# Active I/O Systems: Providing Computing-Ready Data

Dr. Shu Yin
ShanghaiTech University

L.1.O.N
LABORATORY OF I/O SYSTEMS & DATA SCIENCE
并行与分布式I/O系统实验室

# My Research Domains

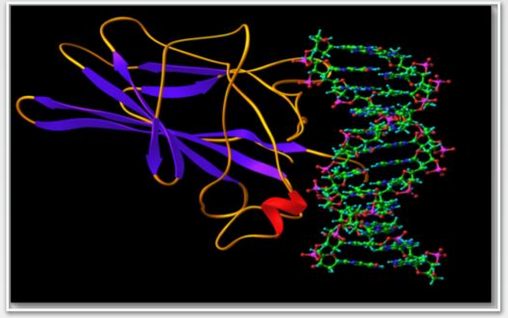
- Reliability Analysis
  - Parallel Disks Storage Systems
  - Cloud Storage Systems
- HPC Storage Systems
- Energy-Efficient of Storage Systems

- ADA: An Application-Conscious Data Acquirer for Visual Molecular Dynamics
- BORA: A Bag Optimizer for Robotic Analysis

L.1.O.N

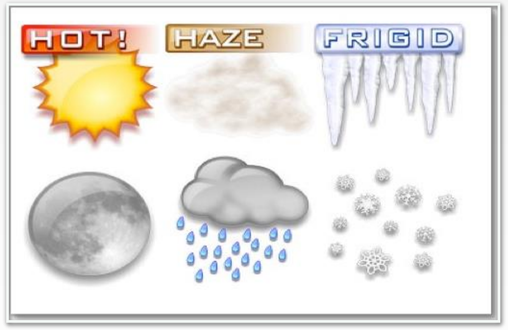
# Motivation


Stream Multimedia



Bioinformatic


3D Graphic

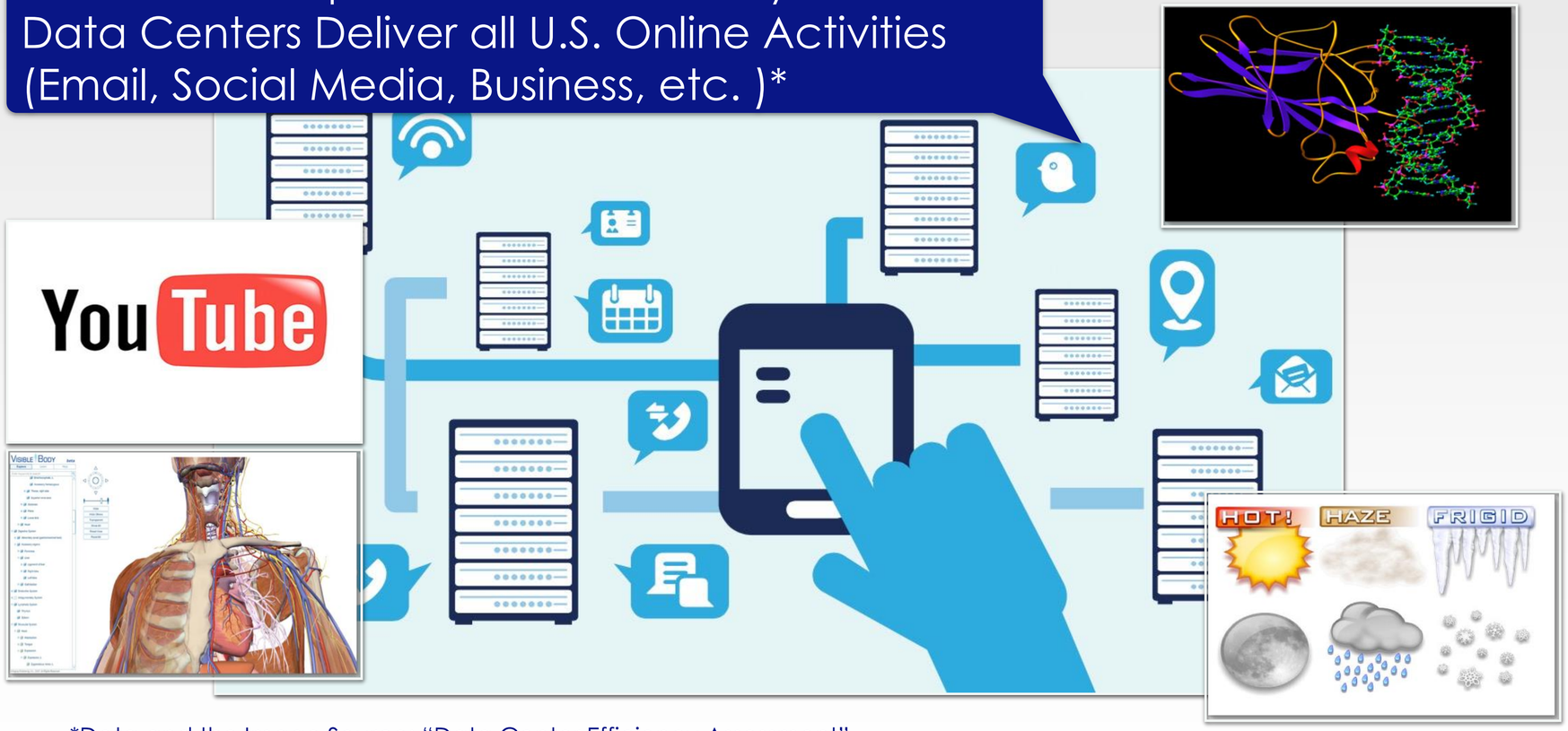



Weather Forecast

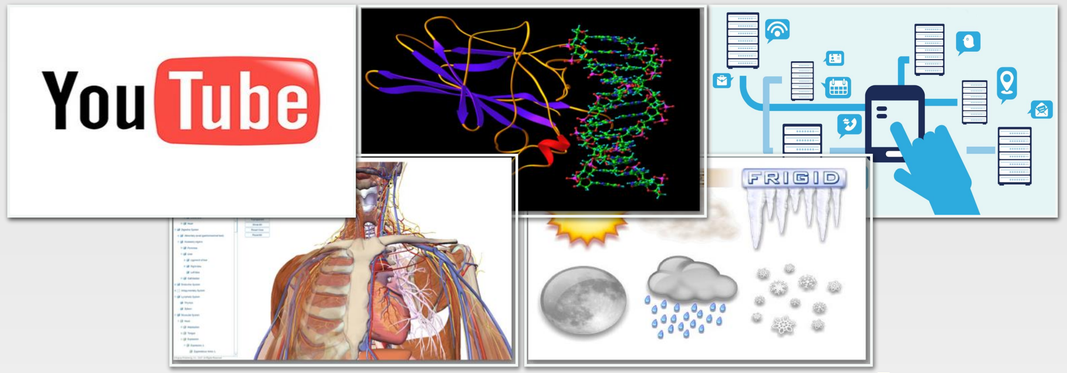
Data Intensive Applications

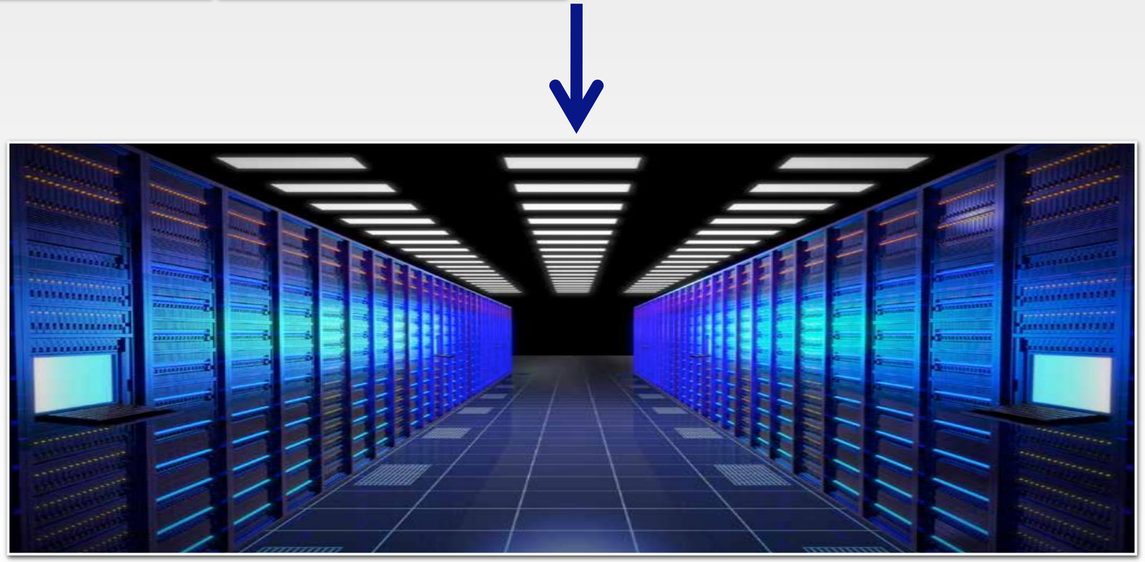
# Motivation

12 Million Computer Servers in Nearly 3 Million Data Centers Deliver all U.S. Online Activities (Email, Social Media, Business, etc. )*

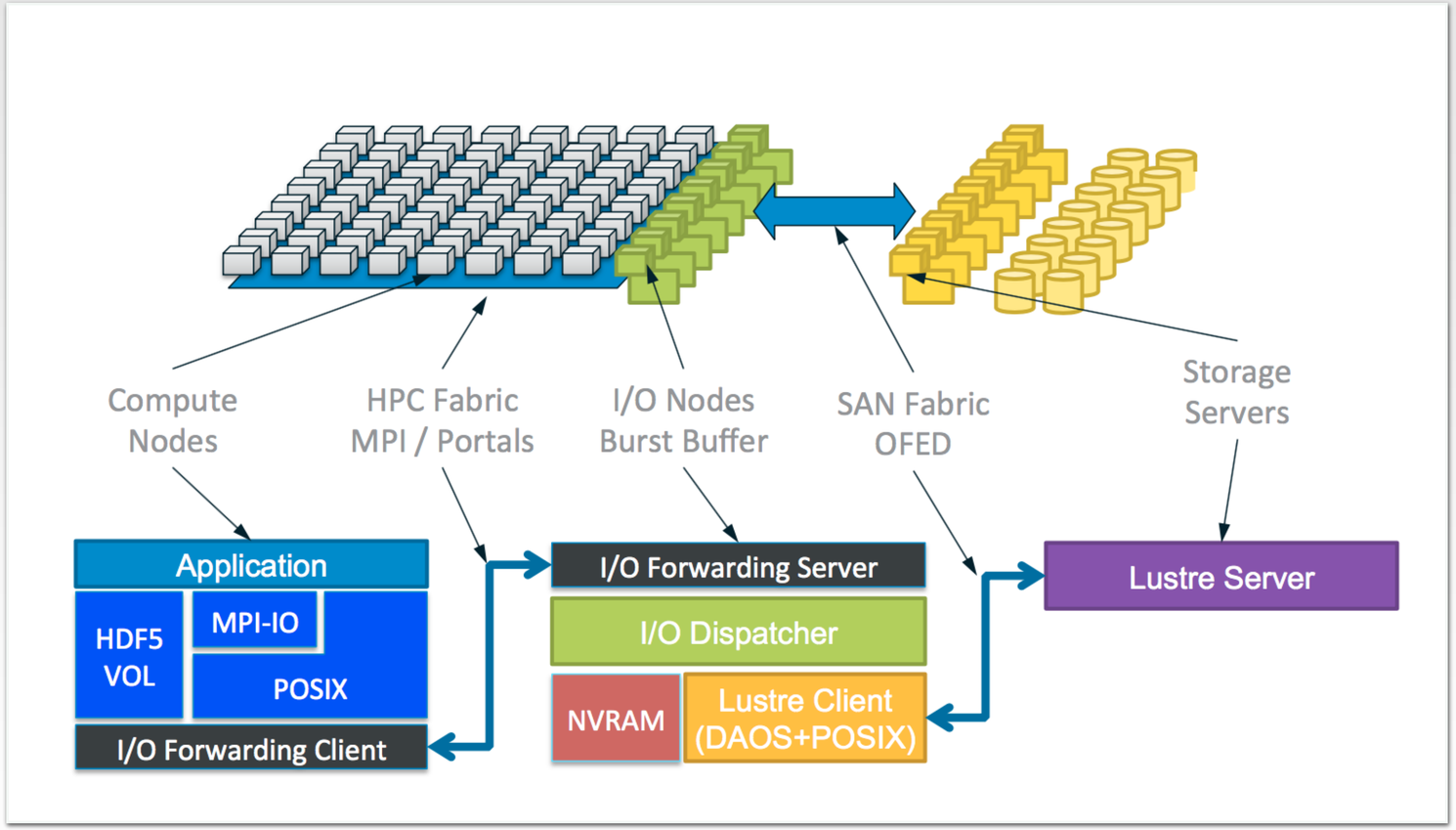



*Data and the Image Source: "Data Center Efficiency Assessment", www.nrdc.org/energy/data-center-efficiency-assessment.asp



*Data and the Image Source: “Data Center Efficiency Assessment”, www.nrdc.org/energy/data-center-efficiency-assessment.asp



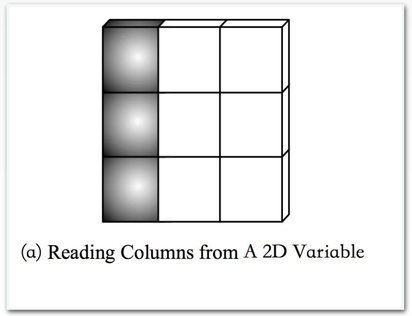

Jay lofstead, et. al. "DAOS and Friends: A Proposal for an Exascale Storage System", SC'16

L.1.O.N

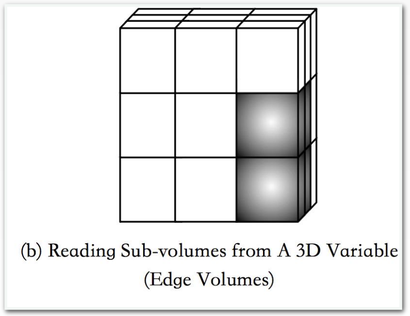
# HPC-IO:
# Makes Data Analysis Better

上海科技大学
ShanghaiTech University

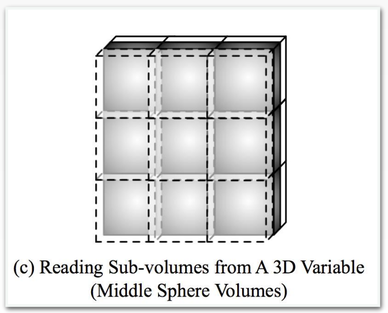
- In Order to Compute Data, Need to Follow A Pattern:
  - Transfer (I/O -> Memory)
  - De-Compress (Compressed Data -> RAW)
  - Distill (RAW -> Useful Datasets)
  - Compute



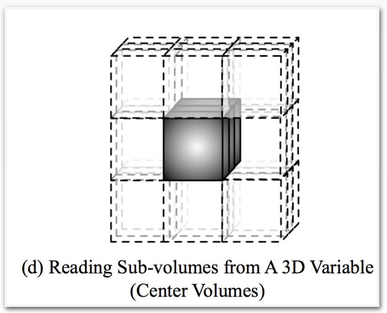
(a) Reading Columns from A 2D Variable



(b) Reading Sub-volumes from A 3D Variable (Edge Volumes)



(c) Reading Sub-volumes from A 3D Variable (Middle Sphere Volumes)



(d) Reading Sub-volumes from A 3D Variable (Center Volumes)

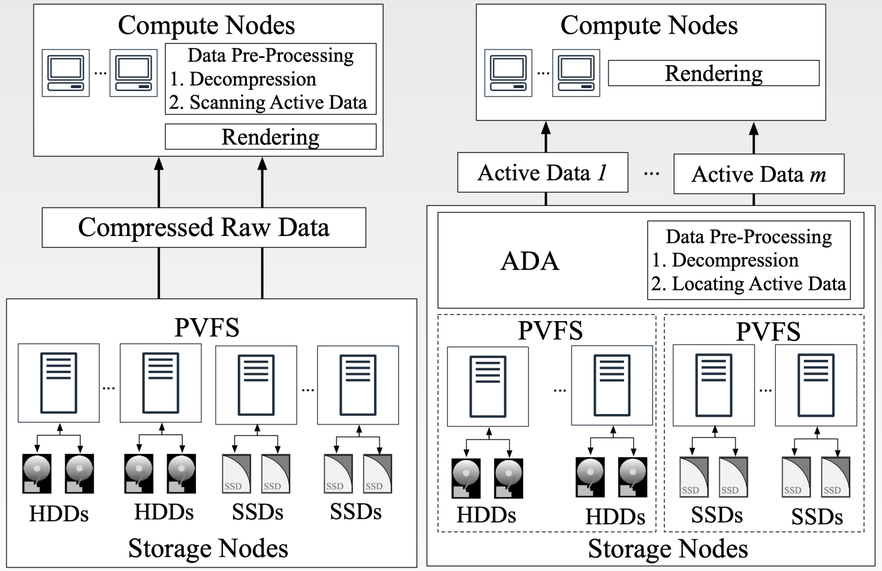
L.1.O.N

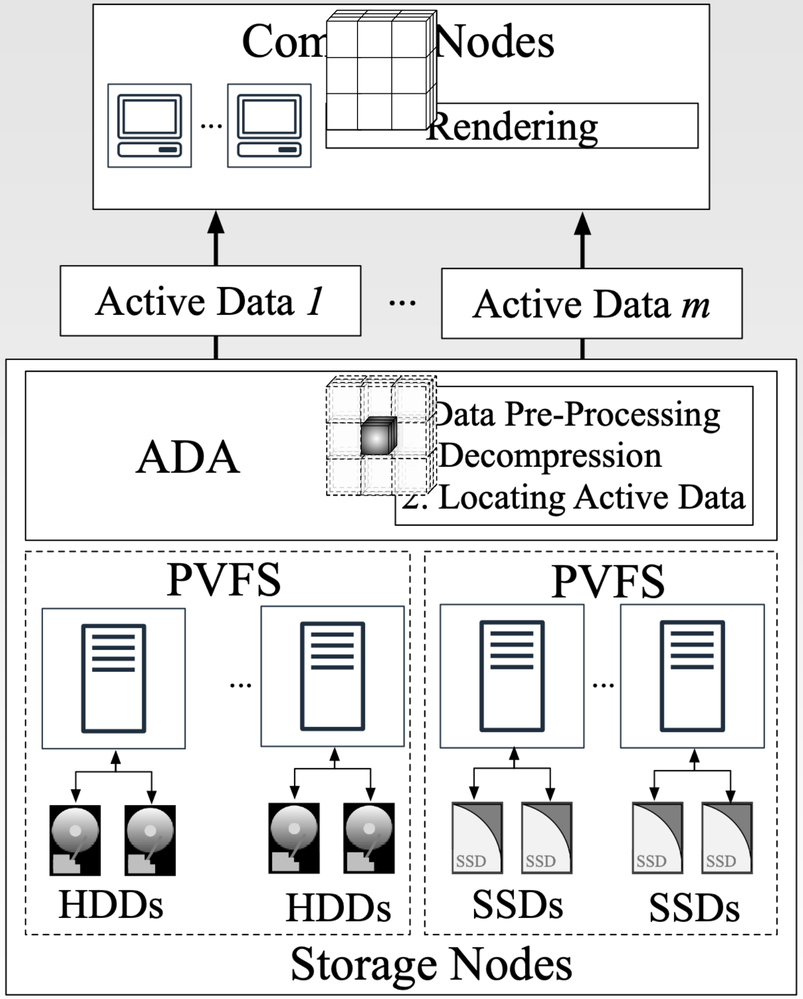
# ADA

- An Application-Conscious Data Acquirer for VMD
- Dedicated Hybrid Storage System
- Application Driven
- Divide Dataset into Few Sub-Datasets
- Provide Data on Current Usage

# Dedicated Hybrid Storage System

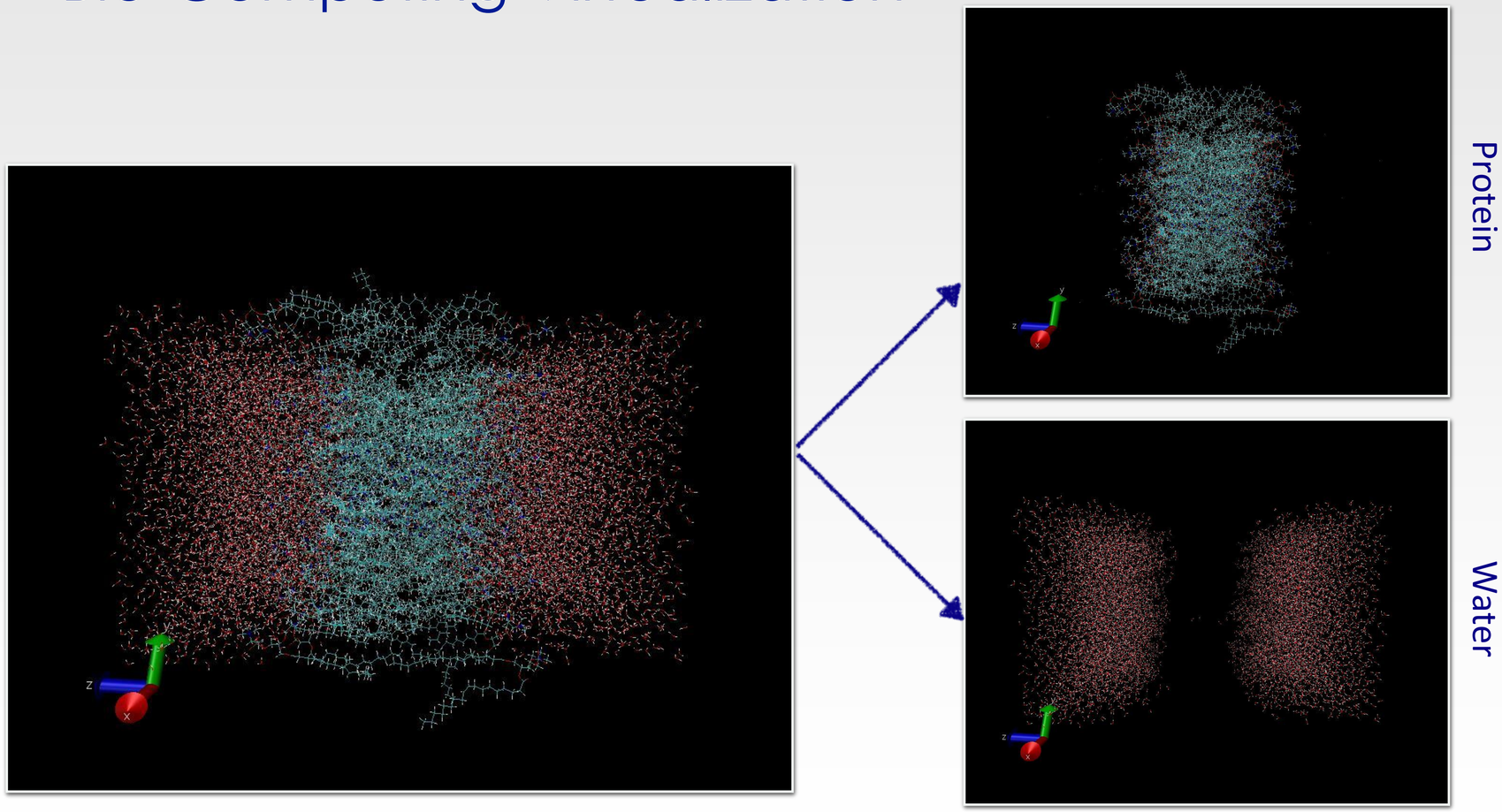
- Duo Parallel Storage Systems
  - NVM
  - HDD pool
  - Independent
- Data Pre-Processing

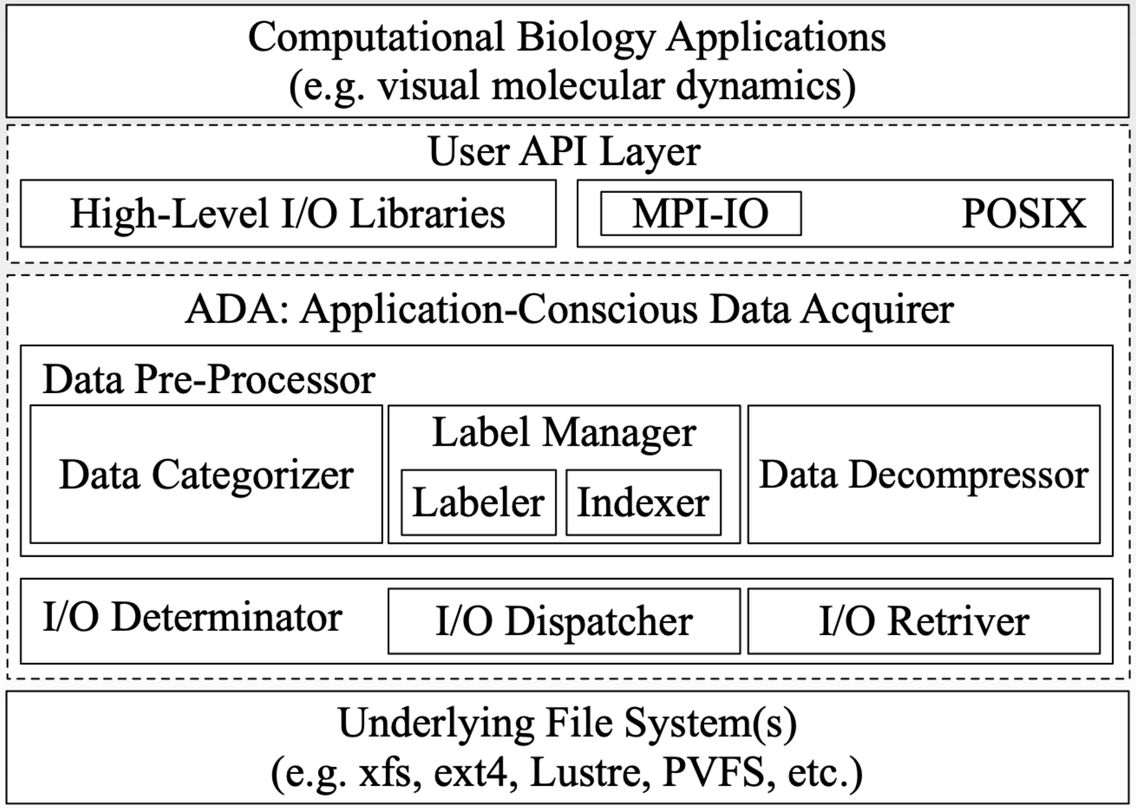


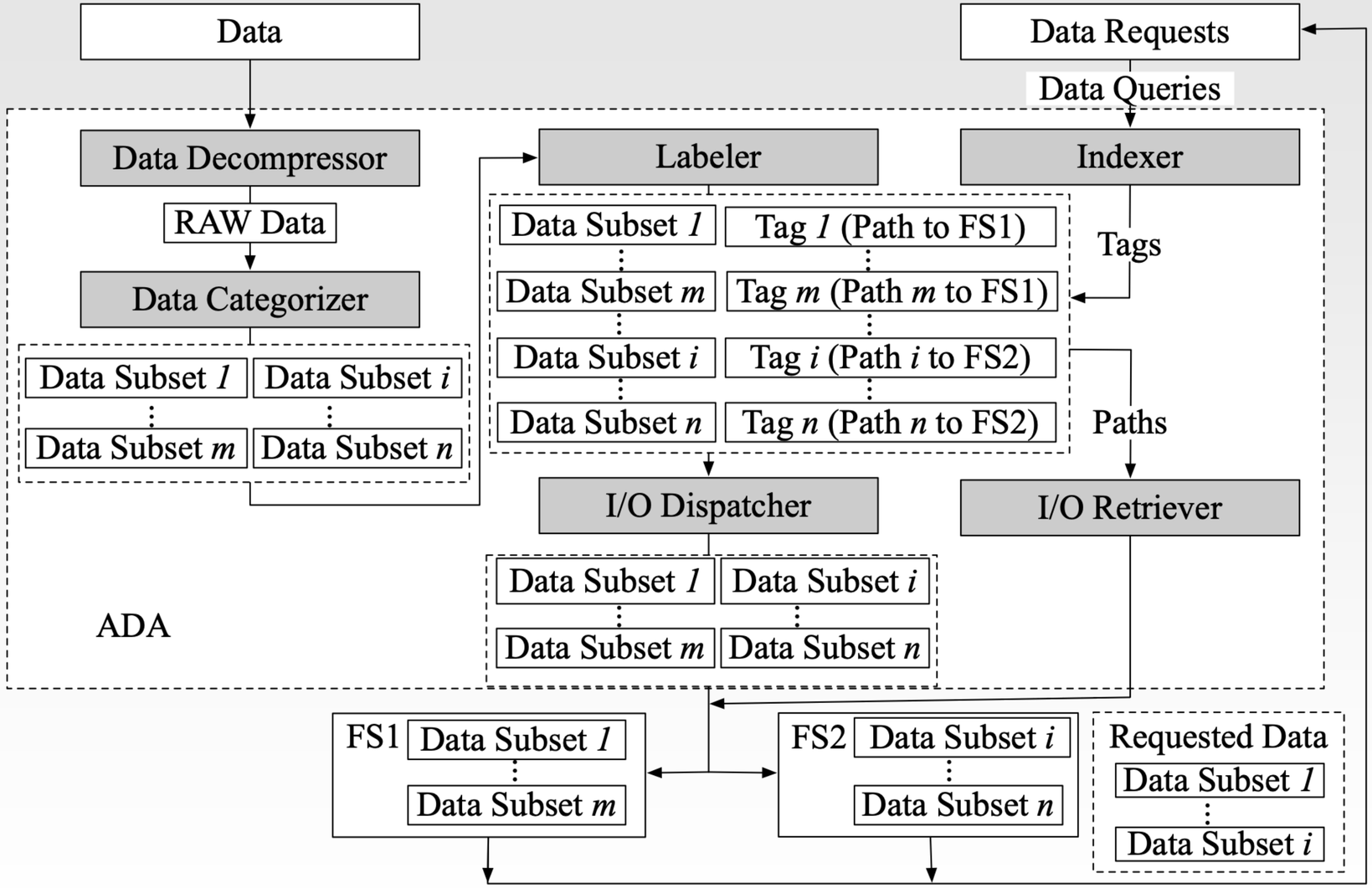

Com Nodes
Rendering
Active Data 1
Active Data m
ADA
ata Pre-Processing
Decompression
. Locating Active Data
PVFS
PVFS
HDDs
HDDs
SSDs
SSDs
SSD
Storage Nodes

# Application Driven

- Bio-Computing Virtualization

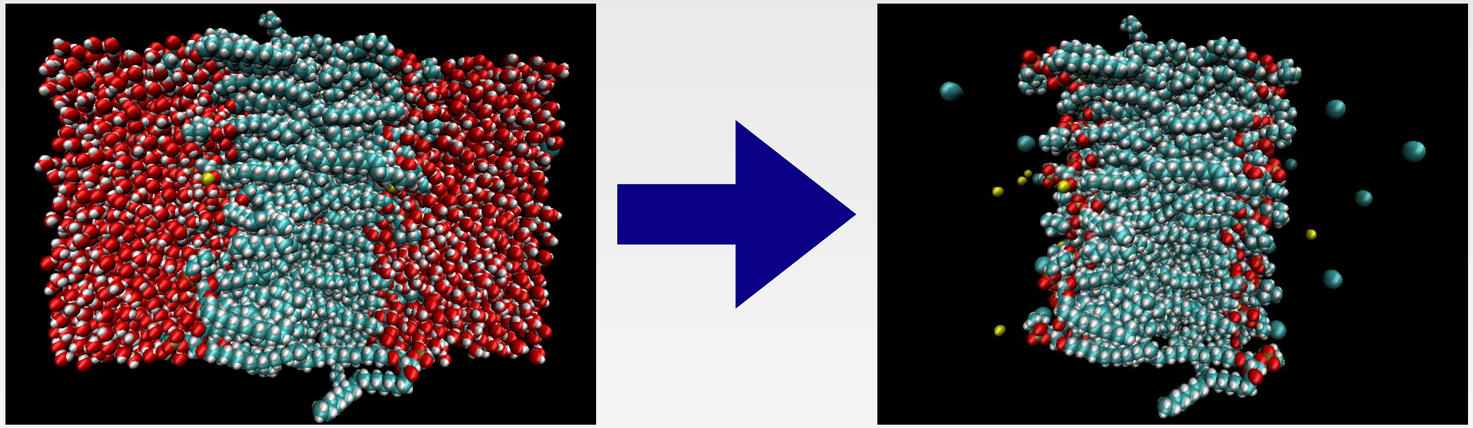


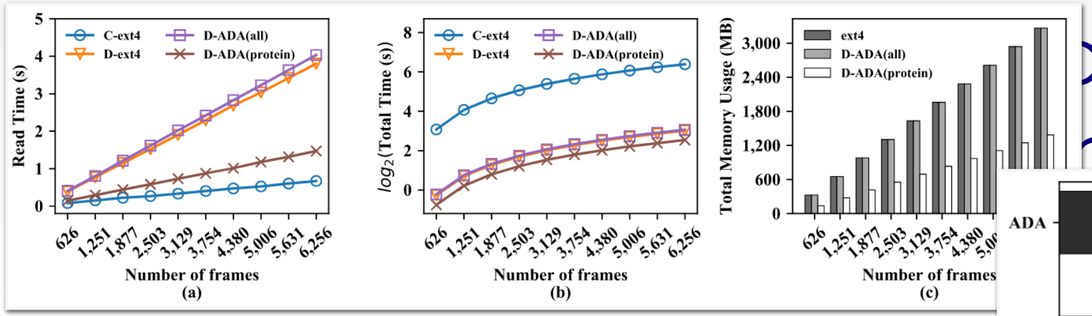

Computational Biology Applications
(e.g. visual molecular dynamics)
User API Layer
High-Level I/O Libraries
MPI-IO
POSIX
ADA: Application-Conscious Data Acquirer
Data Pre-Processor
Data Categorizer
Label Manager
Labeler
Indexer
Data Decompressor
I/O Determinator
I/O Dispatcher
I/O Retriver
Underlying File System(s)
(e.g. xfs, ext4, Lustre, PVFS, etc.)

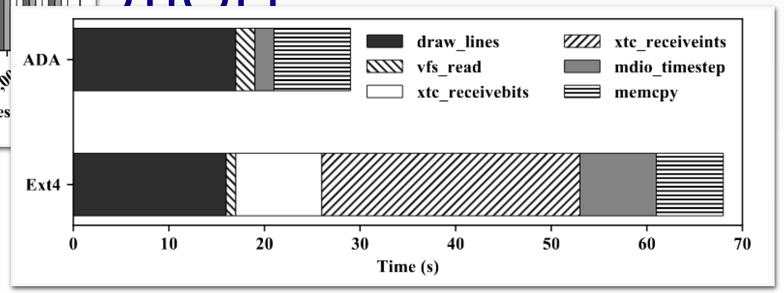

Data
Data Requests
Data Queries
Data Decompressor
Labeler
Indexer
RAW Data
Data Subset 1
Tag 1 (Path to FS1)
Tags
Data Categorizer
Data Subset m
Tag m (Path m to FS1)
Data Subset 1
Data Subset i
Data Subset i
Tag i (Path i to FS2)
Data Subset m
Data Subset n
Data Subset n
Tag n (Path n to FS2)
Paths
I/O Dispatcher
I/O Retriever
Data Subset 1
Data Subset i
ADA
Data Subset m
Data Subset n
FS1
Data Subset 1
FS2
Data Subset i
Requested Data
Data Subset m
Data Subset n
Data Subset 1
Data Subset i

- NVM - Protein Dataset
  - Fast
  - Efficient
    - Energy Consumption
    - I/O Transfer
    - Memory Allocation
- HDD - Water Dataset
  - Large Capacity
  - High Reliability
  - Can be pushed to idle long as needed

- VMD- with iHuman





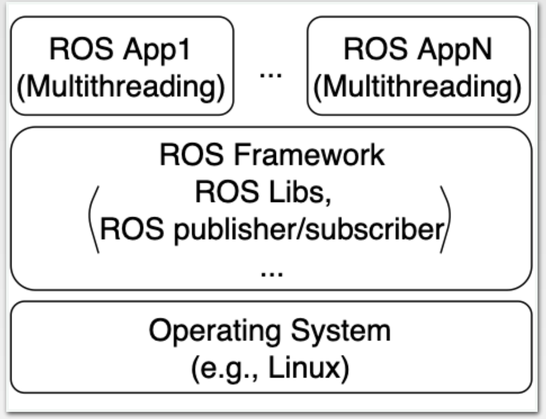



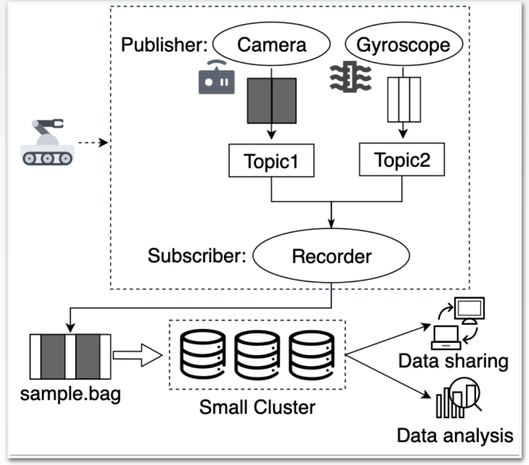
L.1.O.N

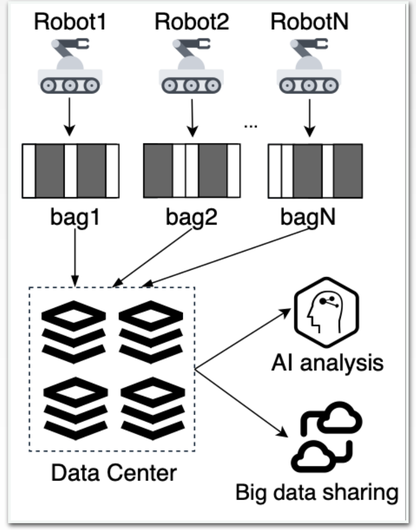
# BORA: A Bag Optimizer for Robotic Analysis

# Background

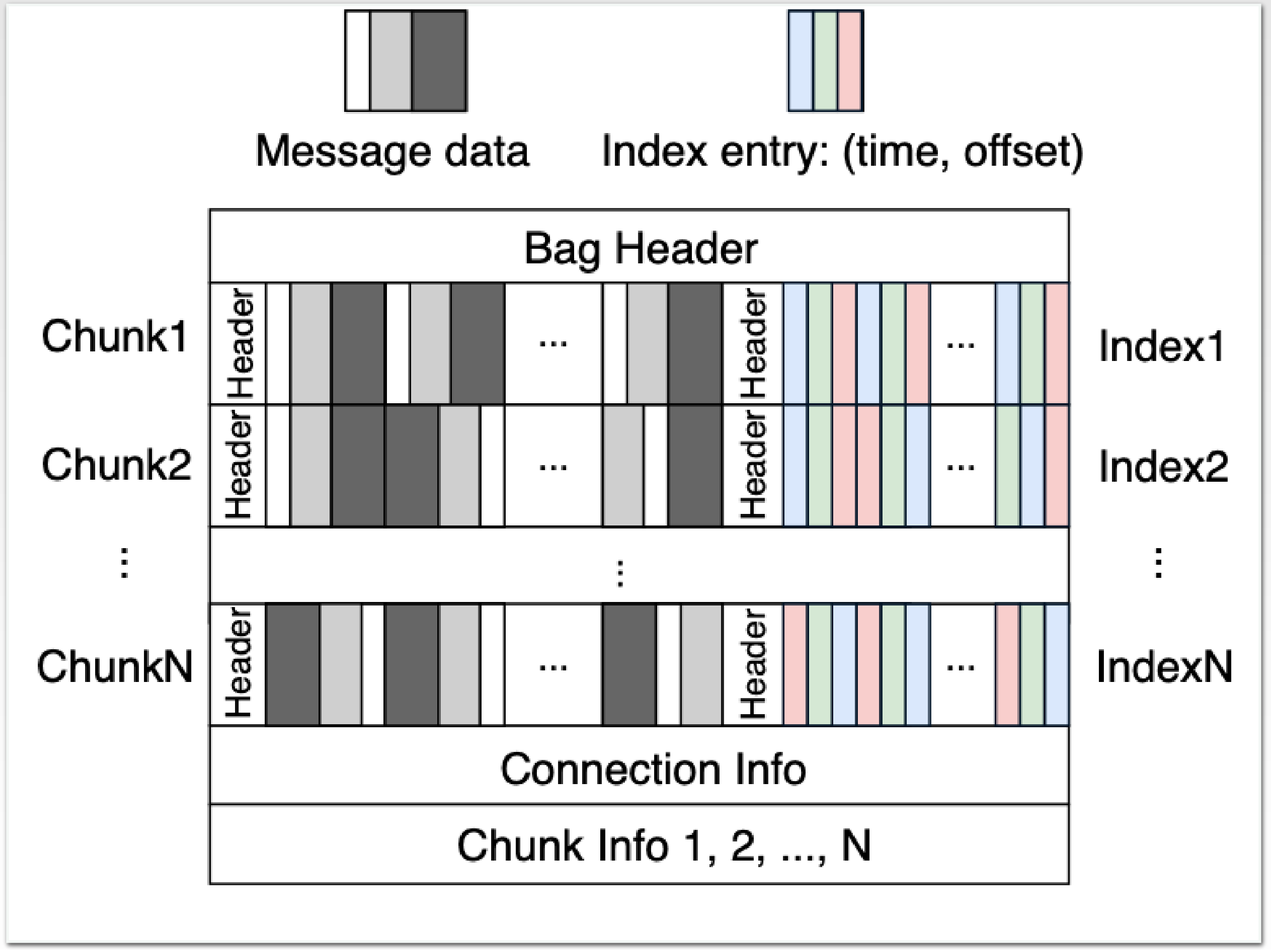
- ROS file format: bag
- `rosbag`: essential bag tool
- Online operation:
  - ROS Compute graph
  - bag replay
- Offline operation
  - Data analysis
  - Rebagging



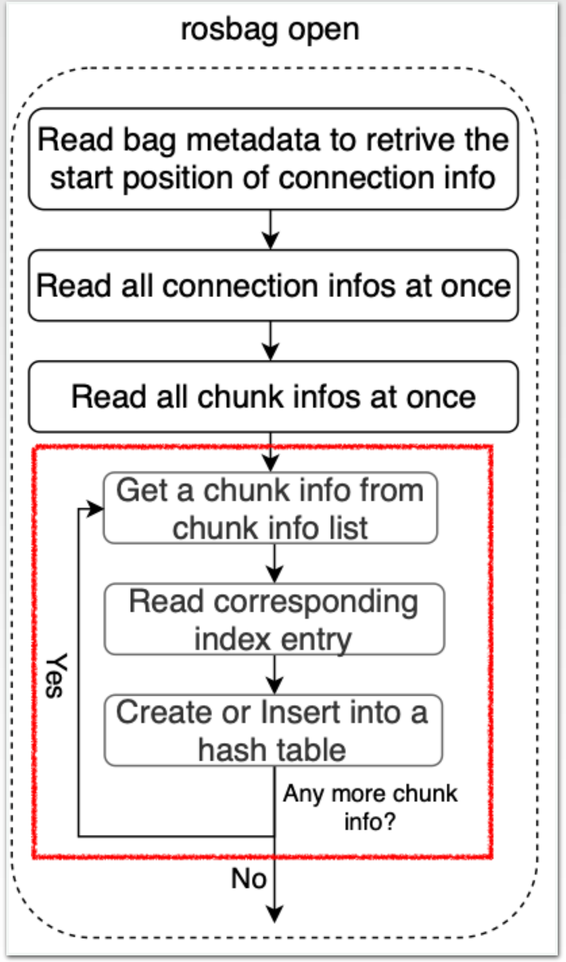
# A bag organization



# Motivation

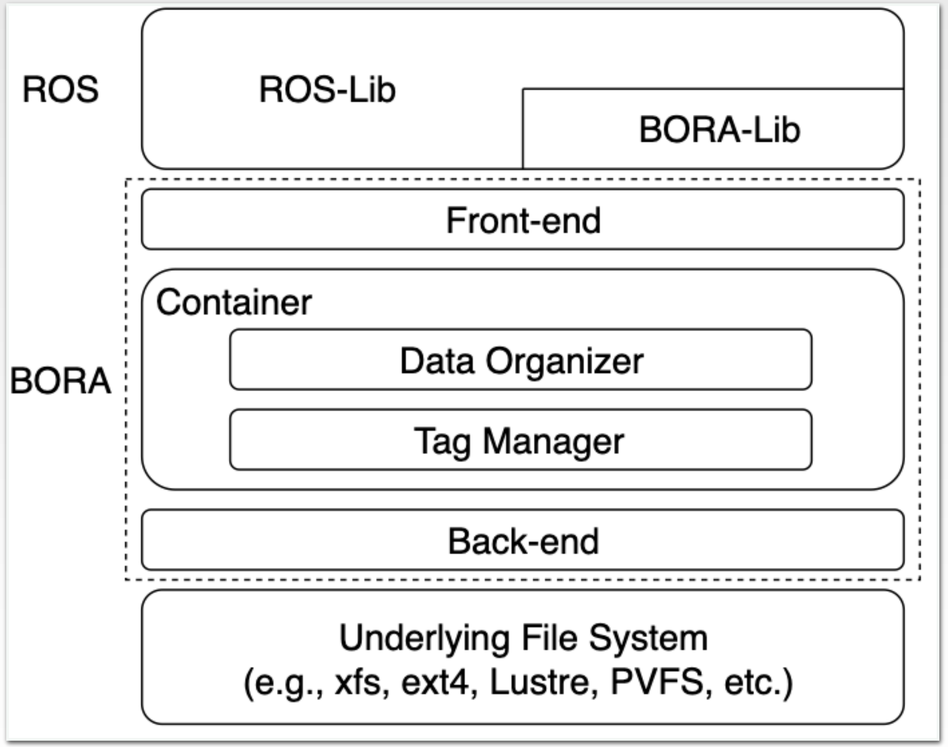
- Pros:
  - Bag can quickly store a large volume of data in a chronological order
  - Bag can support prompt data migration
  - Bag can  store poly-type data (structure& unstructured)
- Cons:
  - Bag data is interleaved, data extraction is not efficient
  - Data index and query is not efficient

# Traditional Rosbag

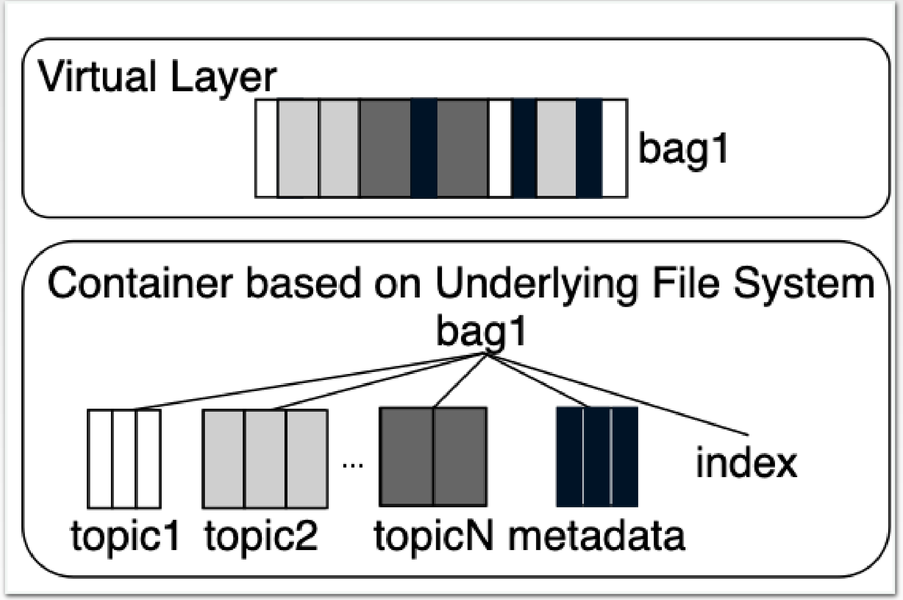


# BORA

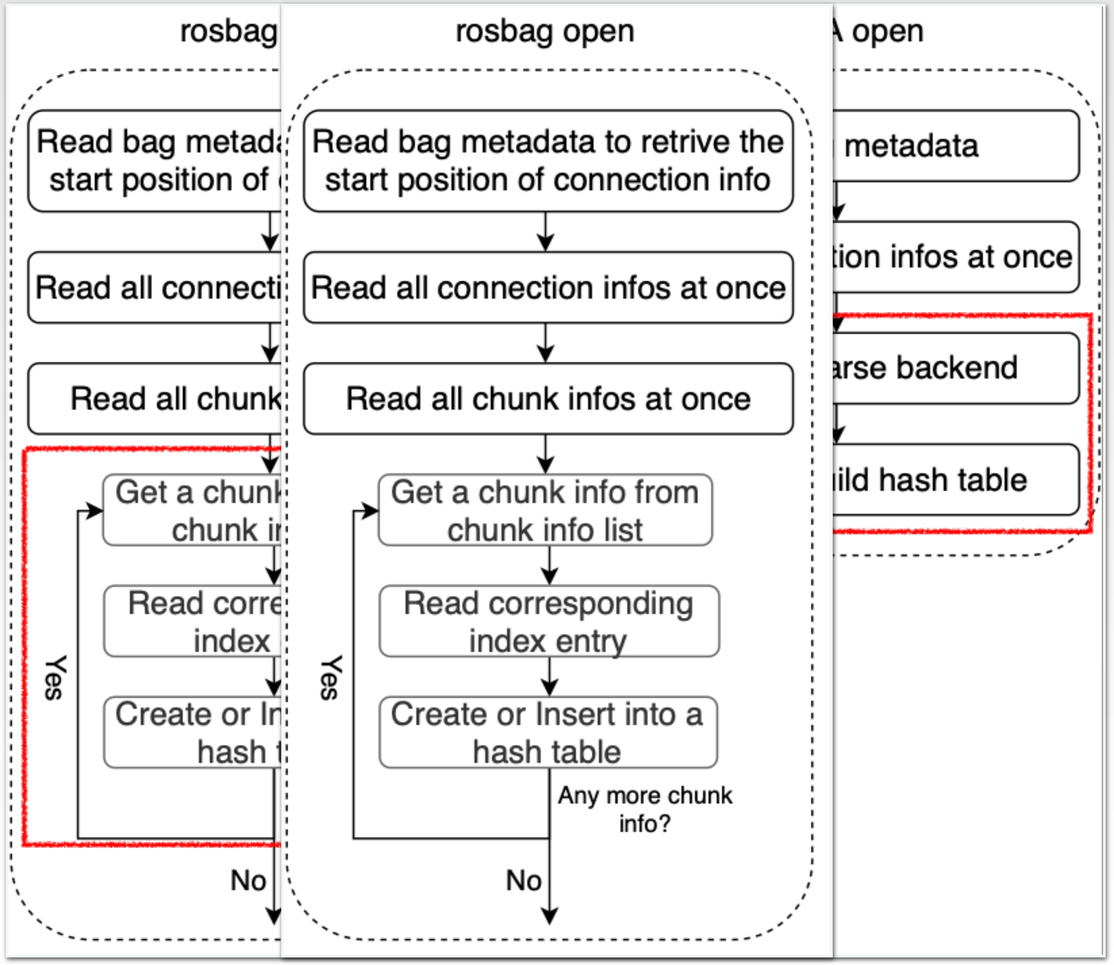
- Bag Organizer for Robotic Analysis



# BORA Container

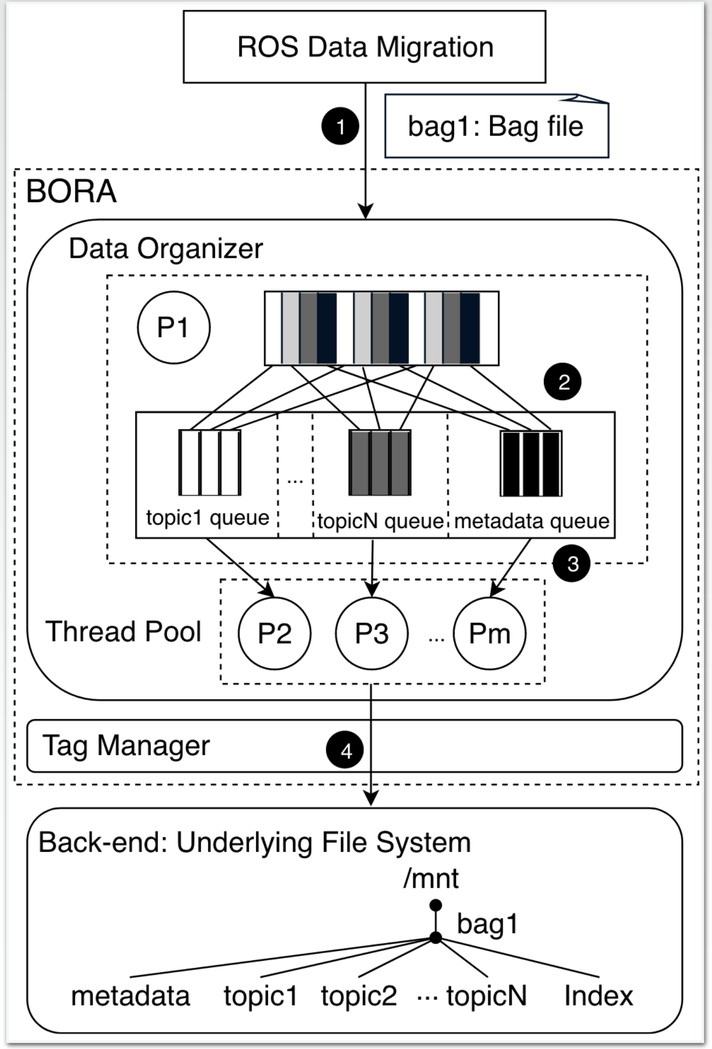


# Traditional Rosbag vs. BORA



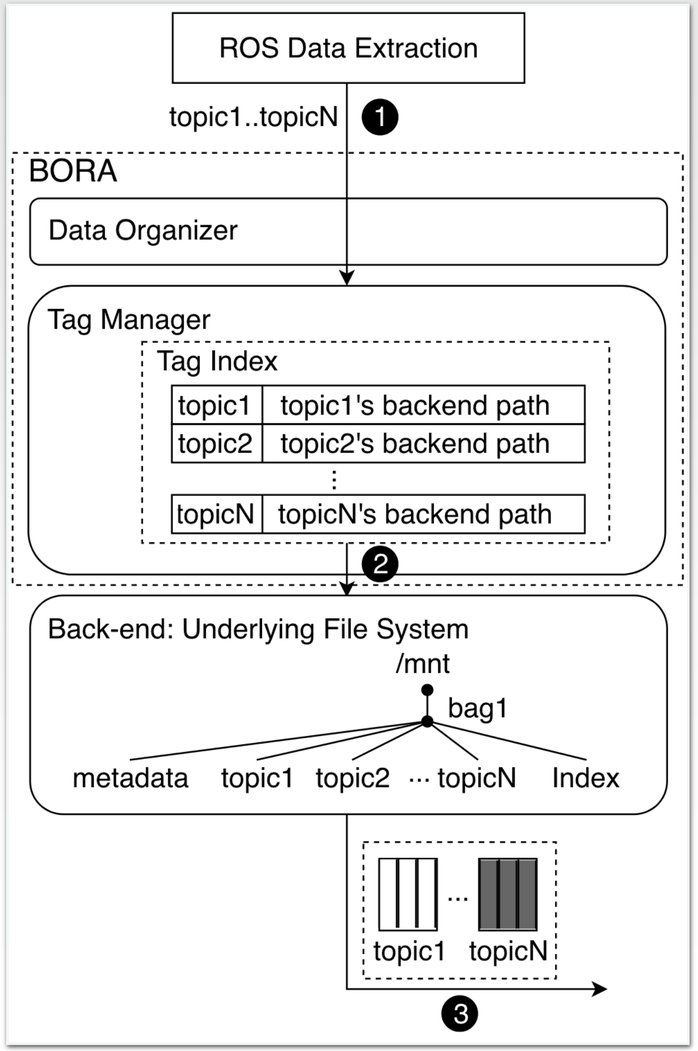
# BORA Data Migration

1. Intercept I/O request `cp`
2. DO scan & divide data to topics
3. DO sends topics to TP
4. TP assigns available thread to write topics to underlying FS



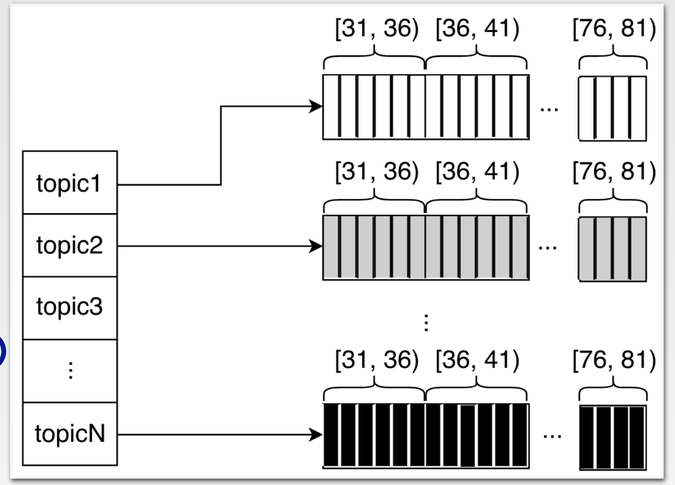
# BORA Data Extraction

1. Intercept query request
   `bag.read_messages (topics)`
2. TM search backend paths from TI by `topic`
3. Underlying FS return topic data to ROS



# BORA Topic-Time Index

- Coarse-Grain Time Index
  1. Divides messages from each topic by T
  2. Calculate $\lfloor T_{start}/T \rfloor$, $\lceil T_{end}/T \rceil$ to target message ranges
  3. Only return the range of messages
- Improve query performance
- Offer a FS-based 2-D index



# BORA Experiment Data

## Data organization of a 2.9 GB bag

| Id | Topic name | Type description | # of Messages | Data size |
|---|---|---|---|---|
| A | /camera/depth/image | Depth Image | 1,429 | 1.64 GB |
| B | /camera/rgb/image_color | RGB Image | 1,431 | 1.23 GB |
| C | /camera/rgb/camera_info | RGB CameraPose Info | 1,432 | 594 KB |
| D | /camera/depth/camera_info | Depth CameraPose Info | 1,430 | 594 KB |
| E | /cortex_marker_array | Primitive Shapes (MarkerArray) | 14,487 | 8.4 MB |
| F | /imu | Inertial Measurement Unit Info (IMU) | 24,367 | 8.4 MB |
| G | /tf | Transform Stamped Message (TF) | 16,411 | 3.6 MB |

## Required Topics in Each Real-world Application

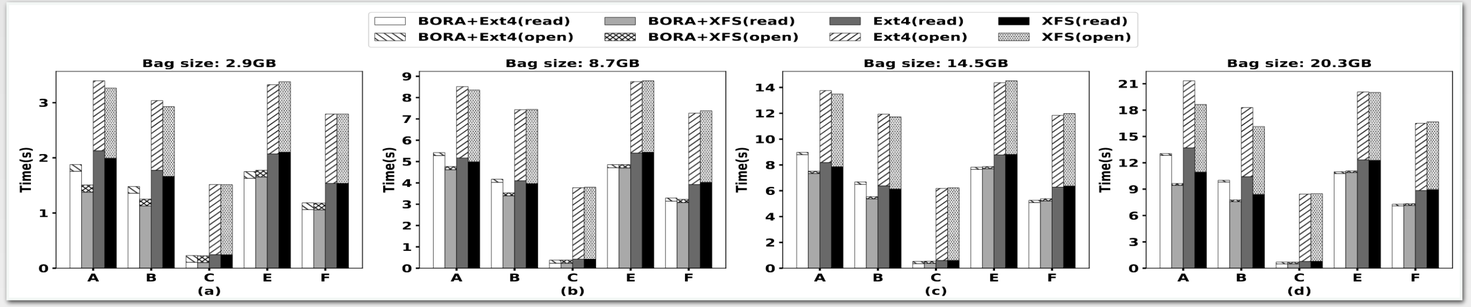
| Application | Required Topics |
|---|---|
| Handheld SLAM (HS) | Depth Image, RGB Image |
| Robot SLAM (RS) | Depth Image, RGB Image, IMU |
| Dynamic Object (DO) | TF, RGB Image<br>CameraPose, MarkerArray |
| Pre-analysis Algorithms(PA) | Randomly Pick |

# BORA Testing Environment

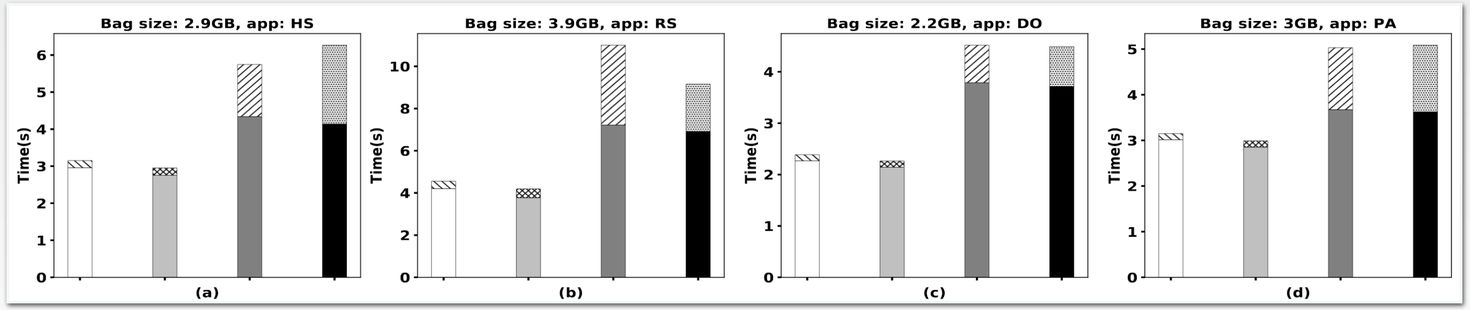
- A Single Server (all SSD)
  - 256GB NVMe SSD*2
- 4-node Cluster (all SSD)
  - 256GB NVMe SSD*8
  - InfiniBand Connection
- A Tianhe-1A Storage Subsystem
  - 12 Compute Nodes, 3 OSSs, 4 MDSs
  - 804TB

# Single Server: query by topics

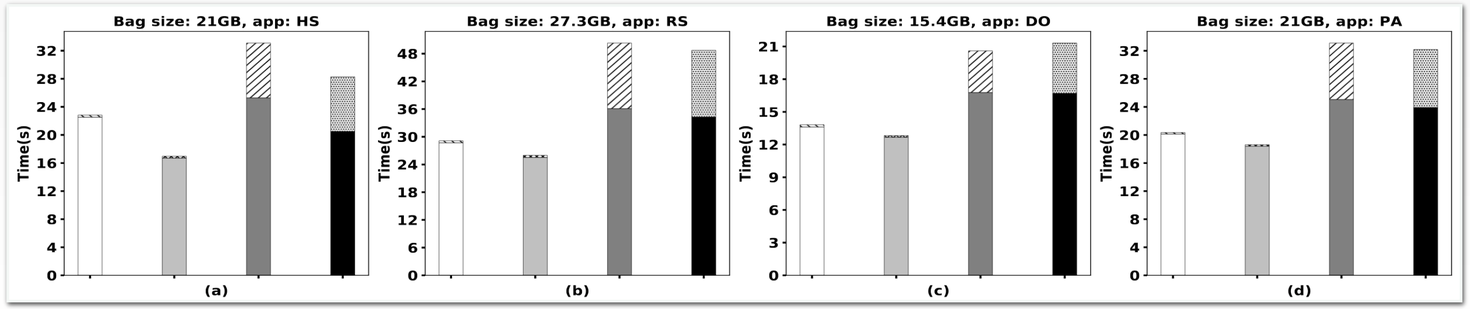
Handheld SLAM (2.9GB bag)




Real-world Apps (small bag)

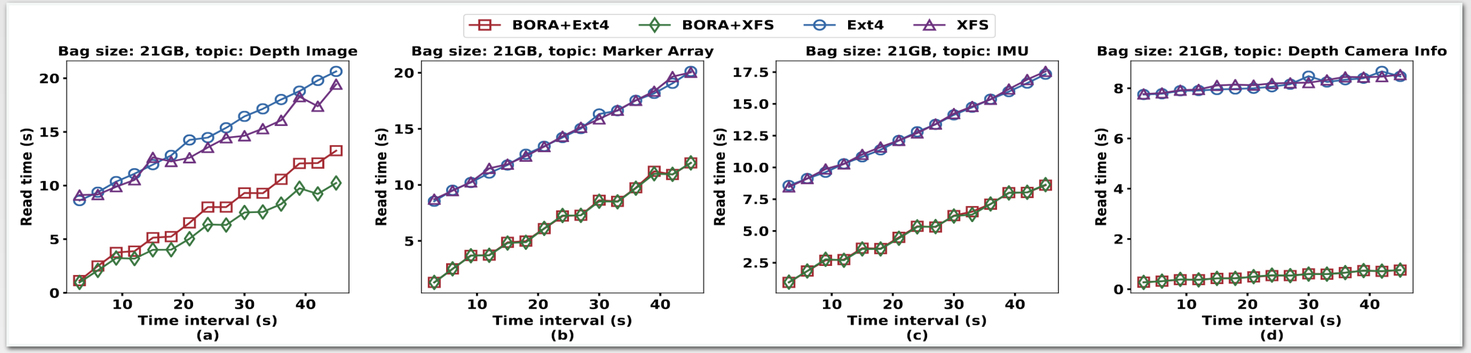



Real-world Apps (large bag)

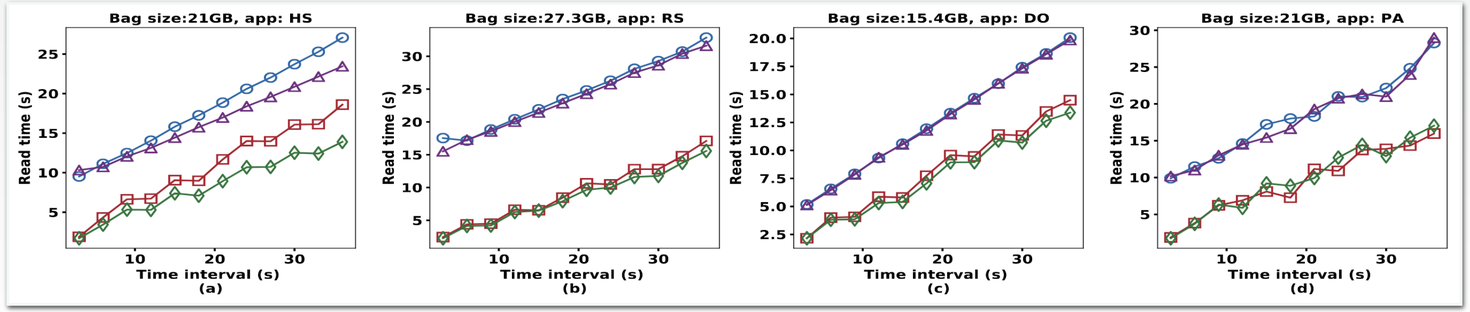


# Single Server: topics + start-stop time

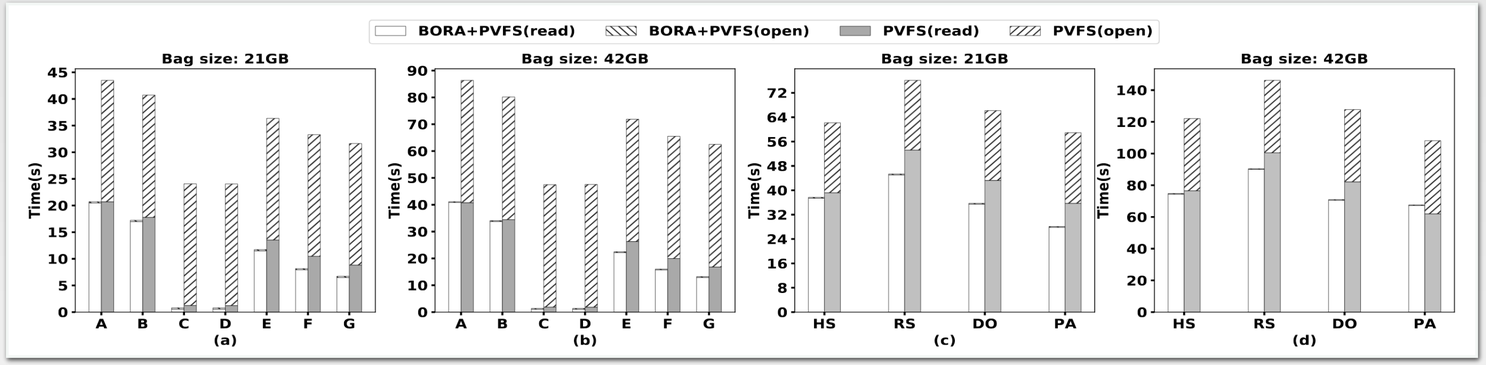
Handheld SLAM (21GB bag)



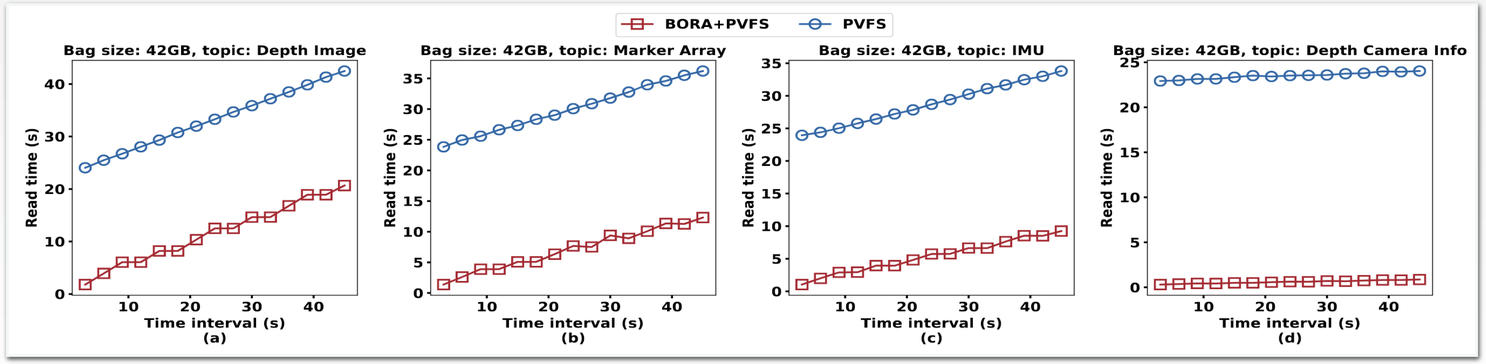

Real-world Apps (large bag)



# 4-node Cluster




(a)(b)Handheld SLAM; (c)(d) real-world apps
(query by topics)

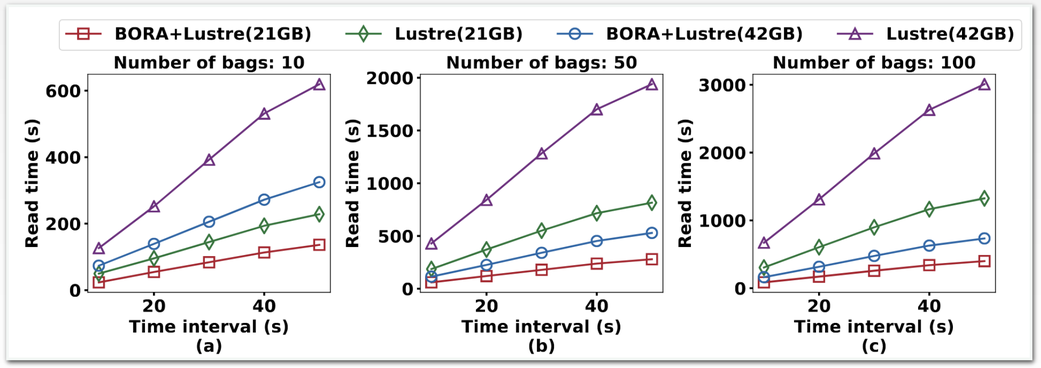



Handheld SLAM
(query by topics and start-stop time)

# Tianhe-1A Storage Subsystem

- Open multiple bags simultaneously
  - Swarm Robot scenario
  - bag quantity: 10, 50, 100
  - bag size : 21GB, 42GB


query by topics




query by topics and start-stop time

# BORA - Results Summary

- Single server
  - 5x-11x improvement
- 4-node server
  - Up to 30x improvement  (small size topics)
- Tianhe-1A Storage
  - Swarm Robot
  - 11x improvement (HDD-based, JBOD)
  - Open operation: 3110x improvement

# Takeaways

- ADA
  - Middleware for VMD
  - Pushes data pre-process on storage
  - 2x performance, 65% less memory
- BORA
  - First FS-based middleware for ROS
  - Improves data query efficiency
  - Provides 2-D indexing for files
  - 5x-11x, 30x, 3110x

L.1.O.N

# Conclusion

- Middleware: Transparent to Both Apps and File Systems
- Application-Driven
- Data Pre-processing
- Computing-Ready Data vs. Entire Data

# Thanks!

# Questions?