arXiv:2603.28783v1 [cs.DC] 19 Mar 2026

Nextflow工作流执行的全面基于插件的监控

Sami Kharma
kharma@zib.de 柏林
蔡司研究所 德国柏林

Tobias Wies
tobias.wies@tu-darmstadt.de
达姆施塔特工业大学 德国达姆施塔特

Florian Sc
hintke
schintke@zib.de 柏林
蔡司研究所 德国柏林

1 背景与目标

Nextflow [2] 是一种工作流管理系统，常用于生物信息学 [6] 和地球观测 [4]等领域。它将各种工具的分布式数据处理编排为无环任务序列，同时运用容器化（例如 Docker）、编排（例如 Kubernetes）或批处理（例如 SLURM）。监控此类工作流执行可能颇具挑战，但有助于性能分析、调试和数据溯源。

除了 Nexflow 内置的基础监控外，wf-commons 工具 [1] 用于创建 wf-instances，被 Nextflow社区广泛视为标准方案。我们提出的监控插件提供了一种更详细、更灵活的替代方案，与 wf-instances 兼容；它利用 Nextflow 的插件机制（21.10 版本），无需定制 Nextflow分支，还可选择直接修改静态制品的 .jar 文件而无需重新编译，并支持执行期间的在线监控。

2 监控方法

我们使用 Groovy[1] 实现了一个 Nextflow 插件，通过回调和 JVM反射访问 Nextflow 的内部状态，从而对工作流执行进行详细监控。该插件可在运行时在线构建（部分）物理执行图。对于每个任务，它都会收集详细的执行信息，例如开始和结束时间、执行方式、容器镜像以及任务的工作目录；调度器等其他工具可以利用这些信息在线处理监控数据。插件本身也会利用这些数据获取更多上下文信息，例如通过观察正在使用的共享文件系统来捕获文件元数据。可通过在 Nextflow 配置文件中启用该插件，或使用命令行覆盖来激活它。它提供有效且扩展的 wf-instance JSON格式。

由于 Nextflow 可用的内部数据或文件无法提供详细的硬件信息，也无法访问这些信息，我们提供了一个可选补丁，以实现更深入的节点级监控。具体做法是将代码或监控工具直接注入 Nextflow 部署任务的节点。这个可选补丁只需在 Nextflow 用于执行任务的包装脚本中添加钩子，即可启动外部节点级监控工具，从而静态获取 CPU、RAM、网络信息和 IP地址，或以时间序列的形式获取资源使用情况，例如使用 collectl。[2]

3 结果

借助该插件，无需修改工作流即可无缝收集真实工作流执行的数据。我们监控了 nf-core代码库中的工作流 [3]。图 1 展示了 rnaseq [5] 工作流一次包含六个节点的执行中，物理任务的分配情况。

[1]https://github.com/cookiephone/nf-bigbrother [2] https://github.com/sharkcz/collectl

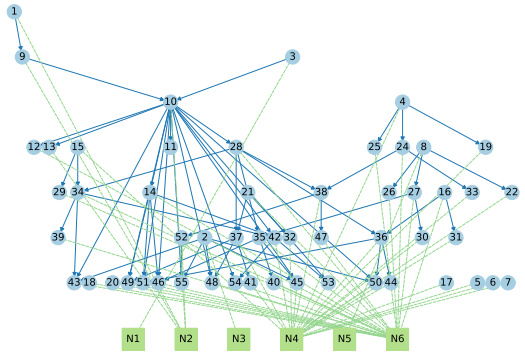

图1：通过我们的插件监控的部分 rnaseq [5] 运行的物理任务图（蓝色）与节点（绿色）分配情况。

与基本的 wf-instances 监控相比，对工作流执行进行更广泛的扩展视图，为实际的工作流分析和优化提供了新机会。

4 结论

本文介绍的插件无需修改受监控的工作流，即可全面追踪 Nextflow 执行过程。所获取的详细数据有助于深入了解工作流本身、其性能及潜在瓶颈，也可用于指导工作流的资源分配和调度，而这些做法并不适用于传统的高性能计算使用场景。

致谢

本研究获得德国研究基金会（DFG）CRC 1404 项目的资助：*FONDA*：大规模科学数据分析工作流的基础。

参考文献

[1] Coleman 等，2022 年。WfCommons：促进科学工作流研究与开发的框架。*FGCS*，128，16 20 21 27 101016 09 043 – 。doi：。/j.future.。。[2] P. Di Tommaso 等，2017 年。Nextflow 支持可复现的计算工作流。en。《自然-生物技术》，35，4，（2017 年 4 月），316–319。[3] P. A. Ewels 等，2020 年。面向社区维护的生物信息学流程的 nf-core 框架。《自然生物技术》，38，3，276–278。doi:10.1038/s41587-020-0439-x。[4] F. Lehmann 等，2021 年。FORCE on Nextflow：在通用集群上对地球观测数据进行可扩展分析。载于第 3052 卷。CEUR-WS.org。https://ceur-ws.or g/Vol-3052/short12.pdf。[5] [SW] H. Patel 等，nf-core/rnaseq：nf-core/rnaseq v3.19.0——Tungsten Turtle 版本 3.19.0，2025 年 6 月。doi:10.5281/zenodo.15631172。[6] B. Van de Sande 等，2020 年。用于单细胞基因调控网络分析的可扩展 SCENIC 工作流。en。《自然-实验指南》，15，7，（2020 年 6 月），2247–2276。

这项工作已被 SCA/HPC Asia 2026 接收为海报展示。