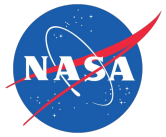


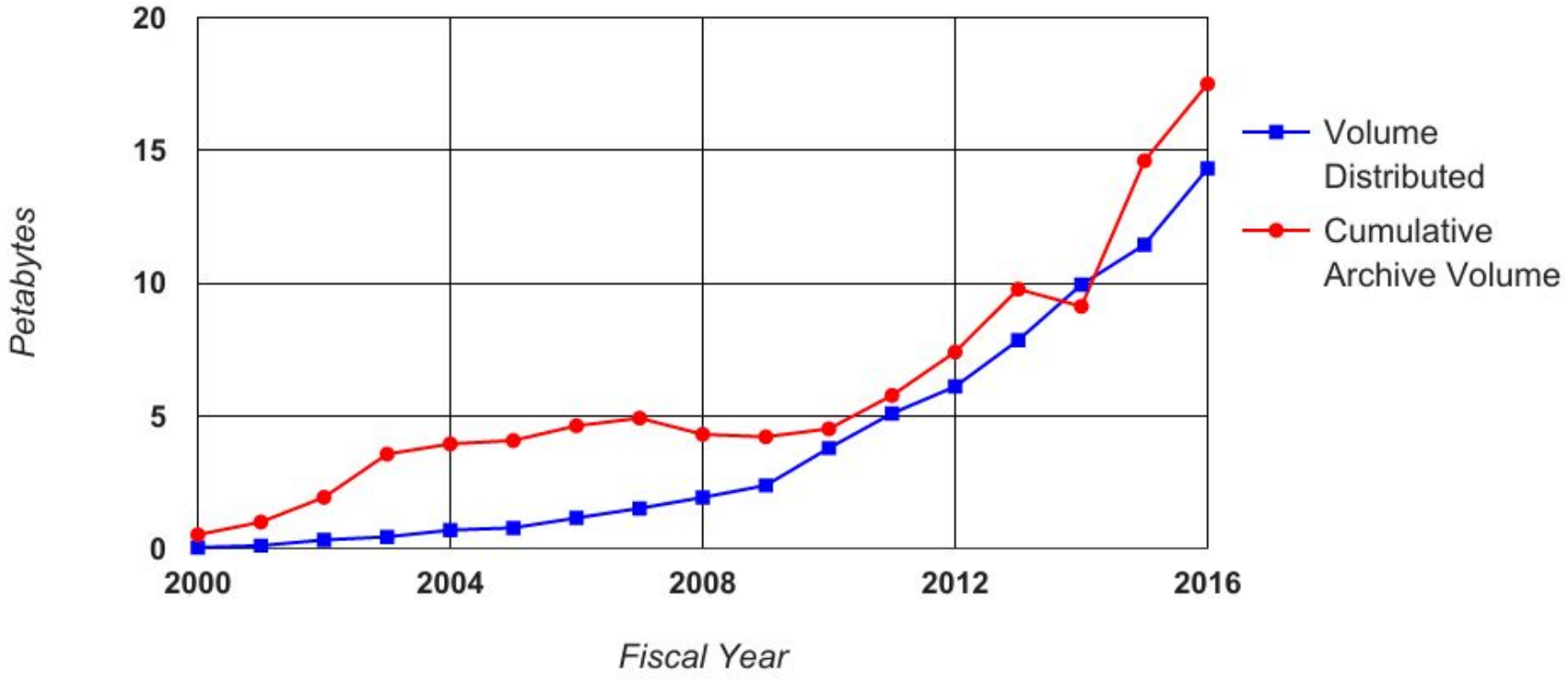
NASA Update for Unidata Stratcomm

Chris Lynnes
EOSDIS System Architect
NASA
chris.lynnes@nasa.gov

Cloud Computing

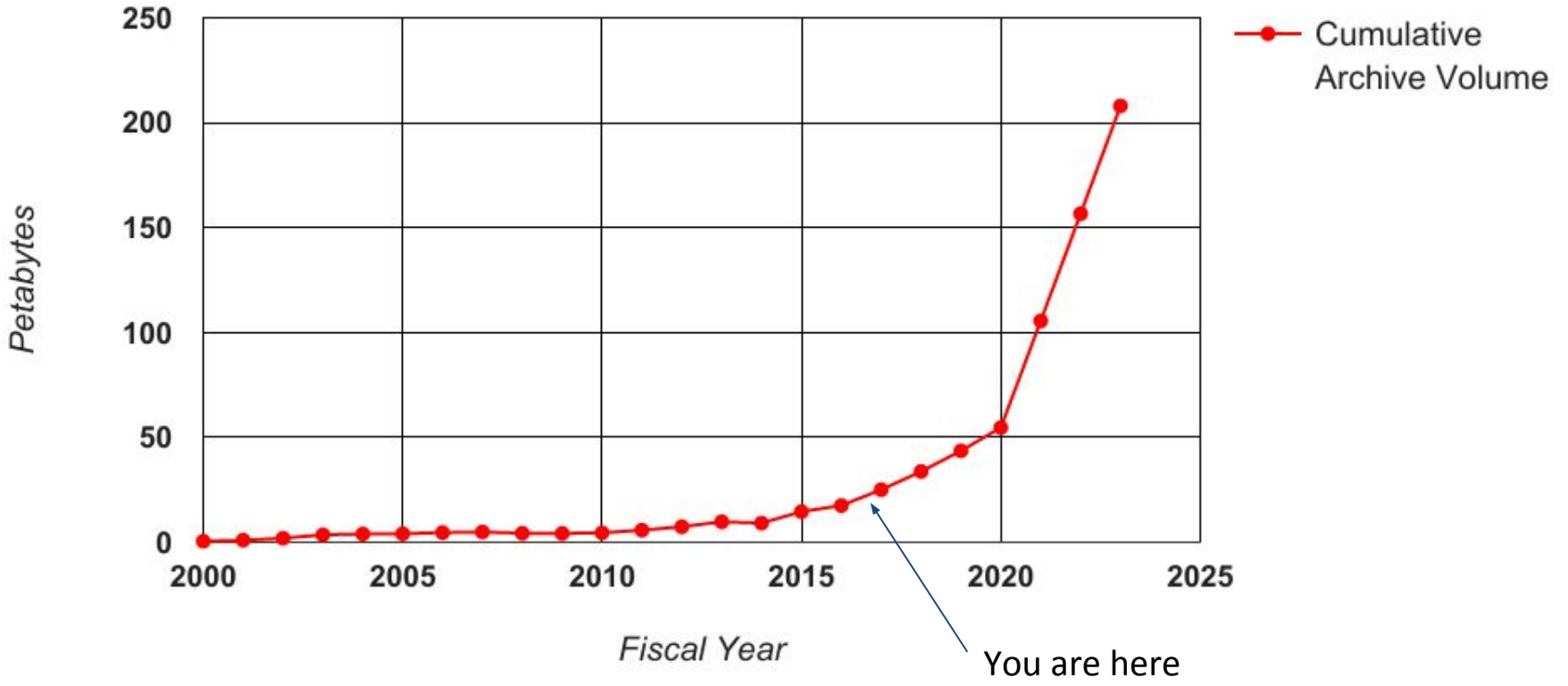
Most of the EOSDIS enterprise has some cloud computing aspect in the works

Annual distribution is on the same order of magnitude as the total archive volume



EOSDIS

And Archive Slated to Grow Substantially...



EOSDIS

How are end users going to handle the volume?

Push analysis computing closer to data

Cloud is the easiest place to do this

Ongoing Archive Prototypes

- Data Ingest + Archive (Cumulus)
  - Experiment with serverless architecture in Amazon Web Services
    - Lambda triggers
    - Step Function workflows
- Data Archive + Production (GRFN)
  - Archive / Production interface
  - On-demand production
- Web Object Storage edge server
- OPeNDAP on Web Object Storage Study*

OPeNDAP - Web Object Storage Trade Study

1. Baseline Hyrax Data Access:
   - Fetch file from S3 to Elastic Block Storage and serve
2. Store files as objects and subset with HTTP range-gets
   - Relies on external index map of chunks in HDF5 file
   - Developed for long term preservation of HDF4 data
3. Store HDF5 Datasets as objects
   - In theory, we could store less-used variables on colder storage

Q: Which is best?
A: If accessing < 20% of the file at once: option 2
Else: option 1

Technical details at https://github.com/OPENDAP/cloudydap/wiki

Cloud Analytics Prototypes

- NEXUS (JPL)
  - Storage: tiles in Cassandra DB
  - Compute: Spark
- Giovanni (GSFC)
  - Storage: netCDF in S3
  - Compute: EC2, GPUs
- Climate Analytics as a Service (GSFC)
  - Storage: HDFS
  - Compute: MapReduce
- Data Containers Study: optimizing cloud storage of data for analytics (poster)
  - NEXUS
  - ClimateSpark: HDFS with spatio-temporal indexes of netCDF files
  - MongoDB

Cloud Analytics Prototypes (cont.)

- "Data Cubes" - from Committee on Earth Observing Satellites (NASA implementation)
  - Momentum growing via WGISS
- Cloud Analytics Toolkit to Enhance Earth Sciences
  - Jupyter notebooks showing how to get, access, analyze, and cloud
  - Trying to find the best way to deliver/deploy: conda | docker | AWS AMI | ?

Interoperability and Usability

- ESDIS is deprecating HDF4
  - Still paying to fix critical bugs affecting EOSDIS datasets
  - No new enhancements
- Dataset Interoperability Recommendations for Earth Science approved by EOSDIS Standards Office
  - The devil in the details of HDF5/netCDF4
- Outreach to data providers
  - 1 hour session on data product design best practices at workshop of data producers in May
  - Workshops at ESIP
  - Data Product Design How-To?

Other Developments

- Asked by HQ to Open-Source new software whenever possible
- Therefore:
  - Common Metadata Repository (catalog and search engine)
  - Earthdata Search Client (search client for CMR)
  - NEXUS cloud analytics (see ESIP workshops)
  - More to come...
- Working on making data more GIS-friendly
  - E.g., recipes for ingesting netCDF into ArcGIS