Imaging Node

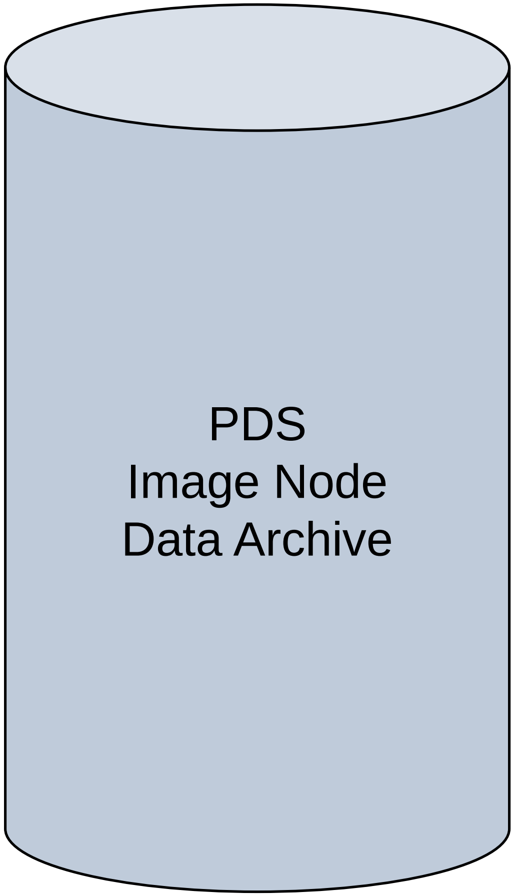
3rd Planetary Data Workshop, 2017

Next Generation Parallelization Systems for Processing and Control of PDS Image Node Assets

Rishi Verma, Jet Propulsion Laboratory, California Institute of Technology

Challenge

Rapid Analysis of PDS Image Node Data Assets



- 650 million files
- 63+ TB (w/o high-res)

Are we sure our data holdings match specifications.. years after publication?

Can we quickly re-generate assets, like metadata or thumbnails, without taking weeks or months to reprocess?

Introducing:

Archive Inventory Management System (AIMS)

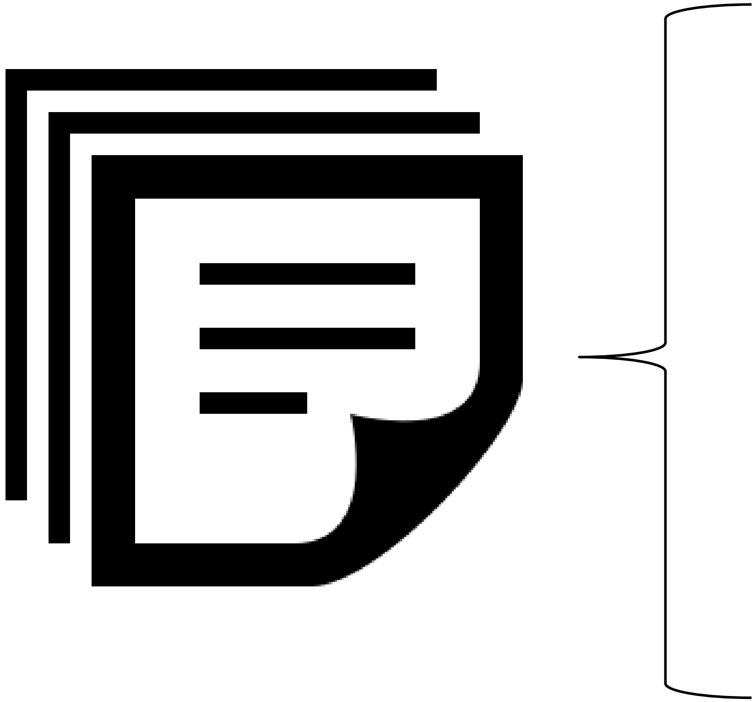
Archive Inventory Management System (AIMS)

Overview

- Objectives:
  1. Validate PDS Image Node Data Assets
     - File size, checksums, permissions, location paths
     - File-system link integrity
     - De-duplication
  2. Offer platform to generate / augment metadata
     - Thumbnail generation
     - Automated image feature detection

Archive Inventory Management System (AIMS)

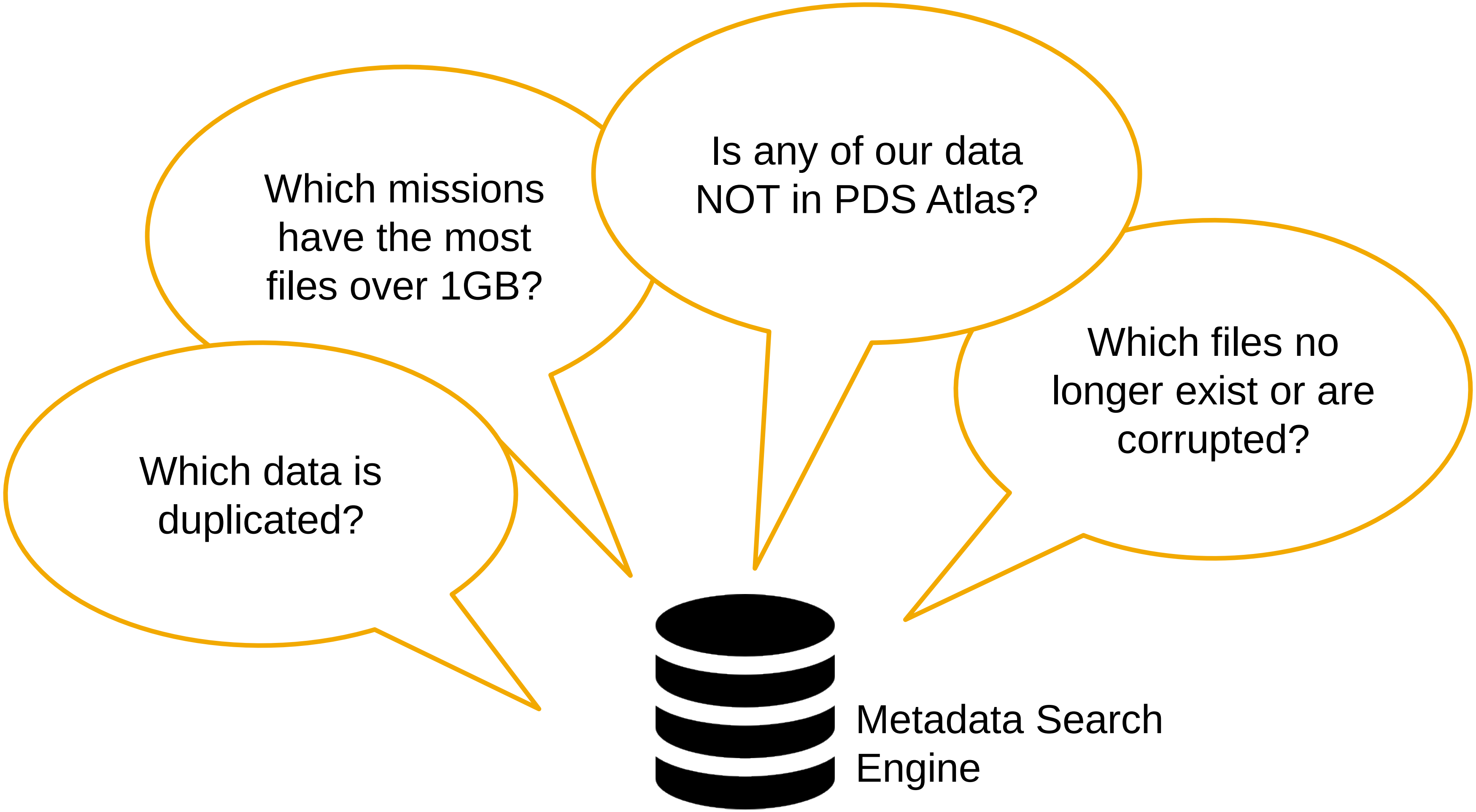
Metadata extraction



- File path
- Size
- MD5
- Mission
- Volume name
- Instrument
- Safed?
- Old volume?
- Staged?
- Extras?
- Text-snippets
- …

Archive Inventory Management System (AIMS)

Data Analytics



Which missions have the most files over 1GB?

Is any of our data NOT in PDS Atlas?

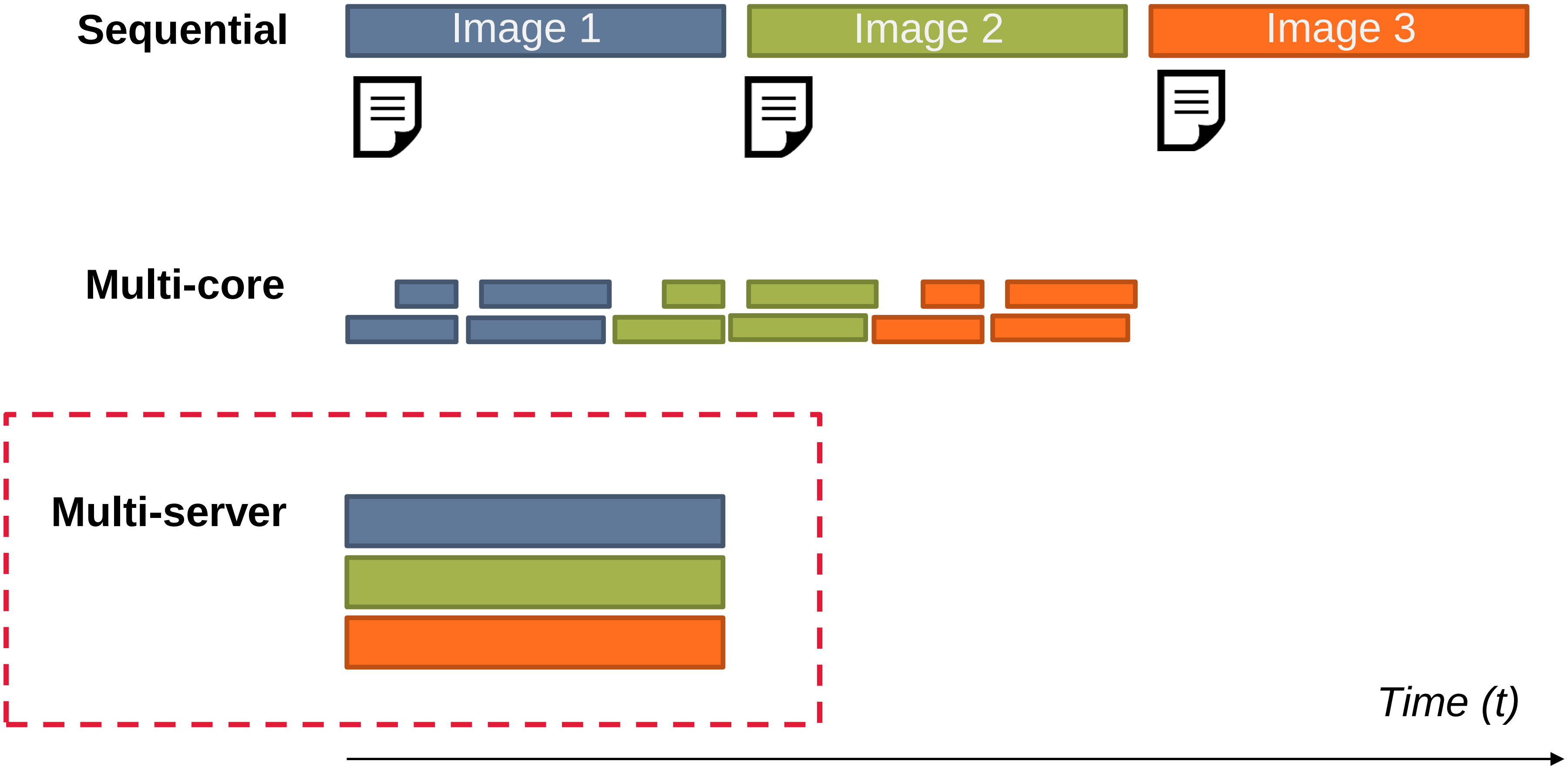
Which data is duplicated?

Which files no longer exist or are corrupted?

Metadata Search Engine

Archive Inventory Management System (AIMS)

Processing strategies



Sequential

Image 1

Image 2

Image 3

Multi-core

Multi-server

Time (t)

Archive Inventory Management System (AIMS)

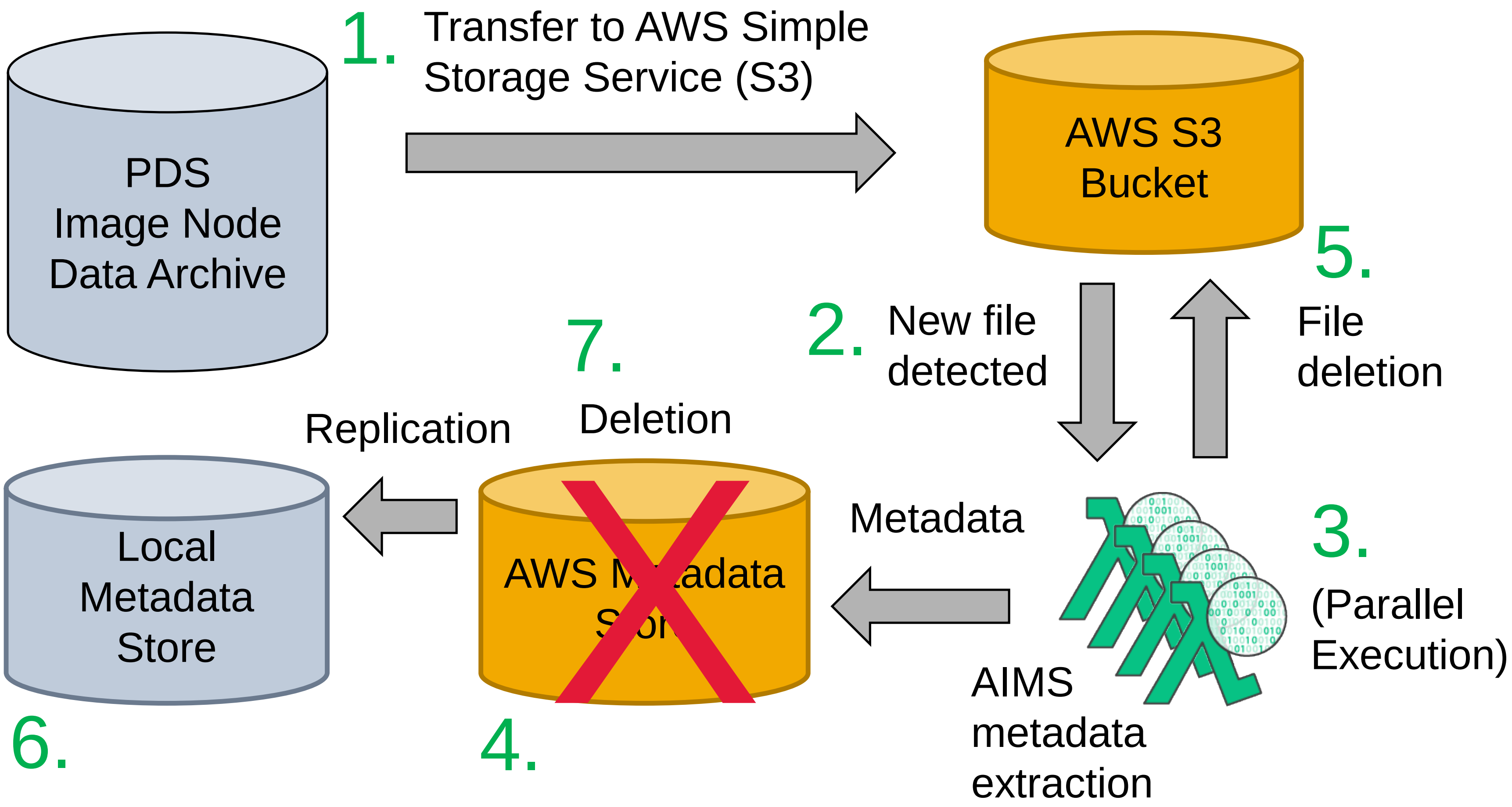
Multi-server process strategy: details

- Rent processors from Amazon Web Services (AWS)
- Continuous scaling, on-demand as file are uploaded
- Billed for execution time only

https://aws.amazon.com/lambda/

Archive Inventory Management System (AIMS)

Multi-server process strategy: architecture



PDS Image Node Data Archive

1. Transfer to AWS Simple Storage Service (S3)

AWS S3 Bucket

2. New file detected

3. (Parallel Execution)

AIMS metadata extraction

Metadata

4. ~~AWS Metadata Store~~

5. File deletion

6. Local Metadata Store

Replication

7. Deletion

Archive Inventory Management System (AIMS)

Multi-server process strategy: pricing

- Free tier
  - 1 million requests free per month
  - 266,667 seconds of processing time free per month (at 1.5 GB RAM)
- Additional costs
  - $0.20 per 1 million requests beyond free tier
  - $0.000002501 per 100 ms (at 1.5 GB memory)

- Rough estimate: ~$150 to reprocess entire archive.
  - Not including egress of metadata

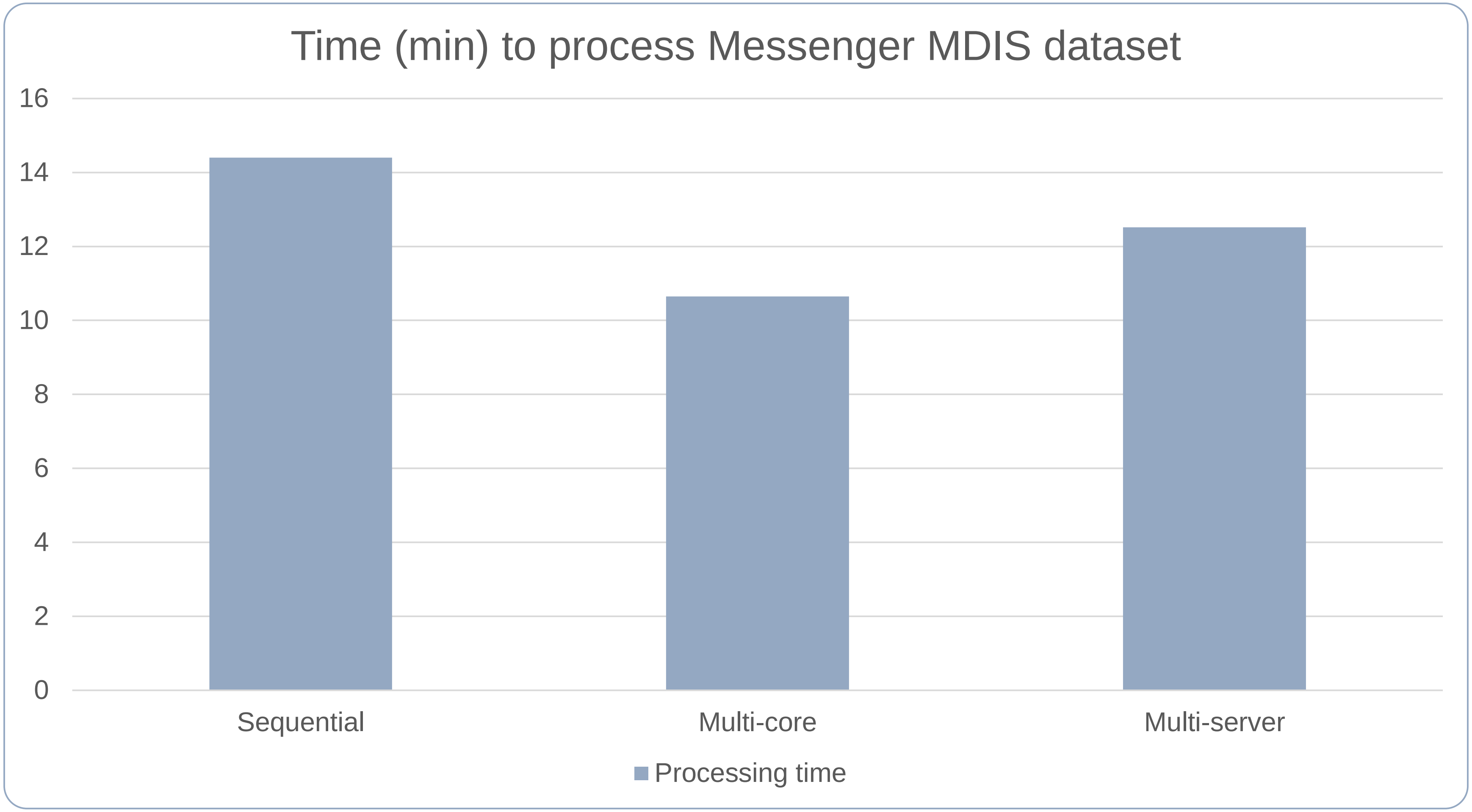
Archive Inventory Management System (AIMS)

Evaluation: experiment setup

- Dataset details:
  - Name: Messenger MDIS (4001)
  - Size: 74 GB
  - Number of files: 345
  - Average file size: 1-2 GB
- Experiment details - for each file:
  - Extract metadata: name, location, size, mission, volume, etc.
  - Generate checksums: MD5

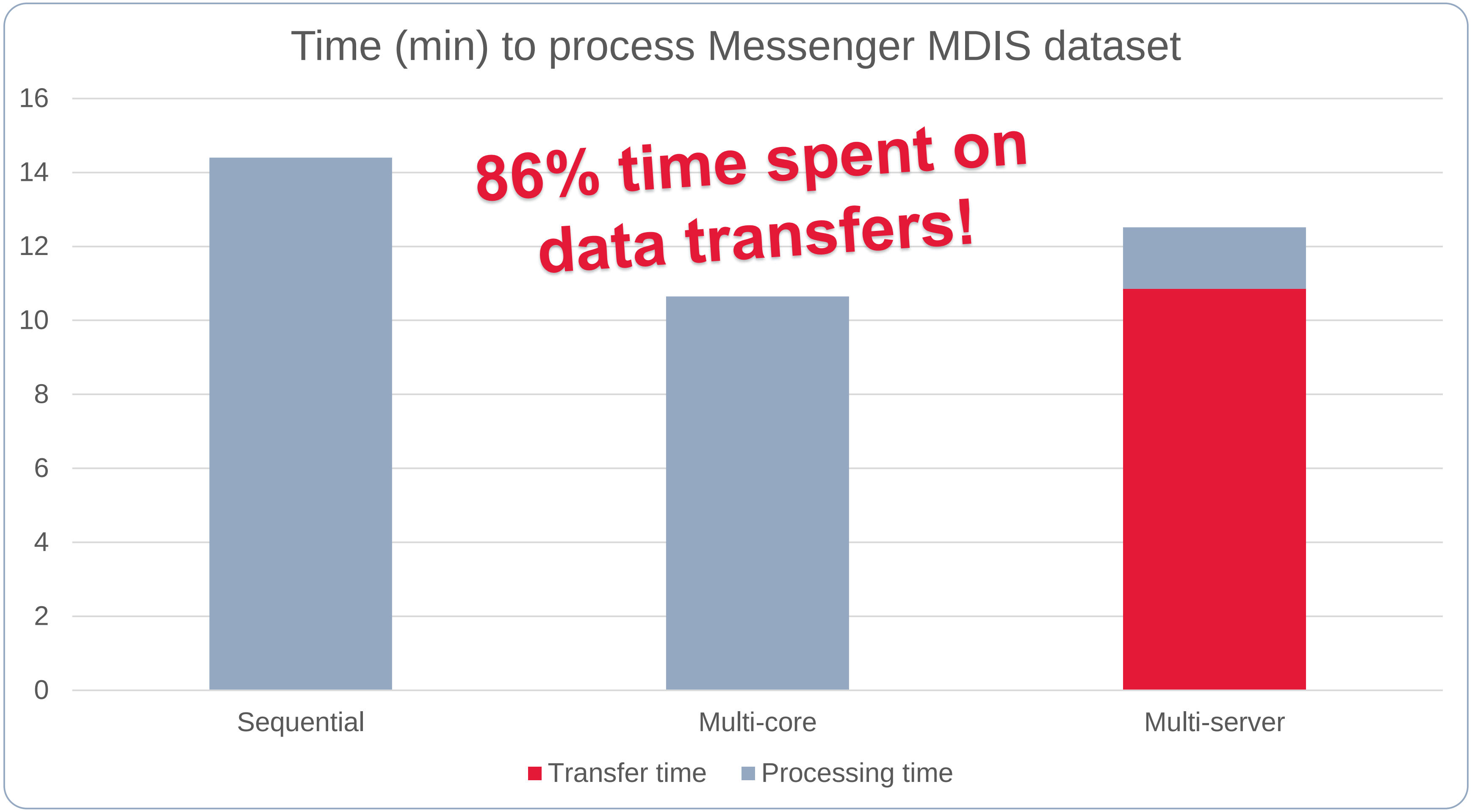
Archive Inventory Management System (AIMS)

Evaluation: experiment results



Archive Inventory Management System (AIMS)

Evaluation: experiment results



86% time spent on data transfers!

Archive Inventory Management System (AIMS)

Next steps

- What was the problem?
  - Even at 100 MBps, terabyte-sized uploads take time
- Can we do better?
  - Parallelized or multi-server uploads show promise
  - Amazon Snowball for an initial data drop
  - OR; just bite the bullet - ETA 7 days for entire archive, which is still faster than alternatives since multi-core does not scale for millions of files.

Questions / Comments / Collaboration?

Contacts:

E-mail: Rishi.Verma@jpl.nasa.gov
Phone: 818-393-5826

E-mail: Jordan.H.Padams@jpl.nasa.gov

Useful links:

- Amazon Lambda: https://aws.amazon.com/lambda/

Jet Propulsion Laboratory
California Institute of Technology