January 2024 Research Grant Applicants

Guidance and Resources for considering cloud computing costs in your Research Grant proposed budget

Cloud computing costs can be unpredictable, so developing a budget to support use of cloud resources can be challenging. The GREGoR Data Coordinating Center (DCC) has compiled the following resources and information to support GREGoR Research Grant applicants in developing their proposal budgets.

For context, the GREGoR Consortium Dataset is available on NHGRI's Analysis, Visualization, and Informatics Lab-space (AnVIL). AnVIL is currently developing multi-cloud capabilities across the Microsoft Azure and the Google Cloud Platforms; however, as of December 2023, the Google Cloud Platform integration is more mature.

We recommend looking at AnVIL's page on Preparing a Cloud Cost Budget to develop your GREGoR Research Grant Proposal Budget. This page provides links to pricing tables that can be used to estimate costs for storage, egress, and computing, as well as example budget justification text that you can adapt for your project. The Google Cloud Pricing Calculator is also available, which enables more precise calculation based on more variables.

Working with GREGoR Data

A researcher can work with GREGoR Consortium data in a variety of ways. If you choose to work on the AnVIL platform, analysis would generally begin by "cloning a workspace", which provides pointers to the data files included in the dataset. With this approach, you would not need to pay for storage of the GREGoR Dataset, but you would be responsible for costs associated with the computing resources used for your analysis. On the other hand, if you choose to download GREGoR data to your local computing resources, you would need to pay for the egress charges associated with downloading the data.

The first GREGoR dataset release is approximately 79.2 TB, and the data set continues to grow. Based on Google Cloud rates as of December 2023, egress charges for downloading all 79.2 TB of data within the USA would be roughly $6,500.

Cloud Computing Costs

Cloud costs generally fall into two categories: storage and compute. Storage costs can be estimated as scaling linearly with the volume of stored data. GREGoR Research Grantees would be responsible for paying for the cost of any persistent disks or intermediate files they may generate on the cloud that would not be part of the GREGoR Consortium Dataset, but they would not need to pay for the storage of the GREGoR Dataset itself.

In our experience, the compute portion of Cloud Costs associated with working on AnVIL generally fall into three categories that have increasing complexity with respect to cost estimation: interactive analysis (such as using a Jupyter notebook or R Studio Session), workflow execution, and other Google Cloud Platform applications.

Interactive analysis on AnVIL uses a defined compute instance, with a fixed hourly cost. This cost varies based on factors such as the amount of memory or number of processors on the virtual machine the user selects for their work. GCP Virtual machines can be suspended when not in use to lower costs. For example, the default RStudio Cloud Environment on AnVIL (4 CPUs; 15GB RAM) costs $0.20/hr when active and the default Jupyter Cloud Environment on AnVIL (1 CPU; 3.75 GB RAM) costs $0.06/hr when active. Both environments cost $0.01/hr when paused, and both environments can be customized. Interactive analyses also incur a storage charge for the persistent disk that can be attached to the virtual machine; the default is $2.00/month for 50GB.

Workflow execution generally scales in predictable ways, but is affected by the details of the computational task, the extent of parallelization, and the volume of data processed. Workflow execution costs can vary widely based on the duration of execution time and complexity of the workflow. These costs are calculated as a function of execution time, with different rates for different instance types. For example, when the GREGoR DCC reprocesses submitted .bam files in preparation of joint calling, we spend approximately $16/sample to run the reprocessing workflow. Thus, we generally recommend a conservative approach to piloting workflow execution to estimate total costs - start with a small test run to establish baseline cost estimates, grow gradually, and monitor costs carefully when executing a large-scale task. The best way to budget for workflow execution costs is to pilot the workflow on AnVIL.

Finally, it is possible to use other Google Cloud resources (such as BigQuery) from AnVIL workspaces. Estimating costs for this kind of application depends on the specific details of the application and services used, and is beyond the scope of this document.

Additional Resources for Developing Budgets

There are a variety of NIH resources to help researchers estimate and reduce the costs associated with cloud computing. NIH-funded researchers are encouraged to engage with the NIH STRIDES initiative, which has negotiated with the major cloud providers to provide cost discounts and training resources. Many universities are currently participating in the STRIDES initiative. You can check whether your institution is participating at https://cloud.nih.gov/about-strides/participants/. If there is an established relationship with STRIDES, your local contact point may be able to provide additional information in compiling your Cloud Computing budget. GREGoR Research Grant applicants are responsible for verifying whether Research Grant award funds would be eligible for the STRIDES discounts.

An additional set of resources has been assembled by the NIH Cloud Computing Interoperability (NCPI) Program at https://www.ncpi-acc.org/resources/cloud-cost-estimation. The links assembled by the NCPI Program include documentation and advice from the AnVIL program about estimating cloud costs and writing a budget justification.