

CSC4160 Final Project Proposal

Title: Analyzing public cloud traces using machine learning to improve resource allocation

Team member:

Chua Qin Di, 125400010, qindichua@link.cuhk.edu.cn

Topic Description & Aim:

This project aims to analyze how cloud computing resources are allocated and utilized in large-scale cloud data centers. Public cloud traces such as Google Cluster Workload datasets will be used to study real-world usage.

The goal is to analyze patterns in resource requests and actual usage, identify inefficiencies, and apply machine learning techniques to predict cloud resource utilization. The findings will be used to propose improvements in scheduling efficiency and resource management in large-scale cloud data centers.

Background:

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of computing resources such as servers, storage and processing power. This allows users to access scalable infrastructure without owning physical hardware.

Large cloud platforms like AWS and Google Cloud process millions of computing tasks daily on thousands of servers. Each job must specify the amount of CPU and memory required before execution, but these resource estimates are often inaccurate. These inaccurate estimates lead to inefficient resource utilization, such as over-allocation which negatively impact performance and increases operational costs.

To address this, major providers such as AWS have introduced Predictive Scaling features that use machine learning to forecast future workload demand, highlighting the ongoing challenge of inefficient resource allocation.

This project seeks to address this issue by analyzing public cloud workload traces and applying machine learning techniques to better predict forecast cloud resource utilization, and propose strategies to improve efficiency in large-scale data centers.