Hey AI, can you allocate resources for me with <cost, time> constraints?

Swathi Vallabhajosyula; Prateek Sharma; Rajiv Ramnath
vallabhajosyula.2@buckeyemail.osu.edu; prateeks@iu.edu; ramnath.6@osu.edu

INTRODUCTION:
High-performance computing (HPC) is crucial in executing resource-intensive scientific workflows such as training, fine-tuning, or inferencing deep neural networks (DNNs).

To accurately allocate resources, the user has to understand:
1. The job’s resource requirements (How much and How long?)
2. The node configurations (CPU/GPU, Memory availability)
3. The Cyberinfrastructure policies (batch limitations vs. queues vs. allocations)

To accurately allocate resources, the user should:
1. Run the job against different nodes (architectures-CPUs, GPUs)
2. Run the job with different configurations (batch size)

HARP Framework Features:
1. Profiles applications against different Systems/Architectures
2. Builds suitable regression Models
3. Applies CI’s queueing batch policies to predictions.

Implementation Challenges:
1. Generating enough profiling data for every job is expensive
2. Availability of the CI’s SLURM job queues to estimate Wait times

Current Work
1. Integrating HARP API calls with an Intelligence and Monitoring Plane to oversee job progress and enable ad-hoc decision-making.
   • Profiling each application workflow presents challenges, especially within DNN training loops.
2. Incorporating HARP with TAPIS for customized job scheduling, tailoring resource allocations to meet specific job requirements.
   • Developing a user-friendly Python module to simplify TAPIS service calls, enhancing functionalities such as estimation and scheduling.
   • The ICICLE Intelligence plane monitors job progress and performance, dynamically adjusting execution strategies.
   • Adapting code to retrieve estimated wait times during job submission.*

References