education and outreach program.

Project website: **https://lanzhiling.github.io/projects/AIMCI/**