geocost: Cost-latency Optimized Multi-zone Cloud Storage Solution

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Haotian Liu
Spring, 2020

Technical Project Team Members
Haotian Liu (student researcher)

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Haotian Liu

Approved _____ Date _____
Haiying Shen, Department of Computer Science

Today's large-scale and multi-region internet applications often use CDNs (Content Delivery Network) to store and serve files and data in multiple data centers across the globe from one or multiple cloud storage service provider. Often, files from the closest data center to the clients are served, thereby minimizing access latency by reducing the physical distance between client and data. However, while minimizing client-side transaction latency, this strategy is often not most cost-optimized, as the closest data center to clients do not always offer the best per-volume data downloading / uploading rates. For example, the S3 cloud storage service offered by Amazon Web Services (AWS) charges nearly twice as much ($0.0405 per GB storage) in its South America data center as it does in its Northern Virginia data center ($0.023 per GB storage). In my research project, we propose geocost, a reinforcement learning-based solution that finds the perfect balance between cloud storage data transaction latency and cost.

To understand the complications of serving data from a cloud storage providers and how data transfer latency would vary on a time-dependent basis, we first set out to benchmark the AWS S3 cloud storage service by recording time used for downloading data of a range of sizes from 7 major AWS data centers repeatedly over a long period of time. The benchmarking results are then used as time-series training data for an array of model / algorithms including common regression models, LSTM and ARIMA (Auto Regressive Integrated Moving Average) in order to obtain a time-series model to forecast future latencies given a specific data center location / data size and make intelligent decisions on data center to upload to / download from. So far, our model is able to reach around 85% validation accuracy over historical data, while this score is expected to improve as size of our training data continues to grow.

Following the benchmarking of S3 services, we also built the geocost module. To simulate real-world usage of CDNs and multiple data centers, we opted for an API-oriented

design of geocost, which functions as a server layer between client side and S3 and exposes two RESTful endpoints that allows for downloading and uploading files to S3 buckets. We intend future users to use geocost as a worry-free “magic box” that intelligently choses the most cost and latency optimized data center to upload / download data, thereby removing such concern from users. Combined with the most optimized cost-latency prediction of the time-series forecast model, geocost is capable of minimizing both the cost and latency under a set of preset hyperparameters such as aggregated threshold cost estimate and maximum latency expectation.