高性能计算平台助手

许谋诤

案例提供部门：管理信息技术与系统办公室

支持部门：西浦学习超市

1. 案例背景

HPC（高性能计算）平台本身具有体系复杂、组件繁多的特点，且主要通过 Linux 命令行进行操作，对大部分普通用户而言使用门槛较高。用户往往需要掌握大量命令、参数和配置方式，才能获取平台状态或提交任务，这在一定程度上限制了平台的易用性。

在实际使用过程中，用户如果想了解诸如当前资源使用情况、各队列负载、GPU/CPU 可用性等信息，需要自行查询多个命令或系统界面，信息分散且难以快速理解。同时，不同类型任务如何选择合适的资源、应提交到哪类节点也缺少便捷的指导。对于缺乏经验的用户来说，很难在短时间内根据自身需求做出最优的资源选择，容易导致资源滥用或等待时间过长。

2. 解决方案

针对用户难以获取 HPC 平台信息、操作门槛高的问题，本项目通过引入 AI agnet，为用户提供实时、智能化的交互方式。AI 可通过 web 访问的方式查询系统的 HPC 各节点、队列、资源使用情况等实时数据，由 AI 自动分析平台状态并生成可读性强的结果反馈给用户。

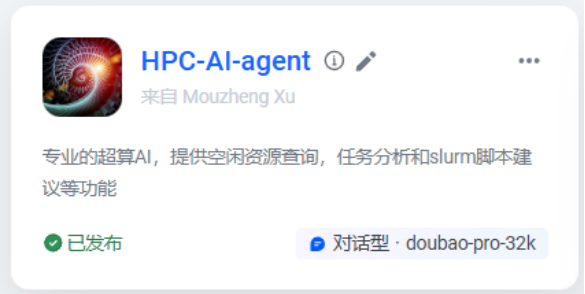

图 7-1 HPC-AI-agent

用户无需了解复杂的 Linux 命令或监控工具，只需以自然语言描述需求，例如“当前 GPU 资源是否紧张？”、“我应该把 cpu 计算任务提交到哪个分区？”，AI 即可基于最新服务器数据提供：

- 实时资源使用情况解析
- 节点或分区的负载分析
- 根据任务需求自动推荐最适合使用的资源
- 简化的操作指导与参数建议

通过这种方式，用户可以更直观、更便捷地了解平台状态，提高任务提交的准确性和资源使用效率。

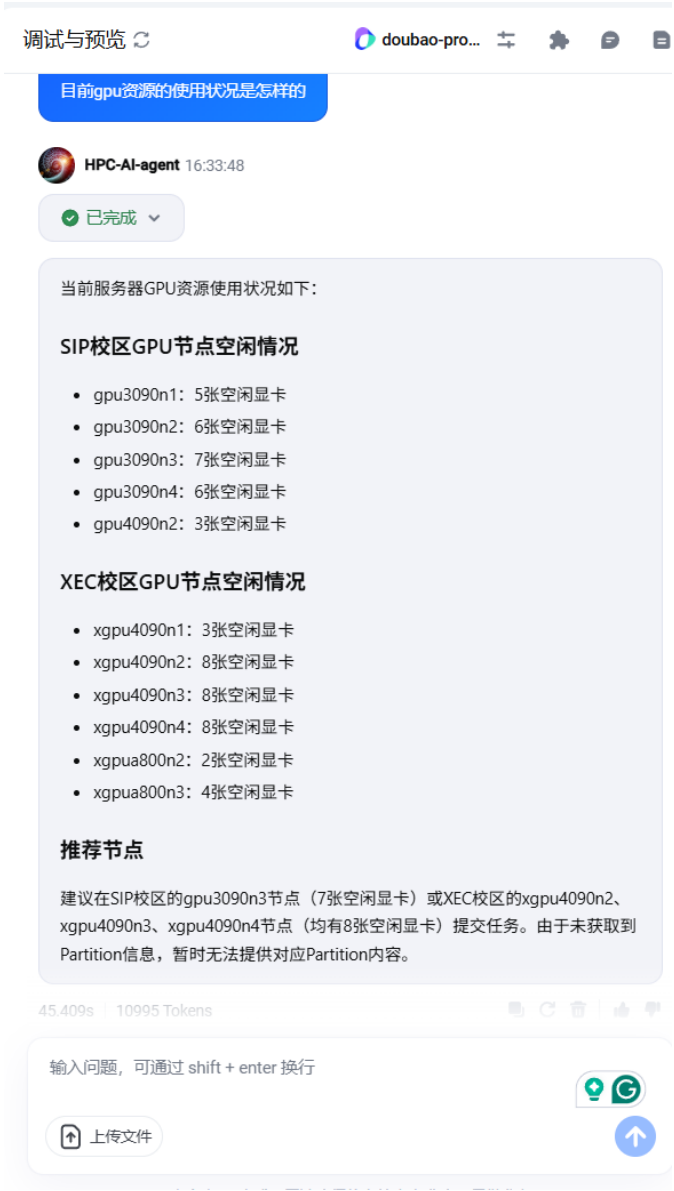

图 7-2 HPC-AI-agent 示例

3. 成果与效益

项目已初步实现预期目标，实现了 AI 读取实时服务器数据并为用户提供资源分析与建议的核心功能。用户无需掌握复杂命令即可获取 HPC 平台关键信息，显著降低了使用门槛，提升了平台的友好性与可访问性。

然而，可能因为模型本身问题，或者设置 AI 的经验不足，导致 AI 的回答稳定性尚不足，部分场景下仍存在理解偏差或建议不够准确的问题，系统调试和

优化也较为困难。目前的成果为智能化方向奠定了基础，后续仍需持续改进和尝试、以进一步提升回答质量和实用性。

4. 下一步计划

为进一步提升 AI 助手的稳定性和专业性，后续将围绕两大方向持续优化：

1）**引入工作流机制，提高回答稳定性**

将 AI 的回答过程纳入可控的工作流，通过预设步骤、逻辑分支和数据校验来约束回答路径，减少随机性与偏差，使 AI 输出更加一致、可靠。

2）**增强任务（Job）分析能力与脚本推荐功能**

新版本将针对用户的作业配置与运行情况，提供更深入的分析，包括排队原因、资源占用评估、潜在优化点等。同时，AI 将能够根据用户需求自动生成或优化 Slurm 脚本，为任务提交提供更准确的参数与结构建议。