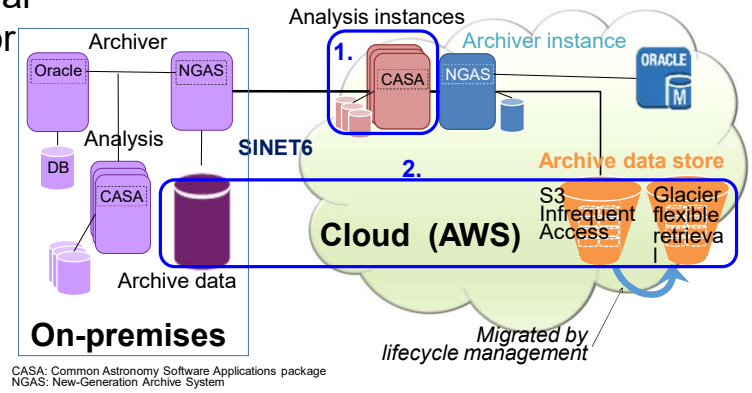
Case Studies of Hybrid Cloud Architectures for Astronomical Observatory Data

Adopting public cloud services for scientific research will reduce the total cost of ownership, allow the use of state-of-the-art technologies (e.g., the latest GPUs), and establish BCP. However, no methodology exists for designing an optimal architecture to realize these advantages. We have been conducting case studies of storing and analyzing ALMA radio telescope data in public cloud services in collaboration with the National Astronomical Observatory of Japan (NAOJ) to demonstrate the best practices and discuss the design of a suitable architecture.

Source: NAOJ

Case Study Overview

NAOJ is considering a hybrid cloud architecture comprising its on-premises system and additional public cloud services. To establish the criteria for optimal data allocation between the cloud and on-premises system, we ported the ALMA data and analysis/archiver software to public cloud services to investigate the cost-effective usage of computing resources and storage.



CASA: Common Astronomy Software Applications package
NGAS: New-Generation Archive System

Examples of Case Study Results

1. Selection of optimal server instances for analyses of ALMA data

■For 7m antenna observations*, we developed a methodology to choose the cloud server instances based on the resource usage estimation (number of cores and memory capacity) according to the observation parameters.

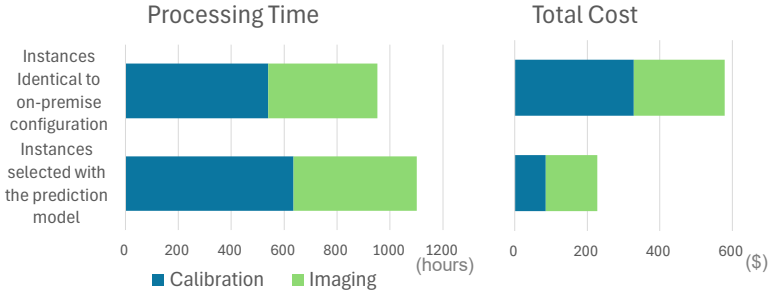
i. First, resource usage was measured in the instances with abundant resources to acquire training data.

ii. Focused on resource-intensive execution stages of whole analysis processes, we created machine learning models based on linear regression and random forest regression. The models estimate necessary resource quantities according to parameters such as data size, total number of channels, etc. We found that 40 samples of resource usage measurement results are enough for training.

iii. We adjusted the estimated resource quantities to compensate for the underestimated cases, rounded up the quantities to the cloud instance specification values (e.g., memory capacity = 4GB/8GB/16GB), and we selected fitting instances.

■The result of 372 analyses on the instances chosen with this methodology showed that the processing time increased by 15%. However, the total cost decreased by 60% compared to the instances identical to on-premise servers (a simple “lift and shift” case).

*ALMA telescope consists of 7m and 12m antennas. As a first step, we tried to establish a model for the observations with 7m ones because the analysis execution times are relatively short, performance is determined mainly by CPU and memory, and the effect of disk I/O is small.



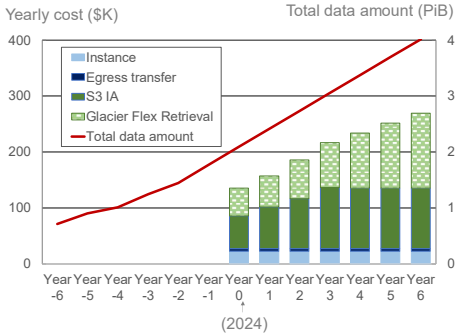
2. Cost estimation of storing ALMA archive data in the tiered storage of the hybrid cloud

■ The estimation result shows the following advantages:

I. Investment in on-premises storage can be kept constant.

II. Using two cloud storage tiers inhibits the cloud cost from increasing in proportion to the total amount of data.

III. The influence of the long restore time for Glacier (200 minutes) can be mitigated (limited to 10% of downloads).



[Assumption based on the status as of 2024/2Q]
- Amount of on-premise storage: **960 TB**
- Amount of archive data: **2092 TB in year 0; increases 320 TB/year.**
- Dataset location: **On-premise for 3 years; S3 IA for 2 years; then migrated to Glacier**
- Amount of downloaded data: **550 TB/year; 12% from S3 IA and 6% from Glacier**

Acknowledgment: We would like to thank the PoC members at NAOJ for providing data and support.

National Institute of Informatics
https://ccrd.nii.ac.jp/sc24/