Comparative Analysis of Historical Job Data Across High Performance Computing Systems

Bachelor Thesis

University of Basel
Faculty of Science
Department of Mathematics and Computer Science
High Performance Computing Group

Advisor and Examiner:
Supervisor:

Author:
Email:

August 4, 2025

Abstract

This document serves as a writing template for all writings (scientific papers, technical reports, master theses, bachelor theses, master project reports, etc.) in the DMI-HPC group.

Contents

Chapter 1

Introduction

1. What are you trying to do? Articulate your objectives using absolutely no jargon.
2. How is it done today, and what are the limits of current practice?
3. What is new in your approach and why do you think it will be successful?
4. Who cares? If you are successful, what difference will it make?
5. What are the risks?
6. How much will it cost?
7. How long will it take?
8. What are the mid-term and final “exams” to check for success?

1.1 Operational Data Analytics

High-Performance Computing (HPC) systems form an important pillar in our modern society by enabling fields like earthquake prediction, oil exploration, and climate analysis. In recent years both the application fields and demand for processing power are growing explosively [4], especially with the emergence of deep learning AI. This results in the emergence of the first exascale systems that face significant operational challenges due to unprecedented complexity. To fulfill this operational demand, improving the speed, efficiency, and reliability of these systems is vital. A term named named Operational Data Analytics (ODA) was coined by Bourassa et al.[1] in the context of optimizing cooling systems at the National Energy Research Scientific Computing Center (NERSC) and expanded on by Netti et al.[2] into a general coneptual Framework with the goal of ”continuous monitoring, archiving, and analysis of near real-time performance data, providing immediately actionable information for multiple operational uses.”[3] In this bachelor thesis we aim to produce descriptive and prescriptive ODA on the application layer across different HPC system datasets by analyzing their submitted job data.

Similar approaches and what we will do differently -¿ analyze multiple datasets

1.2 Motivation

Look for patterns in job data that are useful in enhancing performance. Maybe more results from analysis. Comparative analysis for insights. Batch analysis of job with simulator.

1.3 Challenges

HPC Job Data Datasets are rare and big in size. Existing analysis explore single datasets or just subsets.

1.4 Goals

Find patterns and anomalies across three different HPC systems to gain novel insights. Along with simulator

1.5 Solution and Answer

Batch analyze job data using python and a hpc node and simulator. Look for interesting insights that might enhance HPC Systems.

Chapter 2

Related Work

What are the current practices and their limitations? What is new and what is different in our work? Look at general HPC System analysis Look at similiar job analysis approaches. What was their methodology.

Chapter 3

Methods

3.1 What methods are we using to achieve our goal?

Data analysis with python, parquet file format, processing with HPC. How to set up simulation.

Chapter 4

Analysis

4.1 Marconi M100 Analysis

4.2 Fugaku Analysis

4.3 NREL Eagle Analysis

4.4 Similarities

4.5 Anomalies

Chapter 5

Simulation

https://gitlab.inria.fr/batsim/pybatsim

5.1 Available Simulators

5.2 Open Dc Setup

5.3 Results

Chapter 6

Discussion

6.1 What do the results mean and why does it matter?

6.2 How do we compare to other related work?

6.3 Are there problems or limitations remaining?

6.4 Limitations of Open DC

Chapter 7

Conclusion

7.1 What are the main contributions of our work?

7.2 Is there room for future research?

7.3 Cross Dataset Analysis

7.4 Simulate HPC

Bibliography

[1] Norman Bourassa, Walker Johnson, Jeff Broughton, Deirdre McShane Carter, Sadie Joy, Raphael Vitti, and Peter Seto. Operational data analytics: Optimizing the national energy research scientific computing center cooling systems. In *Workshop Proceedings of the 48th International Conference on Parallel Processing*, pages 1–7, 2019.

[2] Alessio Netti, Woong Shin, Michael Ott, Torsten Wilde, and Natalie Bates. A conceptual framework for hpc operational data analytics. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 596–603. IEEE, 2021.

[3] Michael Ott, Woong Shin, Norman Bourassa, Torsten Wilde, Stefan Ceballos, Melissa Romanus, and Natalie Bates. Global experiences with hpc operational data measurement, collection and analysis. In *2020 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 499–508. IEEE, 2020.

[4] Jia Wei, Mo Chen, Longxiang Wang, Pei Ren, Yujia Lei, Yuqi Qu, Qiyu Jiang, Xiaoshe Dong, Weiguo Wu, Qiang Wang, et al. Status, challenges and trends of data-intensive supercomputing. *CCF Transactions on High Performance Computing*, 4(2):211–230, 2022.

Declaration of Scientific Integrity