An Implementation of International Data Cloud

Moving forward with routine data service between trusted international collaborators

Chia-Chuan Chuang, Hsi-En Yu, Hsui-Mei Chou, Weicheng Huang*

NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING, NARL, TAIWAN

Overview

Data needs to be shared and integrated to create values. The global data sharing between international partners requires a sustainable trust-based platform. Starting April 2018, a 4-year international data cloud effort between AIST Japan and NCHC Taiwan was initiated to prototype a platform for sharing both the data and the computing power.

Since the Data Cloud has the nature of distributed data across organizational boundaries, it has to tackle with data sharing and safety, user management, etc.

There are 4 major work packages that will be implemented simultaneously through out the 4 years period:

- Connecting the data center between international partners, namely
  - Advanced Industrial Science and Technology of Japan
  - National Center for High-performance Computing of Taiwan
- Secured data sharing and exchange
- Accessing data center from IoT devices
- User access control

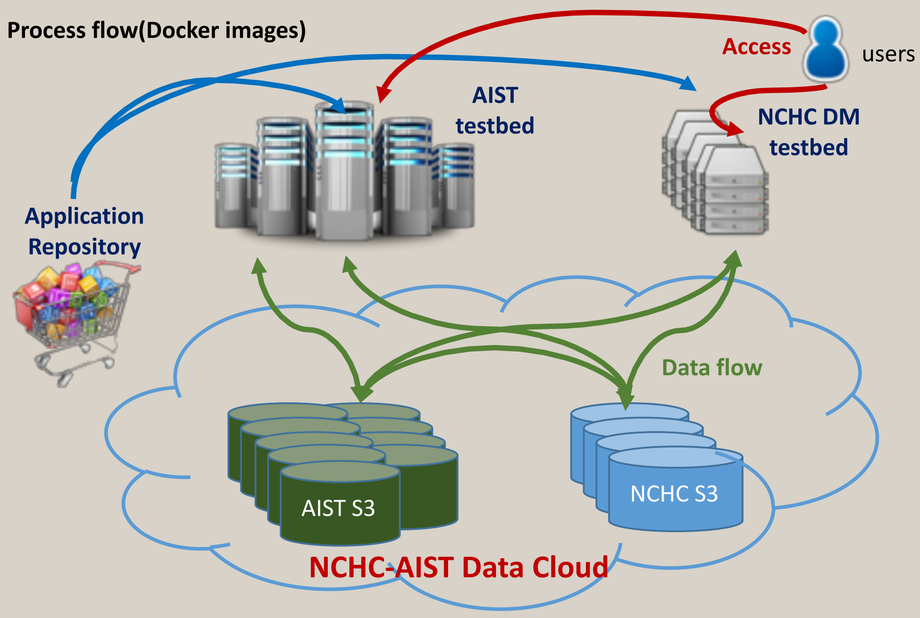

Figure1, Data Cloud between NCHC & AIST, sharing storage, computing, and data

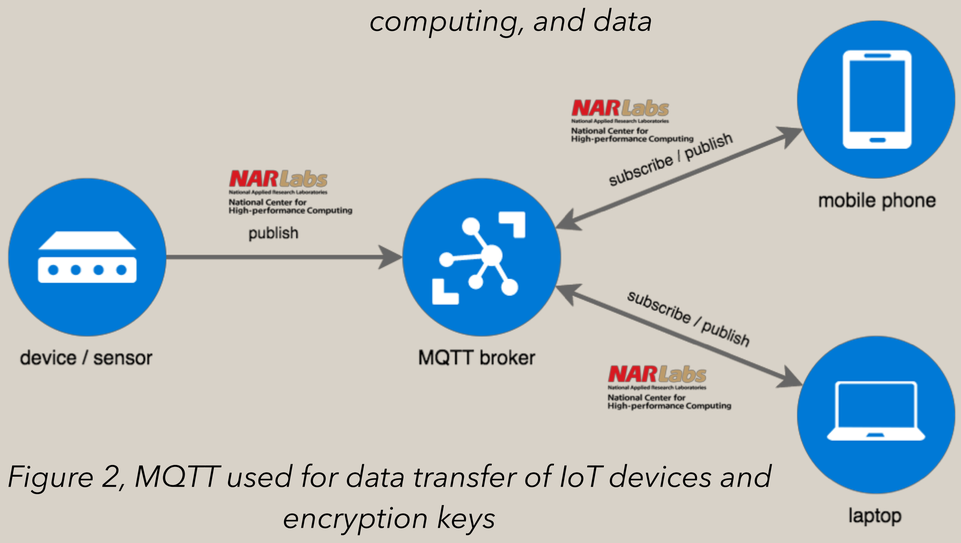

Figure 2, MQTT used for data transfer of IoT devices and encryption keys

Ongoing works

- Data safety & security : In addition to visible watermarking to protect the sensitive image, the image after watermarked will be encrypted before transferred via internet. The keys of encryption and the watermarking will be sent separately to the receiver to decrypt and remove the watermark from the image.
- Persistent identifier (PID) : PID is a long-lasting reference to digital objects that are accessible over the internet. To address the credits of digital contents, PID is implemented to facilitate such goal. In addition, it will be used to trace the origins of digital objects collected in the data cloud.
- User authentication & authorization : Since IoT devices are targeted to be connected to the data cloud, the feature of users' authentication and authorization should consider the needs from field data and mobile devices.
- AI application : To help pushing forward the implementation of the data cloud, AI-oriented solar panel detection is adopted to drive the system implementation. The project is constantly looking for AI-based applications to diversify the scope of implementation. For example, the team is looking into the possibility of using AI in the process of design of geometry with reduced drag.
- Disaster recovery : Generally speaking, data cloud does not need to make replica of remote data. However, in reality, disaster might happen and data will be demanded before infrastructure can be recovered. It is one of core features, to provide Disaster Recovery mechanics, to ensure the safety and availability of data during disaster.
- Catalog service : The catalog service is the central of information sharing, dissemination or information awareness. For systematic and efficient data access, references of existing datasets and services will be published in a catalog.

Implementation

Over the past year, the project has achieved the followings:

- Decentralized storage
  - S3-compliant storage via CEPH
  - Block/object/filesystem storage
  - 100TB storage
- Repository for containerized application
  - Docker
  - Singularity
- Data platform via NextCloud
  - Federated data sharing
  - Web. access
- Generic lossless visible watermarking for image data safety
  - Cover image with watermark selected to protect the image
  - Parallelized with multiple cores
  - Performance : Image size of 500MB w/14000x12500 pixels took 150 sec. w/15 cores
- MQTT for data transfer
- Demonstrative AI application
  - Images from satellites of Japan & Taiwan
  - AI application to solar panel detection
  - AI model training and inference
  - Containerized

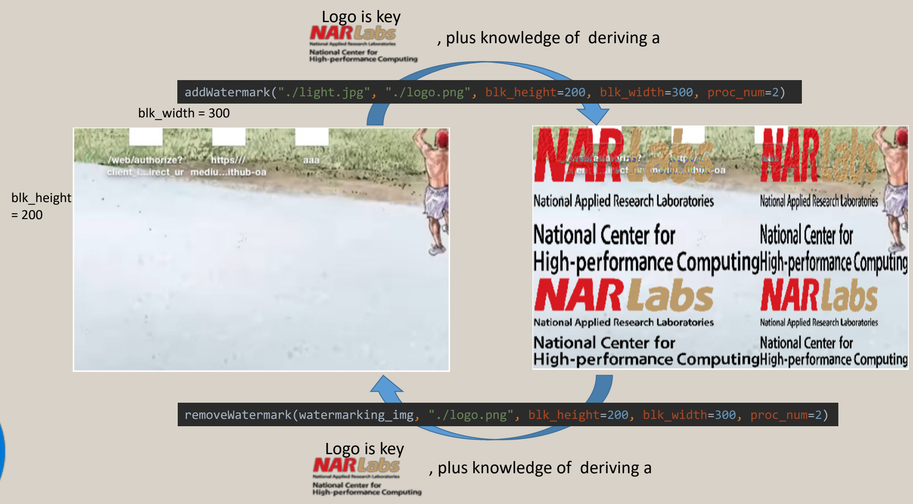

Figure 3, Implementation of watermarking

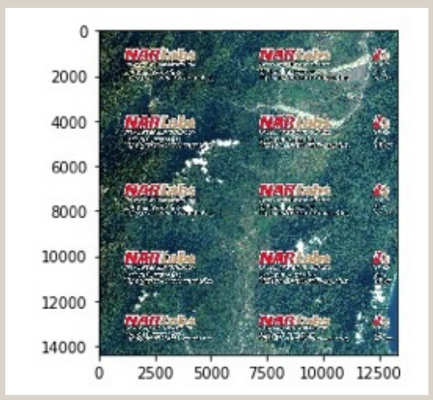

Figure 4, Watermarked image for data protection

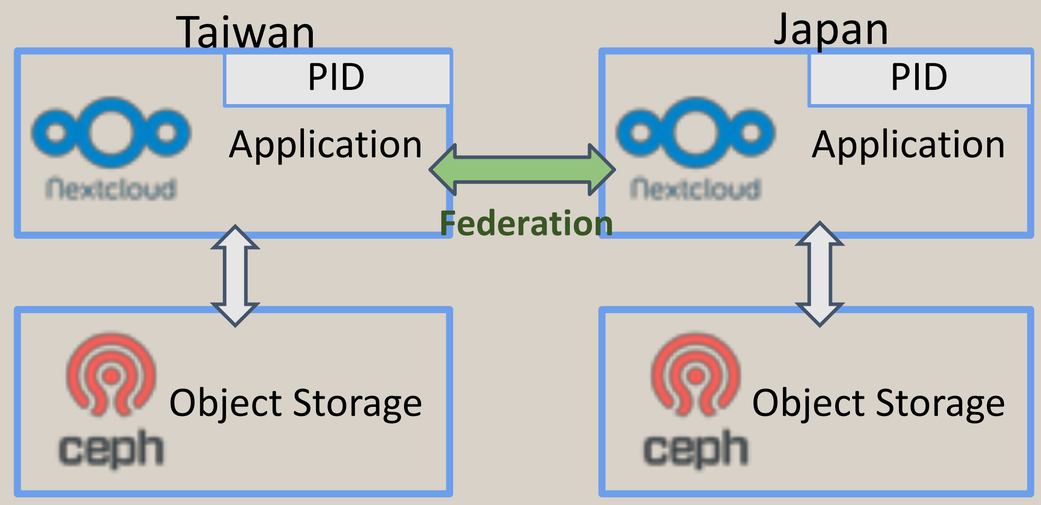

Figure 5, PID mechanism between AIST & NCHC will be implemented and integrated with data federated

Collaborators

- Advanced Industrial Science and Technology, Japan
- National Space Organization, NARL, Taiwan
- National Center for High-performance Computing, NARL, Taiwan

Reference

- NextCloud : https://NextCloud.com
- CEPH : https://ceph.com
- Singularity : https://www.sylabs.io/singularity
- Docker : https://www.docker.com
- Container : https:/github.com/containers/libpod
- *Generic Lossless Visible Watermarking – A New Approach*, IEEE Transactions on Image Processing, Vol. 19, No. 5, May 2010.

***contact : whuang@nchc.narl.org.tw, whuang91362@gmail.com**