# Research Proposal

CS5352 Course Project, Spring 2021

**Student:** Huyen Nguyen
**Mentor:** Jie Li

**Research Topic**

No. 4: Clustering Job Accounting Data

**Problem statement**

*High Performance Computing (HPC)* serves researchers and domain experts on numerous computing tasks regarding science, engineering, security, and commerce. To improve productivity and maximize the computing power of supercomputers, researchers are expected to have system-specific knowledge to develop codes, and job submission scripts, starting with system information such as software libraries supported by the HPC system or the number of computing cores should be for a specific job on the system [1]. It is crucial to understand the current HPC workloads and their evolution to assist informed future scheduling research and enable efficient scheduling in future HPC systems [2]. To investigate the HPC platform performance and whether HPC users used the resources productively, we conduct this usage behavior analysis.

**Research and development activities**

- Examine the data of two months of job accounting data (457,486 records) provided
- Understand the metric in the job accounting information and extract features from it. An accounting record is written to the accounting file for each job having finished.
- Select and apply appropriate clustering algorithms on the job accounting data. Two kinds of clustering algorithms, partitioning clustering, and hierarchical clustering will be analyzed, then we can determine the appropriate number of behavior categories, such as data-intensive jobs and computational-intensive jobs. A Dendrogram can be used in accordance with the algorithm to identify the categories.
- Discuss findings from the cluster analysis. By exploring the usage behavior, we can gain insights into the system and then work out the strategy to use the resource productively.

**References**

[1] Zhang, Hao, Haihang You, Bilel Hadri, and Mark Fahey. "HPC usage behavior analysis and performance estimation with machine learning techniques." In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[2] Rodrigo, Gonzalo P., P-O. Östberg, Erik Elmroth, Katie Antypas, Richard Gerber, and Lavanya Ramakrishnan. "Towards understanding HPC users and systems: a NERSC case study." *Journal of Parallel and Distributed Computing* 111 (2018): 206-221.