ArrayMorph: Efficient Edge-to-Cloud Data Management for Machine Learning

Tinggang Wang, Ruochen Jiang, Rakesh Rajeev, Spyros Blanas
The Ohio State University

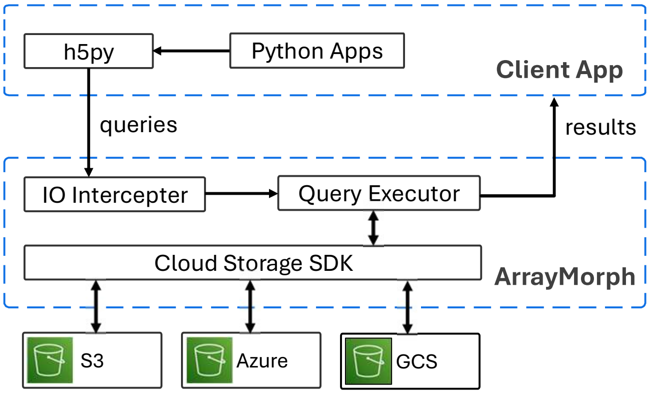
Abstract

Cloud object storage is common for sharing machine learning data, but transferring entire datasets—including irrelevant portions—from the edge is inefficient, increasing costs and transfer times. To address this, we present ArrayMorph, a software for efficiently managing array data on cloud object storage. ArrayMorph streamlines preprocessing at the edge, allowing users to upload only relevant data subsets. This reduces unnecessary transmission, lowers storage overhead, and improves efficiency for distributed ML pipelines.

1. Background and Motivation

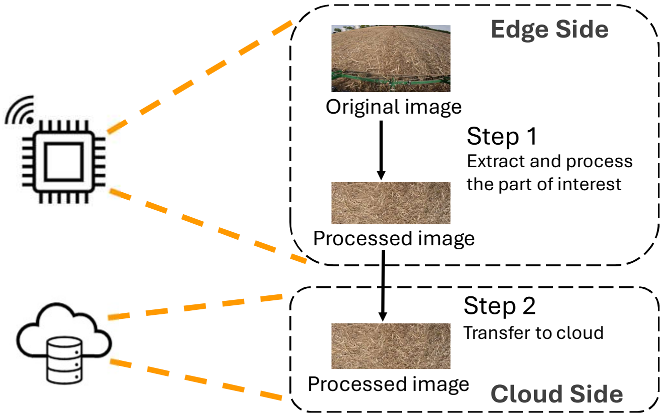
- Cloud object storage
  - Scalable, durable, and pay-as-you-go
  - Often used for storing raw data for ML tasks
- Data transfer from edge to cloud
  - State of the art: transfer the entire dataset
  - ArrayMorph: preprocessing -> transfer subsets
- HDF5
  - Ease of use in storing/retrieving NumPy arrays

2. System Overview



- ArrayMorph is implemented using the Virtual Object Layer (VOL) feature of HDF5 and serves as an I/O interceptor between the HDF5 library and cloud storage services
- ArrayMorph exposes the API of h5py and is loaded as a dynamic plugin. Users can seamlessly connect existing applications to the cloud by using ArrayMorph without code changes

3. Digital Agriculture Use Case

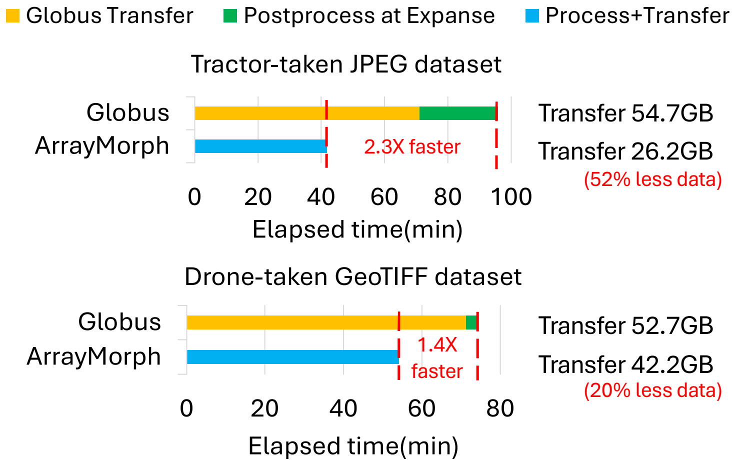


4. Evaluation

- Platform: NVIDIA Jetson
- Baseline: Globus

	ArrayMorph	Globus
Transfer Scope	Edge to Azure Blob Storage	Edge to SDSC Expanse cluster (Lustre filesystem)
Transfer type	Transfer objects	Transfer files
How to use	Via numpy and h5py	Submit a globus job
Processing	Preprocess at the edge	Postprocess on Expanse compute node

- Datasets
  - Tractor-taken JPEG dataset:
    - 55GB, 16,800 images
    - Extract a fixed quadrilateral region + rectify
  - Drone-taken GeoTIFF dataset:
    - 53GB, 180 images
    - Crop each image by 20%


🟨 Globus Transfer
🟩 Postprocess at Expanse
🟦 Process+Transfer
Tractor-taken JPEG dataset
Globus
ArrayMorph
2.3X faster
Transfer 54.7GB
Transfer 26.2GB
(52% less data)
Elapsed time(min)
Drone-taken GeoTIFF dataset
Globus
ArrayMorph
1.4X faster
Transfer 52.7GB
Transfer 42.2GB
(20% less data)
Elapsed time(min)