An Introduction to DosNa

Distributed NumPy Arrays for High-performance cloud computing

Gabryel Mason-Williams

gabryel.mason-williams@rfi.ac.uk

The Rosalind Franklin Institute

Contents

- Introduction to The Rosalind Franklin Institute
- Background to problem
- Problem
- What is DosNa
- DosNa features
- How DosNa works
- How DosNa solves the problem
- Case Study: SAVU
- Features coming
- Summary and Conclusion

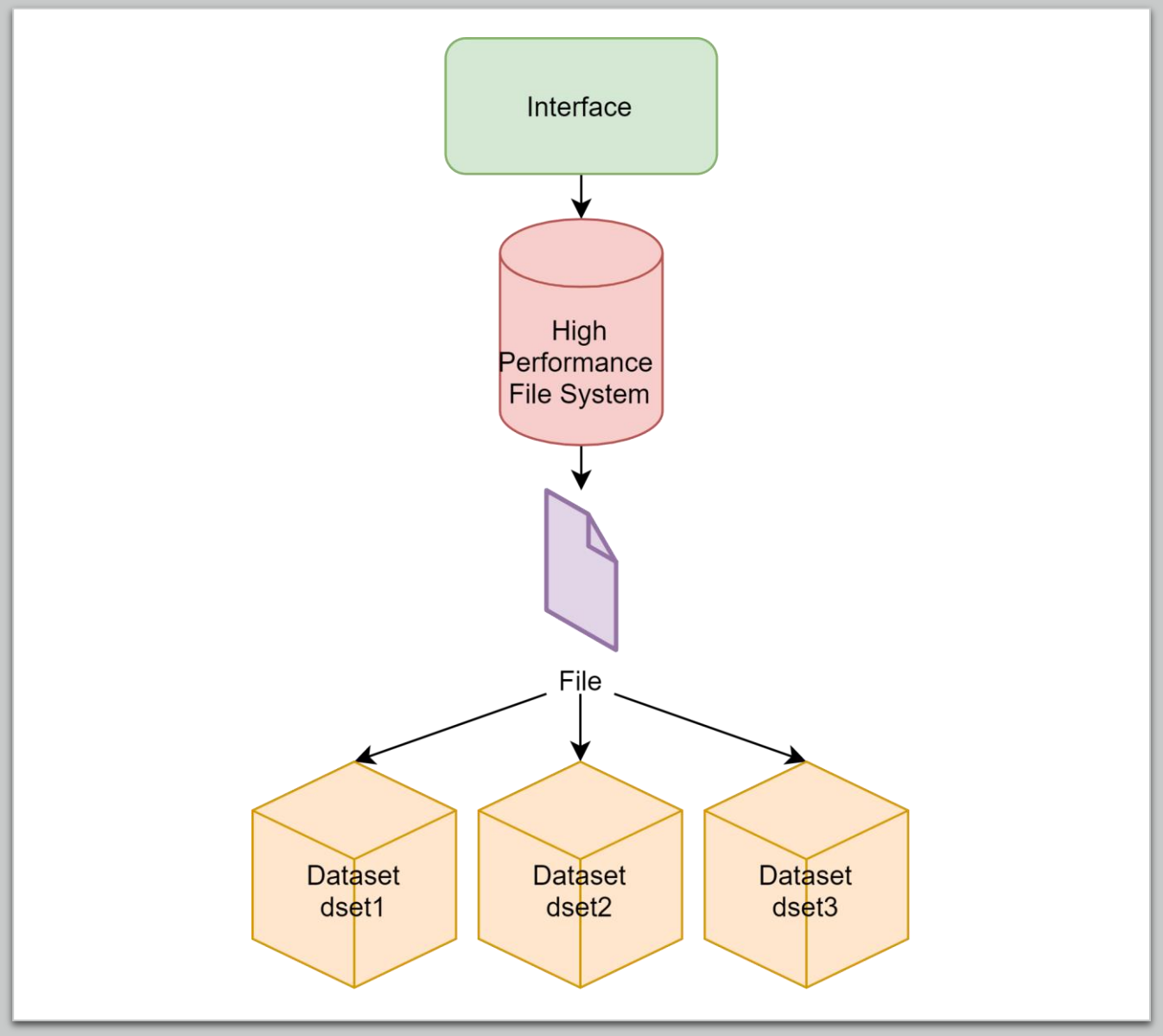
The Rosalind Franklin Institute

Rosalind Franklin Institute

The Rosalind Franklin Institute

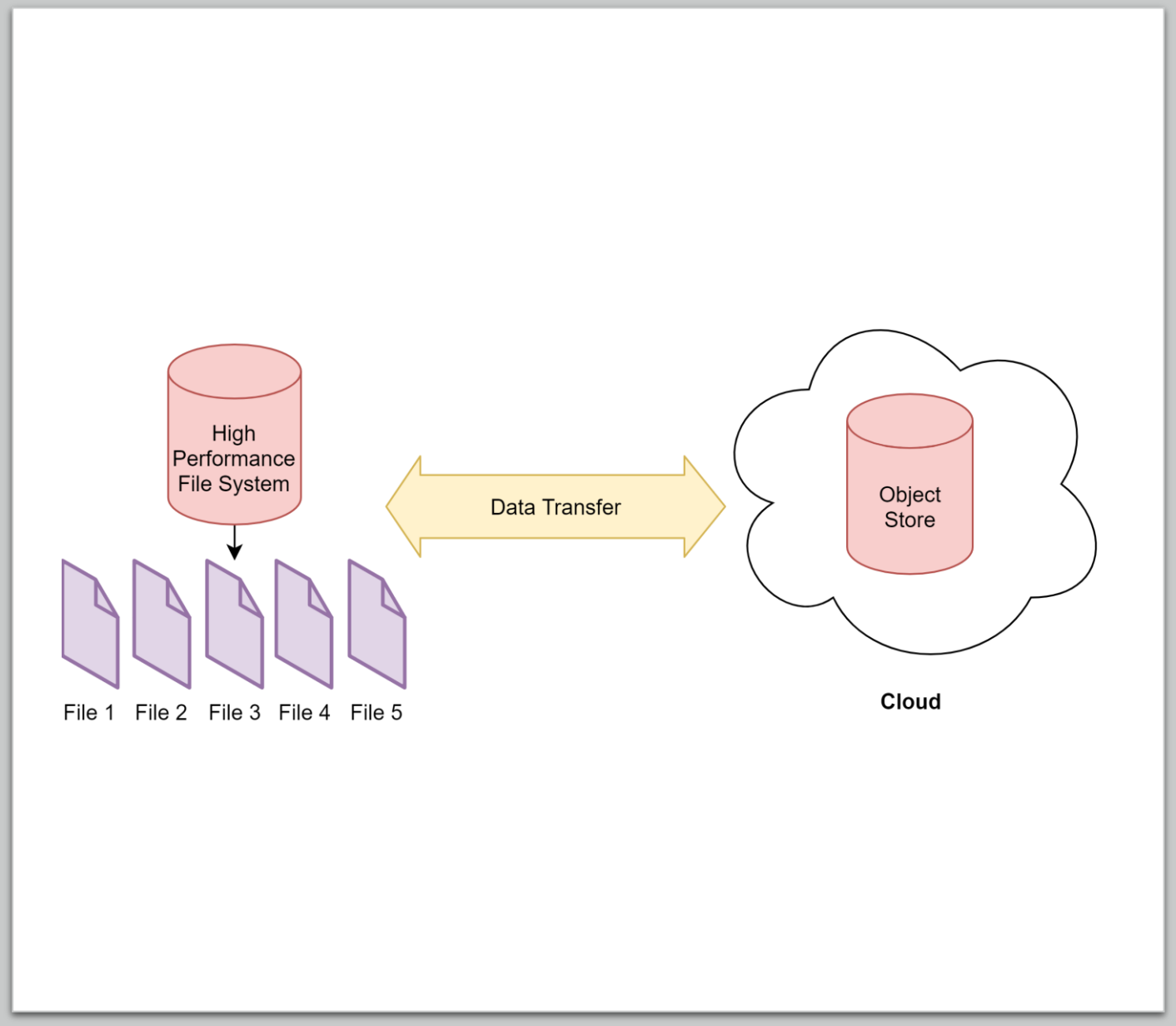
Background To The Problem

- Current user workflow
  - File interaction



Problem

- Moving large data on and off the cloud is difficult
- Migration can be expensive
- Increased storage requirement
- Time consuming



Solution

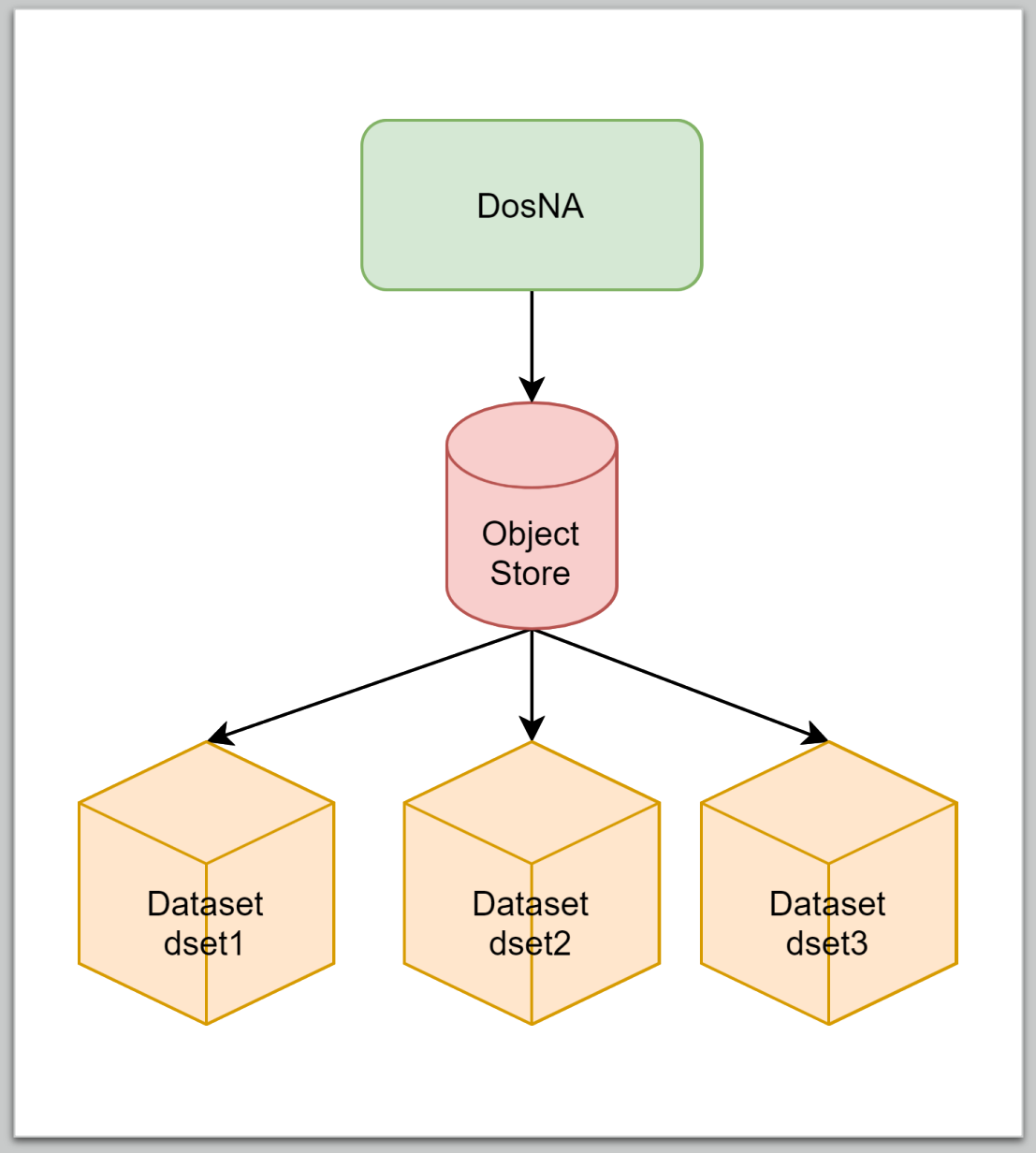
DosNa

https://github.com/rosalindfranklininstitute/DosNa

The Rosalind Franklin Institute

What Is DosNa?

- **D**istributed **O**bject **S**tore **N**umpy **A**rray
- DosNa is a python wrapper that can distribute N-dimensional arrays over an Object Store server
- Storage Backends: Ceph, S3, and in memory RAM



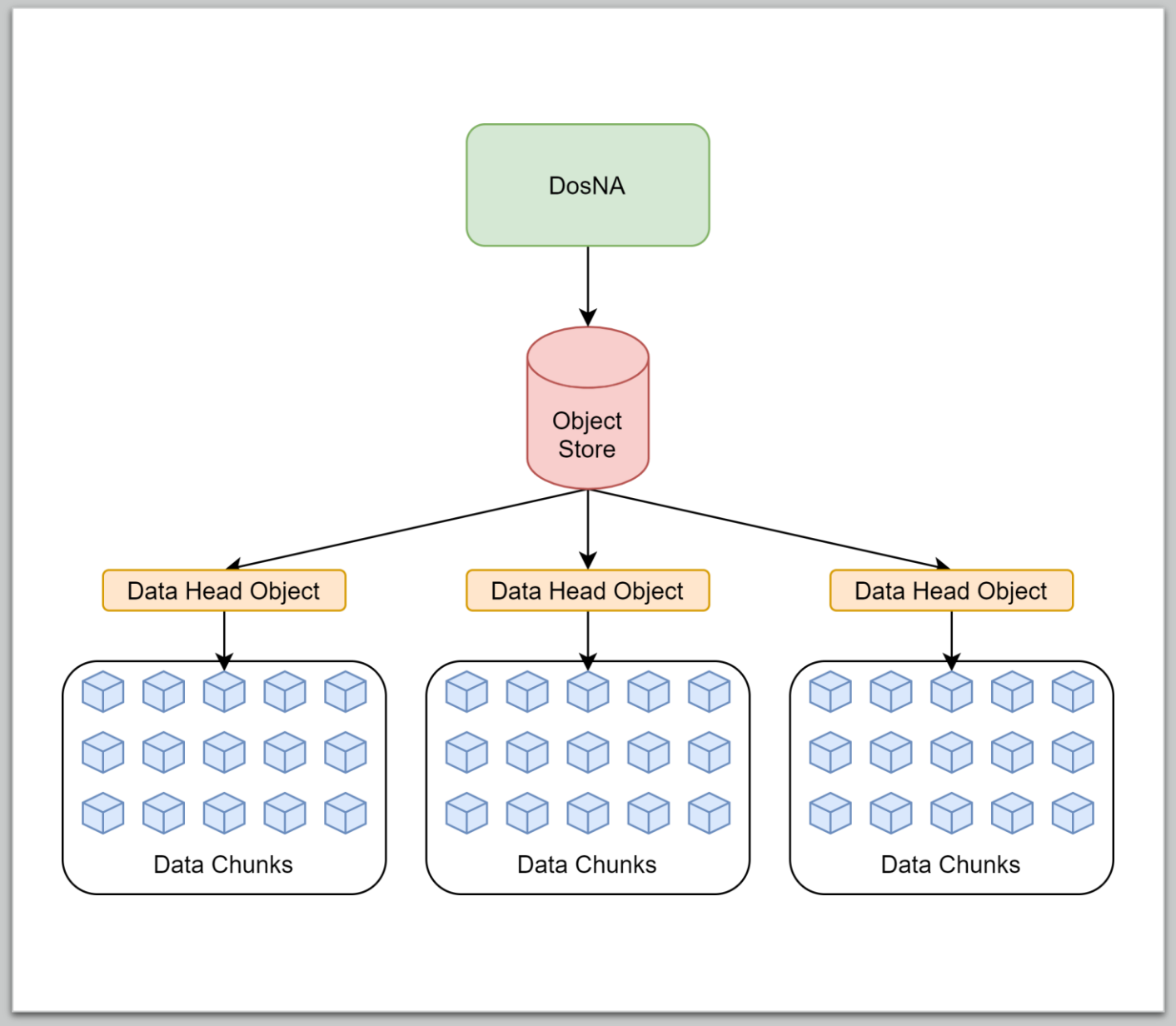
DosNa Core Features

- Chunked Datasets
- Hierarchal Structure
- Parallelism

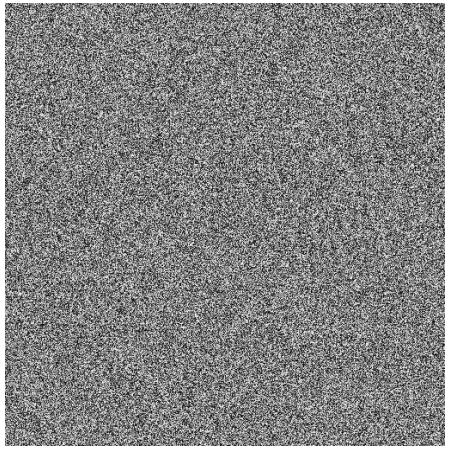
The Rosalind Franklin Institute

Chunked Datasets

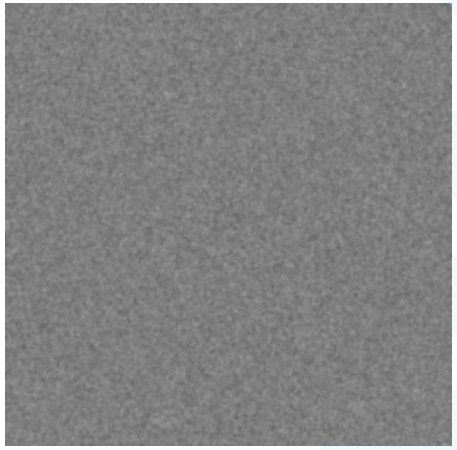
- Creation of Chunked Datasets
- Depending on access pattern will determine chunking size, i.e. horizontal, vertical, gridded



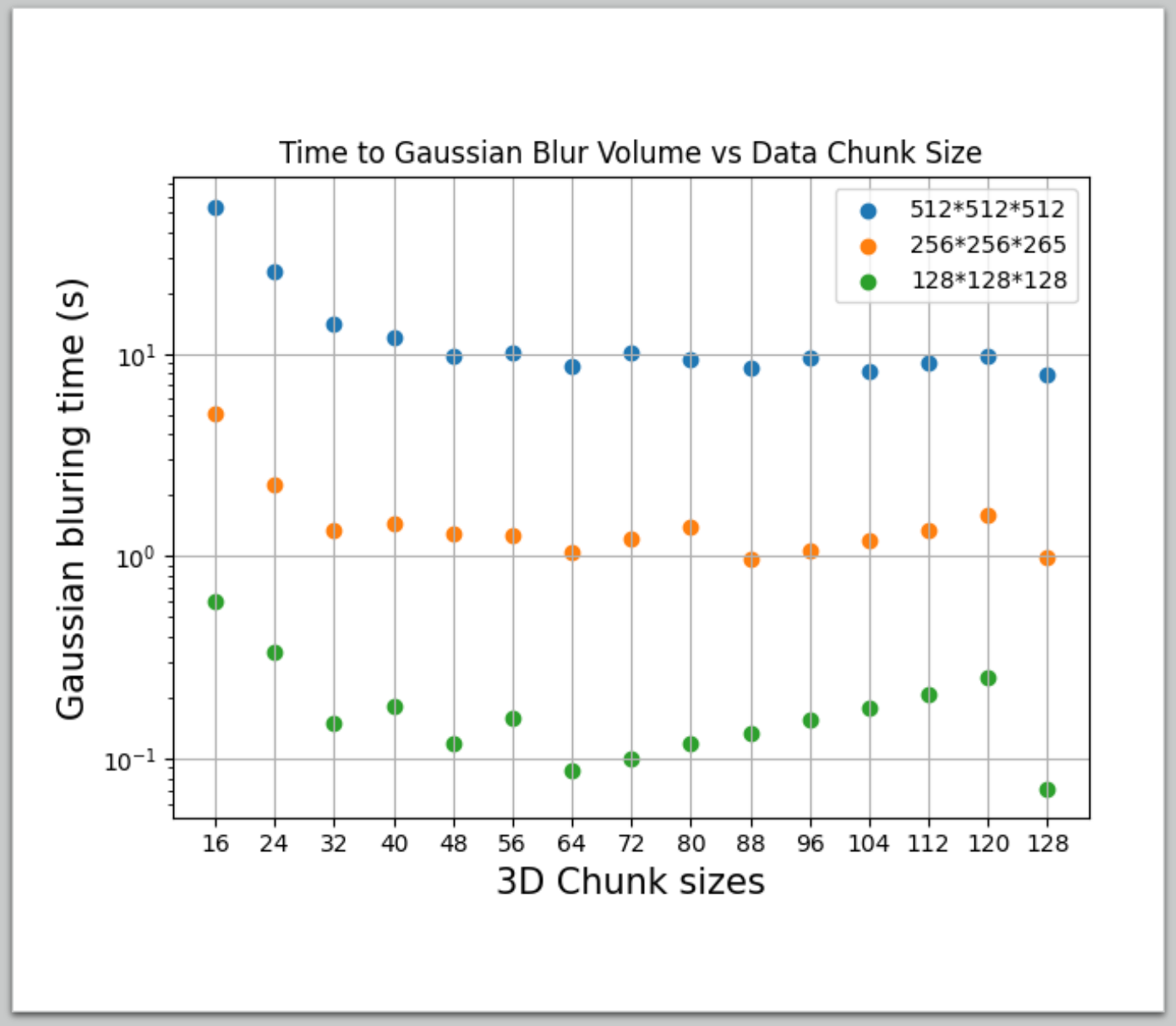
DosNa Chunked Dataset Example: Gaussian Blur



RAW Volume Slice

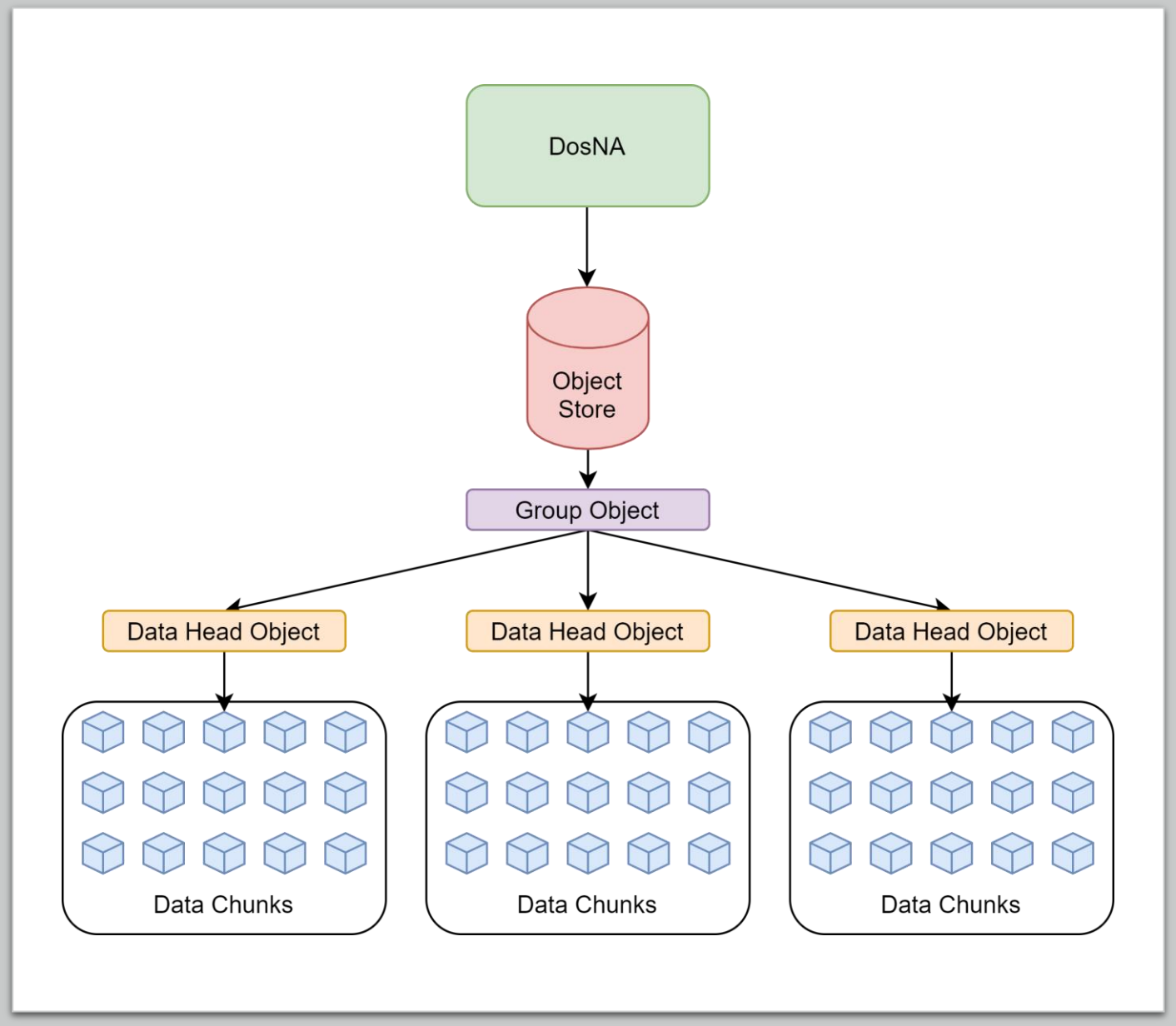


Blurred Volume Slice



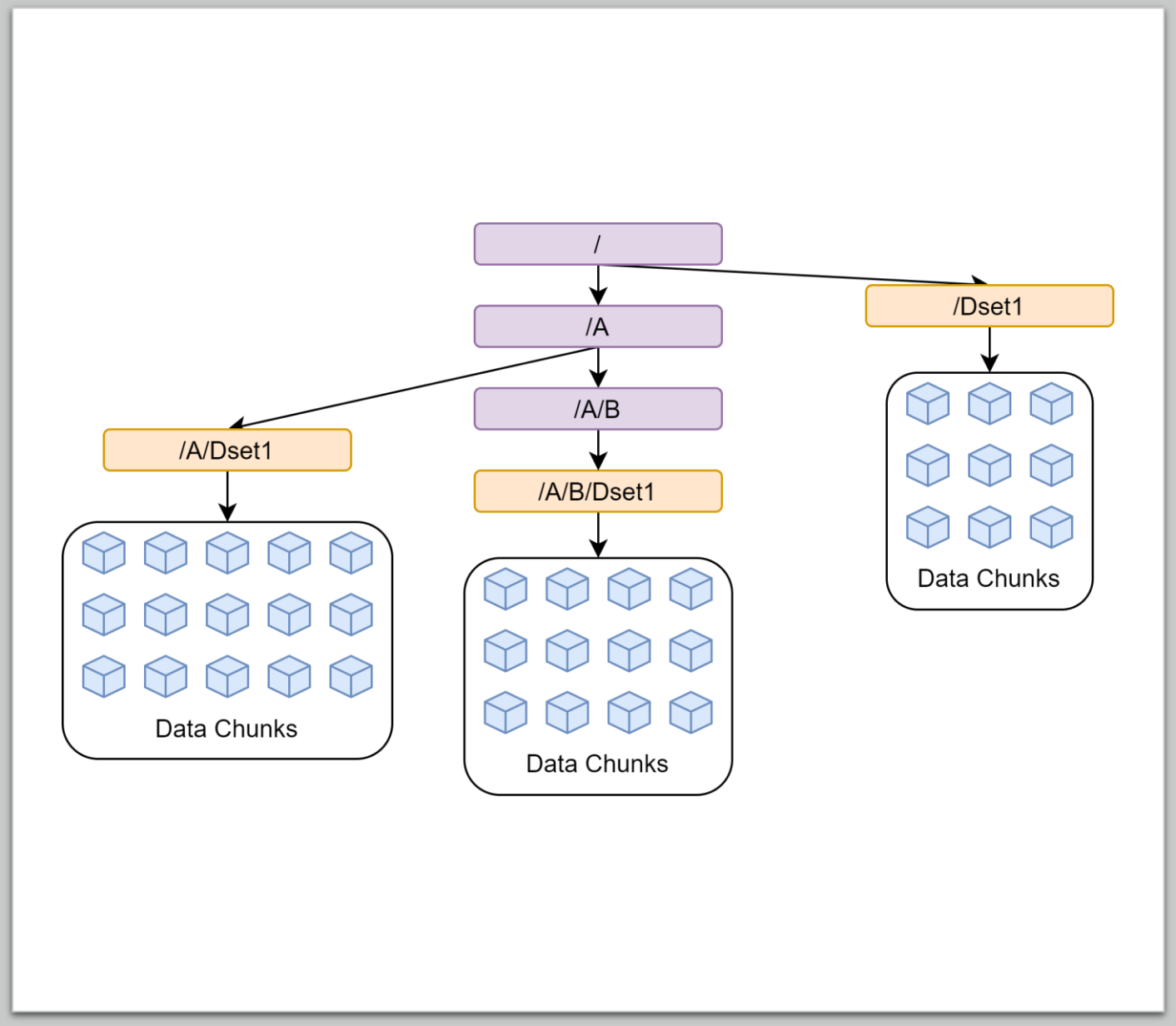
Hierarchical Structures

- Creation of Hierarchical Structures Via Groups
- Linking Datasets and Groups together
- Addition of Group Metadata



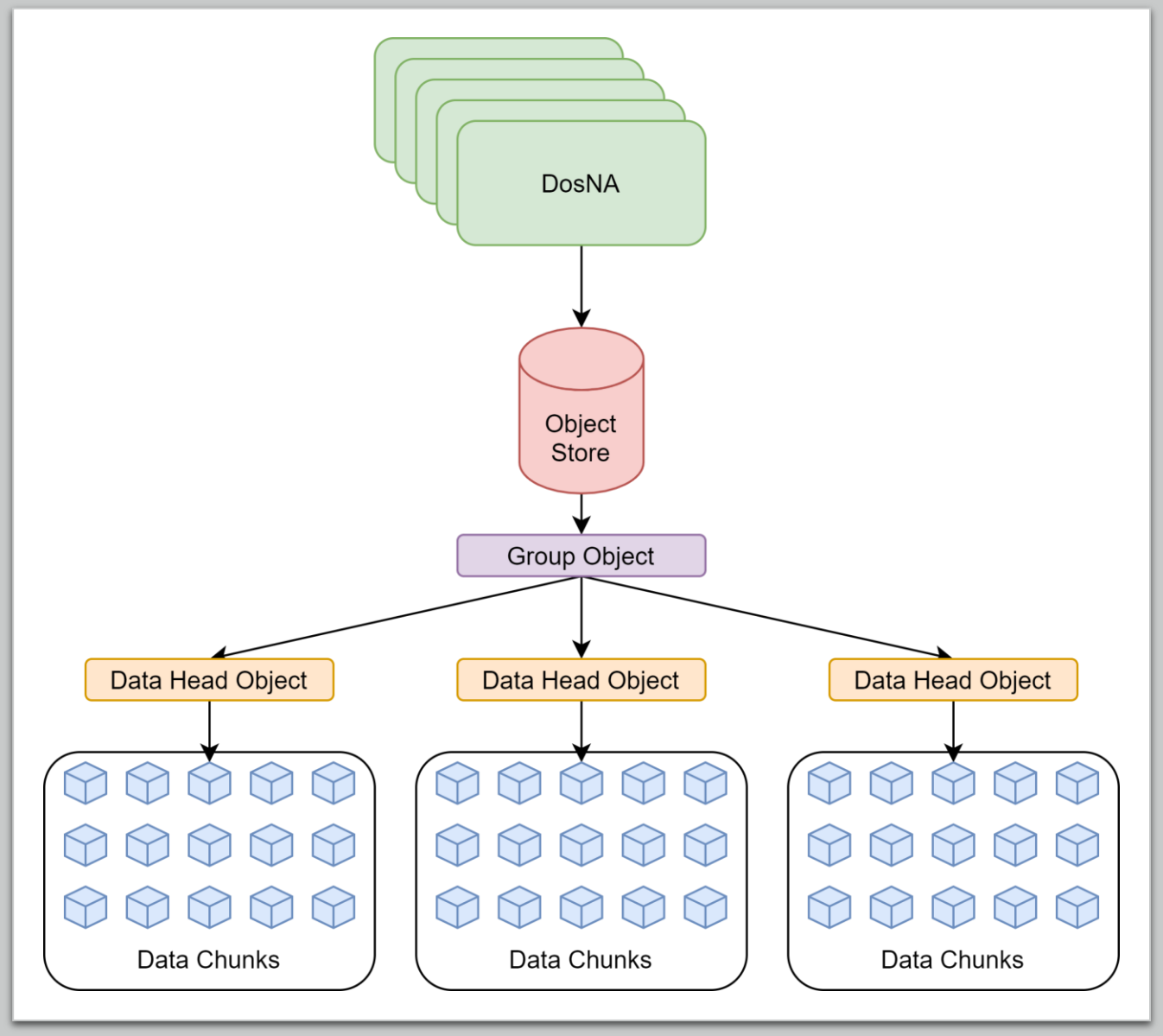
DosNa Hierarchical Structure Example

- This allows for converting HDF5 files to DosNa Objects seamlessly
- DosNa tool to convert files



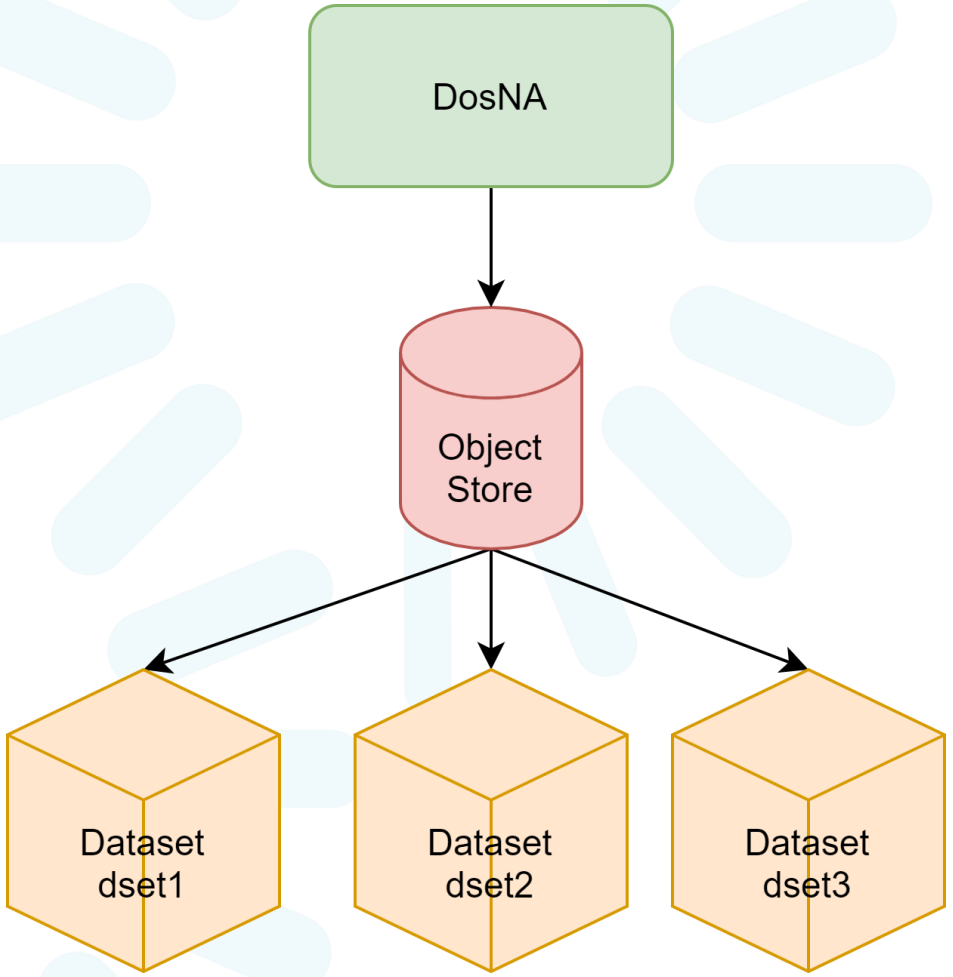
Parallelism

- Multithread and Multiprocessor parallelism through Joblib and MPI



How DosNa Works?

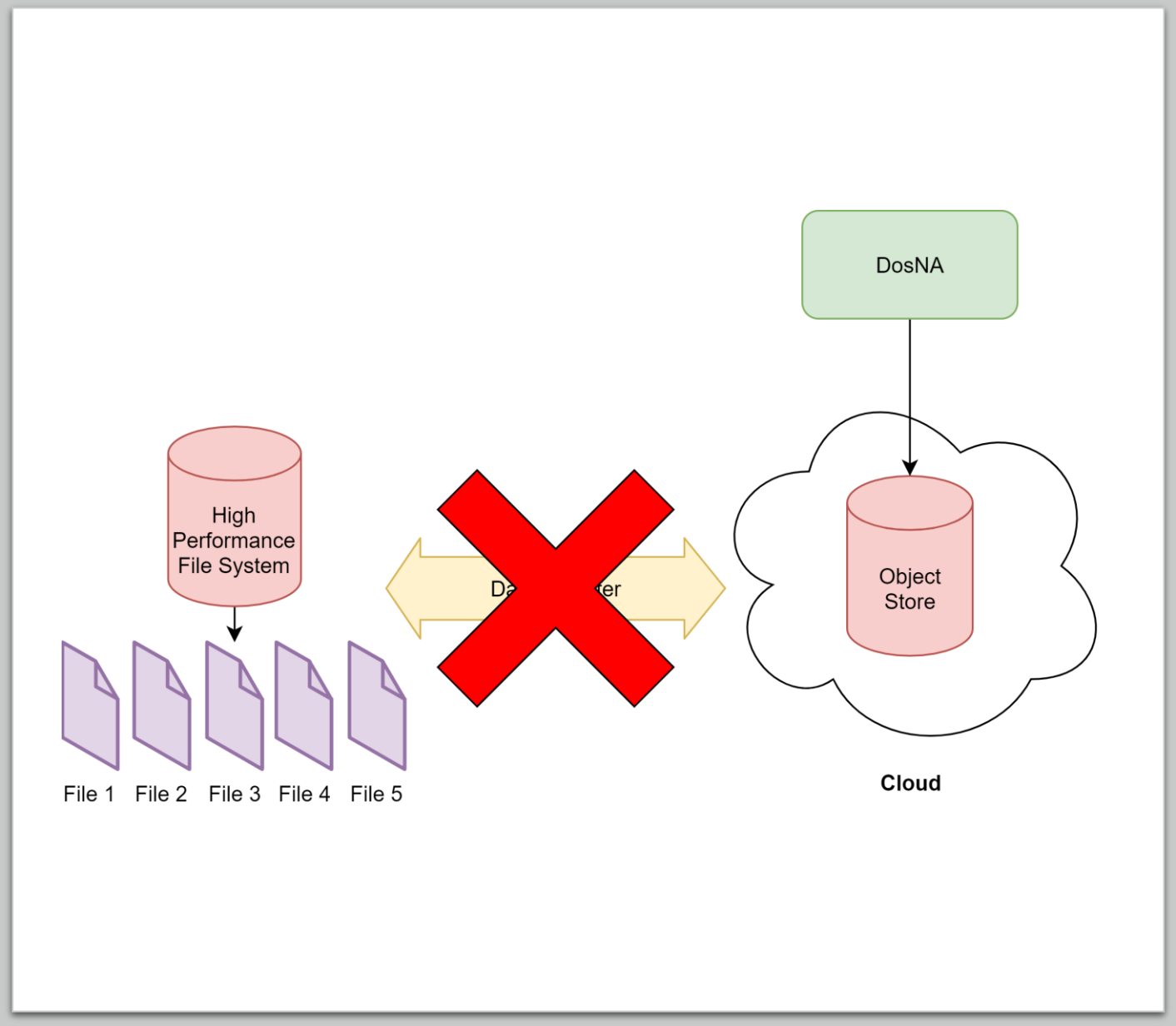
- Connection to object store backend
- DosNa standard numpy slicing, with modifications taken care of behind the scenes.
- An Dosna Dataset can be used as an H5 Dataset object or a Numpy Array.
- Simple/Familiar interface to the user



The Rosalind Franklin Institute

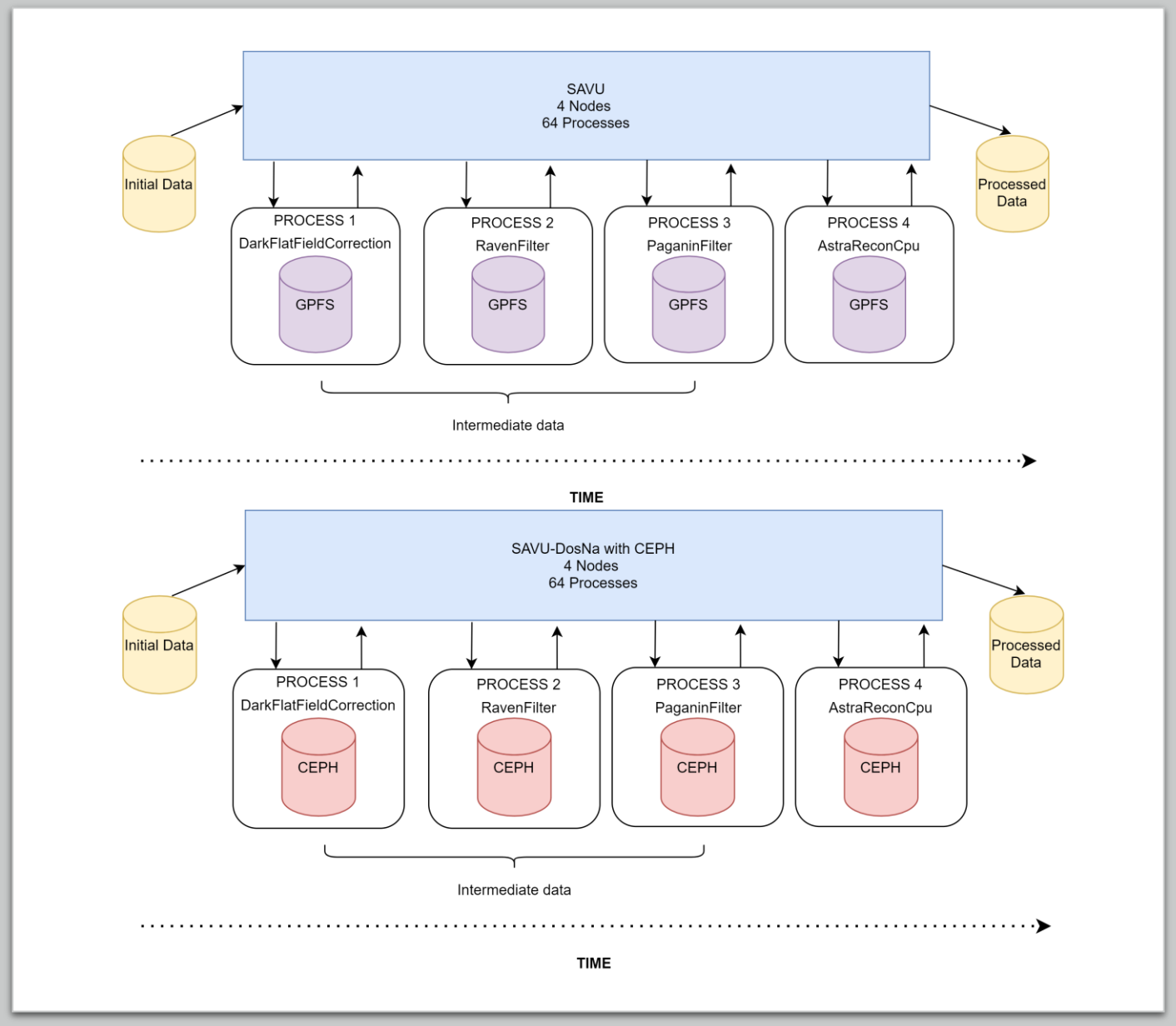
How DosNa Solves This Problem

- Removes data transfer requirement
- Keeps data in the cloud
- Familiar interface
- Drop in replacement



Case Study: SAVU

- SAVU: Tomography Reconstruction and Processing Pipeline
- Drop-in replacement for H5 Files on GPFS



Features to be added

- API to browse DosNa objects and visualize data
- Object locking
- Option for compression mechanism
- Option for checksums

The Rosalind Franklin Institute

Summary

- DosNa is a python wrapper that can distribute N-dimensional arrays over an Object Store server
- Supports Hierarchical Structures allows for converting HDF5 files to DosNa objects
- Parallelizable
- Currently underway API to visualize data, object locking, compression, and checksums
- Release date: July via Pypi

The Rosalind Franklin Institute

Thanks For Listening
Any Questions?

The Rosalind Franklin Institute