Astronomy Data Center
国立天文台 天文データセンター

国立天文台 NAOJ National Astronomical Observatory of Japan

Strategies to support the NAOJ future projects: Astronomy Data Center (ADC)

George KOSUGI
Director
Astronomy Data Center
National Astronomical Observatory of Japan

November 7, 2023
NAOJ Future Planning Symposium 2023

Contents

- Mission of ADC
- Strategy 1: Organization
- Strategy 2: Consolidation of IT infrastructure and Data Services
- Strategy 3: Development and Implementation of a mechanism for handling Huge Data
- Strategy 4: Introduction and Verification of new technology
- Summary

Astronomy Data Center
国立天文台 天文データセンター

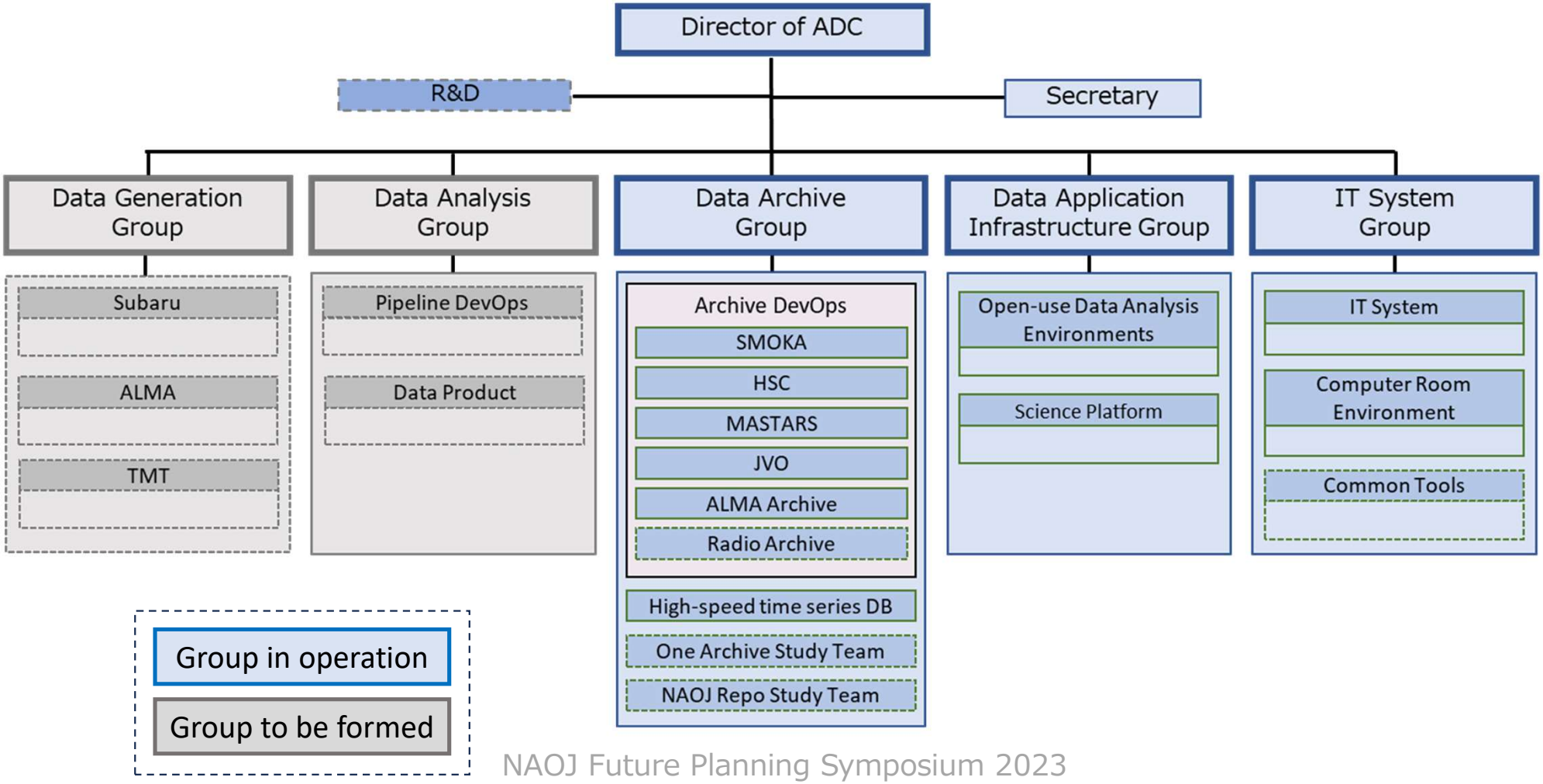
Missions of Astronomy Data Center (ADC)

国立天文台 NAOJ National Astronomical Observatory of Japan

- Develops and operates computer infrastructure and software to **generate, archive, analyze, and release astronomical data** in collaboration with NAOJ projects
- **Promotes data science across wavelengths** by providing the astronomy community with research infrastructure and educational opportunities to utilize astronomical data

Future projects will have to deal with huge observation data which ADC should support: "era of huge data".

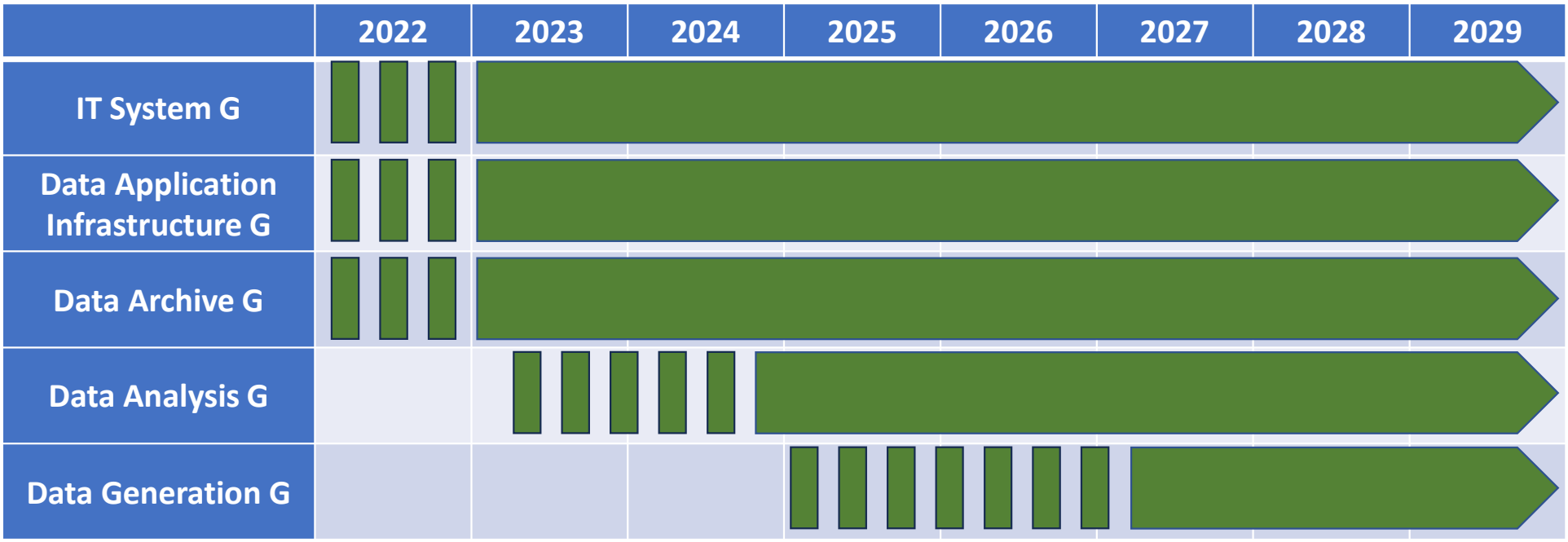
Strategy 1 (1/4): Organization



Strategy 1 (2/4): Role of each Group

- Data Archive Group
  - Data Archive operations and Services in cooperation with projects
- Data Application Infrastructure Group
  - Operation of Open-use data analysis systems with helpdesk and training sessions
  - R&D (Science Platform)
- IT System Group
  - Provides IT infrastructure to ADC groups and observatories (joint operation)
  - Researches and implements new technologies
- Data Analysis Group (to be formed)
  - Develops and operates data reduction tools and pipelines in cooperation with projects
  - Timeline: Starting with Subaru this fiscal year, and gradually expand over the next several years
- Data Generation Group (to be formed in the future)
  - Develops instrument, telescope, and observatory control system for data generation
  - Timeline: No exact plans yet (a few years to start?). Depends on how the demand and need to be.

Strategy 1 (3/4): Timeline



Strategy 1 (4/4): for Future Projects

- NAOJ-wide Groups to collect expertise and human resource

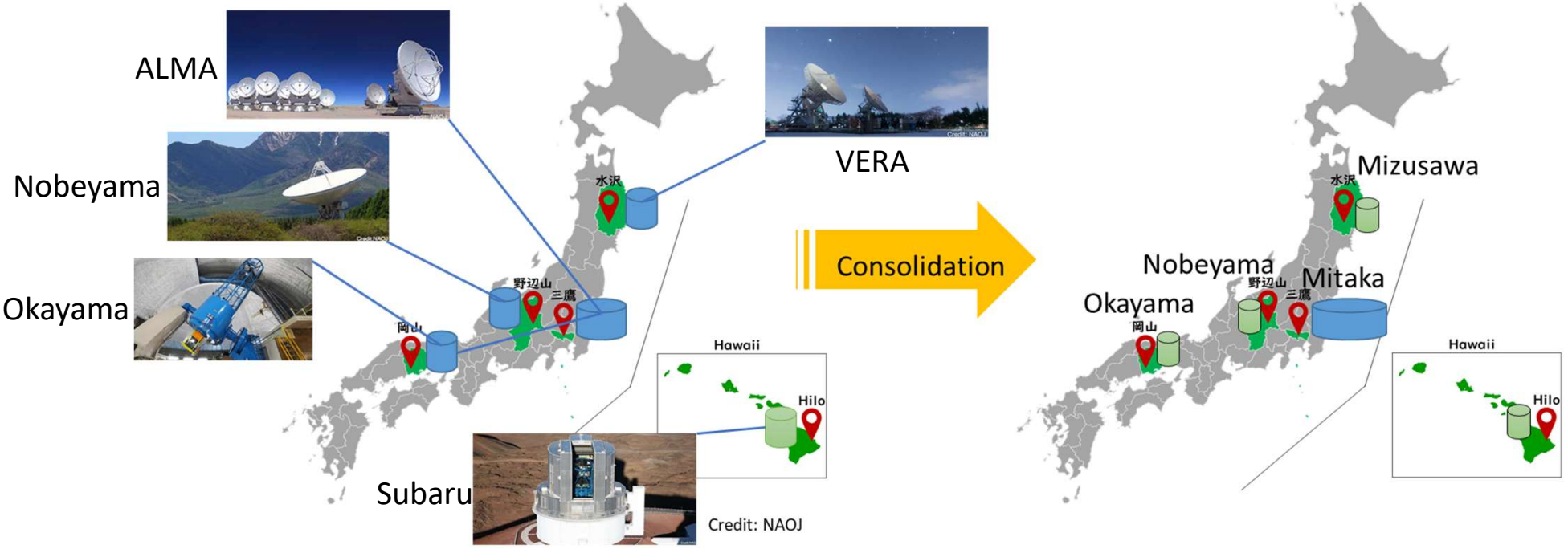
- Experienced & skilled staffs can contribute to NAOJ future projects effectively

Human resource sharing or collaboration status:

- IT System Group
  - OAO, VERA, RISE, Solar, CfCA, KAGRA
- Other Groups or general computing
  - Subaru, ALMA, TMT, NRO(+ASTE)

Strategy 2 (1/3): Consolidation of IT Infrastructure

- Consolidate archive storages in Mitaka and shared with projects
  - At the timing of computer system replacement (by July 2024)
  - Temporary storage (staging area) is located at each site
  - Archive related human resources are also being shared with projects



Strategy 2 (2/3): Consolidation of Data Archives

- ADC is planning to strengthen cooperation among multiple archives operated by ADC and projects
- ADC has future prospect to consolidate multiple archives into One Archive for more efficient use of operation & maintenance manpower and computing resources
- One stop archive service has many advantages also for researchers

Timeline:

- NAOJ Optical/Infrared Archive may be operational around 2027-28
- NAOJ One Archive in 2030

Strategy 2 (3/3): for Future Projects

- Strengthen the coordination between multiple archives and consolidate them (One Archive) in the future
- IT infrastructure (storage system, cluster computers) will be shared by multiple projects

- Future projects can utilize existing IT infrastructure with minimum initial investment (money & human)
- Data taken with the future projects can easily be combined with other multi-wavelength science data stored in NAOJ One Archive

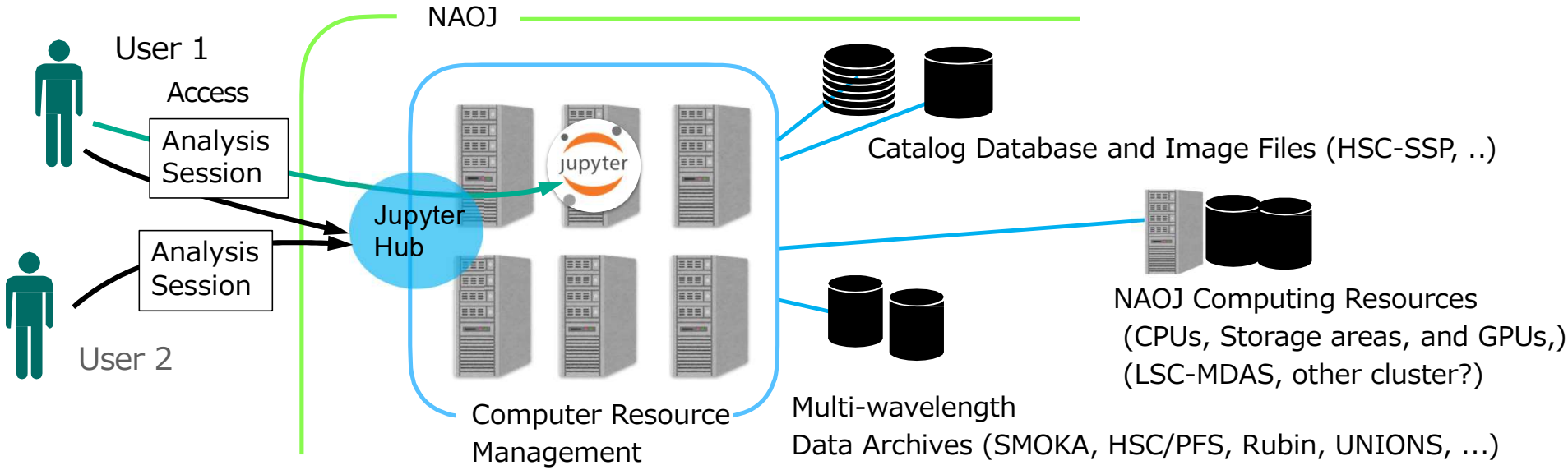
Future Projects contacted for resource sharing: JASMINE, EA-SKA

Strategy 3 (1/4): Mechanism for handling Huge Data

- ADC is developing and implementing a mechanism for researchers to utilize observation data to pursue their research in the "era of Huge Data"
- One of these mechanism is called "Science Platform"
- Many new projects which generates massive observation data are planning to have Science Platform: Vera/Rubin/LSST, SKA, etc.

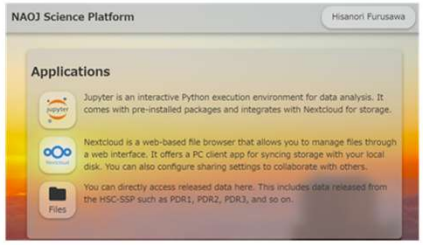
Strategy 3 (2/4): Science Platform (SP)

- As a data utilization service for ADC data archives (raw & science-ready data)
- A prototype system based on Kubernetes + JupyterHub to utilize HSC-SSP data is under way
  - to perform efficient analysis over a mass of products from a distance
  - to make the most use of available computing resources in ADC
- The system will connect to future science products (incl. PFS, Rubin) and public raw data as well

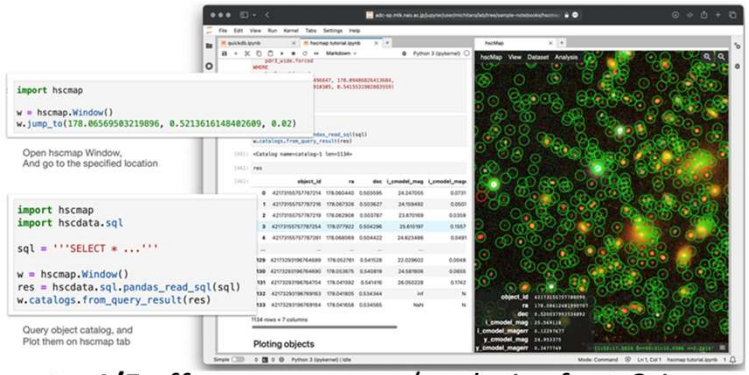


Strategy 3 (3/4): Science Platform (SP)

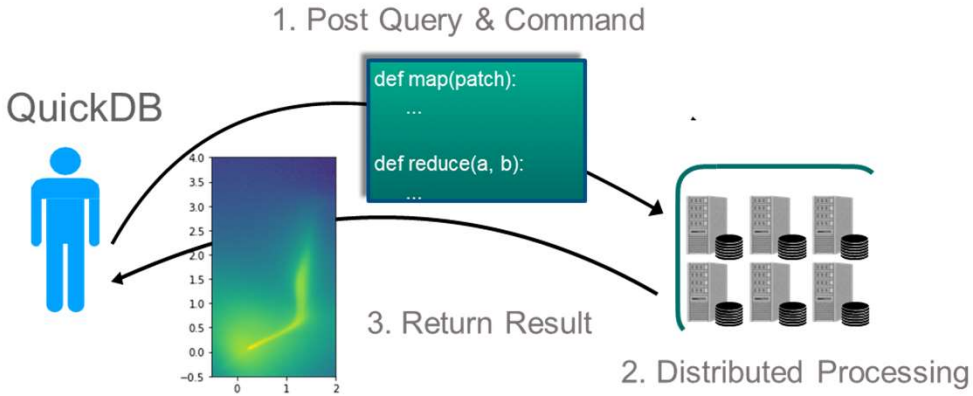
- Snapshots from the On-going HSC-SP Development



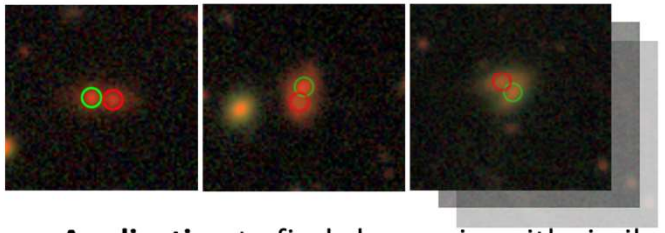
HSC-SP provides 1) computing resources in ADC,
2) Jupyter-notebook I/F for data query & processing,
3) Efficient file sharing mechanisms: Inter-operation w/ various archives (PFS, Rubin, SMOKA...) in the plan



Jupyter I/F offers easy access/analysis of cat & image with Python APIs and interactive HIPS viewer hscMap.



QuickDB a columnar federated database, capable of fast search (<2.5sec for 800M rows) and MapReduce-driven complex query



A Science Application to find close pairs with similar colors by a QuickDB query, obtaining 87k pairs in 5sec for 500M rows. Optimal tools for various science cases to be developed.

Strategy 3 (4/4): for Future Projects

- Open-use Multi-wavelength Data Analysis System (MDAS) and Large Scale data analysis system (LSC) will evolve to or will be integrated into the Science Platform in the “era of Huge Data”

- ADC will support Future projects to implement their own Science Platform
- ADC Science Platform can be connected to the Science Platform for the Future projects

Timeline:

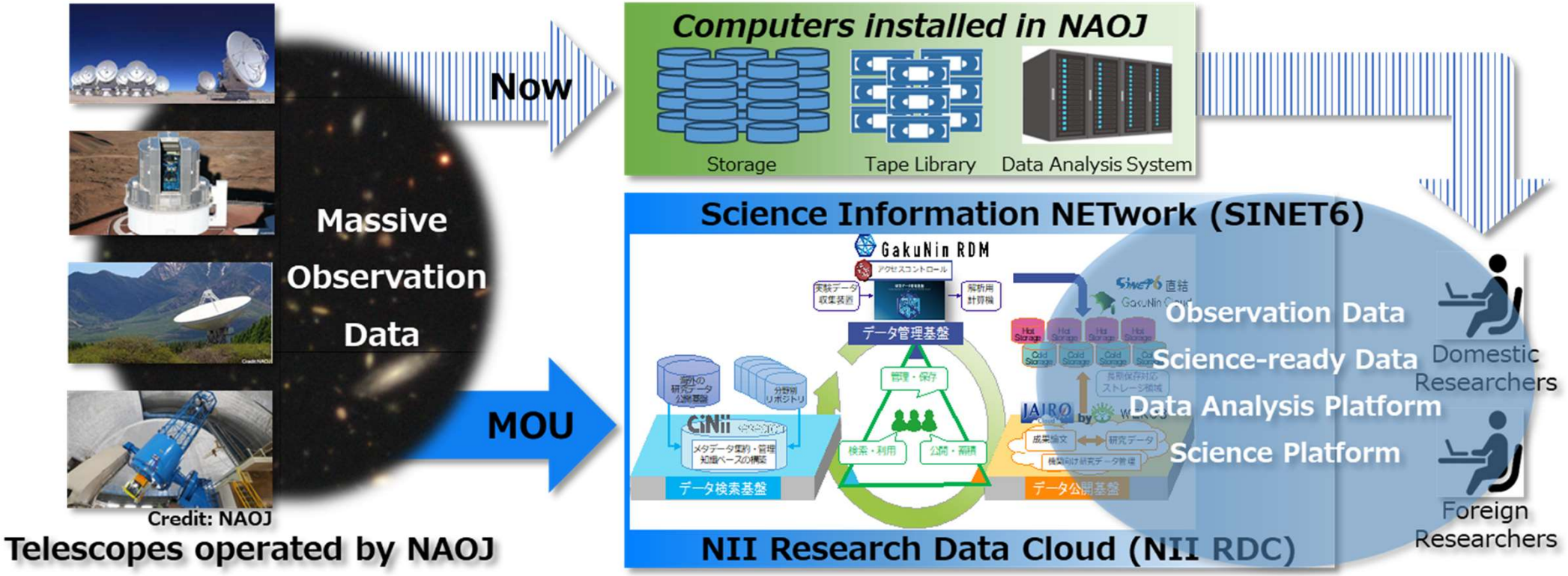
- Science Platform is being built for Subaru HSC/PFS and Rubin/LSST
- Open-use Data Analysis Science Platform may be operational in 2028-29

Strategy 4 (1/3) : Introduction of new Technology

- We are entering an "era of huge data", in which the amount of observation data is exploding with the evolution of observing instruments.
- Cloud computing is considered one of key technologies for data services in the “era of Huge Data”
- To sustain archive operation and data services, ADC has signed an agreement (MOU) with National Institute of Informatics (NII) for experimental implementation to evaluate and utilize cloud infrastructure

Strategy 4 (2/3): MOU with NII

- Check usability, functionality, and long-term cost profile for archives
- Evaluate the operation of data analysis/research environment



Strategy 4 (3/3): for Future Projects

- Cloud computing is considered one of key technologies for data services in the “era of Huge Data”

- Easy to scale (storage volume, computational resources)
- Easy to connect between multiple data and heterogeneous services

Timeline:

- MOU with NII covers 2-year term for demonstration experiment

Summary

- Strategy 1: Organization
  - Cross-organizational project support
  - 組織横断的なプロジェクトサポート
- Strategy 2: Consolidation of IT infrastructure and Data Services
  - Cooperative operation of multiple archives and future consolidation of them
  - アーカイブ機器の三鷹集約、複数アーカイブの連携と将来の融合
- Strategy 3: Development and Implementation of a mechanism for handling Huge Data
  - Preparation of Science Platform, a research environment in the “era of Huge Data”
  - 巨大データ時代の研究環境 = Science Platform
- Strategy 4: Introduction and Verification of new technology
  - Evaluation of data services on Cloud
  - データサービスのクラウド環境移行に関する実証実験

Astronomy Data Center
国立天文台 天文データセンター

Comments & Questions

国立天文台 NAOJ National Astronomical Observatory of Japan