The goal of the Reproducibility Challenge at SC24 is to reproduce the results of a selected paper and gain hands-on experience by producing artifacts to ensure reproducible scientific experiments. The written report (limited to five pages) will be assessed along with deliverables that include plots, tables, and code artifacts. These materials will be reviewed by the Reproducibility Committee, with submissions due by Wednesday, the 20th, at 5 PM EST.

Below is a summary of the software requirements to begin the challenge:

A. Based on the paper *Data Flow Lifecycles for Optimizing Workflow Coordination* from SC23 (https://dl.acm.org/doi/abs/10.1145/3581784.3607104), set up the software environment as detailed in the appendix *Artifact Description/Artifact Evaluation*. It includes the installation of **datalife** (available at https://github.com/pnnl/datalife).
B. Install the main workflow application, the 1000 Genomes Workflow (available at https://github.com/pegasus-isi/1000genome-workflow).

For artifact submission, three specific deliverables are required in addition to the written report:

1. **Reproduction of Table (at Figure 2f)** from the paper, showing a ranked list based on an analysis of data flows captured with **datalife**.
2. **Reproduction of Figure 4a**, visualizing data volumes with the workflow pipeline.
3. **Reproduction of Figure 6**, illustrating optimized performance.

Additional instructions for optimizing performance in Section 3.:

- You are encouraged to explore storage options such as node-local storage, shared file systems, or both. Disregard the legend annotations of Figure 6, please provide your own annotations. Note that the use of 10 compute nodes, as referenced in the paper, is not expected for this challenge, use given number of compute nodes accordingly.
- Four parameters may be adjusted for performance optimization: (1) the number of concurrent tasks, (2) the size of input data, (3) assignment of tasks to nodes, and (4) data to storage.
- Invariant: Initial input (final output) files begin (end) on global storage.
- The first step of the workflow pipeline (individuals) is considered for varying the number of concurrent tasks.
- For input data (10 chromosome VCF files), you may select a subset to maximize throughput by varying the total number of input files.

Additionally, an optional workflow application will be announced at the start of the challenge, offering further opportunities for exploration.

The final report, including the .pdf and .tex files, should be submitted via Linklings (https://submissions.supercomputing.org/) under 'Stage 5: Reproducibility Challenge Consideration' in your team's submission. **Please use the AD/AE template available at https://github.com/jeanbez/sc24-repro.** You should also provide a Zenodo link with all deliverables and code/data artifacts. The deadline is Wednesday, November 20, 2024, at 5 PM EST.

Grading criteria: 70% for the report, 20% for code and data artifacts, and 10% for additional application.