SMC Data Challenge 2021 : Analyzing Resource Utilization and User Behavior on Titan Supercomputer

Sajal Dash[1] dashs@ornl.gov
Arnab K. Paul[1] paula@ornl.gov
Sarp Oral[2] oralhs@ornl.gov
Feiyi Wang[1] fwang2@ornl.gov

[1]Analytics & AI Methods at Scale
[2]Technology Integration

March 26, 2021

1 Introduction

Resource utilization statistics of submitted jobs on a supercomputer can help us understand how users from various scientific domains use HPC platforms and better design a job scheduler. We explore to generate insight regarding workload distribution and usage patterns domains from job scheduler trace, GPU failure information, and project-specific information collected from Titan supercomputer. Furthermore, we want to know how the scheduler performance varies over time and how the users' scheduling behavior changes following a system failure. These observations have the potential to provide valuable insight, which is helpful to prepare for system failures. These practices will help us develop and apply novel machine learning algorithms in understanding system behavior, requirement, and better scheduling of HPC systems.

2 Datasets

Two datasets for this challenge are available at `https://doi.ccs.ornl.gov/ui/doi/334`. Each dataset directory has a ReadMe document that has information on all the features.

- `RUR`: This dataset is the job scheduler traces collected from the Titan supercomputer from 01/01/2015 to 07/31/2019 (*2015.csv - 2019.csv*). These were collected using Resource Utilization Report (RUR), a Cray-developed resource-usage data collection and reporting system. It contains the usage information of its critical resources (CPU, Memory, GPU, and I/O) of each running job on Titan during that period [2].

  `ProjectAreas`: Every job is associated with a project ID. The *ProjectAreas.csv* dataset provides a mapping of the project ID to its domain science.

- `GPU`: There have been some hardware-related issues in the GPUs in Titan that caused some GPUs to fail, sometimes irrecoverably during some job runs. This dataset provides information regarding these failures during the execution of the submitted jobs. GPUs on Titan are uniquely identified by a serial number (SN), and they are installed in a location. A GPU can be installed in a location, then removed from that location following a failure, and then re-installed in a different location after fixing the problem. If the failure can't be recovered, the GPU might be removed entirely from Titan. There are two prominent types of failures that resulted in the removal of GPUs from Titan: *Double Bit Error* (DBE) and *Out of the Bus* (OTB). The dataset (*gc_full.csv*) has the following fields:

  1. SN : Serial number of a GPU
  2. location : The location where it is installed
  3. insert : The time when it was inserted into that location
  4. remove : The time when it was removed from that location
  5. duration : Amount of time the GPU spent in this location
  6. out : If the device was taken out entirely w/o a re-installment into a new location.
  7. event : If the GPU was taken out entirely, the reason for its removal.

  To learn more about this dataset, please refer to the git repository `https://github.com/olcf/TitanGPULife` and the related publication [1].

3 Challenges

There are four challenges with varying degrees of difficulties and openness of the scope. The participants are encouraged to explore the problems in-depth and refine the challenge objectives.

3.1 Challenge 1

Perform exploratory data analysis on the `RUR` dataset to summarize data characteristics. Is there any relationship between client CPU memory usage, GPU memory usage, and the job size (number of compute nodes)?

3.2 Challenge 2

For every job, extract the project information from the *command* feature given in the `RUR` dataset. Use clustering methods to see if there are similarities in the resource usage patterns among jobs based on projects.

3.3 Challenge 3

Given a month's data, can you predict the next seven days' usage (memory, CPU hours (*stime* and *utime*), GPU hours (*gpu_secs*), etc.)? Is there any seasonal impact on such predictions? The predictive model should take various domains, user, season, system failure, etc., into consideration.

3.4 Challenge 4

(a) Can you characterize the time-lagged relationship between the `GPU` dataset and the RUR dataset? How does the change in pattern in `RUR` dataset affect any change in pattern in the `GPU` dataset?

(b) Provide a predictive analysis on how the change in `GPU` dataset has impacted the user behaviors in terms of submitted number of jobs and job sizes. Verify and validate your predictive analysis with the `RUR` dataset. You can consider `GPU` data and the `RUR` data from early 2015 to build and train your predictive models and verify and validate your predictive models with the `GPU` and `RUR` data from 2015-2017.

References

[1] George Ostrouchov, Don Maxwell, Rizwan A Ashraf, Christian Engelmann, Mallikarjun Shankar, and James H Rogers. Gpu lifetimes on titan supercomputer: Survival analysis and reliability. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14. IEEE, 2020.

[2] Feiyi Wang, Sarp Oral, Satyabrata Sen, and Neena Imam. Learning from five-year resource-utilization data of titan system. In *2019 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 1–6. IEEE, 2019.