Contribution ID: **39** Type: **not specified**

基于机器学习的 HPC 集群资源需求预测研究

Tuesday, 26 August 2025 14:00 (20 minutes)

本研究针对高能物理领域中高性能计算（HPC）集群资源调度所面临的低效利用与排队延迟问题，提出了一种基于静态代码特征与作业运行特征融合的智能化资源需求预测与调度优化框架。首先，通过对 Python 编写的高能物理作业源代码进行静态分析，提取代码行数、函数调用、循环深度、并行接口和第三方库调用等关键结构特征；同时，从 SLURM 调度日志中获取作业提交信息、资源申请与实际使用情况等作业基础特征与历史运行特征，并对多源特征进行规范化和降维处理，构建统一的“embedded Slurm job”向量表示。进而，基于历史运行特征采用层次聚类结合 K-Means 方法进行初步无监督聚类，生成类别标签，实现对作业资源使用混合特性的多标签标注。针对标注结果，构建基于 Transformer、Resnet、DNN 的 Ensemble 分类模型，通过加权交叉熵损失与注意力机制缓解类别不平衡问题，实现对作业类型的高精度识别。随后，在分类信息与“embedded Slurm job”向量的基础上，设计多目标回归模型，采用多任务学习策略并结合多分类模型隐藏层提取作业共性特征，并针对 CPU 核数、内存容量与 GPU 数量三维资源需求执行联合预测。同时，引入基于利用率均值与标准差的安全裕量计算，生成符合调度需求的资源推荐标签以防止欠配或过度浪费。最后，构建 SLURM 调度仿真环境，将传统用户申请资源与模型预测资源两种策略进行对比实验，并通过平均排队时间、集群资源利用率与作业吞吐量等指标定量评估所提方法的实际调度收益。实验结果表明，该框架能够显著提高资源利用效率、降低任务等待时长，为高能物理及其他 HPC 应用场景提供了一种可推广的智能化调度优化解决方案。

Summary

Primary authors: 何, 煜; 石, 京燕 (高能物理研究所); 杜, 然 (高能物理研究所)

Presenter: 何, 煜

Session Classification: 科学计算技术

Track Classification: 科学计算技术及用户交流