Contribution ID: **20** Type: **Oral Presentation**

Performance and Cost Evaluation of Public Cloud Cold Storage Services for Astronomy Data Archive and Analysis

Friday, 28 August 2020 09:30 (30 minutes)

Currently major cloud providers provide cold storage services as a part of their public IaaS offerings, targeting users who need to store data with relatively low access frequency for long periods. The adoption of cold storage services should be considered in order to reduce the total cost of ownership and the labor of storage management of maintaining large amounts of scientific research data over a long period of time. However, performance and cost of public cold storage services in scientific applications have not been well studied, and the following issues arise:

 It is difficult to determine whether cold storage services meet the performance requirements of research applications. It is also difficult to assess the feasibility of storing and accessing research data in cold storage services in terms of cost.

In order to address the issues mentioned above and to validate feasibility of adopting cold storage services in the astronomical research area, we present evaluation of cloud cold storage services using astronomical research data and applications. We stored the observation and analysis data of the ALMA radio telescope project[1] in S3 Infrequent Access and Glacier provided by Amazon Web Services (AWS) and we ported the data archive software used in the ALMA project, Next Generation Archive System (NGAS), to AWS.

To solve the first issue, we measured the performance of data retrieval operations of NGAS on AWS. In addition, we conducted performance benchmark tests such as uploading up to 60TB ALMA data. We also conducted the same benchmark tests on other commercially available cold storage services such as Google, Azure, and Oracle to validate the performance requirements can be generally fulfilled.

To solve the second issue, we proposed a cost estimation model of NGAS based on the AWS payment required for storing and retrieving data and estimated the yearly expense of the NGAS on AWS by using the actual values of data amounts and the accesses frequency statistics. Our estimation shows that retrieving data from a cold storage service and analyzing the data outside of the cloud (e.g. an on-premise system) increase the cost because data transfer cost outward the cloud is significantly high. We designed the architecture to analyze the retrieved data inside cloud and estimated cost for running common analysis applications, Common Astronomy Software Applications package (CASA)[2], with NGAS on a variety of instances of AWS.

From those experiments, the following findings are obtained:

 We can obtain practically acceptable performance of data access in the cold storage services by configuring the archive system with appropriate sizing of instances and tuning the system. Although some cold storage services require hours to start data access, this disadvantage can be mitigated by adopting an appropriate tiered storage architecture. The proposed cost estimation model enables to estimate total cost of data archive in the cloud cold storage services and data analysis on the cloud services. The model is also capable to estimate cost on a hybrid system organized by clouds and on-premise systems. Additionally, the practical information which can be used to determine the optimal configuration of the analysis system such as sizing information of AWS instances are acquired.

[1] https://www.nao.ac.jp/en/research/project/alma.html
[2] https://casa.nrao.edu/

Primary author: Mr YOSHIDA, Hiroshi (National Institute of Informatics)

Co-authors: Dr MORITA, Eisuke (National Astronomical Observatory of Japan); Prof. KOSUGI, George (National Astronomical Observatory of Japan); Prof. AIDA, Kento (National Institute of Informatics); Dr MIEL, Renaud (National Astronomical Observatory of Japan); Dr NAKAZATO, Takeshi (National Astronomical Observatory of Japan); Dr HAYASHI, Yohei (National Astronomical Observatory of Japan)

Presenter: Mr YOSHIDA, Hiroshi (National Institute of Informatics)

Session Classification: Infrastructure Clouds and Virtualisation Session

Track Classification: Infrastructure Clouds and Virtualisation