HPC-AI-agent

Mouzheng Xu

Management Information Technology and System Office

Support by LM

1. Background

High Performance Computing (HPC) platforms are inherently complex, consisting of numerous components and subsystems. They are primarily operated via the Linux command line, which presents a relatively high barrier to entry for most general users. To obtain platform status information or submit jobs, users are often required to master a large number of commands, parameters, and configuration options, which limits overall usability.

In practical use, when users want to understand information such as current resource utilization, queue workloads, or GPU/CPU availability, they must manually query multiple commands or system interfaces. This information is scattered and not easy to interpret quickly. In addition, there is limited intuitive guidance on how to select appropriate

resources for different types of workloads or which partitions or nodes are best suited for a given task. For inexperienced users, it is difficult to make optimal resource decisions within a short time, which may result in inefficient resource usage or excessively long waiting times.

2. Solutions

To address the challenges of information accessibility and high operational complexity on the HPC platform, this project introduces an AI agent to provide real-time, intelligent interaction. Through web-based access, the AI can query live data on HPC nodes, queues, and resource utilization. It then automatically analyzes the platform status and presents the results to users in a clear and readable form.

Users no longer need to understand complex Linux commands or monitoring tools. By simply describing their needs in natural language—such as "Are GPU resources currently under heavy load?" or "Which partition should I submit a CPU-intensive job to?"—the AI can, based on the latest server data, provide:

- Real-time analysis of resource utilization
- Load analysis of nodes or partitions
- Automatic recommendations for the most suitable resources based on task requirements

- Simplified operational guidance and parameter suggestions

This approach enables users to understand platform conditions more intuitively and conveniently, improving the accuracy of job submissions and overall resource utilization efficiency.

3. Outcomes and Benefits

The project has preliminarily achieved its intended objectives by implementing the core capability for the AI to read real-time server data and provide resource analysis and recommendations. Users can now obtain key HPC platform information without needing to master complex Linux commands, significantly lowering the barrier to entry and improving the platform’ s usability, accessibility, and overall user experience.

However, due to potential limitations of the underlying model or insufficient experience in AI configuration, the stability of the AI’ s responses is not yet fully consistent. In certain scenarios, misunderstandings or suboptimal recommendations may still occur, and system debugging and optimization remain challenging. Nevertheless, the current results establish a solid foundation for intelligent HPC support, and ongoing refinement and experimentation

will be required to further enhance response quality and practical effectiveness.

4. Next Steps

To further improve the stability and professionalism of the AI assistant, future work will focus on two main directions:

1. **Introducing workflow mechanisms to improve response stability**

   The AI response process will be incorporated into controlled workflows. By defining preset steps, logical branches, and data validation rules, response paths can be constrained, reducing randomness and deviation and ensuring more consistent and reliable outputs.

2. **Enhancing job analysis and script recommendation capabilities**

   Future versions will provide deeper analysis of users' job configurations and runtime behavior, including queueing reasons, resource utilization assessments, and potential optimization opportunities. At the same time, the AI will be able to automatically generate or optimize Slurm scripts based on user requirements, offering more accurate parameter and structure recommendations for job submission.