A. Artifact Appendix

A.1 Abstract

This appendix provides a piece of code and associated data structures used to validate the graph analysis methodology proposed in the paper. The code creates a directed acyclic graph (DAG) consistent with the examples in the paper over networkx and implements several metric computation methods defined in the paper, including the Critical Path, Total Work (TWork), and Modified Critical Path (ModCP) metrics. By running the code, you are able to check the effectiveness of NewLB on different datasets, thus verifying the reproducibility of the paper's conclusions.

A.2 Artifact check-list (meta-information)

- **Algorithm:** Critical path calculation in DAG, TWork (total work) calculation, ModCP (modified critical path) calculation, NewLB calculation.
- **Program:** Python 3 script using networkx library.
- **Compilation:** Python3 Interpreter
- **Transformations:** N/A.
- **Binary:** N/A
- **Model:** Directed Acyclic Graph (DAG) model representing stages and tasks.
- **Dataset:** Google Cluster, Alibaba Cluster
- **Run-time environment:** Python 3.10, Google cloud instance, or Google Colab
- **Hardware:** No specific hardware requirement; any standard machine running Python 3. However, if you want to run the script of alibaba cluster, you will need a google cloud instance.
- **Run-time state:** N/A
- **Execution:** N/A
- **Metrics:** N/A
- **Output:** The computed NewLB
- **Experiments:** Running the script to produce the NewLB for the provided dataset.
- **How much disk space required (approximately):** Depends on the size of the dataset, few MB to 120 GB
- **How much time is needed to prepare workflow (approximately):** Depends on the size of the dataset, few minutes to few hours
- **How much time is needed to complete experiments (approximately):** Depends on the size of the dataset, few minutes to few hours
- **Publicly available:** N/A
- **Code licenses (if publicly available):** N/A
- **Data licenses (if publicly available):** N/A
- **Workflow automation framework used:** N/A
- **Archived (provide DOI):** N/A

A.3 Description

A.3.1 How to access

Please visit our GitHub repository at `https://github.com/JameZ233/cluster-scheduling`

A.3.2 Hardware dependencies

No special hardware dependencies.

A.3.3 Software dependencies

- Python 3.10
- `networkx`
- `matplotlib`

A.3.4 Datasets

We mainly tested NewLB on two datasets.

- Google Cluster
  `https://github.com/google/cluster-data`
- Alibaba Cluster
  `https://github.com/alibaba/clusterdata`

For google cluster, we have a script to extract a specific user's task data for evaluation, a user's data is relatively small, and easy for evaluation. For alibaba cluster, the download link is provided below. `https://github.com/alibaba/clusterdata/blob/master/cluster-trace-v2018/fetchData.sh`

A.3.5 Models

A.4 Installation

No need to install, just clone the whole repository.

A.5 Experiment workflow

Here we give the example of analysis of google cluster, first you need to extract some sample from google cluster, as the data of the whole cluster is extremely big, we provide a jupyter notebook to extract the data from google database.

A.6 Evaluation and expected results

After running google_cluster_analysis.py, you should get the value of NewLb, CPlen, TWork, and the graph after a cut of all of the DAGs. You can also get the visualization of all the graphs, to explicitly see the heiriachy of the DAG.
For the google cluster analysis, the output is the same as above, however, we looked into the alibaba cluster deeper, the scope of analysis of alibaba cluster is the whole cluster, which means the script will run through the whole cluster. This requires few hours to days of computation (due to the limited computing resource we had). Hence the results are given in the folder of results.

A.7 Experiment customization

You can run newlb_calculator.py on data of every DAG scheduling cluster for calculating the NewLB. google_cluster_analysis.py is just a example for fitting the google cluster data into newlb_calculator. All you need to do is convert your data into the format below.
If you want to run alibaba_cluster analysis, you need a google cloud instance (virtual machine). The output file is a set of txt files which contains the CPlen, TWork and NewLB of all the DAGs.

A.8 Notes

Here we introduce the functions in the newlb_calculator.py.

- create_sample_graph() will create a sample DAG that is completely the same as the graph shown in the Graphene paper.
- calc_critical_path() shows the critical path of the input graph which is the longest path length of the input graph.
- calc_twork() calculates the input graph's total work, which is the product of task number, resource and duration.
- cut_dags() is the most complex part of the process of calculate the NewLB, it will properly cut the input DAG into two sub-DAGs. The process of how to doing this is explained in the final report.
- calc_newlb() returns the NewLB of the input DAG, which can be considered as the combination of the above functions.